Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 5609–5616
Marseille, 20-25 June 2022

BRATECA (Brazilian Tertiary Care Dataset): a Clinical Information Dataset for the Portuguese Language

Bernardo S. Consoli[1], **Henrique D. P. dos Santos**[2], **Ana Helena D. P. S. Ulbrich**[2], **Renata Vieira**[3], **Rafael H. Bordini**[1]
[1]Pontifical Catholic University of Rio Grande do Sul (PUCRS), Brazil
[2]Institute for Artificial Intelligence in Healthcare, Brazil
[3]CIDEHUS, University of Évora, Portugal
bernardo.consoli@acad.pucrs.br, {henrique, anahelena}@noharm.ai,
renatav@uevora.pt, rafael.bordini@pucrs.br

Abstract

Computational medicine research requires clinical data for training and testing purposes, so the development of datasets composed of real hospital data is of utmost importance in this field. Most such data collections are in the English language, were collected in anglophone countries, and do not reflect other clinical realities, which increases the importance of national datasets for projects that hope to positively impact public health. This paper presents a new Brazilian Clinical Dataset containing over 70,000 admissions from 10 hospitals in two Brazilian states, composed of a sum 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. In the course of presenting the new dataset, this paper will explore the new dataset's structure, population, and potential benefits of using this dataset in clinical AI tasks.

Keywords: computational medicine, portuguese language, clinical data, tertiary care

1. Introduction

Decision making in healthcare scenarios has been a topic of increasing interest in the field of artificial intelligence (Shamout et al., 2021; Si et al., 2021). Studies into methods for predicting diseases (Song et al., 2018; Xu et al., 2018), mortality (Xu et al., 2018), length-of-stay (Song et al., 2018; Xu et al., 2018), admission (Liu et al., 2019), and interventions (Suresh et al., 2017) have become more common with the widespread use of EHRs (Electronic Health Records) in hospitals worldwide, which in turn led to efforts to deidentify this information, as it is sensitive and extremely personal, in order to make it available for use in related research (Shamout et al., 2021).

These efforts have resulted in several clinical datasets, most of which were constructed from sources in the English language. The MIMIC (Medical Information Mart for Intensive Care) collection, composed of data extracted from the Beth Israel Deaconess Medical Center in the United States, is a part of PhysioNet (Goldberger et al., 2000) and is the foremost example of such datasets in computational medicine and focuses on intensive care patients. Many papers (Song et al., 2018; Xu et al., 2018; Liu et al., 2019; Suresh et al., 2017) make use of this collection to build training and testing datasets for deep learning architectures. MIMIC-IV (Johnson et al., 2020), the latest edition of the collection, possesses comprehensive information about each patient, including laboratory measurements, administered medication, documented vital signs, etc.

The present work's goal is to introduce a large collection of clinical data akin to the MIMIC collection but for Brazilian Portuguese instead. Being focused on clinical notes in the Portuguese language, it can be used for any project focusing on Brazilian clinical scenarios. This new dataset, henceforth referred to as BRATECA (BRAzilian TErtiary CAre dataset), boasts more than 400 million words across over 2.5 million free-text clinical notes from over 70,000 individual admissions in 10 different hospitals located in two Brazilian states. The dataset also possesses patient, prescription, and exam information for these admissions, when available. This data is collected, deidentified and managed by the Institute for Artificial Intelligence in Healthcare[1], a non-profit startup from Brazil composed of an interdisciplinary team of data scientists and practicing healthcare professionals such as pharmacists and physicians that develop smart systems for clinical pharmacy. The dataset has been made available by them for credentialed access.

2. Related Work

One of the most widely used clinical datasets is MIMIC. It has several versions, and its most current iteration, MIMIC-IV, is separated into 6 modules: core, hosp, icu, ed, cxr, and note. The *core* module is composed of patient demographics, hospitalization records, and ward stay records. The *hosp* module is composed of data recorded during the patient's hospital stay such as lab measurements, medication administration and prescription, billing information, etc. The *icu* module is composed of data taken from patients in intensive

[1]https://noharm.ai/en

care units (ICUs), and include intravenous and fluid inputs, patient outputs, procedures, date and time information, etc. The *ed* module is composed of data from emergency department (ED) patients, and includes reason for admission, triage assessment, vital signs, etc. The *cxr* module contains chest x-ray (CXR) images from ED patients from multiple viewpoints. Finally, the *note* module contains patient's deidentified free-text clinical notes for hospitalization, although this module is not yet available to the public.

