# Predicting ICU transfers using text messages between nurses and doctors

**Faiza Khan Khattak**[1,2,3], **Chloé Pou-Prom**[2], **Robert Wu**[4], **Frank Rudzicz**[1,2,3,5]
[1]Department of Computer Science, University of Toronto
[2]St. Michael's Hospital
[3]Vector Institute for Artificial Intelligence
[4]Toronto General Hospital
[5]Surgical Safety Technologies
faizakk@cs.toronto.edu, poupromc@smh.ca,
robert.wu@uhn.ca, frank@cs.toronto.edu

## Abstract

We explore the use of real-time clinical information, i.e., text messages sent between nurses and doctors regarding patient conditions in order to predict transfer to the intensive care unit (ICU). Preliminary results, in data from five hospitals, indicate that, despite being short and full of noise, text messages can augment other visit information to improve the performance of ICU transfer prediction.

## 1 Introduction

'Failure to rescue' is an important aspect of patient safety and can be caused by poor communication, or a lack of situational awareness, in the care team (Brady and Goldenhar, 2014). There has been increased recognition of the importance of acting on deteriorating patients by escalating their care via rapid response and emergency medical teams (DeVita et al., 2006). Established criteria, such as the Modified Early Warning Score (MEWS) (Subbe et al., 2001), identify patients at risk of deterioration. Recently, machine learning approaches have employed electronic patient record data, vital signs, and laboratory results (Zhou et al., 2016; Futoma et al., 2015; Che et al., 2016; Frost et al., 2017), and have typically performed better than MEWS (Churpek et al., 2016; Zhai et al., 2014).

Related work in intensive care unit (ICU) transfer prediction often relies on structured data (i.e., lab results and vitals) taken from the patient's electronic health record. For instance, Tabak et al. (2017) developed a measure that relied on both clinical and administrative data (e.g., diagnosis, length of stay, number of previous discharges) and predicted hospital readmission with c-statistics up to 0.722. Similarly, Genevès et al. (2018) focused on drug prescription data on the day of admission, and predicted various forms of risk, including ICU admissions ($\geq$65% AUC). By contrast, Escudié et al. (2018) represented the *text* of electronic health records based on the Fast Healthcare Interoperability Resources format[1] and used word embedding and random forests to predict disease codes at the time of discharge, with a wide range in accuracies. Miotto et al. (2016) embedded medications, diagnoses, procedures, lab tests, and other structured information in a deep neural net and were able to predict various diseases with an average AUC-ROC of 0.773. Crucially, none of these systems used dynamic real-time data on a patient.

Real-time clinical information, especially communication between nurses and doctors, may be useful in improving the accuracy of detecting deteriorating patients (Rajkomar et al., 2018). In particular, this information may hold vital data not included in other fields, including changes in consciousness, pain, and other symptoms. Often, urgent communication in the hospital still occurs through pagers, limiting analysis of this communication (De Meester et al., 2013; Wu et al., 2013; Johnston et al., 2014). In some hospitals, however, communication occurs through text messaging. This transition from unrecorded messages to text allows for deeper analysis of these potentially crucial information. In this work, we evaluate the impact of using text messages between physicians and nurses to predict ICU transfer.

## 2 Data

Our data consist of 38,373 patients across 49,224 visits, between 2011 and 2017, divided into five groups according to different institutional codes. Messages from 2011 to 2015 are in a different format (from an older system), so we focus our analysis on messages from 2015 to 2017. We also exclude all patients who have missing institutional

[1]https://www.hl7.org/fhir/overview.html



*Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 89–94
Minneapolis, Minnesota, June 7, 2019. ©2019 Association for Computational Linguistics

