MEDINFO 2021: One World, One Health – Global Partnership for Digital Innovation
P. Otero et al. (Eds.)

doi:10.3233/SHTI220239

Annotating Free-Texts in EHRs Towards a Reusable and Machine-Actionable Health Data Resource

Jiayang Wang[a], Lin Yang[a], Xiaoshuo Huang[a], Jiao Li[a]

[a] Institute of Medical Information and Library,
Chinese Academy of Medical Sciences/Peking Union Medical College, Beijing, China

Abstract

Electronic health records provide more insight and possibilities for retrospective studies with the help of data science techniques. However, the free-text in medical records limit the availability and reusability of data. Credentialed data that is contained in free-text also puts more restriction on data sharing and secondary use. We aim to propose a method for guiding EHRs free-text data annotation and sharing from the perspective of data safety and procedure standardization.

Keywords:

Natural Language Processing, Electronic Health Records.

Introduction

In the era of open science, accessible electronic health records (EHRs) are essential for data-driven medical research. It requires the medical data to be available, interoperable, and machine-actionable [1]. Open access EHRs can facilitate individual health management and medical service and population health research, potential medical discovery, health care facility management, and more. Therefore, free accessible EHRs databases like MIMIC (Medical Information Mart for Intensive Care) and eICU are published [2].

Confined by regulations or securities rules of data sharing like the Health Insurance Portability and Accountability Act of 1996 (HIPAA), individuals' privacy recorded in electronic form should be protected. Shared EHRs must be desensitized, and data users must work under relevant regulations. Unstructured data like discharge summaries containing verbal descriptions can still be sensitive after deidentification.

Free-text in EHRs is informative, but its existence reduces the reusability of EHRs. In order to make the text records machine-readable, metadata need to be pre-annotated by domain researchers or trained personnel. Annotation of medical free-text usually requires domain knowledge and is labor-intensive, which makes it expensive and time-consuming. Corpus used in EHRs NLP researches and challenge tasks are usually annotated by members of the group, and permissions of usage were bounded by restrictions.

Despite the powerful tool of NLP that allows retrospective studies to extract information from free-text of EHRs, utilization of EHRs under the frame of credential data protection has not been easy. Medical records publication are restricted for personal information protection, but this also limits secondary use of EHRs. The ever-increasing demand for learning algorithm-targeted medical NLP research urges the open access and sharing of annotated medical records.

Addresings the aforementioned scientific demands, there have been explorations of annotated medical free-text sharing. A successful example is Moseley et al. annotated 15 phenotypes in MIMIC-III notes. This dataset can be used for phenotyping identification based on medical notes. Here we propose a guideline contrapose the acquisition, annotation, and data sharing of medical free-text, based on MIMIC-III notes that focused on National Institute of Health Stroke Scale (NIHSS) extraction.

Methods

MIMIC-III clinical notes were used for NIHSS annotation. This guideline also focused on data safety and procedure normalization during EHRs handling. The flowchart in Figure 1 demonstrates the workflow.

Data acquisition

First, determine the scope of the research project and the corresponding data source that can apply to the project, or conversely, choose the data source first, then the research scope. Medical data should be collected under the restriction of local regulations by the specialized person with patients' consent. Researchers who work with open access databases should sign a corresponding confidentiality agreement. According to the predetermined research object, rearrange the dataset and narrow it down to research focused scope.

We used MIMIC-III intensive care unit data as the data source and focused the research scope down to extract the national institute of health stroke scale (NIHSS) from discharge summaries. All researchers who require handling the data have obtained the Collaborative Institutional Training Initiative training certificate and were granted access by the MIMIC-III committee.

Free-text data annotation

The annotation rule is applied to the linguistic data by annotators. If there is more than one annotator involved, the degree of agreement between annotators should be measured by the Kappa coefficient. The midway corpus can also be fed into ML models for prediction effect evaluation. Corpus evaluation could be iterative during annotation, and the original annotation rule should be adjusted accordingly. All personnel who have access to the data should follow the usage regulation of the data provider. In the end, all free-text metadata should be converted into a corpus with annotated features that are agreed upon between annotators.