Another example is the United Kingdom's National Health Service's (NHS) comprehensive dataset collection[2]. The data is collected in order to support the analysis of specific policies of interest as well as the effects of particular policy initiatives, and it is separated into several different datasets, each with a different focus and different kinds of data.

A more task-focused example of English language clinical dataset can be found in the National NLP Clinical Challenges (n2c2) datasets. These challenges have been proposed since 2006, starting with the i2b2 project, n2c2's predecessor. These two series of challenges have presented datasets for a variety of tasks, such as deidentification, obesity prediction, coreference, temporal relations, heart disease, clinical semantic textual similarity, and family history extraction. The current edition, n2c2 2022[3], proposes three tracks: Contextualized Medication Event Extraction; Extracting Social Determinants of Health; and Progress Note Understanding: Assessment and Plan Reasoning. Task-specific datasets were released alongside each of these challenges, though some, such as the current challenge's third track, make use of already available resources (MIMIC-III in this case) when they are appropriate for the proposed task.

However, these are English language datasets extracted from hospitals in certain anglophone countries, and do not conform to the clinical realities of Brazil. It is thus important to gather national data for local research projects which may be able to positively impact Brazilian public health. The development of national clinical resources has started in earnest in recent years, with work such as SemClinBR (e Oliveira et al., 2020), a dataset with 1000 clinical notes annotated with over 65,000 entities and over 11,000 relations. The dataset was manually annotated and may be used for a variety of tasks, such as clinical named entity recognition and negation detection. It bears more resemblance to the n2c2 challenge datasets than to MIMIC.

BioBERTpt (Schneider et al., 2020) is a fine-tuned BERT model trained on clinical EHR texts as well as texts from the biomedical literature. It has three versions, each trained with a different corpus. The first was trained with more than 2 million clinical notes from Brazilian hospitals collected between 2002 and 2018. The second with titles and abstracts from Portuguese biomedical scientific papers published in PubMed and Scielo. A third version combining both corpora into one was also trained. The clinical note corpus does not seem to have been made available after its use in training the models.

The literature also covers a Brazilian healthcare image dataset, the labeled chest X-ray dataset BRAX (Reis et al., 2022). Although it is not a language resource, that dataset is nonetheless an example of a Brazilian healthcare dataset, and it is similar to MIMIC's CXR (chest X-ray), except that the images are not complemented by text-based healthcare resources like MIMIC's.

Another example of a Portuguese-language health-related dataset was developed by de Melo and Figueiredo (2020). Their Twitter-based dataset is composed of nearly 4 million tweets and about 18,000 news articles related to COVID-19 in Brazil. It has a different domain from the other datasets presented thus far and so has a different overall purpose, being more focused on public discourse and sentiment about public health issues rather than clinical information.

3. BRATECA

BRATECA contains 73,040 admission records of 52,973 unique adults (18 years of age or older) extracted from 10 hospitals located in two Brazilian states. Amongst those admissions, several are associated with specialty treatment wards, as follows: publicly funded wards (12,096 admissions total); intensive care wards (4,666 admissions total); obstetrics wards (5,550 admissions total); COVID-19 wards (1,714 admissions total); surgical wards (25,004 admissions total); emergency wards (37,392 admissions total); and ambulatory wards (3,107 admissions total). The remaining 8,674 admissions associated with any specialty wards.

The median patient age is 54 (Q1 = 38, Q3 = 68), 41.3% of the patients are male, 70.7% are identified as white, 3.8% are identified as mixed, 3.8% are identified as black and 0.2% are identified as yellow, and the mortality rate of patients is 6.5%. Each admission is paired with laboratory exam results (2,374,807 total), prescriptions and their itemized contents (519,318 total), and clinical notes (2,849,572 total). An interactive dashboard has been created to present some details of BRATECA and is linked in the project's GitHub page[4]. Table 1 presents statistics for each admission type.

