Medical Care
Research and Review

Analysis of unstructured text-based data using machine learning techniques: the case of pediatric emergency department records in Nicaragua

Journal:	*Medical Care Research and Review*
Manuscript ID	MCRR-2018-0091-ER.R2
Manuscript Type:	Empirical Research
Keywords:	Emergency Department visits, Low- and middle-income countries, Free text Discharge Diagnosis, Spanish, Random Forest

Analysis of unstructured text-based data using machine learning techniques: the case of pediatric emergency department records in Nicaragua

Abstract

Free text information is still widely used Emergency Department (ED) records. Machine Learning Techniques (MLT) are useful for analyzing narratives, but they have been used mostly for English-language datasets. Considering such a framework, it was tested the performance of an ML classification task of a Spanish-language ED visits database.

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 dataset (1789 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 dataset and tested. The classification accuracies had a median of 0.783 (95% C.I. 0.779-0.796).

MLTs seemed to be a promising opportunity for the exploitation of unstructured information reported in ED records in low- and middle-income Spanish-speaking countries.

Running title: Analysis of text-based Emergency Department records

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, i.e., 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 USA (Geisler, Schuur, and 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 continually being developed, but their use is not straightforward (Biese et al. 2013).

Such barriers in the analysis of ED datasets 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 (Taira et al. 2016; Crouse 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, and Herrera 2016).

Conceptual framework

Given such a framework, besides a progressive development of a standardized data collection system for ED visits, in both high- 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 et al. 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 datasets. Only a few examples are available in the literature about the application of MLT to the Spanish language (Pérez et al. 2017; Cotik, Filippo, and Castaño 2014; Castillo 2010; 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 and Szmrecsanyi 2016) (e.g., Spanish exhibits slightly higher levels of morphological complexity compared to English (Bentz et al. 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 contribution

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 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 datasets. Considering such a framework, it was tested the performance of a Machine Learning classification task of Spanish free-text discharge diagnoses reported in an ED visits database from Nicaragua.

Methods

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, USA), 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 Supplementary Material (Table S1). 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.

For the study, we focused on ED visits reported in the data collection system in 2012 for which discharge diagnosis was available. The full dataset (ED visits collected in 2012) was represented by 2723 ED visits, and those for which discharge diagnosis was available were 1789 (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 ten 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 dataset 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, Pre-processing, 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 pre-processing (Denny and Spirling 2017) consisted in the transformation of all characters in lower-case letters, in the removal of all non-alphabetical 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. It was reported the Term Frequency-inverse Document Frequency (TF-iDF) (Salton and Buckley 1988) in each cell of the DTM. 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 a token appeared in the diagnoses. The most important tokens (including bigrams) are reported in Table S2 of the Supplementary 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 five hundred bootstrap resamples of the dataset. Each training procedure involved the fitting of a set of Random Forests (RFs) MLT (Breiman 2001; Liaw and 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 (CV) 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, 1000 up to the maximum number of variables -*tokens*- available).

Once the optimal *mtry was* chosen (through the five repetitions of the 10-fold CV procedure), a final validated model (i.e., a brand new RF made up of new 500 trees), was trained on the whole bootstrapped dataset (1789 bootstrapped ED visits), and tested on its Out-Of-Bag (OOB) set, i.e., 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 (I and III quartiles) for continuous variables, and percentages (absolute numbers) for categorical variables. Thanks to the bootstrap procedure adopted, the classification task could have been evaluated by the Out-Of-Bag (OOB) classification performance of the final trained RF (Anon n.d.) for each one of the 500 bootstrapped RFs, i.e., the performance of every one of this final set of forests were assessed

on the set of observations not included in the bootstrapped dataset used to train the trees in the forest. 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 (ver. 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 and 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 and Hornik 2017), randomForest (Liaw and Wiener 2001) 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(R) quad Core (TM) i7-6700 CPU @ 3.4GHz with x64-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

