Contents lists available at ScienceDirect

Environment International

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

Evaluation of a semi-automated data extraction tool for public health literature-based reviews: Dextr

Vickie R. Walker[a,*], Charles P. Schmitt[a], Mary S. Wolfe[a], Artur J. Nowak[b], Kuba Kulesza[b], Ashley R. Williams[c], Rob Shin[c], Jonathan Cohen[c], Dave Burch[c], Matthew D. Stout[a], Kelly A. Shipkowski[a], Andrew A. Rooney[a]

[a] *Division of the National Toxicology Program (DNTP), National Institute of Environmental Health Sciences (NIEHS), National Institutes of Health (NIH), Research Triangle Park, NC, USA*

[b] *Evidence Prime Inc, Krakow, Poland*

[c] *ICF, Research Triangle Park, NC, USA*

ARTICLE INFO

Handling Editor: Paul Whaley

Keywords:
Automation
Text mining
Machine learning
Natural language processing
Literature review
Systematic review
Scoping review
Systematic evidence map

ABSTRACT

Introduction: There has been limited development and uptake of machine-learning methods to automate data extraction for literature-based assessments. Although advanced extraction approaches have been applied to some clinical research reviews, existing methods are not well suited for addressing toxicology or environmental health questions due to unique data needs to support reviews in these fields.

Objectives: To develop and evaluate a flexible, web-based tool for semi-automated data extraction that: 1) makes data extraction predictions with user verification, 2) integrates token-level annotations, and 3) connects extracted entities to support hierarchical data extraction.

Methods: Dextr was developed with Agile software methodology using a two-team approach. The development team outlined proposed features and coded the software. The advisory team guided developers and evaluated Dextr's performance on precision, recall, and extraction time by comparing a manual extraction workflow to a semi-automated extraction workflow using a dataset of 51 environmental health animal studies.

Results: The semi-automated workflow did not appear to affect precision rate (96.0% vs. 95.4% manual, $p = 0.38$), resulted in a small reduction in recall rate (91.8% vs. 97.0% manual, $p < 0.01$), and substantially reduced the median extraction time (436 s vs. 933 s per study manual, $p < 0.01$) compared to a manual workflow.

Discussion: Dextr provides similar performance to manual extraction in terms of recall and precision and greatly reduces data extraction time. Unlike other tools, Dextr provides the ability to extract complex concepts (e.g., multiple experiments with various exposures and doses within a single study), properly connect the extracted elements within a study, and effectively limit the work required by researchers to generate machine-readable, annotated exports. The Dextr tool addresses data-extraction challenges associated with environmental health sciences literature with a simple user interface, incorporates the key capabilities of user verification and entity connecting, provides a platform for further automation developments, and has the potential to improve data extraction for literature reviews in this and other fields.

1. Introduction

Systematic review methodology is a rigorous approach to literature-based assessments that maximizes transparency and minimizes bias (O'Connor et al. 2019). Three main assessment formats (systematic reviews, scoping reviews, and systematic evidence maps) use these methods in a fit-for-purpose approach depending on the research question and project goals. Systematic reviews follow a pre-defined protocol to identify, select, critically assess, synthesize, and integrate evidence to answer a specific question and reach conclusions. The best

Abbreviations: SR, Systematic review; SCR, Scoping review; SEM, Systematic evidence map.

* Corresponding author at: NIEHS, P.O. Box 12233, Mail Drop K2–04, Research Triangle Park, NC 27709, USA. Express mail: 530 Davis Drive, Morrisville, NC 27560, USA.

E-mail address: Walker.Vickie@nih.gov (V.R. Walker).

https://doi.org/10.1016/j.envint.2021.107025
Received 1 July 2021; Received in revised form 7 October 2021; Accepted 3 December 2021

Box 1

A comparison of the literature-assessment steps between scoping reviews/systematic evidence maps and systematic reviews[1].

Literature- Assessment Steps		Task Description in Scoping Review (SCR) or Systematic Evidence Mapping (SEM) Workflow	Task Description in Systematic Review (SR) Workflow	Percent Time[2]
Problem Formulation and Protocol Development		Define research question and objectives, typically with a broad PECO[3], and all methods before conducting review. • Best practice to publish protocol before conducting review for transparency; however, only minor impact on bias • Objectives are often open questions to survey broad topics and identify extent of evidence (i.e., areas that are data rich or data poor / data gaps) • Protocol should describe key concepts that will be mapped (e.g., exposures) to support objectives	Define research question, PECO, and all methods before conducting review to reduce bias. • Best practice to publish protocol for transparency • Critical to publish protocol before starting evidence evaluation to reduce bias • Objectives are focused, closed questions (e.g., specific exposure outcome pairs/ hazards)	8%
Identify the Evidence	**Identify Literature**	Develop search strategy to identify evidence relevant to address the question. • Search is biased to address the degree of precision and certainty of the objectives where a comprehensive search may not be necessary • Searches conducted in one or more major literature database, or in a stepwise manner, to address objectives • Search terms are generally broad / topic based with lower specificity • Searches retrieve evidence supportive of multiple decisions and scenarios	Develop comprehensive search strategy to identify all relevant evidence to address the question. • Search is biased toward maximum number of sources to ensure identification of all evidence relevant to synthesis • Search includes literature databases, sources of grey literature, and published data • Search terms are highly resolved and specified for key elements of the objectives	7%
	Screen Studies	Screen studies against eligibility criteria from objectives and PECO. • Inclusion and exclusion criteria are topic-based and may only address PECO at a high level • Included studies likely to address diverse scenarios	Screen studies against eligibility criteria from objectives and PECO. • Inclusion and exclusion criteria specified in detail for all PECO elements • Assure specific research question is efficiently addressed	17%
	Extract Data	Extract study *meta*-data and characteristics to address objectives. • Flexible approach supports fit-for-purpose maps of varying degrees of comprehensiveness • Optional extraction of study findings and other characteristics depending on objectives	Complete extraction of *meta*-data and results to address question. • Entities determined by project objectives	15%
Evaluate the Evidence		Appraisal of studies is optional depending on objectives. • Study characteristics relating to quality of study design and conduct, or internal validity, may be extracted • May include stepwise approach (e.g., methods mapped relative to objectives), or quality only assessed for studies addressing key outcomes	Critical appraisal of included studies is essential to characterizing certainty in bodies of evidence. • Performed as assessment of internal validity (risk of bias) • May include external validity, sensitivity, other factors	9%
Summarize and Synthesize Data		SCRs and SEMs have limited or no synthesis – may only include summaries. • Primary output shows extent of evidence and key characteristics relative to question and objectives • SCRs provide narrative summaries (limited or no synthesis) of evidence relative to objectives	Quantitative synthesis addresses question and objectives where appropriate; qualitative synthesis used if pooling not appropriate. • Synthesis supports a specific decision context • May include *meta*-analysis	5%

(*continued on next page*)

(*continued*)

Literature- Assessment Steps		Task Description in Scoping Review (SCR) or Systematic Evidence Mapping (SEM) Workflow	Task Description in Systematic Review (SR) Workflow	Percent Time[2]
		• SEMs output includes evidence map, database, or tables to support and inform decision making on question • Although data may inform multiple decisions, summary may be specific for decision-making context in objectives	• Synthesis should address key features separately (e.g., evidence streams, exposures, health effects) • Example for environmental health questions would synthesize hazard characterization data	
Integrate Evidenceand Report Findings	**Integrate Evidence**	SCRs and SEMs do not typically include integration or synthesis. • SEMs may identify regions of evidence with study characteristics associated with confidence or certainty	Assessment of confidence or certainty in the results of the synthesis described according to the objectives. • Should address certainty of each body of evidence relative to questions or objectives • Includes integration of the evidence base as a whole • Example for environmental health questions would provide detailed certainty of evidence for hazard or risk conclusions from exposure	8%
	DevelopReport	All review outputs provided in accessible format. • SCRs and SEMs do not typically provide conclusions • Good SEMs are interactive, sortable, and searchable • Outputs support and inform decision making on question • Outputs should inform research and analysis decisions, where data rich areas may support conclusions or data poor areas may serve as areas of uncertainty that could be addressed by research or evidence surveillance	Report all conclusions in clear language and accessible format with answer to review question. • Includes description of certainty of conclusions • Describes limitation in the review and limitations in the evidence base for assessing the question	12%
Project Management[4]		Oversight of team interactions and workflow performed to complete the review. • Develop materials and guidance for steps in the review (screening, data extraction, etc.) and provide training • Manage communication and meetings for workflow, track progress, and address problems • Arrange and conduct pilot testing of review steps and revise approach based on lessons learned • Arrange for workflow integration of new tools, machine learning and AI features • Recruit technical experts and new team members and address conflict of interest • Plan for protocol, data, and document review	Oversight is the same as SCR and SEM with additional steps for critical appraisal and integration of evidence.	19%

[1] Note: For the purpose of the table, scoping reviews and systematic evidence maps are considered to have the same workflow; adapted from James et al. (2016) and Wolffe et al. (2020).

