A Section 508–conformant HTML version of this article
is available at https://doi.org/10.1289/EHP13215.

# Standardizing Extracted Data Using Automated Application of Controlled Vocabularies

*Caroline Foster,[1] Jessica Wignall,[1] Samuel Kovach,[1] Neepa Choksi,[2] Dave Allen,[2] Joanne Trgovcich,[1] Johanna R. Rochester,[1] Patricia Ceger,[2] Amber Daniel,[2] Jon Hamm,[2] Jim Truax,[2] Bevin Blake,[3] Barry McIntyre,[3] Vicki Sutherland,[3] Matthew D. Stout,[3] and Nicole Kleinstreuer[4]*

[1]ICF, Durham, North Carolina, USA
[2]ILS, Research Triangle Park, North Carolina, USA
[3]Division of Translational Toxicology (DTT), NIEHS, NIH, Research Triangle Park, North Carolina, USA
[4]National Toxicology Program Interagency Center for the Evaluation of Alternative Toxicological Methods (NICEATM), DTT, NIEHS, NIH, Research Triangle Park, North Carolina, USA

**Background:** Extraction of toxicological end points from primary sources is a central component of systematic reviews and human health risk assessments. To ensure optimal use of these data, consistent language should be used for end point descriptions. However, primary source language describing treatment-related end points can vary greatly, resulting in large labor efforts to manually standardize extractions before data are fit for use.

**Objectives:** To minimize these labor efforts, we applied an augmented intelligence approach and developed automated tools to support standardization of extracted information via application of preexisting controlled vocabularies.

**Methods:** We created and applied a harmonized controlled vocabulary crosswalk, consisting of Unified Medical Language System (UMLS) codes, German Federal Institute for Risk Assessment (BfR) DevTox harmonized terms, and The Organization for Economic Co-operation and Development (OECD) end point vocabularies, to roughly 34,000 extractions from prenatal developmental toxicology studies conducted by the National Toxicology Program (NTP) and 6,400 extractions from European Chemicals Agency (ECHA) prenatal developmental toxicology studies, all recorded based on the original study report language.

**Results:** We automatically applied standardized controlled vocabulary terms to 75% of the NTP extracted end points and 57% of the ECHA extracted end points. Of all the standardized extracted end points, about half (51%) required manual review for potential extraneous matches or inaccuracies. Extracted end points that were not mapped to standardized terms tended to be too general or required human logic to find a good match. We estimate that this augmented intelligence approach saved >350 hours of manual effort and yielded valuable resources including a controlled vocabulary crosswalk, organized related terms lists, code for implementing an automated mapping workflow, and a computationally accessible dataset.

**Discussion:** Augmenting manual efforts with automation tools increased the efficiency of producing a findable, accessible, interoperable, and reusable (FAIR) dataset of regulatory guideline studies. This open-source approach can be readily applied to other legacy developmental toxicology datasets, and the code design is customizable for other study types. https://doi.org/10.1289/EHP13215

## Introduction

Extraction of toxicological end points from primary sources, including legacy historical datasets, is a central component of systematic review and human health risk assessments.[1] Augmented intelligence approaches can leverage computational resources to enhance and support human intellect in more efficiently and effectively curating and annotating high-quality *in vivo* toxicology studies. Developmental and reproductive toxicity (DART) studies comprise a critical component of hazard and risk assessments aimed at determining whether a chemical poses a risk to the developing fetus, maternal health, or the reproductive system.[2] Animal studies examining DART outcomes have been conducted for decades, many of which are stored in legacy reports and other forms of electronic documents, such as results from National Toxicology Program (NTP) teratology studies, housed within the publicly accessible Chemical Effects in Biologic Systems (CEBS) database.[3] Such historical datasets, especially studies conducted according to specific testing guidelines, are invaluable in contemporary research endeavors, particularly to the computational and predictive toxicology scientific communities. However, extracting and annotating the data in an efficient and standardized manner remains a barrier to harmonizing toxicology data for use in predictive frameworks.

DART study data are particularly crucial for informing on potential hazards and in both predictive research and regulatory contexts, as these studies encompass numerous critical windows of exposure and sensitive periods of development as well as potential impacts on maternal health and reproductive function.[4] The maternal-fetal unit is considered uniquely vulnerable to environmental insults, and exposures during development may adversely impact perinatal and latent health outcomes for both the mother and offspring. Further, offspring exposed to stressors during sensitive periods of development are predisposed to heightened risk for adverse health outcomes later in life.[5] A large body of epidemiological studies support associations between developmental chemical exposures and later life diseases, particularly concerning aspects of neurodevelopment (e.g., prenatal exposures to organophosphate pesticides and altered IQ and cognitive development in children) and impacts from exposure to endocrine disrupting compounds (EDCs) (e.g., exposure to EDCs leading to higher incidence of cardiometabolic diseases, cancers, and reproductive dysfunction).[6,7] Animal testing data from DART studies are considered relevant for understanding potential developmental and other health impacts on humans because overt teratological, reproductive, or systemic effects at higher doses may be sentinels for subtle changes at lower, more human-relevant exposures.[8,9]

Address correspondence to Nicole Kleinstreuer, 530 Davis Dr., Morrisville, NC 27560 USA. Telephone: (984) 287-3150. Email: Nicole.kleinstreuer@nih.gov

Supplemental Material is available online (https://doi.org/10.1289/EHP13215).

The authors declare they have no actual or potential financial conflicts of interest.

Conclusions and opinions are those of the individual authors and do not necessarily reflect the policies or views of EHP Publishing or the National Institute of Environmental Health Sciences.

Received 24 April 2023; Revised 5 January 2024; Accepted 8 January 2024; Published 13 February 2024.

**Note to readers with disabilities:** *EHP* strives to ensure that all journal content is accessible to all readers. However, some figures and Supplemental Material published in *EHP* articles may not conform to 508 standards due to the complexity of the information being presented. If you need assistance accessing journal content, please contact ehpsubmissions@niehs.nih.gov. Our staff will work with you to assess and meet your accessibility needs within 3 working days.

The mammalian reproductive cycle and embryo-fetal development is vulnerable to chemical exposure at multiple stages, and DART studies provide data needed to assess myriad outcomes that include:

- Maternal effects: litter loss/resorptions/fetal death, systemic toxicity (e.g., elevated liver enzymes), postpartum health effects (e.g., cardiovascular disease), lactational impairment (e.g., impacts offspring postnatal development)
- Developmental effects: short-term/immediate health outcomes for offspring: teratogenicity, death, preterm, low birth weight; latent health effects: altered timing of puberty, neurobehavioral outcomes, adult health outcomes (e.g., diabetes, cardiovascular disease, metabolic syndrome)
- Reproductive effects: fecundity/fertility/sterility resulting from *in utero* or postnatal exposures
- Transgenerational effects: developing offspring (F1) impacted by exposure that leads to alterations in biology carried on to F2 (e.g., F1 female *in utero* exposures that effect developing F2 oocytes)

As such, DART study data have the potential to inform policy decisions that have far-reaching public health impact, and the harmonization of these data would provide a strong anchor for understanding variability of traditional mammalian toxicology tests and supporting validation of alternative methods and computational/predictive modeling. Existing efforts to establish scientific confidence in alternatives such as embryonic stem cell-based assays,[10,11] small model organisms,[12] and pathway-based integrated strategies leveraging high-throughput screening data[13] rely heavily upon highly curated reference data that ideally cover a range of chemical exposures and properties.[10,14,15] In the context of risk assessment, using *in vivo* DART studies conducted according to testing guidelines to calculate relevant dose-response parameters (such as reference doses or benchmark doses) is considered the most protective approach by regulatory authorities such as the US Environmental Protection Agency (EPA).[16] This concept is especially pertinent when applied to computational methods that use datasets compiled from methodologically consistent studies to predict developmental toxicity of under- or unstudied chemicals.[17] Consistent language for end point descriptions is required for extracted data to be comparable across studies and to provide robust reference datasets to perform analyses including identification of chemical- or species-specific effects, validation of new approaches, and computational/predictive modeling, but this is often lacking even across studies conducted according to the same methodology.

Thus, standardized extraction of DART study data from legacy reports summarizing studies conducted according to regulatory guidelines or other "high-quality" studies (i.e., confirmed by experts to comply with relevant guideline criteria) into databases using unified end point terminology is a pressing need. Specifically, to ensure that data are optimally useful across human and machine applications, they need to be findable, accessible, interoperable, and reusable (FAIR). The US EPA has annotated a number of DART studies in their ToxRefDB, but broader FAIR efforts have been stymied by the absence of a tool for harmonized, standardized annotation and an agreement on which terms to use. This lack of standardization, stemming from divergent language used to describe study parameters and end points, has so far inhibited crosstalk among individual studies and resources, preventing meaningful synthesis of data across studies.[18,19]

To combat this issue, harmonized vocabularies or ontologies can be implemented to ensure standardized annotation of extracted data and to facilitate semantic interoperability across studies, ultimately generating a computationally friendly database.[20] Ontologies are defined here as common, controlled knowledge representations designed to facilitate knowledge transfer and computer reasoning.[21] Ontologies are built upon "controlled vocabularies," which are defined here as collections of standardized terms that consistently describe data.[22] Examples of controlled vocabularies amenable to describing DART study data include the Unified Medical Language System (UMLS)[23,24] and the Organization for Economic Cooperation and Development (OECD) harmonized templates.[25] A harmonized lexicon that includes hierarchical relationships developed specifically for application to DART data can be found in the German Federal Institute for Risk Assessment (BfR) DevTox Project.[26,27]

Here, we used an augmented intelligence approach to address the challenge of harmonizing extracted data across studies in a two-phase mapping exercise. In the first phase, we developed a crosswalk (an annotation of the overlap between controlled vocabularies) between:

1. UMLS terms relevant to prenatal developmental endpoints
2. OECD harmonized template terminology for the prenatal developmental study design
3. The BfR DevTox database lexicon

In the second phase, we designed annotation code to automate the use of the crosswalk in standardizing primary source language extracted from high-quality DART studies.

We examined the performance of the controlled vocabulary crosswalk and associated annotation code in standardizing end point language from primary source extractions derived from high-quality (guideline or sufficiently compliant with guidelines) DART studies, specifically prenatal developmental toxicity studies performed by the NTP or submitted to the European Chemicals Agency (ECHA). We explored an augmented intelligence approach to automate applying the crosswalk to the extracted end point language to optimize what is conventionally a labor-intensive process and examined the potential utility in future data extraction, computation, and modeling endeavors. This approach allowed for cross-study comparison, integration, and standardized end point annotation of reference *in vivo* studies, with a focus here on DART end points, and aligns with the guidance established by the FAIR Data Principles.[28] The crosswalk, organized terminology related terms lists, annotation code, and resulting standardized developmental toxicity datasets are provided as open-access resources to the scientific community. We also provide a glossary of terms (Table 1) and an acronym list (Table 2) for reference.

