Medical Care Research and Review
2021, Vol. 78(2) 138–145

DOI: 10.1177/1077558719844123
journals.sagepub.com/home/mcr
SAGE

Analysis of Unstructured Text-Based Data Using Machine Learning Techniques: The Case of Pediatric Emergency Department Records in Nicaragua

Giulia Lorenzoni[1], Silvia Bressan[2], Corrado Lanera[1], Danila Azzolina[1], Liviana Da Dalt[2], and Dario Gregori[1]

Abstract

Free-text information is still widely used in emergency department (ED) records. Machine learning techniques are useful for analyzing narratives, but they have been used mostly for English-language data sets. Considering such a framework, the performance of an ML classification task of a Spanish-language ED visits database was tested. ED visits collected in the EDs of nine hospitals in Nicaragua were analyzed. Spanish-language, free-text discharge diagnoses were considered in the analysis. Five-hundred random forests were trained on a set of bootstrap samples of the whole data set (1,789 ED visits) to perform the classification task. For each one, after having identified optimal parameter value, the final validated model was trained on the whole bootstrapped data set and tested. The classification accuracies had a median of 0.783 (95% CI [0.779, 0.796]). Machine learning techniques seemed to be a promising opportunity for the exploitation of unstructured information reported in ED records in low- and middle-income Spanish-speaking countries.

Keywords

emergency department visits, low- and middle-income countries, free-text discharge diagnosis, Spanish, random forest, classification task

Introduction

Monitoring emergency department (ED) visits represents a powerful tool for public health surveillance (Hirshon et al., 2009). It allows for the analysis of frequency (e.g., time trends, seasonality) and distribution of diseases and injuries referred to ED, the early detection of outbreaks (through syndromic surveillance) (Heffernan, 2004; Henning, 2004) which is currently employed in a growing number of application fields other than the ones for which it has been initially developed, that is, the early detection of bioterrorism attack (Lall et al., 2017), the quality assessment of health services, and, not least, the evaluation of the effectiveness of intervention programs.

The availability of computerized and coded patients' information (e.g., signs, symptoms, admission diagnosis) is crucial for the successful monitoring of ED visits with the purpose of epidemiological surveillance. In view of making ED information readily accessible, since the beginning of the 2000s, several signs of progress have been made in the computerization and coding of ED health records, especially in high-income countries (e.g., in the United States; Geisler, Schuur, & Pallin, 2010). However, using information on ED visits for epidemiological research is still challenging (Hirshon et al., 2009). The main barrier is represented by the employment of heterogeneous data collection systems, regarding methods of data collection, type of data collected, data structure, data format, lack of consistency and underuse of coding systems of diseases and injuries, and the widespread use of narrative free text. Particularly, the documentation of ED visits using unstructured free text is still widely used, since several coding systems are available and are

This article, submitted to *Medical Care Research and Review* on June 4, 2018, was revised and accepted for publication on March 18, 2019.

[1]Unit of Biostatistics, Epidemiology and Public Health, Department of Cardiac, Thoracic, Vascular Sciences and Public Health, University of Padova, Padova, Italy
[2]Division of Pediatric Emergency Medicine, Department of Women's and Children's Health, University of Padova, Padova, Italy

Corresponding Author:
Dario Gregori, Unit of Biostatistics, Epidemiology and Public Health, Department of Cardiac, Thoracic, Vascular Sciences and Public Health, University of Padova, Via Loredan, 18, Padova 35131, Italy.
Email: dario.gregori@unipd.it

continually being developed, but their use is not straightforward (Biese et al., 2013).

Such barriers in the analysis of ED data sets for epidemiological research are even more relevant for low- and middle-income countries (LMICs), where the care of acute conditions is not as well established as in high-income countries (Obermeyer et al., 2015). Fortunately, in recent years, several initiatives have been put forward to improve the performance of EDs in LMICs, and especially in Latin American ones (Crouse et al., 2016; Taira et al., 2016). However, the wide use of free-text information instead of coded and computerized data collection systems makes the analysis of ED visits epidemiology difficult. These data are useful to monitor ED performance and to target ad hoc interventions to develop emergency care systems in such countries further (Johnson, Gaus, & Herrera, 2016).

