Original Paper

Stroke Outcome Measurements From Electronic Medical Records: Cross-sectional Study on the Effectiveness of Neural and Nonneural Classifiers

Bruna Stella Zanotto[1,2*], MSc, PharmD; Ana Paula Beck da Silva Etges[1,3*], Eng, MSc, PhD; Avner dal Bosco[3*], MSc; Eduardo Gabriel Cortes[4*], MSc; Renata Ruschel[1*], PT; Ana Claudia De Souza[5*], MD, PhD; Claudio M V Andrade[6*], MSc; Felipe Viegas[6*], MSc; Sergio Canuto[6*], MSc, PhD; Washington Luiz[6*], MSc; Sheila Ouriques Martins[5*], MSc, MD, PhD; Renata Vieira[7*], MSc, PhD; Carisi Polanczyk[1,2*], MSc, MD, PhD; Marcos André Gonçalves[6*], MSc, PhD

[1]National Institute of Health Technology Assessment - INCT/IATS (CNPQ 465518/2014-1), Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil

[2]Graduate Program in Epidemiology, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil

[3]School of Technology, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, Brazil

[4]Graduate Program of Computer Science, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil

[5]Brazilian Stroke Network, Hospital Moinhos de Vento, Porto Alegre, Brazil

[6]Computer Science Department, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

[7]Centro Interdisciplinar de História, Culturas e Sociedades (CIDEHUS), Universidade de Évora, Évora, Portugal

[*]all authors contributed equally

Corresponding Author:
Marcos André Gonçalves, MSc, PhD
Computer Science Department
Universidade Federal de Minas Gerais
Avenue Antônio Carlos, 6627
Belo Horizonte, 31270-901
Brazil
Phone: 55 3134095860
Email: mgoncalv@dcc.ufmg.br

Abstract

Background: With the rapid adoption of electronic medical records (EMRs), there is an ever-increasing opportunity to collect data and extract knowledge from EMRs to support patient-centered stroke management.

Objective: This study aims to compare the effectiveness of state-of-the-art automatic text classification methods in classifying data to support the prediction of clinical patient outcomes and the extraction of patient characteristics from EMRs.

Methods: Our study addressed the computational problems of information extraction and automatic text classification. We identified essential tasks to be considered in an ischemic stroke value-based program. The 30 selected tasks were classified (manually labeled by specialists) according to the following value agenda: tier 1 (achieved health care status), tier 2 (recovery process), care related (clinical management and risk scores), and baseline characteristics. The analyzed data set was retrospectively extracted from the EMRs of patients with stroke from a private Brazilian hospital between 2018 and 2019. A total of 44,206 sentences from free-text medical records in Portuguese were used to train and develop 10 supervised computational machine learning methods, including state-of-the-art neural and nonneural methods, along with ontological rules. As an experimental protocol, we used a 5-fold cross-validation procedure repeated 6 times, along with *subject-wise sampling*. A heatmap was used to display comparative result analyses according to the best algorithmic effectiveness (F1 score), supported by statistical significance tests. A feature importance analysis was conducted to provide insights into the results.

Results: The top-performing models were support vector machines trained with lexical and semantic textual features, showing the importance of dealing with noise in EMR textual representations. The support vector machine models produced statistically superior results in 71% (17/24) of tasks, with an F1 score >80% regarding care-related tasks (patient treatment location, fall risk, thrombolytic therapy, and pressure ulcer risk), the process of recovery (ability to feed orally or ambulate and communicate), health care status achieved (mortality), and baseline characteristics (diabetes, obesity, dyslipidemia, and smoking status). Neural

XSL•FO
RenderX

methods were largely outperformed by more traditional nonneural methods, given the characteristics of the data set. Ontological rules were also effective in tasks such as baseline characteristics (alcoholism, atrial fibrillation, and coronary artery disease) and the Rankin scale. The complementarity in effectiveness among models suggests that a combination of models could enhance the results and cover more tasks in the future.

Conclusions: Advances in information technology capacity are essential for scalability and agility in measuring health status outcomes. This study allowed us to measure effectiveness and identify opportunities for automating the classification of outcomes of specific tasks related to clinical conditions of stroke victims, and thus ultimately assess the possibility of proactively using these machine learning techniques in real-world situations.

(JMIR Med Inform 2021;9(11):e29120) doi: 10.2196/29120

KEYWORDS

natural language processing; stroke; outcomes; electronic medical records; EHR; electronic health records; text processing; data mining; text classification; patient outcomes

Introduction

Background

Stroke is the second leading cause of mortality and disability-adjusted life years globally [1,2]. The outcomes of stroke can vary greatly, and timely assessment is essential for optimal management. As such, there has been an increasing interest in the use of automated machine learning (ML) techniques to track stroke outcomes, with the hope that such methods could make use of large, routinely collected data sets and deliver accurate, personalized prognoses [3]. However, studies applying ML methods to stroke, although published regularly, have focused mostly on stroke imaging applications [4-6] and structured data retrieval [3]. Few studies have addressed the unstructured textual portion of electronic medical records (EMRs) as the primary source of information.

Indeed, the use of EMR data in the last decade has led to promising findings in population health research, such as patient-use stratification [7], treatment-effectiveness evaluation [8], early detection of diseases [9], and predictive modeling [10]. However, dealing with EMR data is often labor intensive [11] and challenging because of the lack of standardization in data entry, changes in coding procedures over time, and the impact of missing information [9,12-14]. The information technology (IT) gap between automated data collection from EMRs and improving the quality of care has been described in the literature as a decelerator of value initiatives [15-18].

With recent advances in IT, several groups have attempted to apply natural language processing (NLP) to the text analysis of EMRs to achieve early diagnosis of multiple conditions, such as peripheral arterial disease [19], asthma [20], multiple sclerosis [21], and heart failure [22]. In these studies, NLP was used to find specific words or phrases in a predefined dictionary that described the symptoms or signs of each disease [14,21,23].

Objectives

Generating value for the patient as the central guide requires advances in strategies to automate the capturing of data that will allow managers to assess the quality of service delivery to patients [24,25]. Accordingly, our research aims to compare the effectiveness of state-of-the-art automatic text classification methods in classifying data to support the prediction of clinical patient outcomes and the extraction of patient characteristics from EMR sentences. With stroke as our case study application, our specific goal is to investigate the capability of these methods to automatically identify, with reasonable effectiveness, the outcomes and clinical characteristics of patients from EMRs that may be considered in a stroke outcome measurement program.

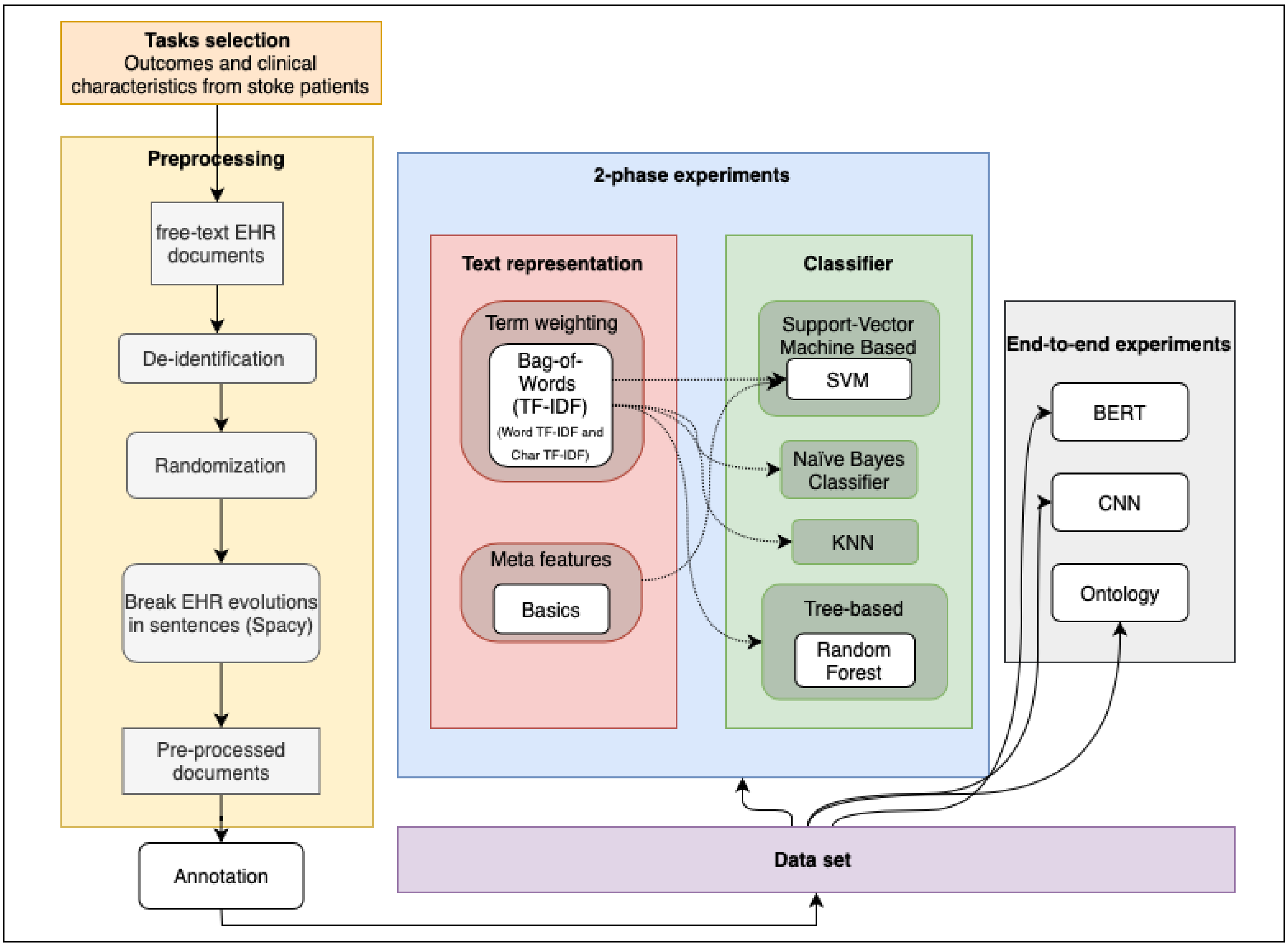
Methods

Overview

This study faced a computational problem related to information extraction and free-text classification. As presented in Figure 1, the dotted lines represent the union of the text representative technique that was used with each classifier in the two-phase experiments. Our study was generally organized into four stages: (1) task selection; (2) study design, preprocessing, and data annotation; (3) definition of automatic text classification methods; and (4) experimental evaluation (experimental protocol, setup, and analysis of results).

XSL•FO
RenderX

Figure 1. Study architecture. BERT: bidirectional encoder representation from transformers; CNN: convolutional neural network; EHR: electronic health record; KNN: K-nearest neighbor; SVM: support vector machine; TF-IDF: term frequency-inverted document frequency.



Task Selection

A literature review and multidisciplinary expert interviews (n=8) were used to define specific outcome dimensions and measures that may be considered in an outcome measurement program for ischemic stroke. The outcome identification step was based on adhering to value agenda element dimensions to cover the tiers of the outcome hierarchy [26], such as functionality dimensions, the recovery process, and outcomes that matter to patients. These dimensions included risk events, achieved health care status, and stroke outcome scales, such as the National Institutes of Health Stroke Scale (NIHSS) and the modified Rankin scale (mRS) [27,28].

Study Design and Data Annotation

We retrospectively built a database of medical records from a digital hospital system. The database covered 2 years of patients hospitalized for ischemic stroke. The hospital is a private institution of excellence in southern Brazil. The EMR system used was the MV Soul (Recife). Since 2017, the hospital has introduced the ICHOM standard sets' data collection routine for different clinical pathways and created an office for institutional values. To examine the stroke pathway, data were collected on October 15, 2015. In 2019, the hospital incorporated the Angel Awards Program [29], which was certified as a platinum category at the end of the first year. This study was approved by the hospital ethics committee (CAAE: 29694720000005330).