3.1. Classes of Data

BRATECA is composed of descriptive data, laboratory data, medication data, intervention data, and clinical notes. Descriptive data includes patient specific information such as dates of birth, admission and discharge, skin color, height, weight, and reasons for dis-

[2] https://digital.nhs.uk/data-and-information/data-collections-and-data-sets/data-sets

[3] https://n2c2.dbmi.hms.harvard.edu/2022-challenge

[4] https://github.com/noharm-ai/brateca

Admission Type	Publicly Funded	Intensive Care	Obstetrics	COVID-19	Surgical	Emergency	Ambulatory	Normal
Median Age (Q1-Q3)	58 (38-69)	64 (52-73)	30 (25-36)	61 (49-73)	56 (40-68)	54 (38-69)	44 (34-59)	56 (40-70)
Median Laboratory Results (Q1-Q3)	25 (0-53)	119 (48-263)	0 (0-17)	117 (54-290)	0 (0-10)	17 (0-29)	0 (0-0)	0 (0-19)
Median Prescriptions (Q1-Q3)	3 (1-10)	28 (15-52)	3 (1-6)	15 (8-36)	2 (1-6)	1 (1-4)	1 (1-2)	3 (1-7)
Median Clinical Notes (Q1-Q3)	12 (3-57)	140 (77-291)	11 (3-42)	106 (54-231)	5 (2-26)	4 (2-14)	2 (2-5)	19 (9-32)
Male Percentage	42.2%	55.48%	0.14%	55.54%	41.92%	43.44%	31.48%	41.71%
Mortality Percentage	5.13%	24.09%	0.07%	17.68%	2.44%	10.62%	0.19%	1.44%
Skin Color Percentages	W: 66.32% B: 8.59% M: 9.22% Y: 0.21% NI: 15.67%	W: 67.10% B: 2.48% M: 3.36% Y: 0.06% NI: 27.00%	W: 68.81% B: 11.68% M: 9.44% Y: 0.13% NI: 9.95%	W: 78.65% B: 4.03% M: 3.79% Y: 0.23% NI: 13.30%	W: 83.25% B: 3.26% M: 2.90% Y: 0.14% NI: 10.46%	W: 60.50% B: 4.25% M: 4.65% Y: 0.11% NI: 30.48%	W: 83.71% B: 3.41% M: 2.22% Y: 0.29% NI: 10.46%	W: 78.20% B: 2.02% M: 1.59% Y: 0.16% NI: 18.03%

Table 1: Rows present each the following information, from top to bottom: Median age (Q1 through Q3) per admission type; Median number of laboratory results per patient per admission type; Median number of prescriptions per patient per admission type; Median number of clinical notes per patient per admission type; Percentage of male patients per admission type; mortality percentage per admission type; percentages for patient skin color identification (W is white, B is black, M is mixed, Y is yellow and NI is no information). Columns each present one type of ward. Wards deemed "normal" are those that do not fall into any of the other categories. Note that a single admission may have a patient move wards one or more times, and a single ward may belong to more than one category.

charge. Laboratory data include data on various laboratory exam results for patients. Medication data includes prescription items, as well as dosage, frequency, and other such administration details specific to each patient and prescription. Intervention data includes notes on whether there were pharmacist interventions on specific prescriptions that may have been mistakenly administered, as identified by the Institute for Artificial Intelligence in Healthcare's NoHarm.ai clinical pharmacy AI system (D. P. dos Santos et al., 2021). Notes are free-text clinical notes describing a patient's evolving hospital admission.

4. Development Methods

BRATECA is an edited and reorganized version of the Institute for Artificial Intelligence in Healthcare's own internal Brazilian tertiary care information database and is intended to be a public[5] edition for use in machine learning research. For this purpose, certain data tables deemed most useful at the time of extraction were reorganized into the 5 datasets of BRATECA. This section presents the process of extraction and deidentification of the database's information into the format presented in Section 5.

