Medical & Biological Engineering & Computing
https://doi.org/10.1007/s11517-024-03251-4

ORIGINAL ARTICLE

Predicting hospitalization with LLMs from health insurance data

Everton F. Baro[1,2] · Luiz S. Oliveira[1] · Alceu de Souza Britto[3]

Received: 9 May 2024 / Accepted: 19 November 2024

Abstract
Predictions of hospitalizations can help in the development of applications for health insurance, hospitals, and medicine. The data collected by health insurance has potential that is not always explored, and extracting features from it for use in machine learning applications requires demanding processes and specialized knowledge. With the emergence of large language models (LLM) there are possibilities to use this data for a wide range of applications requiring little specialized knowledge. To do this, it is necessary to organize and prepare this data to be used by these models. Therefore, in this work, an approach is presented for using data from health insurance in LLMs with the objective of predict hospitalizations. As a result, pre-trained models were generated in Portuguese and English with health insurance data that can be used in several applications. To prove the effectiveness of the models, tests were carried out to predict hospitalizations in general and due to stroke. For hospitalizations in general, F1-Score = 87.8 and AUC = 0.955 were achieved, and for hospitalizations due to stroke, the best model achieved F1-Score = 88.7 and AUC of 0.964. Considering the potential for use, the models were made available to the scientific community.

Keywords Health insurance · Hospitalization · Strokes · Machine learning · Large language models · BERT · RoBERTa · LLaMA

1 Introduction

In recent years, natural language processing (NLP) has experienced significant advances. One of the most notable innovations was the emergence of LLMs (large language models), which have reached the state of the art in several tasks in this area. LLMs use a neural network architecture called transformer that was introduced in 2017 by Vaswani [1], being highly scalable and efficient, allowing LLMs to be trained on large amounts of data.

In the health area, in its most diverse subareas, data accumulates in large quantities, much of which is non-textual, which theoretically would make its use with LLMs unfeasible. However, the relationship between patients and health service providers, such as health insurance, hospitals, laboratories, among others, allows the recording, through their systems, in some way, of patients' health problems over time. The sequence of these contacts with these services, if organized chronologically, can closely resemble the textual structure normally used in LLM training.

Therefore, considering that it is common for health problems and illnesses to be related to previous illnesses and occurrences, organizing this data in this way becomes pertinent when trying to analyze and predict future situations. Furthermore, the transformers' attention mechanisms present in LLMs allow finding these relationships in data organized in this way, just as they do with common text sentences. Therefore, generating LLM models from structured health data, organized chronologically, can serve to solve a significant range of health-related problems, opening up the scope

Luiz S. Oliveira and Alceu de Souza Britto contributed equally to this work

✉ Everton F. Baro
efbaro@inf.ufpr.br

Luiz S. Oliveira
luiz.oliveira@ufpr.br

Alceu de Souza Britto
alceu.junior@pucpr.br

1 Department of Informatics, Federal University of Parana, Rua Francisco H. dos Santos, 100, Curitiba 81530-090, Parana, Brazil

2 Department of Informatics, Federal Institute of Parana, Rodovia PR 323, KM 310, Umuarama 87507-014, Parana, Brazil

3 Postgraduate Program in Informatics, Pontifical Catholic University of Parana, Rua Imaculada Conceicao, 1155 Bloco 8, Curitiba 80215-901, Parana, Brazil

Published online: 19 December 2024

for solving health problems through LLMs trained with data of this type.

A practical application for this type of model is hospitalization prediction, which can use chronologically organized sentences as histories for training LLMs. Hospitalizations represent a significant part of the costs of Brazilian health systems. In 2019, if requests for health insurance, payments for procedures and supplies related to hospitalizations are taken into account, the value exceeds US$ 10 billion.[1] Considering these values, even a small reduction in the number of hospitalizations becomes significant. Furthermore, in many cases, hospitalizations can represent a complication in the patient's health condition. Therefore, when possible, avoiding hospitalizations is beneficial, considering monetary aspects, the quality of life and health of patients.

Predicting hospitalization also allows hospital managers and health insurance to plan and optimize processes to reduce costs. In medicine, knowing in advance possible cases of hospitalization allows preventive actions to be taken in order to avoid them. The analysis of these predictions proves to be an important object of study, allowing the discovery of factors that lead to disease complications, some of which may be unknown.

In this study, data from Brazilian health insurance was used to train the models, resulting in descriptions being in Portuguese. This posed certain challenges, as most pre-trained models available in the literature are trained in English. Therefore, considering these aspects of the data, it was necessary to use pre-trained models in Portuguese. Another approach to address this issue was to train the model from scratch using data in Portuguese, without initially relying on a pre-trained model.

Thus, given all these characteristics and possibilities, this work aims to answer the following research questions:

i How can we train LLM-type machine learning models using structured data from health insurance?
ii Does the size of LLM models trained from health insurance data influence the quality of hospitalization predictions?
iii How effective is the use of generalist hospitalization prediction models, based on LLMs, in predicting hospitalizations related to specific problems?
iv Is using trained LLMs for feature extraction as efficient as using a fully connected layer?

Thus, in seeking to answer our research questions, the main objective of this work is to evaluate LLMs for predicting general and stroke hospitalizations, with a particular focus on the latter due to its predictive difficulty. To achieve this, we introduce a preprocessing technique to make structured health insurance data suitable for training LLMs. Additionally, we explore different strategies, ranging from using pre-trained models to training models from scratch.

In addition to the models built for general hospitalizations, as in our previous work [3], tests were conducted on a dataset of stroke hospitalizations to assess the models' efficiency in specific cases. Stroke hospitalizations were chosen due to their supposed unpredictability, in contrast to childbirth hospitalizations, where prior prenatal care events are directly associated with hospitalization. In tests conducted for both general and stroke hospitalizations, we observed AUCs of 0.955 and 0.964, respectively. The best results in both experiments were achieved by combining the three LLM models.

The contribution of this work is threefold, as follows: (i) the proposal of a preprocessing technique for preparing structured health insurance data for training LLMs, offering an alternative for scenarios where related structured events occur over time; (ii) the evaluation of LLMs trained on preprocessed health insurance data for predicting general and stroke-related hospitalizations; and (iii) making prediction models represented by pre-trained LLMs available to the scientific community.