Medical records of patients were submitted to preprocessing using the spaCy Python library (Python Software Foundation; Python Language Reference, version 2.7) [30] to stratify texts into sentences. A total of 44,206 EMR sentences were obtained from 188 patients. The approach followed a hypothesis for managing unbalanced data, such as electronic health records, which assumes that relevant information to be retrieved from EMRs encompasses a small space of words delimited as sentences, and the residual is noise [31-33]. During the text stratification process, spaCy [30] uses rule-based algorithms that set the sentence limits according to the patterns of characters, thereby delimiting its beginning and end. The names of patients and medical staff were identified, thus removing all confidential information from the data set. The preprocessed textual sentence was represented in a vector of words that disregarded grammar and word order but maintained their multiplicity.

For sentence annotation (intratask class labeling), we developed annotation guidelines that provided an explicit definition of each task, its classes (response options), and examples to be identified in the documents. This guideline is written in Portuguese and is available upon request.

Two annotators independently reviewed the preprocessed text documents (44,206 sentences) and had the percent agreement

XSL•FO
RenderX

between them measured by κ, which was higher than 0.61 (substantial agreement) [34]. Task-level disagreements were resolved by consensus determination by 2 annotators, with assistance from a committee composed of experts (APE, ACS, MP, KBR, and CAP).

Each task could have two or more output answers, depending on the meaning of the sentence. Examples of an EMR and the annotation process can be seen in Multimedia Appendices 1 and 2. Task details in terms of class and sentence distribution are shown in Multimedia Appendix 3 and demonstrate the highly imbalanced nature of the tasks with most of the sentences belonging to the NI (noninformative) class. This makes it a very hard endeavor from an ML perspective. Subsequently, we evaluated the impact of this imbalance in the experimental results.

Automatic Text Classification Methods

As presented in the study design, the ML methods were divided into two categories: two-phase methods and end-to-end (E2E) methods [35]. The first category of methods consisted of approaches whose document (ie, sentence) representation was intrinsically independent of the classification algorithm used to predict the class. In other words, the classifier used to predict the class of documents was not used in the construction phase of the document representation. In terms of text representations, we considered three alternatives, namely traditional term-weighting alternatives (term frequency-inverted document frequency [TFIDF]); weighting based on word and character (n-gram) frequency; and recent representations based on meta-features, which capture statistical information from a document's neighborhood and have obtained state-of-the-art effectiveness in recent benchmarks [35-39].

As two-phase classification algorithms, we exploited support vector machines (SVMs), which are still considered the most robust nonneural network text classification algorithm [35,39,40], random forests (RF), K-nearest neighbor (KNN), and naïve Bayes classifier (NBC), to address the most popular algorithms in terms of classification and retrieval of text information [41-44].

In contrast, E2E methods use a discriminative classifier function to transform the document representation space into a new and more informed (usually more reduced and compact) space and use this classifier to predict the document class. In general, these approaches use an iterative process of representation, classification, evaluation, and parameter adaptation (eg, transform, predict, evaluate loss function, and backpropagate, respectively). For E2E classifiers, we exploited two neural architectures, namely convolutional neural networks (CNNs), which exploit textual patterns such as word co-occurrences, and bidirectional encoder representation from transformers (BERT), which exploits attention mechanisms and constitute the current state-of-the-art in many NLP tasks.

Finally, we exploited a rule-based classifier specialized for the tasks at hand (stroke tasks, represented in the ontology web language [OWL]). The rule-based knowledge model was developed using logical conditions built alongside domain specialists [45]. This technique has shown effectiveness equivalent to that of some ML classification models in certain domains without the need for a large amount of data and training time, which are commonly required by supervised methods [46-49]. In contrast, it is heavily dependent on the specialists and the coverage of the rules on the text expressions. More details about each of the exploited algorithms are provided in Multimedia Appendix 4 [3,35,37,39,41-45,50-63].

The two-phase methods used in this research are referred to as the representation technique combined with the classification algorithm, as follows: word-TFIDF and character-TFIDF combined with SVM (SVM+W+C), Bag-of-Words (BoW) combined with SVM (SVM+BoW), meta-features combined with SVM (meta-features), word-TFIDF combined with SVM (SVM+Word-TFIDF), character-TFIDF combined with SVM (SVM+Chard-TFIDF), Word-TFIDF combined with random forest (RF+Word-TFIDF), word-TFIDF combined with KNN (KNN+Word-TFIDF), and word-TFIDF combined with naïve Bayes (Naïve Bayes+Word-TFIDF). In contrast to TFIDF, BoW explores only the frequency of terms (term frequency) and not the frequency of terms in the collection (IDF component). The E2E methods are simply called CNN and BERT, and the ontological method is called OWL.

Experimental Evaluation

Overview

The experimental process consisted of testing different classification methods with sets of annotated data to assess and compare their performances (effectiveness). The experimental procedure, described in Multimedia Appendix 5, consisted of four phases: (1) representing the free-text sentences as numerical vectors, (2) the training and tuning process (in a validation set) by means of a folded cross-validation procedure, (3) the execution of the classification algorithms in the test set and effectiveness assessment, and (4) the synthesis of the results in a heatmap table.

A classification model was developed for each task. Each task resulted in an individual automatic classification model for the training and testing process of the model. As an experimental protocol, we used a five-fold cross-validation procedure repeated six times (resulting in 30 test samples). We also exploited *subject-wise cross-validation* in the sense that the information from the same patient was always assigned to the same fold to test the ability of the model to predict new data that was not used in the learning process. These procedures address potential problems, such as overfitting and selection bias [64], and produce results that are more reliable.

To evaluate the ability to classify the relevant Brazilian-Portuguese medical free-text records correctly, we used the Macro-F1 score (equation 1). This metric is based on a *confusion matrix* and is defined as follows:

$$F_1 = 2 * \frac{PRECISION * RECALL}{PRECISION + RECALL} = \frac{2TP}{2TP + FP + FN} \qquad \textbf{(1)}$$

where TP is true positive, TN is true negative, FP is false positive, and FN is false negative. Precision (positive predictive

XSL•FO
RenderX

value) = TP / TP + FP = the number of returned hits that were true positive. Recall (sensibility) = TP / TP + FN = is the fraction of the total number of true positives retrieved.

The F1 measure is calculated for each class. Macro-F1 summarizes the classification effectiveness by averaging F1 values for all classes. Macro-F1 is one of the most popular aggregated evaluation metrics for the classifier evaluation of unbalanced or skewed data sets [42,65,66]. Macro-F1 is especially suitable for imbalanced data sets, as the effectiveness of each individual class contributes equally to producing a final score. For instance, in a task with four classes, in which one of them is NI, if all classes are predicted as NI, the Macro-F1 score will be no higher than 0.25 (F1 of 1 for NI and 0 for the three other classes). Accuracy or any other evaluation measure focused on the instance, instead of the class effectiveness, would produce a very high score (close to 1 in this particular case).

To compare the average results of our cross-validation experiments, we assessed statistical significance by using a paired two-tailed *t* test with 95% CIs. To account for multiple tests, we adopted the Friedman-Nemenyi test [67] with Bonferroni correction for multiple comparisons of mean rank sums. The Friedman test was used to compare multiple methods.

We consider that making the data and the code used in our experimental protocol available to others is potentially useful for reproducibility and for use in other studies. Both the code and data will be available upon request. The mood-specific parameter tuning details are presented in Multimedia Appendix 6.

Experimental Analysis

The experiments aimed to provide relationships between the classification methods and the tasks, allowing for connecting the best methods with each outcome measure or patient characteristics. Considering that the model's results can influence health decision-making in some way, the F1 score thresholds may vary depending on the type of class and the imbalance of the data. We reported the results by means of a heatmap, adopting a red color for F1<20%, a gradual color scale from orange to yellow for 21%<F1<79%, and green for F1>80% [68-71]. Tasks (represented by the lines) were ordered by the average of the performed models, whereas the ordering of the columns shows the rank position of each method according to the statistical analysis.

For the sake of the fairness of the comparison, the OWL technique should not be and is not directly compared and ranked herein along with the other ML models described above that require a combination of text representations with trained classification algorithms. OWL rules were designed to work with the entire corpus (including the test) and were not designed for generalization. Instead, they are built to work well in the specific domain or task for which they were created. In any case, for reasons of practical application and as a research exercise, as a secondary analysis, we compared (later) the OWL technique with the ML model ranked as the best based on the Friedman test. This analysis allowed us to identify the weaknesses and strengths of both approaches (generalized ML models vs domain or task-specific ontological rules) in the contrasting tasks.

Moreover, we performed a feature selection analysis [72,73]. This technique is used to rank the most informative features of each task according to the information theory criteria. In particular, we used SelectKBest (Python Software Foundation; Python Language Reference, version 2.7) with the chi-square, which is independent of the classification algorithms used [74]. This final analysis helps in understanding how ML can help with outcome measurements for the stroke care pathway, potentially boosting advances in quality indicator automation.

Finally, to complete the analysis and evaluate the impact of the highly skewed distribution, especially toward the NI class, we ran an experiment in which we performed a random undersampling process for all considered tasks (we used the RandomUnderSampler Phyton library [75]). In detail, we randomly selected the same number of training random examples of the NI as the number of instances of the second largest (non-NI) class of a given task. We then reran all ML classifiers (the ontology method is not affected by this process as it has no training) in all 24 tasks, considering as the training set the reduced (undersampled) NI training samples along with the same (unchanged) previous samples for the other classes. We did that for all six rounds of five-fold cross-validation of our experimental procedure, changing the seed for selection in each round, resulting in six different NI reduced training sets. The test folds in all cases remain unchanged, meaning that we keep the same skewed distribution as in the original data set, as we do not know the class of the test instances.

Results

Tasks Selection

Discussions with experts in the stroke care pathway allowed us to define 30 tasks that were considered feasible to extract from EMRs. For the first tier, the standard sets were usually defined to evaluate the clinical stroke outcomes that were used, including the mRS [27] and the NIHSS scales [76], in addition to traditional outcomes such as mortality and pain level. For tier 2, the ICHOM standard set developed for ischemic stroke was used [77], which considers measures of mobility, ability to communicate, ability to feed orally, the ability to understand, and measures and scales of strength level. Indicators of the hospitalization care process used in the institution were also included, such as rating scales and risk events tracked by fall risk, pressure ulcer risk, fall events during hospitalization, infection indicators, intracranial hemorrhage, therapy care (thrombolytic, thrombectomy, or both), and the location of the patient during the inpatient path [78]. Finally, baseline characteristics important for tracking the population and further risk-adjusted analysis were included [79], such as high blood pressure, smoking status, coronary artery disease, atrial fibrillation, diabetes, prior stroke, active cancer, alcoholism, obesity, and dyslipidemia. Each category, containing the tasks and their respective classes, is presented in Table 1.

XSL•FO
RenderX

Table 1. Eligible tasks for analysis and classification rules.

Tasks	Number of classes	Supporting information for classes
Health care status achieved (tier 1)		
Rankin	8	• 0-6 • NI[a]
National Institutes of Health Stroke Scale	42	• 1-41 • NI
Death	3	• Absence of vital signs • Vital signs present • NI
Process of recovery (tier 2)		
Mobility level	16	• 1-15 • NI
Self-care	3	• Able • Unable • NI
Pain	4	• No pain • Low to intermediate pain • Intense pain • NI
Strength	7	• 0-5 • NI
Paresis	3	• Yes • No • NI
Ability to feed orally	3	• Yes • No • NI
Ability to communicate	4	• Yes • No • Poorly or symptomatic • NI
Ability of understanding	4	• Yes • No • Poorly or symptomatic • NI
Ability to ambulate	4	• Yes • No • Poorly or symptomatic • NI
Treatment or care related		
Thrombolytic therapy	3	• No delta • Yes • NI
Thrombectomy	3	• No delta • Yes • NI

XSL•FO
RenderX

Tasks	Number of classes	Supporting information for classes
Location	4	- Emergency room - ICU[b] - Inpatient unit - NI
Infection indication	3	- Yes - No - NI
Intracranial hemorrhage	3	- Yes - No - NI
Fall risk	4	- Low risk - Moderate risk - High risk - NI
Pressure ulcer risk	4	- Low risk - Moderate risk - High risk - NI
Fall event during inpatient	3	- Yes - No - NI
Baseline characteristics		

XSL•FO
RenderX

Tasks	Number of classes	Supporting information for classes
High blood pressure	3	• Yes • No • NI
Smoking status	4	• Yes • No • Former • NI
Coronary artery disease	3	• Yes • No • NI
Atrial fibrillation	3	• Yes • No • NI
Diabetes	3	• Yes • No • NI
Prior stroke	3	• Yes • No • NI
Cancer	3	• Yes • No • NI
Alcoholism	4	• Yes • No • Former • NI
Obesity	3	• Yes • No • NI
Dyslipidemia	3	• Yes • No • NI

[a]NI: noninformative.