4.1. Dataset Organization

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 were extracted. This time period was set to sometime between 2020 and 2021, but this will not be specified so as to further enhance patient privacy. All admissions that fit within the presented parameters had their IDs extracted and used to gather related data from the database and create the 5 separate but interconnected datasets: Admission, Exam, Clinical Note, Prescription, and Prescription Item. These datasets are further described in Section 5. The SQL scripts used to extract the data are available in the project's GitHub page[6].

4.2. 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.

All names in BRATECA's free text notes were deidentified using state-of-the-art deep learning methods (Bi-LSTM-CRF) (Akbik et al., 2018). Two corpora and three language models were evaluated on a Named Entity Recognition (NER) task focused on person names to evaluate which combination delivered the best performance. The experiments revealed that using domain-specific corpora (focused on deidentification of clinical notes) and a contextualized embedding stacked with word embeddings achieved the best results: an F-measure of 0.94 and Recall of 0.95 (Santos et al.,

[5]Note that BRATECA is property of the Institute for AI in Healthcare and only credentialed access is allowed, but it is freely available for research use.

[6]See footnote 4

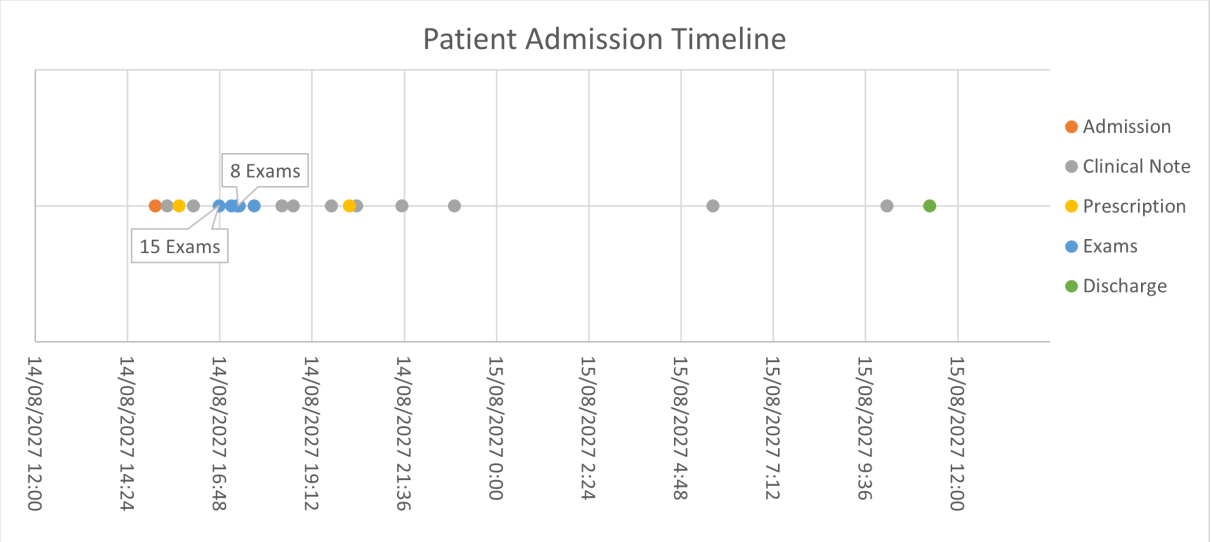

Figure 1: A simple example timeline of an admission, including recorded time of admission, laboratory examination, prescription administration, clinical note writing, and discharge. The two labels represent two instances where events were logged at the same time. In these cases, 15 and 8 exam results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

2021). Dates present in the free text notes were also removed, though not using NER but rather regular expressions. The date removal script is available on the project's GitHub page[7].

Furthermore, all dates not part of free-text notes were shifted randomly 5 to 10 years forward. Dates referring to the same admission were shifted the same amount of days forward (i.e. if admission "1" of the Admission dataset was shifted 100 days forward, all dates of all entries in the other 4 datasets which refer to admission "1" in their Admission ID field were shifted 100 days forward as well). This was done in order to maintain timeline coherence within the same admission. Note that multiple admissions of the same patient may not be in chronological order and do not maintain any sort of temporal relation in order to more thoroughly deidentify such patients.

