Regulatory Toxicology and Pharmacology 137 (2023) 105287

Contents lists available at ScienceDirect

Regulatory Toxicology and Pharmacology

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

Development of benchmark datasets for text mining and sentiment analysis to accelerate regulatory literature review

Leihong Wu [a,*], Si Chen [b], Lei Guo [b], Svitlana Shpyleva [b], Kelly Harris [c], Tariq Fahmi [d,2], Timothy Flanigan [e], Weida Tong [a], Joshua Xu [a], Zhen Ren [b,1,**]

[a] Division of Bioinformatics and Biostatics, National Center for Toxicological Research, U.S. FDA, Jefferson, AR, 72079, USA
[b] Division of Biochemical Toxicology, National Center for Toxicological Research, U.S. FDA, Jefferson, AR, 72079, USA
[c] Division of Genetic and Molecular Toxicology, National Center for Toxicological Research, U.S. FDA, Jefferson, AR, 72079, USA
[d] Office of Scientific Coordination, National Center for Toxicological Research, U.S. FDA, Jefferson, AR, 72079, USA
[e] Division of Neurotoxicology, National Center for Toxicological Research, U.S. FDA, Jefferson, AR, 72079, USA

ARTICLE INFO

Handling Editor: Dr. Martin Van den berg

Keywords:
Benchmark dataset
Literature analysis
Regulatory review
Artificial intelligence
Text mining

ABSTRACT

In the field of regulatory science, reviewing literature is an essential and important step, which most of the time is conducted by manually reading hundreds of articles. Although this process is highly time-consuming and labor-intensive, most output of this process is not well transformed into machine-readable format. The limited availability of data has largely constrained the artificial intelligence (AI) system development to facilitate this literature reviewing in the regulatory process.

In the past decade, AI has revolutionized the area of text mining as many deep learning approaches have been developed to search, annotate, and classify relevant documents. After the great advancement of AI algorithms, a lack of high-quality data instead of the algorithms has recently become the bottleneck of AI system development.

Herein, we constructed two large benchmark datasets, Chlorine Efficacy dataset (CHE) and Chlorine Safety dataset (CHS), under a regulatory scenario that sought to assess the antiseptic efficacy and toxicity of chlorine. For each dataset, ~10,000 scientific articles were initially collected, manually reviewed, and their relevance to the review task were labeled. To ensure high data quality, each paper was labeled by a consensus among multiple experienced reviewers. The overall relevance rate was 27.21% (2,663 of 9,788) for CHE and 7.50% (761 of 10,153) for CHS, respectively. Furthermore, the relevant articles were categorized into five subgroups based on the focus of their content.

Next, we developed an attention-based classification language model using these two datasets. The proposed classification model yielded 0.857 and 0.908 of Area Under the Curve (AUC) for CHE and CHS dataset, respectively. This performance was significantly better than permutation test ($p < 10E-9$), demonstrating that the labeling processes were valid. To conclude, our datasets can be used as benchmark to develop AI systems, which can further facilitate the literature review process in regulatory science.

1. Introduction

Reviewing existing relevant literature is an essential step to retrieve information for many regulatory tasks (Booth et al., 2016; Pare and Kitsiou, 2017; Jensen et al., 2006). In general, literature review begins with a keyword search in multiple bibliographic databases. Keywords such as the name of a certain compound and safety related terms are used to query PubMed or Web-of-Science (WOS) and retrieve a list of potentially related articles. The articles are then screened based on the information including title, publishing journal, authors, abstract, and the full text of articles. It is important for the reviewers to make efforts in identifying as many relevant articles as possible to have a comprehensive understanding and to minimize bias (Booth et al., 2016). Unfortunately, keyword searching in bibliographic databases has low specificity

* Corresponding author.
** Corresponding author.
E-mail addresses: leihong.wu@fda.hhs.gov (L. Wu), zhen.ren@fda.hhs.gov (Z. Ren).
[1] The current address of Zhen Ren is Office of Biotechnology Product, CDER, U.S. FDA, Silver Spring, MD 20993, USA
[2] The current address of Tariq Fahmi is Division of Biochemical Toxicology, National Center for Toxicological Research, U.S. FDA, Jefferson, AR, 72079, USA

https://doi.org/10.1016/j.yrtph.2022.105287
Received 14 March 2022; Received in revised form 18 October 2022; Accepted 6 November 2022
Available online 11 November 2022

(O'Mara-Eves et al., 2015). For example, when using a compound's name in keyword searching, the queried results generally contain a large proportion of articles related to the derivatives of the compound, or other compounds with similar names. It is particularly challenging when a chemical or material of interest has multiple functions and/or has been widely used in different applications, whereas a particular task only focuses on a specific domain of application. Based on our current experience, it would take five full-time reviewers three months to screen ~20,000 articles and identify relevant ones. Similar estimates were reported by other groups conducting systematic reviews (Wallace et al., 2010; Allen and Olkin, 1999).

These challenges further emphasize the need to develop artificial intelligence (AI)-based tools to assist reviewers in literature screening (Wagner et al., 2021). In the past decade, numerous AI language models were developed in the social science fields such as movie reviews, WIKI, news, and product feedbacks. Historically, large-scale benchmark datasets have played critical roles for researchers and developers to train and evaluate AI algorithms (Hermann et al., 2015; Joshi et al., 2017). For example, the Stanford Question Answering Dataset (SQuAD) is a reading comprehension dataset (Rajpurkar et al., 2016, 2018) consisting of over 100,000 answered questions that have been widely used for Question and Answering (Q&A) model training. The Computational Natural Language Learning (CoNLL) is a well-used benchmark dataset for Name Entity Recognition (NER) model training and fine-tuning (Sang and De Meulder, 2003). The Internet Movie Database (IMDb) is one of the most widely used benchmark datasets for sentiment modeling and analysis (Maas et al., 2011). These benchmark datasets have enabled the fast growth of deep learning algorithm development.

