Using Machine-Learning to Facilitate Data Extraction for Human Health Chemical Assessments: Protocol for a case application

Michelle Angrish[1*], Kristina A. Thayer[1], Brittany Schulz[2], Artur Nowak[3], Amanda Persad[1], Allison L. Phillips[4], Glenn Rice[4], Teresa Shannon[4], A. Amina Wilkins[5], Krista Christensen[5], Elizabeth G. Radke[5], Andrew Shapiro[6], Michele M. Taylor[1], Vickie R. Walker[7], Andrew A. Rooney[7], Sean M. Watford[5]

[1] Center for Public Health and Environmental Assessment, Chemical & Pollutant Assessment Division, US Environmental Protection Agency, Durham, NC

[2] Oak Ridge Associated Universities, Environmental Protection Agency National Student Services Contract, Oak Ridge, Tennessee

[3] Evidence Prime Inc., Hamilton, Canada

[4] Center for Public Health and Environmental Assessment, Chemical & Pollutant Assessment Division, US Environmental Protection Agency, Cincinnati, OH

[5] Center for Public Health and Environmental Assessment, Health & Environmental Effects Assessment Division, US Environmental Protection Agency, Washington DC

[6] Center for Public Health and Environmental Assessment, Health and Environmental Effects Assessment Division, US Environmental Protection Agency, Durham, NC

[7] National Institute of Environmental Health Sciences, Division of Translational Toxicology, Durham, NC

Author initials and email address:

Angrish, Michelle (MA) angrish.michelle@epa.gov

Christensen, Krista (KC) christensen.krista@epa.gov

Nowak, Artur (AN) artur.nowak@evidenceprime.com

Persad, Amanda (AP) Persad.amanda@epa.gov

Phillips, Allison (ALP) phillips.allison@epa.gov

Rice, Glen (GR) rice.glenn@epa.gov

Rooney, Andrew (AR) andrew.rooney@nih.gov

Radke, Elizabeth G (ER) radke-farabaugh.elizabeth@epa.gov

Schulz, Brittany (BS) schulz.brittany@epa.gov

Shannon, Teresa (TS) shannon.teresa@epa.gov

Shapiro, Andrew (AS) shapiro.andy@epa.gov
Taylor, Michele (MT) taylor.michelem@epa.gov
Thayer, Kristina (KAT) thayer.kris@epa.gov
Walker, Vickie (VRW) vickie.walker@nih.gov
Watford, Sean (SW) watford.sean@epa.gov
Wilkins, A. Amina (wilkins.amina@epa.gov)

*__Corresponding author:__
Angrish, Michelle (MA) angrish.michelle@epa.gov

Key Words:

systematic review, systematic evidence map, hazard identification, toxicity, artificial intelligence, machine learning, FAIR principles

CRediT Statement:

Michelle Angrish: Conceptualization, Methodology, Formal Analysis, Resources, Supervision, Project Administration, Writing – Original Draft Preparation. **Kristina A. Thayer:** Conceptualization, Supervision, Methodology, Formal Analysis, Project Administration, Writing – Original Draft Preparation. **Brittany Schulz:** Data curation, Visualization, Writing – Reviewing and Editing. **Artur Nowak:** Conceptualization, Methodology, Software, Resources, Validation, Formal Analysis, Writing – Original Draft Preparation. **A. Amina Wilkins** – Methodology, Reviewing and Editing. **Krista Christensen** – Methodology, Reviewing and Editing. **Amanda Persad:** Data Curation, Resources, Writing – Reviewing and Editing. **Allison L. Phillips:** Resources, Writing – Reviewing and Editing. **Elizabeth Radke** – Methodology, Reviewing and Editing. **Glenn Rice:** Supervision, Writing – Reviewing and Editing. **Teresa Shannon:** Resources, Writing – Reviewing and Editing. **Andrew Shapiro:** Conceptualization, Methodology, Software, Resources, Validation, Formal Analysis, Writing – Reviewing and Editing. **Michele Taylor:** Conceptualization, Methodology, Writing – Reviewing and Editing. **Vickie Walker:** Conceptualization, Methodology, Resources, Writing – Reviewing and Editing. **Andrew Rooney:** Conceptualization, Methodology, Resources, Writing – Reviewing and Editing. **Sean Watford:** Conceptualization, Methodology, Software, Resources, Validation, Formal Analysis, Supervision, Project Administration, Writing – Original Draft Preparation.

Acknowledgements:

The authors would like to acknowledge Dale Hoff from US EPA Center for Computational Toxicology and Exposure and Charles Schmitt from National Institute of Environmental Health Sciences for insightful comments on previous versions of this protocol. The authors would also like to acknowledge current US EPA scientists and SRC, Inc. for developing foundational approaches used by the Provisional Peer-Reviewed Toxicity Value program. US EPA scientists include Jay Zhao, Phillip Kaiser, Lucina Lizarraga, Jeff Dean, and Laura Carlson.

Data Availability and Deposition:

The data that support this work are openly available in HAWC and HERO at https://hawc.epa.gov/assessment/100500328/ and https://heronet.epa.gov/heronet/index.cfm/project/page/project_id/3591, respectively.

Funding and Disclaimer:

The authors declare they have no actual or potential competing financial interests. The views expressed are those of the authors and do not necessarily represent the views or policies of the US Environmental Protection Agency. Any mention of trade names, products, or services does not imply an endorsement by the US Government or the US Environmental Protection Agency. Funding was provided by institutional affiliation salary support. Laser AI software and support was procured through order number 68HE0B23P0279.