All internal database IDs, such as those for Patient ID or Admission ID, were also deidentified. Each was assigned a random numerical ID, congruent between datasets (i.e. if Admission ID "123456" is assigned the new ID "789" in the Admission dataset, the Admission ID "123456" was also assigned "789" whenever it appeared in the other 4 datasets).

Finally, ward information[8] was generated using the actual names of the wards of the hospitals from which the information was collected. Ward names were replaced with the aforementioned labels in order to better prevent hospital identification while maintaining some of the more relevant information. The generation was performed with the help of an active healthcare professional.

4.3. Code Availability

All of the code associated with BRATECA (algorithms, sample datasets, language models, and experiments) is available on the project's GitHub page[9] for replicability purposes.

5. Data Records

BRATECA is composed of 5 datasets in the CSV (Comma Separated Values) format. These are as follows: Admission, a dataset of every individual admission, which includes patient demographic data; Exam, a dataset of exams and their respective results performed for each admission; Prescription, a dataset of prescription headers, which includes information such as patient/admission ID for the patient/admission which received the prescription, pharmacy assessments, prescription date, expiration date, ward information, whether the prescription includes special medication such as controlled substances, intravenously administered drugs (IV drugs), and antibiotics; Prescription Item, a dataset of prescribed medications which includes details of each prescribed medication, including name, dosage, and information on how the medication is to be administered, with each entry of this dataset being directly related to a prescription header in the Prescription dataset; and Clinical Note, a dataset of free-text clinical notes on details of the patient's stay and treatment. A simple example of a patient timeline which shows details from all datasets in conjunction can be seen in Figure 1.

[7] See footnote 4

[8] Public, IC, Obstetrics, COVID-19, Surgical, Emergency and Ambulatory. See the Prescription row of Table 2 for further details.

[9] See footnote 4.

