Improving mortality prediction for intensive care unit patients using text mining techniques

Primož Kocbek[1], Nino Fijačko[1], Milan Zorman[2], Simon Kocbek[1,3], Gregor Štiglic[1,2]

[1]Univerza v Mariboru Fakulteta za zdravstvene vede, Maribor, Slovenija, +386 2 300 47 00

[2]Univerza v Mariboru Fakulteta za elektrotehniko, računalništvo in informatiko, Maribor, Slovenija, +386 2 220 7000

[3]Kinghorn Centre for Clinical Genomics, Garvan institute of Medical Research, Sydney, Australia, +61 (02) 9295 8100

{primoz.kocbek, nino.fijacko, milan.zorman, gregor.stiglic}@um.si, skocbek@gmail.com

ABSTRACT

Numerous severity assessment scores for estimation of in-hospital mortality in Intensive Care Unit (ICU) have been developed over the last 40 years. In this study, we predicted 1-month mortality in chronic kidney disease (CKD) patients using the open Medical Information Mart for Intensive Care III (MIMIC III) database. Additionally, we observed the improvement in predictive performance and interpretability of the baseline model used in ICUs to a more complex model using simple features such as unigrams or bigrams, as well as advanced features extracted from textual nursing notes. For the latter, MetaMap extraction tool was used to extract medical concepts based on the Unified Medical Language System (UMLS) terminology. We used a logistic regression based classifier, built using Simplified Acute Physiology Score II (SAPS II), age and gender, as a baseline model. The baseline model was then compared to regularized logistic regression based classifier built using simple and more complex additional features. The Area Under the ROC Curve (AUC) results for the baseline predictive performance improved from of 0.761 to 0.782 when frequency of unigrams and bigrams were used to build the model. In a similar scenario, where unigram and bigram frequency was replaced with Term Frequency–Inverse Document Frequency (TF-IDF) based feature values, AUC further increased to 0.786.

This paper represents an opportunity in extracting new knowledge in the form of unigrams, bigrams or concepts extracted from textual notes accompanied by regression coefficient values that can be interpreted as relations between the features and the outcome. The combination of both can provide added value in decision support systems in ICU departments, where data is collected in electronic medical records (EMRs) in real-time.

Categories and subject descriptors

H.2.6 [Information Systems]: Database Machines

H.2.8 [Information Systems]: Database Applications

General Terms

Algorithms, Measurement, Documentation, Performance, Reliability, Experimentation.

Keywords

Text mining, ICU, database, machine learning, mortality prediction.

1. INTRODUCTION

Predicting the mortality of ICU patients is a complex and dynamic process. Critical illness severity assessment scores, such as Acute Physiology and Chronic Health Evaluation I-IV (APACHE), Sequential Organ Failure Assessment score (SOFA), Mortality Probability Model I-III (MPS), or Simplified Acute Physiology Score I-III (SAPS), help clinicians detect patient problems earlier, thus providing a better holistic treatment for patients and making patient care more cost-effective. The sheer number of different severity scores used is partly because of the quality of recorded data needed to calculate them. An example of such severity score is APACHE IV, which tends to have the best discriminative performance but the data needed to compute the score is complex and hospitals would need to develop a good enough high-quality database for analysis of risk stratification [1, 2].

The MIMIC III database [3], a free public-access intensive care unit repository, is widely used for predicting the mortality of ICU patients, where developers provided several severity scores for the database (e.g., OASIS, SAPS, SAPS II, SOFA), but also noted that for APACHE IV the coding of the diagnostic component is difficult and might lack accuracy [4]. Part of MIMIC III are free text nursing notes, which represent a good candidate source of information for mortality risk prediction, as they contain a detailed and regularly-updated record of the interventions performed, medications administered, vital signs, and physical examination findings, all of which carry highly specific information about the patient's dynamic physiological state and eventual outcome [5]. Because such data is unstructured, our purpose is knowledge discovery where we observe the improvements in predictive performance and interpretability of predictive models based on additional features extracted from nursing notes collected in EMRs. More precisely, we aim to predict one month mortality in CKD (ICD 9 code 585.x) patients and compare the improvement of the baseline model performance by including simple features such as unigrams or bigrams as well as more advanced features extracted from text in the form of medical concepts using MetaMap [6] to define mapping between textual notes and UMLS terminology.