However, most of these benchmark datasets focus on social science and general languages, and few have yet to be built in the field of regulatory science. In recent years, considerable effort has been made in developing machine learning tools for literature review or the systematic review process. In 2014 the European Chemicals Agency converted their large collection of in vitro and in vivo toxicity studies into machine readable and searchable format and generated the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) dataset, which greatly facilitates the research of computational toxicology (Luechtefeld et al., 2016). Howard et al. explored the efficiency of different machine learning models in conducting tagging and annotation based on 15 scientific public datasets with diverse sizes and complexity (Howard et al., 2016). Other large-scale datasets labeled for neuropathic pain, psychosis, and vitamin D were also reported (Liao et al., 2020). Datasets for chemical properties obtained from studies, or chemical structure information are also available (Luechtefeld et al., 2018; Isaacs et al., 2022; Sun et al., 2017). Despite all these developments, there remains a crucial need for large labeled scientific literature datasets (Asmussen and Møller, 2019). To fill in this gap, here we presented two labeled datasets for literature reviewing, each containing ~10,000 articles. Both datasets were conducted using chlorine as a model chemical, as the diversity of its application poses a challenge in extracting context of interests efficiently. The two benchmark datasets focused on the efficacy of chlorine as an antiseptic (*i.e.,* Chlorine Efficacy dataset, CHE), and its toxicity in animals (*i.e.,* Chlorine Safety dataset, CHS). To demonstrate the utility of the datasets, we developed an attention-based classification language model to predict relevant articles, using these benchmark datasets.

2. Methods and materials

2.1. Study overview

The current study overview is shown in Fig. 1. The study consisted of two parts, benchmark dataset construction and language modeling analysis. We established two benchmark datasets, CHE and CHS, in this study. First, articles were queried and acquired from six public scientific databases using pre-defined keywords as described below. The articles were then distributed to reviewers for screening and manual labeling. Each article was reviewed by up to three reviewers for the purpose of consensus and the final relevance labels were determined by the consensus result. In detail, one person reviewed the abstract or full article screening the article for relevancy. The second reviewer, usually the first reviewer's mentor, was involved in this stage of the screening. The third reviewer was involved during the summary report stage, where only relevant articles have been passed into. All three reviewers had options whether an article is relevant or not. When discrepancies arose, such as one article was passed into the review stage, but the third reviewer didn't consider it to be relevant, the article was extensively discussed during the process with additional reviewers involved if

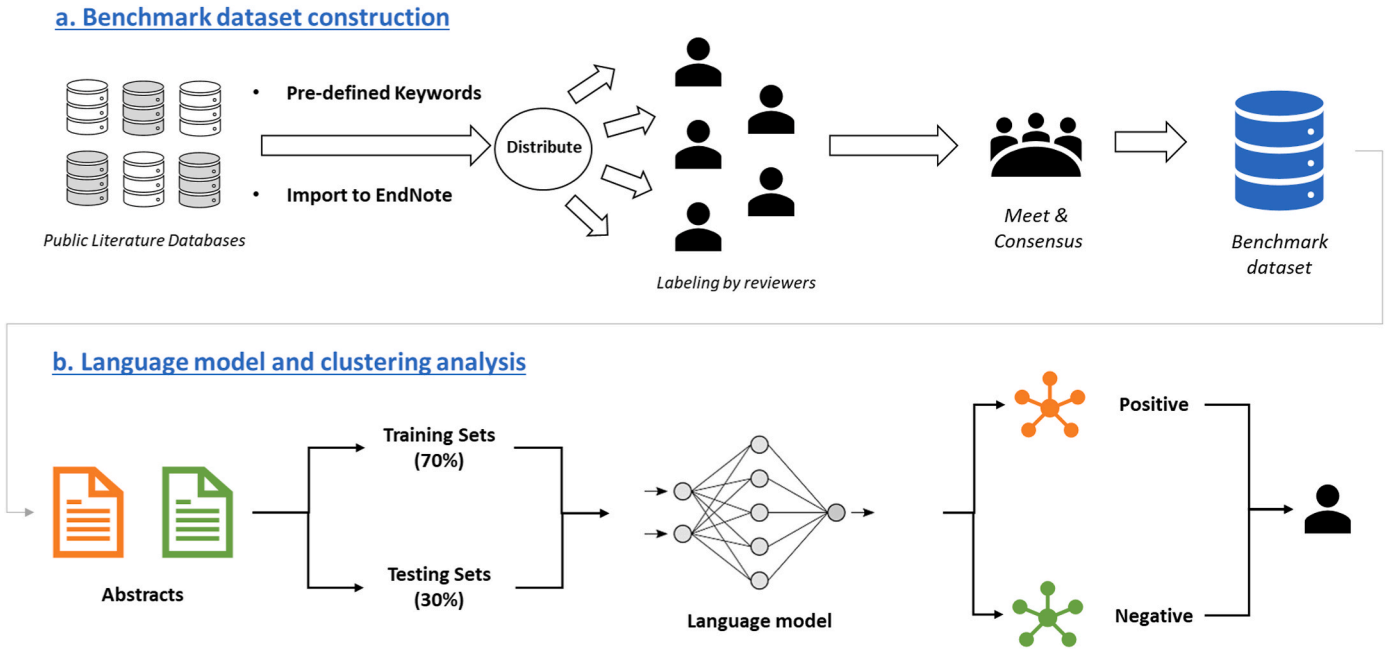


Fig. 1. Study overview. This study consists of two parts of work: (a) Benchmark dataset construction. (b) Language modeling analysis.

necessary. The articles as well as the labeling results were stored in the format of EndNote library. A language model was then trained based on the processed titles and abstracts. The titles and abstracts were combined and converted into numerical vectors via a word embedding approach. A multilayer perceptron model (i.e., densely connected neural network) (Gardner and Dorling, 1998) was developed based on the word embedding features of training datasets and evaluated on the testing datasets. Finally, the modeling result was examined and validated by experienced reviewers for explanations.

