Enhancing Recall Using Data Cleaning for Biomedical Big Data

Priya Deshpande, Alexander Rasin, Roselyne Tchoua,
Jacob Furst, Daniela S. Raicu
DePaul University
Chicago, USA
f pdeshpa1, arasin, rtchoua, jfurst, dstg@depaul.edu

Sameer Antani
National Library of Medicine
Bethesda, USA
santani@mail.nih.gov

Abstract—In clinical practice, large amounts of heterogeneous medical data are generated on a daily basis. This data has the potential to be used for biomedical research and as a diagnostic reference for physicians. However, leveraging heterogeneous data for analysis requires integrating it rst. Integration process includes a pre-processing data cleaning phase that eliminates inconsistencies and errors originating from each data source. In this paper, we describe a work ow for cleaning heterogeneous biomedical data sources. Our novel data cleaning approach can be applied for replacement of missing text and to improve the number of relevant cases retrieved by search queries. When the threshold for missing category replacement is met, our results show that our method achieves a missing content replacement precision of 85%, which represents an improvement of 18% over the baseline state of our datasets.

Index Terms—Data cleaning, Data integration, Medical datasets, Semi-structured data, Information retrieval

I. Introduction

Biomedical research needs data integration techniques to combine available heterogeneous data sources. Some public data sources are available online; all hospitals generate vast amounts of internal data that is used internally and may be partially released after the data is anonymized through de-identi cation techniques. Making these data accessible to researchers and doctors in a uni ed data repository would contribute to progress in the eld of biomedical research. Biomedical data is generated by different experts and through different processes, commonly producing heterogeneous content, such as clinical reports, radiology teaching les, or x-ray datasets. Data may be stored in different formats and assume different terminology; there may be missing values in different data categories. In order to integrate these datasets into a common repository, these inconsistencies need to be reduced with data cleaning approaches. The medical domain has relatively few signi cant public data sources. Even with the recent major efforts such as LIDC [1] and Chest x-ray [2], little realistic biomedical data is available for research. Therefore, the integrated search cannot afford to miss relevant documents. Our main focus is thus on enhancing recall (the fraction of the relevant documents that are successfully retrieved) of the queries executed over an integrated medical data repository.

In this paper, we propose an approach to clean biomedical data and facilitate data source integration. Current research literature estimates that data discovery and integration accounts for 80% of data scientist's work [3]. Data preparation includes nding relevant data sources, extracting data from those data sources, data cleaning, data transformation, and data integration. Our data integration work ow is designed to streamline the data preparation process. We collected datasets from different biomedical data sources and evaluated our data cleaning process (see Section III-A). Our techniques should extend to similar datasets from the biomedical domain. Our experiments validate our approach by measuring the impact of our data cleaning approaches including replacement of missing data, numeric and date correction, and abbreviation expansion based on the precision and recall of queries over medical datasets. We show that our three types of data cleaning approaches improve precision and recall of query retrieval, on average, by 50%.

II. Related work

Data cleaning is the task of detecting and removing errors and inconsistencies from collected data; prior research applied data cleaning primarily for analytics of structured data. We discuss the importance of data cleaning in data integration, including the challenges and solutions from previous studies. Woo et al. [4] describe a data cleaning mechanism that uses OpenRe ne tool with clustering techniques for semi-structured medical reports. Stonebraker et al. [3] talk about data integration challenges in a real-world use case Tamr (https://www.tamr.com/). In our previous work [5] we discussed challenges in designing integrated repository that combines biomedical data sources. Prokoshyna et al. [6] discuss quantitative and logical data cleaning approaches. Proposed work identi ed semantic similarity between attributes using metric functional dependency. This approach is most applicable in relational databases where correlations between attributes are common. Dziadkowiec et al. [7] discuss data integration for electronic health records. Author integrated relational data and have not considered the problems speci c to unstructured data. Mohammed et al. [8] talk about clinical data warehouse challenges . They demonstrate that clinical data is very domain-speci c and requires domain knowledge to apply data cleaning approaches. Kruse et al. [9] discuss complexity of data integration and data cleaning. Authors also discuss structural conflict challenges and proposed a solution for relational data to eliminate such structural conflicts.

From the literature survey we observed that most of data cleaning work is done in the context of relational data. However, few papers discuss the same problem for unstructured heterogeneous data such as biomedical data sources.

Rayhan et al. [10] discuss the use of abbreviations common in medical reports. Authors present the problems associated with using abbreviations and discuss the need for uniformity in medical reports. From our previous research [11], we observed that radiologists commonly use abbreviations in clinical reports (e.g., CT for Computed Tomography, MRI for Magnetic Resonance Imaging). Our analysis also showed that different clinical reports use different names for the same category contents (e.g., Differential Diagnosis category may be named DDX). Data cleaning should therefore synchronize terms and their corresponding abbreviations.