One thousand seven hundred eighty-nine 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 two years) of male gender (56%). According to the gold standard (manual classification), the discharge diagnoses' class most represented was that about 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 three thousand eight hundred ninety-one distinct tokens were considered in the analyses, in particular, they range from two hundred fifty-six distinct tokens for Hospital Juigalpa to one thousand five hundred fifty-two distinct tokens for Hospital La Mascota, and a median of 461 tokens. The overall CPU time (on Intel(R) quad Core (TM) i7-6700 CPU @ 3.4GHz with x64-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% C.I. 0.7792-0.7965) on the dataset 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 ML 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 models was better (p-value <0.001) for male gender (0.788 95% C.I. 0.783-0.785) compared to the female ones (0.777 95% C.I. 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 healthcare system, although hampered by a lack of resources. From the epidemiological point of view, Nicaragua is still considered a pre-transitional country, characterized by a high prevalence of infectious 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 to 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 et al. 2018). It is worth pointing out that the ED visits included in the analyses were the most severe ones corresponding to 1-2% 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 to 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 *(i)* 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, *(ii)* the identification of pediatric population healthcare needs, *(iii)* the monitor of the performance of the EDs, and *(iv)* the evaluation of the effectiveness of public health interventions.

Limitations

The main limitations were represented by the fact that the MLTs was applied to a small (1789 ED records) dataset in the Spanish language, which has been only rarely analyzed using MLTs. The fact that the dataset 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 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 (Pérez et al. 2017; Cotik, Filippo, and Castaño 2014; Castillo 2010; 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 datasets allowing for a more detailed classification, through a strict collaboration between physicians, epidemiologists, and big data specialists.

References

Anon. An Introduction to Statistical Learning - with Applications in R | Gareth James | Springer,

Bentz, C., T. Ruzsics, A. Koplenig, and T. Samardzic. 2016. A Comparison between Morphological Complexity Measures: Typological Data vs. Language Corpora. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pp. 142–153.

Biese, K. J., C. R. Forbach, R. P. Medlin, T. F. Platts-Mills, M. J. Scholer, B. McCall, F. S. Shofer, M. LaMantia, C. Hobgood, and J. S. Kizer. 2013. "Computer-facilitated Review of Electronic Medical Records Reliably Identifies Emergency Department Interventions in Older Adults." *Academic Emergency Medicine*, 20(6): 621–628.

Breiman, L. 2001. "Random Forests." *Machine Learning*, 45(1): 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. Association for Computational Linguistics.

Cotik, V., D. Filippo, and J. Castaño. 2014. "An Approach for Automatic Classification of Radiology Reports in Spanish." *Studies in health technology and informatics*, 216: 634–638.

Crouse, H. L., F. Torres, H. Vaides, M. T. Walsh, E. M. Ishigami, A. T. Cruz, S. B. Torrey, and M. A. Soto. 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(3): 219–224.

Denny, M. J. and A. Spirling. 2017. "Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It," SSRN Scholarly Paper ID 2849145, Rochester, NY: Social Science Research Network [accessed on May 9, 2017]. Available at: https://papers.ssrn.com/abstract=2849145.

Ehret, K. and B. Szmrecsanyi. 2016. "An Information-Theoretic Approach to Assess Linguistic Complexity." *Complexity and isolation. Berlin: de Gruyter*.

Feinerer, I. and K. Hornik. 2017. Tm: Text Mining Package,

Ford, E., J. A. Carroll, H. E. Smith, D. Scott, and J. A. Cassell. 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(5): 1007–1015.

Geisler, B. P., J. D. Schuur, and D. J. Pallin. 2010. "Estimates of Electronic Medical Records in US Emergency Departments." *PLoS One*, 5(2): e9274.

Gerbier, S., O. Yarovaya, Q. Gicquel, A.-L. Millet, V. Smaldore, V. Pagliaroli, S. Darmoni, and M.-H. Metzger. 2011. "Evaluation of Natural Language Processing from Emergency Department Computerized Medical Records for Intra-Hospital Syndromic Surveillance." *BMC medical informatics and decision making*, 11(1): 50.