[2] Estimated percent of work time to complete a systematic review adapted from Clark et al. (2020).

[3] Research questions for scoping and systematic reviews should be stated in terms of the Population, Exposure, Comparator, and Outcome (PECO) of interest. Scoping reviews and evidence maps sometimes do not include a specific comparator.

[4] Although project management is not typically considered a step in the literature review process, it took nearly 20% of time when considered as a separate function by Clark et al. (2020).

SR = Systematic Review, SCR = Scoping Review, SEM = Systematic Evidence Map.

systematic reviews use a comprehensive literature search based on a narrowly focused question to facilitate conclusions. Scoping reviews utilize systematic-review methods to summarize available data on broad topics to identify data-rich and data-poor areas of research and inform evidence-based decisions on further research or analysis. Systematic evidence maps use systematic-review methods to characterize the evidence base for a broad research area to illustrate the extent and types of evidence available via an interactive visual format that may be a stand-alone product or part of a scoping review (Wolffe et al. 2020). All three of these assessment formats are generally time-consuming and resource-intensive to conduct, primarily due to the need to accomplish most steps manually (Marshall et al. 2017), but also driven by the complexity of the data under consideration and amount of relevant literature to evaluate.

The specific steps in a literature-based assessment depend on the goals and approach used, with five basic steps in most assessments: 1) define the question and methods for the review (i.e., problem formulation and protocol development), 2) identify the evidence, 3) evaluate the evidence, 4) summarize the evidence, and 5) integrate the evidence and report the findings. Box 1 compares these steps across a scoping-review/systematic-review map and systematic-review products. Many steps in the literature-based assessment process have repetitive and rule-based decisions, which lend the steps to automated or semi-automated approaches.

The development and use of automation are steadily advancing in literature analysis, with much of its uptake focused on clinical and medical research. This progress may be related to funding advantages and the relative consistency of medical data and publications, or perhaps may be because clinical research has used systematic-review methodology longer. Although data sources are similar for many literature-based assessments, there are differences and unique aspects of the data relevant for addressing toxicology or environmental health questions versus clinical questions. One important difference is that environmental health assessments require the identification of research from multiple evidence streams (i.e., human, animal, and in vitro exposure studies), which necessitates training tools on publications addressing each evidence stream. In contrast, data that are relevant for addressing clinical and medical review questions come primarily from randomized controlled trials in human subjects. Even within the human data there is greater complexity in toxicology or environmental research, where a range of epidemiological study designs are used for investigating the health effects of environmental chemicals. Moreover, experimental animal studies measure more diverse endpoints and may report more data than clinical studies, resulting in longer and more complicated data extraction. Finally, cell-based assays and in vitro exposure studies provide valuable mechanistic insights for the question at hand, but these assays cover an even more diverse range of endpoints, platforms, technologies, and associated data. While clinical and environmental assessments both focus on health-related outcomes, the requirements for environmentally focused reviews expand beyond those considered in the medical field. Therefore, a tool that meets the needs of environmental health assessments could likely be applied successfully to clinical questions, while the opposite may not be true.

The systematic-review toolbox (http://systematicreviewtools.com) provides a catalogue of over 200 tools that address parts of all five assessment steps as well as associated tasks such as *meta*-analysis or collaboration (Marshall and Brereton 2015). Within the toolbox, there are multiple resources that support developing and implementing literature assessments using manual processes. For instance, several tools provide web-based forms for review teams to capture objectives, record search strategies, detail quality-assessment checklists, and record manual steps such as extracting evidence and making risk-of-bias or quality-assessment judgements. The availability of tools that support full- or partial-automation of literature-assessment processes is much more limited, despite recent advances in natural language processing (NLP), machine learning, and artificial intelligence (AI). This is especially true for the process of identifying evidence, where a combination of active learning and linguistic models can successfully predict the relevance of literature based on small samples of manually selected studies (e.g., Brockmeier et al. 2019; Howard et al. 2020; Rathbone et al. 2017; Wallace et al. 2012). These approaches have now been incorporated into several systematic-review tools. In contrast, the development and the adoption of automation methods for steps three through five of a systematic review has been limited (Box 1). The risk-of-bias assessment of individual studies is a critical and time-consuming process in assessments that is generally considered to require subject-matter experts to evaluate complex factors in study design and reporting. While a few models exist to predict risk-of-bias ratings for clinical research studies (Marshall et al. 2017; Millard et al. 2016), such methods have not translated into adoption within mainstream systematic-review tools.

Several assessment steps rely on the extraction of identified data from text, another widely recognized time-consuming process. Recent developments in NLP, including both general extraction of named entities and relationships (surveyed in Yadav and Bethard 2018) as well as specific extraction of biomedical terms (e.g., chemicals, genes, and adverse outcomes (reviewed in Perera et al. 2020)), suggest that machine-based approaches are sufficiently mature for semi-automated data-extraction approaches. Some elements of human- and machine-based data extraction are straightforward, including identifying the species and sex of the experimental animal models. Other elements, such as identifying the results of experimental assays, questionnaires, or statistical analyses, are more complex because publications may report the results from numerous assays and endpoints after multiple exposures, doses, and time periods. Standardization of reporting is also lacking, such that authors may report the experimental details using different measurement units, different names for the same chemical, and other variations in terminology (Wolffe et al. 2020). In addition, this information may be located within the text of the publication or in a table, figure, or figure caption.

In 2018, the Division of the National Toxicology Program (DNTP) participated in the National Institute of Standards and Technology Text Analysis Conference (NIST TAC) challenge by hosting the Systematic Review Information Extraction (SRIE) track to investigate the feasibility of developing machine-learning models to identify, extract, and connect data entities routinely extracted from environmental health experimental animal studies. The data entities included 24 fields such as species, exposure, dose level, time of dose, endpoint. Creating the training and test sets required structured annotation of the various data-extraction entities and labeling them in the text of each article. Developing these datasets required more comprehensive study annotation than typical data extraction workflow because the training dataset needed to capture all entities and endpoints in a research publication rather than the subset that might be relevant for a given systematic review question. These training datasets are critical to providing a fixed format that can be automatically processed and interpreted by a computer for training and model development. Overall, the results of the challenge were promising in that model-derived annotation of design features from the methods section of experimental animal studies achieved results in some extraction fields that neared human-level performance, suggesting that computer-assisted data extraction is a viable option for assisting researchers in the labor- and resource-intensive steps of data extraction in the literature-assessment process (Schmitt et al. 2018).

Given the positive outcome of the NIST TAC challenge, we developed Dextr, a web-based tool designed to incorporate NLP data-extraction models (including but not limited to models developed for the NIST TAC Challenge) into annotation and data-extraction workflows to support literature-based assessments. Many potential features were considered as we established the design requirements for Dextr (Table 1), with three design features considered key for our needs. First, and most importantly, was the ability of the tool to make data-extraction predications automatically, with the user's ability to manually verify the predicted entity or override and modify the extracted information (i.e., a

Table 1
Dextr design requirements.