## Methods

### *Initial Data Curation*

***Study sources.*** The NTP Electronic Library (NELI),[33] an internal NTP resource that requires system access and thus is not accessible to outside users, was searched by database administrators for all studies tagged as prenatal dose range-finding studies and full prenatal developmental toxicity study reports (also referred to as embryo-fetal developmental studies, teratology studies, or Segment II studies) prepared by the NTP (OECD test guideline 414[34]). Modified one-generation-type studies were not included in this evaluation (OECD test guideline 443[35]).

The ECHA database of Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH)[36] study results was searched for chemicals with available prenatal dose range-finding and developmental toxicity studies. Chemical dossiers that were identified to have data published in the open literature were identified, and relevant references were retrieved. References were reviewed by NTP to determine those of sufficient quality that warranted further data extraction (see Table 3).

***Study requirements and extraction.*** Studies that met the study design requirements above were reviewed by subject matter experts (Ceger, Hamm, Kleinstreuer, Blake, McIntyre, and Sutherland) to

**Table 1.** Glossary of terms.

| Term | Description |
|---|---|
| Annotation code | Code written in Python 3 (version 3.7), an open-source scripting language, to match extracted end points to terms in the controlled vocabulary crosswalk. |
| Combination word | A word used in developmental toxicology treatment end points that contains both a localization within the body (e.g., head) and an observation description (e.g., small) in their definition (e.g., microcephaly). One of four types of words found in the user-defined look up lists (defined below). |
| Controlled vocabulary | An authoritative set of terms selected and defined based on the requirements set out by the user group. Used to ensure consistent indexing (human or automated) or description of data or information. Controlled vocabularies do not necessarily have any structure or relationships between terms within the list.[29,30] |
| Crosswalk | A spreadsheet file that includes all three controlled vocabularies and shows how terms in each vocabulary were matched to each other according to the rules described in Table 4. |
| Crosswalk compatible | A term that did not get automatically mapped by the annotation code and to which only one UMLS term was applied manually; only one DevTox and only one OECD term are associated with that UMLS term in the crosswalk and are pulled in by default. If a term is not crosswalk compatible, then the appropriate DevTox and OECD terms are manually applied. |
| Concept unique identifiers (CUIs) | Unique identifiers assigned to concepts in UMLS (e.g., "C0015392" is the CUI for "eye"). A concept represents a single meaning and contains all words/phrases that express that meaning. |
| Data frame | A two-dimensional data structure with rows and columns (often used in Python's most common data analysis and manipulation tools). |
| Developmental toxicology | The study of the potential for substances to cause birth defects and other signs of toxicity during embryo-fetal development.[31] |
| Harmonized vocabulary | Combination of multiple languages into a single comparable view building from the components of each.[32] |
| Localization word | A word used in developmental toxicology treatment end points that describes a place within the body (e.g., head). |
| Natural language processing | A branch of computer science and linguistics that focuses on enabling computers to read, understand, process, and analyze human language. |
| Observation word | A word used in developmental toxicology treatment end points that describes a state of a body part (e.g., the word "small" in "small head"). |
| Ontology | A formal representation of a body of knowledge within a given domain, in a computer-readable format. Ontologies usually comprise a set of terms or concepts with relations that operate between them.[30] |
| Predictive toxicology | Multidisciplinary approach to toxicology that uses innovative approaches to predict human-relevant health effects from exposure to substances. |
| Root words | Words where variations on that word, such as adjectives and plurals, are also expected to be relevant (e.g., duplicat*) |
| Standardized language | Language that uses a common set of terms across datasets or resources. |
| Targeted review | Purposeful reviewing of specific end points known to be more susceptible to errors (e.g., end points related to "small" or "large"). |
| Unique word | A word used in developmental toxicology treatment end point whose meaning is not made up of a localization and observation but is rather a stand-alone concept (e.g., "mortality"). |
| User-defined look-up lists | A collection of common words used in developmental toxicology treatment end points, linked with associated words [e.g., "retina" and "eye" (linked together) and "non-live" and "dead" (linked together)]. There are 4 lists: Localization list, Observation list, Combination list, and Unique words list (defined separately). |
| Whole words | Words within the user-defined look up lists that are expected to stand alone and only be relevant with the current word form (e.g., large), as opposed to root words (defined above). |

Note: OECD, Organization for Economic Cooperation and Development; UMLS, Unified Medical Language System.

ensure that they met all additional minimum criteria for inclusion approved by the NTP, described in Table 3. Studies that met these requirements were considered of sufficient quality to include in further analyses.

Study metadata was extracted from each study that fulfilled all the criteria and included information such as chemical name, study identifiers, administration route and method, and treatment duration. For all treatment-related effects, the description of the end point was copy/pasted ("extracted") to maintain fidelity to the original study report language. While dose-response data (e.g., dose, mean responses, standard deviation, and statistical significance) were extracted into the database, that information was not the focus of this effort. The effort focused on prenatal developmental toxicology end points, though some other DART end points (namely, reproductive and maternal) were extracted if they were present in a study. All extracted information was compiled into a data set and is available for public access through the Data Sets tab on the Integrated Chemical Environment (https://ice.ntp.niehs.nih.gov/datasetdescription) by clicking on Developmental and Reproductive Toxicity (DART) in the Download Data Sets table.

### *Vocabulary Crosswalk*

This section describes the methods that apply to the first mapping phase of the project.

***Selection and curation of individual source vocabularies.*** We selected three sources of controlled vocabularies to use for mapping the prenatal developmental toxicology end points: *a*) Unified Medical Language System (UMLS), *b*) The BfR DevTox Project, and *c*) OECD Harmonized Templates. Further details on each resource and our methods for developing the crosswalk are described below. The UMLS terms were chosen to facilitate joining with resources such as US EPA's ToxRefDB,[18] a digital repository of *in vivo* toxicity study results[24]; OECD terminologies were chosen to facilitate international regulatory harmonization with data submitted under OECD Harmonized Templates through REACH; and the lexicon from the BfR DevTox Project was chosen as a bespoke tool created by experts in the field of teratology and tailored for prenatal developmental toxicology assessments. For ease of reference throughout this paper, from here on out we will refer to all three resources used in the crosswalk as "controlled vocabularies."

**UMLS (Unified Medical Language System).** We used a subset of UMLS that was applied to ToxRefDBv2.0 as described in Watford et al.[24] to facilitate future cross-referencing and merging between the extracted NTP/ECHA developmental toxicology data and ToxRefDB and because the UMLS terms used in Watford et al. are a list curated for the purposes of describing animal toxicology data reported in guideline study designs used in pesticide regulatory programs (i.e., inclusive of systemic,

**Table 2.** Acronyms.

| Acronym | Definition |
|---|---|
| AI | Augmented intelligence |
| AOP-Wiki | Adverse Outcome Pathway Wiki |
| CCNet | Cambridge Cell Networks |
| CEBS | Chemical Effects in Biologic Systems |
| CUIs | Concept Unique Identifiers |
| DART | Developmental and reproductive toxicity |
| DTT | Division of Translational Toxicology |
| ECHA | European Chemicals Agency |
| EDCs | Endocrine disrupting compounds |
| FAIR | Findability, accessibility, interoperability, and reuse |
| BfR | German Federal Institute for Risk Assessment |
| NIH | National Institutes of Health |
| NIEHS | National Institute of Environmental Health Sciences |
| NICEATM | NTP Interagency Center for the Evaluation of Alternative Toxicological Methods |
| NTP | National Toxicology Program |
| NELI | NTP Electronic Library |
| OECD | Organization for Economic Cooperation and Development |
| OHTs | OECD Harmonised Templates |
| REACH | Registration, Evaluation, Authorisation and Restriction of Chemicals |
| UMLS | Unified Medical Language System |

reproductive, and developmental effects).[24] "Terms" used by Watford et al. comprised a list of >3,000 unique combinations of 1,800+ UMLS Concept Unique Identifiers (CUIs), where one to six CUIs were packaged into a single text string to describe a single biological outcome (e.g., a developmental finding of a narrow aortic arch would be represented by the UMLS "term" "UMLS; C0000768;CUI;Congenital Abnormality|UMLS;C0003489;CUI; Aortic arch structure|UMLS;C1261287;CUI;Stenosis").[24] For this project, we used a subset of the full list relevant to the end points found in the study designs described above (360 "terms" related to maternal, reproductive, and developmental effects). Each term in this subset contained CUIs exactly as derived in Watford et al., denoting whether the effect was related to congenital, maternal, gross pathology, histopathology, or organ-specific effects. Congenital terms typically indicated a body localization and an observed effect; however, some terms indicated either just a localization (e.g., UMLS; C0000768;Congenital Abnormality|UMLS;C0015392;CUI;Eye) or just an observed effect (e.g., UMLS;C0000768;CUI;Congenital Abnormality|UMLS;C0019080;CUI;Hemorrhage). Because the goal was to use the UMLS terms as applied to ToxRefDB, we used the controlled vocabulary without any project-specific refinements.

**BfR DevTox Project.** The BfR DevTox Project aims, among other goals, to harmonize nomenclature used to describe developmental effects across laboratory studies to increase consistency across regulatory classifications for teratogenicity.[26,27,37] We used the BfR DevTox vocabulary as available in version 3.1, which includes a hierarchically structured lexicon of harmonized terms. The end points were divided into a three-part hierarchy: structure (e.g., external, visceral, skeletal, and maternal-fetal), localization (e.g., eye, femur, and lung), and observation (e.g., large, misshaped, and absent) Terms were considered inherently congenital-related unless the structure was "maternal-fetal," which facilitated comparison to the "congenital" terms from the ToxRefDB UMLS subset. To allow for an effect that was specific to a localization but not an observation, we added "NULL" terms to this controlled vocabulary by combining each structure and localization and leaving off an observation. This created localization-only terms (e.g., Visceral-Head-NULL).