ABSTRACT

Artificial intelligence (AI) methods including natural language processing, active learning, and large language models are expected to provide workflow advances to reduce risk assessors' time and effort while maintaining the accuracy necessary to meet demand for chemical assessments. A growing suite of modular software applications that integrate AI methods and leverage human-in-the-loop workflows are making operationalization of these advancements feasible. The case application in this protocol supports development of a Provisional Peer-Reviewed Toxicity Value (PPRTV) assessment for 1,3-dinitrobenzene (1,3-DNB). The protocol describes methods to develop a literature inventory and systematic evidence map (SEM) for 1,3-DNB. Along with typical systematic review methods, the protocol applies an active learning approach to screen records at the title and abstract level using AI methods. While active learning has been a routine method used to reduce the resources required to screen records at the title and abstract level, automated processes for data extraction with user verification have evolved slowly. The prolonged evolution of AI for data extraction continues to be a challenge primarily because of insufficient training datasets that limit and or lead to immature models with poor performance. This protocol showcases how software applications like Dextr can be used to address this challenge with the potential to make progress toward a modern workflow stack including data extraction.

INTRODUCTION

Machine learning (ML) methods utilizing natural language processing and active learning have proven invaluable to meet the demand for efficient and rapid development of systematic reviews and systematic evidence maps (SEMs) (Thayer et al. , 2022). Application of artificial intelligence (AI) that utilizes these methods and large language models is expected to address increasing demand for assessments based on systematic methods. Assessments developed utilizing these methods are supported by a growing suite of modular software applications that integrate AI/ML and leverage human-in-the-loop workflows. Operationalization of data, analytics and AI processes greatly facilitates electronic dissemination of information gathered during the course of an assessment in accordance with FAIR principles (Findable, Accessible, Interoperable, Reusable) {Wilkinson, 2016, 5160090}. To date, the ML-assisted workflows implemented with these applications have principally eased the human burden by decreasing the amount of time spent screening studies at the title and abstract (TiAB) level, resulting in a significant reduction in overall time and costs associated with screening {Howard, 2020, 6570105}. Automation of full text screening, study quality evaluation, and data extraction steps, each with user verification (i.e., human-in-the-loop) remain the pinch points in operationalizing ML-assisted steps and there is a need to move away from these costly and complex infrastructures toward modern data stacks and workflows fit for AI. However, the establishment of automated approaches has languished in part due to a lag in the development of training data needed to develop successful natural language processing models for the aforementioned steps and, to a larger extent, due to lack of stackable software applications with the flexibility to test new technology and evolve over time.

The goal of the current case application is to pilot use of an automated data extraction tool with a human-in-the-loop structure (Dextr, https://pubmed.ncbi.nlm.nih.gov/34920276/) into the existing workflow used to conduct an environmental human health assessment. More specifically, Dextr is a web-based data extraction tool that provides a user-verification workflow of ML predictions for data entities pertinent to conducting a human health assessment. The case application is 1,3-dinitrobenzene (1,3-DNB), a chemical slated for a future US Environmental Protection Agency (EPA) Provisional Peer Reviewed Toxicity Value (PPRTV, https://www.epa.gov/pprtv) assessment. PPRTV assessments derive provisional human health toxicity values (e.g., provisional reference values [p-RfVs] and provisional cancer slope factors [p-CSFs]) primarily for use in the US EPA's Superfund program. While the database for PPRTV assessments typically contain relatively few relevant studies (often <30 and mostly animal

toxicology studies as PPRTV chemicals are less commonly studied), data extraction remains a time-consuming and key aspect of assessment development. PPRTV chemicals offer manageable opportunities for testing ML-assisted data extraction for use in developing an environmental human health assessment. Experience gained from this case application can benefit other environmental human health literature assessments exploring use of ML/AI technologies, such as those conducted by the EPA Integrated Risk Information System (IRIS) Program, National Institute for Environmental Health (NIEHS)/Division of Translational Toxicology (DTT), and EPA Office of Chemical Safety and Pollution Prevention (OCSPP).

This protocol outlines a consolidated workflow to screen and extract data from grey and published literature in support of developing a draft PPRTV assessment for 1,3-dinitrobenzene. The workflow incorporates the use of multiple tools with machine learning features and describes data transfer from one step to the next with the goal of data integrity, visibility, and control while operationalizing efficiency through modernization of processes fit for AI and content experts.

METHODS

Populations, Exposures, Comparators, and Outcomes (PECO) Criteria and Supplemental Material Tagging

PECO criteria are used to focus the scope of an evidence map or systematic review by defining the research question(s), search terms, and inclusion/exclusion criteria. The PECO criteria for this case example are presented in Table 1. In addition to PECO-relevant studies, studies that did not meet PECO criteria but contained "potentially relevant" supplemental material are tracked during the literature screening process. Supplemental material was tagged by category, as outlined in Table 2.

Table 1. Populations, Exposures, Comparators, and Outcomes (PECO) criteria

PECO element	Evidence
Populations	**Human:** Any population and lifestage (occupational or general population, including children and other sensitive populations). **Animal:** Nonhuman mammalian animal species (whole organism) of any lifestage (including fetal, early postnatal, adolescents and adults).

PECO element	Evidence
Exposures	**Relevant forms:** "1,3-Dinitrobenzene", "99-65-0" OR "1,3-dinitrobenzene", "1,3-dinitrobenceno", "1,3-dinitrobenzol", "2,4-dinitrobenzene", "3-dinitrobenzene", "benzene, 1,3-dinitro-", "benzene, m-dinitro-", "dinitrobenzene", "m-dinitrobenzene", "meta-dinitrobenzene", "NSC 7189" **Human:** Any exposure to 1,3-dinitrobenzene via oral or inhalation routes. Citations will also be included if biomarkers of exposure are evaluated (e.g., measured chemical or metabolite levels in tissues or bodily fluids) but the exposure route is unclear or likely from multiple routes. Other exposure routes, such as those that are clearly dermal, will be tracked during title and abstract screening and tagged as "Supplemental – Non-PECO route of exposure." **Animal:** Any exposure to 1,3-dinitrobenzene via oral or inhalation routes of ≥14 d duration, or any duration assessing exposure during reproduction or development. Studies involving exposures to mixtures will be included only if they include an experimental arm with exposure to a defined mixture including the chemical of interest. Other exposure routes, including [dermal or injection], are tracked during title and abstract as "Supplemental – Non-PECO route of exposure."
Comparators	**Human:** A comparison or referent population exposed to lower levels (or no exposure/exposure below detection limits), or exposure for shorter periods, or cases versus controls, or a repeated measures design. However, worker surveillance studies are considered to meet PECO criteria even if no statistical analyses using a referent group are presented. Case reports or case series of >3 people will be considered to meet PECO criteria, while case reports describing findings in 1–3 people will be tracked as "supplemental – Case reports or case series l." **Animal:** A concurrent control group exposed to vehicle-only treatment or untreated control (control could be a baseline measurement, e.g., acute toxicity studies of mortality, or a repeated measure design).
Outcomes	All health outcomes (cancer and noncancer). In general, endpoints related to clinical diagnostic criteria, disease outcomes, biochemical, histopathological examination, or other apical/phenotypic outcomes are considered to meet PECO criteria. Mechanistic data are tagged as "Supplemental – Mechanistic endpoints."