Conceptual Framework

Given such a framework, besides a progressive development of a standardized data collection system for ED visits, in both high-income countries and LMICs, it is crucial to adopt approaches of analysis allowing for the exploitation of unstructured, text-based, ED medical records currently available. Data extraction from free-text ED health records might be done through a manual, in-deep, review of individual medical records; however, such a strategy is extremely expensive and time-consuming (Biese et al., 2013). Conversely, the automatic coding of free-text information reported in ED health records through appropriate machine learning techniques (MLTs) would be a promising opportunity (Ford, Carroll, Smith, Scott, & Cassell, 2016), which is increasingly used also for the analysis of ED records, with encouraging results (Gerbier et al., 2011; Metzger et al., 2017). However, the research on the use of MLTs to automatically extract information from medical records is still at an early stage, and it is applied mainly to the English-based data sets. Only a few examples are available in the literature about the application of MLT to the Spanish language (Castillo, 2010; Cotik, Filippo, & Castaño, 2014; Pérez et al., 2017; Tanev et al., 2009), which is one of the most widespread languages worldwide. In addition to that, it is well known that different languages show different levels of linguistic, morphological, and syntactical complexities (Ehret & Szmrecsanyi, 2016) (e.g., Spanish exhibits slightly higher levels of morphological complexity compared with English; Bentz, Ruzsics, Koplenig, & Samardzic, 2016). This inevitably influences how medical information is reported in ED health records and, consequently, the accuracy of automatic classification algorithms. This highlights the need for testing MLTs algorithms on different languages other than the English ones.

New Contributions

Considering the usefulness of ED data for monitoring population's health care needs, but the wide heterogeneity of data collection systems employed in the EDs and, not least, the wide use of free-text information instead of coded ones, it is crucial to develop analysis approaches that are able to exploit the ED data available for deriving useful information to monitor population's health. MLTs would be a promising approach of analysis of free-text medical information, but their use is still limited, and most of the studies have been done on English language data sets. Considering such a framework, the performance of a machine learning classification task of Spanish free-text discharge diagnoses reported in an ED visits database from Nicaragua was tested.

Method

Italy–Nicaragua Cooperation Project

Data were derived from an international cooperation project between Italian and Nicaraguan pediatricians aimed at setting up a pediatric emergency clinical network in Nicaragua. The project started in 2011 and was carried out thanks to the partnership between the Regione Lombardia, the IRCCS Fondazione Ca' Grande—Policlinico Milano, the Department of Women's and Children's Health—University of Padova, the Nicaraguan government, and La Mascota Hospital in Managua.

Nine Nicaraguan hospitals were included in the project: one referral center, La Mascota Hospital located in Managua, the capital city of Nicaragua, and eight referring hospitals located in the towns of Chinandega, Granada, Juigalpa, Jinotega, Matagalpa, Masaya, Bluefields, and Puerto Cabeza. Clinical resources and pediatrician coverage greatly varied between hospitals making pediatric emergency care of acutely ill or injured patients challenging.

Data Source

An electronic data collection system was developed, using FileMaker Pro 11.0v3 (Santa Clara, CA), as part of the international cooperation project to monitor the clinical outcomes of patients presenting to the ED with urgent or emergent clinical conditions based on the inclusion criteria available as Supplemental Material (Table S1, available online). All the ED visits entered in the data collection system, according to the inclusion criteria, were used in the analysis. Such a system, initially developed with the goal to use it as a base for telemedicine communication with the referral hospital, worked within an intranet system between the referring hospitals and the referral center.

Data available in the system were represented by children's demographic characteristics (age and gender) and clinical history, vital signs (body temperature, blood pressure, heart and breathing rates, and oxygen saturation), results of laboratory tests, diagnostic and therapeutic interventions (if performed), discharge diagnosis, outcomes of the ED visit (hospitalization, transfer to another hospital, death, discharge from ED). Most information was reported in Spanish narrative free text.

Table 1. Variables Included in the Data Set With the Corresponding Type and Examples.

Variables	Data type	Examples
Age (in years)	Numerical	6, 13
Gender	Categorical	Male, female
Vital signs (i.e., body temperature, blood pressure, heart rate, breathing rate, oxygen saturation)	Numerical	Body temperature (°C; 37.6, 39, 36.7) Blood pressure (BP; mmHg; Systolic BP: 91, 79, 107; Diastolic BP: 65, 79, 50) Heart rate (bpm; 145, 138, 149) Breathing rate (bpm; 22, 42, 70) Oxygen saturation (%; 97, 100, 78)
Laboratory tests (i.e., white blood cells count, creatinite, glucose, natremia, urea)	Numerical	
Diagnostic and therapeutic interventions (i.e., radiological examinations, respiratory support, medications administered, vascular access)	Free text	Radiological examinations: "rx torax: infiltrado basal derecha"; "eco cardiograma: hap severa, falla cardiaca aguda, derrame pericardio moderado."; "rx de abdomen. radiopoacidad en fid."
Outcome of emergency department visit	Categorical	Ingresado, fallecido
Discharge diagnosis	Free text	"dengue hemorrágico" "crisis convulsiva febril"
Manual classification of the discharge diagnosis (gold standard)	Categorical	Gastrointestinal, cardiovascular, neurological

For the study, we focused on ED visits reported in the data collection system in 2012 for which discharge diagnosis was available. The full data set (ED visits collected in 2012) was represented by 2,723 ED visits, and those for which discharge diagnosis was available were 1,789 (66%).