**Table 3.** Prenatal developmental toxicity study requirements for inclusion.

| Study Parameters | Criteria |
|---|---|
| Species | Rats, mice, or rabbits |
| Route of administration to dams/does | Oral gavage, feed, water, inhalation, or subcutaneous injection |
| Doses tested | At least one dose and concurrent control group |
| Number of dams/does tested | Minimum of six viable pregnant females per dose and control group at the end of the study |
| Dosing window | Dosing occurs between gestational day 0 to parturition (single "day" treatment is acceptable) |
| Maternal end points | Toxicological end points (e.g., weight gain) and clinical observations performed and noted (e.g., clinical signs of toxicity, food and/or water consumption) |
| Fetal end points | Litter measurements performed and noted (e.g., live/dead, malformations, sex ratio) |

**OECD.** The OECD has developed Harmonised Templates (OHTs) for standardized reporting of health effect test results and methods. We translated the template for OHT74 "Developmental toxicity/teratogenicity," the template associated with OECD test guideline 414,[34] into a spreadsheet, preserving the four-part hierarchical nature of the end point reporting. Endpoints were categorized by fetal vs. maternal and abnormality vs. toxicity. Fetal end points were then categorized by structure (external, visceral/soft tissue, skeletal) and then by localization. The OHT74 controlled vocabulary did not include standardized terms for reporting observations.

***Crosswalk of the three controlled vocabularies.*** The set of terms from each controlled vocabulary was represented in a spreadsheet or a template, which we reviewed and adjusted to fit into a hierarchical structure of relevant terms to facilitate comparison across vocabularies. This mostly included separating fetal from maternal effects and localizations (e.g., liver, brain) from observations (e.g., enlarged, small), but the specific details and processing varied by vocabulary.

After processing the controlled vocabularies as described above, we mapped the three of them to one another to create a controlled vocabularies "crosswalk" designed for use by the annotation code (described below in section "Annotation Automation") in extracted end point mapping. Because UMLS was the primary controlled vocabulary of interest and developmental and reproductive terms were most relevant for the prenatal developmental end points in the extracted data, UMLS terms designated under ToxRefDB's end point categories "developmental" and "reproductive" served as the backbone to the crosswalk. The goal was to have each UMLS term separately paired with the best possible BfR DevTox term and best possible OECD term, organized in a spreadsheet. A "best" term ideally was an exact match in meaning. When exact matches for UMLS terms did not exist (using the same terms or direct synonyms), a set of standard crosswalk conventions were followed to consistently select the "best" BfR DevTox and OECD matches, which could be closely related in meaning or of disparate degrees of specificity and result in one UMLS term mapping to several BfR DevTox and/or OECD terms or vice versa, as seen in Figure 1. Whether a term pairing was an "exact" match or a different type of "best" match was captured in labels for each pairing (e.g., the label "related term" was used in the crosswalk to denote matches between related but not exact matches (e.g., pallor vs. discolored). Table 4 specifies the labels used to denote the various types of mappings that can be found in the crosswalk.

BfR DevTox and OECD terms were separately mapped to UMLS developmental and reproductive terms in separate spreadsheets so that each UMLS term was independently matched to a BfR DevTox term and to an OECD term.

In cases where best matches were of disparate degrees of specificity resulting in one-to-many or many-to-one mappings,

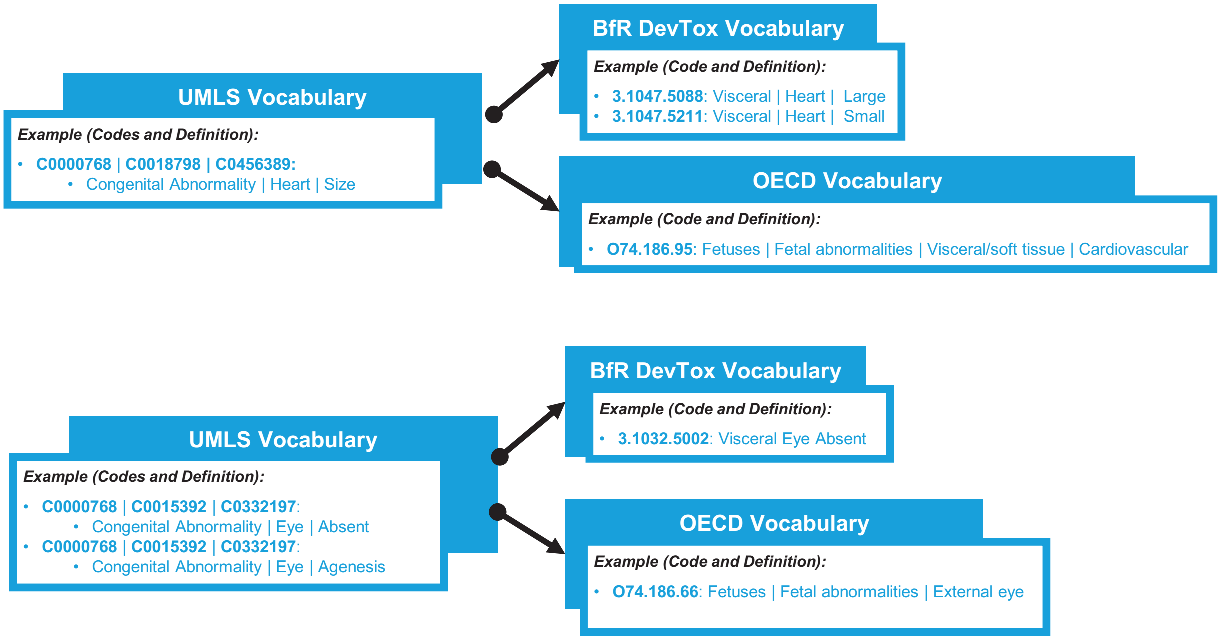

UMLS Vocabulary
Example (Codes and Definition):
• C0000768 | C0018798 | C0456389:
• Congenital Abnormality | Heart | Size
BfR DevTox Vocabulary
Example (Code and Definition):
• 3.1047.5088: Visceral | Heart | Large
• 3.1047.5211: Visceral | Heart | Small
OECD Vocabulary
Example (Code and Definition):
• O74.186.95: Fetuses | Fetal abnormalities | Visceral/soft tissue | Cardiovascular
UMLS Vocabulary
Example (Codes and Definition):
• C0000768 | C0015392 | C0332197:
• Congenital Abnormality | Eye | Absent
• C0000768 | C0015392 | C0332197:
• Congenital Abnormality | Eye | Agenesis
BfR DevTox Vocabulary
Example (Code and Definition):
• 3.1032.5002: Visceral Eye Absent
OECD Vocabulary
Example (Code and Definition):
• O74.186.66: Fetuses | Fetal abnormalities | External eye

**Figure 1.** Conceptualization of one-to-many and many-to-one crosswalk mappings. The top half of Figure 1 illustrates an example of a one-to-many match where a less specific Unified Medical Language System (UMLS) term mapped to multiple, more specific German Federal Institute for Risk Assessment (BfR) DevTox terms. Both BfR DevTox terms are equally relevant to the UMLS term. The bottom half of the figure is an example of a many-to-one match where multiple UMLS terms map to the same BfR DevTox term and to the same Organization for Economic Cooperation and Development (OECD) term.

each individual mapping was made on a separate spreadsheet row, resulting in multiple rows of the crosswalk containing the same UMLS term but different BfR DevTox/OECD term(s) or vice versa (see Figure 2).

**Table 4.** Definitions of labels given to vocabulary crosswalk matches.[a]

| Label | Definition |
|---|---|
| Exact | Terms match identically on localization and observation. |
| Mapped to a higher level | UMLS term was more specific than available BfR DevTox /OECD terms. For BfR DevTox terms, UMLS terms were mapped to "NULL" codes created by combining a BfR DevTox structure and localization with no observation. |
| Mapped to all possibilities | UMLS term included a localization but no observation. The UMLS term was mapped to all BfR DevTox terms for the localization.[b] |
| Mapped to all relevant end points | UMLS term had more than one BfR DevTox/ OECD match. The UMLS term was mapped to all BfR DevTox/OECD matches. |
| One step | The best BfR DevTox/OECD match for the UMLS term is not an exact match and requires one mental step to make the connection.[c] |
| Related term | BfR DevTox term is an exact match to the UMLS term meaning but uses different (synonymous) words. |
| Other | The observation portion of the UMLS term is "other." These UMLS terms were mapped to BfR DevTox "NULL" codes created by combining a BfR DevTox structure and localization with no observation. |

Note: BfR, German Federal Institute for Risk Assessment; OECD, Organization for Economic Cooperation and Development; UMLS, Unified Medical Language System.

[a] Note that additional labels were used to label and track rows of UMLS terms that did not receive BfR DevTox/OECD term matches. These can be found in the supplemental materials.

[b] This method of mapping was designed for use with the annotation code. It ensured extracted end points were matched to the best available UMLS term when a specific BfR DevTox term but no specific UMLS term was available.

[c] These decisions were made using expert subjectivity within the context of this project.

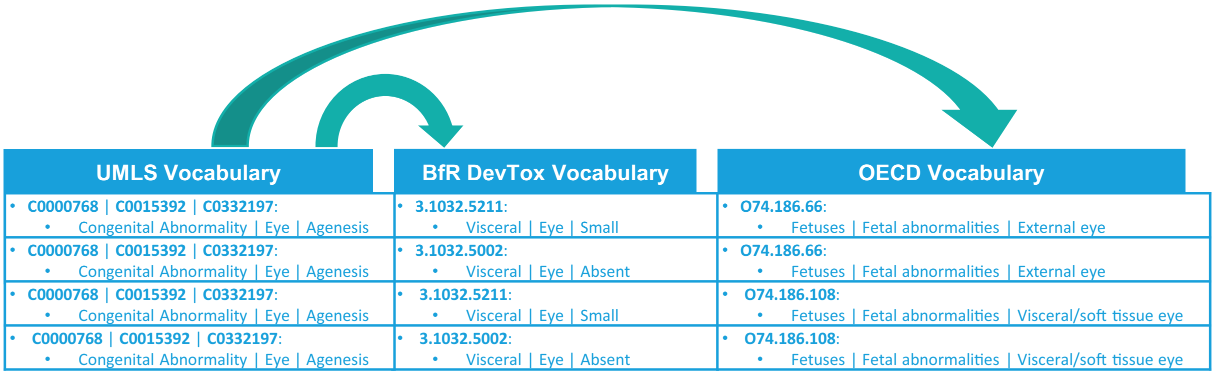
Terms from each vocabulary that did not map to terms from either of the other two vocabularies remained in the crosswalk without matches so as to be available for use in extracted end point mapping. This included UMLS terms designated by ToxRefDB as end point categories other than "developmental" and "reproductive," UMLS developmental and reproductive terms that did not have any appropriate BfR DevTox or OECD matches, and BfR DevTox and OECD terms without a matching UMLS developmental or reproductive term.

### *User-Defined Look-up Lists*

To be able to link the language of the primary source extracted end points to the crosswalk, we developed "user-defined look-up lists" in a spreadsheet format. We identified common words used in the extracted end point descriptions and within the controlled vocabularies, paired them with associated terms [these terms could be different spellings of the same word, direct synonyms, or simply related terms (e.g., leg and limb)], and stored them in four separate lists: Localization, Observation, Combination, and Unique words (see example excerpts in Figure 3):

1. Localization words describe the physical location of the observed effect (e.g., limb).
2. Observation words describe the observed effect, independent of location (e.g., small).
3. Combination words are single words that include a localization and an observation in their meaning (e.g., micromelia).
4. Unique words contain an entire endpoint concept within their definition and do not contain a localization and observation (e.g., resorption).