III. METHODOLOGY

In this section we describe our data sources, the search operations that we perform to evaluate results, and data cleaning issues associated with medical data integration.

A. Data sources

We used 4 different data sources and 2 medical ontologies. Radiology Society of North America Medical Imaging Resource Community (http://mirc.rsna.org/query) RSNA MIRC is a large repository of 2,500 teaching files with cases including patient history, diagnosis, differential diagnosis, findings, discussion as well as external references (e.g., journal articles). Teaching files are used as a learning source by radiology students and doctors. Weinberger et al. developed MyPacs.net [12] webservice that allows radiologists to share (create, upload, and modify) teaching cases. 17,000 teaching cases were publicly available between 2002 and 2019. EURORAD (http://www.eurorad.org/) is a dataset with radiology case reports more than 7,200, operated by the European Society of Radiology. National Institutes of Health provides a dataset [13] with more than 3,500 public clinical reports. Medical ontologies provide definitions, synonyms, and conceptual relation information for medical terms. We used Radiology Lexicon (RadLex, http://radlex.org/) containing more than 45,000 terms and Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT, https://www.nlm.nih.gov/healthit/snomedct/) with more than 300,000 terms.

Integration of these data sources is challenge due to their heterogeneous nature. For example, some data sources (e.g., MIRC, MyPacs) have title, diagnosis, history, differential diagnosis, findings, and discussion categories. Other data sources (e.g., NIH x-ray, EURORAD) do not have findings, history, or differential diagnosis categories. Instead, these sources provide other information fields such as observation, procedure, or image findings. We also observed data inconsistency such as varied use of abbreviation and different date format.

B. Search mechanism

We use three different types of search to evaluate the performance of our data cleaning approach. The first search type is a *diagnostic search* that queries the diagnosis category. Diagnostic search finds relevant cases based on the presence of the search terms in the diagnosis category. In NIH x-ray dataset, impression is the category that discusses the diagnosis of the case, so for the purposes of our analysis we consider impression category as a diagnosis category. The second search type is an *abbreviation search* that inspects the entire text of the clinical case. Both search types apply query expansion – searching for query terms and abbreviations of the query terms. Third type of a search is a *basic content search* search that uses queries related to age or date of modification that inspects the entire text of clinical cases. In practice, users of medical search engines are often interested in the diagnostic across other data sources. If a user searches for a particular diagnosis information and diagnosis category contents are missing, then although other contents from the case provides information relevant to the query, that case might not be ranked as a top relevant case. For example, a case with “cardiomegaly” title (an enlargement of heart) might not have diagnosis information available but other contents from this case (e.g., title, discussion) may show that this is a relevant case. However, because diagnosis information is not available this case might not be retrieved by the “cardiomegaly” query.

b) Errors and inconsistent values: Our data analysis shows that date categories (such as date of modi cation or date of creation) contains different errors and inconsistencies in different datasets. For example, “20000-12/19” has an additional digit in the year eld and separators are not consistent. Based on geographical location of the data sources, date formats can be different (e.g., MM-DD-YYYY or DD-MM-YYYY). In history category, patient age may be recorded erroneously, such as “190 year old female with diabetic history”. Medical data is typically de-identi ed when it is shared publicly; each data source provider applies their own de-identi cation techniques. For example, NIH clinical reports de-identify patient personal details by replacing personal information such as date of birth with “xxx”. MIRC and MyPacs datasets do not provide any personal information but these data sources are meant as a learning source by radiology students and de-identi ed some of the personal details (e.g., name, address).

The bulk of work that addressed data cleaning problems focused on structured data integration. , 631(on)-cleaning problemsp Td [(of)-350(the)-un31(data)-631]TJ lysis "study") using python NLTK library (https://www.nltk.org/), language identi cation, garbage characters removal, and removal of stop-words. Stop-words are the most common words used in a language, removed in natural language processing because term frequency of these words would be higher than other important words in corpus (e.g., "the", "but", "and"). Using medical ontologies (RadLex and SNOMED CT), we created our own list of stop-words that we did not remove from our data. For example, "with" or "no" are stop-words. However, in medical domain these terms are signi cant and may belong to an ontology entry or modify other medical terms. We have identi ed 24 custom stop-words that we keep in our dataset such as most, between, no, below, or with.

3) Abbreviation substitution: We substitute abbreviations for terms in clinical data reports to maintain a uniform terminology. Uses our abbreviation dictionary, we expand the search to both the query term itself and the abbreviation of that query term. We used well-known medical data sources such as American College of Radiology (https://www.acr.org/), Radiologyinfo (https://www.radiologyinfo.org/), Radiopaedia (https://radiopaedia.org/), SNOMED CT, and RadLex ontology to create our abbreviation dictionary. For example, for "Computed Tomography" query our system searches for "CT" and "Computed Tomography".