2 Materials and methods

In this work, three different LLMs were used. The first model was trained using the robustly optimized BERT approach (RoBERTa) [4] structure, a variation of bidirectional encoder representations from transformers (BERT) [5] that enhances performance and reduces training time.

The second LLM was trained using a variation of BERT, called BERTimbau [6]. The third LLM, Open-Cabrita3b [7], is a variation of Large Language Model Meta AI (LLaMA) released by Meta AI [8]. Unlike RoBERTa and BERT, Open-Cabrita3b allows training and prediction with sentences of up to 2048 tokens. Due to the large size of LLaMA structures, extensive computational resources are required. Therefore, we used parameter-efficient fine-tuning (PEFT) to reduce the computational demands. PEFT involves adding a subset of parameters to the model while keeping the pre-trained parameters fixed, significantly reducing computational costs and achieving performance comparable to traditional fine-tuning. Specifically, we employed LoRA PEFT [9] to fine-tune the pre-trained Open-Cabrita3b model [7].

In addition to the LLMs, the ensemble method random forest (RF) [10], trained on embeddings extracted using the Sentence Transformers framework [11], was used to predict hospitalizations. Sentence Transformers is a Python framework that can be used to facilitate the extraction of embeddings from pre-trained LLMs. Another approach used in this work for predicting hospitalizations involved adding a

[1] Calculation carried out based on public data from ANS (National Supplementary Health Agency) [2]

fully connected layer (FC) to the LLM. This layer is termed "fully connected" because it establishes complete linkage within the network. FC layers, typically found at the end of a neural network architecture, are responsible for producing output predictions. In this work, a FC layer was connected to the output embedding of the LLM token classification token ([CLS]) [5] to classify sentences as indicating hospitalization or not. The [CLS] token provides an aggregated representation of the entire sequence, containing a high-level, contextualized representation, making it an ideal candidate for sequence classification. Figure 1 exemplifies the FC connection in an LLM.

It is important to highlight that the fine-tuning step is carried out with a fully connected layer coupled to the network, in which weight adjustments, both in the fully connected layer and in the rest of the network, are carried out based on the errors and successes of hospitalization predictions.

In addition to fine-tuning, another important training carried out in the LLMs of this work was the self-supervised training known as masked language model (MLM) method [12], in which a random number of tokens from the input sequence is selected and replaced by the special token *[MASK]* where the goal of the network is then to predict the masked tokens.

Regarding the metrics, we have used the Area under the ROC Curve (AUC) [13], sensitivity (Eq. 1), specificity (Eq. 2), and F1-Score (Eq. 3), where TP is true positive, TN is true negative, FP is false positive and FN is false negative rate.

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{1}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{2}$$

$$\text{F1-Score} = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{3}$$

It is important to note that, due to data privacy considerations, all pre-trained models utilized in this study are open-source, and training takes place locally on the institution's servers.

3 Health insurance data preprocessing

The dataset utilized in this study is an expanded version of the database introduced by Baro [3], comprising 89,339,526 records associated with 445,199 beneficiaries spanning from January 1990 to December 2021.

To generate the historical sentences, we employed the method outlined in [3], albeit with a modification: only the event descriptions were utilized this time. The International Classification of Diseases (ICD) and specialties were excluded from the process, as the data we are currently using contains less frequent occurrences of this information.

3.1 Organization in historical data

As outlined in the introduction, events associated with beneficiaries registered by health insurance can, when linked chronologically, form a type of historical narrative. Given the textual nature of these events, they can serve as inputs for training LLM models. In the context of this study, to construct these narratives from the database records, the following steps were undertaken: (1) merging the data into a single table; (2) removing noise; (3) aggregating the data; and (4) chaining the data.

In the first step, the data was joined into a single table to simplify the narrative creation. This consolidation involved selecting the event description, gender, beneficiary ID, date of event occurrence, care regime, and ICD. The description of each selected characteristic in this consolidation is outlined in Table 1. It is worth noting that the beneficiary ID is a fictitious identifier created during the anonymization process of the dataset.

During the noise removal process, samples exhibiting inconsistencies in either the date of birth or event occurrence were eliminated. The third step involves aggregating the data. Frequently, when a beneficiary interacts with health services, multiple records of occurrences and procedures are generated on the same date. This scenario is evident in numerous examples from the database. Upon analyzing these instances, it becomes apparent that the fields for care regimen and ICD share identical values, as does the gender, for obvious

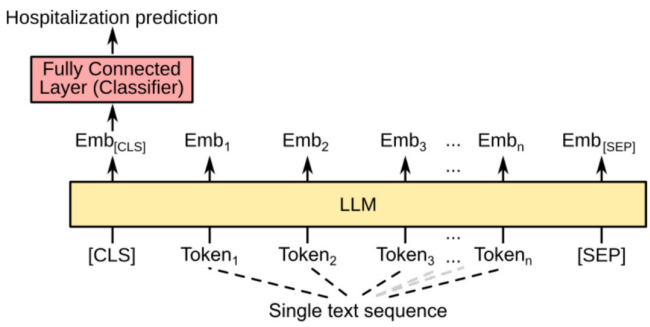


Fig. 1 FC layer example linked to the [CLS] token output embedding of an LLM

Table 1 Features available in the dataset

Feature	Description
Beneficiary ID	Identifier code of the examples of a beneficiary
Gender	Beneficiary's gender identification
Date	Date of occurrence of the procedure
Event Description	Description of the procedure performed
Care Regimen	Whether the care was on an outpatient basis, inpatient, inpatient - day hospital or home care
ICD	Beneficiary ICD-10 code at the time of generating the procedure

reasons. The only variable that differs is the event description, which documents the procedures and inputs related to the beneficiary during a service encounter.

Therefore, in an effort to simplify the process of generating beneficiary event narratives, we restricted each beneficiary to having only one record per date. Achieving this outcome without sacrificing data integrity necessitated aggregating data from event descriptions. Figure 2 provides a concrete example of this aggregation process.

After noise removal and aggregation, the number of records was reduced from 89,339,526 to 20,721,377. After these steps, each example begins to represent the beneficiary's service on a given day in which all descriptions of all events on that day are aggregated into a single example.