Related terms were limited for our purposes to words that can be used in place of the starting word and lead to an accurate controlled vocabulary match. This means that for a word to be a "related term" it cannot be more specific than the starting word or there is a risk of false positive controlled vocabulary matches (e.g., "eye" cannot be replaced with "retina" without risking applying a "retina" vocabulary term to an eye extracted end point



| UMLS Vocabulary | BfR DevTox Vocabulary | OECD Vocabulary |
|---|---|---|
| • C0000768 \| C0015392 \| C0332197:<br>• Congenital Abnormality \| Eye \| Agenesis | • 3.1032.5211:<br>• Visceral \| Eye \| Small | • O74.186.66:<br>• Fetuses \| Fetal abnormalities \| External eye |
| • C0000768 \| C0015392 \| C0332197:<br>• Congenital Abnormality \| Eye \| Agenesis | • 3.1032.5002:<br>• Visceral \| Eye \| Absent | • O74.186.66:<br>• Fetuses \| Fetal abnormalities \| External eye |
| • C0000768 \| C0015392 \| C0332197:<br>• Congenital Abnormality \| Eye \| Agenesis | • 3.1032.5211:<br>• Visceral \| Eye \| Small | • O74.186.108:<br>• Fetuses \| Fetal abnormalities \| Visceral/soft tissue eye |
| • C0000768 \| C0015392 \| C0332197:<br>• Congenital Abnormality \| Eye \| Agenesis | • 3.1032.5002:<br>• Visceral \| Eye \| Absent | • O74.186.108:<br>• Fetuses \| Fetal abnormalities \| Visceral/soft tissue eye |

**Figure 2.** Example section of the completed crosswalk and the first phase of mapping. This figure illustrates the first mapping phase where we developed a crosswalk (an annotation of the overlap between controlled vocabularies) between the three vocabularies. It illustrates both a complex and a simple mapping. The four rows contain the same Unified Medical Language System (UMLS) term, as the mapping for the UMLS term was a one-to-many match for both German Federal Institute for Risk Assessment (BfR) DevTox and Organization for Economic Cooperation and Development (OECD). The result of combining the mappings into one crosswalk is four rows, each with a different combination of BfR DevTox and OECD terms paired to the same UMLS term. The BfR DevTox and OECD terms are not paired to one another but to the UMLS term only. This means the crosswalk should be read left to right but only from UMLS to BfR DevTox and UMLS to OECD separately as indicated by the arrows. This design ensures the annotation code that the crosswalk was designed for will be able to find the best matches from each vocabulary for a given extracted end point. Details on how the code works to do this are presented in section "Annotation Automation."

that is not about retinas, but "retina" can be replaced with "eye" and a resulting controlled vocabulary match of an "eye" term to a "retina" end point would not be inaccurate, just less precise).

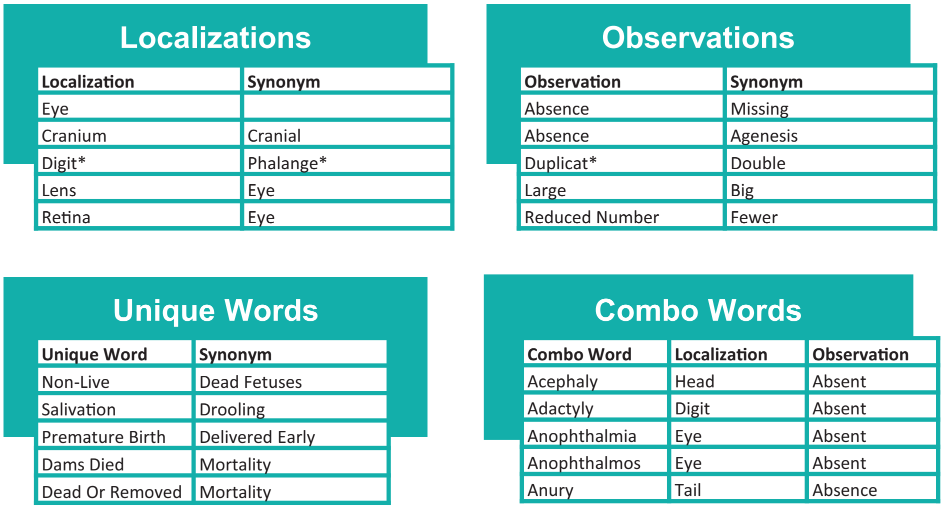
To facilitate various search methods for a word, including stemming, the Observation, Localization, and Combination word lists include two types of words: root words (noted with an asterisk) and whole words (no asterisk). Stemming involves the use of root words, where variations like adjectives and plurals are also considered relevant (e.g., duplicat$^*$), whereas whole words are intended to remain pertinent solely in their current word form (e.g., large).

### *Annotation Automation*

This section describes the methods that apply to the second mapping phase of the project.

We used Python 3 (version 3.7), an open-source scripting language, to develop code to match extracted end points to terms in the controlled vocabulary crosswalk by employing several widely used libraries (e.g., Pandas, re) and applying rule-based logic. The logic underlying the annotation code is illustrated in Figure 4 and described in detail below.

***Underlying logic.*** First, the annotation code reads in the primary source extraction file as a data frame. Extraction data must be formatted such that the treatment-related end points appear in one column, with one row per extracted end point. The code then searches for the user-defined look-up list words within the extracted end points, one list at a time. For root words, the code includes a function that identifies if the word appears regardless of characters surrounding that word. For whole words, the code includes a function to use regular expressions to require nonletter characters (such as spaces and punctuation) before and after the lookup word. After both functions run to find the user-defined look-up list words in an extracted end point, all words found are concatenated into four lists of found words that are specific to that extracted end point row: found words from the localization



Localizations

| Localization | Synonym |
|---|---|
| Eye | |
| Cranium | Cranial |
| Digit* | Phalange* |
| Lens | Eye |
| Retina | Eye |

Observations

| Observation | Synonym |
|---|---|
| Absence | Missing |
| Absence | Agenesis |
| Duplicat* | Double |
| Large | Big |
| Reduced Number | Fewer |

Unique Words

| Unique Word | Synonym |
|---|---|
| Non-Live | Dead Fetuses |
| Salivation | Drooling |
| Premature Birth | Delivered Early |
| Dams Died | Mortality |
| Dead Or Removed | Mortality |

Combo Words

| Combo Word | Localization | Observation |
|---|---|---|
| Acephaly | Head | Absent |
| Adactyly | Digit | Absent |
| Anophthalmia | Eye | Absent |
| Anophthalmos | Eye | Absent |
| Anury | Tail | Absence |

**Figure 3.** Example excerpts from the four User-Defined Look-Up Lists. Excerpts showing words and related terms from each of the four User-Defined Look-Up Lists: Localization words, Observation words, Combination words, and Unique words. These lists serve as input files to the annotation code, which uses these lists to link the extracted end points to the controlled vocabulary terms in the controlled vocabulary crosswalk. The process relies on Boolean logic (i.e., the asterisk represents wildcard searching where any word that includes the letters before the asterisk will be identified).

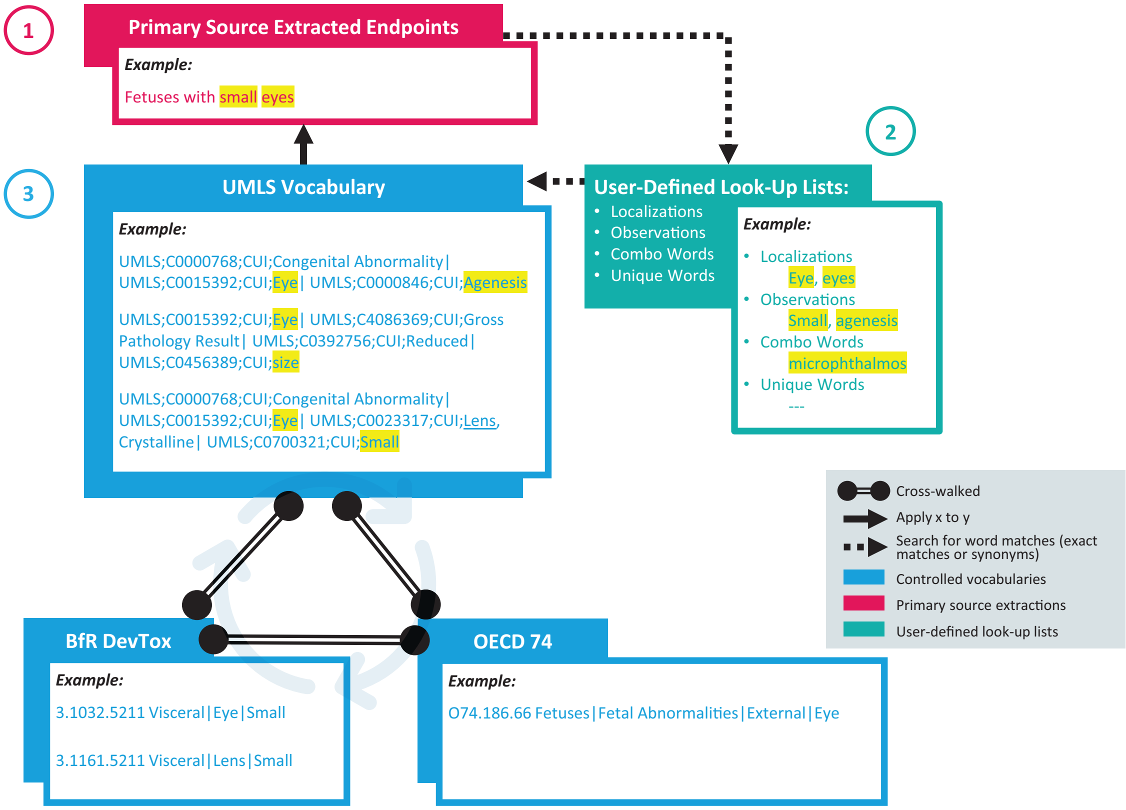

1
Primary Source Extracted Endpoints
Example:
Fetuses with small eyes
2
User-Defined Look-Up Lists:
• Localizations
• Observations
• Combo Words
• Unique Words
Example:
• Localizations
Eye, eyes
• Observations
Small, agenesis
• Combo Words
microphthalmos
• Unique Words
---
3
UMLS Vocabulary
Example:
UMLS;C0000768;CUI;Congenital Abnormality|
UMLS;C0015392;CUI;Eye| UMLS;C0000846;CUI;Agenesis
UMLS;C0015392;CUI;Eye| UMLS;C4086369;CUI;Gross
Pathology Result| UMLS;C0392756;CUI;Reduced|
UMLS;C0456389;CUI;size
UMLS;C0000768;CUI;Congenital Abnormality|
UMLS;C0015392;CUI;Eye| UMLS;C0023317;CUI;Lens,
Crystalline| UMLS;C0700321;CUI;Small
BfR DevTox
Example:
3.1032.5211 Visceral|Eye|Small
3.1161.5211 Visceral|Lens|Small
OECD 74
Example:
074.186.66 Fetuses|Fetal Abnormalities|External|Eye
Cross-walked
Apply x to y
Search for word matches (exact matches or synonyms)
Controlled vocabularies
Primary source extractions
User-defined look-up lists