Challenge	Description	Dextr Features Addressing the Challenge
Interoperability:	Efficiently import and export necessary file types	• Ability to import various file types (i.e., CSV and RIS) • Allows bulk upload of PDFs (click and drag) • Exports as CSV or modified brat file types
Usability features:	User interface that operates with efficient mouse and key stroke options (flexibility for user preferences)	• Selection with mouse click options. • Project management features • Easy to follow user interface • Ability to modify data extraction form
Complex data:	Environmental health sciences publications often report multiple experiments with various chemical exposures and doses and evaluate several endpoints (hierarchical data structure and groupings / connections)	• Ability to extract multiple entities including multiple animal models, exposures, and outcomes • Ability to connect the metadata at various levels (i.e., dose-exposure-outcome pairings)
Annotations:	Capability to annotate studies within a typical data extraction workflow that can be used to develop annotated datasets needed for training or developing new models	• Token-level (i.e., word, phrase, or specific sequence of characters) annotations recorded for each extraction entity • Ability to export annotations in a machine-readable format for model refinement and new model development
Flexibility:	Functionality that emphasizes flexibility for taking advantage of advancements in natural language processing	• Ability to utilize regular expressions to identify a string of text (i.e., #### mg = dose) or keyword searches without models • Ability to add validated models (3rd party models) to the suite of available models

"semi-automated" data-extraction approach where automated predictions are verified by the user). Second, we considered the capability to group the extracted entities (e.g., connect the species, strain, and sex of the animal model or the dose, exposure, and outcomes), critical to supporting hierarchical data extraction and greater utility of the extracted data. The third key feature was the ability of the tool to make token-level annotations (i.e., identifying a word, phrase, or specific sequence of characters) that can be used in either a typical data-extraction workflow as part of a literature review or to annotate studies for developing training datasets. The annotation of studies during data extraction has the potential to create training datasets without a separate, directed effort if the tool includes appropriate machine-readable export options. In addition to the three key features identified above, Table 1 describes unique characteristics of environmental health data and key challenges for data extraction considered in developing Dextr. Given the increasing volume of published studies, we believe that semi-automation of the labor-intensive step of data extraction by Dextr has great potential to improve the speed and accuracy of conducting literature-based assessments and reduce the workload and resources required without comprising the rigor and transparency that are critical to systematic-review methodology.

In this paper, we briefly describe the methods development of Dextr, investigate the tool's usability, and evaluate the tool's impact on performance in terms of recall, precision, and time using this semi-automated approach in DNTP's data-extraction workflow.

2. Methods

2.1. Tool development

The underlying machine-learning model (Nowak and Kunstman 2018) was developed as part of the NIST TAC 2018 SRIE workshop (Schmitt et al. 2018). Briefly, the model is a deep neural network containing more than 31 million trainable parameters. It consists of pre-trained embeddings (Global Vectors for Words Representations: GloVe and Embeddings from Language Models: ELMo), a bidirectional long short-term memory (LSTM) encoder, and a conditional random field (Nowak and Kunstman 2018). This model was developed and trained only on the methods sections of environmental health studies. Therefore, for this project, the model similarly was restricted to the methods section and performed the sequence tagging task by producing a tag (denoting a single data-extraction field) for every token from the input in the methods section. The goal for Dextr was to develop a flexible user interface that met the design requirements in Table 1 and leverage this model within a literature-review workflow. The project was conducted according to Agile software development methodology around two principal teams. A development team (AJN and KK) was formed to outline potential features and functions for Dextr and code the tool. An advisory team (AAR, CPS, RS, ARW, MSW, VRW) of experts with backgrounds in public health, literature analysis, and computational methods was then formed to guide the developers. The development team worked sequentially in "sprints" on clearly defined, testable pieces of functionality that could inform further planning and design. Throughout the process, the development team consulted with the advisory team, demonstrated newly added functionality, and presented mock-ups of the user interface illustrating key functions within the tool (e.g., project management screens, data import, extraction interfaces). As part of the Agile development sprints, each task had a test plan to verify its correctness. These test plans were then performed, first by the testers that were part of the development team, and then (if successful) by the advisory team. Members of the advisory team (ARW and RS) oversaw the project schedule and timeline and managed the development team and evaluation study. The development and advisory teams discussed potential refinements, suggested improvements, and agreed upon the approach to be implemented. When all features had been developed, a test version of the tool was produced and tested by the advisory team (ARW and RS). All issues or bugs identified during testing were addressed by the development team. When both teams agreed that the tool met the design requirements, the development team deployed the Minimum Viable Product (MVP) version of the tool to the Quality Assurance (QA) environment in April 2020. Before applying this initial version of the tool (Dextr v1.0-beta1) in daily work, QA testing and basic performance evaluation were conducted on the MVP version to quantify the potential gains in using a semi-automated workflow with Dextr compared to a manual workflow as described in the following section.

2.2. Evaluation

The aim of the evaluation was to understand how the integration of Dextr, a semi-automated extraction tool employing a machine-learning model would perform in the DNTP literature-review workflow. Specifically, we sought to understand how the tool would impact data extraction recall, precision, and extraction time compared to a manual workflow. The performance of the underlying machine-learning model was evaluated previously and was not within the scope of this evaluation (Schmitt et al. 2018). Although Dextr enables users to connect extracted entities, there is no difference in this aspect of the workflow between manual and semi-automated approaches. Therefore, the connection feature was rigorously tested and subject to QA procedures, but not part of the evaluation. Similarly, as we have continued to refine the user

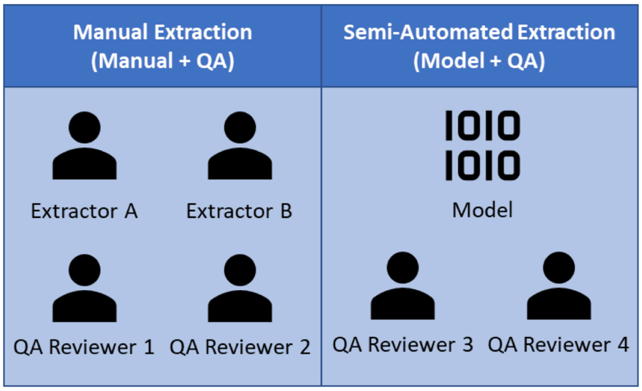

Fig. 1. Evaluation study design with two teams: manual extraction and semi-automated extraction. The manual-extraction team included two extractors who completed the primary extraction, followed by two QA reviewers. The semi-automated extraction team included the algorithm that completed the primary extraction, followed by two different QA reviewers.

interface and develop new capabilities for Dextr, we did not evaluate new features (such as the ability to use controlled vocabularies) that were not expected to negatively impact recall, precision, or extraction time. The evaluation study was designed and conducted independently of the development team and consisted of two teams, a manual extraction team (RS, ARW, RB, KS) and a semi-automated extraction team (JR and JS) (Fig. 1). The manual extraction team included two manual extractors (RS and ARW), who read each study and manually extracted each data element, and two QA reviewers (RB and KS), who reviewed the manual data extractions and made any corrections as needed (Manual + QA). The semi-automated team included two QA reviewers (JR and JS) who reviewed the machine-generated data extractions and made any corrections as needed (Model + QA). The manual extractors had prior experience with Dextr related to development discussions and testing tasks. The QA reviewers on both teams had previous and similar levels of experience with conducting data extraction for literature reviews, received a user guide for Dextr, and completed a pilot test in Dextr prior to the evaluation. Reviewers were told to accept correct data, add missing data, and ignore data incorrectly identified by either the extractor or the model. Incorrect data were ignored to minimize any additional time reviewers would spend clicking to reject incorrect suggestions. This approach reflects use of Dextr for literature-based reviews; however, if Dextr is used to construct new training data for model development, then incorrect data would need to be labeled as such.

Extractor time and reviewer time were recorded within Dextr. Prior to either beginning extraction or review, users started a timer on the tool's user interface and paused or stopped the timer (as needed) until they completed their task. The times between start and stop actions were manually checked against a complete event log to identify potential cases of the timer not being started or stopped and summed for each study to provide a total extraction time.

All the statistical analyses (JC) used in the evaluation were conducted using SAS Version 9.4 (SAS Institute Inc., Cary, NC). The generalized linear mixed model regressions were fitted using the GLIMMIX procedure with maximum likelihood estimation based on the Laplace approximation. Two-sided t tests were used to test the null hypothesis that a given fixed effect coefficient or a linear combination of the coefficients (such as the estimated difference between the log odds of recall for the manual and semi-automated modes averaged across fields and extractors) was zero. Similarly, one-sided F tests were used to test null hypotheses for interactions and contrasts, i.e., that all the corresponding linear combinations of the coefficients were zero. All statistical tests (two-sided t tests for estimated fixed effects and one-sided F tests for interactions and contrasts) were carried out at the 5% significance level. P-values are shown in the tables. There was no missing data and no removal of potential outliers.