[b]ICU: intensive care unit.

After the identification of all tasks and the annotation process, the analysis proceeded only with tasks that had substantial (0.61>κ>0.80) and almost perfect (κ≥0.81) agreement between annotators [34]. A total of six tasks were excluded from the final analysis because of moderate or fair agreement or disagreement: (1) active cancer information, (2) strength level, (3) intracranial hemorrhage, (4) ability to understand, (5) self-care, and (6) fall events during inpatient visits. All documents were labeled by the annotators, and the median κ regarding the 24 remaining tasks was 0.74 (IQR 0.65-0.89; substantial agreement).

Patient Characteristics

The descriptive characteristics of patients, including previous comorbidities, NIHSS score, and clinical care, are presented in Table 2.

XSL•FO
RenderX

Table 2. Descriptive characteristics of the patients.

Characteristics	Patients with ischemic stroke evaluated (n=188)	
	Values, median (range)	Values, n (%)
Age (years)	79 (68-87)	N/A[a]
LOS[b] (days)	6 (4-12)	N/A
Sex		
Female	N/A	100 (53)
Male	N/A	88 (47)
Comorbidities		
Previous stroke	N/A	38 (20)
Previous coronary artery disease	N/A	12 (6)
Atrial fibrillation	N/A	33 (18)
Diabetes	N/A	53 (28)
Hypertension	N/A	125 (66)
Smoking status	N/A	15 (8)
Alcoholism	N/A	4 (2)
Treatment and care related		
Antithrombotic therapy	N/A	131 (70)
Thrombolysis with rtPA[c]	N/A	38 (20)
Thrombectomy	N/A	12 (6)
Thrombolysis and thrombectomy	N/A	7(4)
NIHSS[d]		
<8	N/A	147 (78)
>8 and <15	N/A	24 (13)
>15	N/A	17 (9)

[a] N/A: not applicable.

[b] LOS: length of stay.

[c] rtPA: alteplase.

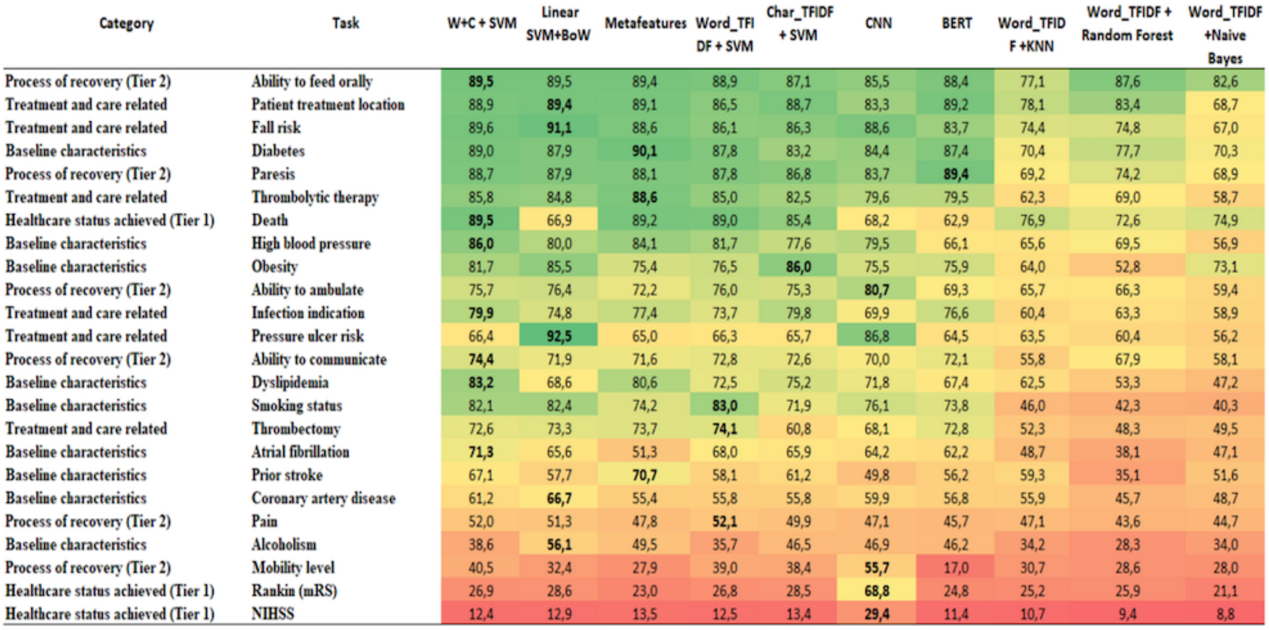
[d] NIHSS: National Institutes of Health Stroke Scale.

Experimental Results

The Macro-F1 values for each of the 24 tasks using the 10 compared models are shown in Figure 2. Considering each task separately, there is no single method that always dominates, and there is no agreement on a unique category of tasks that perform better. The ML models SVM+W+C and SVM+BoW were the best and most consistent techniques used in this data set. Both techniques use term-weighting representations that are used alongside SVM classifiers. The latter simply exploits within-document word term frequencies (term frequency), whereas the former, in addition to exploiting data set–oriented term statistics (IDF), also builds character-based n-gram representations of the words in the vocabulary. The character-based n-grams, despite increasing the vocabulary size and sparsity, help to deal with misspellings and word variations that are common in EMRs, which might explain the SVM+W+C good results.

XSL•FO RenderX

Figure 2. Results of Macro-F1 for each task and comparative models (expressed in percentage). BERT: bidirectional encoder representation from transformers; CNN: convolutional neural network; mRS: Modified Rankin Score; NIHSS: National Institutes of Health Stroke Scale; SVM+BoW: support vector machine plus Bag-of-Words; TFIDF: term frequency-inverted document frequency; W+C+SVM: word-term frequency-inverted document frequency and character-term frequency-inverted document frequency combined with support vector machine.



Category	Task	W+C + SVM	Linear SVM+BoW	Metafeatures	Word_TFIDF + SVM	Char_TFIDF + SVM	CNN	BERT	Word_TFIDF +KNN	Word_TFIDF + Random Forest	Word_TFIDF +Naive Bayes
Process of recovery (Tier 2)	Ability to feed orally	**89,5**	89,5	89,4	88,9	87,1	85,5	88,4	77,1	87,6	82,6
Treatment and care related	Patient treatment location	88,9	**89,4**	89,1	86,5	88,7	83,3	89,2	78,1	83,4	68,7
Treatment and care related	Fall risk	89,6	**91,1**	88,6	86,1	86,3	88,6	83,7	74,4	74,8	67,0
Baseline characteristics	Diabetes	89,0	87,9	**90,1**	87,8	83,2	84,4	87,4	70,4	77,7	70,3
Process of recovery (Tier 2)	Paresis	88,7	87,9	88,1	87,8	86,8	83,7	**89,4**	69,2	74,2	68,9
Treatment and care related	Thrombolytic therapy	85,8	84,8	**88,6**	85,0	82,5	79,6	79,5	62,3	69,0	58,7
Healthcare status achieved (Tier 1)	Death	**89,5**	66,9	89,2	89,0	85,4	68,2	62,9	76,9	72,6	74,9
Baseline characteristics	High blood pressure	**86,0**	80,0	84,1	81,7	77,6	79,5	66,1	65,6	69,5	56,9
Baseline characteristics	Obesity	81,7	85,5	75,4	76,5	**86,0**	75,5	75,9	64,0	52,8	73,1
Process of recovery (Tier 2)	Ability to ambulate	75,7	76,4	72,2	76,0	75,3	**80,7**	69,3	65,7	66,3	59,4
Treatment and care related	Infection indication	**79,9**	74,8	77,4	73,7	79,8	69,9	76,6	60,4	63,3	58,9
Treatment and care related	Pressure ulcer risk	66,4	**92,5**	65,0	66,3	65,7	86,8	64,5	63,5	60,4	56,2
Process of recovery (Tier 2)	Ability to communicate	**74,4**	71,9	71,6	72,8	72,6	70,0	72,1	55,8	67,9	58,1
Baseline characteristics	Dyslipidemia	**83,2**	68,6	80,6	72,5	75,2	71,8	67,4	62,5	53,3	47,2
Baseline characteristics	Smoking status	82,1	82,4	74,2	**83,0**	71,9	76,1	73,8	46,0	42,3	40,3
Treatment and care related	Thrombectomy	72,6	73,3	73,7	**74,1**	60,8	68,1	72,8	52,3	48,3	49,5
Baseline characteristics	Atrial fibrillation	**71,3**	65,6	51,3	68,0	65,9	64,2	62,2	48,7	38,1	47,1
Baseline characteristics	Prior stroke	67,1	57,7	**70,7**	58,1	61,2	49,8	56,2	59,3	35,1	51,6
Baseline characteristics	Coronary artery disease	61,2	**66,7**	55,4	55,8	55,8	59,9	56,8	55,9	45,7	48,7
Process of recovery (Tier 2)	Pain	52,0	51,3	47,8	**52,1**	49,9	47,1	45,7	47,1	43,6	44,7
Baseline characteristics	Alcoholism	38,6	**56,1**	49,5	35,7	46,5	46,9	46,2	34,2	28,3	34,0
Process of recovery (Tier 2)	Mobility level	40,5	32,4	27,9	39,0	38,4	**55,7**	17,0	30,7	28,6	28,0
Healthcare status achieved (Tier 1)	Rankin (mRS)	26,9	28,6	23,0	26,8	28,5	**68,8**	24,8	25,2	25,9	21,1
Healthcare status achieved (Tier 1)	NIHSS	12,4	12,9	13,5	12,5	13,4	**29,4**	11,4	10,7	9,4	8,8

The SVM+W+C model excels in tasks belonging to different categories, such as the ability to feed orally (Tier 2: the process of recovery), with an F1 score of 89.5% (95% CI 89.2%-89.8%); death (tier 1: health care status achieved), with an F1 score of 89.5% (95% CI 87.5%-92.5%); and high blood pressure and dyslipidemia (the baseline characteristics of patients), with F1 scores of 86% (95% CI 83.8%-88.2%) and 83.2% (95% CI 77%-89%), respectively. SVM+BoW, in turn, excels in tasks belonging to the treatment- or care-related categories, such as patient location during treatment (F1 score 89.4%; 95% CI 88%-91%), fall risk (F1 score 91.1%; 95% CI 90.1%-92.1%), and pressure ulcer risk (F1 score 92.5; 95% CI 91.5%-93.5%). The meta-features model, which also exploits SVM as a classifier but uses a completely different text representation, was on average, the third-best placed ML model to cover more tasks with good effectiveness, except in tasks such as diabetes (F1 score 90.1%; 95% CI 88.8%-91.4%) and thrombolytic therapy (F1 score 88.6%; 95% CI 87.5%-90.1%), in which it was the sole winner model (best performer with no ties). The models that used SVM but exploited either only word- or character-based representations came in the fourth and fifth places, losing to methods that exploited both representations in a conjugated way.

The neural methods CNN and BERT were grouped in the middle, with only moderate effectiveness in most tasks. This outcome is mostly due to the lack of sufficient training data for the optimal deployment of these methods. Indeed, previous work has demonstrated that neural solutions are not adequate for tasks with low to moderate training data, and they can only outperform other more traditional ML methods in text classification tasks when presented with massive amounts of training [35,39], which is generally uncommon in the health domain.