**Figure 4.** Illustration of the annotation code logic and the second phase of mapping. This figure illustrates a simplified version of the three-step approach taken by the code to match extracted end points to terms from each of the three controlled vocabularies. The code completes step 3 three times, each time replacing which controlled vocabulary it searches for matches. The code then pulls in the terms from the other two controlled vocabularies that are linked in the crosswalk to any found matches. During this process if a UMLS term is not found, the code will pull in BfR terms based on an OECD match or an OECD term based on a BfR match. After the code runs on the three controlled vocabularies, the code will use the crosswalk between vocabularies to search for matches. After the conclusion of the process, the end points may have more than a single match (e.g., a UMLS term can have many matches in OECD or BfR). Note: BfR, German Federal Institute for Risk Assessment; OECD, Organization for Economic Co-operation and Development; UMLS, Unified Medical Language System.

look-up list, the observation look-up list, the combination look-up list, and the unique word look-up list.

After the code has generated the found words lists, the code brings in the related terms paired to these words by looking in the User-Defined Look-up Lists data frame and appends all of the paired related terms to the list of found words. The initial found words as well as the related terms are appended into a final list of found words that serve as output to the results, again grouped into four lists.

Using the words found in the steps above, the code searches the controlled vocabulary crosswalk for combinations of words found in the lists. The code first searches solely the UMLS terms and identifies a matching combination if one of the following three sets of logic are fulfilled:

1. A term in the controlled vocabulary crosswalk contains a localization and an observation included in the found word list.
2. A term in the controlled vocabulary crosswalk contains a combination word included in the found word list or its equivalent localization and observation.
3. A term in the controlled vocabulary crosswalk contains a unique word included in the found word list.

The processing follows a sequence of three distinct passes, each aligned with one of the controlled vocabularies: UMLS, BfR DevTox, and OECD. In the first pass, if a UMLS term fulfills one of the sets of logic, it is returned as one potential match to the extracted end point. The code will return all UMLS matches it finds BfR DevTox term(s) and OECD term(s) that are mapped in the crosswalk to the UMLS term match(es). To account for the variation of term specificity across controlled vocabularies, once this UMLS pass has been completed, the code repeats the process searching the BfR DevTox terms using the same logic rules and returns any BfR DevTox matches as well as the OECD and UMLS terms crosswalked to the BfR DevTox matches. This entire process is repeated for a third time searching the OECD terms.

Finally, the code deduplicates the matches so that any unique term is not returned as a match more than once. The result is structured so that each extracted end point row has a new column with all unique UMLS term matches, a new column for all unique BfR DevTox term matches, and a new column for all unique OECD term matches. Additional columns are paired with each term match column to indicate the number of terms returned from the associated vocabulary.

The resulting data frame of the process described above is exported for user review and processing.

***Refinement and application.*** We ran the annotation code iteratively on the NTP developmental toxicity primary study extraction file to optimize the user-defined look-up lists for the extracted end points. We manually reviewed each run output for quality and list refinement in order to improve both recall and precision. To refine the lists, words were removed from the user-defined look-up lists if the quality check found that they were resulting in more false positive matches than true matches (e.g.,

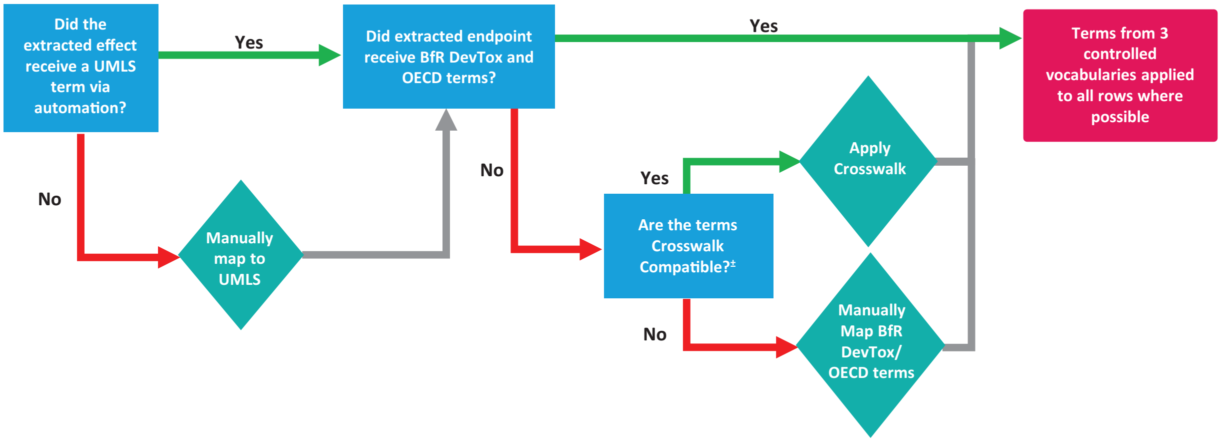

Did the extracted effect receive a UMLS term via automation?
Yes
Did extracted endpoint receive BfR DevTox and OECD terms?
Yes
Terms from 3 controlled vocabularies applied to all rows where possible
No
Manually map to UMLS
No
Yes
Apply Crosswalk
Are the terms Crosswalk Compatible?±
No
Manually Map BfR DevTox/ OECD terms

**Figure 5.** Manual mapping workflow of extracted end points not standardized to controlled vocabulary terms by the automated annotation code. Crosswalk compatible is defined as all three conditions being met: *a*) did not receive a Unified Medical Language System (UMLS) term from code, *b*) has only one UMLS term applied, and *c*) applied UMLS term crosswalks to no more than one German Federal Institute for Risk Assessment (BfR) DevTox and one Organization for Economic Cooperation and Development (OECD) term.

we removed localized alopecia terms to prevent false positive matches while keeping a correct general alopecia term). Additionally, words were added to the user-defined look-up lists if extracted end points were unmatched and the addition of the end point's content (localization, observation, combination, or unique word) to the appropriate list would not result in false positives (e.g., terms "number of females" and "number of males" were added to match with sex ratio). Once the user-defined look-up lists were optimized, we prospectively applied the code to the ECHA developmental toxicity primary study extraction file and evaluated the performance of the automated standardized term mapping workflow.

### *Manual Mapping and Curation of Terms*

Extracted end point rows that did not automatically receive controlled vocabulary terms using the approach described above were manually reviewed and matched to controlled vocabulary terms. As seen in Figure 5, rows that did not receive UMLS terms were addressed first. If those same rows also had not been matched to BfR DevTox and/or OECD terms, the crosswalk was applied to the manually entered UMLS terms where possible to match the extracted end point to BfR DevTox and/or OECD terms. If application of the crosswalk was not possible, BfR DevTox and OECD terms were then manually selected and entered. For consistency of mapping, the following set of decisions was devised and followed:

1. Controlled vocabulary terms were not applied to extracted endpoints if the controlled vocabulary term was more specific than the extracted endpoint.
2. Extracted endpoints that included the phrase "examined for" were mapped according to the phrase that follows the word "for."
3. All fetal extracted endpoints were mapped to at least one congenital UMLS term. If UMLS terms less specific than the extracted endpoints needed to be applied to meet this requirement, the less specific UMLS terms were applied according to the following logic:
    - If there was not a congenital UMLS term for either the localization or the observation of the fetal extracted endpoint, the general congenital abnormality "other" term (UMLS;C0000768;CUI;Congenital Abnormality|UMLS;C0205394;CUI;Other) was applied along with any applicable gross pathology and/or histopathology terms.
    - If a fetal extracted endpoint included both a specified localization component and a specified observation component but the observation component could not be matched to a congenital UMLS term for that localization, the UMLS term that pairs "congenital" with the given localization and the observation "other" was applied (e.g., UMLS;C0000768;CUI;Congenital Abnormality|UMLS;C0015392; CUI;Eye|UMLS;C0205394;CUI;Other).
    - If a fetal extracted endpoint included a specified localization component but not an observation component, the UMLS term that pairs just a "congenital" with the given localization (i.e. there is no observation component) was applied (e.g., UMLS;C0000768;CUI;Congenital Abnormality|UMLS;C0015392;CUI;Eye).
    - If a fetal extracted endpoint did not specify a localization or an observation, the plain congenital UMLS term was applied (UMLS;C0000768;CUI;Congenital Abnormality).

### *Quality Assessment of Results*

Extracted end points that received controlled vocabulary terms from the annotation code were reviewed for accuracy. This quality assessment included both targeted and nontargeted review. Targeted review was intentional examination of specific types of end points that were known to result in superfluous mappings (such as maternal end points, end points related to size, weight, or pregnancy) and of results that contained UMLS gross pathology or histopathology terms. Rows selected for targeted review were methodically revisited by type to make standard changes (e.g., removing all congenital vocabulary terms from maternal end points). Nontargeted review was a quality check on a random 5% of the end points not included in the targeted review. Both targeted and nontargeted rows were reviewed and assessed to ensure they followed the mapping decisions outlined above. Inaccurate and/or extraneous controlled vocabulary terms were removed. Correct but missing controlled vocabulary terms were added where necessary.

## Results

### *Study Extractions*

A total of 111 studies, representing 58 different substances, were retrieved from the NELI database of NTP study reports. Based on the methods and study criteria detailed above, information from

Table 5. Number of terms and associated terms for the four user-defined lookup lists.

| Lists | Number of terms | Number of associated terms |
|---|---|---|
| Localizations | 295 | 75 |
| Observations | 302 | 129 |
| Unique words | 96 | 72 |
| Combo words | 55 | 55 |

106 developmental toxicity studies (23 of which were range-finding studies) was extracted. These studies represented a total of 57 different chemicals. Extraction of all study information and animal data produced a spreadsheet containing >30,000 data rows. A total of 66 articles in the ECHA database were reviewed by NTP subject matter experts, again to determine adherence to minimum study protocol criteria for guideline-like prenatal developmental toxicity studies. Forty-eight articles were identified as having sufficient study quality (i.e., meeting all criteria in Table 3) to warrant extraction. These articles represented 57 different single chemicals and five mixtures containing two components. Extraction of all ECHA study information and animal data produced a spreadsheet containing >6,000 data rows. All extracted information is available at https://ice.ntp.niehs.nih.gov/datasetdescription by clicking on "Data" and then the Developmental and Reproductive Toxicity (DART) file in the Download Data Sets table.