| | *Group A* | *Group B* | *Group C* | *Group D* | *Group E* |
|---|---|---|---|---|---|
| **Patient info.** | | | | | |
| # Patients (M/F) | 4,536 / 4,031 | 3488 / 3363 | 206 / 202 | 21 / 19 | 17 / 10 |
| Age at admission | 63.45 (18.55) | 70.01 (18.86) | 72.75 (14.45) | 72.59 (17.35) | 72.66 (12.07) |
| # mheaders/patient | 13.86 (23.07) | 15.54 (24.96) | 21.81 (38.58) | 15.07 (19.68) | 11.82 (9.17) |
| # mreplies/patient | 14.27 (23.52) | 16.46 (26.49) | 22.68 (40.86) | 21.61 (23.39) | 12.32 (8.74) |
| **Visit info.** | | | | | |
| # Visits | 10,001 | 8,586 | 527 | 57 | 30 |
| # visits/patient | 1.35 (0.89) | 1.41 (1.02) | 1.37 (0.86) | 1.48 (0.85) | 1.29 (0.60) |
| # days/visit | 9.80 (19.01) | 9.91 (21.47) | 15.22 (19.87) | 12.76 (14.75) | 9.22 (6.47) |
| # mheaders/visit | 9.85 (16.58) | 10.64 (17.84) | 15.48 (27.02) | 14.40 (18.14) | 8.67 (7.97) |
| # mreplies/visit | 10.18 (16.98) | 11.31 (18.95) | 16.18 (28.46) | 15.07 (19.68) | 9.03 (7.99) |
| **Messages info.** | | | | | |
| # mheaders | 98,468 | 91,330 | 8,159 | 821 | 260 |
| # mreply | 99,456 | 95,654 | 8,395 | 844 | 271 |
| ICU% | 16.75% | 0.36% | 35.86% | 22.12% | 2.01% |
| # tokens/mheader | 22.31 (14.22) | 22.90 (14.28) | 22.68 (14.62) | 22.38 (13.47) | 23.70 (13.84) |
| # tokens/mreply | 7.34 (7.97) | 7.55 (8.07) | 7.47 (7.89) | 7.05 (7.22) | 7.42 (7.17) |

Table 1: Patient, visit, and messages information of data between years 2015 and 2017 used to train models for predicting ICU transfer. We indicate standard deviation in parentheses. ICU % is the ratio of mheaders resulting in ICU transfer within 3 days of the message send date.

code in their record. Data include patient and visit information, and text messages.

**Patient information** includes patient ID, date of birth, gender, date admitted, most recent medication, and most recent diagnosis.

**Visit information** includes visit number, discharge date time, diagnosis made during the visit[2], visit type ("Emergency" or "Inpatient"), doctors' notes, lab results, institutional code, and an Admission/Discharge/Transfer (ADT) code indicating to where the patient was admitted, discharged, or transferred to. Of the 539 ADT values, 19 correspond to an ICU transfer.

**Text messages** are collected from the hospital network system and split into *message headers* and *message replies*. *Message headers* consist of text messages sent from nurses to physicians. These messages include information such as medication and status of patient. Some message headers have a corresponding *message reply*, which consists of text responses from doctors. The database system in which these text messages are stored only allows for replies from doctor but not a reply back from nurse. If a nurse replies back, it is considered a new message header, making it difficult to track a "conversation thread". Sometimes the message reply gets sent more than one time and many other times it is empty. In our experiments, we only look at the *message header*, as most *message replies* are short and uninformative. The top most frequent replies are: *"thanks"*, *"ok"* and *"noted"* across all groups. We split the data by institutional code and report a summary of the demographics, visits, and messages in Table 1.

[2]This is not the same as the diagnosis in the patient information.

| |
|---|
| **mheader:** *"hb=65, cr=123 & more lab res up from last nights bldwork. Ping if anything you want me to follow up."*<br>**mreply:** *"informed."* |
| **mheader:** *"dc hep drip on epr. Pls see chart order. Thnx."*<br>**mreply:** *"done thanks"* |
| **mheader:** *"hey are icu recommends to be cosigned. thx."*<br>**mreply:** *"Ok. Pls run one l of ringers wide & then one more"* |

Table 2: Examples of *message header* (mheader) and *message reply* (mreply) pairs. Modified for anonymity.

Text messages can be challenging to analyze, given spelling mistakes, abbreviations specific to the medical domain, missing punctuation, and other challenges. Open-source spelling correction

| | Group *A* | | Group *C* | |
|---|---|---|---|---|
| | $--$ | $+Ling$ | $--$ | $+Ling$ |
| $Visit$ | 0.47 ± (0.01) | 0.50 ± (0.01) | 0.44 ± (0.02) | 0.53 ± (0.02) |
| $Visit + TFIDF$ | **0.51** ± **(0.01)** | 0.48 ± (0.01) | 0.56 ± (0.04) | 0.56 ± (0.04) |
| $Visit + w2v_{SMS}$ | **0.51** ± **(0.02)** | 0.48 ± (0.01) | **0.57** ± **(0.05)** | **0.57** ± **(0.04)** |
| $Visit + w2v_{Pubmed}$ | **0.51** ± **(0.01)** | 0.49± (0.01) | 0.54 ± (0.04) | 0.54 ± (0.04) |
| $Visit + TFIDF + w2v_{SMS}$ | **0.51**± **(0.01)** | 0.48 ± (0.01) | 0.56 ± (0.04) | 0.56 ± (0.04) |
| $Visit + TFIDF + w2v_{Pubmed}$ | 0.49± (0.01) | 0.47± (0.01) | 0.54± (0.03) | 0.51 ± (0.04) |

Table 3: Macro $F_1$- scores on the logistic regression model for Group *A* and *C*. We report the macro $F_1$ metric averaged over 5-fold cross-validation (with standard deviations in parentheses).

software[3] provides little improvement, due to the domain-specific nature of the words. E.g., the message *'pls add prn pain med, not PO. thx'* gets corrected to *'ls add pr pain mod, not PO. tax'*. We provide examples of message header and message reply pairs in Table 2.