Discharge Diagnosis Classification: The Gold Standard

The free-text discharge diagnoses were manually revised and classified by an independent peer-review group of expert pediatricians. The classification comprised 10 different classes, including diseases of the cardiovascular, gastrointestinal, metabolic, neurological, respiratory systems, tropical diseases, injuries, poisonings, burns, and others. Such classification was considered as the gold standard. Table 1 reports the variables available in the data set after the manual classification. The variable reporting the final discharge diagnosis (i.e., discharge diagnosis) was the basis to create the set of tokens used as predictors. The variable reporting the manual classification (i.e., manual classification, which represents the gold standard) was used as the target variable in the classification procedure.

Data Import, Preprocessing, and Management

Original data were available in Excel file format. For the analysis using MLT, they were converted in CSV using the UTF-8 character's encoding. Data preprocessing (Denny & Spirling, 2018) consisted in the transformation of all characters in lower case letters, the removal of all nonalphabetical characters and extra white spaces, and the transformation of each word to its corresponding *lemmata* (i.e., term reported in the dictionary). Every single word and every consecutive sequence of two words (bigrams) were considered as *tokens*.

A document-term matrix (DTM) was then built up. Each column in a DTM corresponds to a *token* and each row to a discharge diagnosis. The Term Frequency–inverse Document Frequency (TF-iDF; Salton & Buckley, 1988) in each cell of the DTM was reported. The TF-iDF consists in the product between the TF (number of times that a token was reported in a free-text diagnosis record) and the inverse of the logarithm of DF (number of free-text diagnosis records in which a token appeared), thus providing information on the frequency of a token in the diagnoses. The most important tokens (including bigrams) are reported in Table S2 of the Supplemental Material.

Data Analysis and MLT Training

To obtain a fair estimation of the performance ranges, the strategy adopted for the analyses was to repeat the whole training procedure on 500 bootstrap resamples of the data set. Each training procedure involved the fitting of a set of random forests (RFs) MLT (Breiman, 2001; Liaw & Wiener, 2002). The classification task was to classify the manual-identified diagnoses' classes (i.e., the gold standard) using only the text of discharge diagnoses. Each RF was trained considering a forest with 500 trees. The number was set large enough to reach the stability of the votes in the classification model (Figure 1). For each RF, the optimal number of variables (*tokens*) to be sampled and selected for the training procedure, namely *mtry* parameter, was established independently for each one. The *mtry* selection strategy was to perform five repetitions of a 10-fold cross-validation procedure (Kim, 2009). This was the optimal *mtry* selected to guarantee the optimal trade-off between bias and variance of the models estimated. As a set of options for the *mtry* search, the procedure considered a pseudo-exponential sequence of possible values (i.e., 3; 10; 30; 100; 300; 1,000 up to the maximum number of variables—*tokens*—available).

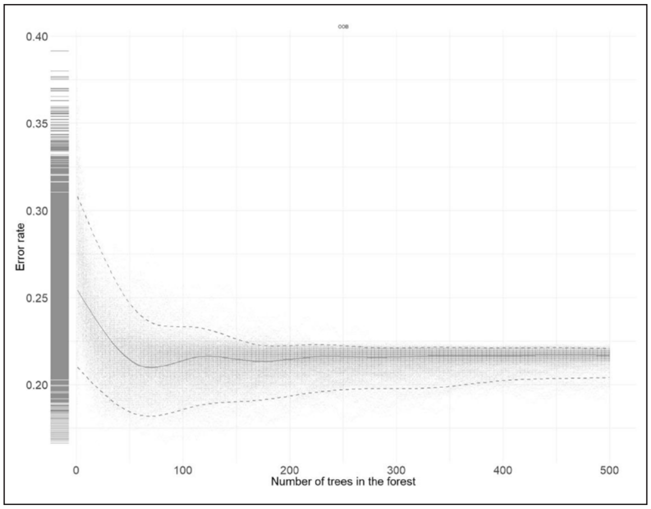

Figure 1. Out-of-bag error of the final validated models (calculated considering the out-of-bag performance of 500 bootstrap repetitions of evaluation a 500-tree random forest classifier by 5 repetition of 10-fold cross-validation procedure).
Note. Dashed lines represent the performance corresponding to the 95% confidence interval borders for the bootstrapped classifiers, the solid line represents the median one, and each semitransparent dot corresponds to the performance of a single random forest into the pool created by the bootstrapped procedure.

Once the optimal *mtry* was chosen (through the five repetitions of the 10-fold cross-validation procedure), a final validated model (i.e., a new RF made up of new 500 trees) was trained on the whole bootstrapped data set (1,789 bootstrapped ED visits), and tested on its out-of-bag (OOB) set, that is, the observation initially excluded by the bootstrap selection and hence never seen by the whole training procedure. The strategy is reported in Figure 2.