1

1 **Table 2. Categories of Supplemental Material in the PPRTV Program[1]**

Category	Description
Pharmacokinetic • **ADME (absorption, distribution, metabolism, and excretion)** • **Classical pharmacokinetic (PK)** • **Physiologically based pharmacokinetic (PBPK) model studies**	The category of pharmacokinetic studies includes ADME (absorption, distribution, metabolism, and excretion) studies, classical pharmacokinetic (PK) or dosimetry model studies, and physiologically based pharmacokinetic (PBPK) or mechanistic dosimetry model studies. ADME studies are primarily controlled experiments, where defined exposures usually occur by intravenous, oral, inhalation, or dermal routes, and the concentration of particles, a chemical, or its metabolites in blood or serum, other body tissues, or excreta are then measured. These data are used to estimate the amount absorbed (A), distributed (D), metabolized (M), or excreted (E). ADME data can also be collected from human subjects who have had environmental or workplace exposures that are not quantified or fully defined. ADME data, especially metabolism and tissue partition coefficient information, can be generated using in vitro model systems. Although in vitro data may not be as definitive as in vivo data, these studies should also be tracked as Pharmacokinetic. Classical PK or dosimetry modeling usually divides the body into just one or two compartments, which are not specified by physiology, where movement of a chemical into, between, and out of the compartments is quantified empirically by fitting model parameters to ADME data. PBPK models represent the body as various compartments (e.g., liver, lung, slowly perfused tissue, richly perfused tissue) to quantify the movement of chemicals or particles into and out of the body (compartments) by defined routes of exposure, metabolism, and excretion, and thereby estimate concentrations in blood or target tissues. Pharmacokinetic studies are further subtagged as either in vivo, in vitro, or model. *Studies describing environmental fate and transport or metabolism in bacteria or model systems that are not applicable to humans or animals are not considered relevant and should be excluded.
Mechanistic endpoints	Studies that do not meet PECO criteria but report measurements that inform the biological or chemical events associated with phenotypic effects related to a health outcome. Experimental design may include in vitro, in vivo (by various routes of

Category	Description
	exposure; includes all transgenic models), ex vivo, and in silico studies in mammalian and nonmammalian model systems. Studies using new approach methodologies (NAMs; e.g., high throughput testing strategies, read-across applications) are also categorized here. Studies where the chemical is used as a laboratory reagent (e.g., as a chemical probe used to measure antibody response) generally are not considered relevant and should be excluded. For PPRTV assessments, genotoxicity mechanistic studies are specifically sub-tagged. *Several chemicals that don't meet PECO for 1,3-dinitrobenzene are commonly used as model haptens to explore aspects of mechanisms of immune function, allergenicity, and dermal sensitization [e.g., dinitrochlorobenzene (DNCB), dinitrofluorobenzene (DNFB)]. Abstract were excluded that described the use of a specific immune/allergenicity/dermal sensitization assay tested on reference compounds but did not specifically mention 1,3-dinitrobenzene.
Non-PECO route of exposure	Epidemiological or animal studies that use a non-PECO route of exposure, (e.g., injection studies or dermal studies if the dermal route is not part of the exposure criteria). *This categorization generally does not apply to epidemiological studies where the exposure route is unclear; such studies are considered to meet PECO criteria if the relevant route(s) of exposure are plausible, with exposure being more thoroughly evaluated at later steps.
Non-PECO exposure duration	PPRTV assessments focus on subchronic and chronic exposure durations. Thus, short-term and acute exposure durations (defined as animal studies of less than 14 d in duration) are considered supplemental.
Case reports or case series	Human studies that present an investigation of a single exposed individual or group of ≤3 subjects that describe health outcomes after exposure but lack a comparison group (i.e., do not meet the "C" in the PECO criteria) and typically do not include reliable exposure estimates.
Records with no original data	Records that do not contain original data, such as other regulatory agency assessments, informative scientific literature reviews

[1]This approach is a modification of the tagging scheme for supplemental material in EPA IRIS assessments {U.S. EPA, 2022, 10367891}. This scaled back tagging scheme reflects the data poor nature and reduced evidence base

Category	Description
complexity of PPRTV assessments in comparison to IRIS assessments, e.g., PBPK modeling is not pursued in PPRTV assessments.	

Literature Search and Screening Strategies

Database Search Term Development

The literature search focuses only on chemical name (and synonyms) with no additional limits. Chemical synonyms are identified through EPA's CompTox Chemicals Dashboard. A chemical search using the CAS registration number and any synonyms indicated as "valid" or "good" were included in the search strings . The PubMed search is shared with information specialists to develop search strategies tailored for each of the databases below because each database has its own search syntax.

Literature search updates are performed periodically during draft development, with the last full literature search update conducted within six months prior to the planned release of the finalized document. The results of the last literature update are screened against the PECO criteria and citation information available in HERO (see HERO description below). Newly identified studies are only incorporated into the assessment if they impact toxicity values.

