The prediction of hospital length of stay using unstructured data

Jan Chrusciel, François Girardon, Lucien Roquette, David Laplanche, Antoine Duclos, Stephane Sanchez

Chrusciel *et al.*
BMC Medical Informatics and Decision Making (2021) 21:351
https://doi.org/10.1186/s12911-021-01722-4

BMC Medical Informatics and Decision Making

RESEARCH

Open Access

The prediction of hospital length of stay using unstructured data

Jan Chrusciel[1], François Girardon[2], Lucien Roquette[2], David Laplanche[1], Antoine Duclos[3,4] and Stéphane Sanchez[1*]

Abstract

Objective: This study aimed to assess the performance improvement for machine learning-based hospital length of stay (LOS) predictions when clinical signs written in text are accounted for and compared to the traditional approach of solely considering structured information such as age, gender and major ICD diagnosis.

Methods: This study was an observational retrospective cohort study and analyzed patient stays admitted between 1 January to 24 September 2019. For each stay, a patient was admitted through the Emergency Department (ED) and stayed for more than two days in the subsequent service. LOS was predicted using two random forest models. The first included unstructured text extracted from electronic health records (EHRs). A word-embedding algorithm based on UMLS terminology with exact matching restricted to patient-centric affirmation sentences was used to assess the EHR data. The second model was primarily based on structured data in the form of diagnoses coded from the International Classification of Disease 10th Edition (ICD-10) and triage codes (CCMU/GEMSA classifications). Variables common to both models were: age, gender, zip/postal code, LOS in the ED, recent visit flag, assigned patient ward after the ED stay and short-term ED activity. Models were trained on 80% of data and performance was evaluated by accuracy on the remaining 20% test data.

Results: The model using unstructured data had a 75.0% accuracy compared to 74.1% for the model containing structured data. The two models produced a similar prediction in 86.6% of cases. In a secondary analysis restricted to intensive care patients, the accuracy of both models was also similar (76.3% vs 75.0%).

Conclusions: LOS prediction using unstructured data had similar accuracy to using structured data and can be considered of use to accurately model LOS.

Keywords: Emergency department, Length of stay, Data mining, Health services research

Introduction

Length of stay (LOS) is a critical indicator for hospital management and has direct consequences on hospital costs and patient satisfaction. Moreover, LOS is correlated with disease severity and mortality [1]. When a patient is in the emergency department (ED), some predictors of hospital LOS are known before hospital admission. Studies have found patients at an ED are associated with a longer LOS [2–5] and patients who develop further complications in intensive care units (ICU) have a longer LOS beforehand at the ED [6]. For stroke patients, however, there is a significant inverse linear association between LOS at the ED and hospital LOS [7]. ED crowding and hospital occupancy at entry are predicted to have longer LOS [5, 8], however, there are other hospital characteristics that play a role in determining it [9, 10].

Patient characteristics also influence LOS, such as demographic characteristics and comorbidities which are

*Correspondence: stephane.sanchez@ch-troyes.fr
[1] Pôle Territorial Santé Publique et Performance, Centre Hospitalier de Troyes, 101 Avenue Anatole France CS 10718, 10003 Troyes Cedex, France
Full list of author information is available at the end of the article

often available at admission [11]. Depending on the medical specialty, physicians can predict LOS [12] although they tend to underestimate LOS in some cases such as patients with heart failure with LOS > 3 days [13, 14]. In psychiatry, patients have their own predictors such as a history of attempted suicide, which was negatively associated with LOS in a sample of 385 patients in Brazil [15]. These predictors are different according to age, where isolation plays a greater role for geriatric patients [16], but remains difficult to predict for psychiatric patients [17].

Although indicators can be compiled in bedside clinical scores like ALICE [18], statistical models can offer more flexibility for predictions of patient LOS. To date, logistic regression models have been used to predict discharge [19]. Cubist models have shown LOS prediction results [20] and tree-based models have presented improved performance and interpretability [21]. However, these models are usually run on structured data in tabular databases.

Most clinical data in electronic health records (EHRs) are presented in unstructured text form such as patient history narratives written by physicians. Although this data contains valuable information, it has rarely been used for automated predictions, particularly in the context of an ED. To date, these methods for knowledge extraction are not widely available in the medical field. Moreover, manual chart abstraction is time-consuming and expensive [22]. Automated information extraction from unstructured text can be simplified using controlled vocabularies [23, 24] like the Unified Medical Language System (UMLS). The objective of this study was to assess performance improvement for LOS prediction when accounting for clinical information written in text compared to the traditional approach of solely considering structured information.

Methods

Inclusion criteria

This retrospective study included patients admitted to the *Centre Hospitalier de Troyes,* a large French hospital situated in a rural region, between 1 January 2019 and 24 September 2019. Patient were included if they were admitted to the hospital through the ED and stayed more than two days for the subsequent hospital ward. Patients not admitted through the ED and patients with very short subsequent hospital stays (< 2 days) were excluded. The hospital under study had a Short Stay Emergency Ward. This unit has the capacity to host patients for several days, therefore, it is treated as any other medical ward. The time spent in the Short Stay Emergency Ward was accounted for in the total LOS.