Statistical Analysis and Estimation of MLT Performance

Descriptive statistics were reported as median (first and third quartiles) for continuous variables, and percentages (absolute numbers) for categorical variables. Thanks to the bootstrap procedure adopted, the classification task could evaluated by the out-of-bag (OOB) classification performance of the final trained RF (James, Witten, Hastie, & Tibshirani, 2013) for each one of the 500 bootstrapped RFs. The quality of the classification task was assessed by computing the accuracy (rate of discharge diagnosis correctly classified, according to the gold standard, by the algorithm) overall and stratified by each class of discharge diagnosis. The set of accuracies of the 500 bootstrapped RFs was computed and reported with their median and the corresponding 95% confidence interval.

Software

R software Version 3.4.2 (R Core Team, 2017) was used for the analyses, within the packages rms (Harrell, 2014) for the statistical analyses, tidyverse (Wickham, 2017b) for the data management, lubridate (Grolemund & Wickham, 2011) for the date–time data management. Packages stringr (Wickham, 2017a) and glue (Hester, 2017) were used for the text management, while tm (Feinerer & Hornik, 2017), randomForest (Liaw & Wiener, 2002) and caret (Wing et al., 2017) were employed for text analyses and machine learning interface. All the analyses run on a Windows 10 Enterprise desktop computer powered by an Intel® quad Core™ i7-6700 CPU at 3.4 GHz with ×64-based operating system and processor, equipped with 40 GB of RAM. The scripts were implemented to train the trees of the RFs in parallel on 3 (i.e., $n - 1$) cores.

Results

A total of 1,789 pediatric ED records reported in 2012 in the data collection system set up in the context of the *Italy–Nicaragua Cooperation Project* were considered in the analysis. Most of the children admitted to ED were young children (median age of 2 years) of male gender (56%). According to the gold standard (manual classification), the discharge diagnoses class that was most represented was related to the respiratory system (mainly pneumonia), followed by that of the gastrointestinal tract (diarrhea) (Table 2). The male gender was the most prevalent in all the discharge diagnoses classes except for the metabolic and the poisoning ones. Children admitted to ED with diagnoses about the metabolic system and affected by tropical diseases were the oldest (median age of 13 and 9 years, respectively).

Machine Learning Classification Task Performance

Overall, 3,891 distinct tokens were considered in the analyses, in particular, they range from 256 distinct tokens for Hospital Juigalpa to 1,552 distinct tokens for Hospital La Mascota, with a median of 461 tokens. The overall CPU time (on Intel® quad Core™ i7-6700 CPU at 3.4 GHz with ×64-based operating system and processor, equipped with 40 GB of RAM) to train all the models was of 3968.68 seconds, ranging from 35.33 seconds for Hospital Puerto Cabezas to 3090.29 seconds for Hospital La Mascota, and a median CPU time of 95.56 seconds.

Looking at the classification task, it showed an accuracy of 0.7831 (95% CI [0.7792, 0.7965]) on the data set overall (Table 3). The analysis of the accuracy of the RF according to discharge diagnoses' classes generally showed good performance. Figure 1 shows the trend of the OOB error from 1 to 500 trees considered for each of the validated bootstrap RF models, showing very good performance of the machine learning algorithm, with a very low and stable error rate at 500 trees.

The analysis of the RF performance according to the sample characteristics (age and gender) showed a good performance for age (Figure 3). Conversely, the accuracy of the