Finally, the last step involves chaining the data. Following data aggregation, each beneficiary now possesses a maximum of one sample per day. This data organization allows for the chronological arrangement of each beneficiary's event descriptions, thereby constructing a historical narrative detailing the events they experienced while utilizing the health insurance. Given that one of the objectives of this work is to forecast hospitalizations, these historical sentences were organized by including events up to the day preceding a hospitalization event. Since a beneficiary may have experienced multiple hospitalizations, several examples can be generated for a single beneficiary. In instances where the beneficiary is hospitalized, the example is labeled with the value 1; otherwise, it is labeled with 0.

To illustrate the data-chaining process, Fig. 3 provides examples of historical data with either two hospitalizations or none.

As a result of this pre-processing stage, 880,193 historical sentences were formatted, of which 451,649 were for cases of hospitalization and 428,544 for cases of non-hospitalization. To exemplify the result of organizing the data into historical sentences, the Appendix A presents a concrete example of a generated historical sentence, with the gender of the beneficiary added at the end.

In this study, we examined models capable of processing sentences with a maximum of 512 tokens (RoBERTa and BERTimbau) and 2048 tokens (Open-Cabrita3B). It is crucial that our dataset includes sentences surpassing the maximum token limit of the smaller models (512 tokens). This enables us to assess the relevance of longer sentences for this particular application. In the dataset employed in this study, based on the number of tokens generated by the BERTimbau tokenizer, over 50% of the historical sentences comprise more than 1854 tokens. In LLM training, sentences that exceed the model's maximum limit are truncated from the left. This approach ensures that the most recent events prior to hospitalization are preserved, preventing the generation of identical sentences for different hospitalization cases. Additionally, to

Fig. 2 Example of data aggregation

Beneficiary	Date	Gender	Care Regime	Event Description	ICD
28739	2011-08-22	M	INPATIENT	DISPOSABLE NEEDLE 25 X 07	J32.9
28739	2011-08-22	M	INPATIENT	DISPOSABLE NEEDLE 40 X 12	J32.9
28739	2011-08-22	M	INPATIENT	PERIPHERAL VENOUS CATHETER Nº 18	J32.9
28739	2011-08-22	M	INPATIENT	GLASSES TYPE NASAL CATHETER	J32.9
28739	2011-08-22	M	INPATIENT	3-WAY MULTI-INFUSION DEVICE	J32.9

Beneficiary	Date	Gender	Care Regime	Event Description	ICD
28739	2011-08-22	M	INPATIENT	DISPOSABLE NEEDLE 25 X 07 DISPOSABLE NEEDLE 40 X 12 PERIPHERAL VENOUS CATHETER Nº 18 GLASSES TYPE NASAL CATHETER 3-WAY MULTI-INFUSION DEVICE	J32.9

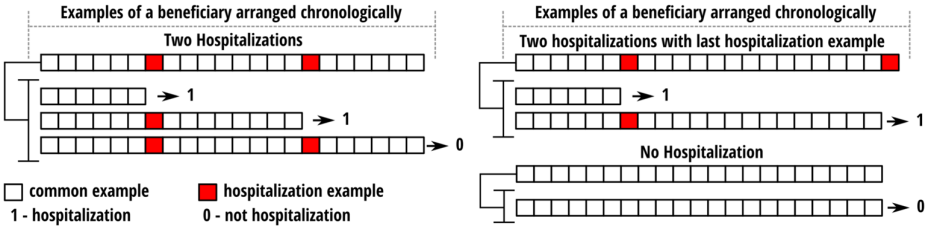


Fig. 3 Example of generating historical sentences per beneficiary for cases of two and no hospitalizations

gain insights into the dataset's profile, Fig. 4a illustrates the distribution histograms of service data by age, portraying the beneficiaries' age profile, and Fig. 4b the temporal distribution of data. It is notable that the volume of data before 2014 is negligible. Moreover, the dataset has 8399 distinct ICDs, with 6913 associated with hospitalization events, including the stroke ICD among them.

3.2 Splitting data for training and testing

The 880,193 historical sentences were divided into three parts. The largest portion was used for training and fine-tuning LLMs. The remaining portions were used for testing and for training a RF model with features extracted from the LLMs. Figure 5 illustrates this data split.

4 Training LLMs in health insurance data

To address research questions on training LLMs with health insurance data, specifically examining the impact of model size on hospitalization prediction quality and the capability of these models to predict hospitalizations for specific conditions, four models were trained. The first model was based on the RoBERTa structure; the second and third were derived from the pre-trained BERTimbau model; and the fourth was derived from the pre-trained OpenCabrita3B model. Figure 6 summarizes the training sequence.

4.1 RoBERTa

For the experiment involving training from scratch, we used RoBERTa due to its optimized structure and modest hardware requirements. In this case, all model weights were adjusted based on historical health insurance data that constitute the corpus of this work. We generated our own tokenizer for this experiment, trained on aggregated data from event descriptions, ensuring it is strongly tailored to this domain.

The model was pre-trained using the self-supervised MLM method on the historical data of beneficiaries. We used 837,159 examples (dataset I) over two epochs, resulting in a model we call RoBERTa-MLM. Fine-tuning was then performed over two epochs for the NLP task sequence for classification, using 10% of the historical data (10% of dataset I) labeled for predicting hospitalization one day in advance. This final model is referred to as RoBERTa-MLM-FT. Both training sessions used sequences with a maximum of 512 tokens. Figure 7a summarizes the training sequence used to develop the model.

4.2 BERTimbau

BERTimbau is a BERT model trained in Portuguese. According to Souza et al. (2023) [14], it has achieved state-of-the-art performance in several NLP tasks, surpassing multilingual models for Portuguese. Given that our work uses Portuguese data, BERTimbau was chosen for training to compare the effectiveness of using domain-specific data.