Dataset	*Column*	Description	*Column*	Description
Admission	*Hospital_ID*	The identification code for the hospital from which the data originated.	*Patient_ID*	The identification code for the patient for whom the admission was registered.
	Admission_ID	The identification code for the admission to which the information belongs.	*Date_of_Birth*	Patient's date of birth.
	Gender	Patient's gender.	*Admission_Date*	Date patient was admitted to hospital.
	Skin_Color	Patient's skin color.	*Height*	Patient's height.
	Weight	Patient's weight.	*Height_Date*	Date patient's height was measured.
	Weight_Date	Date the patient was weighted.		
Exam	*Hospital_ID*	The identification code for the hospital from which the data originated.	*Patient_ID*	The identification code for the patient for whom the admission was registered.
	Admission_ID	The identification code for the admission to which the information belongs.	*Exam_Name*	Name of the exam that was performed.
	Exam_Date	Date the exam was performed	*Value*	Numerical value of the result of the exam.
	Unit	Unit of measurement the exam's Value is in.		
Clinical Note	*Hospital_ID*	The identification code for the hospital from which the data originated.	*Patient_ID*	The identification code for the patient for whom the admission was registered.
	Admission_ID	The identification code for the admission to which the information belongs.	*Note_Date*	Date the note was written.
	Note Text	The contents of the note.	*Notetaker_Position*	Notetaker's job title.
Prescription	*Hospital_ID*	The identification code for the hospital from which the data originated.	*Patient_ID*	The identification code for the patient for whom the admission was registered.
	Admission_ID	The identification code for the admission to which the information belongs.	*Prescription_ID*	The identification code for the prescription to which prescription items are associated.
	Prescription_Date	Date the prescription note was written.	*Pharmacy_Assessment*	Whether the prescription was revised by a pharmacist.
	Expiration_Date	Prescription expiration date.	*Assessment_Date*	Date the pharmacy assessment was performed.
	Allergy	Whether patient is allergic to one or more of the prescribed medications.	*Prescription_Score*	Score generated by artificial intelligence (the higher the score, the more unusual the prescription).
	Alerts	Prescription alerts. A complete list of alerts is shared in the documentation.	*Score_One*	The quantity of prescription items given a "1" score by the AI.
	Antibiotics	Number of antibiotics prescribed.	*Score_Two*	The quantity of prescription items given a "2" score by the AI.
	High_Alert	Number of high alert medication prescribed.	*Score_Three*	The quantity of prescription items given a "3" score by the AI.
	Controlled	Number of controlled medication prescribed.	*Tube*	Number of IV drugs prescribed.
	Not_Default	Number of non-standard medications prescribed.	*Different_Drugs*	Number of prescribed medications not previously reviewed by a pharmacist.
	Alert_Exams	Alerts related to exams. Examples can be found in the documentation.	*Interventions*	Number of interventions related to the prescription.
	Complication	Number of complications detected in clinical notes related to the prescription.	*Public*	Whether or not the prescription is for a publicly funded ward.
	IC	Whether or not the prescription is for an Intensive Care ward.	*Obstetrics*	Whether or not the prescription is for a obstetrics ward.
	COVID-19	Whether or not the prescription is for COVID-19 ward.	*Surgical*	Whether or not the prescription is for a surgical recovery ward.
	Emergency	Whether or not the prescription is for an emergency ward.	*Ambulatory*	Whether or not the prescription is for an ambulatory ward.
Prescription Item	*Hospital_ID*	The identification code for the hospital from which the data originated.	*Patient_ID*	The identification code for the patient for whom the admission was registered.
	Admission_ID	The identification code for the admission to which the information belongs.	*Prescription_ID*	The identification code for the prescription to which prescription items are associated.
	Drug_Name	Name of the drug.	*Dosage*	Dosage of each administration.
	Daily_Frequency	Number of times a drug is administered per day.	*Administration_Route*	Route of drug administration.
	Note	Medical observations related to the prescription.	*Normalized_Dosage*	Dose converted to a single numerical unit.
	Time	Time each dose is to be administered.	*Source*	Whether it is nutrition, a drug, a procedure drug or a solution.
	Suspension_Date	Date the medication is to be suspended.	*(Solution)_Group*	Group to which the solution belongs.
	(Solution)_at_Medical_Discretion	Medical observations related to the prescribed solution.	*(Solution)_Steps*	Frequency of solution adminitration.
	(Solution)_Hour	Time each solution dose is to be administered.	*(Solution)_App_Time*	How long a solution is to be administered for.
	(Solution)_Dosage	The dosage of the solution.	*(Solution)_Unit*	The unit of measurement of the dosage.
	Administration_Period	The period during which the item is to be administered.	*Allergy*	Whether the patient is allergic to the prescription item.
	Tube	Whether the prescription item is administered intravenously.	*(Intervention)_Date*	Date of the intervention
	(Intervention)_Note	Medical observations related to the intervention.	*(Intervention)_Status*	Resolution of the intervention request.
	(Intervention)_Update	Date of the final intervention update.	*(Intervention)_Motive*	Motive of the intervention.
	(Intervention)_Error	Intervention considered a prescription error.	*(Intervention)_Cost*	Intervention that generated a reduction of costs.

Table 2: Columns and descriptions of columns for each of the 5 datasets.

All datasets have IDs that are used for identification of relations between entries in each file. These are: Hospital ID, the identification for the hospital from which the raw data was collected; Patient ID, the ID for a given patient in the database; Admission ID, the ID for the patient's admissions, of which a single patient might have many; and Prescription ID, specific to the Prescription and Prescription Item datasets, which identifies prescription items as belonging to specific prescriptions.

5.1. Datasets

The datasets were developed in the way described above so that they can be used separately as well as in conjunction. Each is composed of several columns from tables in the original database, organized for ease of use. The information in each of the 5 datasets are presented in Table 2.

6. Usage Notes

6.1. Data Access

As mentioned previously, BRATECA is distributed by the Institute for Artificial Intelligence in Healthcare through Physionet credentialed access. In order to receive access, the researcher must complete the following steps:

1. sign in to and confirm your identity in the Physionet platform
2. complete a course on protecting human research participants;
3. if the requester is a student, their supervisor must also agree to the terms of confidentiality;
4. access the BRATECA page in Physionet[10] and request access to the dataset;
5. wait for approval by the Institute for Artificial Intelligence in Healthcare.

