Standardizing Developmental Toxicology Study Extractions Using Automated Application of Ontologies

C. Foster[1], J. Wignall[1], S. Kovach[1], N. Choksi[2], D. Allen[2], J. Trgovcich[1], J. Rochester[1], P. Ceger[2], A. Daniel[2], J. Hamm[2], J. Truax[2], B. Blake[3], B. McIntyre[3], V. Sutherland[3], M. Stout[3], N. Kleinstreuer[4]

[1]ICF, Durham, NC, United States; [2]ILS, RTP, NC, United States;[3]NIH/NIEHS/DNTP, RTP, NC, United States; [4]NIH/NIEHS/DNTP/NICEATM, RTP, NC, United States

**Presenting author*

Extraction of toxicological data from primary sources is a central component of systematic reviews in human health risk assessment. Data obtained from disparate sources must be standardized to ensure that endpoints are identified in a harmonized manner. This standardization process, which can require large labor resources if done manually, is critical to downstream data analyses such as calculating chemical-specific effects, establishing reference datasets for validation of new approaches, and computational modeling purposes.

To expedite the process, and reduce overall level of effort required, we developed a Python script that automates application of pre-existing ontologies and controlled vocabularies to extracted endpoints. Our approach created a harmonized controlled vocabulary crosswalk comprising Unified Medical Language System codes, German Federal Institute for Risk Assessment DevToxDB ontology terms, and Organisation for Economic Co-operation and Development endpoint vocabularies. The crosswalk was applied to roughly 36,000 extractions from prenatal developmental toxicology studies conducted by the National Toxicology Program and 6,400 extractions from prenatal developmental toxicology studies submitted by registrants to the European Chemicals Agency (ECHA).

Our script automatically applied standardized terms to 76% of the NTP extracted endpoints and 60% of the ECHA extracted endpoints. About half (53%) of the standardized terms required manual review after automation to ensure accuracy. Extracted endpoints that were not mapped to standardized terms were too generalized (e.g., “number of fetuses with abnormal organ”) or required human logic to find an adequate match. We estimate that our automated language standardization saved ~375 hours of time while still yielding a valuable computationally accessible dataset. This open-source approach can be applied to other developmental toxicology datasets or customized for other study types to leverage legacy datasets for use in modeling or other analyses.

This project was funded with federal funds from the NIEHS, NIH under Contract No. HHSN273201500010C.