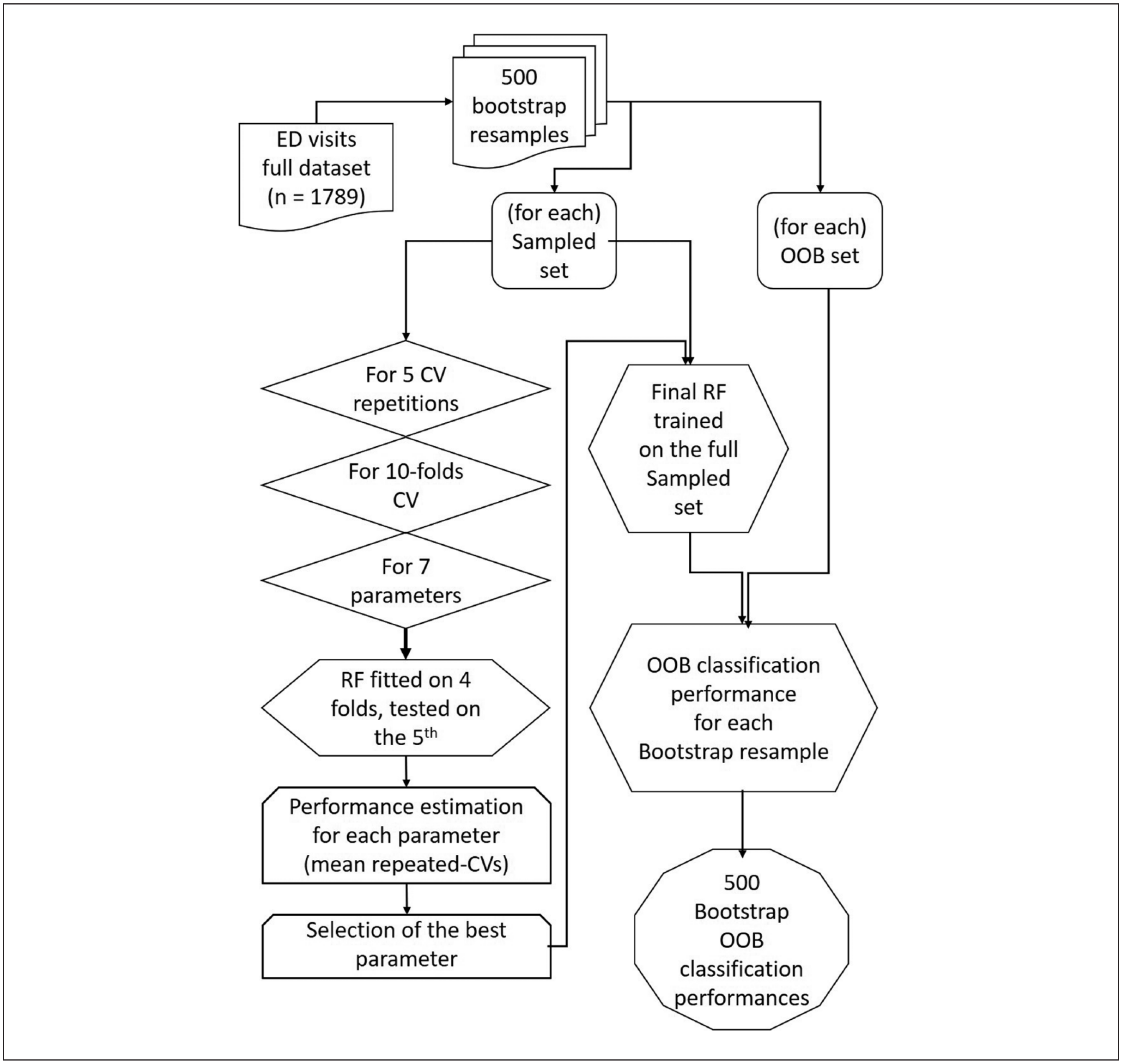

Figure 2. Training procedure.
Note. ED = emergency department; CV = cross-validation; RF = random forest; OOB = out-of-bag. For each of the 500 bootstrap resampled data set the performance estimation was calculated on its OOB set, which was never seen by the training procedure and was different for every sample. For the final model trained on each bootstrap sample, the optimal parameter was selected by 5 repetition of 10-fold CV estimation.

models was better ($p < .001$) for male gender (0.788 95% CI [0.783, 0.785]) compared with the female ones (0.777 95% CI [0.772, 0.773]).

Discussion

The present study aimed at assessing the performance of RF-based classification strategy in the automatic classification of free-text discharge diagnoses reported in pediatric ED records from the country of Nicaragua.

Nicaragua is one of the poorest countries in the Western world. In recent years, several efforts have been put forward to try to improve the Nicaraguan health care system, although hampered by a lack of resources. From the epidemiological point of view, Nicaragua is still considered a pretransitional country, characterized by a high prevalence of infectious

Table 2. Children's Characteristics According to Diagnosis Category.

	Gold standard	Age (years)	Gender, male
Burn	1 (20)	4.0 [4.0; 7.5]	70 (14)
Cardiovascular	5 (98)	1.5 [1.0; 9.0]	57 (55)
Gastrointestinal	12 (208)	2.0 [1.0; 7.0]	52 (107)
Injury	4 (80)	8.0 [5.0; 11.0]	68 (54)
Metabolic	2 (29)	13.0 [10.0; 15.0]	31 (9)
Neurological	8 (141)	4.0 [2.0; 9.0]	59 (82)
Poisoning	1 (13)	4.0 [3.0; 7.0]	46 (6)
Respiratory	56 (1,003)	1.0 [1.0; 3.0]	56 (560)
Tropical disease	6 (104)	9.0 [6.0; 11.0]	51 (53)
Other	5 (93)	4.0 [2.0; 9.0]	57 (52)
Overall	100 (1,789)	2.0 [1.0; 6.0]	56 (992)

Note. Data are expressed as medians [first quartile; third quartile] for continuous data and percentages (absolute number) for categorical ones.

Table 3. Median Accuracy (Rate of Diagnosis Correctly Classified by the Final Validated Model) of the Machine Learning Algorithms Together With 95% Confidence Interval (CI).