### *Vocabulary Crosswalk*

In the output of the first mapping phase, the two controlled vocabulary mappings (from UMLS to BfR DevTox and UMLS to OECD) were combined into one spreadsheet to form the complete crosswalk such that the BfR DevTox and OECD terms that relate to the same UMLS term were on the same row(s), though effort was not made to ensure that the BfR DevTox and OECD terms on the same individual row logically relate to one another. This design means the crosswalk reads by UMLS term rather than by row and reads from UMLS on the left across to BfR DevTox in the center and OECD on the right separately rather than straight across from left to right (see Figure 2 and Excel Tables S1 and S2).

This design allows the annotation code to match an extracted end point to all exact or related terms from anywhere in the crosswalk, as well as to match an extracted end point to the best possible term from a controlled vocabulary that lacked an exact match, if there was an exact or related term match from another controlled vocabulary.

Of the 360 UMLS terms associated with developmental and reproductive end points, we were able to map all to at least one BfR DevTox term or at least one OECD term; the vast majority (81%) were mapped to both. Of the UMLS terms, all received OECD mappings while 22 reproductive terms and 48 general terms without specified localizations did not receive BfR DevTox mappings. To accomplish the mappings, 582 unique BfR DevTox terms were used, including 42 of the created "NULL" terms (localizations lacking observations), and 66 unique OECD terms were used; 1,395 BfR DevTox terms and 44 OECD terms were left unmapped to UMLS terms due to differences in specificity between the end points contained within each list.

### *User-Defined Look-up Lists*

The four user-defined look-up lists, Localizations, Observations, Unique Words, and Combo Words, were of varying lengths as seen in Table 5. The lists can be found in supplemental material (Excel Tables S3–S7). These lists were reviewed by subject matter experts and found to cover a comprehensive range of expected developmental, teratogenic, and reproductive end points and are therefore useful for future study extraction efforts. Further, the content of the user-defined look-up lists can be readily adapted to correspond to other study types.

### *Annotation Code*

The code, developed with readability and reuse in mind, was written to be an intuitive workflow that follows human-readable logic and where the inputs can be adjusted as needed. As detailed in "Methods" and represented in Figure 4, the code begins with reading into Python the Microsoft Excel spreadsheet input files from the primary source extractions, and then converts the inputs into three separate data objects: *a*) a data frame for the extractions, *b*) the four user-defined look-up lists, and *c*) the controlled vocabulary crosswalk. Then, in short, the code *a*) reads the extraction file row by row, *b*) searches each extracted end point for words found in the user-defined look-up lists, and then *c*) matches the found words in the extracted end point with words in the controlled vocabulary crosswalk. Due to the generalizable nature of the code, each of these three data objects/processes could be adapted for future projects, e.g., using the existing user-defined lookup lists and crosswalk to automate mapping of extracted information from published developmental toxicity studies identified via systematic literature review, expanding the crosswalk to map terminologies from another developmentally relevant controlled vocabulary, or adjusting all three inputs to cover other toxicological study types.

The annotation code is a .py file that was tested in Python 3.7, utilizing several mainstream libraries. The .py file "DevTox_AutomationPilot_02_27_2020" is available in a github repository (see https://github.com/NIEHS/DevToxAutoStandardization) and is also included in supplemental materials (see Supplemental Material, DevTox_AutomationPilot_02_27_2020_python code).

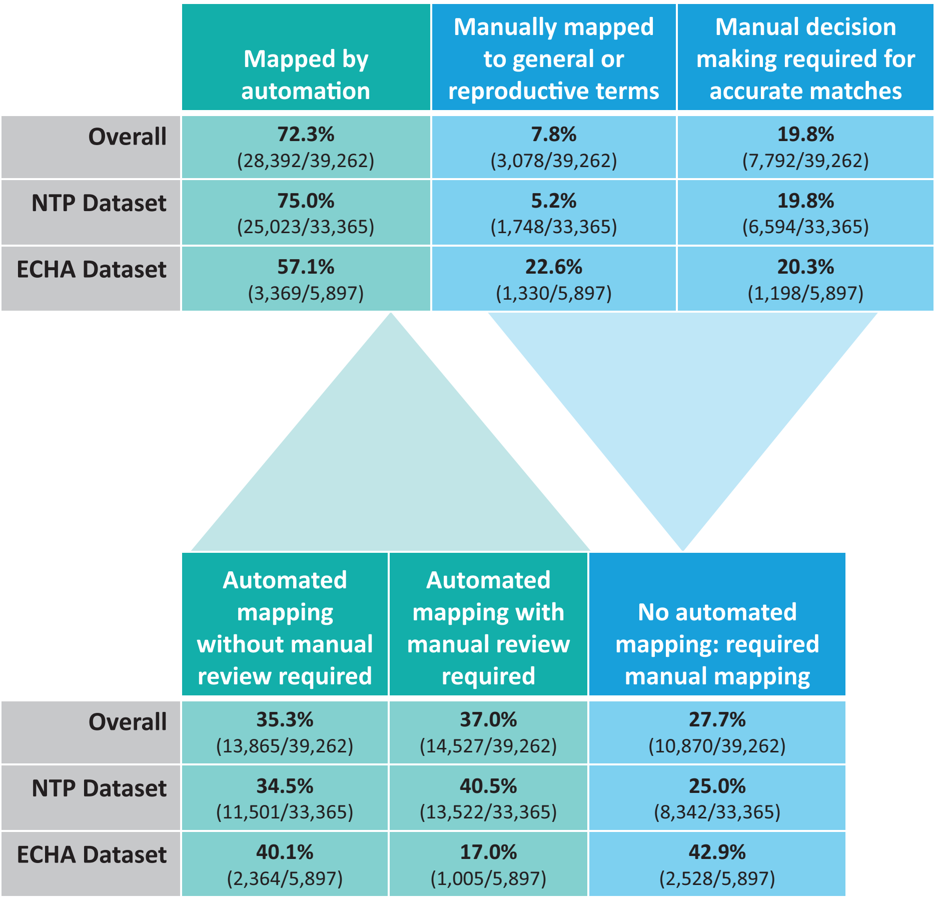
### *Controlled Vocabulary Application*

Of all extracted end points from both datasets, 3% (1,269 out of 40,531) were not developmental or reproductive end points (e.g., "number of does treated"). Of the extracted end points remaining, 75% (25,023 out of 33,365) from the NTP dataset received standardized controlled vocabulary mappings via the annotation code, where 40.5% (13,522) were extracted end point types that required manual review through targeted review and 34.5% (11,501) did not require manual review. Of the 25% (8,342) not mapped by the code, 19.8% (6,594) were able to be manually mapped to specific terms using manual decision making and 1.8% (589) were mapped to reproductive terms, while the remaining 3.4% (1,159) were at best able to be manually mapped to generic terms.

For the ECHA dataset, 57.1% (3,369 out of 5,897) of extracted end points received standardized controlled vocabulary mappings via the annotation code, where 17% (1,005) were extracted end point types that required manual review through targeted review and 40.1% (2,364) did not require manual review. Of the 42.9% (2,528) not mapped by the code, 20.3% (1,198) were able to be manually mapped to specific terms using manual decision making and 5.2% (305) were mapped to reproductive terms, while the remaining 17.4% (1,025) were at best able to be manually mapped to generic terms. The results for both datasets combined, as well as the breakdown by dataset, can be found in Figure 6. We estimate that the application of the automated term standardization workflow reduced time spent standardizing the language for these two datasets by 54%, from 656 to 304 h, assuming 1 min per manual extraction and 30 s per manual review.

### *Standardized Datasets*

The high-quality developmental toxicity animal study datasets resulting from the second mapping phase, the application of the



| | Mapped by automation | Manually mapped to general or reproductive terms | Manual decision making required for accurate matches |
|---|---|---|---|
| Overall | **72.3%**<br>(28,392/39,262) | **7.8%**<br>(3,078/39,262) | **19.8%**<br>(7,792/39,262) |
| NTP Dataset | **75.0%**<br>(25,023/33,365) | **5.2%**<br>(1,748/33,365) | **19.8%**<br>(6,594/33,365) |
| ECHA Dataset | **57.1%**<br>(3,369/5,897) | **22.6%**<br>(1,330/5,897) | **20.3%**<br>(1,198/5,897) |

| | Automated mapping without manual review required | Automated mapping with manual review required | No automated mapping: required manual mapping |
|---|---|---|---|
| Overall | **35.3%**<br>(13,865/39,262) | **37.0%**<br>(14,527/39,262) | **27.7%**<br>(10,870/39,262) |
| NTP Dataset | **34.5%**<br>(11,501/33,365) | **40.5%**<br>(13,522/33,365) | **25.0%**<br>(8,342/33,365) |
| ECHA Dataset | **40.1%**<br>(2,364/5,897) | **17.0%**<br>(1,005/5,897) | **42.9%**<br>(2,528/5,897) |

**Figure 6.** Results of automated controlled vocabulary term application (annotation code) and impact on manual effort requirements. The top table in Figure 6 provides the breakdown of extracted end points automatically mapped to controlled vocabulary terms by the code (green) vs. extracted end points that were not automatically mapped and instead required manual mapping (blue). Extracted end points required manual mapping because the extracted end points did not have precise controlled vocabulary term matches and could either only be mapped to general and reproductive terms (8%) or could be mapped to more specific terms but only by using human logic (20%). The bottom table in Figure 6 illustrates the breakdown of manual effort required when using the automated annotation code. Manual effort was eliminated for the 35% of extractions that were automatically mapped via the annotation code and did not require manual review, halved for the 37% of extractions that were automatically mapped via the annotation code but did require manual review, and unaffected for the 28% of extractions that required manual mapping.

automated annotation tool, are now computationally accessible due to the standardized terminologies used. These datasets are provided as a resource to the scientific community and may be accessed via the Integrated Chemical Environment (https://ice.ntp.niehs.nih.gov/). Work is underway to combine information from these studies with other databases (e.g., ToxRefDB) and apply them to projects such as computational elucidation of adverse outcome pathways, identification of reference compound sets for validation of new approach methodologies, and analysis of animal study variability and effect incidence.