Grolemund, G. and H. Wickham. 2011. "Dates and Times Made Easy with Lubridate." *Journal of Statistical Software*, 40(3): 1–25.

Harrell, F. E. J. 2014. "Rms: Regression Modeling Strategies. R Package Version 4.1-3." Available at: http://CRAN.R-project.org/package=rms.

Heffernan, R. 2004. "Syndromic Surveillance in Public Health Practice, New York City."

Henning, K. J. 2004. "What Is Syndromic Surveillance?" *Morbidity and Mortality Weekly Report*: 7–11.

Hester, J. 2017. Glue: Interpreted String Literals,

Hirshon, J. M., M. Warner, C. B. Irvin, R. W. Niska, D. A. Andersen, G. S. Smith, and L. F. McCaig. 2009. "Research Using Emergency Department–Related Data Sets: Current Status and Future Directions." *Academic emergency medicine*, 16(11): 1103–1109.

Johnson, T., D. Gaus, and D. Herrera. 2016. "Emergency Department of a Rural Hospital in Ecuador." *Western Journal of Emergency Medicine*, 17(1): 66.

Kim, J.-H. 2009. "Estimating Classification Error Rate: Repeated Cross-Validation, Repeated Hold-out and Bootstrap." *Computational Statistics & Data Analysis*, 53(11): 3735–3745, doi:10.1016/j.csda.2009.04.009.

Lall, R., J. Abdelnabi, S. Ngai, H. B. Parton, K. Saunders, J. Sell, A. Wahnich, D. Weiss, and R. W. Mathes. 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. and M. Wiener. 2002. "Classification and Regression by RandomForest." *R news*, 2(3): 18–22.

Liaw, A. and M. Wiener. 2001. Classification and Regression by RandomForest,

Liu, N., Z. Zhang, A. F. Wah Ho, and M. E. Hock Ong. 2018. "Artificial Intelligence in Emergency Medicine." *Journal of Emergency and Critical Care Medicine*, 2.

Metzger, M.-H., N. Tvardik, Q. Gicquel, C. Bouvry, E. Poulet, and V. Potinet-Pagliaroli. 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).

Obermeyer, Z., S. Abujaber, M. Makar, S. Stoll, S. R. Kayden, L. A. Wallis, and T. A. Reynolds. 2015. "Emergency Care in 59 Low-and Middle-Income Countries: A Systematic Review." *Bulletin of the World Health Organization*, 93(8): 577–586.

Pérez, A., R. Weegar, A. Casillas, K. Gojenola, M. Oronoz, and H. Dalianis. 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, Vienna, Austria: R Foundation for Statistical Computing.

Salton, G. and C. Buckley. 1988. "Term-Weighting Approaches in Automatic Text Retrieval." *Inf. Process. Manage.*, 24(5): 513–523, doi:10.1016/0306-4573(88)90021-0.

Sequeira, M., H. Espinoza, J. J. Amador, G. Domingo, M. Quintanilla, and T. De los Santos. 2011. "The Nicaraguan Health System." *Seattle, WA: PATH*, 201(1).

Taira, B. R., A. Orue, E. Stapleton, L. Lovato, S. Vangala, L. S. Tinoco, and O. Morales. 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.

Tanev, H., V. Zavarella, J. Linge, M. Kabadjov, J. Piskorski, M. Atkinson, and R. Steinberger. 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,

Wickham, H. 2017b. Tidyverse: Easily Install and Load the "Tidyverse,"

Wing, M. K. C. from J., S. Weston, A. Williams, C. Keefer, A. Engelhardt, T. Cooper, Z. Mayer, B. Kenkel, the R. C. Team, M. Benesty, R. Lescarbeau, A. Ziem, L. Scrucca, Y. Tang, C. Candan, and T. Hunt. 2017. Caret: Classification and Regression Training,

Worster, A., R. D. Bledsoe, P. Cleve, C. M. Fernandes, S. Upadhye, and K. Eva. 2005. "Reassessing the Methods of Medical Record Review Studies in Emergency Medicine Research." *Annals of emergency medicine*, 45(4): 448–451.