Data source

Patient stays and related features were selected and extracted all at once using the Dr Warehouse platform [25]. The information used for modelling was all information that was available to the ED staff at the time of the patient's transfer to another ward of the hospital. This information included: i) personal information such as age, gender and zip/postal code, ii) context information such as entry date, LOS at the ED, triage (CCMU and GEMSA) codes, iii) ICD-10 primary diagnosis code and iv) unstructured information such as the UMLS concepts extracted from the text documents uploaded during the stay at the ED.

Ethical and regulatory considerations

The study was declared to the French registry of studies using healthcare data (N° F20210719114017). The study was conducted in compliance with French MR004 regulation (*Commission Nationale Informatique et Libertés*). Since the study was retrospective and was based on pseudonymized data and purely observational, it was exempt from Institutional Review Board approval according to the French Public Health Code (L1121-1, Law number 2012-300, 5 March 2012).

UMLS concept extraction

UMLS is a meta-thesaurus and ontology of medical concepts created by the National Library of Medicine (USA) covering a broad range of concepts from anatomy to physiology and medical semiology. It includes vocabularies from SNOMED CT, RxNorm, LOINC, MeSH, CPT, ICD-10-CM, MedDRA, the Human Phenotype Ontology and other sources. We used the UMLS detection module of the Dr Warehouse platform [26] to extract UMLS concepts from free text in the EHRs at the ED. The main computation steps used for the extraction were [27]: i) to split the free-text into a collection of sub-text (sentences, or propositions) using punctuation and text structure, ii) to classify each sub-text within the following categories: "patient related—affirmation", "patient related—negation", "family related—affirmation", "family related—negation", where affirmation stood also for neutral context, and iii) for each sub-text labelled as "patient related—affirmation" to find the most precise concepts that exactly match concepts of the UMLS thesaurus within this sub-text. The UMLS tree-structure was leveraged to reach the most precise concept that is a concept leaf of the tree.

To address the issue of high prevalence of some concepts, a variation of the relevance frequency concept [28] was used to filter out non-relevant concepts. For each

extracted concept we computed the *srf* (symmetric relevance frequency) score as follows:

$$srf = \log_2(2 + \max\ (\mathrm{a}/\max(1,\mathrm{c}),\ \mathrm{c}/\max(1,\mathrm{a}))$$

where a is the number of long stays (≥ 7 days) in which the concept is found and c is the number of short stays (<7 days) in which the concept is found. All concepts for which the prevalence of one class over the other was under the 45% threshold, meaning $srf \leq \log_2(2+55/45) = 1.688$, were marked as non-relevant.

ICD-10 diagnostic codes

Numerous diagnoses have a low number of occurrences. To tackle this issue, the hierarchical structure of the ICD-10 diagnosis code was leveraged in our study. For each diagnosis code, if the number of occurrences was lower than five, we replaced it with its parent in the hierarchy, stopping at the three characters level (such as C00) to avoid losing too much information. For example, if M6289 appeared in less than five stays then it was replaced by M628. If this code still appeared in less than five stays, it was then replaced by M62.

CCMU classification code

The CCMU classification code consists of either "P" if the patient presents psychiatric symptoms, "D" if the patient is deceased on arrival or a number between 1 and 5 depicting the patient's condition (1: stable and 5: vital prognosis engaged). The numbers were left unchanged; however, the letters had to be replaced by numerical values. The letter D was replaced by the number 6 and the letter P was replaced with the number 0.

Added features

To improve model performance, several features were built using the available data. Firstly, the "recent prior visit" feature was built by looking at previous admissions in the ED for each patient. The "recent prior visit" value was defined as 1 if a patient had already been admitted at least once in the seven days prior to the given stay and 0 otherwise. We obtained a total of 656 (13%) stays with the flag set to 1 out of the total stays.

Another added feature was the short-term ED activity index since ED crowding has been shown to help predict patients' overall LOS [29]. Although there were other determinants of crowding (for example, the number of beds available outside of the ED), crowding was expected to occur with increased frequency when the number of incoming patients was unusually high. For each ED admission, we counted the number of admissions that occurred during the seven previous days, and then: i) if the count was under the 1st decile of prior-admission counts then the index was set to 0, ii) if the count was over the 9th decile of prior admission counts then the index was set to 1 and iii) if the count was between the two values, then the index was linearly interpolated. This indicator also captures seasonal effects, being low in periods of the year during which patients are less likely to come to the ED of the hospital under study.

To produce a fair comparison, the following three sets of features were chosen: i) features common to both sets including age, gender, zip/postal code, LOS at the ED, recent visit flag, short-term ED activity index, hospital service after ED stay, ii) "structured data only" set including CCMU, GEMSA & ICD-10 codes, iii) "unstructured data included" set: UMLS concepts.