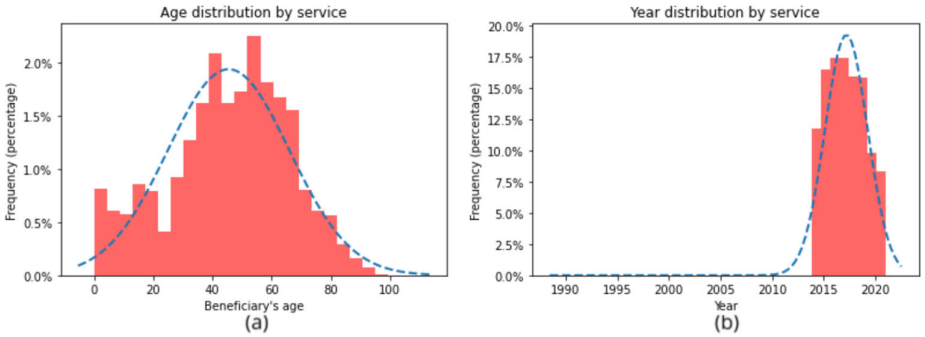


Fig. 4 Histogram with data distribution by age of beneficiaries

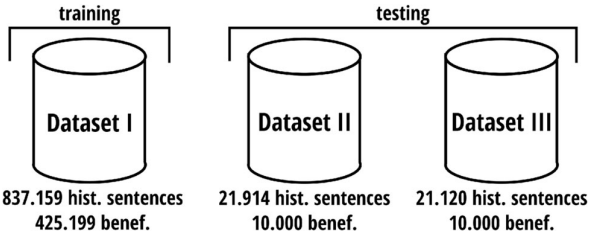


Fig. 5 Splitting data for training and testing

Three training sessions were conducted. The first was self-supervised MLM training using the BERTimbau pre-trained model, resulting in what we call BERTimbau-MLM. The second and third sessions involved fine-tuning for the NLP classification task: one on the original BERTimbau model, resulting in BERTimbau-FT, and the other on BERTimbau-MLM, resulting in BERTimbau-MLM-FT.

The MLM training was performed over one epoch using the same 837,159 examples (dataset I) as in the RoBERTa training. The fine-tuning sessions were also conducted over one epoch, utilizing 10% of dataset I, which consists of historical data labeled for predicting hospitalization one day in advance. All training was performed with sequences of up to 512 tokens. Figure 7c summarizes the training sequences used to develop the models.

4.3 Open-Cabrita3B

Considering that 50% of the historical sentences in the dataset exceed the 512-token limit of both BERT and RoBERTa, a significant portion of the sentences are discarded during hospitalization inference. This limitation suggests the need for models that can handle longer sentences to determine if older data are important for inferring hospitalizations.

An alternative for longer sentences would be to use larger models like the LLaMA model, which in its first version supports sequences of up to 2048 tokens but is primarily trained on English data. Therefore, as a Portuguese alternative, we adopted Open-Cabrita3B [7] in this work. Open-Cabrita3B, derived from OpenLLaMA [15], supports sequences of up to 2048 tokens and has 3 billion parameters.

Due to the high computational cost of pre-training the model, only fine-tuning was performed for the NLP classification task. This was done using the LoRA PEFT technique with sequences of up to 2048 tokens and 10% of the labeled historical data (10% of dataset I). The resulting model is called Open-Cabrita3B-FT. The fine-tuning was completed in one epoch. Figure 7b presents a summary of the training sequence used to obtain this model.

5 Experiments

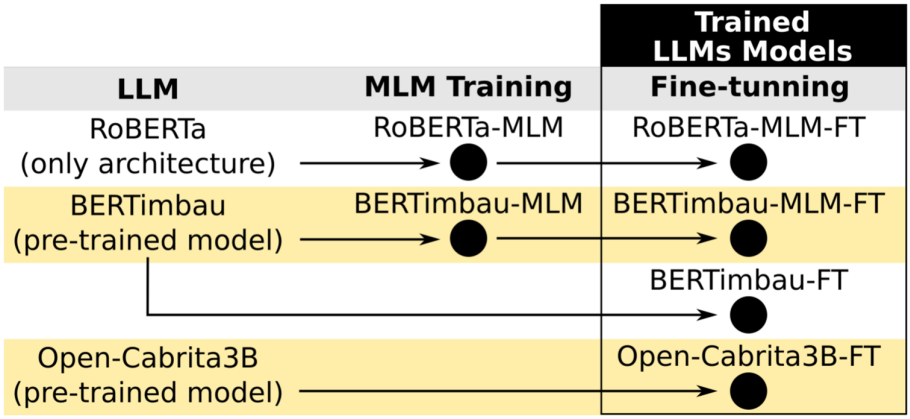
In this section, we present the experiments and results related to hospitalization predictions in general and for stroke tested from trained models presented in Section 4.

5.1 General hospitalization prediction

To assess the quality and identify the best method for predicting hospitalizations, experiments were conducted with the models described in Section 4. Models based on RoBERTa and BERTimbau were evaluated with embeddings extracted using the Sentence Transformers framework and applied to RF. Additionally, direct inference was performed by incorporating a fully connected layer for sequence classification. However, the Open-Cabrita3B model was solely assessed using the fully connected layer due to the challenges in extracting embeddings through the Sentence Transformers framework caused by its PEFT training method. Figure 8 summarizes the sequence of steps performed.

Additionally, two types of classifier combinations were conducted as illustrated in Fig. 9. The first approach (Fig. 9a) involves combining the RF classifier with one of the models utilizing the fully connected layer. The second approach