	Accuracy [95% CI]	
Burn	0.900	—
Cardiovascular	0.683	[0.663, 0.704]
Gastrointestinal	0.759	[0.745, 0.769]
Injury	0.837	[0.825, 0.850]
Metabolic	0.758	—
Neurological	0.602	[0.588, 0.624]
Poisoning	0.692	—
Respiratory	0.801	[0.797, 0.826]
Tropical disease	0.971	[0.971, 0.980]
Other	0.752	[0.731, 0.763]
Overall	0.783	[0.779, 0.796]

Note. Median accuracy was calculated considering the out-of-bag performance of 500 bootstrap repetitions of evaluation a 500-tree random forest classifier by 5 repetition of 10-fold cross-validation procedure. The CI was not estimated for burn, metabolic, and poisoning discharge diagnosis classes because of the small size of the sample of children in these classes.

diseases and adverse maternal and neonatal outcomes (Sequeira et al., 2011). This is consistent with the present analysis since most of the children were admitted to the ED with respiratory and gastrointestinal diseases (mainly respiratory infections and diarrhea).

The analysis of RFs accuracy according to sample characteristics showed that the performance of the classification algorithm was stable over children's age, even though the age group most represented was that of young children. Conversely, the RFs performance varied according to gender. The accuracy of the classification task was better for boys compared with girls. One potential explanation of such finding could be represented by the fact that the algorithm was unsuitable to classify discharge diagnoses in female children. However, this seems very unlikely, given the good performance of the classification algorithm for the overall sample. The lower accuracy in reporting the diagnoses for female children compared with males is more likely to explain our finding. However, there are no available data to support either hypothesis.

Overall, the algorithm's performance was found to be very good, providing new insights about the application of such techniques to ED data. MLTs have been increasingly used in the field of emergency medicine, as it has been shown by a recent literature review (Liu, Zhang, Wah Ho, & Hock Ong, 2018). It is worth pointing out that the ED visits included in the analyses were the most severe ones corresponding to 1% to 2% of all the ED visits. This is even more relevant from the public health perspective since the most severe ED visits are those that require the most careful monitor and the most complex clinical management since they are related to higher morbidity and mortality compared with the less severe ones. For this reason, an accurate classification of such ED visits is essential to allow for careful planning of the ED activities and resources, especially in LMIC where the care of acute conditions is not as well established as in high-income countries.

The main applications of such techniques to emergency medicine data are the development of predictive risk models, the patients' monitoring, and the integration of such techniques with EDs activities (e.g., in the triage; Liu et al., 2018). Present findings further improve our knowledge about the potentials of the application of MLTs to emergency medicine data. Such an algorithm would be a promising tool to automatically classify information from ED health records for the Nicaraguan government since the only requirement for MLTs use is that the ED records are extractable. This means that the application of the algorithm to free-text information might improve (a) the epidemiological surveillance of ED visits (e.g., seasonality, identification of infectious diseases outbreaks) to allow for a better plan of ED activities and resources' allocation, (b) the identification of pediatric population health care needs, (c) the monitor of the performance of the EDs, and (d) the evaluation of the effectiveness of public health interventions.

Limitations

The main limitations were represented by the fact that the MLTs were applied to a small (1,789 ED records) data set in the Spanish language, which has been only rarely analyzed using MLTs. The fact that the data set was small represents the main reason why the actual discharge diagnosis categories were broader than those identified by the manual classification (gold standard) and, as a consequence, some discharge diagnosis categories were underrepresented. However, the performance of the machine learning algorithm in classifying the discharge diagnoses was very good, both overall and by discharge diagnoses' groups. This in line with

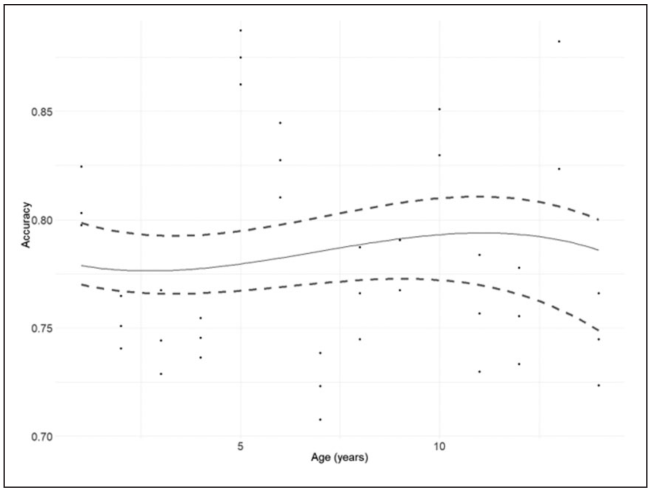

Figure 3. Accuracy according to children's age.
Note. Dashed lines represent 95% confidence interval (CI; calculated considering 500 bootstrap repetitions), solid line represents the median.

the very few studies available from international literature about the application of MLTs to the Spanish language, suggesting a good performance of MLT also in this linguistic context (Castillo, 2010; Cotik et al., 2014; Pérez et al., 2017; Tanev et al., 2009). Looking specifically at the studies on the analysis of free-text ED records using MLT, our results are in line with those of previous studies, showing good performance of RF (Metzger et al., 2017) and the usefulness of analyzing free-text information to enhance information from medical records (Worster et al., 2005).