Regarding the effectiveness of the tasks, patient characteristics and care-related process tasks produced better effectiveness. Five of them are examples of good adherence with multiple models, including patient treatment location, fall risk, thrombolytic therapy, diabetes, and paresis, all with multiple models with high effectiveness. Tasks related to measures of mobility, ability to communicate, ability to ambulate, and pain did not achieve high Macro-F1 values in most models.

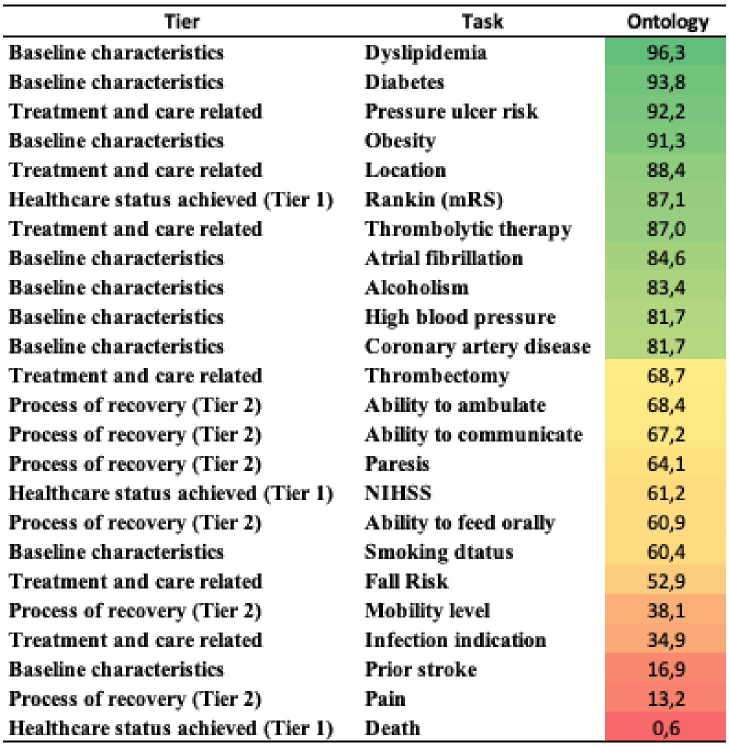
The tasks with many classes, such as NIHSS (42 classes), mobility level (n=16), and Rankin (n=8), performed worse, regardless of the model. This outcome is mostly due to issues related to the very skewed distribution (high imbalance) found in our unstructured real-life data set. Indeed, the high percentage of NI in the document penalizes effectiveness, mainly for the minor classes, which are captured more faithfully by the Macro-F1 score. However, properly dealing with such an imbalance is not a simple task, as discussed next. Finally, as the sentence length was very similar across tasks and classes, this factor did not affect the results, that is, we could not infer any significant relationship between the mean number of words per sentence and the Macro-F1 scores of the models.

Figure 3 provides information regarding the effectiveness of the OWL classifier. In general, the OWL effectiveness is similar to that of the best ML models, with 11 tasks having a Macro-F1 score higher than 80%. The most interesting issue is that most of the best-performing tasks by OWL *do not coincide* with the best ones produced by the ML models in Figure 2. For instance, the OWL classifier performed very well on the patient's baseline characteristics tasks, such as NIHSS and mRS scale, precisely the ones in which the ML models performed poorly. Overall, the OWL strategy was more robust in the tasks in which the ML models suffered from a scarcity of examples and high imbalance. On the contrary, OWL suffered on tasks that were much more passible in interpretation and had more text

XSL•FO
RenderX

representations from those for which they were built [49,80]. For instance, in the *death* task, despite good within-annotator agreement, we believe that due to a variety of clinical terms in the clinical text used to describe multiple clinical concepts, the rules initially created failed to reflect the understanding of a noninformative sentence versus a sentence that reports the vital signs of patients, which penalized the OWL model.

Figure 3. Effectiveness results for the ontology-based model. mRS: Modified Rankin Score; NIHSS: National Institutes of Health Stroke Scale.



Tier	Task	Ontology
Baseline characteristics	Dyslipidemia	96,3
Baseline characteristics	Diabetes	93,8
Treatment and care related	Pressure ulcer risk	92,2
Baseline characteristics	Obesity	91,3
Treatment and care related	Location	88,4
Healthcare status achieved (Tier 1)	Rankin (mRS)	87,1
Treatment and care related	Thrombolytic therapy	87,0
Baseline characteristics	Atrial fibrillation	84,6
Baseline characteristics	Alcoholism	83,4
Baseline characteristics	High blood pressure	81,7
Baseline characteristics	Coronary artery disease	81,7
Treatment and care related	Thrombectomy	68,7
Process of recovery (Tier 2)	Ability to ambulate	68,4
Process of recovery (Tier 2)	Ability to communicate	67,2
Process of recovery (Tier 2)	Paresis	64,1
Healthcare status achieved (Tier 1)	NIHSS	61,2
Process of recovery (Tier 2)	Ability to feed orally	60,9
Baseline characteristics	Smoking dtatus	60,4
Treatment and care related	Fall Risk	52,9
Process of recovery (Tier 2)	Mobility level	38,1
Treatment and care related	Infection indication	34,9
Baseline characteristics	Prior stroke	16,9
Process of recovery (Tier 2)	Pain	13,2
Healthcare status achieved (Tier 1)	Death	0,6

A direct comparison between OWL and the best ML method is presented in Figures 4 and 5, in which Figure 4 represents the tasks in which OWL performed better than the best ML model for the same tasks and Figure 5 represents the tasks with higher F1 scores in the ML model against OWL. SVM+W+C has a considerable advantage over the other ML strategies, as the strategy of choice to be compared in the vast majority of cases. The best tasks performed by the best model in each case, either SVM+W+C or OWL, do not coincide. Indeed, there is a potential complementarity between ML and alternatives.

XSL•FO
RenderX

Figure 4. Best performed tasks in Ontology versus top-ranked model. mRS: Modified Rankin Score; NIHSS: National Institutes of Health Stroke Scale; SVM: support vector machine; W+C+SVM: word- term frequency-inverted document frequency and character- term frequency-inverted document frequency combined with support vector machine.

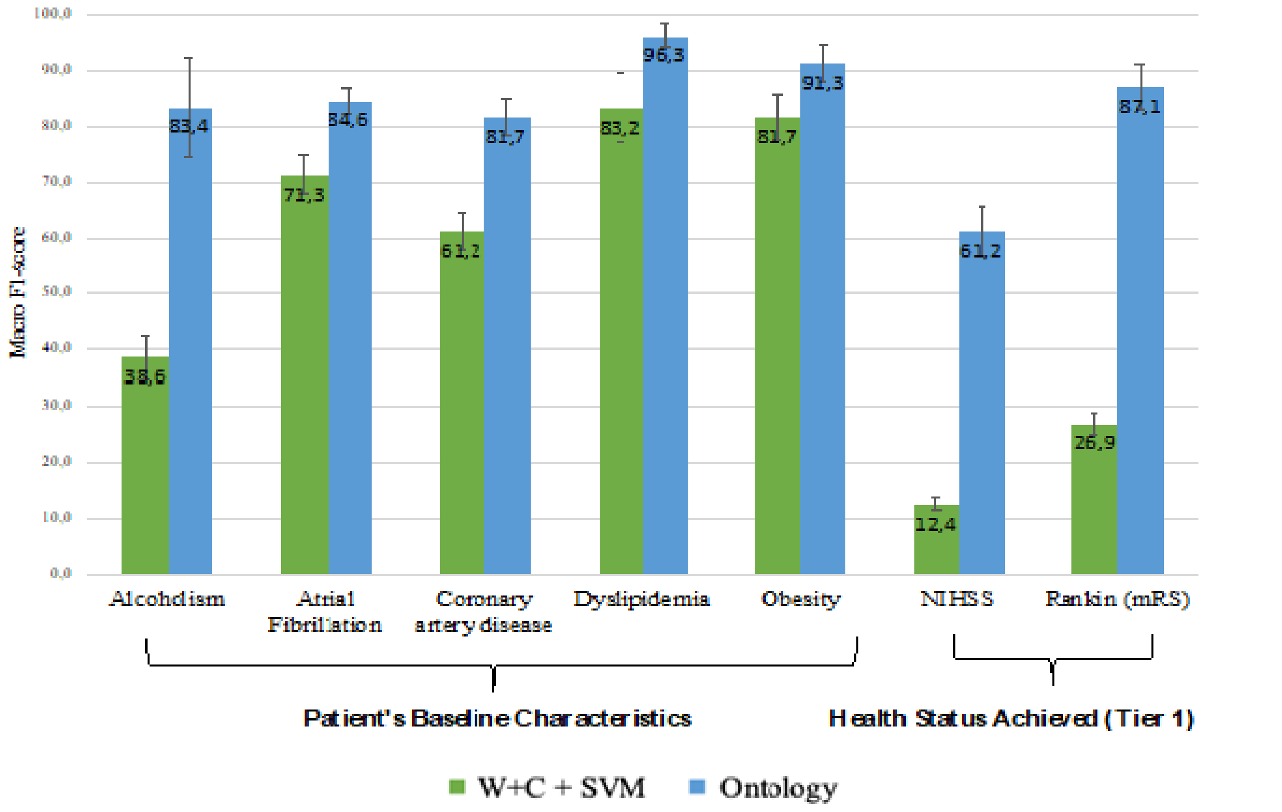
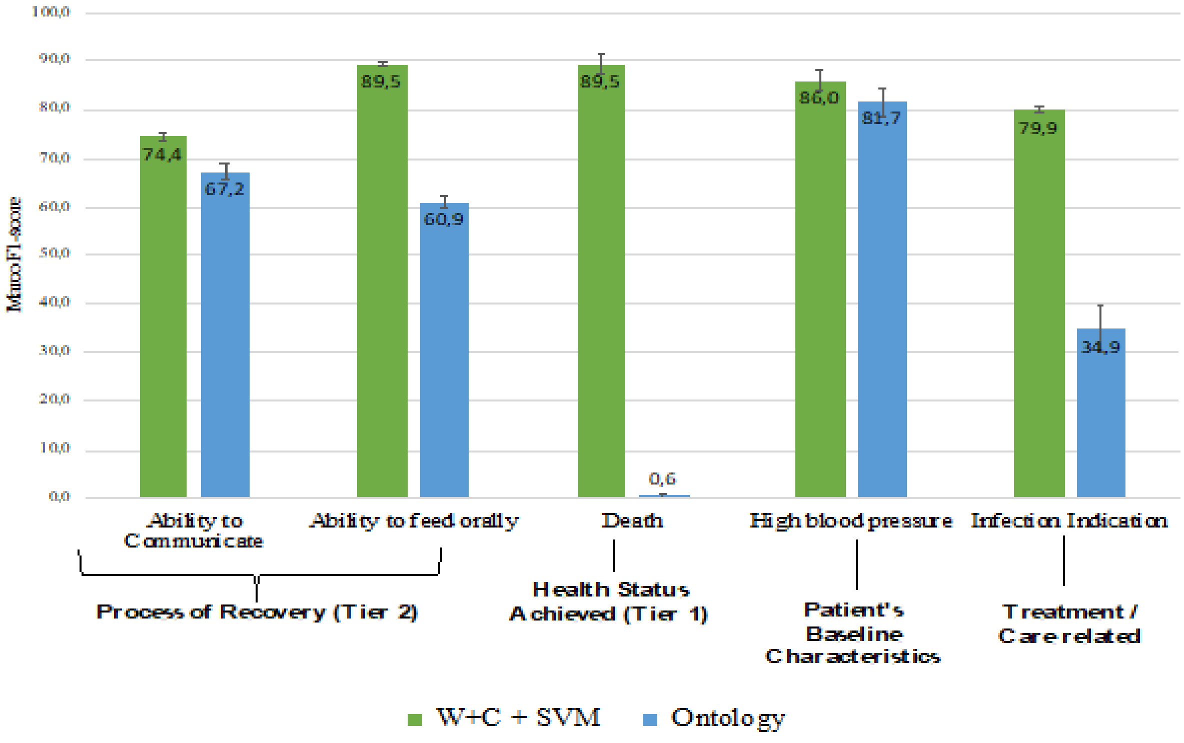


Figure 5. Best performed tasks in the top-ranked model versus Ontology. SVM: support vector machine; W+C: word-term frequency-inverted document frequency and character-term frequency-inverted document frequency.