2. METHODS

The data were obtained from the MIMIC III database, version 1.3, to select 58,976 hospitalizations for 46,520 patients who stayed in critical care units of the Beth Israel Deaconess Medical Center in Boston between 2001 and 2012. The database included 26 linked tables, which can be merged mostly by patient or hospitalization identification numbers. Our focus were nursing notes (i.e., free text

notes from patients with CKD diagnosis), where we excluded hospitalization of patients that died within 24 hours of admission and nursing notes that were not fully updated, where duplication of data was likely. That left us with 10,867 nursing notes from 4,381 hospitalizations. The first nursing note was taken on average 7.8 hours after admission (85.2 % hospitalizations have at least two notes), second taken on average 14.7 hours after admission (38.1 % had at least 3 taken) and the third one was taken on average 17.5 hours after admission. A slight majority of hospitalized patients were male (59.4 %), with an average age of 65.6 (Standard Deviation (SD) 15.2)) and a 13.4 % mortality rate (death during or up to one month after the hospitalization was recorded).

Developers of the database also included source code for calculation of several severity scores (i.e., OASIS, SAPS, SAPS II and SOFA), and we selected SAPS II as the main feature in the baseline model, since it is used on daily bases in hospitals. The input features of our baseline model consisted of SAPS II score, age and gender.

Nursing notes were initially processed using traditional text extraction algorithms with stemming and removal of stop words, which produced 51,680 unique unigrams and 363,055 unique bigrams. Both frequency and TF-IDF tables were prepared. The text from the nursing notes was also processed using the MetaMap tool from the US National Library of Medicine. MetaMap identifies and normalizes biomedical terminology from the Unified Medical Language System (UMLS). Binary representations of Bags of Phrases (BOP) identified by MetaMap, and their UMLS (Concept Unique Identifiers) CUIs were used as features for the classifier. Space characters in phrases were replaced with underline character. Word sense disambiguation was used to distinguish similar words with same structure. In addition, phrases were marked with whether the associated concepts were found in a positive or negative context. To identify the polarity of phrases (negative or positive), the NegEx module [7] of MetaMap was enabled in order to identify the polarity of phrases (negative or positive). NegEx implements a simple algorithm that contains several regular expressions indicating negation, filters out sentences containing phrases that falsely appear to be negation phrases, and limits the scope of the negation phrases.

For better understanding, we provide a short example of MetaMap-annotated phrases from part of the sentence "URINE MICROSCOPY" (Table 1).

Table 1. Example of MetaMap-annotated phrases from part of the sentence "URINE MICROSCOPY"

Meta Candidates	
Score	**Matched concept**
1000	C0430397: Urine microscopy (Microscopic urinalysis) [Laboratory Procedure]
861	C0026018: Microscopy [Laboratory Procedure]
789	C0205288: Microscopic [Qualitative Concept]
694	C0042036: Urine [Body Substance]
694	C0042037: Urine (In Urine) [Functional Concept]
694	C2963137: Urine (Portion of urine) [Body Substance]
Meta Mapping	
Score	**Matched concept**
1000	C0430397: Urine microscopy (Microscopic urinalysis) [Laboratory Procedure]

The *meta candidates* are all discovered mappings that are ordered according to an evaluation metric described in [7], while *meta mappings* represent the selected phrases which finally represent features in our model. Please note that several meta mappings may be found in a sentence. Parentheses contain the concept's preferred name while square brackets contain the concept's semantic type.