We focus our experiments on Group *A* since it has the most amount of data, and on Group *C* since it has the most number of messages per visit and the longest messages. We use the ADT (Admission/Discharge/Transfer) code in the patients' records to determine transfer to the ICU. A *mheader* is determined to have the outcome if an ICU transfer occurs within the next 3 days of the message send date (Table 1).

## 3 Methods

For each text message, we include the patient's age and gender, the total number of days spent in hospital at the time the message is sent[4], their prescribed medication at the time of the message, and their diagnosis. The medication and diagnosis are encoded with one-word TF-IDF.

We then look at the following representations of text messages. For each representation, we use at most 20 words and zero-pad if necessary:

**TF-IDF**: We represent each text message with its TF-IDF representation. We experiment with word, $n$-gram, and character-level TF-IDF, as well as combinations. We use $n$-gram TF-IDF ($n = 1, 2, 3$) in our final models.

**Word2Vec**: We use 1) pre-trained word embeddings (Mikolov et al., 2013) trained on publicly available PubMed articles (Moen and Ananiadou, 2013), as well as 2) our own word embeddings, trained on the text messages data. We train word embeddings of dimension size 100, with a context window equal to 5 for training (Bojanowski et al., 2017). We explore different combinations of the text message word embeddings through concatenation, summing, and averaging. We report results using a combination of all three types. More specifically, we concatenate twenty 100-dimensional word embeddings (2000 dimensions), a sum of the word embeddings (100 dimensions), and an average of the 20 words (100 dimensions), for a total of 2200-dimensional feature vector.

**Linguistic features**: We represent each text message as a vector containing 9 linguistic features. We compute lexical features (character and word count, word density[5]), syntactic features (counts of nouns, verbs, adjectives, and adverbs), and positive and negative polarity extracted from `nltk`'s sentiment analyzer (Loper and Bird, 2002).

We use an ANOVA-based feature selection (Pedregosa et al., 2011), and we train a logistic regression model. We report the macro $F_1$ metric averaged across 5-fold cross-validation.

## 4 Results

We experiment with *Visit* (i.e., age, gender, total number of days spend in hospital, medication, and diagnosis), *TFIDF*, *Ling* (i.e., linguistic), $w2v_{SMS}$ (i.e., word vectors trained on text messages), and $w2v_{Pubmed}$ (i.e., pre-trained word vectors from PubMed) features. When multiple text representations are used (e.g., TF-IDF and w2v), we concatenate them together. Typically, the addition of linguistic features does not seem to improve performance.

We then look at performance across data and report results in Table 4 on the logistic regression model using visit information only, visit fea-

[3] `https://github.com/rfk/pyenchant`

[4] this includes the number of days spent in the hospital from *previous* visits

[5] *Word density* is the number of words in a message divided by the number of characters in a message.

tures augmented with text message representations ($w2v_{SMS}$, and $TFIDF$).

| | visit | visit + $\mathbf{w_{sms}}$ | visit + tfidf |
|---|---|---|---|
| *A* | 0.47 (0.01) | **0.51 (0.02)** | **0.51 (0.01)** |
| *B* | 0.48 (0.01) | 0.46 (0.07) | **0.50 (0.01)** |
| *C* | 0.44 (0.02) | **0.57 (0.05)** | **0.56 (0.04)** |
| *D* | 0.44 (0.03) | **0.46 (0.07)** | 0.44 (0.04) |
| *E* | 0.69 (0.28) | 0.69 (0.28) | 0.69 (0.28) |

Table 4: Model performance across the different data. $F_1$ macro results on a logistic regression model using 1) visit features only, and 2) visit features, word2vec embeddings (i.e., $w2v_{SMS}$) and 3) visit features, TF-IDF features.