Personal information of patients and the context were kept for both featured sets. In the first set, the structured diagnosis information was added, whereas in the second set, only the clinical data directly extracted from the text notes was added. The variables used for the structured data model and for the unstructured model are summarized in Additional file 1: Table S5. Additional information on how the data was encoded for the Random Forest Model can be found in Additional file 2: Appendix 2. Although the same kind of model was used for both featured sets, each set's model had its own set of hyper-parameters. Both sets of hyper-parameters were computed independently to optimize performance in each case. This allowed us to compare the best achievable model for both set of features.

Model

To alleviate the problems inherent to the modelling of long-tail distributions such as LOS, we decided to reduce the inference scenario to a binary classification defined by the ad-hoc threshold of seven days. This threshold represented the median LOS for our dataset ensuring balanced classes in the classification outcomes. A random forest model was used to predict the "long stay" and "short stay" classes of the LOS variable. Long stays were defined as lasting longer than the LOS median of six days. One motivation for the choice of a random forest model was the distribution of the classes shown in Additional file 1: Figure S1 (Appendix 1). Indeed, the decision region shapes needed to correctly encompass each class were too complex for linear or kernel-based models. A tree-based model, however, could sufficiently produce complex decision regions. Moreover, random forests have unique properties like the reduction of overfitting by averaging multiple decision trees [30].

It is worth noting that the ICD-10 diagnosis codes and UMLS Concepts are categorical features, meaning that to be used by the machine learning models, they had to be encoded using One-Hot Encoding (each category value is

converted into a new column and assigned a 1 or 0 (notation for true/false) value to the column. In this study, there were 969 UMLS concepts and each stay had 17.23 associated UMLS concepts on average. Whereas there were 222 ICD10 diagnostic codes with 1 per stay showing numerous co-occurrences of UMLS concepts. This translated to a high multi-collinearity between concepts, which is a problem for linear and kernel-based models.

Model hyperparameter tuning

The method used to choose the optimal set of hyperparameters for each feature set were described. This method consisted of using a random search to go through hyperparameter combinations (within previously defined bounds) and evaluate the model's performance with each one using cross-validation. The set of hyperparameters used for the final model was the set that produced the best model accuracy. The list of all possible values for each hyperparameter can be found in Additional file 1: Table S1 (Appendix 1).

The key hyperparameters used were: i) n_estimators: This parameter determined the number of decision trees that constituted our forest. Additional trees, up to a certain point, improve model performance, ii) min_samples_split: This controlled the minimum number of samples required for a split to be able to happen on a node. Too high values led to under fitting as trees were not able to split enough times to achieve high-purity leaves. *Note:* We placed the lower bound to 5 to allow splitting even when considering infrequent diagnosis codes and iii) min_samples_leaf. Similar to ii), this parameter set a minimum number of samples required for a leaf node after splitting. The minimum value for this one was set low enough to account for the very infrequent diagnosis codes.

With the hyperparameter ranges of values defined, a random search was used to go through combinations of hyperparameters. Through this method, each iteration produced a different, randomly-chosen combination of hyperparameter values. Each combination was then evaluated using cross validation. The process of evaluating a model through cross validation started with partitioning the dataset into several "folds", in other words subsets of equal size. For every such fold, the model was fitted on the union of all the other folds and its score was evaluated on the given fold (which was left out in the model fitting). The mean of the scores obtained in that manner constituted the score of the set of hyperparameters.

Primary outcome

The primary outcome of this study was accuracy = TP + TN/(P + N) as the score, where TP is True Positives, TN is True Negatives, P is Positives and N is Negatives with long stays being considered positives. We used three folds and also recorded other indicators: i) sensitivity: proportion of actual long stays (≥ 7 days) predicted as such, ii) specificity: proportion of actual short stays (<7 days) predicted as such, iii) precision: proportion of correct predictions among predicted long stays and iv) accuracy: proportion of correct predictions.

Both models were fitted on a subset made of 80% of the dataset (training set) and evaluated on the rest (test set). Furthermore, both models were fitted on the exact same training set and evaluated on the exact same test set. Each model used its own set of hyperparameter values obtained using the method described earlier with the same number of iterations on the random search.

Results

Patient characteristics

In total, 5,006 patients were included in the study. Patient characteristics are presented in Table 1. The admission rate in the ED of the hospital under study was 28.7% in 2019. The types of stays registered in the ED of this hospital were very diverse: even the most prevalent diagnoses had a relatively low number of occurrences. This was not the case for the UMLS concepts, with the top two most frequent UMLS concepts being present in more than 91% of all stays (Additional file 1: Figure S2, Appendix 1).

Model results and performance

Overall, model performance of the two models (unstructured data vs structured data) were similar. The set of hyperparameter values chosen for each model are presented in Additional file 1: Table S2. Examples using other parameters are presented in Additional file 1: Figure S3 (Appendix 1). These values were produced using 50 iterations on the random search and 3 folds in the cross validation.

Table 2 shows the performance of each model. Including the clinical data extracted from text notes produced in the ED led to a small increase in predictive performance, from 74.1% to 75.0% (with an F1-score change from 75.7% to 76.4%). The two models concurred in 86.6% of predictions. The number of records for which the two model predictions differ or concur is highlighted in Additional file 1: Table S3. As shown in Additional file 1: Table S4 (Appendix 1), there was a distinction between the characteristics of EHRs which the models produced the same prediction and those for which the predictions were different.