Conclusions

Results of the present study showed a good performance of a machine learning approach for the automatic classification of ED free-text discharge diagnoses in the Spanish language, providing insights for the use of MLT for the exploitation of unstructured information reported in ED records for epidemiologic surveillance in LMICs Spanish-speaking countries and communities. Clearly, further work should be done in testing the algorithm on wider pediatric ED data sets allowing for a more detailed classification, through a strict collaboration between physicians, epidemiologists, and big data specialists.

Authors' Note

The work has been presented at the American Public Health Association Annual Meeting (Session—Oral 3381.0: Evaluating Emergency Systems: Using Various Types of Data Techniques, Monday, November 6, 2017, 2:30 p.m. to 4:00 p.m.)

Acknowledgments

The authors acknowledge Prof. Fabio Sereni, Professor Emeritus of the University of Milan Medical School, Milan, Italy, for his invaluable contribution to the setup of the project; Dr. Myriam Chamorro and Dr. Raquel Abarca—La Mascota Hospital de Managua, Nicaragua, for the coordination and supervision in data entry within their hospital and the network; Dr. Franklin Baltodano—Hospital Escuela Cesar Amador Molina de Matagalpa, Dr. Luisa Torres—Hospital General Espana de Chinandega, Dr. Edgar Bojorge—Hospital Ernesto Sequeira de Blufields, Dr. Perla Raudes—Hospital Humberto Alvarado de Masaya, Dr. Ana F. Castro Lopez—Hospital Asuncion de Juigalpa, Dr. Marisol Luis Gazo—Hospital Amistad Japon de Granada, Dr. Ismael Carranza Vidaurre—Hospital Victoria Motta de Jinotega, Dr. Denise Stigol—Hospital Nuevo Amanecer de Puerto Cabezas, for data entry; Ms. Silvana Loi and Mr. Stefano Bellani—Department of Clinical Sciences and Community Health, University of Milan, Milan, Italy, for their technical support with the electronic data collection system.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Supplemental Material

Supplemental material for this article is available online.

ORCID iDs

Corrado Lanera https://orcid.org/0000-0002-0520-7428
Dario Gregori https://orcid.org/0000-0001-7906-0580

References

Bentz, C., Ruzsics, T., Koplenig, A., & Samardzic, T. (2016). A comparison between morphological complexity measures: Typological data vs. language corpora. *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*. https://www.aclweb.org/anthology/W16-4117

Biese, K. J., Forbach, C. R., Medlin, R. P., Platts-Mills, T. F., Scholer, M. J., McCall, B., . . . Kizer, J. S. (2013). Computer-facilitated review of electronic medical records reliably identifies emergency department interventions in older adults. *Academic Emergency Medicine*, *20*, 621-628.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5-32. doi:10.1023/A:1010933404324

Castillo, J. J. (2010). A machine learning approach for recognizing textual entailment in Spanish. In *Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas* (pp. 62-67). Stroudsburg, PA: Association for Computational Linguistics.

Cotik, V., Filippo, D., & Castaño, J. (2014). An approach for automatic classification of radiology reports in Spanish. *Studies in Health Technology and Informatics*, *216*, 634-638.

Crouse, H. L., Torres, F., Vaides, H., Walsh, M. T., Ishigami, E. M., Cruz, A. T., . . . Soto, M. A. (2016). Impact of an Emergency Triage Assessment and Treatment (ETAT)-based triage process in the paediatric emergency department of a Guatemalan public hospital. *Paediatrics and International Child Health*, *36*, 219-224.

Denny, M. J., & Spirling, A. (2018). Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *Political Analysis*, *26*, 168-189.

Ehret, K., & Szmrecsanyi, B. (2016). *An information-theoretic approach to assess linguistic complexity*. https://pdfs.semanticscholar.org/c4c9/d7dce2d1733f98ae268ca80eb35a3cc6e71b.pdf

Feinerer, I., & Hornik, K. (2017). tm: Text mining package (Version 0.7-1). https://CRAN.R-project.org/package=tm

Ford, E., Carroll, J. A., Smith, H. E., Scott, D., & Cassell, J. A. (2016). Extracting information from the text of electronic medical records to improve case detection: A systematic review. *Journal of the American Medical Informatics Association*, *23*, 1007-1015.

Geisler, B. P., Schuur, J. D., & Pallin, D. J. (2010). Estimates of electronic medical records in U.S. emergency departments. *PLoS One*, *5*, e9274. doi:10.1371/journal.pone.0009274

Gerbier, S., Yarovaya, O., Gicquel, Q., Millet, A.-L., Smaldore, V., Pagliaroli, V., . . . Metzger, M.-H. (2011). Evaluation of natural language processing from emergency department computerized medical records for intra-hospital syndromic surveillance. *BMC Medical Informatics and Decision Making*, *11*, 50. doi:10.1186/1472-6947-11-50

Grolemund, G., & Wickham, H. (2011). Dates and times made easy with lubridate. *Journal of Statistical Software*, *40*, 1-25.