E. Evaluation

In this section, we describe the search methodology used to evaluate our approach, including our measure of search result relevance. To evaluate our data cleaning approach, we perform query search on the datasets before and after cleaning. We used queries collected from radiologists at a well-known medical hospital and from an extensive literature survey [16]. We split the set of queries into two different parts: diagnostic queries (diagnosis-related terms) and queries for which there is a medical abbreviation. We evaluated 14 diagnostic queries (cardiomegaly, chiari, angiosarcoma, varicocele, acl tear, appendicitis, hepatic adenoma, annular pancreas, perthe, splenic hemangioma, CCAM, pseudohypoparathyroidism, congenital indifference, ameloblastoma) and 5 (cystic brosis, brocystic, ff - free uids, study of bladder function, plain x-ray) abbreviations queries. These queries are the most representative queries from our collection (28 queries) of different queries that represents diagnostic and abbreviation terms. We consider the 10 most recent results (based on date of modi cation of the case) for this evaluation.

Query retrieval results were evaluated with the help of experts in NLP, databases, and information retrieval but with no medical training. Retrieved results relevance evaluation was based on a coding standards document. This coding standards document was created with all relevant de nitions such as medical term synonyms and pertinent information about the diseases. We created the coding standard based on medical ontologies RadLex and SNOMED CT as well as other reference sources. Evaluators scored search results on a binary scale of 0 ("not relevant") and 1 ("relevant"). Our evaluators were given detailed instructions on what constitutes each of the dataset categories based the coding standard document. We present our results in terms of precision and recall. We evaluated the precision of substitution by based on randomly chosen 50 documents from each dataset where replacement was performed; we computed precision using Equation 2.

$$Precision = \frac{found\_and\_relevant}{total\_found} \quad (2)$$

found and relevant is the total number of documents where manual evaluation shown the replaced diagnosis to be relevant to the contents of the clinical report. *total found* is the total number out of 50 documents that we evaluated from a set of documents that was changed by content replacement approach. We compute recall as shown in Equation 3.

$$Recall = \frac{found\_and\_relevant}{total\_relevant} \quad (3)$$

IV. Experiments and Results

In this section, we evaluate the improvements achieved by our approach. We rst performed an analysis using word cloud generation for each individual category. From this analysis, we observed that many of the cases were missing contents in some categories. The amount of missing category contents varies among different data sources: for example, in MIRC 3.7% of diagnoses are missing, while in MyPacs 31% of diagnoses are missing (see Section III-C). We computed the sequence similarity (using Equation 1) between different categories to identify a replacement source for the missing content. For example, in MIRC the average similarity between title and history: 0.14, between differential diagnosis and discussion: 0.23, and between ndings and diagnosis: 0.51. The average similarity between title and diagnosis categories for the four datasets: MIRC similarity is 0.76, MyPacs is 0.65, EURORAD is 0.72, and NIH chest x-ray dataset is 0.20. NIH dataset has a very low similarity between title and diagnosis because the title often contains case details and hospital name. MyPacs has a lower similarity ratio compared to MIRC and EURORAD because several of case titles use a sequential number (e.g., CASE102, CASE1005) instead of diagnosis-related terms. Only the title and diagnosis categories have an average similarity above the threshold of 0.6. Some of the categories (e.g., history, discussion) contain a lot of general text making it dif cult to replace category contents with other values. For categories that are missing data, our current algorithm replaces category contents with "NA".

1) Evaluation of missing content replacement: We performed a manual evaluation of the cases where missing diagnosis contents were replaced by data from title category. Our precision for MIRC dataset is 84%, for MyPacs the precision is 82%, and for EURORAD the precision is 88%. In order to evaluate the similarity threshold, we calculated the average precision across all four datasets for different similarity thresholds. As shown in Figure 1, the knee of the curve can be found at the threshold of 0.6. Increasing the threshold (i.e., requiring a higher similarity between title and diagnosis) exhibits a small marginal improvement in the

precision of the replacement. However, lowering the threshold to 0.5 or below signi cantly decreases the average precision.

Figure 1. Average precision across 4 datasets.

V. Evaluation analysis

We divided the evaluation of out results based on the speci c tasks described in Section III.