Our results indicate that the addition of information from text messages improves results in ICU transfer prediction three days before the event happens. Our best results are in the model consisting of visit and word2vec features trained on our data (i.e., $w2v_{SMS}$). We look more closely at the performance of this model with this subset of features in Table 5. As expected, the model does better on messages that don't result in ICU transfer. We obtain recall of 0.22 and 0.43 in ICU transfer messages for Groups *A* and *C*, respectively.

| | | Group A | Group C |
|---|---|---|---|
| **No ICU transfer** | **P** | 0.86 (0.01) | 0.78 (0.05) |
| | **R** | 0.80 (0.05) | 0.73 (0.07) |
| **ICU transfer** | **P** | 0.16 (0.02) | 0.36 (0.06) |
| | **R** | 0.22 (0.06) | 0.43 (0.13) |
| **Micro $\mathbf{F_1}$** | | 0.72 (0.04) | 0.65 (0.05) |
| **Macro $\mathbf{F_1}$** | | 0.51 (0.02) | 0.57 (0.05) |
| **Weighted $\mathbf{F_1}$** | | 0.74 (0.03) | 0.66 (0.05) |

Table 5: Results for logistic regression model using Visit and $w2v_{SMS}$.

## 5 Discussion

Table 4 shows that the addition of text message representations yields to improvements in Groups *A*, *B*, *C*, and *D*. The greatest improvement in in Group *C*. The proportion of messages which are followed by an ICU transfer three days later is much higher in Group *C*, which could reasonably explain the difference in performance. However, we also note that text messages in Group *C* tend to be longer than other data, and that nurses in Group *C* send more messages per visit. Across all data, the best model performance is for Group *E* but no improvement after adding text messages. However, it consists of the smallest number of messages and the highest variance in performance across validation folds. The ratio of messages which are followed by an ICU transfer in the next 3 days are 16.75%, 0.36%, 35.86%, 22.12% and 2.01% for Groups *A*, *B*, *C*, *D*, and *E*, respectively (Table 2). The differences in performance could be attributed to the number of messages.

| Word | $\mathbf{w2v_{SMS}}$ | $\mathbf{w2v_{Pubmed}}$ |
|---|---|---|
| dr | dr., doctor, md, resident, oncologist | 99:1, diastereoselectivities, ee, =98:2, 98:2 |
| bld | blood, bloood, blod , frozen, pt.iv | whi, bldB, EPS-deficient, transposon-generated, A-factor-deficient |
| med | medication, pill, lactulose, risperidone, hypoglycemics | Nicolae, Delores, Dres, habil., CSc. |
| bp | b/p, bp=, bp-, bpm, pulse | nt, bps, nts, bp-long, bp-long |
| icu | msicu, emerg, er, cvicu, gim | bag/mask, Patient-initiated, extra-hospital, patient-cycled, airway-management |

Table 6: Comparison of word embeddings. Top five similar words for common abbreviated medical terms. $\mathbf{w2v_{SMS}}$ denotes the word embeddings trained on our text message data and $\mathbf{w2v_{Pubmed}}$ denotes the word embeddings trained on publicly available PubMed articles.

Using word embeddings trained on our data performs better than the pre-trained ones. We dig deeper and report the top 5 similar words of some common medical terms in Table 6. Word embeddings trained on text messages do a much better job of capturing different spellings (e.g., *"bp"* and *"b/p"*) as well as common misspellings (e.g., *"bloood"* and *"blod"*). These results further highlight the need for context-specific word embeddings (Chiu et al., 2016).

## 6 Conclusion & Future Work

In this work, we look at the added value of text messages sent from nurses to doctors in predicting transfer to the ICU within three days of the message send date. We find that including messages from information - through linguistic features, TF-IDF features, and word vector representations - improves performance. This finding is consistent in 4 of the 5 datasets divided by institutional codes. The best performance was observed in the data with the most ICU transfers, the longest text messages and the most text messages per visit and per patient. We find that using word vectors trained on the text messages results in the best model performance, and a closer look shows that the embeddings do a better job at capturing misspellings and abbreviations unique to text messages.

In future work, we want to investigate differences across the data, and hope to identify key features of the text messages that are relevant in identifying ICU transfer. Other than that, we will also investigate the utility of adding the message replies, along with the message headers, as features. In this work, we have only looked at predictions for a given text message. Exploring how the prediction probabilities change over time would also be of interest. We will also consider different word embeddings (Peters et al., 2018), as we hypothesize that character-level word embeddings could better capture the unique vocabulary of text messages. To address class imbalance, we will explore undersampling/oversampling methods such as SMOTE (Chawla et al., 2002). Furthermore, we want to look at the added value of text messages in a more complex set of features (i.e., lab results and vitals), as we believe that this would provide a complete picture of the patient's visit profile.

## Acknowledgments

Rudzicz is an Inaugural CIFAR Chair in Artificial Intelligence.