Harrell, F. E. J. (2014). rms: Regression modeling strategies (Version 4.1-3) [R Package]. http://CRAN.R-project.org/package=rms

Heffernan, R. (2004). Syndromic surveillance in public health practice, New York City. *Emerging Infectious Diseases*, *10*, 858-864.

Henning, K. J. (2004). What is syndromic surveillance? *Morbidity and Mortality Weekly Report, 53*(Suppl.), 5-11. https://www.cdc.gov/mmwr/preview/mmwrhtml/su5301a3.htm

Hester, J. (2017). glue: Interpreted string literals. https://cloud.r-project.org/package=glue

Hirshon, J. M., Warner, M., Irvin, C. B., Niska, R. W., Andersen, D. A., Smith, G. S., & McCaig, L. F. (2009). Research using emergency department–related data sets: Current status and future directions. *Academic Emergency Medicine*, *16*, 1103-1109.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in R*. New York, NY: Springer.

Johnson, T., Gaus, D., & Herrera, D. (2016). Emergency department of a rural hospital in Ecuador. *Western Journal of Emergency Medicine*, *17*(1), 66-72.

Kim, J.-H. (2009). Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational Statistics & Data Analysis*, *53*, 3735-3745. doi:10.1016/j.csda.2009.04.009

Lall, R., Abdelnabi, J., Ngai, S., Parton, H. B., Saunders, K., Sell, J., . . . Mathes, R. W. (2017). Advancing the use of emergency department syndromic surveillance data, New York City, 2012-2016. *Public Health Reports, 132*(1_suppl), 23S-30S.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news*, *2*(3), 18-22.

Liu, N., Zhang, Z., Wah Ho, A. F., & Hock Ong, M. E. (2018). Artificial intelligence in emergency medicine. *Journal of Emergency and Critical Care Medicine*, *2*, 82. doi:10.21037/jeccm.2018.10.08

Metzger, M.-H., Tvardik, N., Gicquel, Q., Bouvry, C., Poulet, E., & Potinet-Pagliaroli, V. (2017). Use of emergency department electronic medical records for automated epidemiological surveillance of suicide attempts: A French pilot study. *International Journal of Methods in Psychiatric Research*, *26*(2). doi:10.1002/mpr.1522

Obermeyer, Z., Abujaber, S., Makar, M., Stoll, S., Kayden, S. R., Wallis, L. A., & Reynolds, T. A. (2015). Emergency care in 59 low- and middle-income countries: A systematic review. *Bulletin of the World Health Organization*, *93*, 577-586.

Pérez, A., Weegar, R., Casillas, A., Gojenola, K., Oronoz, M., & Dalianis, H. (2017). Semi-supervised medical entity recognition: A study on Spanish and Swedish clinical corpora. *Journal of Biomedical Informatics*, *71*, 16-30.

R Core Team. (2017). R: A language and environment for statistical computing (Version 3.4.2). Vienna, Austria: R Foundation for Statistical Computing.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, *24*, 513-523. doi:10.1016/0306-4573(88)90021-0

Sequeira, M., Espinoza, H., Amador, J. J., Domingo, G., Quintanilla, M., & De los Santos, T. (2011). *The Nicaraguan health system*. https://path.azureedge.net/media/documents/TS-nicaragua-health-system-rpt.pdf

Taira, B. R., Orue, A., Stapleton, E., Lovato, L., Vangala, S., Tinoco, L. S., & Morales, O. (2016). Impact of a novel, resource appropriate resuscitation curriculum on Nicaraguan resident physician's management of cardiac arrest. *Journal of Educational Evaluation for Health Professions*, *13*, 25. doi:10.3352/jeehp.2016.13.25

Tanev, H., Zavarella, V., Linge, J., Kabadjov, M., Piskorski, J., Atkinson, M., & Steinberger, R. (2009). Exploiting machine learning techniques to build an event extraction system for Portuguese and Spanish. *Linguamática*, *1*(2), 55-66.

Wickham, H. (2017a). stringr: Simple, consistent wrappers for common string operations. https://stringr.tidyverse.org/

Wickham, H. (2017b). tidyverse: Easily install and load the "tidyverse." https://tidyverse.tidyverse.org/

Wing, M., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., . . . Hunt, T. (2017). caret: Classification and regression training. https://rdrr.io/cran/caret/

Worster, A., Bledsoe, R. D., Cleve, P., Fernandes, C. M., Upadhye, S., & Eva, K. (2005). Reassessing the methods of medical record review studies in emergency medicine research. *Annals of Emergency Medicine*, *45*, 448-451.