XSL•FO
RenderX

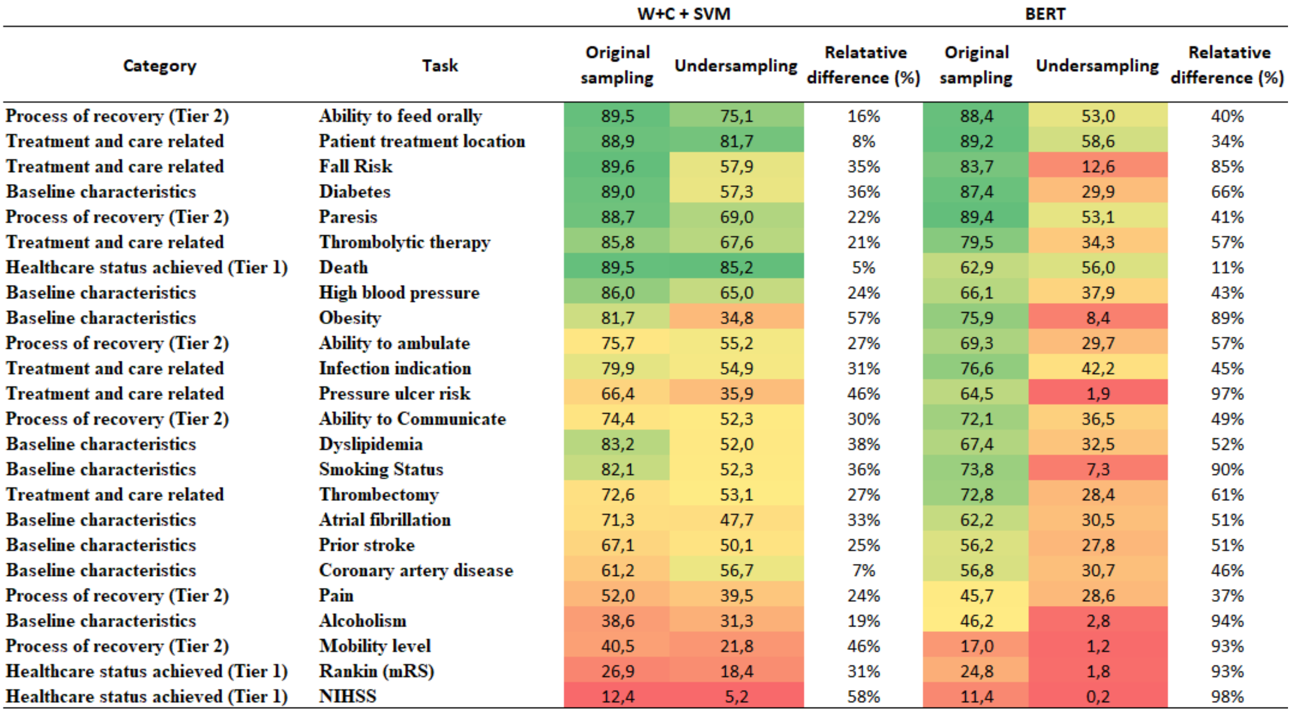
Effect of Class Imbalance on the Results—Undersampling

As we have discussed, all our tasks are extremely skewed, in the sense that the NI (noninformed; majority) class dominates over the other (minority) classes, where the useful information really lies. This imbalance occurs in a proportion that can achieve 1:1000 examples in the minority class to the majority class for some tasks.

This imbalance may cause bias in the training data set influencing some of the experimented ML algorithms toward giving priority to NI class, ultimately undermining the classification of the minority classes on which predictions are most important. One approach to addressing the problem of class imbalance is to randomly resample the training data set. A simple, yet effective approach to deal with the problem is to randomly delete examples from the majority class, a technique known as random undersampling [81].

The results of this experiment are shown in Figure 6, which compares the performance of the classifiers in scenarios with and without undersampling. For the sake of space, we only show the results for the best nonneural (W+C+SVM) and neural (BERT) classifiers, but the results are similar for all tested classifiers (Multimedia Appendix 7).

Figure 6. Results of Macro-F1 score in the undersampling sample, expressed by percentage. mRS: Modified Rankin Score; NIHSS: National Institutes of Health Stroke Scale; SVM: support vector machine; W+C: word- term frequency-inverted document frequency and character- term frequency-inverted document frequency.



		W+C + SVM			BERT		
Category	Task	Original sampling	Undersampling	Relatative difference (%)	Original sampling	Undersampling	Relatative difference (%)
Process of recovery (Tier 2)	Ability to feed orally	89,5	75,1	16%	88,4	53,0	40%
Treatment and care related	Patient treatment location	88,9	81,7	8%	89,2	58,6	34%
Treatment and care related	Fall Risk	89,6	57,9	35%	83,7	12,6	85%
Baseline characteristics	Diabetes	89,0	57,3	36%	87,4	29,9	66%
Process of recovery (Tier 2)	Paresis	88,7	69,0	22%	89,4	53,1	41%
Treatment and care related	Thrombolytic therapy	85,8	67,6	21%	79,5	34,3	57%
Healthcare status achieved (Tier 1)	Death	89,5	85,2	5%	62,9	56,0	11%
Baseline characteristics	High blood pressure	86,0	65,0	24%	66,1	37,9	43%
Baseline characteristics	Obesity	81,7	34,8	57%	75,9	8,4	89%
Process of recovery (Tier 2)	Ability to ambulate	75,7	55,2	27%	69,3	29,7	57%
Treatment and care related	Infection indication	79,9	54,9	31%	76,6	42,2	45%
Treatment and care related	Pressure ulcer risk	66,4	35,9	46%	64,5	1,9	97%
Process of recovery (Tier 2)	Ability to Communicate	74,4	52,3	30%	72,1	36,5	49%
Baseline characteristics	Dyslipidemia	83,2	52,0	38%	67,4	32,5	52%
Baseline characteristics	Smoking Status	82,1	52,3	36%	73,8	7,3	90%
Treatment and care related	Thrombectomy	72,6	53,1	27%	72,8	28,4	61%
Baseline characteristics	Atrial fibrillation	71,3	47,7	33%	62,2	30,5	51%
Baseline characteristics	Prior stroke	67,1	50,1	25%	56,2	27,8	51%
Baseline characteristics	Coronary artery disease	61,2	56,7	7%	56,8	30,7	46%
Process of recovery (Tier 2)	Pain	52,0	39,5	24%	45,7	28,6	37%
Baseline characteristics	Alcoholism	38,6	31,3	19%	46,2	2,8	94%
Process of recovery (Tier 2)	Mobility level	40,5	21,8	46%	17,0	1,2	93%
Healthcare status achieved (Tier 1)	Rankin (mRS)	26,9	18,4	31%	24,8	1,8	93%
Healthcare status achieved (Tier 1)	NIHSS	12,4	5,2	58%	11,4	0,2	98%

As it can been seen, the undersampling process caused major losses in both classifiers. Such losses occurred across all tasks, varying from 5% of Macro-F1 score reduction (death) to 58% (NIHSS) for W+C+SVM, and 11% (death) to 98% (NIHSS) of Macro-F1 effectiveness loss in BERT. The largest losses for the neural method were expected, as this type of classifier is more sensitive to the amount of training. However, to a certain degree, all the classifiers suffered major losses after the undersampling process. These results may be attributed to the largest difference in class distribution between training and testing and the inevitable loss of information that comes after the removal of training instances after undersampling.

These phenomena can be better seen when we look at the individual values of F1, precision, and recall of the classes of the tasks. Table 3 shows an example of the tasks of infection indication, thrombolytic therapy, and ability to communicate with the W+C+SVM classifier. As we can see, all classes have a reduced F1 in the undersampling scenario. This is mainly due to a large reduction in the precision of the classes. This happens because W+C+SVM misclassifies NI instances as belonging to some of the relevant classes. As the classifier is obliged to categorize a sentence in one of the existing classes, the lack of information about the fact that a sentence does not have useful information for assigning the sentence in one of the classes of interest confounds the classifier. In other words, the negative information about the NI (eg, frequent words in NI sentences that help to characterize this class but that are also shared by some non-NI instances, and whose frequency was altered by the undersampling) is in fact useful information for avoiding false positives, which may cause many problems in a real scenario, including false alarms, waste of resources, and distrust of the automatic methods.

XSL•FO
RenderX

Table 3. Comparison of undersampling and original sampling in terms of precision, recall, and Macro-F1 score (W+C+SVM model).

Class	Undersampling			Original sampling		
	Precision	Recall	F1 (%)[a]	Precision	Recall	F1 (%)[a]
Infection indicative						
−1	1	0.96	98	0.99	1	99
0	0.39	0.89	54	0.88	0.75	81
1	0.28	0.82	42	0.68	0.53	59
Thrombolytic therapy						
−1	1	0.98	99	1	1	100
0	0.32	0.62	42	0.69	0.52	59
1	0.31	0.91	47	0.89	0.91	90
Ability to communicate						
−1	1	0.96	98	0.99	1	100
0	0.34	0.63	44	0.9	0.26	40
1	0.35	0.81	49	0.76	0.64	69
2	0.32	0.93	48	0.82	0.8	81

[a] Macro-F1 score (W+C+SVM model).

Feature Importance

For the tasks presented in Textbox 1 (alcoholism, atrial fibrillation, coronary artery disease, dyslipidemia, obesity, NIHSS, Rankin [mRs], infection indicators, high blood pressure, death, ability to feed orally, and ability to communicate), we present the top 10 clinical features (ie, words) used in the task prediction in Textbox 1, which means the 10 features with higher contribution to task prediction. This analysis helps to better understand the divergence between approaches. It is worth noting that in the tasks in which the ML models performed better (second column), the top-ranked features were all related to the semantics of the task. For instance, considering the *death* task as an example, the ML model was able to identify important features for the task, which produced a higher information gain than the OWL model. Indeed, for *death*, only three features of the 10 most relevant explicitly use the word *death*, but most features are somewhat related to this outcome. This finding suggests data quality issues (vocabulary coverage) that may drastically influence the effectiveness of the OWL strategy, which exploits only rules that explicitly contain the word *death* (or related ones) but no other terms. However, for the features in the first column, in which the OWL models were better, there were still features with considerable contributions that were not directly related to the information sought. For example, to mention the NIHSS task, rule-based knowledge models built alongside clinical domain vocabulary specialists may be the best alternative.

XSL•FO
RenderX

Textbox 1. Top 10 clinical indicators for task prediction models and feature importance. In parenthesis, the translation to English language is indicated, where there may be misspellings in the original writing that are also indicated.