Database Searches and Filtering for Human Health Records

An information specialist searched PubMed, Web of Science, and Scopus through May 2023, and the results were stored and are maintained in the Health and Environmental Research Online (HERO) database 1,3-dinitrobenzene. The HERO database is used to provide access to the references used in US EPA's scientific assessments, including this effort. Full details of the search strategy are presented in Supplemental Material, Appendix A. The search will be updated during the conduct of the case application with the last pre-manuscript submission update occurring ~ one month ahead of anticipated submission.

The records returned undergo deduplication in HERO using unique identifiers (e.g., PMID, WoSID, or DOI) and citation content. Following deduplication of the records, SWIFT-Review software (https://www.sciome.com/swift-review/searchstrategies/) is used to prioritize references for screening, based on use of the pre-set literature search strategies ("filters"): "animal (human health models)", "human", and "in vitro". These strategies have been developed by

information specialists and can be applied to distinguish studies containing human health-relevant information from studies with less relevant content (e.g., environmental fate).

Studies not retrieved or prioritized using the search strategies are not considered further. Studies that include one or more of the search terms in the title, abstract, keyword, or medical subject headings (MeSH) fields are exported as a RIS file for screening as described below.

Other Resources Consulted for Primary Research Reports

The literature search strategies described above are designed to be broad, but like any search strategy, studies may be missed (e.g., cases where the specific chemical is not mentioned in title, abstract, or keyword content; ability to capture "gray" literature that is not indexed in the databases listed above). Thus, in addition to the database searches, the resources in Table 3 are used to identify studies that may have been missed based on the database search. These sources are searched using customized processes described (Supplemental Material, Appendix B). References that appear to meet the PECO criteria are uploaded into the screening software, annotated with respect to source of the record, and screened as described below according to PECO. Searching of these sources is summarized to include the source type or name, the search string (when applicable), the URL (when available and applicable), number of results, and number of unique references not otherwise identified from database searching. To identify unique references, a citation for each identified study is generated in HERO and verified that it is not already identified from the database searches (e.g., PubMed, WoS, etc) prior to moving forward to screening. Note, the Defense Technical Information Center (DTIC) is searched only for compounds pertinent to the Department of Defense. Therefore, 1,3-dintrobenzene was considered DOD-relevant due to its use in making explosives.

Table 3. Grey Literature Sources for Primary Research Reports

Resource	Comments
Bibliographies of included studies (studies meeting PECO)	Manual review (at the title/abstract) level of reference list in studies screened as PECO-relevant after full-text review.
Reference list from human health focused assessments	Review of the reference list from final or publicly available draft or finalized assessments (e.g., EPA IRIS [Integrated Risk Information System], ATSDR [Agency for Toxic Substances and Disease Registry] Toxicological Profile) or from published

Resource	Comments
	journal review specifically focused on human health. Reviews and assessments can be identified from the database search or surveys for existing assessments (Table 4).
References identified by technical consultants or during peer-review	
AEGLs	Acute Exposure Guideline Levels for Airborne Chemicals (AEGL) https://www.epa.gov/aegl/access-acute-exposure-guideline-levels-aegls-values#chemicals
AICIS	https://services.industrialchemicals.gov.au/search-assessments/
ChemView	EPA ChemView database to identify unpublished studies, information submitted to EPA under Toxic Substances Control Act (TSCA) Section 4 (chemical testing results), Section 8(d) (health and safety studies), Section 8(e) (substantial risk of injury to health or the environment notices), and FYI (For Your Information, voluntary documents). Other databases accessible via ChemView include EPA's High Production Volume (HPV) Challenge database and the Toxic Release Inventory database.
DTIC	Defense Technical Information Center. Searched if topic is pertinent to the Department of Defense
ECETOC	http://www.ecetoc.org/publications
ECHA	European Chemicals Agency (ECHA) registration dossiers to identify data submitted by registrants. Registration dossiers contain data on substances such as hazardous properties, safe uses, classifications, environmental fate, and ecotoxicological and toxicological information. The amount of information provided for each substance varies and is obtained directly from companies' REACH registrations. ECHA does not give any guarantees or warranties regarding the quality and correctness of the published information. The information in the portal is published 'as provided' by industry, and its accuracy has not been verified by ECHA (https://echa.europa.eu/information-on-chemicals/registered-substances).
JECDB	http://dra4.nihs.go.jp/mhlw_data/jsp/SearchPageENG.jsp

Resource	Comments
NTP	National Toxicology Program (NTP) toxicology testing results and literature analysis reports. https://ntp.niehs.nih.gov/
NTRL	https://ntrl.ntis.gov/NTRL/
OECD	http://webnet.oecd.org/hpv/ui/Search.aspx The *Organisation* for Economic Cooperation and Development (OECD) eChemPortal to retrieve results for OECD Screening Information DataSet (SIDS) and High Production Volume (HPV) Chemicals (https://www.echemportal.org/echemportal/).
US EPA's CompTox Chemicals Dashboard ToxValDB	References from US EPA's CompTox Chemicals Dashboard ToxValDB (Toxicity Values Database) to identify studies or assessments that present point of departure (POD) information. ToxValDB collates publicly available toxicity dose-effect related summary values typically used in risk assessments. Many of the PODs presented in ToxValDB are based on gray literature studies or assessments not available in databases such as PubMed, WoS, etc. It is important to note that ToxValDB entries have not undergone quality control to ensure accuracy or completeness and may not include recent studies.
EFSA	http://www.efsa.europa.eu/

Survey of Previous Assessments, Regulatory Reference Values, Risk Thresholds or Assessment Based Points of Departure

"Toxicity value" is a broad term that encompasses reference values and cancer risk estimates (i.e., slope factors and unit risk estimates). The term reference value applies to values designed to provide a "benchmark" or exposure limit below which adverse effects on human health are not expected to occur. Reference values are the most common final output from the dose response assessment component of the risk assessment paradigm set forth by the National Research Council and are based on an observed or estimated threshold for an effect, usually noncancer. Table 4 presents a list of organizations that disseminate toxicity values queried to conduct a survey.