## Discussion

The creation and implementation of an automated process to annotate extracted DART data from select sources, using a terminology crosswalk which unified three independent sets of controlled vocabularies, allowed for efficient end point standardization of primary source extracted end points. Data for this study came from two reference databases containing guideline range-finding and prenatal developmental toxicity animal studies. The methods developed here resulted in high-fidelity extracted end point mappings that reduced time spent standardizing the language for the datasets by an estimated 352 h, assuming 1 min per manual extraction and 30 s per manual review. In particular, the tools we built enabled the direct comparison of end points across studies through unified controlled vocabularies and provide a powerful framework upon which future data extractions can be organized and standardized.

The controlled vocabulary crosswalk was developed with a subset of the UMLS terms used in annotating ToxRefDB2.0 serving as the primary vocabulary, with BfR DevTox and OECD terms mapped onto the UMLS terms. All UMLS terms, which represented maternal, reproductive, and developmental end points, were mapped to at least one BfR DevTox or one OECD term, and nearly all (81%) were mapped to a term from both BfR DevTox and OECD. The 19% not mapped to both are a combination of 22 reproductive terms that are out of the scope of the BfR DevTox vocabulary and 48 terms consisting of observations without specific localizations (e.g., bent skeletal bone). While BfR DevTox was specific to developmental toxicology and tended to be more specific to fetal abnormalities, OECD terms were structured around test guideline 414 and therefore included effects beyond those measured in fetuses (e.g., maternal toxicity, reproductive effects). OECD also included less-specific terms, tending to combine multiple effects in broader categories. These features allowed for mapping to the reproductive and less-specific UMLS terms.

Due to the logic of the annotation code built to account for variance in specificity of terms across the three controlled vocabularies, no BfR DevTox or OECD term that was more specific than the UMLS term of interest was mapped to the UMLS term, making the specific nature of BfR DevTox terms prohibitive in mapping with these non-location-specific UMLS terms. The specificity of the BfR DevTox terms, however, lead to more exact or related mappings between UMLS and BfR DevTox terms than found between UMLS and OECD, which often had broader OECD terms mapped to more specific UMLS terms. A nuance to the rule about specificity was that it did not apply to scenarios where the concept of a UMLS term could be captured by a combination of BfR DevTox or OECD terms (e.g., UMLS terms for "size" were paired with BfR DevTox terms for both "large" and "small"). Given the difference in numbers (e.g., BfR DevTox has 1,621 more terms than the UMLS subset used in this effort) and specificity of terms within each controlled vocabulary, not all BfR DevTox and all OECD terms were needed to map all UMLS terms, but they remained in the crosswalk for use by the annotation code in mapping the primary source extractions. Keeping these terms in the crosswalk combined with the cyclical nature of the code ensured that any exact or related matches to extracted end points from any of the three controlled vocabularies would be found regardless of whether there was a UMLS match for the extracted end point while limiting extraneous matches where the controlled vocabulary term was too specific for the extracted end point (e.g., a controlled vocabulary term for "arm" vs. an extracted end point that only specifics "limb").

The resulting crosswalk that unified and standardized the three controlled vocabularies is a powerful tool that could be used to combine datasets that have already been mapped to individual terminologies contained therein, such as studies that have been described and submitted to regulatory authorities using the OHT74 form or studies in EPA's ToxRefDB coded using UMLS terms. Using UMLS as the backbone for the crosswalk also extended the utility of the other two controlled vocabularies since UMLS cross-referencing is included in the NCI Thesaurus project supporting the largest collection of open biomedical ontologies (http://www.obofoundry.org/), which is almost entirely lacking in toxicological ontologies.

Standardized language is critical for the ability to compare studies within or across databases and ultimately facilitates subsequent analyses and data applications. Such downstream applications include calculating chemical- and species-specific effects, interrogating variability of mammalian toxicity studies, establishing reference datasets for validation of new approaches, and computational modeling, among others. Standardized nomenclature has been implemented in numerous other scientific disciplines including genetics, systems biology, histopathology, and clinical medicine. Within the field of toxicology, the ToxWiz ontology developed by Cambridge Cell Networks (CCNet) harmonized multiple controlled vocabularies to unify three major categories of interest (effects/outcomes, target proteins/genes, and chemical structures). In doing so, the ToxWiz ontology is well-suited for predicting toxicological effects as well as interrogating modes of action.[38] The approach described here is complementary to ToxWiz in that it similarly harmonized multiple controlled vocabularies but is specific to developmental toxicity. However, the challenges highlighted and addressed here are common across many study types and applications, i.e., variation in primary source language describing end points and specificity of end points, outcomes, or treatment-related effects, which can stymie automated data extraction mapping workflows. The approach outlined in this work successfully implemented multiple strategies to overcome these challenges (illustrated by the numerous mapping conventions presented in Table 4) and minimize the extent of manual labor required to generate data that is fit for use.

One of the strengths of this work is its adherence to the FAIR Guiding Principles[28] for data science stewardship, meaning that the methodology itself and the extracted data are findable, accessible, interoperable, and reusable. Thus, the annotation code used here could be applied as-is to other databases of regulatory guideline DART study end points (e.g., OECD TG 414 studies) and would be expected to yield similarly successful results. The structure of the code also enables easy expansion beyond developmental toxicology end points for application to other study types, where the user would simply utilize appropriate user-defined look-up lists with related terms and desired controlled vocabularies for mapping. For example, localization and observation lists can be used for any toxicity data that can be described in terms of localizations and observations (e.g., neoplastic and nonneoplastic lesions such as hepatocellular adenoma or cardiomyocyte atrophy, respectively), while the "unique list" could be used to describe other types of toxicological end points (e.g., functional end points such as locomotor activity or acoustic startle response). The interoperability and reusability of the methods used here imparts a high degree of versatility that expands beyond the field of developmental and reproductive toxicity.

Of the two datasets used here, the NTP dataset realized a higher percentage of extracted end points mapped by the code to controlled vocabulary terms than the ECHA dataset did (75% and 57%, respectively). This was expected as the annotation code developed for automated end point mapping was optimized for the NTP dataset. Extracted end points from both datasets that were not standardized by the code were either reproductive or too general to be mapped to a single or closely matched controlled vocabulary end point (e.g., "number of fetuses with abnormal organ") or required the use of human logic to make a match (e.g., deciding the best match for a "globular shaped heart" extraction was a controlled vocabulary term for "misshapen heart" or deciding the best match for a "blue nose" extraction was a controlled vocabulary term for "skin discoloration") and therefore not expected to be standardized. The code performance on the ECHA dataset indicates the potential for its utility across other developmental toxicity databases as-is or with some modifications.

When adapting the annotation code, steps can be taken to improve upon results seen in this work. First and foremost, the user-defined look-up lists that feed into the code should be tailored to the content of the extracted data if the content veers from prenatal developmental toxicology. As mentioned, the approach is not constrained by the study type, but expert knowledge is required to develop or identify the most appropriate vocabularies or ontologies for the study type and dataset of interest. Familiarity with the study designs and how the original toxicity end points are represented in the database is also helpful. For example, if the original data were collected from guideline studies, then it is more likely that the end point language is standardized than if the data were collected from a variety of journal articles. Experts can also advise if other elements of the study design are relevant for interpreting the end point language, such as a consideration of life stage during which the end point data was collected, dosing strategies, or genetic background of the animal strains. Taken together, future expansion of this approach would need to consider the complexity of toxicological data and unique study designs that require customization of the user-defined lookup lists.

Also, key to the success of the code is the quality of the extracted end points, which we found to be the driving factor for determining the amount of manual review needed. Misspelled words in the extractions were not able to be automatically mapped. Spell checks of the extracted end points, or incorporating common misspellings of end points as related terms, could overcome this limitation. In the same vein, eliminating nonendpoints and out-of-scope end points from the dataset before beginning would decrease the time spent to resolve those end points. Ideally, toxicological study authors would utilize controlled vocabularies or ontologies proactively when publishing datasets to minimize errors and maximize the utility of the data for use in other applications, instead of retroactively applying to published datasets. This along with approaches such as publishing supplemental materials in structured data formats, building and using high-quality data dictionaries, and preparing data for publication in open access repositories would increase adherence with FAIR principles and are all parts of the consideration of data management and sharing that is now required for all National Institutes of Health (NIH)-funded research as of January 2023.

Another way to increase manual hours saved would be to limit results that require manual review through targeted application, such as by using the code only on fetal outcomes and not on maternal end points. Alternatively, additional logic could be added to distinguish between maternal and fetal extractions and between maternal and fetal controlled vocabulary terms. Even within prenatal toxicology, more words in the user-defined look-up lists used here could be adapted into root words or paired with more related terms. In addition, the code could be appended to only use crosswalk matches labeled as "exact" if a stricter standardization was desired. Logic could also be added to determine appropriate application of gross pathology and histopathology terms. Improvements could also be made to the logic for mapping organ weight end points and pregnancy-related end points. For this effort, improved logic to eliminate the need for manual review would have saved an additional ~180 h (assuming 1 min/30 s for extraction/review).

Contexts in which this approach could be applied, adapted or as-is, might include when legacy data has already been extracted into a database or system of some kind, and there is interest in making that data interoperable with other systems. Using a crosswalk to link one knowledge system to another can facilitate semantic interoperability across systems, which is essential to leverage existing legacy data for analyses and decision making. For example, US EPA's ToxRefDB includes legacy prenatal developmental toxicity study data, which can now be combined with the NTP and ECHA prenatal developmental toxicity data reviewed in this effort, to create a larger integrated dataset. Datasets like this can be used to better understand the developmental effects of chemicals and to make more informed decisions to protect public health. Adding more controlled vocabularies to the crosswalk could link these data to other systems, such as the Collaborative Adverse Outcome Pathway Wiki (AOP-Wiki) (http://aopwiki.org), so that legacy testing data could be linked to specific key events of an AOP to better understand mechanisms of developmental toxicity. Additional vocabularies could also expand the application of this approach to other knowledge domains, such as by linking to genes. Based on our experience in this effort, these expansions and applications would benefit from an augmented intelligence approach integrating subject matter expertise to ensure that the connections are scientifically accurate and not perpetuating errors or mismatches.

Here, we present methods for automatically standardizing the annotation of developmental toxicity end points extracted from primary sources and a resulting FAIR dataset that is computationally accessible. This work developed and implemented a controlled vocabulary crosswalk to ensure consistent language in end point descriptions in the final dataset and automated over 70% of the term mappings using annotation code that is intuitive, extensible, and reusable. The resulting dataset allowed for direct comparison across studies, which will serve as an indispensable tool in establishing reference datasets for validation of new approaches, computational modeling, and systematic reviews for human health risk assessment. Importantly, the engineering of the annotation code can be easily modified for application to other study types to support improved utility of legacy datasets in future analyses.