Figures 1 and 2 present the relative importance of the features for both the unstructured data and structured data models. Age was the most determining factor in predicting LOS. Another important feature was the short-term ED activity index. Regarding UMLS concepts,

Table 1 Patient characteristics for the study to assess the prediction of hospital length of stay (LOS) using unstructured data at the emergency department (ED)

Characteristic	Total
n	5,006
Age—mean ± SD	64.3 ± 26.3
Age category—n (%)	
< 18	494 (9.9)
Age ≥ 18	4512 (90.1)
Gender—n (%)	
Male	2,333 (46.6)
Female	2,673 (53.4)
Emergency LOS (hours)—Median (Q1–Q3)	7.2 (4.8–9.6)
Total (ED + hospital) LOS (days)—Median (Q1–Q3)	6.1 (3.7–11.0)
Intensive care patients—n (%)	378 (7.6)
Most frequent diagnoses—n (%)	
Pneumonia (J189)	212 (4.2)
Altered general health (R53 + 0)	188 (3.8)
Shortness of breath (R060)	174 (3.5)
Abdominal pain (R104)	122 (2.4)
Femoral bone fracture (S7200)	121 (2.4)
Most frequent concepts—n (%)	
Pain	4,921 (98.3)
Blood pressure	4,568 (91.3)
Capillary	3,521 (70.3)
Abdomen	2,155 (43.0)
Face	2,046 (40.9)
Type of hospital stay, n (%)	
Pulmonology	871 (17.4)
Digestive system	761 (15.2)
Cardiovascular medicine (except cardiovascular catheterization)	503 (10.0)
Trauma and orthopaedics	467 (9.3)
Diseases of the nervous system (including stroke)	465 (9.3)
Urology, nephrology	332 (6.6)
Rheumatology	313 (6.3)
Endocrinology	195 (3.9)
Hematology	193 (3.9)
Diagnostic or therapeutic catheterization	161 (3.2)
Dermatology	134 (2.7)
ENT, stomatology	128 (2.6)
Toxicology, alcohol-related disease	122 (2.4)
Psychiatry	107 (2.1)
Multidisciplinary stays and known disease follow-up	89 (1.8)
Obstetrics	51 (1.0)
Infectiology	44 (0.9)
Gynecology	38 (0.8)
Other (chronic pain, ophtalmology, complex trauma, burn injury)	32 (0.6)

the presence of "capillary" in the ED health record was associated with the presence of a standardized vital parameters surveillance chart (which included the measure of capillary glycaemia) and in turn influenced the probability of a long stay. A secondary analysis measured the performance of the two models for LOS prediction of ICU patients. The training set of 378 ICU patients was used to train the model, which was tested on the

Table 2 Model performance for the "structured-data only" and "unstructured-data added" feature sets

	Structured	Unstructured	Difference (pts)	All features
Recall	77.3%	77.1%	−0.19	76.6%
Specificity	70.4%	72.7%	2.31	71.1%
Precision	74.2%	75.7%	1.48	74.4%
Accuracy	74.1%	75.0%	1.0	75.0%
F1 Score	75.7%	76.4%	0.68	75.5%

remaining 76 patients. In this analysis, the unstructured data model achieved better accuracy than the structured data model (76.3% versus 75.0%) (Table 3). Feature importance for the two models limited to intensive care stays are presented in Additional file 1: Figure S4 and Figure S5 (Appendix 1).

Discussion

This study showed that UMLS-based one-hot vector word-embedding within an affirmative patient-centric context from EHRs is an effective way to predict LOS