Table 4. Sources Searched for Existing Human Health Reference Values and Cancer Descriptors

Source	Query and/or link
ACGIH	ACGIH. 2007. 2007 TLVs and BEIs: Based on documentation of the threshold limit values for chemical substances and physical agents and biological exposure indices. Cincinnati, OH: American Conference of Governmental Industrial Hygienists.
AIHA	AIHA. 2019. 2019 ERPG/WEEL Handbook. Fairfax, VA: American Industrial Hygiene Association. [List of values.] AIHA. 2002 (and updates). 2002 Emergency Response Planning Guidelines. Fairfax, VA: American Industrial Hygiene Association. [Details used in deriving values.]
ATSDR	http://www.atsdr.cdc.gov/toxprofiles/index.asp
	https://wwwn.cdc.gov/TSP/MRLS/mrlsListing.aspx
US EPA's CompTox Chemicals Dashboard ToxValDB	https://comptox.epa.gov/dashboard *Note: ToxValDB collates publicly available toxicity dose-effect related summary values typically used in chemical assessments. Many of the PODs presented in ToxVal are based on gray literature studies or assessments not available in databases such as PubMed and Web of Science (WoS). Although many of the resources included in this table are represented in ToxValDB, they are also manually searched because most of the ToxValDB entries have not undergone quality control to ensure accuracy or completeness and might not include recent studies*.

Source	Query and/or link
CT DEEP	https://eregulations.ct.gov/eRegsPortal/Browse/getDocument?guid={00D6A654-0300-CC47-9B95-397D2AD21304}
DFG	https://onlinelibrary.wiley.com/doi/10.1002/9783527826889.oth
OW (US EPA)	https://www.epa.gov/system/files/documents/2022-01/dwtable2018.pdf
	https://www.epa.gov/dwstandardsregulations/drinking-water-contaminant-human-health-effects-information#dw-standards
EPA/NRC AEGL	https://www.epa.gov/aegl/access-acute-exposure-guideline-levels-aegls-values#chemicals
European Commission	https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32017L0164&from=EN
Health Canada	https://www.canada.ca/en/services/health/publications/healthy-living.html
	https://publications.gc.ca/collections/collection_2021/sc-hc/H129-108-2021-eng.pdf
	http://publications.gc.ca/site/archivee-archived.html?url=http://publications.gc.ca/collections/Collection/H46-2-96-194E.pdf
HEAST	https://epa-heast.ornl.gov/heast.php
	https://cfpub.epa.gov/ncea/risk/hhra/recordisplay.cfm?deid=2877
HSA	https://www.hsa.ie/eng/publications_and_forms/publications/chemical_and_hazardous_substances/chemical_agents_and_carcinogens_code_of_practice_2021.html

Source	Query and/or link
IARC	https://monographs.iarc.who.int/monographs-available/
	http://monographs.iarc.fr/ENG/Classification/List_of_Classifications.pdf
IDEM	https://www.in.gov/idem/toxic/2343.htm
ID DEQ	https://adminrules.idaho.gov/rules/current/58/580101.pdf
IFA	https://limitvalue.ifa.dguv.de/WebForm_gw2.aspx
IPCS	https://www.inchem.org/pages/cicads.html
IRIS	http://www.epa.gov/iris/
JSOH	https://www.sanei.or.jp/?mode=view&cid=328
MassDEP	https://www.mass.gov/service-details/massdep-ambient-air-toxics-guidelines
MDH	https://www.health.state.mn.us/communities/environment/risk/guidance/air/table.html
MI EGLE	https://www.michigan.gov/documents/deq/deq-rrd-chem-CleanupCriteriaTSD_527410_7.pdf
NATICH	https://nepis.epa.gov/Exe/ZyPDF.cgi/2000NS7S.PDF?Dockey=2000NS7S.PDF
NC DEQ	https://files.nc.gov/ncdeq/Air%20Quality/rules/rules/D1104.pdf
NDEP	https://ndep.nv.gov/resources/risk-assessment-and-toxicology-basic-comparison-levels

Source	Query and/or link
NIOSH	http://www.cdc.gov/niosh/npg/npgdcas.html
	https://www.cdc.gov/niosh/pubs/criteria_date_desc_nopubnumbers.html
	https://www.cdc.gov/niosh/idlh/intridl4.html
NJ DEP	https://dep.nj.gov/boss/risk-screening-tools/
NYSDEC	https://www.dec.ny.gov/docs/remediation_hudson_pdf/techsuppdoc.pdf
OAQPS (US EPA)	https://www.epa.gov/fera/dose response-assessment-assessing-health-risks-associated-exposure-hazardous-air-pollutants
OEHHA	http://www.oehha.ca.gov/tcdb/index.asp
Ontario MOL	https://www.labour.gov.on.ca/english/hs/pubs/oel_table.php
OPP (US EPA)	https://iaspub.epa.gov/apex/pesticides/f?p=chemicalsearch:1
OR DEQ	https://secure.sos.state.or.us/oard/displayDivisionRules.action?selectedDivision=1556
OSHA	https://www.osha.gov/chemicaldata/
PAC Database	https://pacteels.pnnl.gov/#/
PPRTV	https://www.epa.gov/pprtv/provisional-peer-reviewed-toxicity-values-pprtvs-assessments

Source	Query and/or link
Publications Quebec	http://legisquebec.gouv.qc.ca/en/showdoc/cr/S-2.1,%20r.%2013?csi_scan_9222d36c6a354dc6=BO9xyrMZ+270UP3j0MGuOD0kZjgFAAAAXrM3HA==&bcsi_scan_filename=S-2.1,%20r.%2013&bcsi_scan_9222d36c6a354dc6=KXzmpPueuN0L1AjnJOB1Zerr85YMAAAAyhrPTg==&bcsi_scan_filename=S-2.1,%20r.%2013
RI DEM	http://www.dem.ri.gov/programs/benviron/air/pdf/airtoxgl.pdf
RIVM	https://www.rivm.nl/bibliotheek/rapporten/711701092.pdf
	https://www.rivm.nl/bibliotheek/rapporten/609021044.pdf
	https://www.rivm.nl/bibliotheek/rapporten/711701025.pdf
Safe Work Australia	https://www.safeworkaustralia.gov.au/exposure-standards#exposure-standards-in-australia
SWCAA	http://www.swcleanair.org
TCEQ	https://www.tceq.texas.gov/toxicology/dsd/final
	https://www.tceq.texas.gov/remediation/trrp/trrppcls.html
USAPHC	https://phc.amedd.army.mil/topics/envirohealth/hrasm/Pages/TG230.aspx
VT DEC	https://dec.vermont.gov/sites/dec/files/aqc/laws-regs/documents/AQCD%20Regulations%20ADOPTED_Dec132018.pdf#page=127
WAC	https://apps.leg.wa.gov/WAC/default.aspx?cite=173-460-150