2.3. Pilot evaluation

An initial pilot evaluation was conducted on 10 studies and the results were used to calculate the sample size for the number of studies to be included in the evaluation. All software requires a general understanding of the functionality and features to navigate the user interface and perform the tasks it is designed for – in this case to conduct data extractions. The goals of the pilot were to prepare for evaluating Dextr's performance, not to assess the learning process of new users. Therefore, an extraction guidance document was written so participants would better understand how Dextr worked and minimize the impact of the learning curve. The guidance was developed and reviewed by the advisory team and the development team prior to sharing it with the extraction team. Extraction team members provided usability feedback after the pilot; however, no changes were implemented to Dextr before the evaluation study. Since no changes were made to the tool based on the pilot, the results from the 10 pilot studies were included in the main evaluation study.

The evaluation sample size was selected using a statistical power analysis based on statistical models fitted to data from the pilot study. These statistical models used similar but less complicated formulations than the final models fitted to the final data.

Using the same notation as in the Evaluation Metrics section, the pilot study statistical model for recall was of the form:

$$\text{Logit}(\text{recall}) = intercept + \alpha_i + \beta c + \theta_s,$$

where α_i is a fixed factor for the mode, θ_s is a random factor for the study, drawn from a normal distribution with mean zero, and c is the quantitative complexity score (i.e., the calculated score divided by 100). The logit is the log odds. The same model formulation was used for precision. The pilot study statistical model used for time was of the form:

$$\text{Log}(\text{time}) = intercept + \alpha_i + \beta c + \theta_s + error,$$

where *error* is normally distributed with mean zero and is independent of the random factor θ_s. For several candidate values of K, data for K studies were simulated from each fitted model 100 times each under the alternative hypotheses, and the same statistical model was refitted to the simulated data. 100 simulations were used since the iterative method for fitting the models is computer intensive. For recall, each field was assumed to have 4 gold standard tags (the average number in the pilot data, rounded to the nearest integer). For precision, each field was assumed to have 3 tags (the average number in the pilot data, rounded to the nearest integer). The simulated complexity scores were equally likely to be any of the 10 pilot study complexity scores. For the manual mode, the simulated data used the fitted statistical models for the manual mode. For the semi-automated mode, the simulated data for recall and precision used the same model but increased the log odds by a fixed amount, delta. For the semi-automated mode, the simulated data for time used the same model but decreased the geometric mean time by a fixed percentage, perc. The estimated statistical power was the proportion of the simulated models where the difference between the two modes was statistically significant at the 5% significance level.

Based on 100 simulations from the fitted models and using a 5% significance level, we found that a sample of 50 studies would be sufficient to have an estimated statistical power of 100% (95% confidence interval (96.4, 100) %) to detect an increase or decrease of 1 in the log odds of recall, or a decrease of 1 in the log odds of precision; 98% (95% confidence interval (93.0, 99.8) %) to detect an increase of 1 in the log odds of precision, and 97% (95% confidence interval (91.4, 99.4) %) to detect a 20% decrease in median time. The confidence intervals account for the uncertainty due to the fact that only 100 simulations were used. Therefore, a final sample size of at least 50 studies was selected and 51 studies were chosen for the gold-standard dataset.

Table 2
Respiratory outcome examples and key terms.

Category	Examples and Key Terms
Organ/tissue	nasal cavity and paranasal sinus, nose (including olfactory), larynx, trachea, pharynx, pleura, lung/pulmonary (including bronchi, alveoli), glottis, epiglottis
Signs/symptoms	sneezing/sniffling, nasal congestion, nasal discharge (e.g., rhinorrhea), coughing, increased mucus/sputum/phlegm, breathing abnormalities (e.g., wheezing, shortness of breath, unusual noises when breathing)
Respiratory-related diseases or conditions	fibrosis, asthma, emphysema, chronic obstructive pulmonary disease (COPD), pneumonia, sinusitis, rhinitis, granuloma (or other inflammation)
Lung function measurements	forced expiratory volume (FEV), forced vital capacity (FVC), peak expiratory flow (PEF), expiratory reserve volume (ERV), functional residual capacity (FRC), vital capacity (VC), total lung capacity (TLC), airway resistance, mucociliary clearance

[1]Examples of outcomes or endpoints and key terms in this table provided in a guidance document for training of extractors and QA reviewers.

2.3.1. Data-extraction fields

We selected five data-extraction fields (test article, species, strain, sex, and endpoint) for the evaluation study from the full list of 24 extraction fields included in the NIST TAC SRIE challenge dataset (Schmitt et al. 2018). The challenge evaluated models using the F1 metric, which is a harmonic mean of precision (i.e., positive predictive value) and recall (i.e., sensitivity). The five data fields represented a mix of fields with high (species, sex), medium (strain), and low (test article, endpoint) F1 scores across all models previously evaluated in the challenge.

2.3.2. Gold-standard dataset

The gold-standard dataset comprised respiratory endpoints associated with exposure to biocides manually extracted from 51 experimental animal studies (Supplemental File S1). Although in vitro, experimental animal, and epidemiological study designs are of interest, we focused on experimental animal studies for the evaluation of Dextr because the model used had been developed and trained on experimental animal studies. Guidance on respiratory outcomes or endpoints provided to the extractors and QA reviewers is shown in Table 2. The teams were told that the table was not an exhaustive list and were instructed to identify any respiratory effect evaluated. The gold-standard dataset was developed by a separate extractor (PH), not included in either extraction team, who read the papers and extracted information on test article, species, strain, sex, and endpoints. A QA review was performed on the resulting dataset, or gold-standard dataset, by an independent QA reviewer (VW), who was also not on either extraction team.

2.3.3. Evaluation criteria

The results from Dextr were manually assessed by a single grader (RS or ARW) and compared to the gold-standard dataset. In brief, final data from the manual mode (Manual + QA) and from the semi-automated mode (Model + QA) were exported from Dextr into a CSV file. The results of each study by mode (manual or semi-automated) were graded separately. Each extracted data element was compared to the gold standard and marked as either a “true positive” (TP), if a match with an element in the gold standard, or a “false positive” (FP), if an additional element was not included in the gold standard. Out of 3334 results, 985 were identified as FPs. Fifty-five of these were manually flagged for further investigation for reasons such as a possible gold standard match, duplicate finding, or human error. Gold-standard data elements that were not included in the Dextr results were marked as a “false negative” (FN). Endpoints in the Dextr exports that were more specific than those in the gold standard were considered a TP. For example, if a Dextr result had an endpoint of “lung myeloid cell distribution” and “lung CD4 + T cell numbers,” but the gold-standard endpoint was “lung myeloid cell distribution (B and T cells),” then both Dextr-identified endpoints were considered TPs. Out of 1561 TP results, 540 were exact matches while 1021 were not exact matches for reasons such as plural versus singular, abbreviated or not, order of terms differed, or more detail provided in one source or the other. Out of these non-exact matches, 282 did not exactly match due to plural/singular/abbreviation discrepancies while 739 had a slight difference in wording, but were still considered a match. The graders consulted a tertiary grader (VW) to make a final decision in cases where a data element was identified in the Dextr results and missed in the gold standard. For QA, four studies were independently graded by a separate grader (either ARW or RS) and compared to the initial grading. Changes to the grading were made based on discussions between the graders. If questions between graders remained, then a tertiary grader (VW) was consulted to provide clarity and a final decision. Duplicate data elements in the Dextr results were graded only once.

Table 3
Assigned weights to data extraction fields within Dextr.[1]

Field	Weight
Endpoint(s)	0.5
Sex	1
Species	2
Strain	1
Test article	2

[1] Weights were developed using expert judgment to capture how additional extraction elements introduced complexity into the extraction task.