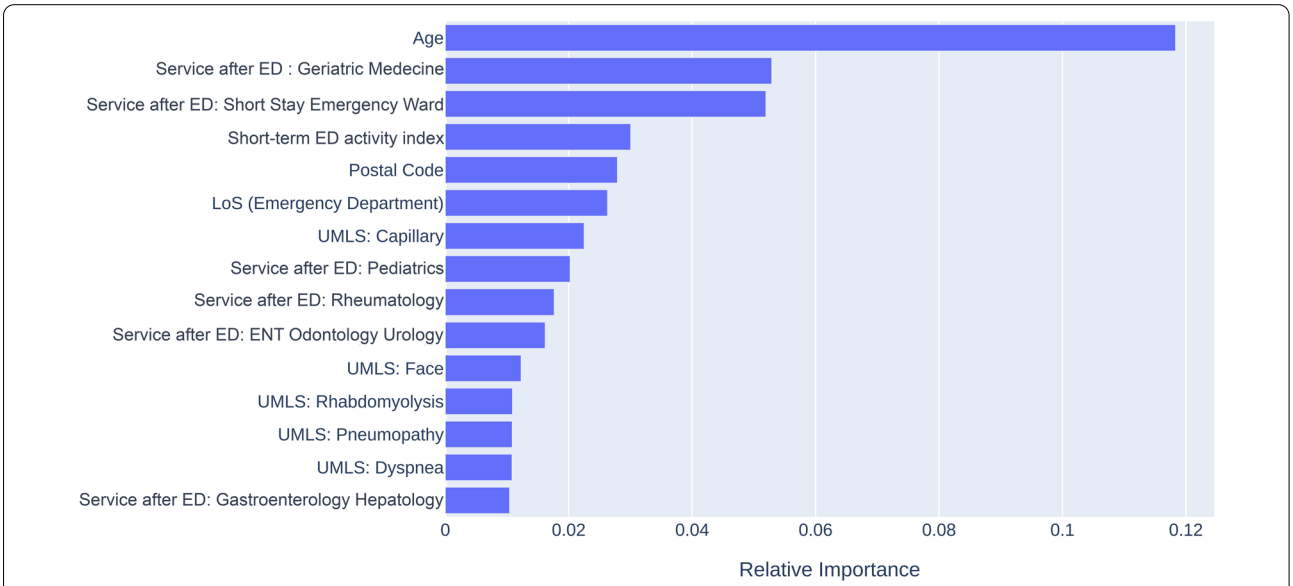

Fig. 1 Feature importance for the unstructured data model to predict hospital length of stay (LOS)

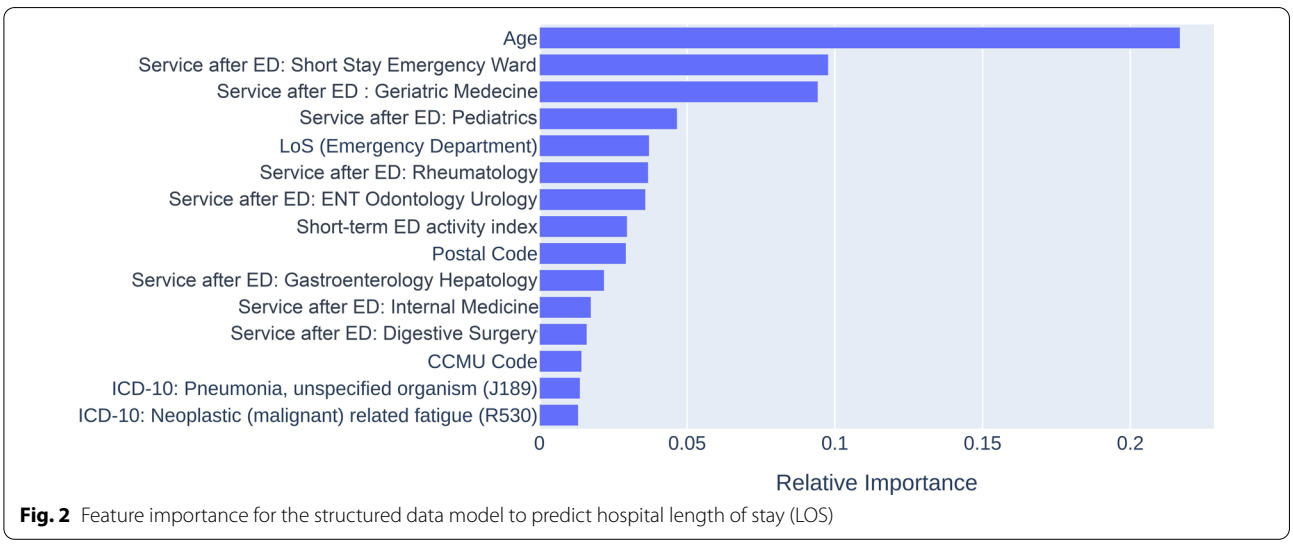

Fig. 2 Feature importance for the structured data model to predict hospital length of stay (LOS)

Table 3 Model performance of structured and unstructured data to predict hospital length of stay (LOS) when trained on intensive care unit stays

	Structured data	Unstructured data	Difference (points)	All features
Recall	77.6%	75.9%	− 1.72	84.5%
Specificity	66.7%	77.8%	11.11	72.2%
Precision	88.2%	91.7%	3.43	90.7%
Accuracy	75.0%	76.3%	1.32	76.3%
F1 score	82.6%	83.1%	0.49	87.5%

at an ED when using machine learning (random forests). The accuracy of the model using unstructured data was similar to the accuracy obtained using structured data. Therefore, this shows that unstructured data should also be considered in its use to obviate the need for resource-intensive data abstracting conducted by humans. The accuracy remained adequate despite the exclusion of very short stays, which could be easier to predict in some cases. Even though the increase in accuracy when unstructured data was used was small, it should be noted that this data set did not contain any of the structured diagnostic information (ICD-10 codes) of the structured model.

Moreover, the unstructured data model performed similarly or better than the structured data model for intensive care stays. ICU patients often have a highly standardized management that involves numerous medical examinations and procedures. Data pertaining to these elements are often recorded in the patient's EHR and thus contains relevant information for the determination of LOS.

We used random forest models to predict LOS at the ED since this model is well suited for treating data with complex interactions between variables and other nonlinear effects. Models based on deep neural networks are another option that could be explored in further studies. Such models have been used to predict admission or discharge of ED patients with better F1-score performance than logistic regression (31), although the obtained F1-score of 0.674 seems low compared to our findings.

In the literature, Roquette et al. used deep neural networks with their *text2num* embedding method (in the context of pediatric ED prediction admission using unstructured text data) and obtained results very similar to ours with a recall and specificity of approximately 80% and a 1.8 point increase in the Area Under the Curve after adding unstructured data [32]. However, in this design it could have been possible that the endpoint was in some cases directly encoded in the training data in the form of emergency physician recommendations regarding admission or discharge.

In another study by Zhang et al. [33] unstructured text improved predictions only when used in conjunction with structured data. Joseph [34] used free text to identify critically ill patients, enabling an increase in the Area Under Curve compared to structured only data models (with an AUC of 0.851 [95% CI: 0.849 to 0.852]). Choi [35] used random forests and gradient boosting to predict ED triage status and enriched their model with free text nursing triage notes. Both models had comparable performance with an Area Under Curve of 0.92 and in each case the best performance was achieved after the addition of text data. Other studies processed unstructured text data in an automated manner to make healthcare predictions regarding mortality [36], disease association patterns [37, 38], or risk areas in medication administration [39].

Regarding accuracy, the accuracy of predictions depends on the quantity and relevance of variables included. At our hospital, socio-economic status is not routinely extracted in health records and were not recorded in this study. Further research into unstructured text-mining methods could extract concepts relevant to this characteristic type. The use of unstructured data in predictive models based on generic, automated and replicable extraction pipelines is of primary interest for scalability purposes of such models on EHR systems, though this desirable scalability property comes with an additional technicality cost.

Two key limitations to this study were the high dimensionality of data and the signal-to-noise ratio within extracted semantic concepts. The first limitation was a common issue and could be tackled with regularization methods such as L1-penality (such as LASSO). The second was intrinsically linked to the retained extraction pipeline. In this study, we leveraged the well-established UMLS terminological system and extraction pipeline embedded in Dr Warehouse [25–27]. The UMLS meta-thesaurus is the richest collection of terminologies available with over 4.4 million medical concepts. This choice may guarantee cross-applications forecasting capabilities, however the signal to noise ratio may not be fully optimized for the specific LOS prediction problem. This observation motivates our pre-processing method using the symmetric relevance score. More sophisticated word-embedding methods could improve the performance of machine learning algorithms by using contextual information, including diagnostic hypotheses, patient comorbidity and patient history, to filter only the most relevant concepts and relations among them.

Although the sample size in our study was adequate, the addition of other centers could have enhanced the generalizability of our results. Single-center studies, however, provide locally actionable insights that could be used to inform quality improvement interventions and other hospitals could train similar models on their own data which could provide results tailored to their needs.

While LOS was considered as a categorical variable to maximise the power of the model, prediction of the LOS as a continuous variable could be a target for future studies. Only the presence of UMLS concepts are considered, and not the context surrounding these concepts, which might warrant investigation in future research.

Conclusions

This study shows that unstructured data (free text) can be used to predict LOS with acceptable predictive performance. The performance was similar to the performance of the model using structured data. Structured data, however, may have the drawback of being more time-consuming to extract. In many applications, unstructured text data contains valuable insights that are yet to be explored. As the methods to automatically extract knowledge evolve, they will undoubtedly give more accurate predictions. Modules to extract specific information like the primary complaint [40] or presence of pain [41] are currently being developed and could be combined or added to already existing software [42–44]. Future research needs to determine how these methods can ultimately improve healthcare outcomes while complying with privacy laws and maintaining high ethical standards.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12911-021-01722-4.

Additional file 1. Supplementary figures and tables related to model specifications, modeling parameters and modeling results.

Additional file 2. Data formats used for modeling.

Acknowledgements
The authors would like to thank Sarina Yaghobian from AcaciaTools for her medical writing and editing services.

Authors' contributions
Conceptualization: SS. Data acquisition: DL, SS. Statistical analysis: FG. Initial manuscript: JC, LR, FG. Revision for critical intellectual content: LR, DL, SS, AD. All authors approved the final manuscript.

Funding
None.

Availability of data materials
All data material can be accessed upon request to the last author Dr Stéphane Sanchez at the following email address stephane.sanchez@hcs-sante.fr.

Declarations

Ethics approval and consent to participate
The study was declared to the French registry of studies using healthcare data (Declaration No F20210719114017). Since the study was retrospective, purely observational, and was based on pseudonymized data, it was exempt from the Institutional Review Board approval according to the French Public Health Code (L1121-1, Law number 2012-300, 5 March 2012).

Competing interests
The authors would like to declare that they have no competing interest in relation to this study.

Author details
[1]Pôle Territorial Santé Publique et Performance, Centre Hospitalier de Troyes, 101 Avenue Anatole France CS 10718, 10003 Troyes Cedex, France. [2]Research and Consulting, CODOC SAS, 75008 Paris, France. [3]Research on Healthcare Performance Lab, INSERM U1290 RESHAPE, Université Claude Bernard Lyon 1, Villeurbanne, France. [4]Health Data Department, Hospices Civils de Lyon, Lyon, France.

Received: 24 July 2021 Accepted: 13 December 2021
Published online: 18 December 2021

References

1. Paterson R, MacLeod DC, Thetford D, Beattie A, Graham C, Lam S, et al. Prediction of in-hospital mortality and length of stay using an early warning scoring system: clinical audit. Clin Med Lond Engl. 2006;6(3):281–4.
2. Krochmal P, Riley TA. Increased health care costs associated with ED overcrowding. Am J Emerg Med. 1994;12(3):265–6.
3. Liew D, Liew D, Kennedy MP. Emergency department length of stay independently predicts excess inpatient length of stay. Med J Aust. 2003;179(10):524–6.
4. Chalfin DB, Trzeciak S, Likourezos A, Baumann BM, Dellinger RP, DELAY-ED study group. Impact of delayed transfer of critically ill patients from the emergency department to the intensive care unit. Crit Care Med. 2007;35(6):1477–83.
5. Kobayashi KJ, Knuesel SJ, White BbA, Bravard MA, Chang Y, Metlay JP, et al. Impact on length of stay of a hospital medicine emergency department boarder service. J Hosp Med. 2020;15(03):147–53.
6. García-Gigorro R, de la Cruz VF, Andrés-Esteban EM, Chacón-Alves S, Morales Varas G, Sánchez-Izquierdo JA, et al. Impact on patient outcome of emergency department length of stay prior to ICU admission. Med Intensiva. 2017;41(4):201–8.
7. Jain M, Damania D, Jain A, Kanthala A, Stead L, Jahromi B. Does prolonged length of stay in the emergency department affect outcome for stroke patients? West J Emerg Med. 2014;15(3):267–75.
8. Derose SF, Gabayan GZ, Chiu VY, Yiu SC, Sun BC. Emergency department crowding predicts admission length-of-stay but not mortality in a large health system. Med Care. 2014.
9. Driesen BEJM, van Riet BHG, Verkerk L, Bonjer HJ, Merten H, Nanayakkara PWB. Long length of stay at the emergency department is mostly caused by organisational factors outside the influence of the emergency department: a root cause analysis. PLoS ONE. 2018;13(9):e0202751.
10. Vicendese D, Marvelde LT, McNair PD, Whitfield K, English DR, Taieb SB, et al. Hospital characteristics, rather than surgical volume, predict length of stay following colorectal cancer surgery. Aust N Z J Public Health. 2020;44(1):73–82.
11. Tsai P-F (Jennifer), Chen P-C, Chen Y-Y, Song H-Y, Lin H-M, Lin F-M, et al. Length of hospital stay prediction at the admission stage for cardiology patients using artificial neural network. J Healthc Eng. 2016;2016:1–11.
12. Robinson GH, Davis LE, Leifer RP. Prediction of hospital length of stay. Health Serv Res. 1966;1(3):287–300.

13. Durstenfeld MS, Saybolt MD, Praestgaard A, Kimmel SE. Physician predictions of length of stay of patients admitted with heart failure. J Hosp Med. 2016;11(9):642–5.
14. Lucas BP, Kumapley R, Mba B, Nisar I, Lee K, Ofori-Ntow S, et al. A hospitalist-run short-stay unit: features that predict length-of-stay and eventual admission to traditional inpatient services. J Hosp Med. 2009;4(5):276–84.
15. Baeza FL, da Rocha NS, Fleck MP, Baeza FL, da Rocha NS, Fleck MP. Predictors of length of stay in an acute psychiatric inpatient facility in a general hospital: a prospective study. Braz J Psychiatry. 2018;40(1):89–96.
16. Ismail Z, Arenovich T, Grieve C, Willett P, Sajeev G, Mamo DC, et al. Predicting hospital length of stay for geriatric patients with mood disorders. Can J Psychiatry. 2012;57(11):696–703.
17. Wolff J, McCrone P, Patel A, Kaier K, Normann C. Predictors of length of stay in psychiatry: analyses of electronic medical records. BMC Psychiatry. 2015;15(1):238.
18. Wilding D, Evans K. Predicting length of stay for acute medical admissions using the ALICE score: a simple bedside tool. Acute Med. 2017;16(2):60–4.
19. Barnes S, Hamrock E, Toerper M, Siddiqui S, Levin S. Real-time prediction of inpatient length of stay for discharge prioritization. J Am Med Inform Assoc. 2016;23(e1):e2-10.
20. Turgeman L, May JH, Sciulli R. Insights from a machine learning model for predicting the hospital Length of Stay (LOS) at the time of admission. Expert Syst Appl. 2017;78:376–85.
21. Pendharkar PC, Khurana H. Machine learning techniques for predicting hospital length of stay in pennsylvania federal and specialty hospitals. Int J Comput Sci Appl. 2014;11(3):45–56.
22. Malmasi S, Hosomura N, Chang L-S, Brown CJ, Skentzos S, Turchin A. Extracting healthcare quality information from unstructured data. AMIA Annu Symp Proc. 2018;16(2017):1243–52.
23. Soibelman L, Wu J, Caldas C, Brilakis I, Lin K-Y. Management and analysis of unstructured construction data types. Adv Eng Inform. 2008;22(1):15–27.
24. Chen R, Ho JC, Lin J-MS. Extracting medication information from unstructured public health data: a demonstration on data from population-based and tertiary-based samples. BMC Med Res Methodol. 2020;20(1):1–11.
25. Garcelon N, Neuraz A, Salomon R, Faour H, Benoit V, Delapalme A, et al. A clinician friendly data warehouse oriented toward narrative reports: Dr. Warehouse J Biomed Inform. 2018;80:52–63.
26. Garcelon N, Neuraz A, Salomon R, Bahi-Buisson N, Amiel J, Picard C, et al. Next generation phenotyping using narrative reports in a rare disease clinical data warehouse. Orphanet J Rare Dis. 2018;13(1):85.
27. Garcelon N, Neuraz A, Benoit V, Salomon R, Burgun A. Improving a full-text search engine: the importance of negation detection and family history context to identify cases in a biomedical data warehouse. J Am Med Inform Assoc. 2017;24(3):607–13.
28. Lan M, Tan CL, Su J, Lu Y. Supervised and traditional term weighting methods for automatic text categorization. IEEE Trans Pattern Anal Mach Intell. 2009;31(4):721–35.
29. Derose SF, Gabayan GZ, Chiu VY, Yiu SC, Sun BC. Emergency department crowding predicts admission length-of-stay but not mortality in a large health system. Med Care. 2014;52(7):602–11.
30. Cheng C-H, Chen H-H. Sentimental text mining based on an additional features method for text classification. PLOS ONE. 2019;14(6):e0217591.
31. Chen C-H, Hsieh J-G, Cheng S-L, Lin Y-L, Lin P-H, Jeng J-H. Emergency department disposition prediction using a deep neural network with integrated clinical narratives and structured data. Int J Med Inf. 2020;139:104146.
32. Roquette BP, Nagano H, Marujo EC, Maiorano AC. Prediction of admission in pediatric emergency department with deep neural networks and triage textual data. Neural Netw. 2020;126:170–7.
33. Zhang X, Kim J, Patzer RE, Pitts SR, Patzer A, Schrager JD. Prediction of emergency department hospital admission based on natural language processing and neural networks. Methods Inf Med. 2017;56(05):377–89.
34. Joseph JW, Leventhal EL, Grossestreuer AV, Wong ML, Joseph LJ, Nathanson LA, et al. Deep-learning approaches to identify critically Ill patients at emergency department triage using limited information. J Am Coll Emerg Physicians Open. 2020;1(5):773–81.
35. Choi SW, Ko T, Hong KJ, Kim KH. machine learning-based prediction of Korean triage and acuity scale level in emergency department patients. Healthc Inform Res. 2019;25(4):305.
36. Ye J, Yao L, Shen J, Janarthanam R, Luo Y. Predicting mortality in critically ill patients with diabetes using machine learning and clinical notes. BMC Med Inform Decis Mak. 2020;20(S11):295.
37. Mullins IM, Siadaty MS, Lyman J, Scully K, Garrett CT, Greg Miller W, et al. Data mining and clinical data repositories: Insights from a 667,000 patient data set. Comput Biol Med. 2006;36(12):1351–77.
38. Chen ES, Sarkar IN. Mining the electronic health record for disease knowledge. Methods Mol Biol Clifton NJ. 2014;1159:269–86.
39. Härkänen M, Paananen J, Murrells T, Rafferty AM, Franklin BD. Identifying risks areas related to medication administrations—text mining analysis using free-text descriptions of incident reports. BMC Health Serv Res. 2019;19(1):1–9.
40. Tootooni MS, Pasupathy KS, Heaton HA, Clements CM, Sir MY. CCMapper: An adaptive NLP-based free-text chief complaint mapping algorithm. Comput Biol Med. 2019;113:103398.
41. Vu T, Nguyen A, Brown N, Hughes J. Identifying Patients with Pain in Emergency Departments using Conventional Machine Learning and Deep Learning. In: Proceedings of the The 17th Annual Workshop of the Australasian Language Technology Association [Internet]. Sydney, Australia: Australasian Language Technology Association; 2019 [cited 2021 Mar 29]. p. 111–9. Available from: https://www.aclweb.org/anthology/U19-1015
42. Sohn S, Clark C, Halgrim SR, Murphy SP, Chute CG, Liu H. MedXN: an open source medication extraction and normalization tool for clinical text. J Am Med Inform Assoc JAMIA. 2014;21(5):858–65.
43. Thomas A, Sangeetha S. An innovative hybrid approach for extracting named entities from unstructured text data. Comput Intell. 2019;35(4):799–826.
44. Scheurwegs E, Luyckx K, Luyten L, Daelemans W, Van den Bulcke T. Data integration of structured and unstructured sources for assigning clinical codes to patient stays. J Am Med Inform Assoc JAMIA. 2016;23(e1):e11–9.

Publisher's Note