MEDINFO 2023 — The Future Is Accessible
J. Bichel-Findlay et al. (Eds.)

doi:10.3233/SHTI231042

Extracting Spatio-Temporal Trends in Medical Research Prioritization Through Natural Language Processing of Case Report Abstracts

Lean Franzl Lim YAO[a], Kongmeng LIEW[a], Shoko WAKAMIYA[a] and Eiji ARAMAKI[a,1]
[a] *Nara Institute of Science and Technology, Japan*
ORCiD ID: Lean Franzl Lim Yao https://orcid.org/0000-0003-3184-9368, Kongmeng Liew https://orcid.org/0000-0002-0755-7173, Shoko Wakamiya https://orcid.org/0000-0002-9371-1340, Eiji Aramaki https://orcid.org/0000-0003-0201-3609

Abstract. Medical research prioritization is an important aspect of decision-making by researchers and relevant stakeholders. The ever-increasing availability of technology and data has opened doors to new discoveries and new questions. This makes it difficult for researchers and relevant stakeholders to make well-informed decisions about the research areas they want to support and the nations they should look for collaborations. It is, therefore, useful to look at the spatio-temporal trends of medical research prioritization to gain insight into popular and neglected areas of research as well as the allocation of prioritization of each nation. In this study, we develop a system that collects, classifies, and summarizes case report abstracts according to the location, time, and disease category of the report. The additional classifications allow us to visualize and monitor the trends in medical research prioritization by location, time, and disease category.

Keywords. Medical research prioritization, natural language processing, named entity recognition

1. Introduction

Advances in information technology paved the way for the modernization of healthcare systems and medical research. Computational tools now available to medical researchers, like machine learning techniques, are widely used in a wide range of tasks and research [1]. This allows researchers to explore questions and research areas that would have otherwise been difficult or impossible. Medical researchers are now faced with an ever-growing number of choices on the types of diseases they want to study, and likewise, institutions are faced with the same number of choices to which to allocate their research funds. While some categories grow more popular, other diseases

[1] Eiji Aramaki, PhD, email: aramaki@is.naist.jp; Social Computing Lab, Nara Institute of Science and Technology, 8916-5 Takayamacho, Ikoma, Nara, 630-0192, Japan.

are faced with neglect [2]. Medical research prioritization is thus a problem faced across various disciplines and usually involves discussions among various stakeholders sharing their domain expertise [3,4]. These discussions and decisions affect the lives of many individuals and take time and resources to reach a conclusion. With the use of technology and the ever-growing amount of data available, the speed and quality of these discussions can be improved. In this study, we propose a system that utilizes natural language processing (NLP) tools to provide a data-driven macro perspective of trends in medical research to gain additional insight that will help researchers and institutions make better-informed decisions.

To achieve this, we rely on advances in named entity recognition (NER). NER refers to the identification of entities in the unstructured text that belongs to predefined categories. In medical research, NER can be used to identify names of diseases, drugs, and procedures, among others (see [5,6]).

In this research, we propose the use of NER models on publicly available databases of case reports. Case reports are written to inform colleagues about important and novel information about symptoms, diagnoses, and treatments of diseases [7,8]. These collections of reports provide a rich source of information on the research prioritization of various countries according to each time period.

2. Methods

The data used in our system consists of case report abstracts from the PubMed archive [9]. PubMed provides free access to indexes, titles, abstracts, and author information from biomedical and life science literature. We extracted N = 298,303 articles using the search term “case report” and a date range from 2016-01-01 to 2021-10-31 (Figure 1). We obtained each article's abstract, publication date, and author affiliation. For the publication date, we took the earliest date recorded if there were multiple dates.

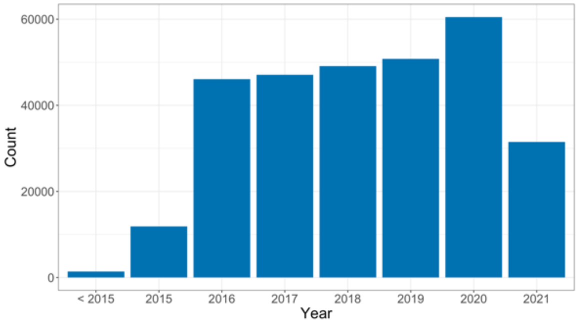


Figure 1. Number of extracted case reports per year.

The extraction of disease names mentioned in the abstract can be classified as an NER task in NLP. A common method to approach NER tasks is to treat it as a sequential labeling problem where labels are assigned to each token (usually words) in a sequence of tokens (usually a sentence). There are many approaches to NER, with state-of-the-art approaches utilizing deep learning-based models to achieve the best performance. Other deterministic approaches, such as conditional random fields (CRF), are still used or utilized by modern techniques. CRFs are a class of statistical modeling methods that perform well for labeling sequential data because the context of a target token can be taken into consideration compared to methods that treat each token as the sole input. In lieu of state-of-the-art models for NER, we used CRF for its better explainability and understanding, ease of training, and good performance [10,11].