We recruited two annotators with medical informatics backgrounds to mark NIHSS item and item relation. First, they annotated the same 100 discharge summaries back to back following the preliminary rule. Then the discrepancy was measured by Cohen's Kappa and adjudicated by a third researcher. All researchers in the project obtained an access certificate of MIMIC-III. If a clinician's opinion was needed, he/she was only consulted from a medical perspective. This process of annotation and adjudication was iteratively repeated until the Kappa coefficient reaches 0.901. Afterward, the two annotators finished the remaining annotation work to form a golden standard corpus.

Data sharing

Annotated corpus can only be handled by people inside the group with metadata access permission. The annotation process is usually carried out on different platforms, and tends to have different annotation file format. Suppose the research team intends to share the corpus with the community for secondary use. They should ensure the original data owner allows such conduct and the data license still applies. Shared data should be rearranged into an open-source format like .txt or .csv for machine-actionable purposes and open source requirements.

The degree of disagreement was measured, and within an acceptable range. Therefore, the final corpus can be considered as the golden standard. Then the single annotated files were rearranged into one single well structured .txt file for following ML task and data sharing. After consulted with the MMIC team, we were encouraged to re-contribute the corpus to the scientific community following the open science practice. The corpus was shared on PhysioNet, documented with data description, usage instruction, and most importantly, under the same license as MIMIC-III [3]

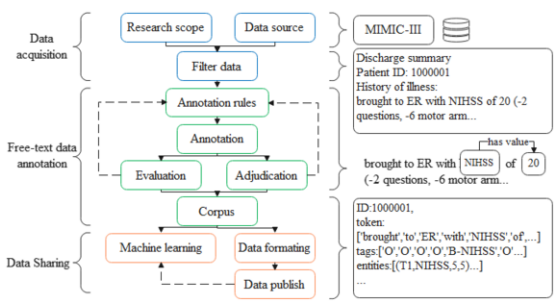

Figure 1– EHRs annotation flowchart and outcome

Results

By applying the EHRs free-text annotation guideline, we developed an NIHSS corpus that can be used for ML and DL tasks. The corpus was tested on a conditional random field followed by random forest models for an entity recognition task. The final f1 score of 0.9628 proves the corpus is suitable for this type of ML mission and may be applied for further DL exploit. Potential NLP approaches of the corpus include NIHSS identification, stroke evidence quantification, and disease phenotyping. With the consent from the team of the data source, the corpus was shared on PhysioNet: a free access physiological and clinical data platform where MIMIC-III has initially been published (https://physionet.org/content/stroke-scale-mimic-iii/1.0.0/). Overall, the NIHSS corpus construction was conducted under open science ideology and full concern of data safety. Furthermore, the data source of the corpus was originated from an open-access database and was enriched its reusability via the scientific communities' curation.

Conclusions

Unlike other scientific data, EHRs contains credentialed information of patients, which restricts its availability and reusability. Annotated free-text that only contains linguistic features can help desensitize data and increase reproducibility. By implementing the guideline we proposed for EHRs free-text annotation, the data were able to be handled with confidentiality. A re-contribution of originally credentialed corpus also fit open science requirements: availability and actionability of data.

Acknowledgements

This research is supported by the Beijing Natural Sci-ence Foundation (Grant No. Z200016), Chinese Academy of Medical Sciences (Grant No. 2017PT63010, 2018PT33024, 2018-I2M-AI-016).

References

[1] P. Wittenburg, Open Science and Data Science, *Data Intelligence* **3** (2021), 95-105.

[2] A.E. Johnson, T.J. Pollard, L. Shen, L.W. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L.A. Celi, and R.G. Mark, MIMIC-III, a freely accessible critical care database, *Sci Data* **3** (2016), 160035.

[3] A.L. Goldberger, L.A. Amaral, L. Glass, J.M. Hausdorff, P.C. Ivanov, R.G. Mark, J.E. Mietus, G.B. Moody, C.K. Peng, and H.E. Stanley, PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals, *Circulation* **101** (2000), E215-220.

Address for correspondence

Jiao Li, Institute of Medical Information, Chinese Academy of Medical Sciences, 3rd Yabao Road, Chaoyang District, Beijing 100020, China. Email: li.jiao@imicams.ac.cn