We calculated a complexity score for each study to account for the additional effort an extraction would take based on the number of variations of an experiment. For example, we anticipated a study with multiple test articles would be more difficult to extract than a study with only one test article. We were unable to find an established method to address complexity, and therefore complexity scores were developed using expert judgment from experienced extractors. Although the complexity scores were designed to address study characteristics overall, the score is based on study characteristics that relate to the specific data extraction elements for this paper. The number of data-extraction elements by field in the gold-standard dataset was multiplied by the weights shown in Table 3 and summed across the five data fields to calculate the score. The advisory team developed the weights for each field based on judgement related to how complex an extraction task was given multiple test articles, species, strains, sexes, and endpoints. Additional test articles and species were identified as introducing complexity to an extraction, while most studies examined multiple endpoints and did not dramatically add time to an extraction.

2.4. Usability feedback

After completion of the evaluation study, we asked manual and semi-automated QA reviewers to provide qualitative feedback on their user experience with Dextr. To summarize the assessment of usability across the reviewers, six open-ended user-experience questions were developed and responses for each question were recorded and compiled (Table S1). Note, that the feedback reflects user experience during the pilot and evaluation phases.

2.5. Evaluation metrics

We evaluated the utility of Dextr in DNTP’s workflow on three key metrics: recall, precision, and extraction time. The recall rate is the probability, prob(recall), that a gold-standard tag was correctly recalled. The precision rate is the probability, prob(precision), that an identified tag was a gold-standard tag. In the main paper, we compare arithmetic means of the recall rate or precision rate or the median total extraction time with predictions from fitted statistical models "unstratified by field" that estimate the rates or medians as a single function of the mode, field, and other explanatory variables. In the Supplemental Materials, we present alternative models for the recall and precision rates "stratified by field," where for each field the recall or precision rate is separately modeled.

The log odds of recall is defined as logit(recall) = log {prob(recall) / (1 - prob(recall)) }, where "log" is the natural logarithm. The fitted statistical model is a version of the model used in Saldanha et al. (2016). The statistical model assumes that the log odds of recall for a given gold-standard tag is a function of the mode (i = manual or semi-automated), complexity score (c), field (f = endpoint, sex, species, strain, or test article), primary extractor (p = primary1 or primary2 for manual mode, NULL for semiautomated mode), quality assurance reviewer (q = q1 or q2 for the two semi-automated mode reviewers, q3 or q4 for the two manual mode reviewers), and study (s = different values for each of the 51 studies). For each combination of study, mode, and field, the recall outcomes for each of the gold-standard tags are independent and have the same log odds, giving a binomial distribution for the number of gold-standard tags correctly recalled. For all these statistical analyses, the complexity score defined above was divided by 100 to improve model convergence without changing the underlying model formulation. The general model used the equation:

$$\text{Logit(recall)} = intercept + \alpha_i + \beta c + \gamma_f + \delta_p + \varepsilon_q + (\alpha\gamma)_{if} + (\alpha\beta)_i c + (\beta\gamma)_f c + \theta_s.$$

In this general model, $\alpha_i$andγ_f are fixed factors for the mode and field; δ_p, ε_q, and θ_s are random factors for the primary extractors, QA reviewers, and study, drawn from independent normal distributions with mean zero; and *c* is the quantitative complexity score (i.e., the calculated score divided by 100). The terms $(\alpha\gamma)_{if}, (\alpha\beta)_i, \text{and} (\beta\gamma)_f$ are interaction terms for mode × field, mode × complexity score, and complexity score × field. Thus, the model allows the effect of the field to vary with the mode or with the complexity score and allows the effect of the mode to vary with the complexity score.

We were unable to fit this general model to the data due to problems with extremely high standard errors for the mode × field interaction and some convergence issues, although there were no problems with complete or quasi-complete separation of the logistic regression models. For example, in the initial model with random factors, the estimated variance for the QA reviewer was zero, but the corresponding gradient of minus twice the log-likelihood was over 150 instead of being at most 0.001, the convergence criterion. For the final model we therefore removed the mode × field interaction and replaced the random factors for the primary and QA reviewers by fixed factors. Replacing the random factors by fixed factors might limit the generalizability of these results to other potential reviewers. We also removed main effects and interactions that were not statistically significant at the 5% level. It is possible that excluding interactions and replacing random factors by fixed factors could have introduced some bias and might limit the generalizability of the study results.

The final model was of the form:

$$\text{Logit(recall)} = intercept + \alpha_i + \beta c + \gamma_f + \delta_p + \theta_s$$

where the only random factor is the study effect. In particular, this model does not have an interaction between mode and field, so the estimated differences in log odds between modes are the same for every field. Additionally, as noted above, to evaluate differences in the mode effect across different fields we fitted alternative models stratified by field, and those results are shown in the Supplemental Materials. In particular, the stratified models show large differences between the estimated study variances for different fields.

The precision rate is the probability, prob(precision), that an identified tag was a gold standard tag. The log odds of precision is defined as logit(precision) = log {prob(precision) / (1 - prob(precision)) }. The general statistical model for precision was the same formulation as the above model for recall. As before, the final model did not include the mode × field interaction due to extremely high standard errors, and we replaced the random factors for the primary and QA reviewers by fixed factors. After removing non-significant main effects and interactions, the final model (using the same notation) was of the form:

$$\text{Logit(precision)} = intercept + \alpha_i + \beta c + \gamma_f + (\alpha\beta)_i c + \theta_s$$

where the only random factor is the study effect. In particular, this model does not have an interaction between mode and field, so the estimated differences in log odds between modes are the same for every field. For each study and mode, the total extraction time, including the primary and QA reviews, was recorded. The time taken for each field was not recorded. The general model for time taken assumes that the natural logarithm of the time taken is the following function of the mode, complexity score, primary extractor, and QA reviewer. Using the same notation as before, the general model is of the form:

$$\text{Log(time)} = intercept + \alpha_i + \beta c + \delta_p + \varepsilon_q + (\alpha\beta)_i c + \theta_s + error$$

where *error* is normally distributed with mean zero and is independent of the random factors δ_p, ε_q, and θ_s. The interaction term for mode × complexity score was not statistically significant, and again it was necessary for convergence to replace the random factors for the primary and QA reviewer by fixed factors. The primary extractor effect was not statistically significant at the 5% level. The final model was of the form:

$$\text{Log(time)} = itercept + \alpha_i + \beta c + \varepsilon_q + \theta_s + error$$

3. Results

3.1. Dextr functionality

The first version of the tool that we evaluated in this study fulfills the five design principles outlined at the tool's inception. Specifically, the tool's set-up feature provides interoperability within the existing DNTP workflow where users can upload .ris and .pdf files and export the extracted data in two forms, as .csv or .zip files. The .zip format allows exported data to be uploaded into brat (an open-source annotation software tool; https://brat.nlplab.org/). The ability to export data in a structure readable by brat allows users to leverage project data for future model development.

In terms of usability requirements, Dextr enables users to select text using their mouse or type a phrase into the extraction form. The default extraction form consists of the five extraction fields, all powered by the underlying model to provide predictions. Users can customize the data-extraction form; however, only the fields on which the model was trained are supported by automation.

We designed the default form and the tool to be able to handle relationships between the extraction entities, called "connections." This allows the user to specify a hierarchy between fields (e.g., multiple animal models can be defined, each with a species, strain, and sex). The animal model and endpoints can then be connected to a test article to create a separate experiment within the study, satisfying the third design principle related to handling complex, hierarchical data.

After project set-up, users and team members can begin extracting data elements (manually or semi-automatically) via the "My Tasks" page. Users then claim available pdfs (i.e., select a given pdf as part of the user's tasks) and access the full-text pdf within the tool to facilitate data extraction. The user can highlight the text and associate it with an extraction field. Additionally, if the exact text or phrase is not within the article itself, the user can type the appropriate text into the extraction field. These flexible options for highlighting text to populate the extraction form are useful for data extraction within a typical literature-assessment workflow, or for a more detailed annotation workflow by generating a dataset that may be used by model developers, thus satisfying the fourth design principle related to annotations. In both the manual or semi-automated workflows, the primary extractor or machine predictions are populated in the extraction form before a user accesses the study. The user then has the option to accept, ignore, or reject the extracted data or add additional data if it is missing within the form.

The last design requirements, the ability for other models to be easily incorporated into the tool and the ability to adapt to new model developments over time, were both addressed in the initial version of Dextr but not tested in the evaluation study.