2.2. Data acquisition

When generating both the Efficacy and Safety datasets, six different scientific databases were used in retrieving literature data. Detailed keywords and search strings used are presented in the supplemental file 1. Chlorine (CAS#7782-50-5) was selected as a model chemical for establishing the datasets due to its multiple applications in different fields, mimicking a particularly challenging scenario for reviewers. Chlorine gas and other chlorine compounds such as hypochlorous acid, calcium hypochlorite and sodium hypochlorite are used in the sanitizing of fresh produce, medical equipment, and general cleaning (Bermudez-Aguirre and Barbosa-Canovas, 2013; Behrsing et al., 2000; Kreske et al., 2006; Dosti et al., 2005; Baek et al., 2012; Fu et al., 2007). At the same time, chlorine is a basic element necessary for normal function of both animals and plants and is broadly used in chemical synthesis and other processes (Naumann, 2000; Chellan and Sadler, 2015; Fang et al., 2019). The low specificity of keyword searching makes it very difficult to effectively differentiate articles from different applications, even after intensive keyword refinement.

The literature search was conducted by using the web interface of six scientific databases: PubMed (https://pubmed.ncbi.nlm.nih.gov/), EMBASE (https://www.embase.com/#search), Toxnet (https://www.nlm.nih.gov/toxnet/index.html), Web of Science (https://www.webofknowledge.com), SciFinder (https://scifinder.cas.org), and ScienceDirect (https://www.sciencedirect.com/). In addition to the keyword "chlorine" and its CAS number, three additional synonyms from PubChem Compound database, "bertholite", "cloro", and "chlorinum", were also used in the literature search. In addition to the shared keywords, specified querying keywords were pre-defined for the Efficacy and Safety datasets, and used in the data acquisition process. Detailed keywords and searching strategies used in this step are provided in **Supplementary file 1**. The literature search and retrieval were conducted in February 2018. All literature was manually collected by reviewers and imported into EndNote, and then converted into machine-readable format such as the Pandas DataFrame. In addition, contents like titles and abstracts were parsed from the xml data file. All data processing steps were performed in Python (v3.6).

These articles were distributed to the reviewer team of domain experts for manually screening and reviewing the relevance of each document. Only articles on chlorine gas and other compounds directly releasing chlorine in its application, such as sodium hypochlorite, were considered relevant in both datasets. For the CHE dataset, articles describing the disinfecting efficacy and microbial resistance were considered relevant, whereas articles on the safety and toxicity studies of chlorine were relevant for the CHS dataset.

2.3. Language modeling

Attention-based model (Vaswani et al., 2017), or Transformers, have recently become the most popular language model framework in text-mining research and many other areas (Wolf et al., 2020; Touvron et al., 2021; Liu et al., 2019). Attention-based model has also been widely applied in many text-mining activities in drug research (Liu et al., 2021). One significant advantage of the attention-based model compared to previous language models such as Long Short Term Memory (LSTM) is the transfer learning. In other words, pre-trained models can be utilized in the current study to transfer knowledge and thus accelerate the model training process. In our study, to extract features from chlorine related literature, we used the pre-trained language model that had been trained on scientific publications. Particularly, the SPECTER (Cohan et al., 2020) pretrained model developed by Allen Institute for AI was used to generate this abstract-level representation. Other pre-trained models also showed similar performance (Supplementary Table S6). A final output layer was added to the pre-trained models to fine-tune the model and produce literature relevance classification results.

The language modeling process used in this study is described as follows: first, the titles and abstracts of the articles were extracted from the dataset. Articles with unavailable title and abstract were discarded and not used in modeling process. Then, the texts were transformed into numerical feature vectors via the SPECTER model. Next, a Multilayer-Perceptron (MLP) model was added to develop a classification model for literature relevant classification. Scikit Learn (*i.e.,* sklearn) library (Pedregosa et al., 2011) was used for the MLP model, with a maximum iteration set to 1,000 and other parameters as defaults.

During the model development, the whole dataset was randomly split into training and testing sets with a ratio of 70:30. For the CHE dataset, the training and testing size were 6,741 and 2,774. For the CHS dataset, the training and test size were 6,251 and 2,680, respectively. We used sub-sampling validation and the process was repeated 100 times. Each time the training and testing data were randomly selected. The datasets were further balanced by down-sampling before training the predictive model. The model performance was evaluated by the confusion matrix (TP, FP, FN, TN) and the AUC-ROC (Area Under the Curve of ROC) score.

To further evaluate the performance of the developed classification language model for CHE and CHS, we conducted two permutation tests for CHE and CHS, by randomly shuffling the labeling of all articles. The other steps in model development (such as the model algorithm, hyperparameters) were kept the same. The permutation test was run 100 times. All computational experiments such as data processing and modeling were using Python (v3.6) and the scripts are provided in **Supplementary File 7**.