One of our goals was interpretability and avoidance of over-fitting, therefore we restricted model building to regularized linear models, further we narrowed the selection to L1 regularization models or least absolute shrinkage and selection operator (lasso), which includes feature selection functionality which was needed due to high number of features in our datasets [8]. We expanded on the work of Marafino et al. [4], where they used the MIMIC II dataset and predicted mortality via stochastic gradient descent-based classifiers with TF-IDF on the extracted unigrams and bigrams for patients that died during the given ICU stay. All experiments were implemented in R language and environment for statistical computing [9] using glmnet package [10] to build and validate predictive models.

3. RESULTS

Results presented in this section were obtained from four scenarios where we compared different combinations of two types of extracted features (n-grams versus concepts) and two types of the extracted feature values (frequency vs. TF-IDF).

The baseline classifier with basic features (SAPS II score, age and gender) to predict mortality was used to evaluate the performance gain when more complex classifiers were built. Initially, we were interested in measuring the improvement of the baseline predictive performance by adding unigram and bigram features to SAPS II, age and gender of the patients. At the same time, we were observing the complexity of the predictive models by observing the number of features that were included in the models.

Second row in Table 2 presents the results when frequency of unigrams and bigrams were used to build the model. It can be seen that baseline predictive performance improved from the AUC of 0.761 to 0.782 when frequency information of unigrams and bigrams was used to build the model. With an improvement of more than 2% in AUC it is also important to note that in this scenario we obtained the simplest models in terms of interpretation with only 9 features on average. In a similar scenario where unigram and bigram frequency was replaced with TF-IDF based feature values resulted in further improvement with AUC of 0.786. In the next two experiments, we replaced unigrams and bigrams with UMLS concepts that were extracted from free text (nursing notes). Table 2 demonstrates further improvement of the classification performance as the AUC in case of TF-IDF increased to 0.789, while even further improvement with a mean AUC of 0.791 was measured in frequency based features experiment. Tables 3 and 4 provide more detailed overview of selected features along with the number of times a feature was included in a predictive model during 100 cross-validation runs. It can be seen that TF-IDF produced predictive models with a larger number of features and therefore represents a richer set of concepts that can be used to warn a medical expert of a potential threat to a patient. On the other hand, a complex model (in case of TF-IDF experiment, more than 35 features were used in a model on average) might represent a challenge for medical experts when interpretation of models is needed.

4. DISCUSSION AND CONCLUSIONS

In this paper, we observed the improvements in predictive performance and interpretability of predictive models based on new features extracted from nursing notes collected in EMRs. More precisely, we predicted one month mortality, at the end of 24-hours spent in the ICU, for CKD patients. The improvement of the

Table 2. Summary of predictive performance measures for different experiments using features extracted from nursing notes

	AUC	Sensitivity	Specificity	PPV	NPV	Selected features
Baseline (SAPS II)	0.761 [0.757-0.766]	0.712 [0.704-0.720]	0.687 [0.680-0.695]	0.283 [0.277-0.288]	0.933 [0.931-0.935]	1.0 [1.0-1.0]
Unigrams and bigrams (frequency)	0.782 [0.778-0.786]	0.727 [0.721-0.734]	0.714 [0.707-0.722]	0.306 [0.300-0.313]	0.939 [0.937-0.940]	9.1 [6.6-11.6]
UMLS concept mapping (frequency)	0.791 [0.787-0.795]	0.736 [0.728-0.745]	0.716 [0.708-0.724]	0.310 [0.304-0.316]	0.941 [0.939-0.943]	17.6 [14.8-20.5]
Unigrams and bigrams (TF-IDF)	0.786 [0.782-0.790]	0.733 [0.725-0.740]	0.712 [0.704-0.720]	0.306 [0.300-0.313]	0.940 [0.938-0.941]	25.1 [21.7-28.6]
UMLS concept mapping (TF-IDF)	0.789 [0.785-0.793]	0.747 [0.741-0.754]	0.700 [0.692-0.709]	0.302 [0.296-0.308]	0.942 [0.94-0.943]	35.5 [31.0-40.0]

baseline model (SAPS II, gender and age) in comparison to predictive models that included unigrams, bigrams as well as more advanced features extracted from text in the form of medical concepts using the MetaMap extraction tool was also observed.
The results show high level of predictive performance that can be compared to a similar study by Brabrand et al. [11] where it was shown that using clinical intuition of the admission staff produced comparable predictions in terms of AUC when identify patients at risk of dying. However, it has to be noted that Brabrand and colleagues did not focus on a specific group of patients.