Table 1. Variables included in the dataset with the corresponding type and examples

Variables	Data type	Examples
Age (in years)	Numerical	6; 38; 13
Gender	Categorical	Male; Female
Vital signs (i.e., body temperature, blood pressure, heart rate, and 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 ED 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

Table 2. Children's characteristics according to diagnosis category. Data are expressed as medians [I; III quartile] for continuous data and percentages (absolute number) for categorical ones

	N, *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% (1003)	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% (1789)	2.0 [1.0; 6.0]	56% (992)

Table 3. Median accuracy (rate of diagnosis correctly classified by the final validated model) of the ML algorithms together with 95% Confidence Interval (C.I.) (calculated considering the Out-Of-Bag performance of 500 bootstrap repetitions of evaluation a 500-trees RF classifier by 5 repetition of 10-fold CV procedure). The C.I. was not estimated for Burn, Metabolic and discharge diagnosis' classes because of the small size of the sample of children in these classes.

	Accuracy	**(95% C.I.)**
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)

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-trees RF classifier by 5 repetition of 10-fold CV procedure). 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 semi-transparent dot corresponds to the performance of a single RF into the pool created by the bootstrapped procedure

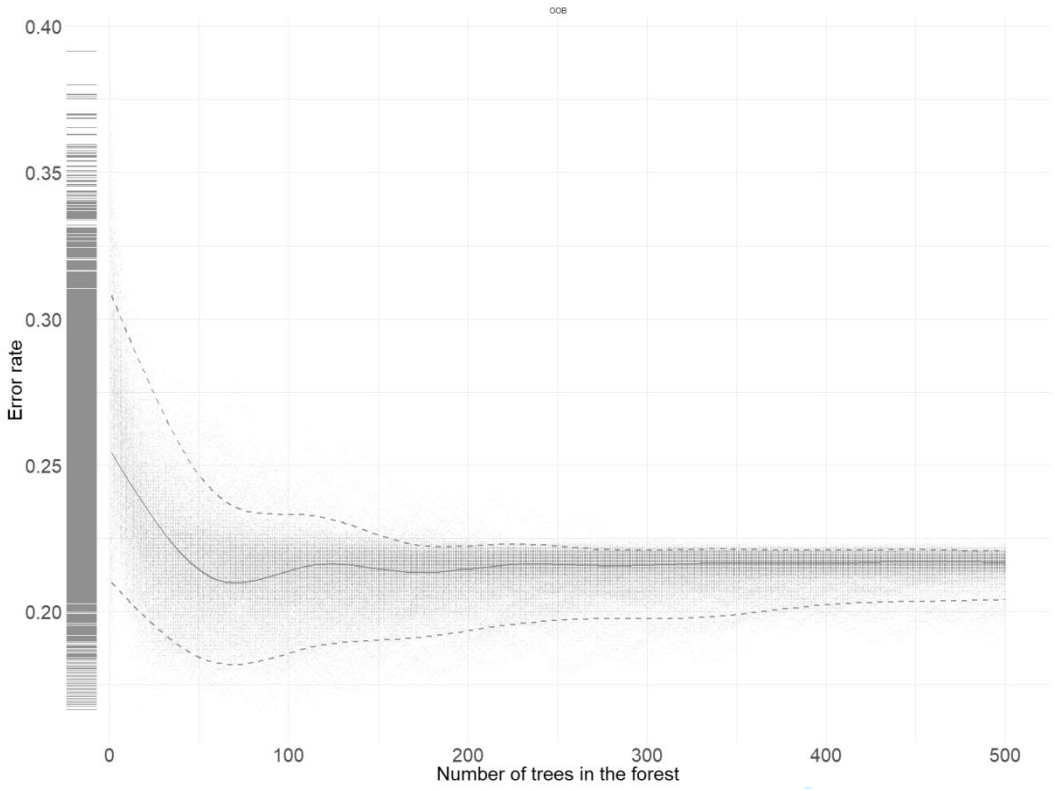


Figure 2 Training Procedure: (ED: Emergency Department; CV: Cross-Validation; RF: Random Forest; OOB: Out-Of-Bag) For each of the 500 bootstrap resampled dataset the performance estimation was calculated on its OOB set, which was never seen by the training procedure and different for every sample. For the final model trained on each bootstrap sample, the optimal parameter was selected by 5 repetition of 10-fold CV estimation.

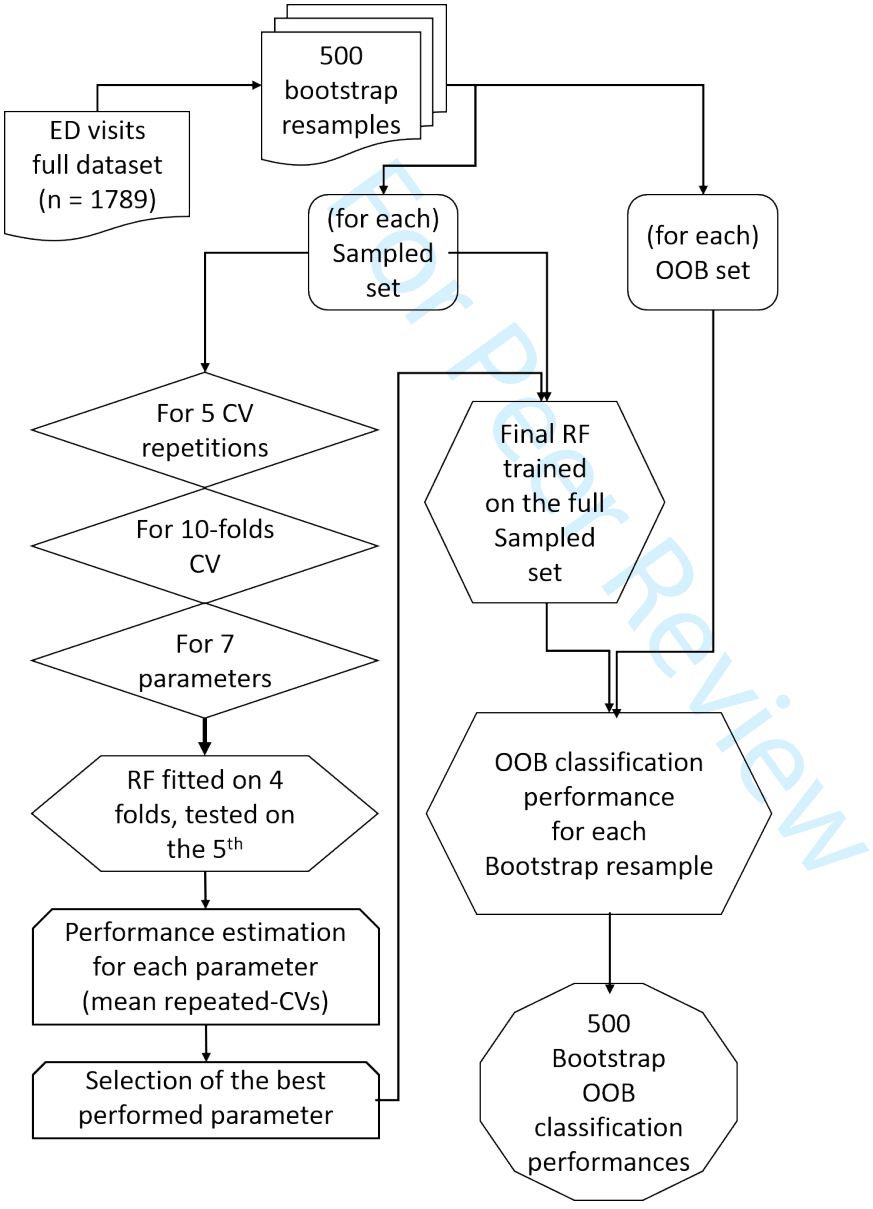
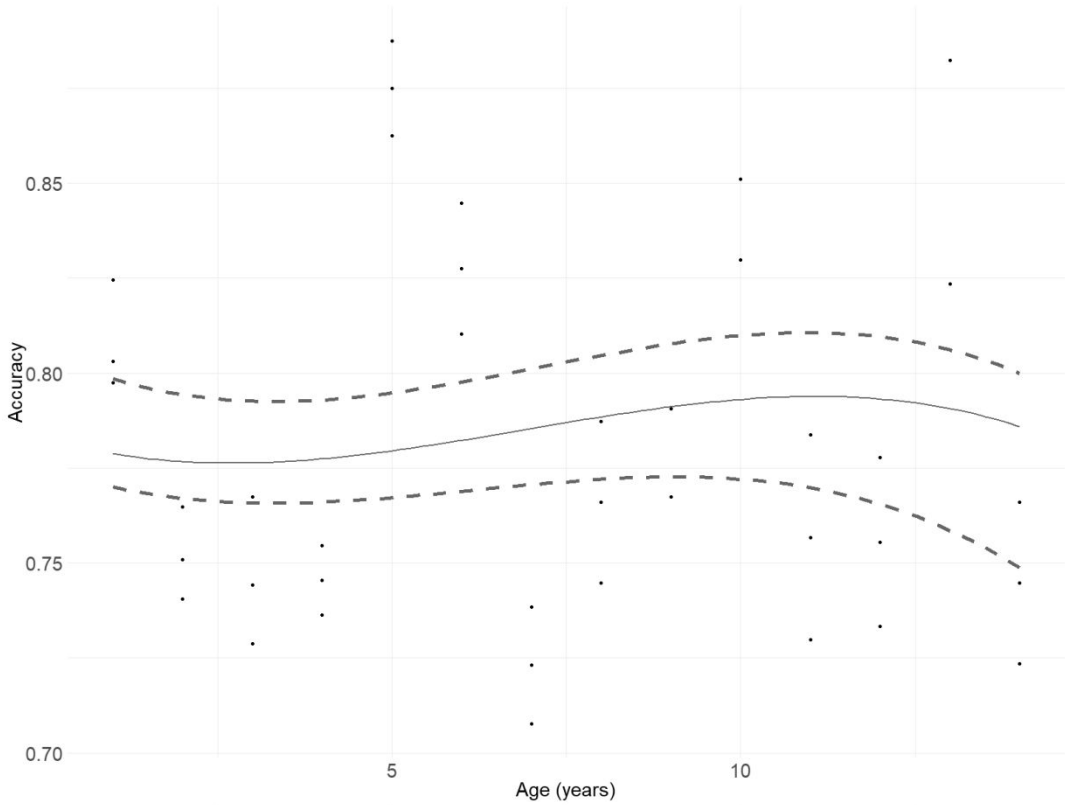