Source	Query and/or link
WHO	https://inchem.org/pages/ehc.html
Worksafe	https://worksafe.govt.nz/topic-and-industry/work-related-health/monitoring/exposure-standards-and-biological-exposure-indices/

ACGIH = American Conference of Governmental Industrial Hygienists; AEGL = Acute Exposure Guideline Levels; AIHA = American Industrial Hygiene Association; ATSDR = Agency for Toxic Substances and Disease Registry; BEI = biological exposure index; CT DEEP = Connecticut Department of Energy & Environmental Protection; DFG = Deutsche Forschungsgemeinschaft, German Research Foundation; EPA = Environmental Protection Agency; ERGP = emergency response planning guideline; HEAST = Health Effects Assessment Summary Tables; HSA = Health and Safety Authority; IARC = International Agency for Research on Cancer; IDEM = Indiana Department of Environmental Management; ID DEQ = Idaho Department of Environmental Quality; IFA = Institut für Arbeitsschutz, The Institute for Occupational Safety and Health; IPCS = International Programme on Chemical Safety; IRIS = Integrated Risk Information System; JSOH = Japan Society for Occupational Health; MassDEP = Massachusetts Department of Environmental Protection; MDH = Minnesota Department of Health; MI EGLE = Michigan Environment, Great Lakes & Energy; MOL = Ministry of Labour; NATICH = National Air Toxics Information Clearinghouse; NC DEQ = North Carolina Department of Environmental Quality; NDEP = Nevada Division of Environmental Protection; NIOSH = National Institute for Occupational Safety and Health; NJ DEP = New Jersey Department of Environmental Protection; NRC = National Research Council; NYSDEC = New York State Department of Environmental Conservation; OAQPS = Office of Air Quality Planning and Standards; OEHHA = California Office of Environmental Health Hazard Assessment; OEL = occupational exposure level; OPP = Office of Pesticide Programs; OR DEQ = Oregon Department of Environmental Quality; OSHA = Occupational Safety and Health Administration; OW = Office of Water; PAC = Protective Action Criteria; PPRTV = Provisional Peer-Reviewed Toxicity Value; RfD = reference dose; RI DEM = Rhode Island Department of Environmental Management; RIVM = Rijksinstituut voor Volksgezondheid en Milieu, The Netherlands Institute for Public Health and the Environment; RSL = regional screening level; SWCAA = Southwest Clean Air Association; TCEQ = Texas Commission on Environmental Quality; TEEL = temporary emergency exposure limit; TLV = threshold limit value; TWA = time-weighted average; USAPHC = United States Army Public Health Center;

VT DEC = Vermont Department of Environmental Conservation; WAC = Washington Administrative Code; WEEL = workplace environment exposure level; WHO = World Health Organization.

Screening and Tagging Process

As described below, two different software applications are used for screening. Regardless of the software application, both TiAB and full-text screening are conducted by two independent reviewers and any conflicts in screening are resolved by discussion between the two independent reviewers; a third reviewer is consulted if any conflicts remained thereafter. Conflicts between screeners in applying the supplemental tags are resolved by discussion at both the TiAB and full-text levels, erring on the side of over-tagging at the TiAB level. At the TiAB level, articles without an abstract are screened based on title (title should indicate clear relevance) and number of pages (articles two pages in length or less are assumed to be abstracts or short communications rather than full study reports and tracked as supplemental).

The records identified from the evidence streams "animal (human health models)", "human", and "in vitro" in SWIFT-Review are deduplicated by "Title" and imported into SWIFT-Active Screener, for TiAB screening. SWIFT-Active Screener is a web-based collaborative software application that utilizes active learning approaches to reduce the screening effort. The screening process is designed to prioritize records that appear to meet PECO criteria or include supplemental material content based on TiAB content (i.e., both types of records were screened as "include" for active learning purposes). Key words are used to facilitate the screening process (Supplemental Material, Appendix A). Records are screened in SWIFT-Active Screener at the TiAB level until the software indicates a likelihood of 95% that all relevant records had been captured. This threshold is comparable to human error rates and is used as a metric to evaluate machine-learning performance (Bannach-Brown et. al. 2018; Howard et al., 2016; Cohen et al., 2006). Any records in "partially screened" status at the time of reaching the 95% threshold are considered fully screened by accepting the predicted decision from the model.

All TiAB screening decisions (human and machine) made in SWIFT Active are imported into US EPA's version of the Health Assessment Workspace Collaborative (HAWC), a free and open source web-based software application that facilitates the management of assessments for environmental pollutants. The TiAB-specific tags are used to separate records meeting PECO criteria (which undergo full-text review in HAWC) versus supplemental content (Table 2) and to tag certain specific categories of supplemental content (pharmacokinetic, genotoxicity, non-PECO route, non-PECO exposure duration, case reports or series, and records with no original data). For records meeting PECO criteria at the TiAB level, full text articles are retrieved through US EPA's HERO database and linked to the record in HAWC. Most tagging of supplemental content occurs during TiAB screening, but tagging can also be done during full-text screening for studies that meet PECO criteria and also contain supplemental content. References that are not able to be retrieved within 45 days are not considered further. Records identified via the gray literature searches are imported directly into HAWC where screening begins at the TiAB phase.