Once the process is complete, and if the request is accepted, the researcher will be granted access to the dataset files.

6.2. Example Usage

There are many tasks which could benefit or even require datasets such as BRATECA. Prediction tasks, such as those mentioned in Section 1, can use these datasets for training purposes. Mortality prediction can use discharge information as mortality annotation, for example.

Researchers with access to the original database have already published several papers with the information which is to be released in BRATECA. Santos et al. (2021), for example, used state-of-the-art methods to identify and remove names from clinical texts. These were the methods were used to deidentify all free-text notes made available as part of BRATECA, as mentioned in Section 4.2. Other examples of previous use of the data are listed below:

- Evaluation of a Prescription Outlier Detection System in Hospital's Pharmacy Services (D. P. dos Santos et al., 2021);
- Case Report of Drug-Induced Liver Injury in a Patient with Covid-19 (Senter et. al, 2021);
- Analysis of Pharmaceutical Interventions Performed with Decision Support Using Artificial Intelligence in Brazilian Hospitals (D. P. S. Ulbrich et al., 2021).

Besides published papers, much research work making use of BRATECA is well under way. Some examples are listed below:

- A machine learning-based clinical decision support system to identify possible drug intervention;
- Detection of Drug-Induced Liver Injury;
- Trends in the use of corticosteroids during the Pandemic.

Finally, several other usages of the dataset are being investigated or set to be explored in the near future. The large amount of free text notes, for example, permits the training of domain-specific language models with word embedding architectures such as Word2Vec (Mikolov et al., 2013) and fastText (Grave et al., 2017), and also contextual embedding models such as ELMO (Peters et al., 2018) and BERT (Devlin et al., 2019). Embeddings like these can be even more specific, using only certain parts of the data, such as limiting training to texts about elderly patients or intensive care patients.

Another avenue of research being explored is the use of the information to create real-time digital twins of patients by utilizing representation learning technology. These digital twins could be used to predict patient developments and aid medical workers keep track of the most important information for each of their patients via alerts, data organization, and information retrieval (Shamout et al., 2021).

6.3. Continuous Development

BRATECA will undergo continuous maintenance and development to ensure the high quality of the data made available for researchers to further the national public good. The maintenance team will encourage the user community to aid in this important endeavor as well, so as to continue to improve the quality of available clinical data for Portuguese language research.

These continuous efforts will, for example, be aimed at ensuring that the data is fully deidentified. Researchers granted access to the data will be asked to report any

[10] https://doi.org/10.13026/v8a6-mr20

deidentification errors they might find so that the patient data is kept as private as it can be while still being useful to data scientists and machine learning experts.
The team will also look into creating annotated subsets from the free-text notes for several tasks relevant for clinical text research, such as named entity recognition, text labeling, coreference, semantic textual similarity, and others. The user community will also be encouraged to allow any such annotated subsets they might themselves develop to be distributed via credentialed access by the Institute for Artificial Intelligence in Healthcare alongside the original version of BRATECA.
Finally, the team will endeavor to develop further versions of BRATECA by adding more data as it becomes available, and making improvements to the dataset according to community feedback. These future versions will be released alongside the original version of BRATECA, similarly open to credentialed access.

7. Ethical Concerns

BRATECA has been deidentified according to the Health Insurance Portability and Accountability Act (HIPAA) standards using structured data cleansing and date shifting. The NoHarm.ai system, developed by the Institute for Artificial Intelligence in Healthcare, gathers no identifiable information from patients.
The data used in the experiments we conducted for this article came from a research project developed with several hospitals in Brazil. Also, all data sharing was approved by each hospital participating in that research. Ethical approval to use the hospitals' datasets in this research was granted by the National Research Ethics Committee under the number 46652521.9.0000.5530.

Acknowledgements

We gratefully acknowledge partial financial support by CNPq under project 25/2020, CAPES, the Institute of Artificial Intelligence in Healthcare, and the FCT under project UIDB/00057/2020 (Portugal).

