BRATECA: a Clinical Information Dataset for the Portuguese Language

Bernardo S. Consoli*, Henrique D.P. dos Santos†, Ana Helena D.P.S Ulbrich†, Renata Vieira‡, Rafael H. Bordini*

* Pontifical Catholic University of Rio Grande do Sul, Brazil
† Institute for Artificial Intelligence in Healthcare, Brazil; ‡ University of Évora, Portugal

Abstract

This work presents a new Brazilian Clinical Dataset containing over 70,000 admissions from 10 hospitals in two Brazilian states, composed of a total of over 2.5 million free-text clinical notes alongside data pertaining to patient information, prescription information, and exam results. This data was collected, organized, deidentified, and is being distributed via credentialed access for the use of the research community by the Institute for Artificial Intelligence in Healthcare.

Data Records

BRATECA is an edited and reorganized version of the Institute for Artificial Intelligence in Healthcare's own internal information database. It contains 73,040 admission records of 52,973 unique patients. Amongst those admissions, several are associated with specialty treatment wards, as follows: publicly funded wards (12,096 admissions); intensive care wards (4,666 admissions); obstetrics wards (5,550 admissions); COVID-19 wards (1,714 admissions); surgical wards (25,004 admissions); emergency wards (37,392 admissions); and ambulatory wards (3,107 admissions). The remaining 8,674 admissions we unassociated.

As it is intended to be an accessible edition for use in machine learning research, the most useful internal data were reorganized into 5 modules. These are: Admission (includes administrative information and patient demographic data), Clinical Note (includes free-text clinical notes on details of the patient's stay and treatment), Exam (includes medical exams and their respective results), Prescription (includes pharmacy assessments, prescription date, expiration date, administrative details about the prescribed drugs), and Prescription Item (includes details of each prescribed medication, including name, dosage, and information on how the medication is to be administered).

All entries in the 5 modules of BRATECA have IDs referring to which admission, patient and hospital the entry relates to. All IDs are numerical in nature and do not reveal any personal medical information.

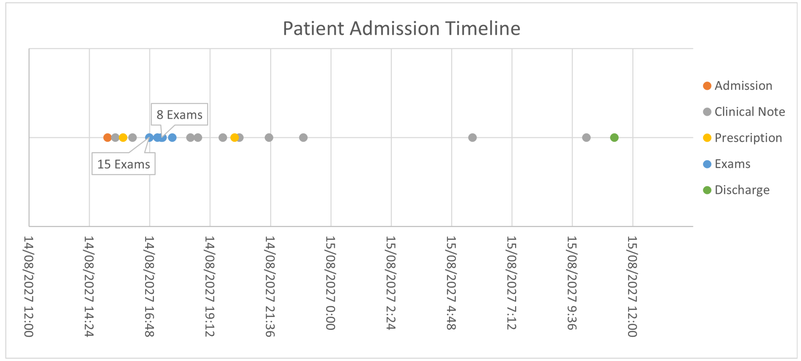

Figure 1. A simple example timeline of an admission, including recorded time of admission, laboratory examination, prescription administration, clinical note writing, and discharge. The two labels represent two instances where events were logged at the same time. In these cases, 15 and 8 exam results were logged simultaneously at two separate points in the timeline.

Data Acquisition

The Institute for Artificial Intelligence in Healthcare's database is centered around its prescription tables. This resulted in only admissions with prescriptions being extracted, as the prescription tables contained ward information and were the best way to ascertain that only adult patients from the desired wards were extracted from the database.

Beyond those requirements, only admissions which both began and ended during a delimited time period of nine months between 2020 and 2021 were extracted.

Deidentification

Though most columns in the datasets provide the exact information present in the original database, some had to be modified to further protect patient's sensitive information and attempt to prevent reverse engineering of identities from the provided data.

Names in the free text notes were deidentified using state-of-the-art Bi-LSTM-CRF deep neural networks (Akbik et al. 2018). Dates were shifted randomly 5 to 10 years forward. Dates referring to the same admission were shifted the same number of days forward. All internal database IDs, such as those for Patient ID or Admission ID, were replaced with unrelated numeral IDs coherent between admissions. Finally, non-identifiable ward labels were generated using the actual names of the wards of the hospitals from which the information was collected.

Data Access

BRATECA is distributed by the Institute for Artificial Intelligence in Healthcare through PhysioNet credentialed access. In order to receive access, the researcher must acquire PhysioNet credentials, request access to the dataset, and wait for approval by the Institute for Artificial Intelligence in Healthcare.

Example Usage

Several papers have been published using information now available through BRATECA. Examples includes “Evaluation of a Prescription Outlier Detection System in Hospital's Pharmacy Services” by Santos et al. 2021 and “Case Report of Drug-Induced Liver Injury in a Patient with Covid-19” by Senter et al. 2021.

The information can also be used for clinical prediction tasks such as mortality prediction and length-of-stay prediction. Furthermore, the free-text clinical notes can, for example, be used to train language models.

References

- Akbik, A., Blythe, D., and Vollgraf, R. (2018). Contextual string embeddings for sequence labeling. In COLING 2018, 27th International Conference on Computational Linguistics, pages 1638–1649.
- D. P. dos Santos, H., D. P. S. Ulbrich, A. H., and Vieira, R. (2021). Evaluation of a prescription outlier detection system in hospital’s pharmacy services. In 12th International Workshop on Biomedical and Health Informatics (BHI).
- Senter et. al, E. (2021). Case report of drug-induced liver injury in a patient with covid-19. In Exhibition of Successful Experiences and Research on the Rational Use of Medicines and Health Education.