Fig. 6 LLM training sequences



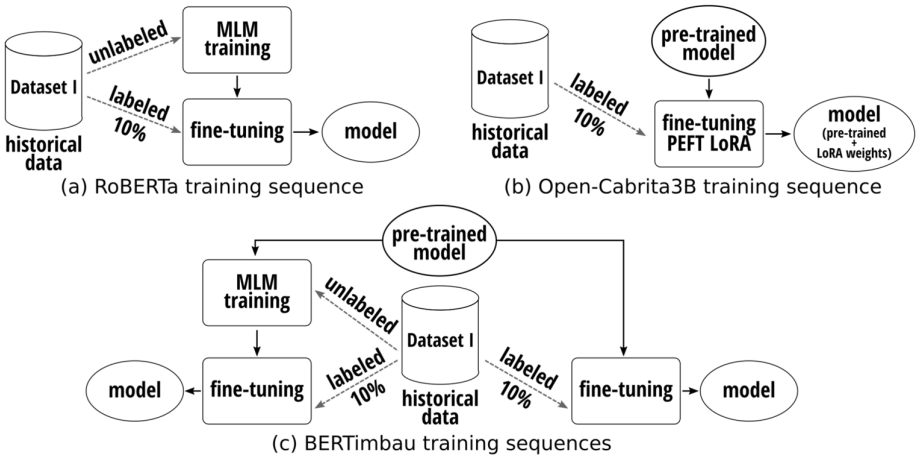


Fig. 7 LLMs training sequences

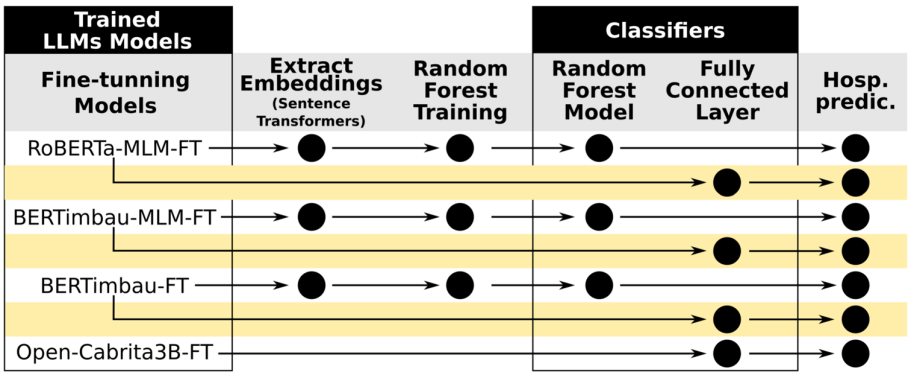


Fig. 8 Sequence of steps for experiments

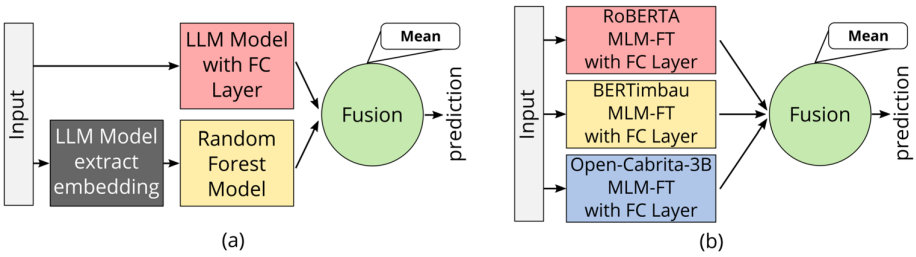


Fig. 9 Combination methods used in experiments

Table 2 Results of the classifiers trained with RoBERTa structure from scratch

Model	Metric	FC	RF	CB
RoBERTa-MLM-FT	F1-Score	86.4±0.5	86.5±0.6	**87.1±0.4**
	Sensitivity	84.5±0.7	**89.8±0.6**	87.7±0.4
	Specificity	**89.0±0.8**	82.9±1.3	86.7±0.9
	AUC	94.8±0.4	94.2±0.3	**95.0±0.3**

Bold shows the best results

(Fig. 9b) combines all three classifiers using the fully connected layer. Both approaches employ the average of the inference probabilities as the fusion method.

The data used for pre-training and fine-tuning, along with the training methodologies, were outlined in Sections 2, 3, and 4. Additionally, a distinct subset of examples was selected, not employed in either LLM training or fine-tuning. This subset, denoted as *dataset I* in Fig. 5, was partitioned into 15% for testing, totaling 3287 examples. For cases where LLMs were utilized solely as feature extractors, the remaining 85%, equivalent to 18,626 examples, were employed for this extraction and subsequent RF training.

The implementations relied on PyTorch [16], Transformers [17], and Scikit-learn [18] as the primary libraries, utilized for both model training and testing. For classification, RF and the FC layer for sequence classification tasks were employed from the Transformers library implementation. In RF, default hyperparameters of Scikit-learn were used, with the exception of the number of estimators, which was set to 200.

5.1.1 RoBERTa results

Table 2 presents the results of the experiments for the RoBERTa-MLM-FT model using RF and FC as prediction methods in addition to the combination (CB) of both as shown in Fig. 9a.

While both F1-Score and AUC yield similar results, an interesting observation emerges regarding the inversion of values between sensitivity and specificity for RF and FC. RF tends to achieve higher values for sensitivity, whereas FC excels in specificity. Moreover, the CB of both methods balances these metrics, enhancing both the F1-Score and AUC.

5.1.2 BERTimbau results

Table 3 presents the results of the experiments for the BERTimbau-FT and BERTimbau-MLM-FT models using RF and FC as prediction methods in addition to the CB of both as shown in Fig. 9a.

Table 3 Results of the classifiers trained with BERTimbau

Model	Metric	FC	RF	CB
BERTimbau-FT	F1-Score	85.2±0.6	85.6±0.6	86.0±0.6
	Sensitivity	81.8±0.9	87.9±0.7	85.0±0.8
	Specificity	89.6±1.0	83.2±0.9	87.6±0.9
	AUC	94.1±0.4	93.9±0.5	94.3±0.4
BERTimbau-MLM-FT	F1-Score	85.7±0.5	86.2±0.9	**86.4±0.6**
	Sensitivity	82.4±0.7	**89.1±0.8**	86.2±0.8
	Specificity	**90.0±0.8**	83.1±1.2	86.9±0.7
	AUC	94.4±0.4	94.2±0.4	**94.7±0.4**

Bold shows the best results

Table 4 Results of the classifier trained with OpenCabrita3B

F1-Score	Sensitivity	Specificity	AUC
87.8±0.7	92.4±0.4	82.6±1.3	95.4±0.3

These experiments also reveal an almost inversion of sensitivity and specificity values. Moreover, the BERTimbau-MLM-FT model achieved slightly superior results, indicating that MLM training contributed to improvements in the final outcome.

Considering the findings from the last two Sections 5.1.1 and 5.1.2, research question iv is addressed. Utilizing LLMs as feature extractors from historical sentences for subsequent RF training yielded marginally better results in predicting hospitalization cases, as evidenced by the sensitivity metric. It is important to note that sensitivity is closely linked to positive cases for hospitalizations.