3. Results and discussion

3.1. Benchmark dataset construction

3.1.1. Querying articles and data acquisition

After retrieving the querying result, the titles, abstracts, and other meta-information of the articles were obtained from the source databases or other sources such as the publishing journals. The collected data was manually stored in the format of EndNote libraries by reviewers for further analysis. The results of literature retrieval from the six databases are summarized in Table 1. The total number of articles were aggregated from all data sources with an additional clinical search performed by the University of Arkansas for Medical Sciences (UAMS) clinical team.

3.1.2. Relevance labeling by reviewers

After data retrieval, there were 9,788 and 10,153 articles collected

Table 1
The literature retrieving results for chlorine (CAS#7782-50-5).

Database	Efficacy	Safety
PubMed	2,131	3,928
EMBASE	2,507	3,726
TOXNET	395	687
Web of Science	3,934	2,106
SciFinder	2,066	2,712
ScienceDirect	3,315	2,943
Additional Clinical Search	139	150
Total (deduplicated)	**9,788**	**10,153**

for the CHE and CHS datasets, respectively. There were 2,663 articles labeled as relevant in the CHE dataset, and 761 relevant in the CHS dataset. The relevance rates of 27.21% and 7.50%, respectively, showcase the low specificity of simple keyword searching in scientific databases. An important reason is that the queried results from keyword searching included articles on various derivatives of chlorine. A few examples include chlorine dioxide, oxalyl chloride, and chloropyrimidines. Articles such as "Acute inhalation toxicology of oxalyl chloride" (Barbee et al., 1995), "Chlorine dioxide inactivation of Cryptosporidium parvum oocysts and bacterial spore indicators" (Chauret et al., 2001), and "Chloropyrimidines as a new class of antimicrobial agents" (Agarwal et al., 2002) were all included in the queried result, but they were irrelevant under the current criteria. Moreover, the multiple applications and areas that chlorine are involved in, as discussed above, contributed to the complexity of the datasets. This is a challenge in regulatory work, when a thorough and comprehensive profiling of the efficacy and/or toxicity of a certain compound/chemical is needed.

The manual screening and labeling process took approximately three months. The CHE and CHS benchmark datasets with relevant labels, titles and abstracts are provided as **Supplementary file 2** and **Supplementary file 3**, respectively. In addition, we added metadata such as DOI, PubMed id, or URL for further research use. Note that the metadata of some articles were found by title or abstract match. In details, the title was used as the query input to PubMed API (implemented by pymed library, https://github.com/gijswobben/pymed). Since the search may return inaccurate result, Spacy (en_core_web_md) (Honnibal et al., 2020) was used to conduct the similarity match between record title/-titles and the query results, where only similarity score higher than 0.95 was considered as a good match. If there is a good match, DOI and PubMed ID were then fetched from the query result and incorporated into the dataset. In addition, the matching result is provided in the "Match" column of the supplementary files.

3.1.3. Subgroup annotation of relevant articles by reviewers

During the labeling process, the reviewers further categorized the relevant articles into subgroups based on their topics in each dataset. The subgroups could provide a second layer of information on the labeling, and may help the user/audience of these datasets to better understand what we consider as a "relevant" paper in particular research areas. For CHE, the five subgroups were: clinical studies, microbial resistance studies, food related non-clinical studies, microbial related non-clinical studies, and water treatment related non-clinical studies. For CHS, the five subgroups were: clinical studies, government documents, in vitro studies, non-mammal animal studies, and mammal animal studies. This subgroup annotation was conducted for all the 3,424 (*i.e.,* 2,663 + 761) relevant articles and provided in Supplementary Tables 2 and 3 The detailed numbers of articles in each subgroup are summarized in Table 2.

Table 2
Statistics for labeling results of two benchmark datasets.

Category	Chlorine Efficacy (CHE)	Category	Chlorine Safety (CHS)
Relevant	**2,663**	**Relevant**	**761**
• *Clinical*	*24*	• *Clinical*	*84*
• *Microbial Resistance*	*282*	• *Gov. Docs*	*46*
• *NonClinical – Food*	865	• *NonClinical – Invitro*	*226*
• *NonClinical – Microbial*	769	• *NonClinical – Mammal*	*243*
• *NonClinical - Water*	*723*	• *NonClinical - Nonmammal*	*162*
Irrelevant	**7,125**	**Irrelevant**	**9,392**
Total	**9,788**	**Total**	**10,153**

3.2. Language model

Once the CHE and CHS datasets were established, to establish the validity of the datasets to be used for AI modeling development, we conducted a "concept-of-proof" data analysis using classification language models. Particularly, we used the abstracts to develop the model. After further data processing to filter out unavailable articles, 9,245 and 8,931 articles remained and were used in the following model development. A 70:30 ratio was used to randomly split each dataset into training and testing sets (Table 3).

As detailed in the Materials and Methods, we applied the SPECTER model developed by Cohen et al. from Allen Institute for AI (Cohan et al., 2020). This model incorporated inter-document relatedness into the Transformer language model and could be readily adapted to the current study for our dataset validation purpose. The model performance for CHE and CHS is shown in Fig. 2. Overall, the Area Under the Curve??? (AUC) score for CHE and CHS, was 0.857 and 0.908 on the two testing datasets, respectively (Fig. 2a). The averaged modeling performance of 100 runs is provided in Fig. 2b.

To further evaluate the performance of the model, permutation tests for CHE and CHS were conducted by randomly shuffling the labeling of all articles. The permutation test was run 100 times. As shown in Fig. 2c and d, the AUC scores for the two datasets were significantly higher than those of the permutation test ($P < 10E\text{-}9$), further confirming the performance of the model.