## Supplemental Materials

The annotation python code is available as supplemental material (DevTox_AutomationPilot_02_27_2020_python code) and on the National Institute of Environmental Health Sciences (NIEHS)'s public GitHub in the DevToxAutoStandardization repository (https://github.com/NIEHS/DevToxAutoStandardization). The repository also contains a Microsoft Excel document with user-defined look-up lists and the vocabulary crosswalk utilized in this study. This Excel file is also included as Excel Tables S1–S7. The Excel crosswalk columns are defined in the included README. Each of the user-defined look-up lists has its own sheet. The format of these sheets is compatible with the code.

## Acknowledgments

This research was supported by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences (NIEHS), Intramural Research projects ES103386, ES103380 and ES013385, and NIEHS Contracts HHSN273201500010C (ILS) and GS00Q14OADU417, Order HHSN273201600015U (ICF).

Data can be accessed online at the Integrated Chemical Environment (https://ice.ntp.niehs.nih.gov/).

## References

1. WHO (World Health Organization). 2021. *Framework for the Use of Systematic Review in Chemical Risk Assessment*. Geneva, Switzerland: WHO, Chemical Safety and Health Unit. https://www.who.int/publications/i/item/9789240034488 [accessed 8 October 2021].
2. Hood RD. 2016. *Developmental and Reproductive Toxicology: A Practical Approach*. Boca Raton, FL: CRC Press.
3. Lea IA, Gong H, Paleja A, Rashid A, Fostel J. 2016. CEBS: a comprehensive annotated database of toxicological data. Nucleic Acids Res 45(D1):D964–D971, PMID: 27899660, https://doi.org/10.1093/nar/gkw1077.
4. US EPA (US Environmental Protection Agency). 1991. Guidelines for developmental toxicity risk assessment. Fed Reg 56(234):63798–63826.
5. Barker DJ. 2007. The origins of the developmental origins theory. J Intern Med 261(5):412–417, PMID: 17444880, https://doi.org/10.1111/j.1365-2796.2007.01809.x.
6. Heindel JJ, Vandenberg LN. 2015. Developmental origins of health and disease: a paradigm for understanding disease cause and prevention. Curr Opin Pediatr 27(2):248–253, PMID: 25635586, https://doi.org/10.1097/MOP.0000000000000191.
7. Heindel JJ, Skalla LA, Joubert BR, Dilworth CH, Gray KA. 2017. Review of developmental origins of health and disease publications in environmental epidemiology. Reprod Toxicol 68:34–48, PMID: 27871864, https://doi.org/10.1016/j.reprotox.2016.11.011.
8. Knudsen TB, Fitzpatrick SC, De Abrew KN, Birnbaum LS, Chappelle A, Daston GP, et al. 2021. FutureTox IV workshop summary: predictive toxicology for healthy children. Toxicol Sci 180(2):198–211, PMID: 33555348, https://doi.org/10.1093/toxsci/kfab013.
9. Foster PM. 2017. Influence of study design on developmental and reproductive toxicology study outcomes. Toxicol Pathol 45(1):107–113, PMID: 27708197, https://doi.org/10.1177/0192623316671608.
10. Zurlinden TJ, Saili KS, Rush N, Kothiya P, Judson RS, Houck KA, et al. 2020. Profiling the ToxCast library with a pluripotent human (H9) stem cell line-based biomarker assay for developmental toxicity. Toxicol Sci 174(2):189–209, PMID: 32073639, https://doi.org/10.1093/toxsci/kfaa014.
11. Seiler AE, Spielmann H. 2011. The validated embryonic stem cell test to predict embryotoxicity in vitro. Nat Protoc 6(7):961–978, PMID: 21720311, https://doi.org/10.1038/nprot.2011.348.

12. Brannen KC, Panzica-Kelly JM, Danberry TL, Augustine-Rauch KA. 2010. Development of a zebrafish embryo teratogenicity assay and quantitative prediction model. Birth Defects Res B Dev Reprod Toxicol 89(1):66–77, PMID: 20166227, https://doi.org/10.1002/bdrb.20223.
13. Kleinstreuer NC, Ceger P, Watt ED, Martin M, Houck K, Browne P, et al. 2017. Development and validation of a computational model for androgen receptor activity. Chem Res Toxicol 30(4):946–964, PMID: 27933809, https://doi.org/10.1021/acs.chemrestox.6b00347.
14. Daston GP, Beyer BK, Carney EW, Chapin RE, Friedman JM, Piersma AH, et al. 2014. Exposure-based validation list for developmental toxicity screening assays. Birth Defects Res B Dev Reprod Toxicol 101(6):423–428, PMID: 25475026, https://doi.org/10.1002/bdrb.21132.
15. Browne P, Judson RS, Casey WM, Kleinstreuer NC, Thomas RS. 2015. Screening chemicals for estrogen receptor bioactivity using a computational model. Environ Sci Technol 49(14):8804–8814, PMID: 26066997, https://doi.org/10.1021/acs.est.5b02641.
16. US EPA (US Environmental Protection Agency). 2002. *A Review of the Reference Dose and Reference Concentration Process. EPA/630/P-02/002F*. Washington, DC: US Environmental Protection Agency. https://www.epa.gov/risk/review-reference-dose-and-reference-concentration-processes-document [accessed 22 December 2023].
17. Ekins S, Puhl AC, Zorn KM, Lane TR, Russo DP, Klein JJ, et al. 2019. Exploiting machine learning for end-to-end drug discovery and development. Nat Mater 18(5):435–441, PMID: 31000803, https://doi.org/10.1038/s41563-019-0338-z.
18. Whaley P, Edwards SW, Kraft A, Nyhan K, Shapiro A, Watford S, et al. 2020. Knowledge organization systems for systematic chemical assessments. Environ Health Perspect 128(12):125001, PMID: 33356525, https://doi.org/10.1289/EHP6994.
19. Hardy B, Apic G, Carthew P, Clark D, Cook D, Dix I, et al. 2012. A toxicology ontology roadmap. Altex 29(2):129–137, PMID: 22562486, https://doi.org/10.14573/altex.2012.2.129.
20. Kleinstreuer NC, Ceger PC, Allen DG, Strickland J, Chang X, Hamm JT, et al. 2016. A curated database of rodent uterotrophic bioactivity. Environ Health Perspect 124(5):556–562, PMID: 26431337, https://doi.org/10.1289/ehp.1510183.
21. Robinson PN, Bauer S. 2011. *Introduction to Bio-Ontologies*. Boca Raton, FL: CRC Press.
22. Courtot M, Juty N, Knüpfer C, Waltemath D, Zhukova A, Dräger A, et al. 2011. Controlled vocabularies and semantics in systems biology. Mol Syst Biol 7:543, PMID: 22027554, https://doi.org/10.1038/msb.2011.77.
23. Bodenreider O. 2004. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic Acids Res 32:D267–D270, PMID: 14681409, https://doi.org/10.1093/nar/gkh061.
24. Watford S, Ly Pham L, Wignall J, Shin R, Martin MT, Friedman KP. 2019. ToxRefDB version 2.0: improved utility for predictive and retrospective toxicology analyses. Reprod Toxicol 89:145–158, PMID: 31340180, https://doi.org/10.1016/j.reprotox.2019.07.012.
25. OECD (Organisation for Economic Co-operation and Development). 2018. *OECD Template #74: Developmental Toxicity/Teratogenicity (Version [6.3])*. https://www.oecd.org/ehs/templates/OHT-74-endpoint-study-record-developmental-toxicity-teratogenicity-v8.5-Sept-2020.doc [accessed 8 October 2021].
26. Solecki R, Heinrich V, Rauch M, Chahoud I, Grote K, Wölffel B, et al. 2010. The DevTox site: harmonized terminology and database. In: *Comprehensive Toxicology*. McQueen CA, ed. Amsterdam, Netherlands: Elsevier Science, 339–346.
27. Solecki R, Rauch M, Gall A, Buschmann J, Kellner R, Kucheryavenko O, et al. 2019. Update of the DevTox data database for harmonized risk assessment and alternative methodologies in developmental toxicology: report of the 9th Berlin Workshop on Developmental Toxicity. Reprod Toxicol 89:124–129, PMID: 31288076, https://doi.org/10.1016/j.reprotox.2019.07.003.
28. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, et al. 2016. The FAIR guiding principles for scientific data management and stewardship. Sci Data 3:160018, PMID: 26978244, https://doi.org/10.1038/sdata.2016.18.
29. National Cancer Institute. 2023. *NCI Thesaurus: Controlled Vocabulary (Code C48697)*. https://ncit.nci.nih.gov/ncitbrowser/ConceptReport.jsp?dictionary=NCI_Thesaurus&ns=ncit&code=C48697 [accessed 22 December 2023].
30. American Society for Indexing. 2023. *Taxonomies & Controlled Vocabularies SIG*. https://www.taxonomies-sig.org/about.htm#cv [accessed 22 December 2023].
31. NTP (National Toxicology Program). 2023. *Developmental & Reproductive Toxicity*. https://ntp.niehs.nih.gov/whatwestudy/testpgm/devrepro [accessed 22 December 2023].
32. Data Sharing for Demographic Research. 2023. *Data Harmonization*. https://www.icpsr.umich.edu/web/pages/DSDR/harmonization.html [accessed 22 December 2023].
33. NTP (National Toxicology Program). 2020. *NTP Electronic Library: NELI*. Research Triangle Park, NC: National Toxicology Program.
34. OECD (Organisation for Economic Co-operation and Development). 2018. *Test No. 414: Prenatal Developmental Toxicity Study*. https://doi.org/10.1787/9789264070820-en [accessed 8 October 2021].
35. OECD (Organisation for Economic Co-operation and Development). 2018. *Test No. 443: Extended One-Generation Reproductive Toxicity Study*. https://doi.org/10.1787/9789264185371-en [accessed 8 October 2021].
36. ECHA (European Chemical Agency). 2023. *REACH Program - Registration, Evaluation, Authorisation and Restriction of Chemicals*. Helsinki, Finland: European Chemical Agency. https://echa.europa.eu/regulations/reach/understanding-reach [accessed 8 October 2021].
37. Makris SL, Solomon HM, Clark R, Shiota K, Barbellion S, Buschmann J, et al. 2009. Terminology of developmental abnormalities in common laboratory mammals (version 2). Reprod Toxicol 28(3):371–434, PMID: 19729062, https://doi.org/10.1016/j.reprotox.2009.06.010.
38. Hardy B, Apic G, Carthew P, Clark D, Cook D, Dix I, et al. 2012. Toxicology ontology perspectives. Altex 29(2):139–156, PMID: 22562487, https://doi.org/10.14573/altex.2012.2.139.