5.1.3 Open-Cabrita3B results

Table 4 presents the results of the experiments for the OpenCabrita3B-FT model using FC as a prediction method.

Open-Cabrita3B-FT attained comparable F1-Score and AUC results compared to the other models examined in this study. However, it demonstrated the highest sensitivity among all the models tested. This observation suggests that the inclusion of longer sentences can have a positive impact on predicting hospitalizations.

5.1.4 Combination results

Table 5 presents the performance obtained from combining the three models: RoBERTa-MLM-FT, BERTimbau-MLM-FT, and OpenCabrita3B-FT with the fully connected layer, as illustrated in Fig. 9b.

The results in Table 5 indicate that combining the three models produces a more balanced model, as evidenced by the close proximity of sensitivity and specificity values.

Table 5 Results from combining the classifiers

F1-Score	Sensitivity	Specificity	AUC
87.5±0.8	87.8±0.9	87.5±0.9	95.5±0.3

Additionally, this combination enables us to achieve the best overall performance among the generalist models.

5.2 Stroke hospitalization prediction

To assess how well the models perform for a specific problem, we examined datasets II and III, which consist of data not utilized in model training, to identify instances of hospitalizations related to stroke (ICD I64). This type of hospitalization was chosen because it plays a critical role in delivering timely and effective care to individuals who have experienced a stroke, with the aim of minimizing brain damage, maximizing recovery, and preventing future strokes.

From this data, we selected 134 beneficiaries who were hospitalized due to stroke and had records dating back at least 360 days before hospitalization. Using event descriptions, we constructed historical sentences for these beneficiaries, generating 200 examples of stroke-related hospitalizations occurring one day in advance. The difference between the number of examples and beneficiaries suggests instances of beneficiaries being readmitted for the same problem. Additionally, from the same test data, we randomly selected 200 examples of non-hospitalizations to compose the test dataset, resulting in a total of 400 test examples. These tests were conducted on previously trained LLMs without specific fine-tuning for stroke. It is worth noting that the number of tokens generated by the BERTimbau tokenizer exceeds 3214 tokens for more than 50% of the examples in this dataset. Consequently, a significant number of examples exceed the maximum token limit of the largest model tested, Open-Cabrita3B.

This data was then applied to the three models trained with the fully connected layer. These three classifiers were also combined as shown in Fig. 9b. Table 6 presents the results achieved.

Table 6 Results of test with Stroke hospitalizations

Metric	a	b	c	d
F1-Score	84.8	88.4	**88.7**	**88.7**
Sensitivity	**93.5**	89.0	88.5	90.5
Specificity	76.5	88.0	**89.0**	87.0
AUC	96.4	96.4	95.6	**96.5**

Bold shows the best results
(a) OpenCabrita3B-FT, (b) RoBERTa-MLM-FT, (c) BERTimbau-MLM-FT, (d) Combination

The same pattern observed between specificity and sensitivity in Section 5.1 is reiterated for stroke-related hospitalizations. Specifically, Open-Cabrita3B-FT exhibits higher sensitivity and lower specificity compared to the other models. The combination of classifiers yields an F1-Score of 88.7%.

Anticipating stroke-related hospitalizations, even if only a few cases, can be beneficial for various applications, including implementing preventive measures by adjusting treatments or providing increased patient care. To explore these potential applications, tests were conducted for stroke-related hospitalizations at different advance periods: 5, 15, 30, and 60 days. We aimed to assess how the model performs across these advanced prediction periods. Figure 10 displays the results for the three models examined in this study, as well as the combination of classifiers.

Considering the findings depicted in Fig. 10, the superiority of Open-Cabrita3B-FT over other models becomes evident. While it initially achieves a similar AUC to the others for predictions a few days in advance, it notably outperforms them as the prediction period extends. This model excels particularly in identifying hospitalization cases, as reflected in its sensitivity values. Given that Open-Cabrita3B-FT is a larger model than the others, capable of handling 2048 tokens, the results presented in this section address research question ii, as the model size positively influences the outcomes.

Furthermore, to address research question iii, fine-tuning of Open-Cabrita3B was exclusively conducted with data pertaining to stroke-related hospitalizations. The training dataset was extracted from dataset I and comprised 7808 examples, consisting of 3904 cases of stroke-related hospitalizations and 3904 cases of non-hospitalizations. The training configurations mirrored those detailed in subsection 4.3, except for the training data. For testing, data on stroke-related hospitalizations occurring 60 days in advance were employed. Table 7 presents the results obtained from this test.

The results presented in Table 7 show the efficacy of the generalist models trained in this study in predicting hospitalizations for specific issues, thereby addressing research question iii. Additionally, it is noteworthy that research question i is also answered by considering the results of all experiments in this study. The approach employed to prepare structured health insurance data has proven to be effective for training LLMs.

6 Discussion

While Open-Cabrita3B outperforms other models, its superiority does not render the results of the other models insignificant. Given the higher hardware demands for training and inference with Open-Cabrita3B, the other models still yield noteworthy results while requiring less computational