HAWC is used for full-text review. The full text articles found for records meeting PECO are available in HERO, which is linked to the record in HAWC. For full-text review, the methods and results sections from records are used to further determine PECO-relevance and to apply more tags to the record. Like TiAB level review, two reviewers independently determine inclusion with conflicts resolved in the same manner.

Literature Inventory

Extraction Fields

Records meeting PECO criteria after full-text review are imported into Dextr (Walker et al., 2022). HAWC-compatible data extraction forms are available in Dextr for mammalian bioassay and epidemiology data extraction. This means that the data extracted into

Dextr can be easily imported into HAWC. For mammalian bioassay studies, the following study summary information was captured in a literature inventory: study type [acute (< 24 hours), short term (1–30 days), subchronic (>30–90 days), chronic (>90 days), developmental, peripubertal, multigenerational], route of exposure, species, sex, and health/organ system(s) (i.e. respiratory, cardiovascular) assessed. Additional terminology resources (such as picklists and controlled vocabularies) were used to facilitate standardization of the author reported extracted data (see Appendix C, Tables S-2-4). Endpoints reported from experimental data were mapped to the Environmental Health Vocabulary (EHV). The EHV comprises standardized endpoint/outcome terminologies. For epidemiology studies, the following study summary information was captured in a literature inventory: sex, population, country, study design, exposure measurement (e.g., blood, feces), and health system(s) assessed. The full list of fields for both sets of forms including definitions and available picklist options is available in (Appendix C, Supplemental Tables 2-4).

Machine-learning (ML) models can be attached to fields in the Dextr forms for automated extraction. When the full text for a record is imported, the full text is processed, and the ML models will populate corresponding form fields. A primary extractor will review the model predictions to verify, edit, or delete the model predictions. The primary extractor will also extract data into the fields that were not automatically populated by a model and fields that did not have a model attached to it. Using the quality assurance (QA) review module in Dextr, extracted data is then reviewed by a QA reviewer for completeness and accuracy. The QA module in Dextr provides the primary extraction data to the QA reviewer for review with the ability to delete, edit, or add extraction information while conducting the quality check step.

The data extractions are exported from Dexter using the export feature. The HAWC client is used to import the datasets into the HAWC project (HAWC, 2023). Evidence maps displaying record counts for any combination of fields are available as visualizations in HAWC. All data are also available for download.

Evaluations of Workflow and Model Performance

The aim of this evaluation is to understand how the integration of Dextr would perform in an assessment workflow (e.g. PPRTV) and to also understand specifically how the automated extraction recall, precision, and extraction time compared to manual extraction. The evaluation plan described below is similar to the approaches previously described by (Walker et al., 2022; Nowak and Kunstman, 2018).

Data Extraction

Data extraction is divided into two study arms: manual data extraction (ME) and human-in-the-loop data extraction (HE) (Figure 1). Every reference will be extracted twice. In the ME arm, one of the data extractors reads each study and manually extracts text fields. In the ME arm, extractors review the model-generated extractions, makes corrections to the model-generated extractions, and add missing ones as needed. The results of extractions from ME and HE are then reviewed by the same quality assurance (QA) reviewer. The corrections applied by the QA reviewer are used to calculate accuracy metrics. QA reviewers evaluate both results separately because we expect that some of the differences in the results across study arms will need to be assessed manually to determine if one of the results is incorrect or if they constitute alternative correct answers. The QA reviewers will be blinded in regards to study arm allocation. To control for differences between extractors' performance, the same extractors participate both in ME and HE study arms. For each extractor, the references assigned to them will be randomized across ME and HE in 1:1 ratio. No extractor will extract or review the same reference across both arms. Each study arm has a separate Dextr project with identical data extraction forms and fields with picklists enabled. The ME project will not have any models attached to the fields in the data extraction forms. Dextr will track all user metrics including time to complete extraction. Team members are instructed to not remain idle while signed into Dextr while completing data extractions to ensure

accurate time to complete extractions is logged. Additionally, the user action logs from Dextr will be filtered for any periods of inactivity longer than 5 minutes and excluded from the calculation of time metrics.

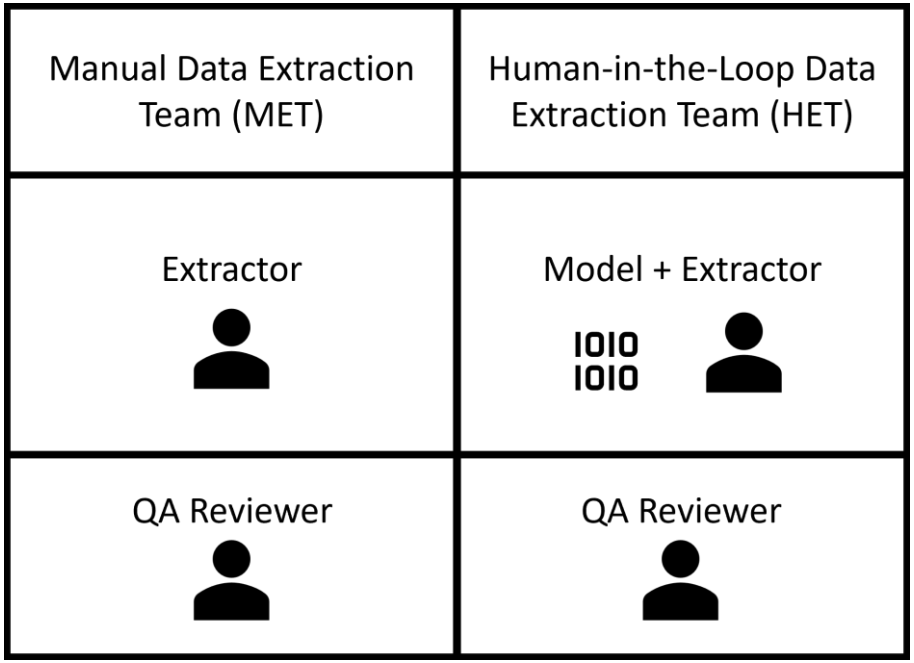

Figure 1: Data Extraction Teams. Two data extractors willwill independently extract data from references into forms using Dextr. The manual data extraction (ME) will manually extract all information while the human-in-the-loop data extraction (HE) will manually extract all non-automated fields, while reviewing and correcting the model-generated

extraction. The same quality assurance (QA) reviewer will then correct both results for a given study to ensure accuracy of data extractions andand calculate quality metrics.