3.1.1. Usability feedback

Three of the reviewers were available to participate in a feedback discussion of the tool's usability (two of the semi-automated QA reviewers and one of the manual QA reviewers). Two reviewers rated the usability of Dextr a 5 out of 10, while the other reviewer rated it an 8 out of 10. The semi-automated reviewers provided feedback on how the tool could be improved related to automatic page navigation and organization on the user interface. The semi-automated QA reviewers liked how the tool organized the extractions and agreed that the tool helped them stay organized. Once they were comfortable with the tool, they were able to work smoothly and efficiently. Reviewers identified one drawback regarding how the tool handled endpoints. All three of the reviewers found it difficult to keep track of endpoints identified by either the machine or a primary extractor. It was a challenge to find previously reviewed and accepted endpoints as reviewers continued searching for new endpoints in the extraction list. This issue was more noticeable when multiple, similar endpoints had been identified. They suggested that a more organized process for listing and tracking endpoints will improve the tool's usability.

3.1.2. Statistical models for recall, unstratified by field

A total of 51 toxicological studies were included in the final dataset. The ability of the model to correctly apply the gold-standard tag, or modeled overall recall rate, was 97.0% for the manual mode and 91.8% for the semi-automated mode. The difference in recall rates for the manual mode compared to the semi-automated mode was observed to be statistically significant ($p < 0.01$) (Table 4). These results are comparable to the arithmetic means of the recall rates across all studies and fields, which were 91.8% for the manual mode and 83.8% for the semi-automated mode. Table 4 also provides the estimate, standard error, and p-value for the difference between the log odds of the two modes. Table 5 shows the estimated log odds and recall probabilities as well as the very similar arithmetic mean recall rates for each field for the manual and semi-automated modes. Note that because there is no interaction term for mode × field, the estimated differences in log odds between the two modes are the same for every field and equal the values in the last row of Table 4. Estimates and standard errors for the fixed effects and random effects related to recall are shown in Table S2.

3.1.3. Statistical models for precision, unstratified by field

The modeled overall precision rate was 95.4% for the manual mode and 96.0% for the semi-automated mode. The precision rate for the semi-automated mode was higher, but the difference was not statistically significant (Table 6). These results can be compared with the arithmetic means of the precision rates across all studies and fields, which were 92.5% for the manual mode and 93.2% for the semi-automated mode. Table 6 gives the estimated log odds and precision probabilities for the manual and semi-automated modes, weighting each field equally, along with their standard errors and p-values. Table 6 also provides the estimate, standard error, and p-value for the difference

Table 4
Recall comparison between manual and semi-automated modes when averaged across fields and extractors, based on the model unstratified by field.[3]

Extraction Mode	Log Odds (Standard Error)	P-value of Log Odds	Probability (Standard Error)	Arithmetic Mean Recall Rate
Manual	3.483 (0.287)	<0.0001	0.970 (0.008)	0.918
Semi-automated[1]	2.418 (0.278)	<0.0001	0.918 (0.021)	0.838
Comparison[2]	−1.065 (0.109)	<0.0001	–	–

[1] Dextr predictions confirmed by QA reviewer.
[2] Comparison between manual and semi-automated extraction modes.
[3] Assumes average study complexity scores (0.175).

Table 5
Recall comparison between manual and semi-automated modes for each mode and field, averaged over evaluators, based on the model unstratified by field.[1.]

Extraction Mode	Field	Log Odds (Standard Error)	P-value of Log Odds	Probability (Standard Error)	Arithmetic Mean Recall Rate
Manual	Endpoint	1.133 (0.115)	<0.0001	0.756 (0.021)	0.744
Manual	Sex	4.856 (0.726)	<0.0001	0.992 (0.006)	0.980
Manual	Species	5.444 (1.014)	<0.0001	0.996 (0.004)	1.000
Manual	Strain	3.893 (0.477)	<0.0001	0.980 (0.009)	0.980
Manual	Test article	2.091 (0.180)	<0.0001	0.890 (0.018)	0.883
Semi-automated	Endpoint	0.068 (0.106)	0.5259	0.517 (0.027)	0.523
Semi-automated	Sex	3.791 (0.721)	<0.0001	0.978 (0.016)	0.990
Semi-automated	Species	4.379 (1.011)	<0.0001	0.988 (0.012)	0.980
Semi-automated	Strain	2.828 (0.470)	<0.0001	0.944 (0.025)	0.922
Semi-automated	Test article	1.026 (0.168)	<0.0001	0.736 (0.033)	0.773

[3]Assumes average study complexity scores (0.175).

Table 6
Precision comparison between manual and semi-automated modes when averaged across fields and extractors, based on the model unstratified by field.[3]

Extraction Mode	Log Odds (Standard Error)	P-value of Log Odds	Probability (Standard Error)	Arithmetic Mean Precision Rate
Manual	3.040 (0.281)	<0.0001	0.954 (0.012)	0.925
Semi-automated[1]	3.174 (0.287)	<0.0001	0.960 (0.011)	0.932
Comparison[2]	0.134 (0.151)	0.3765	–	–

[1] Dextr predictions confirmed by QA reviewer.
[2] Comparison between manual and semi-automated extraction modes.
[3] Assumes average study complexity scores (0.175).

Table 7
Precision comparison between manual and semi-automated modes for each mode and field, averaged over evaluators, based on the model unstratified by field.[1]

Extraction Mode	Field	Log Odds (Standard Error)	P-value of Log Odds	Probability (Standard Error)	Arithmetic Mean Precision Rate
Manual	Endpoint	1.460 (0.152)	<0.0001	0.812 (0.023)	0.794
Manual	Sex	5.073 (1.021)	<0.0001	0.994 (0.006)	0.980
Manual	Species	3.366 (0.494)	<0.0001	0.967 (0.016)	0.990
Manual	Strain	3.330 (0.490)	<0.0001	0.965 (0.016)	0.978
Manual	Test article	1.972 (0.224)	<0.0001	0.878 (0.024)	0.883
Semi-automated	Endpoint	1.594 (0.168)	<0.0001	0.831 (0.024)	0.818
Semi-automated	Sex	5.207 (1.022)	<0.0001	0.995 (0.006)	1.000
Semi-automated	Species	3.500 (0.496)	<0.0001	0.971 (0.014)	0.967
Semi-automated	Strain	3.464 (0.492)	<0.0001	0.970 (0.014)	0.940
Semi-automated	Test article	2.106 (0.236)	<0.0001	0.891 (0.023)	0.934

[3] Assumes average study complexity scores (0.175).

Table 8
Comparison of predicted mean logarithm and median for total time (seconds) between manual and semi-automated modes, averaged over evaluators[3].

Extraction Mode	Mean Log Time Modeled (Standard Error)	P-value of Mean Log Time (Modeled)	Median Time (Modeled)	Arithmetic Mean Time
Manual	6.838 (0.058)	<0.0001	933	971
Semi-automated[1]	6.079 (0.059)	<0.0001	436	517
Comparison[2]	−0.760 (0.071)	<0.0001	–	–

[1] Dextr predictions confirmed by QA reviewer.
[2] Comparison between manual and semi-automated extraction modes.
[3] Assumes average study complexity scores (0.175).

between the log odds of the two modes. Table 7 shows the estimated log odds and precision probabilities as well as the very similar arithmetic mean precision rates for each field for the manual and semi-automated modes. Note that because there is no interaction term for mode × field in the final model, the estimated differences in log odds between the two modes are the same for every field and equal the values in the last row of Table 6. Estimates and standard errors for the fixed effects and random effects related to precision are shown in Table S3.

3.1.4. *Statistical models for time*

The modeled median time was 933 s for the manual mode and 436 s for the semi-automated mode. The median time, which is the exponentiated mean log(time), was significantly lower for the semi-automated mode ($p < 0.01$). These results can be compared with the arithmetic means of the time across all studies: 971 s for the manual mode and 517 s for the semi-automated mode (Table 8). For each mode, Table 8 gives the estimated means for log(time), the standard errors of the means, and the estimated medians for time. For this model, the median time is the same as the geometric mean time. Table 8 also provides the estimate, standard error, and p-value for the difference between the mean log(time) for the two modes. Estimates and standard errors for the fixed effects and random effects related to total time are shown in Table S4.