Alcoholism

- etilismo (alcoholism)
- etilista (alcoholic)
- fumo (smoke)
- históira (story with misspelling in the original)
- álcool (alcohol)
- cart
- osteoartrose (osteoarthritis)
- ttu (short for transurethral resection of the prostate)
- tabagismo (smoking)
- cesária (cesarean)

Atrial fibrillation

- fa (short for atrial fibrillation)
- comorbidades (comorbidities)
- acfa (short for atrial fibrillation)
- paroxística (paroxysmal)
- has (short for high blood pressure)
- anticoagulado (anticoagulated)
- depressão (depression)
- indeterminado (indeterminate)
- digoxina (digoxin)
- institucionalizada (institutionalized)

Coronary artery disease

- cardiopatia (heart disease)
- isquêmica (ischemic)
- actp (short for percutaneous transluminal coronary angioplasty)
- dp
- crm (short for myocardial revascularization surgery)
- iam (short for acute myocardial infarction)
- 2014
- infarto (short for acute myocardial infarction)
- mm
- sf

Dyslipidemia

- dislipidemia (dyslipidemia)
- comorbidades (comorbidities)
- 1hora
- cesária (cesarean)
- morbidades (morbidities)
- puerpera (puerperal)
- has (short for high blood pressure)

- fêmur (fêmur)
- tep
- previas (previous)

Obesity

- BMI (short for body mass index)
- obesidade (obesity)
- m²
- 1994
- lipschitz
- eutrofia
- altura (height)
- peso (weight)
- estatura (stature)
- obesa (obese)

National Institutes of Health Stroke Scale

- nihss
- súbito (sudden)
- asistolia (asystolia)
- sens
- territ
- suboclusiva (subocclusive)
- perg
- mecania (mecanic with mispelling in the original)
- severo (severe)
- visto (seen)

Ability to communicate

- afasia (afasia)
- comunicativa (talkative)
- disartria (dysarthria)
- comunicativo (talkative)
- colóquio (colloquium)
- verbalizando (verbalizing)
- alerta (alert)
- verbaliza (verbalizes)
- expressão (expression)
- hemiparesia (hemiparesis)

Ability to feed orally

- vo (short for orally)
- sne (short for nasoenteral probe)
- dieta (diet)
- pastosa (pasty)
- gastrostomia (gastrostomy)

XSL•FO
RenderX

- enteral (enteral)
- aceitação (acceptance)
- semi (semi)
- exclusiva (exclusive)
- polimérica (polymeric diet)

Death

- óbito (death)
- constato (i’ve verified)
- leito (bed)
- ar (air)
- estável (stable)
- ambiente (environment or room)
- no
- doação (donation)
- obito (death with misspelling in the original)
- óbito (death with misspelling in the original)

High blood pressure

- has (short for high blood pressure)
- dm (short for diabetes)
- dislipidemia (dyslipidemia)
- dm2 (short for diabetes type 2)
- comorbidades (comorbidities)
- fa (short for atrial fibrillation)
- artrite (arthritis)
- definitivo (definitive)
- reumatoide (rheumatoid)
- demencial (dementia)

Infection indication

- afebril (afebrile)
- flogísticos (phlogistic)
- sinais (signs)
- cefuroxima (cefuroxime)
- inserção (insertion)
- tax
- klebsiella (klebsiella)
- d0 (short for day 0)
- atb (short for antibiotics)
- azitromicina (azithromycin)

Modified Rankin Score

- rankin
- mrankin
- demência (dementia)

XSL•FO
RenderX

- caminha (walks)
- corversa (talks)
- alimenta (feed)
- alzheimer
- aparentes (apparent)
- comer (eat)
- mrk (mrs with misspelling in the original)

Discussion

Principal Findings

The study intended to recognize the path and opportunities that may be advanced in terms of the technological capacity to support the outcome measurement process for the stroke care pathway. Real-world sentences from ischemic stroke EMRs were used to develop automatic models using ML and NLP techniques. It was possible to identify that SVM+W+C and SVM+BoW were the most effective models to be used to classify characteristics of a patient and process of care based on the extraction of Brazilian-Portuguese free-text data from the EMRs of patients. Ontological rules were also effective in this task, and perhaps even more importantly, most of the best-performing tasks with the OWL and ML models did not coincide. This outcome opens up the opportunity to exploit such complementarities to improve the coverage of tasks when implementing a real solution for outcome management or even to improve the individual effectiveness of each alternative by means of ensemble techniques such as stacking [82].

One of the good practices that the literature has demonstrated to increase the success of ML algorithms applied to health care is the inclusion of a clinical background in the annotation process [83]. The availability of training data is critical in obtaining good results, thus indicating that variations in clinical terms found in the clinical text could be specific to the type and source of clinical notes that may not have been captured in an available resource. The results from our feature importance analysis are consistent with other study results [21,68,76,83-85] concerning many clinical terms applied to multiple clinical concepts, although there are specific patterns based on semantic types that can help. In general, it is difficult to determine the correct concept when a clinical term normalizes to multiple concepts, and this issue can penalize the effectiveness of the model [86,87].

Our effectiveness results agree with the literature [83,88], in which a Macro-F1 score >80% is considered a successful extraction of medical records. Even though there is still a need to cover more tasks related to ICHOM patient-reported outcome measures [3,74,76,85], we hypothesized that these tasks comprise a feeling state, and the lack of normalization of data contained in EMRs may explain the fact that these task categories did not perform very well [70,89]. Medical records related to baseline characteristics and care processes typically contain much more structured data (eg, numerical values for tasks) than medical patient-reported outcomes, which focus more on unstructured data [83,90]. This issue has been explored in previous studies on EMR-based clinical quality measures [22,82], in which it is suggested that these kinds of data (for baseline characteristics and care-related processes) have the potential to be scaled in other clinical conditions, such as cardiovascular and endocrine conditions [83].

Previous studies have found various advantages of EMR compared with traditional paper records [91]. However, as reported by Ausserhofer et al [12], care workers do not find them useful for guaranteeing safe care and treatment because of the difficulty of tracking clinical and quality measures. The same authors have discussed the importance of having IT capability to track care workers' documentation while increasing safety and quality of care. They emphasized that this approach is important for addressing EMR data collection issues that have been historically extracted via manual review by clinical experts, leading to scalability and cost issues [83,85,90]. In our study, it was possible to demonstrate that for the stroke care pathway, the use of ML models to measure clinical outcomes remains a challenge, but the technology has the potential to support the extraction of relevant patient characteristics and care-process information.

Despite the challenges regarding the accuracy of the outcome measures, promising approaches regarding baseline characteristics and care-related process data have been achieved. This may be the first step toward unlocking the full potential of EMR data [83]. The usefulness of having baseline characteristics tracked is to assist disease prevalence studies and identify opportunities to guide political decisions about the public health sector [13,92,93], automatize eligibility of patients for clinical research [84], and feed risk assessment tools [94]. On the contrary, care-related process metrics boost the opportunity to improve decision-making with new technologies, maintain the effectiveness of treatments, and encourage alternative remuneration models [17,92,95].

The next step would be to invest in the automation of tasks at the patient level that support the control of the progression of patients in real-time during stroke episodes. In a similar manner, it would be useful to identify opportunities to improve the EMR data quality, such as the implementation of quality software with dynamic autocompletes with normalized terms register. The use of NLP for quality measures also adds to the capture of large amounts of clinical data from EMRs [82]. The products of NLP and mixed methods pipelines could potentially impact a number of clinical areas and could facilitate appropriate care

XSL•FO
RenderX

by feeding hospital outcome indicators and data to support epidemiological studies or value-based programs [82].

Limitations

This study had several limitations. For clinical NLP method development to advance further globally and to become an integral part of clinical outcome research or have a natural place in clinical practice, there are still challenges ahead. Our work is based on the EMR of a single center, with a limited number of annotated patients. Thus, further work is needed to test this approach in EMRs from different centers with different patients, who may use different languages for clinical documentation. We have no access to data from exams or hospital indicators, which is the reason why our infection identification, for example, was based on any report of antibiotic use, typical symptoms of infection, or tests described. We were unable to find data samples that included all the risk factors that were discovered in the literature. It would be worth conducting a future study with a larger and different data set with more features to examine whether the findings of this research are still valid. Finally, the design focused on sentences can be significantly influenced by the NI data volume—if a patient smokes, this will probably be reflected in just one sentence, maybe two, and for all of the others, you will have NI. One possible approach would be to use hierarchy models to first classify whether a sentence is relevant and then evolve to classification algorithms to predict classes. Then, the entire record can inform the prediction of the outcome of patients, instead of saying whether a specific sentence indicates a task.

Regarding the undersampling experiment, more intelligent strategies such as choosing the *most positive of the negative samples* or Tomek links [81] should be tested for better effectiveness. We leave this for future work and suggest practical purposes to maintain the original distribution, whereas more effective strategies are not further studied.

Conclusions

This study is innovative in that it considered many and diverse types of automatic classifiers (neural, nonneural, and ontological) using a large real-world data set containing thousands of textual sentences from real-world EMRs and a large number of tasks (n=24) with multiple classes using Brazilian-Portuguese unstructured free-text EMR databases. The effectiveness of these models demonstrated a better result when used to classify care processes and patient characteristics than patient-reported outcomes, which suggests that advances in intelligence in informational technology for clinical outcomes are still a gap in the scalability of outcome measurements in health care. Future research should explore the development of mixed methods to increase task effectiveness. Advances in IT capacity have proved to be essential for the scalability and agility of the ability to measure health outcomes and how it reflects on its external validation to support health real-time quality measurement indicators.

Conflicts of Interest

The authors disclose receipt of the following financial support for the research, authorship, and publication of this paper: this research was supported by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior–Brasil (CAPES)–Finance Code 001 and National Council for Scientific and Technological Development (CNPq 465518/2014-1 and others), Research Support Foundation of the State of Minas Gerais (FAPEMIG), Google, and NVIDIA Corporation.

Multimedia Appendix 1

Example of an evolution on the electronic medical record.
[DOCX File , 14 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Example of the annotation process.
[DOCX File , 19 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Data set characteristics.
[DOCX File , 20 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Details of the automatic text classification methods.
[DOCX File , 28 KB-Multimedia Appendix 4]

Multimedia Appendix 5

Experimental procedure.
[PNG File , 97 KB-Multimedia Appendix 5]

Multimedia Appendix 6

Experimental protocol details—specific parameter tuning.
[DOCX File , 15 KB-Multimedia Appendix 6]

Multimedia Appendix 7

Results of F1 score from the random undersampling experiment. BERT: bidirectional encoder representation from transformers; BoW: Bag-of-Words; KNN: K-nearest neighbor; mRS: Modified Rankin Score; NIHSS: National Institutes of Health Stroke Scale; SVM: support vector machine; TFIDF: term frequency-inverted document frequency; W+C: word- term frequency-inverted document frequency and character- term frequency-inverted document frequency.
[PNG File , 308 KB-Multimedia Appendix 7]

References

1. GBD 2016 Stroke Collaborators. Global, regional, and national burden of stroke, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet Neurol 2019 May;18(5):439-458 [FREE Full text] [doi: 10.1016/S1474-4422(19)30034-1] [Medline: 30871944]
2. Findings From the Global Burden of Disease Study 2017. Institute for Health Metrics and Evaluation (IHME). 2018. URL: http://www.healthdata.org/sites/default/files/files/policy_report/2019/GBD_2017_Booklet.pdf [accessed 2021-10-11]
3. Wang W, Kiik M, Peek N, Curcin V, Marshall I, Rudd A, et al. A systematic review of machine learning models for predicting outcomes of stroke with structured data. PLoS One 2020 Jun 12;15(6):e0234722 [FREE Full text] [doi: 10.1371/journal.pone.0234722] [Medline: 32530947]
4. Kamal H, Lopez V, Sheth SA. Machine learning in acute ischemic stroke neuroimaging. Front Neurol 2018 Nov 8;9:945 [FREE Full text] [doi: 10.3389/fneur.2018.00945] [Medline: 30467491]
5. Feng R, Badgeley M, Mocco J, Oermann EK. Deep learning guided stroke management: a review of clinical applications. J Neurointerv Surg 2018 Apr;10(4):358-362 [FREE Full text] [doi: 10.1136/neurintsurg-2017-013355] [Medline: 28954825]
6. Lee E, Kim Y, Kim N, Kang D. Deep into the brain: artificial intelligence in stroke imaging. J Stroke 2017 Sep;19(3):277-285 [FREE Full text] [doi: 10.5853/jos.2017.02054] [Medline: 29037014]
7. Wodchis WP, Austin PC, Henry DA. A 3-year study of high-cost users of health care. Can Med Asso J 2016 Feb 16;188(3):182-188 [FREE Full text] [doi: 10.1503/cmaj.150064] [Medline: 26755672]
8. Markatou M, Don PK, Hu J, Wang F, Sun J, Sorrentino R, et al. Case-based reasoning in comparative effectiveness research. IBM J Res Dev 2012 Sep;56(5):4:1-4:12. [doi: 10.1147/JRD.2012.2198311]
9. Wu J, Roy J, Stewart WF. Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches. Med Care 2010 Jun;48(6 Suppl):106-113. [doi: 10.1097/MLR.0b013e3181de9e17] [Medline: 20473190]
10. Chechulin Y, Nazerian A, Rais S, Malikov K. Predicting patients with high risk of becoming high-cost healthcare users in Ontario (Canada). Health Care Policy 2014 Feb 26;9(3):68-79. [doi: 10.12927/hcpol.2014.23710]
11. Rotmensch M, Halpern Y, Tlimat A, Horng S, Sontag D. Learning a health knowledge graph from electronic medical records. Sci Rep 2017 Jul 20;7(1):5994 [FREE Full text] [doi: 10.1038/s41598-017-05778-z] [Medline: 28729710]
12. Ausserhofer D, Favez L, Simon M, Zúñiga F. Electronic health record use in Swiss nursing homes and its association with implicit rationing of nursing care documentation: multicenter cross-sectional survey study. JMIR Med Inform 2021 Mar 02;9(3):e22974 [FREE Full text] [doi: 10.2196/22974] [Medline: 33650983]
13. Casey JA, Schwartz BS, Stewart WF, Adler NE. Using electronic health records for population health research: a review of methods and applications. Annu Rev Public Health 2016;37:61-81 [FREE Full text] [doi: 10.1146/annurev-publhealth-032315-021353] [Medline: 26667605]
14. Fernandes M, Sun H, Jain A, Alabsi HS, Brenner LN, Ye E, et al. Classification of the disposition of patients hospitalized with COVID-19: reading discharge summaries using natural language processing. JMIR Med Inform 2021 Mar 10;9(2):e25457 [FREE Full text] [doi: 10.2196/25457] [Medline: 33449908]
15. Porter M, Lee T. The strategy that will fix health care. Harvard Business Review. 2013. URL: https://hbr.org/2013/10/the-strategy-that-will-fix-health-care [accessed 2021-09-07]
16. Golas SB, Shibahara T, Agboola S, Otaki H, Sato J, Nakae T, et al. A machine learning model to predict the risk of 30-day readmissions in patients with heart failure: a retrospective analysis of electronic medical records data. BMC Med Inform Decis Mak 2018 Jun 22;18(1):44 [FREE Full text] [doi: 10.1186/s12911-018-0620-z] [Medline: 29929496]
17. Glaser J. It's time for a new kind of electronic health record. Harvard Bussiness Review. 2020. URL: https://hbr.org/2020/06/its-time-for-a-new-kind-of-electronic-health-record [accessed 2021-09-07]
18. Carberry K, Landman Z, Xie M, Feeley T, Henderson J, Fraser C. Incorporating longitudinal pediatric patient-centered outcome measurement into the clinical workflow using a commercial electronic health record: a step toward increasing value for the patient. J Am Med Inform Assoc 2016 Jan;23(1):88-93 [FREE Full text] [doi: 10.1093/jamia/ocv125] [Medline: 26377989]