1) Relevance of substitution of missing contents and improvement in the diagnostic search: We evaluated our missing value replacement algorithm by performing the diagnostic search. For this evaluation, we ran 14 queries against the diagnosis category to see the difference between results with data cleaning applied (DC=YES) and with no data cleaning (DC=NO). We chose query terms related to diagnosis terminology which is why we excluded some of the previously collected queries (e.g., “study”, “toxic” are too general and not diagnosis-related). Figure 2 shows the increase in the number of found documents (average for all 14 queries) using the diagnostic search with 3 datasets. For MIRC dataset 8 queries out of 14 (57%) have shown some improvement; similarly, for MyPacs 71% and for EURORAD 50% of the queries shown an increase in the number of found documents. Some

Figure 2. Diagnostic Search: with (YES) and without (NO) data cleaning

of the queries search for the term for which our data does not contain a relevant diagnosis. For example, in EURORAD “cardiomegaly” is not present and our algorithm returns zero results even after replacing missing contents. MyPacs is one of the largest datasets with a variety of cases and search results show the most overall improvement. We observed a moderate improvement in MIRC and EURORAD because many of our queries (45%) do not match relevant cases in our dataset. Our analysis shows that our missing contents replacement approach improves diagnosis retrieval performance across all datasets where the replacement threshold is met.

2) Improvement in the basic content search – removing errors and replacing inconsistent contents: As described in Section III, we removed errors and inconsistencies using natural language processing techniques. Without data cleaning, search queries were resulting in UTF data encoding errors; we clean these errors using python string library. Even without

medical dataset, evaluating all documents relevant to a search is prohibitively expensive. To compute relevance of documents

Figure 4. Average precision graph

to context of search queries, we applied the following method. We have the set of documents for DC=NO (we call this set N) and then assume the recall is R. We further assume that the set for DC=YES (we call this set Y) has a recall of S. We then look at the documents in the set Y N (where Y is a strict superset of N). We examine the documents in the set Y N and if any are relevant then we conclude that recall must have gone up and we show that $S > R$. As shown in

Figure 5. Average recall graph

Figure 4 and Figure 5 (accuracy graphs for three data cleaning approaches: DSNO: Diagnostic Search DC=NO, DSYES: Diagnostic search DC=YES, OSASNO: Overall Search Abbreviation Substitution DC=NO, OSASYES: Overall Search Abbreviation substitution DC=YES), our missing data insertion approach improves the average precision by 0.17 and average recall by 0.21.

Replacing abbreviations in query search improves the average precision by 0.34 and recall by 0.31. This analysis shows that our data cleaning approaches improves the quality of search results.

VI. Conclusion

Data cleaning and missing content replacement for heterogeneous biomedical data is a challenging task. Research work presented in this paper proposes and evaluates technique for removing errors and inconsistencies in medical datasets, decreasing the amount of missing contents, and improving query result in terms of number of found cases. Our analysis demonstrates that our data cleaning methods achieves a missing content replacement precision of 85%, which represents an improvement of 18% over the baseline state of our datasets.

Acknowledgments

This research was supported in part by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM), and Lister Hill National Center for Biomedical Communications (LHNCBC). We are also thankful to Dr. Andrew Trotman from The University of Otago, New Zealand, for his valuable suggestions and feedback.

References

[1] S. G. Armato et al., "The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans," Medical physics, vol. 38, no. 2, pp. 915–931, 2011.

[2] X. Wang et al., "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2097–2106.

[3] M. Stonebraker and I. F. Ilyas, "Data integration: The current status and the way forward." IEEE Data Eng. Bull., vol. 41, no. 2, pp. 3–9, 2018.

[4] H. Woo et al., "Application of efficient data cleaning using text clustering for semistructured medical reports to large-scale stool examination reports: Methodology study," Journal of medical Internet research, vol. 21, no. 1, p. e10013, 2019.

[5] P. Deshpande et al., "Diis: A biomedical data access framework for aiding data driven research supporting fair principles," Data, vol. 4, no. 2, p. 54, 2019.

[6] N. Prokoshyna, J. Szlichta, F. Chiang, R. J. Miller, and D. Srivastava, "Combining quantitative and logical data cleaning," Proceedings of the VLDB Endowment, vol. 9, no. 4, pp. 300–311, 2015.

[7] O. Dziadkowiec et al., "Using a data quality framework to clean data extracted from the electronic health record: A case study," eGEMs, vol. 4, no. 1, 2016.

[8] R. O. Mohammed and S. A. Talab, "Clinical data warehouse issues and challenges," International Journal of u-and e-Service, Science and Technology, vol. 7, no. 5, pp. 251–262, 2014.

[9] a. N. Kruse, Papotti, "Estimating data integration and cleaning effort." in EDBT, 2015, pp. 61–72.

[10] R. A. Tariq and S. Sharma, "Inappropriate medical abbreviations," in StatPearls [Internet] StatPearls F[([8.966 Tw2 0 G [-625(R.)-435(A.)-43145.93

[3]