Our NER model was trained using the NCBI corpus, which contains PubMed abstracts that are annotated with mentions of disease names [12]. We started by tokenizing the abstracts first into sentences, then into words [13,14]. We obtain the parts-of-speech tags for each word [13] with the reformatted BIO/IOB tags (classify each token to be at the "beginning," "inside," or "outside" of a named entity) of the annotations to train the model.

Two variables need to be normalized in our system: author affiliations and mentions of disease names. The author affiliations were normalized into country codes using a rule-based approach utilizing a list of country names and lists of states, provinces, and prefectures of some countries. Normalizing disease names required further measures.

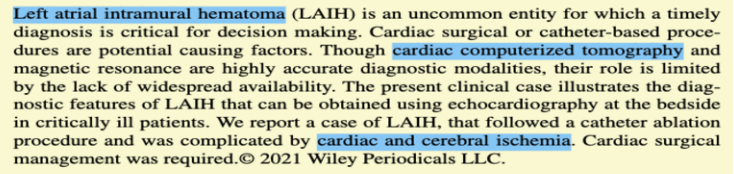
We normalized disease names to the January 2021 release of the International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM) codes [15]. This was done by calculating the Levensthein similarity score between mentions and each of the disease names in ICD-10-CM. Additionally, we also used the Orphanet dataset, which includes ICD-10 disease names, codes, as well as synonyms of rare diseases [16].

3. Results

Our NER model achieved an F1-score of 0.7661 on the NCBI corpus's test set, which suggests that the model performs moderately well in identifying disease names (Table 1). Disease names were detected in 264,399 abstracts (88.63%), whereas 205,939 (69.04%) abstracts contained entities that were successfully normalized to ICD-10-CM codes. A sample of an abstract with annotations is shown in Figure 2.

Table 1. Named entity recognition model performance.

Tag	Precision	Recall	F1 Score	Support
Disease	0.8222	0.7172	0.7661	838
micro avg	0.8222	0.7172	0.7661	838
macro avg	0.8222	0.7172	0.7661	838
weighted avg	0.8222	0.7172	0.7661	838



Left atrial intramural hematoma (LAIH) is an uncommon entity for which a timely diagnosis is critical for decision making. Cardiac surgical or catheter-based procedures are potential causing factors. Though cardiac computerized tomography and magnetic resonance are highly accurate diagnostic modalities, their role is limited by the lack of widespread availability. The present clinical case illustrates the diagnostic features of LAIH that can be obtained using echocardiography at the bedside in critically ill patients. We report a case of LAIH, that followed a catheter ablation procedure and was complicated by cardiac and cerebral ischemia. Cardiac surgical management was required.© 2021 Wiley Periodicals LLC.

Figure 2. Sample abstract with identified disease names in highlights.

Each abstract was tagged according to the first author's affiliated country, publication date, and the ICD-10-CM codes of mentioned diseases. These three variables allowed us to summarize the data and visualize the trends and distributions (Figures 3-5).

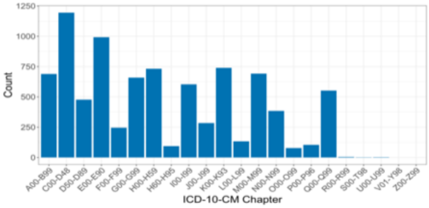


Figure 3. Bar chart of the distribution of published case reports for Japan in 2017.

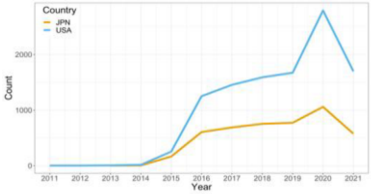


Figure 4. Line chart of published case reports for ICD Chapter A00-B99 from 2011 to 2021.

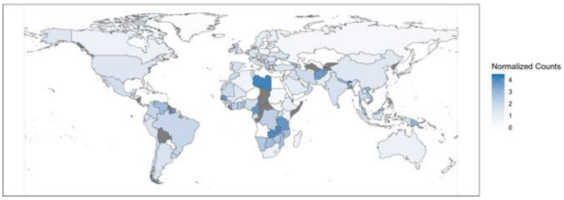


Figure 5. Heat map of normalized counts of published case reports for ICD Chapter A00-B99: Certain infectious and parasitic diseases in 2017.

4. Discussion

The motivation for this paper was to monitor trends and patterns in medical research prioritization using the number of published case reports as a metric. This was a preliminary study that will be expanded further, so we did not aim to achieve state-of-the-art performance for our model's components. Instead, we relied on methods that provide good enough performance that is easy to implement and understand.