8. Bibliographical 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)*.

D. P. S. Ulbrich, A. H., Aline Maciel dos Santos, K., Dias Pereira dos Santos, H., and Zanella Lazaretto, F. (2021). Analysis of pharmaceutical interventions performed with decision support using artificial intelligence in brazilian hospitals. In *XIII Brazilian Congress of Hospital Pharmacy*.

de Melo, T. and Figueiredo, C. M. (2020). A first public dataset from brazilian twitter and news on covid-19 in portuguese. *Data in Brief*, 32:106179.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, et al., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

e Oliveira, L. E. S., Peters, A. C., da Silva, A. M. P., Gebeluca, C. P., Gumiel, Y. B., Cintho, L. M. M., Carvalho, D. R., Hasan, S. A., and Moro, C. M. C. (2020). Semclinbr - a multi institutional and multi specialty semantically annotated corpus for portuguese clinical NLP tasks. *CoRR*, abs/2001.10071.

Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., and Stanley, H. E. (2000). Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220.

Grave, E., Mikolov, T., Joulin, A., and Bojanowski, P. (2017). Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 427–431.

Johnson, A., Bulgarelli, L., Pollard, T., Horng, S., Celi, L., and Mark, R. (2020). Mimic-iv (version 1.0).

Liu, L., Li, H., Hu, Z., Shi, H., Wang, Z., Tang, J., and Zhang, M. (2019). Learning hierarchical representations of electronic health records for clinical outcome prediction. In *AMIA 2019, American Medical Informatics Association Annual Symposium, Washington, DC, USA, November 16-20, 2019*. AMIA.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 27th Annual Conference on Neural Information Processing Systems*, pages 3111–3119.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 16th Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technologies*, pages 2227–2237.

Reis, E. P., Paiva, J., Bueno da Silva, M. C., Sousa Ribeiro, G. A.and Fornasiero Paiva, V., Bulgarelli, L., Lee, H., dos Santos, P. V., Brito, B., Amaral, L., Beraldo, G., Haidar Filho, J. N., Teles, G., Szarf, G., Pollard, T., Johnson, A., Celi, L. A., and Amaro, E. (2022). Brax, a brazilian labeled chest

x-ray dataset (version 1.0.0). *PhysioNet*.

Santos, J., dos Santos, H. D., Tabalipa, F., and Vieira, R. (2021). De-identification of clinical notes using contextualized language models and a token classifier. In *Brazilian Conference on Intelligent Systems*, pages 33–41. Springer.

Schneider, E. T. R., de Souza, J. V. A., Knafou, J., Oliveira, L. E. S. e., Copara, J., Gumiel, Y. B., Oliveira, L. F. A. d., Paraiso, E. C., Teodoro, D., and Barra, C. M. C. M. (2020). BioBERTpt - a Portuguese neural language model for clinical named entity recognition. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 65–72, Online, November. Association for Computational Linguistics.

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*.

Shamout, F., Zhu, T., and Clifton, D. A. (2021). Machine learning for clinical outcome prediction. *IEEE Reviews in Biomedical Engineering*, 14:116—-126.

Si, Y., Du, J., Li, Z., Jiang, X., Miller, T., Wang, F., Zheng, W. J., and Roberts, K. (2021). Deep representation learning of patient data from electronic health records (ehr): A systematic review. *Journal of Biomedical Informatics*, 115.

Song, H., Rajan, D., Thiagarajan, J. J., and Spanias, A. (2018). Attend and diagnose: Clinical time series analysis using attention models. In Sheila A. McIlraith et al., editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 4091–4098. AAAI Press.

Suresh, H., Hunt, N., Johnson, A. E. W., Celi, L. A., Szolovits, P., and Ghassemi, M. (2017). Clinical intervention prediction and understanding with deep neural networks. In Finale Doshi-Velez, et al., editors, *Proceedings of the Machine Learning for Health Care Conference, MLHC 2017, Boston, Massachusetts, USA, 18-19 August 2017*, volume 68 of *Proceedings of Machine Learning Research*, pages 322–337. PMLR.

Xu, Y., Biswal, S., Deshpande, S. R., Maher, K. O., and Sun, J. (2018). RAIM: recurrent attentive and intensive model of multimodal patient monitoring data. In Yike Guo et al., editors, *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*, pages 2565–2573. ACM.