## References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *TACL*, 5:135–146.

Patrick W. Brady and Linda M. Goldenhar. 2014. A qualitative study examining the influences on situation awareness and the identification, mitigation and escalation of recognised patient risk. *BMJ Qual Saf*, 23(2):153–161.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Zhengping Che, Sanjay Purushotham, Robinder Khemani, and Yan Liu. 2016. Interpretable deep models for icu outcome prediction. In *AMIA Annual Symposium Proceedings*, volume 2016, page 371. American Medical Informatics Association.

Billy Chiu, Gamal Crichton, Anna Korhonen, and Sampo Pyysalo. 2016. How to train good word embeddings for biomedical nlp. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, pages 166–174.

Matthew M. Churpek, Trevor C. Yuen, Christopher Winslow, David O. Meltzer, Michael W. Kattan, and Dana P. Edelson. 2016. Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards. *Critical care medicine*, 44(2):368.

Koen De Meester, M. Verspuy, Koen G. Monsieurs, and Peter Van Bogaert. 2013. Sbar improves nurse–physician communication and reduces unexpected death: A pre and post intervention study. *Resuscitation*, 84(9):1192–1196.

Michael A. DeVita, Rinaldo Bellomo, Kenneth Hillman, John Kellum, Armando Rotondi, Dan Teres, Andrew Auerbach, Wen-Jon Chen, Kathy Duncan, Gary Kenward, et al. 2006. Findings of the first consensus conference on medical emergency teams. *Critical care medicine*, 34(9):2463–2478.

Jean-Baptiste Escudié, Alaa Saade, Alice Coucke, and Marc Lelarge. 2018. Deep representation for patient visits from electronic health records. *arXiv preprint arXiv:1803.09533*.

David W. Frost, Shankar Vembu, Jiayi Wang, Karen Tu, Quaid Morris, and Howard B. Abrams. 2017. Using the electronic medical record to identify patients at high risk for frequent emergency department visits and high system costs. *The American journal of medicine*, 130(5):601–e17.

Joseph Futoma, Jonathan Morris, and Joseph Lucas. 2015. A comparison of models for predicting early hospital readmissions. *Journal of biomedical informatics*, 56:229–238.

Pierre Genevès, Thomas Calmant, Nabil Layaïda, Marion Lepelley, Svetlana Artemova, and Jean-Luc Bosson. 2018. Scalable machine learning for predicting at-risk profiles upon hospital admission. 12:23–34.

Maximilian Johnston, Sonal Arora, Dominic King, Luke Stroman, and Ara Darzi. 2014. Escalation of care and failure to rescue: a multicenter, multiprofessional qualitative study. *Surgery*, 155(6):989–994.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. In *In Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. Philadelphia: Association for Computational Linguistics*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

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.

SPFGH Moen and Tapio Salakoski2 Sophia Ananiadou. 2013. Distributional semantics resources for biomedical text processing. In *Proceedings of the 5th International Symposium on Languages in Biology and Medicine, Tokyo, Japan*, pages 39–43.

Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *CoRR*, abs/1802.05365.

Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, et al. 2018. Scalable and accurate deep learning with electronic health records. *npj Digital Medicine*, 1(1):18.

C.P. Subbe, M. Kruger, P. Rutherford, and L. Gemmel. 2001. Validation of a modified early warning score in medical admissions. *Qjm*, 94(10):521–526.

Ying P. Tabak, Xiaowu Sun, Carlos M. Nunez, Vikas Gupta, and Richard S. Johannes. 2017. Predicting readmission at early hospitalization using electronic clinical data: an early readmission risk score. *Medical care*, 55(3):267.

Robert C Wu, Vivian Lo, Dante Morra, Brian M Wong, Robert Sargeant, Ken Locke, Rodrigo Cavalcanti, Sherman D. Quan, Peter Rossos, Kim Tran, et al. 2013. The intended and unintended consequences of communication systems on general internal medicine inpatient care delivery: a prospective observational case study of five teaching hospitals. *Journal of the American Medical Informatics Association*, 20(4):766–777.

Haijun Zhai, Patrick Brady, Qi Li, Todd Lingren, Yizhao Ni, Derek S Wheeler, and Imre Solti. 2014. Developing and evaluating a machine learning based algorithm to predict the need of pediatric intensive care unit transfer for newly hospitalized children. *Resuscitation*, 85(8):1065–1071.

Huaqiong Zhou, Phillip R. Della, Pamela Roberts, Louise Goh, and Satvinder S. Dhaliwal. 2016. Utility of models to predict 28-day or 30-day unplanned hospital readmissions: an updated systematic review. *BMJ open*, 6(6):e011060.