XSL•FO
RenderX

19. Afzal N, Sohn S, Abram S, Liu H, Kullo IJ, Arruda-Olson AM. Identifying peripheral arterial disease cases using natural language processing of clinical notes. IEEE EMBS Int Conf Biomed Health Inform 2016 Feb;2016:126-131 [FREE Full text] [doi: 10.1109/BHI.2016.7455851] [Medline: 28111640]
20. Wi C, Sohn S, Rolfes MC, Seabright A, Ryu E, Voge G, et al. Application of a natural language processing algorithm to asthma ascertainment. An automated chart review. Am J Respir Crit Care Med 2017 Aug 15;196(4):430-437 [FREE Full text] [doi: 10.1164/rccm.201610-2006OC] [Medline: 28375665]
21. Chase HS, Mitrani LR, Lu GG, Fulgieri DJ. Early recognition of multiple sclerosis using natural language processing of the electronic health record. BMC Med Inform Decis Mak 2017 Feb 28;17(1):24 [FREE Full text] [doi: 10.1186/s12911-017-0418-4] [Medline: 28241760]
22. Garvin JH, Kim Y, Gobbel GT, Matheny ME, Redd A, Bray BE, et al. Automating quality measures for heart failure using natural language processing: a descriptive study in the department of veterans affairs. JMIR Med Inform 2018 Jan 15;6(1):e5 [FREE Full text] [doi: 10.2196/medinform.9150] [Medline: 29335238]
23. Dai H, Lee Y, Nekkanti C, Jonnagaddala J. Family history information extraction with neural attention and an enhanced relation-side scheme: algorithm development and validation. JMIR Med Inform 2020 Dec 01;8(12):e21750 [FREE Full text] [doi: 10.2196/21750] [Medline: 33258777]
24. Lee TH. Putting the value framework to work. N Engl J Med 2010 Dec 23;363(26):2481-2483. [doi: 10.1056/NEJMp1013111] [Medline: 21142527]
25. Blumenthal D, Tavenner M. The "meaningful use" regulation for electronic health records. N Engl J Med 2010 Aug 5;363(6):501-504. [doi: 10.1056/NEJMp1006114] [Medline: 20647183]
26. Porter ME, Larsson S, Lee TH. Standardizing patient outcomes measurement. N Engl J Med 2016 Feb 11;374(6):504-506. [doi: 10.1056/NEJMp1511701] [Medline: 26863351]
27. Wilson JL, Hareendran A, Grant M, Baird T, Schulz UG, Muir KW, et al. Improving the assessment of outcomes in stroke: use of a structured interview to assign grades on the modified Rankin Scale. Stroke 2002 Sep;33(9):2243-2246. [doi: 10.1161/01.str.0000027437.22450.bd] [Medline: 12215594]
28. Lyden PD, Lu M, Levine SR, Brott TG, Broderick J, NINDS rtPA Stroke Study Group. A modified National Institutes of Health Stroke Scale for use in stroke clinical trials: preliminary reliability and validity. Stroke 2001 Jun;32(6):1310-1317. [doi: 10.1161/01.str.32.6.1310] [Medline: 11387492]
29. Caso V, Zakaria M, Tomek A, Mikulik R, Martins S, Nguyen T, et al. Improving stroke care across the world: the ANGELS Initiative. CNS - Oruen Ltd. 2018. URL: https://www.oruen.com/wp-content/uploads/2018/12/Review-article-4.pdf [accessed 2021-09-07]
30. Honnibal M, Montani I. Industrial-strength natural language processing. spaCy. URL: https://spacy.io [accessed 2021-09-07]
31. Klie J, Bugert M, Boullosa B, de Castilho RE, Gurevych I. The INCEpTION platform: machine-assisted and knowledge-oriented interactive annotation. In: Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations. 2018 Aug 01 Presented at: 27th International Conference on Computational Linguistics: System Demonstrations; August, 2018; Santa Fe, New Mexico p. 5-9 URL: https://aclanthology.org/C18-2002/ [doi: 10.18653/v1/d18-2022]
32. Manning C, Raghawan P, Schutze H. Introduction to Information Retrieval. Cambridge: Cambridge University Press; 2008:1-506.
33. Manning C, Schutze H. Foundations of Statistical Natural Language Processing. Cambridge, MA: MIT Press; 1999:1-720.
34. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. Fam Med 2005 May;37(5):360-363 [FREE Full text] [Medline: 15883903]
35. Cunha W, Mangaravite V, Gomes C, Canuto S, Resende E, Nascimento C, et al. On the cost-effectiveness of neural and non-neural approaches and representations for text classification: a comprehensive comparative study. Inf Process Manag 2021 May;58(3):102481. [doi: 10.1016/j.ipm.2020.102481]
36. Canuto S, Salles T, Rosa TC, Couto T, Gonçalves MA. Similarity-based synthetic document representations for meta-feature generation in text classification. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2019 Jan 01 Presented at: SIGIR '19: The 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval; Jul 21-25, 2019; Paris France p. 355-364. [doi: 10.1145/3331184.3331239]
37. Canuto S, Salles T, Gonçalves M, Rocha L, Ramos G, Gonçalves G. On efficient meta-level features for effective text classification. In: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. 2014 Jan 01 Presented at: CIKM '14: 2014 ACM Conference on Information and Knowledge Management; Nov 3-7, 2014; Shanghai China p. 1709-1718. [doi: 10.1145/2661829.2662060]
38. Canuto S, Sousa DX, Goncalves MA, Rosa TC. A thorough evaluation of distance-based meta-features for automated text classification. IEEE Trans Knowl Data Eng 2018 Mar 27;30(12):2242-2256. [doi: 10.1109/tkde.2018.2820051]
39. Cunha W, Canuto S, Viegas F, Salles T, Gomes C, Mangaravite V, et al. Extended pre-processing pipeline for text classification: on the role of meta-feature representations, sparsification and selective sampling. Inf Process Manag 2020 Jul;57(4):102263. [doi: 10.1016/j.ipm.2020.102263]

XSL•FO
RenderX

40. Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, et al. Clinical information extraction applications: a literature review. J Biomed Inform 2018 Jan;77:34-49 [FREE Full text] [doi: 10.1016/j.jbi.2017.11.011] [Medline: 29162496]
41. Breiman L. Random forests. Mach Learn 2001 Oct 1;45(1):5-32. [doi: 10.1023/A:1010933404324]
42. Kowsari K, Meimandi KJ, Heidarysafa M, Mendu S, Barnes L, Brown D. Text classification algorithms: a survey. Information 2019 Apr 23;10(4):150. [doi: 10.3390/info10040150]
43. Larson RR. Introduction to information retrieval. J Am Soc Inf Sci Technol 2009 Oct 19;61(4):852-853. [doi: 10.1002/asi.21234]
44. Li L, Weinberg CR, Darden TA, Pedersen LG. Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method. Bioinformatics 2001 Dec;17(12):1131-1142. [doi: 10.1093/bioinformatics/17.12.1131] [Medline: 11751221]
45. Almeida MB, Bax MP. Uma visão geral sobre ontologias: pesquisa sobre definições, tipos, aplicações, métodos de avaliação e de construção. Ci Inf 2003 Dec;32(3):7-20. [doi: 10.1590/s0100-19652003000300002]
46. Allahyari M, Kochut K, Janik M. Ontology-based text classification into dynamically defined topics. In: Proceedings of the IEEE International Conference on Semantic Computing. 2014 Jan 01 Presented at: IEEE International Conference on Semantic Computing; Jun 16-18, 2014; Newport Beach, CA, USA p. 273-278. [doi: 10.1109/icsc.2014.51]
47. Chi N, Lin K, Hsieh S. Using ontology-based text classification to assist job hazard analysis. Adv Eng Inf 2014 Oct;28(4):381-394. [doi: 10.1016/j.aei.2014.05.001]
48. Garla VN, Brandt C. Ontology-guided feature engineering for clinical text classification. J Biomed Inform 2012 Oct;45(5):992-998 [FREE Full text] [doi: 10.1016/j.jbi.2012.04.010] [Medline: 22580178]
49. Wang B, McKay R, Abbass H, Barlow M. A comparative study for domain ontology guided feature extraction. Australian Computer Society. 2003. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.13.3384&rep=rep1&type=pdf [accessed 2021-09-07]
50. Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. Inf Process Manag 1988 Jan;24(5):513-523. [doi: 10.1016/0306-4573(88)90021-0]
51. Andrade CM, Gonçalves MA. Combining representations for effective citation classification. In: Proceedings of The International Workshop on Mining Scientific Publications. 2020 Presented at: The International Workshop on Mining Scientific Publications; Aug 2020; Wuhan, China.
52. Cortes EG, Woloszyn V, Barone DA. When, where, who, what or why? A hybrid model to question answering systems. In: Computational Processing of the Portuguese Language. Cham: Springer; 2018.
53. Viegas F, Rocha L, Resende E, Salles T, Martins W, Freitas MF, et al. Exploiting efficient and effective lazy Semi-Bayesian strategies for text classification. Neurocomput 2018 Sep 13;307:153-171. [doi: 10.1016/j.neucom.2018.04.033]
54. Fei Y. Simultaneous Support Vector selection and parameter optimization using Support Vector Machines for sentiment classification. In: Proceedings of the 2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS). 2016 Presented at: 2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS); Aug 26-28, 2016; Beijing, China. [doi: 10.1109/ICSESS.2016.7883015]
55. Shen Y. Selection incentives in a performance-based contracting system. Health Serv Res 2003 Apr;38(2):535-552 [FREE Full text] [doi: 10.1111/1475-6773.00132] [Medline: 12785560]
56. Georgakopoulos SV, Tasoulis SK, Vrahatis AG, Plagianakos VP. Convolutional neural networks for toxic comment classification. In: Proceedings of the 10th Hellenic Conference on Artificial Intelligence. 2018 Presented at: SETN '18: 10th Hellenic Conference on Artificial Intelligence; Jul 9-12, 2018; Patras Greece. [doi: 10.1145/3200947.3208069]
57. Johnson R, Zhang T. Deep pyramid convolutional neural networks for text categorization. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2017 Presented at: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Jul 30 - Aug 4, 2017; Vancouver, Canada. [doi: 10.18653/v1/P17-1052]
58. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the conference of the North American chapter of the association for computational linguistics: Human language technologies. 2019 Presented at: Proceedings of the conference of the North American chapter of the association for computational linguistics: Human language technologies; Jun,2019; Minneapolis, Minnesota.
59. Gomez-Perez A, Corcho O, Fernández-López M. Ontological Engineering With Examples from the Areas of Knowledge Management, E-Commerce and the Semantic Web. London: Springer; 2004.
60. Han EH, Karypis G. Centroid-based document classification: analysis and experimental results. In: Principles of Data Mining and Knowledge Discovery. Berlin, Heidelberg: Springer; 2000.
61. Manevitz LM, Yousef M. One-class svms for document classification. J Mach Learn Res 2002 Jan 3;2:139-154. [doi: 10.5555/944790.944808]
62. Layeghian Javan S, Sepehri MM, Aghajani H. Toward analyzing and synthesizing previous research in early prediction of cardiac arrest using machine learning based on a multi-layered integrative framework. J Biomed Inform 2018 Dec;88:70-89 [FREE Full text] [doi: 10.1016/j.jbi.2018.10.008] [Medline: 30389440]
63. Salles T, Gonçalves M, Rodrigues V, Rocha L. Improving random forests by neighborhood projection for effective text classification. Inf Syst 2018 Sep;77:1-21. [doi: 10.1016/j.is.2018.05.006]