Figure 3. Accuracy according to children's age. Dashed lines represent 95% C.I. (calculated considering 500 bootstrap repetitions), solid line represents the median.


0.85
0.80
0.75
0.70
Accuracy
5
10
Age (years)

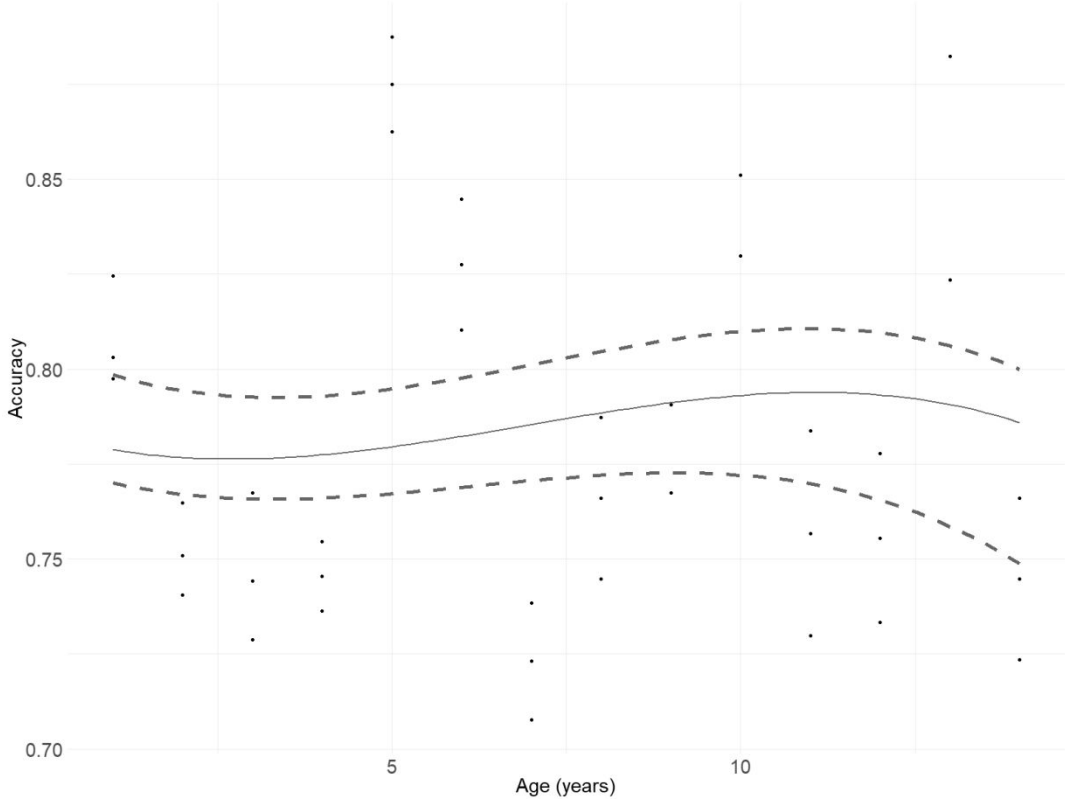

0.85
0.80
0.75
0.70
Accuracy
5
10
Age (years)

Table S1. Criteria for inclusion in the study registry of urgent-emergent paediatrics visits to Paediatric Emergency Departments in Nicaragua

Neurologic	• Persistent altered mental status (GCS < 15) • Signs of raised intracranial pressure • Signs of severe neuroinfection • Active seizures on arrival • Acute focal neurological signs
Respiratory	• Signs of airway obstruction • Severe respiratory distress (based on PALS 2015) • Bradipnoea/apnoea
Cardiovascular	• Cardiac arrest • Signs of shock (based on PALS 2015) • Tachycardia/bradycardia • Signs of cardiac failure • Suspected sepsis • Hypoxic spells
Gastrointestinal	• Acute gastroenteritis (vomiting and/or diarrhea) with severe dehydration (based on clinician judgment) • Gastrointestinal bleeding • Acute abdomen
Metabolic	• Diabetic ketoacidosis
Injury	• Potentially severe isolated (single site) or multiple trauma • Burns > 20% BSA • Venomous snake bite • Poisoning (high risk- based on respiratory, neurologic, cardiovascular, and/or gastrointestinal signs or symptoms
Suspected Dengue with warning signs	

BSA= Body Surface Area; GCS= Glasgow Coma Scale;

Table S2. Common tokens (including bigrams) among the 500 bootstrapped final validated model appearing in the top 100 of each model (according to the TF-iDF weight)

Tokens (including bigrams)
cetoacidosis
dengue
diabetes
grado
intoxicacion
“intoxicacio por”
quemadura
sepsis
shock
sustancia
trauma