4. Discussion

Data extraction is a time- and resource-intensive step in the literature-assessment process. Machine-learning methods for automating data extraction have been explored to address this challenge; however, the use of machine learning for data extraction has been limited to date, particularly in the field of environmental health sciences. Development and uptake of advanced approaches for extraction lag behind other steps in the review process such as literature screening, where automated screening tools have been established and used more widely. In this paper, we introduced Dextr, a web-based data-extraction tool that pairs machine-learning models that automatically predict data-extraction entities with a user interface that enables manual verification of extracted information (i.e., a semi-automated method). This powerful tool does more than provide a convenient user interface for extracting data; the tool's extraction scheme supports complex data extraction from full-text scientific articles with methods to capture data entities as well as connections between entities. With this advanced approach, Dextr supports hierarchical data extraction by allowing users to identify relationships (e.g., the connections between species, strain, sex, exposure, and endpoints) necessary for efficient data collection and synthesis in literature reviews. When evaluated relative to manual data extraction of environmental health science articles, Dextr's semi-automated extraction performed well, resulting in time savings and comparable performance in both recall and precision.

O'Connor et al. (2019) provides a framework to describe the degree of independence or "levels of automation" across tools and discusses potential barriers to adoption of automation for use in literature reviews. The degree of automation can range from tools that improve file management (Level 1), tools that leverage algorithms to assist with reference prioritization (Level 2), tools that perform a task automatically but require human supervision to approve the tool's decision resulting in a semi-automated workflow (Level 3), and tools that perform a task automatically without human oversight (Level 4). In developing Dextr, we intentionally chose to develop a Level 3 tool because we wanted a workflow that would allow expert judgment in a manual verification step to provide users the flexibility to accommodate entities where existing models may have an error rate that is too high to achieve the necessary performance. The decision to develop a semi-automated tool also addresses limited uptake of automation tools (van Altena et al. 2019) and expected barriers to adoption (O'Connor et al. 2019) of automation within the systematic-review community (e.g., providing a user verification option to address mistrust by an end-user of the automation tool, supporting transparency to demonstrate ability of the tool to perform the task, and providing a verification step similar to manual QA to lessen potential disruption of adding automation to current workflows). The work presented in this paper supports widespread adoption of a semi-automated data extraction approach because Dextr has been tested on complex study designs, in an existing workflow, and provides the user the ability to confirm the machine-predicted values, thereby increasing transparency and demonstrating compatibility with current practices.

While systematic reviews, scoping reviews, and systematic evidence maps have different formats and goals, all literature-based assessments are used to inform evidence-based decisions. Therefore, the testing of new procedures and automated approaches is essential to assess both the impact on workflow and the accuracy of the results. Given that Dextr was developed to address the time-intensive step of data extraction, its performance was evaluated in terms of recall, precision, and extraction time. Although the precision rates for the manual mode and semi-automated modes were similar, we found an unexpected and intriguing statistically significant reduction in the recall rate (arithmetic mean recall rate 0.918 for manual and 0.834 for semi-automated). Recall reflects the ability of the data-extraction approach to identify all relevant instances of an entity, and although 84% recall is good, we explored potential reasons for this decrease. While the recall for “sex,” “species,” and “strain” were comparable, the semi-automated recall rate was lower for the “endpoint” and “test article” fields. We hypothesize that the large number of endpoints predicted by Dextr may have been difficult or distracting for the user to sort through compared to manual identification. This is supported by feedback from the reviewer-usability questions and is a target for refining the user interface in future versions of Dextr to avoid this potential distraction by adding search functionality to provide a list of predicted endpoints to help extractors systematically sort through potential endpoints. The differences in recall by field (see Table 5) are also correlated with the recall rates achieved by the model on the TAC SRIE dataset (Nowak and Kunstman 2018). The fields were chosen purposefully to observe the impact of the model performance on the results. While the differences reflect the relative difficulty of the fields, we believe that model improvements will lead to closing the gap between the manual and semi-automated approaches. In terms of time, Dextr added clear efficiencies to our workflow, providing an approximately 50% reduction (53% lower predicted median time and 47% lower average time) in the time required for data extraction. This finding indicates that Dextr has the potential to provide similar recall and precision with substantial time-savings and reduced manual workload for data extraction by integrating semi-automated extraction and QA in a single step and replacing the conventional 2-step data-extraction process (a manual extractor and a manual QC check).

Although primarily developed as a tool to improve data-extraction workflow for literature-based reviews, Dextr can also be used to annotate published studies and produce training datasets for future model development. Using the tool as part of a literature review, Dextr captures token level annotations during the data-extraction workflow; these annotations are part of a machine-readable export that can potentially support model development and refinement. This feature provides an alternative to the current option of a dedicated workflow (i.e., outside of a normal literature review) required to generate training datasets and offers a reduction in cost for developing them. However, the annotations captured on each study during a literature review may have some limitations as the topic of the review could direct the extractors towards endpoints of interest rather than capturing all exposures or endpoints in a study. The lack of applicable datasets is a major impediment to model development for literature reviews (Jonnalagadda et al. 2015), and Dextr provides the potential for important advances to the field.

There are several limitations in the evaluation of Dextr that should be noted. First, we only used a single dataset to test performance. The dataset used to evaluate the tool focused on identifying and extracting respiratory health outcomes only. In contrast, the endpoint entity algorithm was not trained with this specification, and the model predicted all potential health outcomes (or endpoints) in each reference and not the respiratory subset. As noted earlier, the extractors noted in responses to reviewer-feedback questions that non-target endpoints identified by Dextr were a distraction. This limitation could have contributed to the lower recall rate observed because all non-respiratory endpoints had to be reviewed to identify relevant respiratory endpoints. Second, there are limitations associated with the models used, even though the models were not evaluated for this paper. The models currently in Dextr were developed and trained only on the methods section of environmental health animal studies. For this reason, the tool automatically identified and used only the methods section. However, detailed data extraction requires the full text of a reference because entities are commonly identified in the abstract, methods, and results sections. Similarly, information on some endpoints may be available only in tables, which Dextr currently does not process. Third, we evaluated the key performance features of Dextr (recall, precision, and time); however, we acknowledge that other aspects of the tool were beyond the scope of this project and were not tested. For example, the ability of users to establish connections was not directly tested nor a focus for user feedback. Last, this project was intended to develop a user interface designed to incorporate NLP data-extraction models. Evaluation and potential improvement of the models used were outside the scope of the work described in this paper. Therefore, it is likely that our evaluation metrics (e.g., recall of the endpoints field) will improve in conjunction with focused efforts to address model improvements.

Dextr was developed to add automation and machine-learning functions to the data-extraction step in DNTP’s literature-based assessment workflow. Although developed to address a DNTP need, we believe it is important that the new tool be available to others in the research community and be stable (i.e., have technical support) over 2–5 years. We are in the process of obtaining Federal Risk and Authorization Management Program (FedRAMP) authorization for the cloud deployment of Dextr, which will be available at (https://ntp.niehs.nih.gov/go/Dextr) when completed. The current version of Dextr (v1.0-beta1) provides a solid foundation for us to continue to refine and incorporate new features that improve workflow and enable faster and more effective data extraction. Although this publication is paired with the initial release of the tool, we are already working to expand functionality of Dextr, with planned improvements to the user interface, use of controlled vocabularies, and additional data-extraction entities. Testing the tool for data extraction on more diverse datasets is also underway. We are also working to identify existing models and develop new models that can be integrated into Dextr to expand the data-extraction capabilities to other evidence streams (e.g., epidemiological and in vitro studies). Other potential targets include the ability to extract more detailed entities (e.g., results, standard error, confidence interval) and information from tables, figures, and captions of scientific literature. As new features are developed, the design requirements of usability, flexibility, and interoperability will be periodically re-evaluated.