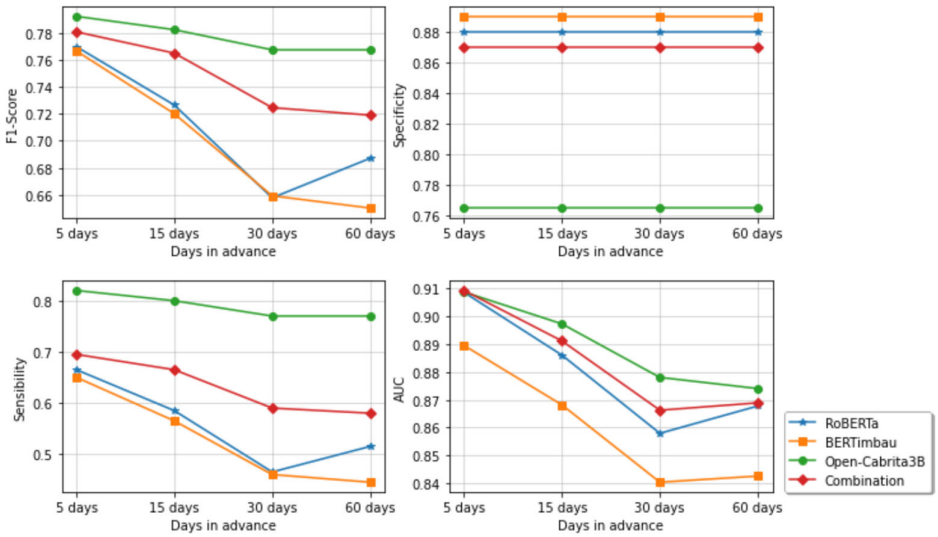


Fig. 10 Predictions of stroke hospitalizations for different periods in advance

power. Moreover, RoBERTa-MLM is tailored specifically to the health insurance data domain, featuring its own tokenizer for this vocabulary. These models, including RoBERTa-MLM and others, hold potential effectiveness in various applications with similar data structures, such as cost predictions, days in hospital, disease and health complication forecasting, analysis of factors related to health complications, treatment evaluations, and more.

Therefore, recognizing the potential utility of these models, they have been made accessible to the scientific community. Additionally, several other models were trained in English, employing translations of the health insurance data utilized in this study. The data translation was facilitated by the *translation-pt-en-t5* model detailed in the work by Lopes et al. (2020) [19]. These models in English and Portuguese, along with the test data and hyperparameters used in training, are accessible through the link https://huggingface.co/efbaro, thus allowing the reproducibility of results. Furthermore, it is important to highlight that data privacy and security are guaranteed through anonymization and strict access controls, so only test data is made available.

Table 7 Results of the test with hospitalizations for stroke 60 days in advance, of the Open-Cabrita3B model trained only with hospitalizations and with all data

Metric	OpenCabrita3B only Stroke data	OpenCabrita3B all data
F1-Score	70.7	**76.7**
Sensitivity	56.0	**77.0**
Specificity	**87.0**	76.5
AUC	81.2	**87.4**

Bold shows the best results

6.1 Limitations

The models trained from scratch, RoBERTa-MLM and RoBERTa-MLM-EN (model trained with English data), exclusively with data from health insurance beneficiary historical sentences, were not evaluated using traditional data sentences. Since they were not exposed to examples of traditional text sentences during training, it is not anticipated that these models would perform well on this type of data.

7 Conclusion

Given the intended objectives, this study determined that LLMs are effective for predicting hospitalizations using chronologically organized health insurance data. Tests conducted with models based on BERT, RoBERTa, and LLaMA yielded significant results, with the combination of models achieving F1-Score percentages of 87.5

For predictions of hospitalizations due to stroke, Open-Cabrita3B-FT demonstrated superior performance, accurately predicting hospitalizations with a significant lead time and high probability for a considerable portion of cases. This highlights the effectiveness of the methodology employed in this study for predicting hospitalizations related to specific health issues. Furthermore, the results attained underscore the potential of utilizing these models to predict other types of hospitalizations, thereby opening up a wide array of research possibilities.

It is important to note that among the pre-trained models investigated in this research, analyses of the results focused on applications of the RoBERTa, BERTimbau, and

Open-Cabrita3B models. Both these analyzed models and others trained in English are accessible via a link provided in this paper, facilitating their utilization in future research endeavors.

Appendix A: Example of a generated historical sentence

'vitamina d 25 hidroxi pesquisa e/ou dosagem vitamina d3 ácido úrico - pesquisa e/ou dosagem colesterol hdl - pesquisa e/ou dosagem colesterol ldl - pesquisa e/ou dosagem colesterol total - pesquisa e/ou dosagem gama-glutamil transferase - pesquisa e/ou dosagem glicose - pesquisa e/ou dosagem hemoglobina glicada a1 total - pesquisa e/ou dosagem triglicerídeos - pesquisa e/ou dosagem colesterol vldl - pesquisa e/ou dosagem creatinina - pesquisa e/ou dosagem uréia - pesquisa e/ou dosagem hemograma com contagem de plaquetas ou frações eritrograma leucograma plaquetas tireoestimulante hormônio tsh - pesquisa e/ou dosagem tiroxina t4 - pesquisa e/ou dosagem triiodotironina t3 - pesquisa e/ou dosagem taxas de alugueis de equipamentos sódio - pesquisa e/ou dosagem transaminase oxalacética amino transferase aspartato - pesquisa e/ou dosagem transaminase pirúvica amino transferase de alanina - pesquisa e/ou dosagem creatinina - pesquisa e/ou dosagem potássio - pesquisa e/ou dosagem uréia - pesquisa e/ou dosagem hemograma com contagem de plaquetas ou frações eritrograma leucograma plaquetas hemograma com contagem de plaquetas ou frações eritrograma leucograma plaquetas medicamentos em geral dengue - igg e igm cada - pesquisa e/ou dosagem dengue - igg e igm cada - pesquisa e/ou dosagem proteína c reativa quantitativa - pesquisa e/ou dosagem consulta em consultório no horário normal ou preestabelecido consulta em consultório no horário normal ou preestabelecido consulta em consultório no horário normal ou preestabelecido consulta em consultório no horário normal ou preestabelecido consulta em consultório no horário normal ou preestabelecido mapeamento de retina oftalmoscopia indireta - monocular consulta em consultório no horário normal ou preestabelecido tc - face ou seios da face us - mamas us - transvaginal útero ovário anexos e vagina us - abdome superior fígado vias biliares vesícula pâncreas e baço paquimetria ultrassônica - monocular campimetria computadorizada - monocular curva tensional diária - binocular feminino'

Acknowledgements This work has been supported partially by the Coordination for the Improvement of Higher Education Personnel (CAPES) - Program of Academic Excellence (PROEX).

Declarations

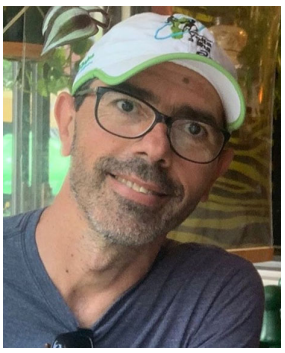
Conflict of interest The authors declare no competing interests.