3.3. Literature validation

Based on the modeling performance, we considered both CHE and CHS datasets that were reviewed and labeled by our reviewing experts as valid, and our developed models can reflect the intrinsic interests of the reviewers for determining whether an article is of interest to their reviewing mission. Furthermore, we were more interested in those wrongly predicted, or sometimes mis-labeled articles that showed discrepancies between the reviewer and our model outcome.

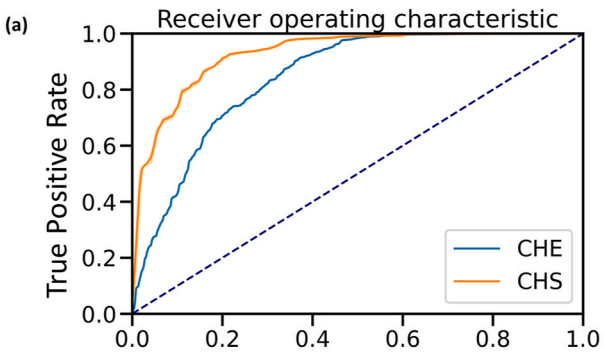
Therefore, we extracted these inconsistent articles, for both false positive (FP) and false negative (FN) types. Specifically, FP articles meant these articles were not labeled as "relevant" but were predicted as "relevant" by our model. On the other hand, FN articles meant these articles were labeled as "relevant" but were not predicted so by our models. Based on the results of 100 runs with random train-test split experiments, each article has a chance to be in the testing dataset and thus being predicted by the model, therefore, we listed the most wrongly predicted articles among 100 runs in both CHE and CHS datasets, in **Supplementary files 4 and 5**.

Particularly, we defined a novel term to denote the articles with less than 20% of corrected prediction ratio as Strong Wrongly Predicted Cases (SWPCs). The threshold of 20% was determined by considering the review expert's opinion and their capability of re-examining the literature. For example, a "*Relevant*" article that was predicted 3 times out of the 20 times would be considered as SWPC, since its corrected prediction ratio (15%, or 3/20) was below the threshold (20%). A total of 843 out of 9,245 CHE articles were considered as SWPCs. Similarly, a total of 321 out of 8,932 CHS articles were defined as SWPCs.

We further selected some typical SWPCs to explore the potential

Table 3
Numbers of articles with abstracts in CHE and CHS.

Endpoint	CHE	CHS
Relevant	2,467	557
Irrelevant	6,778	8,374
Total	**9,245**	**8,931**
• *Training (70%)*	6,741	6,251
• *Testing (30%)*	2,774	2,680



(b) Avg. Performance of 100 runs

	CHE	CHS
tp	623	178
fp	190	38
fn	157	32
tn	589	173
AUC-ROC	0.857	0.908

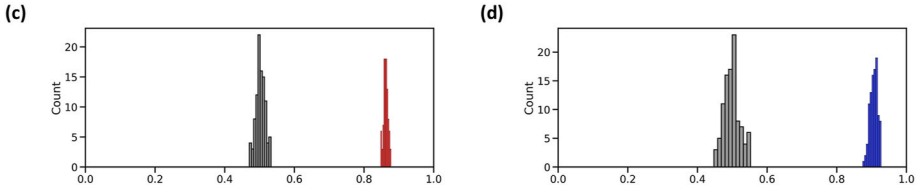


Fig. 2. Modeling performance. (a) The top half of figure shows the ROC curve for CHE and CHS; (b) the averaged modeling performance of 100 runs. (c, d) The bottom half of figure shows the AUC score of 100-run permutation test (gray) compared to (c, red) CHE and (d, gray) CHS model performance (which was also ran 100 times with different train-test splits). Both CHE and CHS showed significant higher AUC than the permutation tests ($p < 10E\text{-}9$). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

reasons for their inconsistent predictions in comparison to the reviewers' opinion. For the CHE database, there were 356 articles predicted to be '*Relevant*', but reviewers categorized them as irrelevant. One example is "A Comparison of Different Chemical Sanitizers for Inactivating *Escherichia coli* O157:H7 and *Listeria monocytogenes* in Solution and on Apples, Lettuce, Strawberries, and Cantaloupe" (Rodgers et al., 2004). This study compared the effectiveness of ozone, chlorine dioxide, chlorinated trisodium phosphate, and peroxyacetic acid on reducing bacteria. The study was very similar to the relevant articles in the database, but the chemical they used was chlorine dioxide, which was not included under our criteria. Therefore, the article was categorized as irrelevant by the reviewers. Similarly, studies using other chlorine releasing compounds or methods such as acidified sodium chlorite, electrolyzed water, or simply unspecified in the study, were categorized as irrelevant as well (Kumar et al., 2007; Nei et al., 2009; Cozad and Jones, 2003; Hricova et al., 2008; Templeton et al., 2009). Some articles explored alternative methods of disinfection, or chemicals to stabilize free chlorine for disinfection purpose, without information on the efficacy of chlorine on its own (Ramos et al., 2014; Shields et al., 2009). For example, the study "Balsamic vinegar from Modena: An easy and effective approach to reduce *Listeria monocytogenes* from lettuce" (Ramos et al., 2014), which explored the efficacy of balsamic vinegar, white wine vinegar and acetic acid in reducing *Listeria monocytogenes*. The setting of the study was comparable to relevant studies on the efficacy of chlorine, which might be the reason that it was predicted to be relevant by our model. However, chlorine was not included in the study, so it was designated as irrelevant by reviewers.