This data can be used to describe the distribution and trends of prioritization of a nation as well as which nations are leading in research output for each disease category. These insights can be used to help make decisions on research focus, funding, and collaboration by looking at the trends and distributions. Incorporating external data of interest to researchers or stakeholders, such as economic or epidemiological data, can uncover underlying relationships relevant to decision-making.

We also note some limitations of our system. Our system uses only the abstracts of published case reports. Although it is possible that some disease names are mentioned in the main text and not in the abstract, we rely on the assumption that relevant disease names would be included in the abstract. Our use of Levensthein distance does not handle acronyms or alternative names, but this can be improved with more state-of-the-art approaches. We rely on another assumption that official names would be used in the abstract at least once. Lastly, our country name normalization does not include every country's subregions and would, therefore, underrepresent some countries.

5. Conclusions

In this paper, we presented an end-to-end system that utilizes NLP techniques to collect, classify, and summarize abstracts of published case reports according to publication date, first author's country of affiliation, and ICD-10-CM codes of identified disease names. This system serves to help medical researchers decide on which diseases to prioritize based on their circumstances and resources. This can also be used by relevant stakeholders to decide on funding allocation and collaboration efforts. Without the use of state-of-the-art tools, we demonstrated the capable performance of simpler methods that still yield good results and usable insights.

Acknowledgements

This work was supported by JST CREST, Grant Number JPMJCR22N1, JST SICORP, Grant Number JPMJSC2107, and MHLW Program Grant Number JPMH21AC500111, Japan.

References

[1] Sidey-Gibbons JAM, Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction. BMC Med Res Methodol. 2019 Mar;19(1):1-18, doi: 10.1186/s12874-019-0681-4.

[2] Burgos LM, Farina J, Liendro MC, Saldarriaga C, Liprandi AS, Wyss F, Mendoza I, Baranchuk A. Neglected Tropical Diseases and Other Infectious Diseases Affecting the Heart. The NET-Heart Project: Rationale and Design. Global Heart. 2020 sep;15(1):60, doi: 10.5334/gh.867.

[3] Carlson JJ, Thariani R, Roth J, Gralow J, Henry NL, Esmail L, Deverka P, Ramsey SD, Baker L, Veenstra DL. Value-of-Information Analysis within a Stakeholder-Driven Research Prioritization Process in a US Setting: An Application in Cancer Genomics. Med Decis Making. 2013 may;33(4):463-71, doi: 10.1177/0272989X13484388.

[4] Dey CJ, Rego AI, Midwood JD, Koops MA. A review and meta-analysis of collaborative research prioritization studies in ecology, biodiversity conservation and environmental science. Proc Royal Soc B. 2020 Mar; 287(1923):20200012, doi: 10.1098/rspb.2020.0012.

[5] Percha B. Modern Clinical Text Mining: A Guide and Review. Annu Rev Biomed Eng. 2021 Jul;4:165-87, doi: 10.1146/annurev-biodatasci-030421-030931.

[6] Yu X, Lu S, Hu W, Sun X, Yuan Z. BioBERT based named entity recognition in electronic medical Record. In2019 10th international conference on information technology in medicine and education (ITME). 2019 Aug;49-2, doi: 10.1109/ITME.2019.00022.

[7] Frontera WR. Scientific research and the case report. Am J Phys Med Rehabil. 2012 Aug;91(8):639, doi:10.1097/ PHM.0b013e31825f1bf7.

[8] Guidelines To Writing A Clinical Case Report. Heart Views: The Official Journal of the Gulf Heart Association. 2017;18:104, doi: 10.4103/1995-705X.217857.

[9] National Center for Biotechnology Information (NCBI);Available from: https://www.ncbi.nlm.nih.gov/.

[10] Lafferty J, Mccallum A, Pereira F. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data Abstract. 2001 Jun;1999:282-9,

[11] Okazaki N. CRFsuite: a fast implementation of Conditional Random Fields (CRFs); 2007. Available from: http://www.chokkan.org/software/crfsuite/.

[12] Doˇgan RI, Leaman R, Lu Z. NCBI disease corpus: A resource for disease name recognition and concept normalization. J Biomed Inform. 2014 Feb;47:1-10, doi: 10.1016/j.jbi.2013.12.006.

[13] Bird S, Klein E, Loper E. Natural Language Processing with Python. O'Reilly Media Inc. 2009.

[14] Honnibal M, Montani I, Landeghem SV, Boyd A. spaCy: Industrial-strength Natural Language Processing in Python. 2020.

[15] World Health Organisation. International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM); 2021. Available from: https://www.cdc.gov/nchs/icd/icd10cm.htm.

[16] Orphanet: an online database of rare diseases and orphan drugs, (2021). Available from: https://www.orphadata.com/alignments/.