As described in the key design requirements, we considered it critical for Dextr to: 1) make data-extraction predictions automatically with user verification; 2) integrate token-level annotations in the data-extraction workflow; and 3) connect extracted entities to support hierarchical data extraction. This third feature, the connection of data entities, is helpful for efficient data collection and essential to enable effective synthesis in literature reviews. Controlled vocabularies and ontologies provide a hierarchical structure of terms to define conceptual classes and relations needed for knowledge representation for a given domain. Controlled vocabularies provide semantics and terminology to normalize author-reported information and support a conceptual framework when evaluating results (de Almeida Biolchini et al. 2007). Efforts are ongoing to develop field structures in Dextr compatible with integrating ontologies and controlled vocabularies. These efforts include the capability of selecting an ontology or vocabulary at the entity level with the ability to select multiple vocabularies when setting up the data-extraction form in Dextr. We are also exploring the ability of an ontology to support data extraction for specific domains or questions based on the sorting, aggregating, and association context of terms in the ontology (i.e., identifying only cardiovascular endpoints from a search of environmental exposure references).

5. Conclusions

Dextr is a semi-automated data extraction tool that has been transparently evaluated and shown to improve data extraction by substantially reducing the time required to conduct this step in supporting environmental health sciences literature-based assessments. Unlike other data extraction tools, Dextr provides the ability to extract complex concepts (e.g., multiple experiments with various exposures and doses within a single study) and properly connect or group the extracted elements within a study. Furthermore, Dextr limits the work required by researchers to generate training data by incorporating machine-readable annotation exports that are collected as part of the data-extraction workflow within the tool. Dextr was designed to address challenges associated with environmental health sciences literature; however, we are confident that the features and capabilities within the tool are applicable to other fields and would improve the data-extraction process for other domains as well.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the Intramural Research Program (Contract GS00Q14OADU417, Task Order HHSN273201600015U) at NIEHS, NIH. DNTP initiated and directed the project providing guidance on tool requirements to support data extraction for literature-analysis as well as the evaluation plan. Robyn Blain, Jo Rochester, and Jennifer Seed of ICF worked under contract for DNTP and completed Dextr testing, manual extraction, semi-automated extractions, evaluation results grading, and statistical analysis, while Pam Hartman of ICF developed the gold-standard dataset. ICF staff also provided project management for the software development and tool evaluation. Dextr programming and development were conducted by Evidence Prime as subcontractor to ICF. The underlying machine-learning model was also developed by Evidence Prime. Kelly Shipkowski (now with DNTP) worked for ICF at the beginning of the project. We appreciate the helpful comments and input on the draft manuscript provided by Keith Shockley and Nicole Kleinstreuer.

Competing Financial Interests: AJN and KK are employed by, and AJN is also a shareholder of, Evidence Prime, a software company that plans to commercialize the results of this work. To mitigate any potential conflicts of interest, these authors excluded themselves from activities that could influence the results of the evaluation study. The remaining authors declare that they have no actual or potential competing financial interests.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envint.2021.107025.

References

Brockmeier, A.J., Ju, M., Przybyła, P., Ananiadou, S., 2019. Improving reference prioritisation with PICO recognition. BMC Med. Inf. Decis. Making 19 (1), 256. https://doi.org/10.1186/s12911-019-0992-8.

Clark, J., Glasziou, P., Del Mar, C., Bannach-Brown, A., Stehlik, P., Scott, A.M., 2020. A full systematic review was completed in 2 weeks using automation tools: a case study. J. Clin. Epidemiol. 121, 81–90. https://doi.org/10.1016/j.jclinepi.2020.01.008.

de Almeida Biolchini, J.C., Mian, P.G., Natali, A.C.C., Conte, T.U., Travassos, G.H., 2007. Scientific research ontology to support systematic review in software engineering. Adv. Eng. Inf. 21 (2), 133–151. https://doi.org/10.1016/j.aei.2006.11.006.

Howard, B.E., Phillips, J., Tandon, A., Maharana, A., Elmore, R., Mav, D., Sedykh, A., Thayer, K., Merrick, B.A., Walker, V., Rooney, A., Shah, R.R., 2020. SWIFT-Active Screener: Accelerated document screening through active learning and integrated recall estimation. Environ. Int. 138, 105623. https://doi.org/10.1016/j.envint.2020.105623.

James, K.L., Randall, N.P., Haddaway, N.R., 2016. A methodology for systematic mapping in environmental sciences. Environ. Evid. 5 (1), 7. https://doi.org/10.1186/s13750-016-0059-6.

Jonnalagadda, S.R., Goyal, P., Huffman, M.D., 2015. Automating data extraction in systematic reviews: a systematic review. Syst. Rev. 4 (1), 78. https://doi.org/10.1186/s13643-015-0066-7.

Marshall, C., Brereton, P., 2015. Systematic review toolbox: a catalogue of tools to support systematic reviews. In: Paper presented at: Proceedings of the 19th International Conference on Evaluation and Assessment in Software Engineering. Association for Computing Machinery; Nanjing, China. https://doi.org/10.1145/2745802.2745824.

Marshall, I.J., Kuiper, J., Banner, E., Wallace, B.C., 2017. Automating biomedical evidence synthesis: RobotReviewer. In: Paper presented at: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics-System Demonstrations. Vancouver, Canada. https://dx.doi.org/10.18653/v1/P17-4002.

Millard, L.A.C., Flach, P.A., Higgins, J.P.T., 2016. Machine learning to assist risk-of-bias assessments in systematic reviews. Int. J. Epidemiol. 45 (1), 266–277. https://doi.org/10.1093/ije/dyv306.

Nowak, A., Kunstman, P., 2018. Team EP at TAC 2018: Automating data extraction in systematic reviews of environmetal agents. In: Paper presented at: National Institute of Standards and Technology Text Analysis Conference. Gaithersburg, MD.

O'Connor, A.M., Tsafnat, G., Thomas, J., Glasziou, P., Gilbert, S.B., Hutton, B., 2019. A question of trust: Can we build an evidence base to gain trust in systematic review automation technologies? Syst. Rev. 8 (1), 143. https://doi.org/10.1186/s13643-019-1062-0.

Perera, N., Dehmer, M., Emmert-Streib, F., 2020. Named entity recognition and relation detection for biomedical information extraction. Front. Cell Dev. Biol. 8, 673. https://doi.org/10.3389/fcell.2020.00673.

Rathbone, J., Albarqouni, L., Bakhit, M., Beller, E., Byambasuren, O., Hoffmann, T., Scott, A.M., Glasziou, P., 2017. Expediting citation screening using PICO-based title-only screening for identifying studies in scoping searches and rapid reviews. Syst. Rev. 6 (1), 233. https://doi.org/10.1186/s13643-017-0629-x.

Saldanha, I.J., Schmid, C.H., Lau, J., Dickersin, K., Berlin, J.A., Jap, J., Smith, B.T., Carini, S., Chan, W., De Bruijn, B., Wallace, B.C., Hutfless, S.M., Sim, I., Murad, M.H., Walsh, S.A., Whamond, E.J., Li, T., 2016. Evaluating Data Abstraction Assistant, a novel software application for data abstraction during systematic reviews: protocol for a randomized controlled trial. Syst. Rev. 5 (1) https://doi.org/10.1186/s13643-016-0373-7.

Schmitt, C., Walker, V., Williams, A., Varghese, A., Ahmad, Y., Rooney, A., Wolfe, M., 2018. Overview of the TAC 2018 systematic review information extraction track. In: Paper presented at: National Institute of Standards and Technology Text Analysis Conference. Gaithersburg, MD.

Altena, A.J., Spijker, R., Olabarriaga, S.D., 2019. Usage of automation tools in systematic reviews. Res. Synth. Methods. 10 (1), 72–82. https://doi.org/10.1002/jrsm.1335.

Wallace, B.C., Small, K., Brodley, C.E., Lau, J., Trikalinos, T.A., 2012. Deploying an interactive machine learning system in an Evidence-based Practice Center: Abstrackr. In: Paper presented at: IHI '12: Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium. ACM Press, New York, NY. https://doi.org/10.1145/2110363.2110464.

Wolffe, T.A.M., Vidler, J., Halsall, C., Hunt, N., Whaley, P., 2020. A survey of systematic evidence mapping practice and the case for knowledge graphs in environmental health and toxicology. Toxicol. Sci. 175 (1), 35–49. https://doi.org/10.1093/toxsci/kfaa025.

Yadav, V., Bethard, S., 2018. A survey on recent advances in named entity recognition from deep learning models. In: Paper presented at: Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe, NM.