References

1. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A.N, Kaiser Ł, Polosukhin I (2017) Attention is all you need. Advances Neural Inf Process Syst 30. https://doi.org/10.48550/arXiv.1706.03762
2. Portal Brasileiro de Dados Abertos - Procedimentos Hospitalares por UF (2021). https://dados.gov.br/dataset/procedimentos-hospitalares-por-uf. Accessed 26 Nov 2021
3. Baro EF, Oliveira LS, Souza Britto Junior A (2022) Predicting hospitalization from health insurance data. In: 2022 IEEE International conference on systems, man, and cybernetics (SMC), pp 2790–2795. https://doi.org/10.1109/SMC53654.2022.9945601
4. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019) RoBERTa: a robustly optimized BERT pretraining approach. https://doi.org/10.48550/arXiv.1907.11692
5. Devlin J, Chang M-W, Lee K, Toutanova K (2018) BERT: pre-training of deep bidirectional transformers for language understanding. https://doi.org/10.48550/ARXIV.1810.04805 arXiv:1810.04805 [cs.CL]
6. Souza F, Nogueira R, Lotufo R (2020) BERTimbau: pretrained BERT models for Brazilian Portuguese. In: Cerri R, Prati RC (eds) Intelligent systems, pp 403–417. Springer, Cham. https://doi.org/10.1007/978-3-030-61377-8_28
7. Larcher C, Piau M, Finardi P, Gengo P, Esposito P, Caridá V (2023) Cabrita: closing the gap for foreign languages.https://doi.org/10.48550/ARXIV.2308.11878 arXiv:2308.11878 [cs.CL]
8. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M.-A, Lacroix T, Rozière B, Goyal N, Hambro E, Azhar F, Rodriguez A, Joulin A, Grave E, Lample G (2023) Llama: open and efficient foundation language models. https://doi.org/10.48550/arXiv.2302.13971 arXiv:2302.13971 [cs.CL]
9. Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, Wang L, Chen W (2021) Lora: low-rank adaptation of large language models. https://doi.org/10.48550/ARXIV.2106.09685 arXiv:2106.09685 [cs.CL]
10. Breiman L (2001) Random forests. Mach Learn 45(1):5–32. https://doi.org/10.1023/A:1010933404324
11. Reimers N, Gurevych I (2019) Sentence-BERT: sentence embeddings using Siamese BERT-networks. arXiv. https://doi.org/10.48550/ARXIV.1908.10084
12. Salazar J, Liang D, Nguyen TQ, Kirchhoff K (2019) Masked language model scoring. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (2020), 2699–2712 https://doi.org/10.18653/v1/2020.acl-main.240 arXiv:1910.14659 [cs.CL]
13. Fawcett T (2006) An introduction to ROC analysis 27(8):861–874. https://doi.org/10.1016/j.patrec.2005.10.010
14. Souza FC, Nogueira RF, Lotufo RA (2023) BERT models for Brazilian Portuguese: pretraining, evaluation and tokenization analysis. Appl Soft Comput 149:110901. https://doi.org/10.1016/j.asoc.2023.110901
15. Geng X, Liu H (2023) OpenLLaMA: an open reproduction of LLaMA. https://github.com/openlm-research/open_llama
16. Paszke A, Gross S, Chintala S, Chanan G, Yang E, DeVito Z, Lin Z, Desmaison A, Antiga L, Lerer A (2017) Automatic differentiation in PyTorch. In: NIPS-W
17. Wolf T, Debut L, Sanh V, Chaumond J et al (2020) Transformers: state-of-the-art natural language processing. In: Liu Q, Schlangen D (eds) Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations, pp 38–45. Association for Computational Linguistics, Online. https://doi.org/10.18653/v1/2020.emnlp-demos.6

18. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V et al (2011) Scikit-learn: machine learning in python. J Mach Learn Res 12:2825–2830. https://doi.org/10.48550/arXiv.1201.0490
19. Lopes A, Nogueira R, Lotufo R, Pedrini H (2020) Lite training strategies for Portuguese-English and English-Portuguese translation. https://doi.org/10.48550/ARXIV.2008.08769 arXiv:2008.08769 [cs.CL]

Everton F. Baro holds a bachelor's and master's degree in Computer Science from the State University of Maringá, is currently a PhD student in Computer Science at the Federal University of Paraná and a professor at the Federal Institute of Paraná. His scientific interests are Machine Learning, Natural Language Processing, Deep Learning, LLMs and Health Insurance.



Luiz S. Oliveira received his Ph.D. degree in Engineering from the École de Technologie Supérieure, Université du Quebec, Canada in 2003. From 2004 to 2009 he was professor of the Computer Science Department at Pontifical Catholic University of Paraná, Curitiba, PR, Brazil. In 2009, he joined the Federal University of Paraná, Curitiba, PR, Brazil, where he is associate professor of the Department of Informatics. His research focuses primarily on pattern recognition and machine learning.

Alceu de Souza Britto Master's degree in industrial informatics from CEFET-PR (Curitiba, 1996), PhD in Informatics from the Pontifical Catholic University of Paraná (PUCPR, 2001) with an internship at École de Technologie Supérieure (ÉTS, Canada). Post-doctorate in Machine Learning at ÉTS (Canada, 2013). Full professor and researcher in the Postgraduate Program in Informatics (PPGIa, PUCPR) since 2001, where he supervises master's, doctoral, and post-doctoral students. Productivity research grantee from CNPQ level 1D. Also acts as an associate professor at UEPG since 1989. He coordinates international research projects in the area of Machine Learning in partnership with researchers from École de Technologie Supérieure (ÉTS, Canada), the University of Rouen (UR, France), and Ingolstadt Technische Hochschule (THI, Germany). He is an ad-hoc consultant for INEP for evaluating higher education institutions and served as coordinator of the Exact and Earth Sciences area of Fundação Araucária (PR) from 2017-2019. Currently, he is the CISIA (Integrated Center for Artificial Intelligence Solutions) coordinator at PUCPR. Dr. Britto has published over 170 scientific articles in international journals and conferences. Areas of interest include Artificial Intelligence, Machine Learning, and Pattern Recognition.