Table 3. Frequency of specific features selected in the UMLS concept mapping (Frequency) experiment

Single_count_all_CKD_all_feat	N
DNR_(DNR_-_Do_not_resuscitate)_[Finding]	100
Map_(Functional_Map)_[Conceptual_Entity]	100
SAPSII	100
Meeting_(Meetings)_[Health_Care_Activity]	98
Coccyx_(Entire_coccyx)_[Body_Part_Organ_or_Organ_Component]	93
PICC_line_(Peripherally_inserted_central_catheter_(physical_object))_[Medical_Device]	86
Anuria_[Disease_or_Syndrome]	85
CMO_(Chronic_multifocal_osteomyelitis)_[Disease_or_Syndrome]	82
Family_[Family_Group]	70
vascular_(Blood_Vessel)_[Body_Part_Organ_or_Organ_Component]	51
Bilirubin_[Biologically_Active_SubstanceOrganic_Chemical]	45
Prognosis_(Forecast_of_outcome)_[Health_Care_Activity]	45
error_[Qualitative_Concept]	35
Necrotic_(Necrosis)_[Organ_or_Tissue_Function]	34
Pleural_effusion_(Pleural_effusion_fluid)_[Body_Substance]	28
Poor_prognosis_(Prognosis_bad)_[Finding]	28
Thick_[Qualitative_Concept]	28
Hypotensive_[Pathologic_Function]	27
dysfunction_(physiopathological)_[Functional_Concept]	25
Brain_[Body_Part_Organ_or_Organ_Component]	23

When providing the predictive models for healthcare experts to support their work in clinical practice, we should also pay attention to interpretability of such models. As already noted in [12] in case of similar predictive performance on training set, the simplest models often also perform the best on the test set. Therefore, we should also take the complexity of models with similar performance into account. In case of our study, the complexity of the four proposed models ranges from 9 up to approximately 35 selected features. In case of both unigram and bigram based models, it would perhaps make sense to use the simpler model as it does not significantly differ in predictive performance at a significantly lower complexity of the simpler, frequency based model.

Table 4. Frequency of specific features selected in the UMLS concept mapping (TF-IDF) experiment

Single_tfidf_all_CKD_all_feat	N
DNR_(DNR_-_Do_not_resuscitate)_[Finding]	100
Meeting_(Meetings)_[Health_Care_Activity]	100
SAPSII	100
CMO_(Chronic_multifocal_osteomyelitis)_[Disease_or_Syndrome]	98
PICC_line_(Peripherally_inserted_central_catheter_(physical_object))_[Medical_Device]	97
Family_[Family_Group]	96
Coccyx_(Entire_coccyx)_[Body_Part_Organ_or_Organ_Component]	88
Heels_(Heel)_[Body_Location_or_Region]	88
Pressors_[Pharmacologic_Substance]	83
Map_(Functional_Map)_[Conceptual_Entity]	74
Levophed_[Organic_ChemicalPharmacologic_Substance]	73
loosen_(Loosening)_[Functional_Concept]	70
neg_DNR_(DNR_-_Do_not_resuscitate)_[Finding]	68
Worsening_(Worse)_[Qualitative_Concept]	68
error_[Qualitative_Concept]	67
dysfunction_(physiopathological)_[Functional_Concept]	64
Bilirubin_[Biologically_Active_SubstanceOrganic_Chemical]	62
Anasarca_[Pathologic_Function]	59
Coccyx_(Bone_structure_of_coccyx)_[Body_Part_Organ_or_Organ_Component]	51
Poor_prognosis_(Prognosis_bad)_[Finding]	50

From the most frequently selected features (Table 3 and 4) we can observe some very general concepts, like “do not resuscitate