XSL•FO
RenderX

64. Cawley G, Talbot N. On over-fitting in model selection and subsequent selection bias in performance evaluation. J Mach Learn Res 2010;11:2079-2107 [FREE Full text]
65. Velupillai S, Suominen H, Liakata M, Roberts A, Shah AD, Morley K, et al. Using clinical Natural Language Processing for health outcomes research: overview and actionable suggestions for future advances. J Biomed Inform 2018 Dec;88:11-19 [FREE Full text] [doi: 10.1016/j.jbi.2018.10.005] [Medline: 30368002]
66. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One 2015 Mar 4;10(3):e0118432 [FREE Full text] [doi: 10.1371/journal.pone.0118432] [Medline: 25738806]
67. Zar JH. Biostatistical Analysis, 5th Edition. London, UK: Pearson; 2010.
68. Reys AD, Silva D, Severo D, Pedro S, de Sousa e Sá MM, Salgado GA. Predicting multiple ICD-10 codes from Brazilian-Portuguese clinical notes. In: Cerri R, Prati RC, editors. Intelligent Systems. Cham: Springer; 2020.
69. Lee GH, Shin S. Federated learning on clinical benchmark data: performance assessment. J Med Internet Res 2020 Oct 26;22(10):e20891 [FREE Full text] [doi: 10.2196/20891] [Medline: 33104011]
70. Kate RJ. Clinical term normalization using learned edit patterns and subconcept matching: system development and evaluation. JMIR Med Inform 2021 Jan 14;9(1):e23104 [FREE Full text] [doi: 10.2196/23104] [Medline: 33443483]
71. Lee DH, Yetisgen M, Vanderwende L, Horvitz E. Predicting severe clinical events by learning about life-saving actions and outcomes using distant supervision. J Biomed Inform 2020 Jul;107:103425. [doi: 10.1016/j.jbi.2020.103425] [Medline: 32348850]
72. Sheikhpour R, Sarram MA, Gharaghani S, Chahooki MA. A survey on semi-supervised feature selection methods. Pattern Recognit 2017 Apr;64:141-158. [doi: 10.1016/j.patcog.2016.11.003]
73. Diao X, Huo Y, Yan Z, Wang H, Yuan J, Wang Y, et al. An application of machine learning to etiological diagnosis of secondary hypertension: retrospective study using electronic medical records. JMIR Med Inform 2021 Jan 25;9(1):e19739 [FREE Full text] [doi: 10.2196/19739] [Medline: 33492233]
74. Zhang Y, Zhou Y, Zhang D, Song W. A stroke risk detection: improving hybrid feature selection method. J Med Internet Res 2019 Apr 02;21(4):e12437 [FREE Full text] [doi: 10.2196/12437] [Medline: 30938684]
75. Guillaume LF, Christos K, Aridas CK. Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning. J Mach Learn Res 2017 Jan;18(1):559-563. [doi: 10.5555/3122009.3122026]
76. Kogan E, Twyman K, Heap J, Milentijevic D, Lin JH, Alberts M. Assessing stroke severity using electronic health record data: a machine learning approach. BMC Med Inform Decis Mak 2020 Jan 08;20(1):8 [FREE Full text] [doi: 10.1186/s12911-019-1010-x] [Medline: 31914991]
77. Healthcare Improvement - Patient-Reported Outcomes. ICHOM. URL: https://www.ichom.org/ [accessed 2021-09-07]
78. Freeman D, Barret K, Nordan L, Spaulding A, Kaplan R, Karney M. Lessons from Mayo clinic’s redesign of stroke care. Harvard Business Review. 2018. URL: https://hbr.org/2018/10/lessons-from-mayo-clinics-redesign-of-stroke-care [accessed 2021-09-07]
79. Feigin VL, Krishnamurthi R. Stroke is largely preventable across the globe: where to next? Lancet 2016 Aug 20;388(10046):733-734. [doi: 10.1016/S0140-6736(16)30679-1] [Medline: 27431357]
80. Zhou P, El-Gohary N. Ontology-based multilabel text classification of construction regulatory documents. J Comput Civ Eng 2015 Sep;30(4):04015058. [doi: 10.1061/(asce)cp.1943-5487.0000530]
81. Chawla NV. Data mining for imbalanced datasets: an overview. In: Maimon O, Rokach L, editors. Data Mining and Knowledge Discovery Handbook. US: Springer; 2010.
82. Weiskopf NG, Khan FJ, Woodcock D, Dorr DA, Cigarroa JE, Cohen AM. A mixed methods task analysis of the implementation and validation of EHR-based clinical quality measures. AMIA Annu Symp Proc 2017 Feb 10;2016:1229-1237 [FREE Full text] [Medline: 28269920]
83. Sheikhalishahi S, Miotto R, Dudley JT, Lavelli A, Rinaldi F, Osmani V. Natural language processing of clinical notes on chronic diseases: systematic review. JMIR Med Inform 2019 Apr 27;7(2):e12239 [FREE Full text] [doi: 10.2196/12239] [Medline: 31066697]
84. Ling A, Kurian A, Caswell-Jin J, Sledge G, Shah N, Tamang S. Using natural language processing to construct a metastatic breast cancer cohort from linked cancer registry and electronic medical records data. JAMIA Open 2019 Sep 18;2(4):528-537 [FREE Full text] [doi: 10.1093/jamiaopen/ooz040] [Medline: 32025650]
85. Wang SV, Rogers JR, Jin Y, Bates DW, Fischer MA. Use of electronic healthcare records to identify complex patients with atrial fibrillation for targeted intervention. J Am Med Inform Assoc 2017 Mar 01;24(2):339-344. [doi: 10.1093/jamia/ocw082] [Medline: 27375290]
86. Ali A, Shamsuddin S, Ralescu A. Classification with class imbalance problem: a review. Int J Advance Soft Compu Appl 2013 Nov;5(3):176-204 [FREE Full text]
87. Li D, Liu C, Hu SC. A learning method for the class imbalance problem with medical data sets. Comput Biol Med 2010 May;40(5):509-518. [doi: 10.1016/j.compbiomed.2010.03.005] [Medline: 20347072]
88. Geng W, Qin X, Yang T, Cong Z, Wang Z, Kong Q, et al. Model-based reasoning of clinical diagnosis in integrative medicine: real-world methodological study of electronic medical records and natural language processing methods. JMIR Med Inform 2020 Dec 21;8(12):e23082 [FREE Full text] [doi: 10.2196/23082] [Medline: 33346740]

XSL•FO
RenderX

89. Ridgway JP, Uvin A, Schmitt J, Oliwa T, Almirol E, Devlin S, et al. Natural language processing of clinical notes to identify mental illness and substance use among people living with HIV: retrospective cohort study. JMIR Med Inform 2021 Mar 10;9(3):e23456 [FREE Full text] [doi: 10.2196/23456] [Medline: 33688848]
90. Liao KP, Ananthakrishnan AN, Kumar V, Xia Z, Cagan A, Gainer VS, et al. Methods to develop an electronic medical record phenotype algorithm to compare the risk of coronary artery disease across 3 chronic disease cohorts. PLoS One 2015 Aug 24;10(8):e0136651 [FREE Full text] [doi: 10.1371/journal.pone.0136651] [Medline: 26301417]
91. Kruse CS, Mileski M, Alaytsev V, Carol E, Williams A. Adoption factors associated with electronic health record among long-term care facilities: a systematic review. BMJ Open 2015 Jan 28;5(1):e006615 [FREE Full text] [doi: 10.1136/bmjopen-2014-006615] [Medline: 25631311]
92. Beam AL, Kohane IS. Big data and machine learning in health care. JAMA 2018 Apr 03;319(13):1317-1318. [doi: 10.1001/jama.2017.18391] [Medline: 29532063]
93. Bugnon B, Geissbuhler A, Bischoff T, Bonnabry P, von Plessen C. Improving primary care medication processes by using shared electronic medication plans in Switzerland: lessons learned from a participatory action research study. JMIR Form Res 2021 Jan 07;5(1):e22319 [FREE Full text] [doi: 10.2196/22319] [Medline: 33410753]
94. Nakatani H, Nakao M, Uchiyama H, Toyoshiba H, Ochiai C. Predicting inpatient falls using natural language processing of nursing records obtained from Japanese electronic medical records: case-control study. JMIR Med Inform 2020 Apr 22;8(4):e16970 [FREE Full text] [doi: 10.2196/16970] [Medline: 32319959]
95. Dafny L, Lee T. Health care needs real competition. Harvard Business Review (Competitive Strategy). 2016. URL: https://hbr.org/2016/12/health-care-needs-real-competition [accessed 2021-09-07]

Abbreviations

BERT: bidirectional encoder representation from transformers
BoW: Bag-of-Words
CNN: convolutional neural network
EMR: electronic medical record
IT: information technology
KNN: K-nearest neighbor
ML: machine learning
NIHSS: National Institutes of Health Stroke Scale
NLP: natural language processing
OWL: ontology web language
SVM: support vector machine
TFIDF: term frequency-inverted document frequency

Edited by T Hao; submitted 29.03.21; peer-reviewed by M Fernandes, X Diao; comments to author 30.05.21; revised version received 27.06.21; accepted 05.08.21; published 01.11.21

Please cite as:
Zanotto BS, Beck da Silva Etges AP, dal Bosco A, Cortes EG, Ruschel R, De Souza AC, Andrade CMV, Viegas F, Canuto S, Luiz W, Ouriques Martins S, Vieira R, Polanczyk C, André Gonçalves M
Stroke Outcome Measurements From Electronic Medical Records: Cross-sectional Study on the Effectiveness of Neural and Nonneural Classifiers
JMIR Med Inform 2021;9(11):e29120
URL: https://medinform.jmir.org/2021/11/e29120
doi: 10.2196/29120
PMID:

©Bruna Stella Zanotto, Ana Paula Beck da Silva Etges, Avner dal Bosco, Eduardo Gabriel Cortes, Renata Ruschel, Ana Claudia De Souza, Claudio M V Andrade, Felipe Viegas, Sergio Canuto, Washington Luiz, Sheila Ouriques Martins, Renata Vieira, Carisi Polanczyk, Marcos André Gonçalves. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 01.11.2021. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX