Combining Structured and Free-text Electronic Medical Record Data for Real-time Clinical Decision Support

Emilia Apostolova[1], Tony Wang[2], Ioannis Koutroulis[3], Tim Tschampel[4], Tom Velez[4]

[1] Language.ai, Chicago, IL *emilia@language.ai*

[2] Imedacs, Ann Arbor, MI *xwang@imedacs.com*

[3] Children's National Health System, Washington, DC *ikoutrouli@childrensnational.org*

[4] Computer Technology Associates, Ridgecrest, CA *tim.tschampel@cta.com, tom.velez@cta.com*

Abstract

The goal of this work is to utilize Electronic Medical Record (EMR) data for real-time Clinical Decision Support (CDS). We present a deep learning approach to combining in real time available diagnosis codes (ICD codes) and free-text notes: *Patient Context Vectors*. Patient Context Vectors are created by averaging ICD code embeddings, and by predicting the same from free-text notes via a Convolutional Neural Network. The Patient Context Vectors were then simply appended to available structured data (vital signs and lab results) to build prediction models for a specific condition. Experiments on predicting ARDS, a rare and complex condition, demonstrate the utility of Patient Context Vectors as a means of summarizing the patient history and overall condition, and improve significantly the prediction model results.

1 Introduction

A key goal in critical care medicine is the early identification and timely treatment of rapidly progressive, life-threatening conditions, such as Sepsis, Septic Shock, and Acute Respiratory Distress Syndrome (ARDS). Such life-threatening conditions, are both rare, and at the same time, complex and heterogeneous, involving the interaction of multiple risk factors, comorbidities, and current symptoms. Hospital alert systems typically rely on screening of structured data such as vital signs and lab results, and, in the case of such rare conditions, are often associated with "alert fatigue" and require manually entered clinical judgement.

The information needed for a reliable risk evaluation of such rare and complex conditions is typically dispersed across the patient EMR, and available at different times throughout the patient stay. The patient demographics, past medical and visit history, chronic conditions, risk factors, current signs and symptoms can be found in the form of clinical notes (e.g. nursing notes, radiology reports, etc.), diagnosis and procedure codes, vital signs, lab orders and results. The challenge of real-time CDS systems is the variability and the availability of real-time EMR data, resulting from different charting behaviors, health care delivery models, hospital settings, etc.

The goal of this work is to utilize all available EMR patient information for real-time predictive modelling. While our experiments are focused on identifying ARDS cases, the described method is applicable to a variety of use cases needing information dispersed across the EMR patient record. The primary contribution of this work is the use of low-dimensional representation of the patient's history, current symptoms and conditions, which we refer to as *Patient Context Vector*. At prediction time, Patient Context Vectors are generated from the combination of available up-to-date ICD codes (if any) and available nursing notes. Patient Context Vectors (vectors of real numbers) are then simply added to the list of existing structured data variables (vital signs and lab results) and used to identify patients at risk of developing life-threatening conditions that require rapid intervention.

2 Method

In this work, we combine ICD codes, clinical notes, vital signs, lab results, and demographic information to build a real-time ARDS prediction model. Low-dimensional representation of ICD codes (ICD embeddings) is generated from a large corpus of patient ICD records. Patient visit EMR data is used to look up recorded up-to-date ICD codes, clinical notes, vital signs, and lab results. The visit ICD codes are converted to embeddings and averaged to produce Patient Context Vectors.

Pertinent patient information might not be necessarily "ICD-coded" during prediction time, but

Proceedings of the BioNLP 2019 workshop, pages 66–70
Florence, Italy, August 1, 2019.

can be available in the form of nursing notes. A deep learning model was trained to predict the patient's Patient Context Vector from nursing notes. The Patient Context Vectors obtained from available in the system ICD codes, and from free-text notes are then used in conjunction with vital signs, and lab results to predict the patient's outcome. Details for each step of the approach are provided in subsequent sections.

2.1 Dataset

We utilized the freely available database comprising deidentified health-related data associated with over 40,000 patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012: the MIMIC3 Intensive Care Unit (ICU) database (Johnson et al., 2016). The dataset contains over 2 million free-text clinical notes and over 650,000 diagnosis codes for over 58,000 visits. Included ICUs are medical, surgical, trauma-surgical, coronary, and cardiac surgery recovery units. EMR data includes vital signs, laboratory results, diagnosis codes, free text nursing notes, radiology reports, medications, discharge summaries, treatments, etc.

2.2 ICD Embeddings and Patient Context Vectors

Clinicians viewing properly coded patient diagnosis codes (ICD9 and ICD10 codes[1]) are typically capable of deducing the overall condition, history, and risk factors associated with a patient. Intuitively, the totality of patient's diagnosis codes represent a meaningful medical summary of the patient. Diagnosis codes are used to describe both current diagnoses (e.g. *Community-acquired Pneumonia*), but also a variety of additional facts. For example, ICD codes can describe patient's history and chronic conditions (e.g. *Chronic kidney disease; Personal history of traumatic fracture;* etc.); information regarding past and current treatments and procedures (e.g. *Infection due to other bariatric procedure*). In some cases, ICD codes contain information such as the patient age group (e.g. *Sepsis of newborn; Elderly multigravida*); expected outcome (*Encounter for palliative care*); patient's social history (e.g. *Adult emotional/psychological abuse*); the reason for the visit, (e.g. *Railway accidents; Motor Vehicle accidents*, etc).

While there are a large number of ICD codes (around 15,000 ICD9 codes and around 68,000 ICD10 codes), they tend to be interdependent, and to co-occur. For example, *Pneumonia* ICD codes are often accompanied with ICD codes describing *Cough, Fever, Pleural effusion*, etc. Inspired by word embeddings (Mikolov et al., 2013), it has been suggested that this medical code co-occurence can be exploited to generate low-dimensional representations of ICD codes: ICD Embeddings (Choi et al., 2016b,a; Kartchner et al., 2017).

All available MIMIC3 patient data was used to generate the ICD embeddings following the approach of (Choi et al., 2016b). In our approach, we attempted to generate a low-dimensional representation of the patient history, symptoms, risk factors, diagnosis, etc, by averaging the patient ICD code embeddings (creating Patient Context Vectors). The optimum size of the vectors was determined to be 50.

2.3 Predicting Patient Context Vectors from Clinical Texts

While averaged ICD embeddings appear to be a useful summary of the overall patient history, condition, symptoms, and risk factors, ICD code data is not necessarily available for real-time CDS systems. Some ICD codes associated with patients' history and symptoms might be entered early on in the EMR system. However, diagnosis ICD codes are typically obtained after tests and lab results and might not be available during prediction time. Similarly, not all relevant patient history and symptoms are necessarily ICD-coded.

At the same time, nursing notes typically contain all currently available information, even if not present in the form of ICD codes. Nursing notes include information such as past medical history, reason for visit, current symptoms, summary of test outcomes, etc.

In order to capture information present in free-text notes, we also built a word-level CNN model that predicts the patient Patient Context Vector from the note text. The model was trained on available nursing and discharge notes and achieved a mean squared error of 0.179 on the validation set. The network was trained on 1,081,176 free-text notes, with pre-trained word-embeddings of size 100. The texts were truncated/padded to the 90th percentile length (785 tokens). The network consists of a Convolutional, Max Pooling layers, followed by 2 hidden layers of size 500. The last layer uses linear activation with loss func-

[1] The International Classification of Diseases, ©The World Health Organization.

tion of mean squared error to predict the Patient Context Vector[2].

2.4 Patient Context Vectors in Prediction Models

In order to test the utility of the Patient Context Vectors for predicting patient outcomes, we focused on building a real-time ARDS prediction model. ARDS is a rare and life-threatening condition that require an early intervention (Fan et al., 2017).

ARDS patients were limited to adult patients only (age 18 or older). The patients inclusion criteria consist of the presence of acute respiratory failure and continuous mechanical ventilation, excluding patients with acute exacerbation of asthma or chronic obstructive pulmonary disease (Bime et al., 2016) [3]. This resulted in 4,624 ARDS admissions from a total of 48,399 admissions.

An ARDS prediction model was built utilizing a combination of vital signs, lab results, ICD codes and free-text notes. Features considered in the baseline predictive model building include: 1) vital signs: heart rate, respiratory rate, body temperature, systolic blood pressure, diastolic blood pressure, mean arterial pressure, oxygen saturation, tidal volume, BMI; 2) laboratory tests: white blood cell count, bands, hemoglobin, hematocrit, lactate, creatinine, bicarbonate, pH, PT, INR, BUN, blood gas measurements (partial pressure of arterial oxygen, fraction of inspired oxygen, and partial pressure of arterial carbon dioxide); 4) motor, verbal, and eye sub-score of Glasgow Coma Scale ; and 5) demographics: gender and age.

In addition to the baseline features (available in structured format in MIMIC), we also included as features the patient's Patient Context Vectors computed from ICD codes and from notes. In real-time CDS systems, it is likely that not all ICD or nursing notes will be available at prediction times. To test this most realistic scenario, we also built a Patient Context Vector by averaging the first half of the patient's ICD codes, and the first half of the patient's nursing notes CNN model predictions.

A Gradient Boosting Machine (GBM) model (Friedman, 2001) and a Distributed Random Forrest Model (DRF) (Geurts et al., 2006) were used to predict ARDS patients from the total population of adult patients. In all cases default model parameters were used (h2o). All results were produced via 10-fold cross evaluation. Table 1 shows the result from the experiments.

GBM				
Features	**AUC**	**P**	**R**	**F1**
Baseline	90.42	41.76	67.80	51.68
Baseline + ICD Patient Context Vector	93.30	53.02	68.44	59.75
Baseline + Notes Patient Context Vector	91.88	48.25	64.25	55.11
Baseline + first half of notes/ICD	**93.59**	56.35	66.52	**61.01**
DRF				
Features	**AUC**	**P**	**R**	**F1**
Baseline	89.14	38.58	66.43	48.81
Baseline + ICD Patient Context Vector	92.08	51.87	63.75	57.20
Baseline + Notes Patient Context Vector	91.18	47.89	62.11	54.08
Baseline + first half of notes/ICD	**92.61**	57.02	61.08	**58.98**

Table 1: 10-fold cross-validation GBM and DRF results of predicting ARDS patients. P=Precision, R=Recall, F1= F1-score for the positive (ARDS) class. The Baseline set of features consists of vital signs, lab results, Glasgow Coma Scale score, gender and age, in the form of structured data. "Baseline + ICD Patient Context Vector" includes all baseline features, plus the Patient Context Vector (of size 50). "Baseline + Notes" includes all baseline features, plus Patient Context Vectors predicted from all visit nursing notes. "Baseline + first half of notes/ICD" includes the average of the first half of entered visit ICD codes embeddings, and Patient Context Vectors predicted from the first half of the visit nursing notes.

Introducing information from both ICD codes and nursing notes data significantly increased the overall performance. Most importantly, the combination of the use of half of the visit notes (used to predict Patient Context Vectors) and the first half of the patient ICD codes produced the best results in both models (GBM and DRF), and proves the utility of the method for combining structured and free-text data for prediction models.

The benefit of averaged ICD-code embeddings, and using notes to predict the same embedding vectors is also illustrated by the model variable importances shown in Figures 1 - 4. As shown, the predictive value of certain embedding dimensions is on a par with important vital signs, such as Tidal Volume, Glasgow Coma Scale, and Mean Respiratory Rate. Intuitively, clinicians' experience utilizes all information present in nursing notes (also coded as ICD codes) to evaluate a patient's condition. Our approach demonstrates that it is possible to summarize that knowledge by combining nursing and ICD codes in the form of predicted and averaged ICD embeddings.

[2] https://github.com/ema-/patient-context-vectors

[3] Inclusion ICD9 Codes: 51881, 51882, 51884, 51851, 51852, 51853, 5184, 5187, 78552, 99592, 9670, 9671, 9672; Exclusion ICD9 Codes: 49391, 49392, 49322, 4280

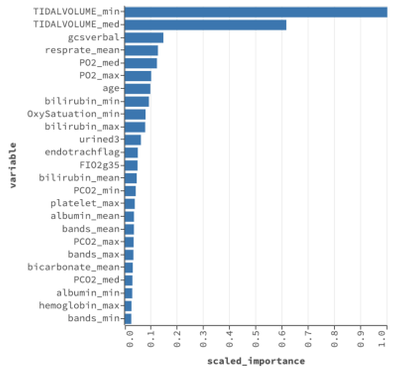

Figure 1: GBM scaled variable importance of Baseline model features.

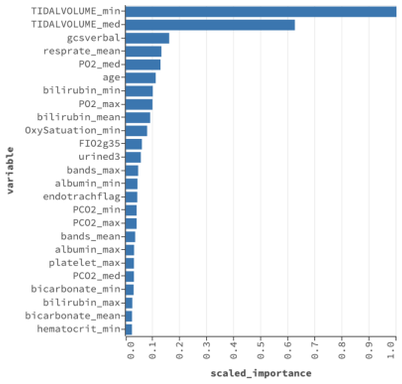

Figure 2: DRF scaled variable importance of Baseline model features.

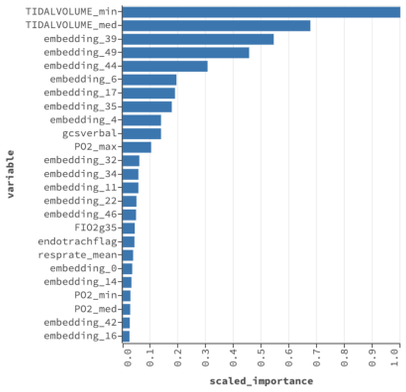

Figure 3: GBM scaled variable importance of Baseline model features plus Patient Context Vectors from first half of ICD codes/notes.

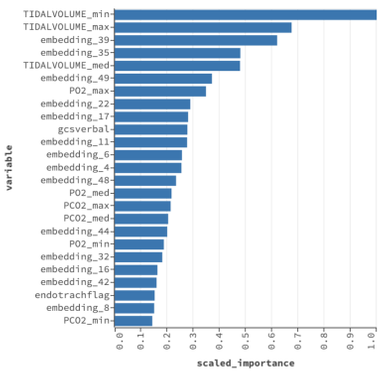

Figure 4: DRF scaled variable importance of Baseline model features plus Patient Context Vectors from first half of ICD codes/notes.

3 Related Work

A large volume of literature on combining structured and free-text EMR data pre-processes the free-text data by applying some information extraction (IE) technique (most frequently, Medical Concept detection). For example, DeLisle et al.(2010) and Zheng et al. (2014) apply free-text search on the notes to find a set of hand-crafted non-negated symptoms, later used as variables in their ML models. Ford et al. (2016) present a review of various approaches to IE from free-text notes for the purpose of detecting cases of a clinical condition, often in conjunction with structured data. The majority of approaches extract UMLS[4] or SNOMED-CT[5] concepts from free-text with their negation status with various off-the-shelf tools (Gundlapalli et al., 2008; Carroll et al., 2011; Karnik et al., 2012; Ananthakrishnan et al., 2013; Zheng et al., 2014).

More recently, deep learning has been used to combine free-text and structured EMR data. Relevant ICD embeddings work was mentioned in Section 2.2. Shickel et al. (2018) present a survey of various deep learning techniques. Most notably, Miotto et al. (2016) convert notes to concepts, which are then used in conjunction with structured data to build a *Deep Patient* representation in an unsupervised manner via denoising autoencoders.

4 Conclusion

Intuitively, the information available in notes and ICD codes, enhances the knowledge of the overall patient condition, which is indicative of the patient outcome. Results show that Patient Context Vectors can be easily combined with structured data in the form of vital signs an lab results and improve significantly the prediction model results. Results also indicate that Patient Context Vectors are suitable for real-time CDS as they perform equally well when only the first half of available ICD codes and notes is used.

Acknowledgements

Research reported in this publication was supported by a NIH SBIR award to CTA by NIH National Heart, Lung, and Blood Institute, of the National Institutes of Health under award number 1R43HL135909-01A1.

[4] Unified Medical Language System, ©The U.S. National Library of Medicine.

[5] Systematized Nomenclature of Medicine - Clinical Terms, ©IHTSDO.

References

h2o.ai. https://www.h2o.ai/. Accessed: 2019-01-30.

Ashwin N Ananthakrishnan, Tianxi Cai, Guergana Savova, Su-Chun Cheng, Pei Chen, Raul Guzman Perez, Vivian S Gainer, Shawn N Murphy, Peter Szolovits, Zongqi Xia, et al. 2013. Improving case definition of crohn's disease and ulcerative colitis in electronic medical records using natural language processing: a novel informatics approach. *Inflammatory bowel diseases*, 19(7):1411–1420.

Christian Bime, Chithra Poongkunran, Mark Borgstrom, Bhupinder Natt, Hem Desai, Sairam Parthasarathy, and Joe GN Garcia. 2016. Racial differences in mortality from severe acute respiratory failure in the united states, 2008–2012. *Annals of the American Thoracic Society*, 13(12):2184–2189.

Robert J Carroll, Anne E Eyler, and Joshua C Denny. 2011. Naïve electronic health record phenotype identification for rheumatoid arthritis. In *AMIA annual symposium proceedings*, volume 2011, page 189. American Medical Informatics Association.

Edward Choi, Mohammad Taha Bahadori, Elizabeth Searles, Catherine Coffey, Michael Thompson, James Bost, Javier Tejedor-Sojo, and Jimeng Sun. 2016a. Multi-layer representation learning for medical concepts. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1495–1504. ACM.

Youngduck Choi, Chill Yi-I Chiu, and David Sontag. 2016b. Learning low-dimensional representations of medical concepts. *AMIA Summits on Translational Science Proceedings*, 2016:41.

Sylvain DeLisle, Brett South, Jill A Anthony, Ericka Kalp, Adi Gundlapallli, Frank C Curriero, Greg E Glass, Matthew Samore, and Trish M Perl. 2010. Combining free text and structured electronic medical record entries to detect acute respiratory infections. *PloS one*, 5(10):e13377.

Eddy Fan, Lorenzo Del Sorbo, Ewan C Goligher, Carol L Hodgson, Laveena Munshi, Allan J Walkey, Neill KJ Adhikari, Marcelo BP Amato, Richard Branson, Roy G Brower, et al. 2017. An official american thoracic society/european society of intensive care medicine/society of critical care medicine clinical practice guideline: mechanical ventilation in adult patients with acute respiratory distress syndrome. *American journal of respiratory and critical care medicine*, 195(9):1253–1263.

Elizabeth Ford, John A Carroll, Helen E Smith, Donia Scott, and Jackie 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.

Jerome H Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Pierre Geurts, Damien Ernst, and Louis Wehenkel. 2006. Extremely randomized trees. *Machine learning*, 63(1):3–42.

Adi V Gundlapalli, Brett R South, Shobha Phansalkar, Anita Y Kinney, Shuying Shen, Sylvain Delisle, Trish Perl, and Matthew H Samore. 2008. Application of natural language processing to va electronic health records to identify phenotypic characteristics for clinical and research purposes. *Summit on translational bioinformatics*, 2008:36.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035.

Shreyas Karnik, Sin Lam Tan, Bess Berg, Ingrid Glurich, Jinfeng Zhang, Humberto J Vidaillet, C David Page, and Rajesh Chowdhary. 2012. Predicting atrial fibrillation and flutter using electronic health records. In *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*, pages 5562–5565. IEEE.

David Kartchner, Tanner Christensen, Jeffrey Humpherys, and Sean Wade. 2017. Code2vec: Embedding and clustering medical diagnosis data. In *Healthcare Informatics (ICHI), 2017 IEEE International Conference on*, pages 386–390. IEEE.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Riccardo Miotto, Li Li, Brian A Kidd, and Joel T Dudley. 2016. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Scientific reports*, 6:26094.

Benjamin Shickel, Patrick James Tighe, Azra Bihorac, and Parisa Rashidi. 2018. Deep ehr: a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. *IEEE journal of biomedical and health informatics*, 22(5):1589–1604.

Hongzhang Zheng, Holly Gaff, Gary Smith, and Sylvain DeLisle. 2014. Epidemic surveillance using an electronic medical record: an empiric approach to performance improvement. *PloS one*, 9(7):e100845.