(DNR)" and high SAPS II score, indicating higher mortality. Also, family related concepts can be easily interpreted by a fact that physicians usually call family members to discuss the severity of the situation, especially when the situation is critical or life threatening. Additional features indicating higher mortality are concepts related to bones like "coccyx" and "heel", which could indicate specific problems related to calcification, frequently observed in CKD patients. Medical terms such as "peripherally inserted central catheter", "chronic multifocal osteomyelitis" and "use of pressors (pharmacologic_substance)" could be interpreted as signs of worsening health situation. We plan to investigate these conclusions in more detail in future work. It is also interesting to note that the UMLS frequency model's most selected variable was the medical term "Anuria (non-passage or less than 100 milliliters passage of urine a day)", which was not selected in the UMLS TF-IDF model.

Further development of our model will include extensions where a shorter (e.g., 6 or 12 hours) period would be used to provide "early warning" signal to healthcare experts working in the ICU. Additional features can be extracted from MIMIC-III that would further improve the predictive performance and possibly also the interpretability of the models.

5. ACKNOWLEDGEMENT

The authors would like to acknowledge financial support from the Slovenian Research Agency (research core funding No. P2-0057 and bilateral grant ARRS-BI-US/16-17-064).

6. REFERENCES

[1] Keegan, M. T., Gajic, O., and Afessa, B. 2012. Comparison of APACHE III, APACHE IV, SAPS 3, and MPM0III and Influence of Resuscitation Status on Model Performance. *Chest.* 142, 4 (April 2012), 851–858. DOI= 10.1378/chest.11-2164.

[2] Wu, V. C., Tsai, H. B., Yeh, Y. C., Huang, T. M., Lin, Y. F., Chou, N. K., ... and Wu, M. S. 2010. Patients supported by extracorporeal membrane oxygenation and acute dialysis: acute physiology and chronic health evaluation score in predicting hospital mortality. *Artificial organs*. 34, 10 (May 2010), 828–835. DOI= 10.1111/j.1525-1594.2009.00920.x.

[3] Johnson, A.E., Pollard, T.J., Shen, L., Lehman, L.W.H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L.A. and Mark, R.G., 2016. MIMIC-III, a freely accessible critical care database. Scientific data, 3.

[4] Marafino, B. J., Boscardin, W. J., and Dudley, R. A. 2015. Efficient and sparse feature selection for biomedical text classification via the elastic net: Application to ICU risk stratification from nursing notes. *Journal of biomedical informatics*. 54 (April 2015), 114–120. DOI= 10.1016/j.jbi.2015.02.003.

[5] MIT Laboratory for Computational Physiology-mimic-code: 2017. https://github.com/MIT-LCP/mimic-code/tree/master/concepts/severityscores. Accessed: 2017- 09- 04.

[6] Metamap: Mapping text to the umls metathesaurus. 2006. https://pdfs.semanticscholar.org/e262/a22134cca0e484be1095160cc4ec8d9e7624.pdf. Accessed: 2017- 09- 04.

[7] Chapman, W. W., Bridewell, W., Hanbury, P., Cooper, G. F., and Buchanan, B. G. 2001. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics*. 34, 5, (October 2001), 301–310. DOI= 10.1006/jbin.2001.1029.

[8] Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), pp.267-288.

[9] R Core Team, 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

[10] Friedman, J., Hastie, T. and Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. Journal of statistical software, 33(1), p.1.

[11] Brabrand, M., Hallas, J., and Knudsen, T. 2014. Nurses and physicians in a medical admission unit can accurately predict mortality of acutely admitted patients: a prospective cohort study. *PloS one*, 9, 7, (July 14), e101739. DOI= 10.1371/journal.pone.0101739.

[12] Stiglic, G., Kocbek, S., Pernek, I., and Kokol, P. 2012. Comprehensive decision tree models in bioinformatics. *PloS one*, 7, 3, (March 30), e33812. DOI= 10.1371/journal.pone.0033812