There were 486 articles that were predicted as 'Irrelevant' but recognized by reviewers as relevant in the CHE dataset. Some of them were studies using sodium hypochlorite or other related materials but did not have the keyword 'chlorine' in their title or abstract, which probably resulted in the negative prediction (Norwood and Gilmour, 2000; Hung et al., 2010; Aider et al., 2012; Macnish et al., 2010). In addition, for studies investigating the efficacy of alternative methods including ozone and UV light, sometimes chlorine treatment was used as a control group (Hu et al., 2004; Mukhopadhyay et al., 2015; Nei et al., 2011). For example, in the study "Effects of integrated treatment of nonthermal UV-C light and different antimicrobial wash on *Salmonella enterica* on plum tomatoes" (Mukhopadhyay et al., 2015), chlorine solution was used as one treatment group to compare with the antiseptic property of UV-C light. These studies could be categorized as 'Irrelevant' using our machine learning model, but because they contained valid information of the efficacy of chlorine, our reviewers considered them as relevant.

For the CHS dataset, there were 49 references predicted to be 'Relevant' but labeled as irrelevant by the reviewers. Similar to the CHE dataset, some of these references related to chlorine releasing chemicals such as chlorine dioxide that were not considered as relevant in current standard (Svecevicius et al., 2005; Itoh et al., 2001; Couri and Abdel-Rahman, 1979). Some references were inhalation or other toxicological studies on chemicals unrelated to chlorine, but probably was predicted to be relevant due to the similarity in experiment settings (Mauderly et al., 2014; Thomas et al., 1987; Traczewska et al., 2007). Another 272 references were predicted to be 'Irrelevant' but labeled relevant by reviewers. At least some of this difference resulted from a lack of proper keywords in the title or abstract (Itoh et al., 2006; Al-Salem and Al-Fadhlee, 2007; Cheng et al., 2010).

It should be noted that the large amount of irrelevant articles represented a high workload of manual screening and could increase human error in the labeling process. Our machine learning model also discovered some articles that could be reconsidered for their relevancy for chlorine efficacy and safety (Hoyle and Svendsen, 2016; Gustavino et al., 2005; Marshall, 1989; Zhou et al., 2017; Gragg and Brashears, 2010).

4. Conclusion

In summary, we manually reviewed and labeled ~20,000 scientific articles of antiseptic efficacy and safety of chlorine, and provided two fully labeled datasets, CHE and CHS. The relevance rates were 27.21% and 7.50%, respectively. The relevant articles were further categorized into five categories in each dataset based on their topics. To demonstrate the utility of these two benchmark datasets for AI system development, we applied classification language models to predict relevant articles of reviewers' interest and analyzed the model performance. This approach also demonstrated that AI approaches could potentially improve the screening process in literature review and thus reduce the workload for researchers and reviewers.

The development of AI tools in scientific literature field is highly restrained by the limited availability of large scale, well curated datasets. These two datasets we present here could fill in this gap and serve as benchmark for further language modeling development and analysis to facilitate the literature reviewing process.

Funding

This research project was funded and supported by NCTR, FDA (project ID: E0777801).

CRediT authorship contribution statement

Leihong Wu: Formal analysis, Data curation. **Si Chen:** Writing – review & editing. **Lei Guo:** Writing - review & editing. **Svitlana Shpyleva:** Writing – review & editing. **Kelly Harris:** Writing – review & editing. **Tariq Fahmi:** Writing – review & editing. **Timothy Flanigan:** Writing – review & editing. **Weida Tong:** Formal analysis, Data curation. **Joshua Xu:** Formal analysis, Data curation. **Zhen Ren:** Writing – review & editing.

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.

Data availability

The data that used in this study was shared in supplementary files

Acknowledgments

This study is an extension of a literature review requested by the Division of Nonprescription Drug Products, Center for Drug Evaluation and Research, FDA. We thank the whole CDER/NCTR monograph team and especially Howell Foster, Pharm.D., Rachael McCaleb, Pharm.D., Daniel C. Spadaro, Pharm.D. from University of Arkansas Medical School for their helps in labeling clinical related articles during this project. We also thank Drs. Xiaoqing Guo, Qiang Shi and Weizhong Zhao for their comments and discussions during the research.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.yrtph.2022.105287.

References

Agarwal, N., et al., 2002. Chloropyrimidines as a new class of antimicrobial agents. Bioorg. Med. Chem. 10 (4), 869–874.
Aider, M., et al., 2012. Electro-activated aqueous solutions: Theory and application in the food industry and biotechnology. Innovat. Food Sci. Emerg. Technol. 15, 38–49.
Al-Salem, S., Al-Fadhlee, A., 2007. Ambient levels of primary and secondary pollutants in a residential area: population risk and hazard index calculation over a three years study period. Am. J. Environ. Sci. 3 (4), 225–229.
Allen, I.E., Olkin, I., 1999. Estimating time to conduct a meta-analysis from number of citations retrieved. JAMA 282 (7), 634–635.
Asmussen, C.B., Møller, C., 2019. Smart literature review: a practical topic modelling approach to exploratory literature review. Journal of Big Data 6 (1), 93.
Baek, S.B., Kim, S.W., Ha, S.D., 2012. Reduction of Escherichia coli on surfaces of utensils and development of a predictive model as a function of concentration and exposure time of chlorine. Foodb. Pathog. Dis. 9 (1), 1–6.
Barbee, S.J., Stone, J.J., Hilaski, R.J., 1995. Acute inhalation toxicology of oxalyl chloride. Am. Ind. Hyg. Assoc. J. 56 (1), 74–76.
Behrsing, J., et al., 2000. Efficacy of chlorine for inactivation of Escherichia coli on vegetables. Postharvest Biol. Technol. 19 (2), 187–192.
Bermudez-Aguirre, D., Barbosa-Canovas, G.V., 2013. Disinfection of selected vegetables under nonthermal treatments: chlorine, acid citric, ultraviolet light and ozone. Food Control 29 (1), 82–90.
Booth, A., Sutton, A., Papaioannou, D., 2016. Systematic Approaches to a Successful Literature Review. Sage.
Chauret, C.P., et al., 2001. Chlorine dioxide inactivation of Cryptosporidium parvum oocysts and bacterial spore indicators. Appl. Environ. Microbiol. 67 (7), 2993–3001.
Chellan, P., Sadler, P.J., 2015. The elements of life and medicines. Philosophical transactions. Series A. Math. Phys. Eng. Sci. 373 (2037), 20140182.
Cheng, Y.S., et al., 2010. Exposing animals to oxidant gases: nose only vs. whole body. Proc. Am. Thorac. Soc. 7 (4), 264–268.
Cohan, A., et al., 2020. Specter: document-level representation learning using citation-informed transformers. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.
Couri, D., Abdel-Rahman, M.S., 1979. Effect of chlorine dioxide and metabolites on glutathione dependent system in rat, mouse and chicken blood. J. Environ. Pathol. Toxicol. 3 (1–2), 451–460.
Cozad, A., Jones, R.D., 2003. Disinfection and the prevention of infectious disease. Am. J. Infect. Control 31 (4), 243–254.
Dosti, B., Guzel-Seydim, Z., Greene, A.K., 2005. Effectiveness of ozone, heat and chlorine for destroying common food spoilage bacteria in synthetic media and biofilms†. Int. J. Dairy Technol. 58 (1), 19–24.
Fang, W.-Y., et al., 2019. Synthetic approaches and pharmaceutical applications of chloro-containing molecules for drug discovery: a critical review. Eur. J. Med. Chem. 173, 117–153.
Fu, E., McCue, K., Boesenberg, D., 2007. In: Johansson, I., Somasundaran, P. (Eds.), *F.1-Chemical Disinfection Of Hard Surfaces – Household, Industrial And Institutional Settings*, in *Handbook For Cleaning/Decontamination Of Surfaces*. Elsevier Science B.V., Amsterdam, pp. 573–592.
Gardner, M.W., Dorling, S., 1998. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. Atmos. Environ. 32 (14–15), 2627–2636.
Gragg, S.E., Brashears, M.M., 2010. Reduction of Escherichia coli O157:H7 in Fresh Spinach, using lactic acid bacteria and chlorine as a multihurdle intervention. J. Food Protect. 73 (2), 358–361.
Gustavino, B., et al., 2005. Modulating effects of humic acids on genotoxicity induced by water disinfectants in Cyprinus carpio. Mutat. Res. 587 (1–2), 103–113.
Hermann, K.M., et al., 2015. Teaching machines to read and comprehend. Adv. Neural Inf. Process. Syst. 28.
Honnibal, M., et al., 2020. spaCy: Industrial-Strength Natural Language Processing in Pytho. https://github.com/explosion/spaCy/blob/master/CITATION.cff.
Howard, B.E., et al., 2016. SWIFT-Review: a text-mining workbench for systematic review. Syst. Rev. 5 (1), 87.
Hoyle, G.W., Svendsen, E.R., 2016. Persistent effects of chlorine inhalation on respiratory health. Ann. N. Y. Acad. Sci. 1378 (1), 33–40.
Hricova, D., Stephan, R., Zweifel, C., 2008. Electrolyzed water and its application in the food industry. J. Food Protect. 71 (9), 1934–1947.
Hu, S.-H., et al., 2004. Antimicrobial effect of extracts of cruciferous vegetables. Kaohsiung J. Med. Sci. 20 (12), 591–599.
Hung, Y.C., et al., 2010. Effect of electrolyzed oxidizing water and chlorinated water treatments on strawberry and broccoli quality. J. Food Qual. 33 (5), 578–598.
Isaacs, K.K., et al., 2022. A harmonized chemical monitoring database for support of exposure assessments. Sci. Data 9 (1), 314.
Itoh, S., et al., 2001. Changes of activity inducing chromosomal aberrations and transformations of chlorinated humic acid. Water Res. 35 (11), 2621–2628.
Itoh, S., Nakano, A., Araki, T., 2006. Reevaluation of the toxicity of chlorinated water and the usefulness of MX as an index. J. Water Health 4 (4), 523–531.
Jensen, L.J., Saric, J., Bork, P., 2006. Literature mining for the biologist: from information retrieval to biological discovery. Nat. Rev. Genet. 7 (2), 119–129.
Joshi, M., et al., 2017. Triviaqa: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension arXiv preprint arXiv:1705.03551.
Kreske, A.C., Ryu, J.H., Beuchat, L.R., 2006. Evaluation of chlorine, chlorine dioxide, and a peroxyacetic acid-based sanitizer for effectiveness in killing Bacillus cereus and Bacillus thuringiensis spores in suspensions, on the surface of stainless steel, and on apples. J. Food Protect. 69 (8), 1892–1903.
Kumar, M., et al., 2007. Mode of Salmonella and Escherichia coli O157:H7 inactivation by a stabilized oxychloro-based sanitizer. J. Appl. Microbiol. 102 (5), 1427–1436.
Liao, J., et al., 2020. *Automation of citation screening in pre-clinical systematic reviews.* Biorxiv 280131.
Liu, Y., et al., 2019. Roberta: A Robustly Optimized Bert Pretraining Approach arXiv preprint arXiv:1907.11692.
Liu, Z., et al., 2021. AI-based language models powering drug discovery and development. Drug Discov. Today 26 (11), 2593–2607.
Luechtefeld, T., et al., 2016. Global analysis of publicly available safety data for 9,801 substances registered under REACH from 2008-2014. ALTEX 33 (2), 95–109.
Luechtefeld, T., et al., 2018. Machine learning of toxicological big data enables read-across structure activity relationships (RASAR) outperforming animal test reproducibility. Toxicol. Sci. 165 (1), 198–212.
Maas, A., et al., 2011. Learning word vectors for sentiment analysis. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies.
Macnish, A.J., et al., 2010. Sodium hypochlorite: a promising agent for reducing Botrytis cinerea infection on rose flowers. Postharvest Biol. Technol. 58 (3), 262–267.
Marshall, V.C., 1989. The predictions of human mortality from chemical accidents with especial reference to the lethal toxicity of chlorine. J. Hazard Mater. 22 (1), 13–56.
Mauderly, J.L., et al., 2014. The National Environmental Respiratory Center (NERC) experiment in multi-pollutant air quality health research: II. Comparison of responses to diesel and gasoline engine exhausts, hardwood smoke and simulated downwind coal emissions. Inhalation Toxicol 26 (11), 651–667.
Mukhopadhyay, S., Ukuku, D.O., Juneja, V.K., 2015. Effects of integrated treatment of nonthermal UV-C light and different antimicrobial wash on Salmonella enterica on plum tomatoes. Food Control 56, 147–154.
Naumann, K., 2000. Influence of chlorine substituents on biological activity of chemicals: a review. Pest Manag. Sci. 56 (1), 3–21.
Nei, D., et al., 2009. Efficacy of chlorine and acidified sodium chlorite on microbial population and quality changes of spinach leaves. Foodb. Pathog. Dis. 6 (5), 541–546.
Nei, D., et al., 2011. Disinfection of radish and alfalfa seeds inoculated with Escherichia coli O157:H7 and Salmonella by a gaseous acetic acid treatment. Foodb. Pathog. Dis. 8 (10), 1089–1094.
Norwood, D.E., Gilmour, A., 2000. The growth and resistance to sodium hypochlorite of Listeria monocytogenes in a steady-state multispecies biofilm. J. Appl. Microbiol. 88 (3), 512–520.