Dextr Model Performance

Results will be calculated by automatically comparing the extracted data after primary extraction and QA. A collection of metrics is generated including, but not limited to, precision, recall, and F1 score. Precision is defined as the fraction of correct data points from the primary extraction among all the primary extraction data points. Recall is defined as the fraction of correct data points from primary extraction among all instances of a fields being correctly populated by either the primary extraction or the QA step F1 score is defined as the harmonic mean of precision and recall.

The manually curated results can be incorporated back into the models to refine the model's performance and then redeployed in Dextr. In subsequent projects, the new model performance can be reevaluated. Over time, it is expected that model performance will improve given that more data are available for training.

User metrics for data extraction across platforms

User metrics include the number of fields for each form, study type, and time spent reviewing a data extraction form. User metrics will be collected from Dextr. Analyses will be performed with the user metrics compared to past projects. Each extractor and reviewer will report on user experience while using Dextr both for user experience and user interactions.

REFERENCES

1. Bannach-Brown, A; Przybyła, P; Thomas, J; Rice, ASC; Ananiadou, S; Liao, J; Macleod, MR. (2018). Machine learning algorithms for systematic review: reducing workload in a preclinical review of animal studies and reducing human screening error. bioRxiv (pp. 1-16). https://doi.org/10.1101/255760.
2. Cohen, AM; Hersh, WR; Peterson, K; Yen, PY. (2006). Reducing workload in systematic review preparation using automated citation classification. Journal of the American Medical Informatics Association 12: 206-219. http://dx.doi.org/10.1197/jamia.M1929.
3. HAWC. (2023) Python HAWC Client. Technical Report. https://hawc.readthedocs.io/client/#python-hawc-client
4. Howard, BE; Phillips, J; Miller, K; Tandon, A; Mav, D; Shah, MR; Holmgren, S; Pelch, KE; Walker, V; Rooney, AA; Macleod, M; Shah, RR; Thayer, K. (2016). SWIFT-Review: A text-mining workbench for systematic review. Systematic Reviews 5:87. http://dx.doi.org/10.1186/s13643-016-0263-z
5. Howard, BE; Phillips, J; Tandon, A; Maharana, A; Elmore, R; Mav, D; Sedykh, A; Thayer, K; Merrick, BA; Walker, V; Rooney, A; Shah, RR. (2020). SWIFT-Active Screener: Accelerated document screening through active learning and integrated recall estimation. Environment International 138: 105623. https://doi.org/10.1016/j.envint.2020.105623.
6. Kristina A. Thayer, Rachel M. Shaffer, Michelle Angrish, Xabier Arzuaga, Laura M. Carlson, Allen Davis, Laura Dishaw, Ingrid Druwe, Catherine Gibbons, Barbara Glenn, Ryan Jones, J. Phillip Kaiser, Channa Keshava, Nagalakshmi Keshava, Andrew Kraft, Lucina Lizarraga, Kristan Markey, Amanda Persad, Elizabeth G Radke, Glenn Rice, Brittany Schulz, Teresa Shannon, Andrew Shapiro, Shane Thacker, Suryanarayana Vulimiri, George Woodall, Erin Yost. (2022). Use of systematic evidence maps within the US Environmental Protection Agency (EPA) Integrated Risk Information System (IRIS) program: Advancements to date and looking ahead. Environment International 169:107363. https://doi.org/10.1016/j.envint.2022.107363.

7. Nowak, A; Kunstman, P. (2018). Team EP at TAC 2018: Automating data extraction in systematic reviews of environmental agents. Technical Report. Cornell University Library, arXiv.org. Ithaca, NY. https://doi.org/10.48550/arXiv.1901.02081.
8. U.S. EPA. (U.S. Environmental Protection Agency). (2022). ORD staff handbook for developing IRIS assessments. (EPA 600/R-22/268) Washington, DC. https://cfpub.epa.gov/ncea/iris_drafts/recordisplay.cfm?deid=356370
9. U.S. EPA. (U.S. Environmental Protection Agency). Health & Environmental Research Online (HERO). [EPA Report]. Washington, DC. https://heronet.epa.gov/heronet/index.cfm/content/home
10. Walker, VR; Schmitt, CP; Wolfe, MS; Nowak, AJ; Kulesza, K; Williams, AR; Shin, R; Cohen, J; Burch, D; Stout, MD; Shipkowski, KA; Rooney, AA. (2022). Evaluation of a semi-automated data extraction tool for public health literature-based reviews: Dextr. Environment International 159: 107025. https://doi.org/10.1016/j.envint.2021.107025.
11. Wilkinson, MD; Dumontier, M; Aalbersberg, IJ; Appleton, G; Axton, M; Baak, A; Blomberg, N; Boiten, JW; da Silva Santos, LB; Bourne, PE; Bouwman, J; Brookes, AJ; Clark, T; Crosas, M; Dillo, I; Dumon, O; Edmunds, S; Evelo, CT; Finkers, R; Gonzalez-Beltran, A; Gray, AJ; Groth, P; Goble, C; Grethe, JS; Heringa, J; 'T Hoen, PA; Hooft, R; Kuhn, T; Kok, R; Kok, J; Lusher, SJ; Martone, ME; Mons, A; Packer, AL; Persson, B; Rocca-Serra, P; Roos, M; van Schaik, R; Sansone, SA; Schultes, E; Sengstag, T; Slater, T; Strawn, G; Swertz, MA; Thompson, M; van Der Lei, J; van Mulligen, E; Velterop, J; Waagmeester, A; Wittenburg, P; Wolstencroft, K; Zhao, J; Mons, B. (2016). Scientific Data 3:160018. http://dx.doi.org/10.1038/sdata.2016.18.

1