O'Mara-Eves, A., et al., 2015. Using text mining for study identification in systematic reviews: a systematic review of current approaches. Syst. Rev. 4 (1), 5.
Pare, G., Kitsiou, S., 2017. Methods for literature reviews. In: Handbook Of eHealth Evaluation: an Evidence-Based Approach. Victoria (BC): University of Victoria.
Pedregosa, F., et al., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.
Rajpurkar, P., et al., 2016. Squad: 100,000+ Questions for Machine Comprehension of Text arXiv preprint arXiv:1606.05250.
Rajpurkar, P., Jia, R., Liang, P., 2018. Know what You Don't Know: Unanswerable Questions for SQuAD arXiv preprint arXiv:1806.03822.
Ramos, B., et al., 2014. Balsamic vinegar from Modena: an easy and effective approach to reduce Listeria monocytogenes from lettuce. Food Control 42, 38–42.
Rodgers, S.L., et al., 2004. A comparison of different chemical sanitizers for inactivating Escherichia coli O157:H7 and Listeria monocytogenes in solution and on apples, lettuce, strawberries, and cantaloupe. J. Food Protect. 67 (4), 721–731.
Sang, E.F., De Meulder, F., 2003. Introduction to the CoNLL-2003 Shared Task: Language-independent Named Entity Recognition arXiv preprint cs/0306050.
Shields, J.M., et al., 2009. The effect of cyanuric acid on the disinfection rate of Cryptosporidium parvum in 20-ppm free chlorine. J. Water Health 7 (1), 109–114.
Sun, J., et al., 2017. ExCAPE-DB: an integrated large scale dataset facilitating Big Data analysis in chemogenomics. J. Cheminf. 9 (1), 17.
Svecevicius, G., et al., 2005. Acute and chronic toxicity of chlorine dioxide (ClO2) and chlorite (ClO2-) to rainbow trout (Oncorhynchus mykiss). Environ. Sci. Pollut. Res. Int. 12 (5), 302–305.
Templeton, M.R., et al., 2009. Chlorine and UV disinfection of ampicillin-resistant and trimethoprim-resistant Escherichia coli. Can. J. Civ. Eng. 36 (5), 889–894.
Thomas, E.L., et al., 1987. Mutagenic activity of chloramines. Mutat. Res. 188 (1), 35–43.
Touvron, H., et al., 2021. Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning. PMLR.
Traczewska, T.M., et al., 2007. Mutagenic activity of the by-products obtained during water chlorination. Environ. Protect. Eng. 33 (4), 87–95.
Vaswani, A., et al., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30.
Wagner, G., Lukyanenko, R., Paré, G., 2021. Artificial intelligence and the conduct of literature reviews. J. Inf. Technol. 37 (2), 209–226.
Wallace, B.C., et al., 2010. Semi-automated screening of biomedical citations for systematic reviews. BMC Bioinf. 11, 55.
Wolf, T., et al., 2020. Transformers: state-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations.
Zhou, Z., et al., 2017. Inactivation of viruses and bacteria on strawberries using a levulinic acid plus sodium dodecyl sulfate based sanitizer, taking sensorial and chemical food safety aspects into account. Int. J. Food Microbiol. 257, 176–182.