Predicting Unplanned Hospital Admissions using Clinical Notes

Isabelle E. Amick, Ada J. Guan
UC Berkeley School of Information
W266: Natural Language Processing
{iengelberg, adajguan}@ischool.berkeley.edu

Abstract

Predicting and preventing unplanned hospital admissions (UHAs) is a valuable task both to improve patient health and prevent overuse of emergency department resources. While traditional UHA models make use of structured health data including diagnosis and procedure codes and demographic factors, there is untapped potential to improve our models by making use of the unstructured clinical notes provided by healthcare professionals. Here, we demonstrate the utility of these notes by using natural language processing to classify UHAs from clinical text.

1 Introduction

The ability of artificial intelligence (AI) to accurately and fairly predict patient outcomes from electronic health records (EHR) has become a hot topic in recent years. While the potential reward is large, the risk is quite literally a matter of life or death. With the well-established medical coding system, it is not difficult to create structured data from patient claims, however, the unstructured clinical notes pose a unique challenge for interpretation by natural language processing (NLP) (Dreisbach et al., 2019; Li et al., 2022; Hossain et al., 2023; Ohno-Machado, 2011). Clinicians report key information including symptoms, family histories, environmental factors, and care plans in their notes–all of which is lost when models only make use of the structured claims data.

One key area where AI has been successful in improving patient outcomes is in predicting (and, subsequently, preventing) unplanned hospital admissions (UHAs). UHAs are costly for both patients and hospitals, and are often avoidable with early detection and high-quality preventative care (Starfield et al., 2005; Kozak et al., 2017). While both regression and deep learning models have made use of the structured health data to predict UHAs (Abel et al., 2018; González-Colom et al., 2023; Klunder et al., 2022), these models typically do not incorporate clinical notes due to the challenge of extracting meaningful data from the unstructured text.

2 Background

While direct prediction of UHAs from clinical notes has not been studied, NLP has been used to make sense of clinical notes in various ways. Some examples include extracting symptoms (Koleck et al., 2019) and predicting diagnostic codes (Hsu et al., 2020; Chen et al., 2021) using Named Entity Recognition (NER), and creating structured text from the unstructured clinical notes (Liang et al., 2017) using summarization.

3 Methods

3.1 Data

To analyze the impact of clinical notes on unplanned hospital admissions models, we made use of the publicly available MIMIC-III (Medical Information Mart for Intensive Care III) dataset (`https://mimic.mit.edu/docs/iii/`), which consists of health data from over 40,000 deidentified patients who were admitted to the critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012 (Johnson et al., 2016). The database is made up of 26 tables encoding unique aspects of health data. In this work, we have used the `ADMISSIONS`, `DIAGNOSES_ICD`, `PROCEDURES_ICD`, and `NOTEEVENTS` tables to incorporate admission information, diagnosis and procedure codes, and corresponding clinical notes (unstructured text), respectively. All tables were joined using the `SUBJECT_ID` (unique patient id) and `HADM_ID` (unique hospital stay id).

3.2 Data Preprocessing

After joining the admissions, diagnoses, procedures, and notes tables, the resulting dataset

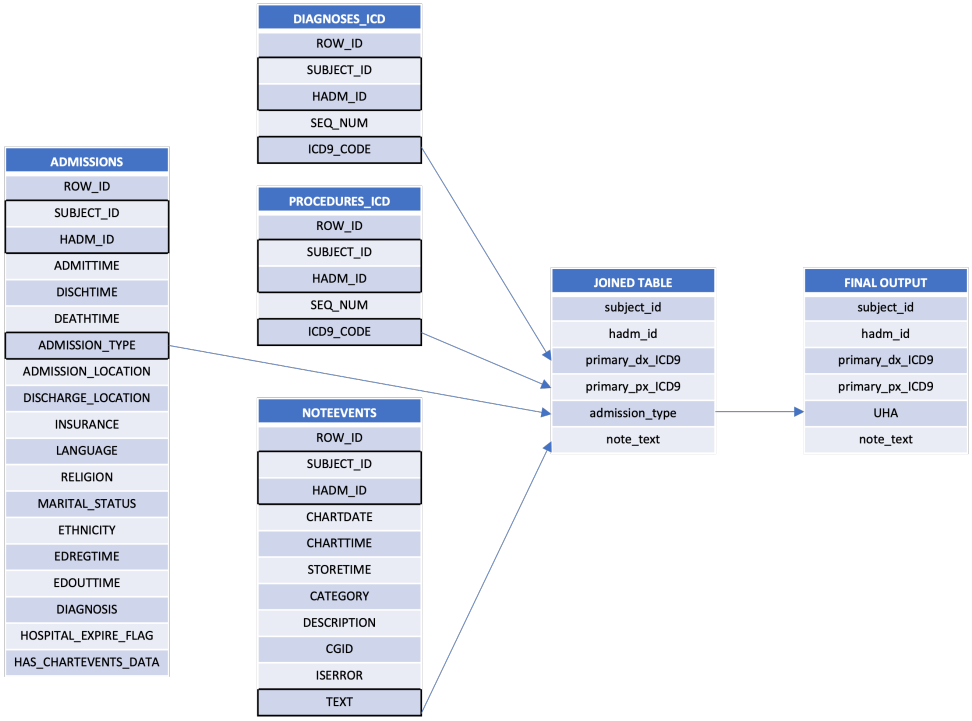

Figure 1: MIMIC-III Data processing pipeline. SUBJECT_ID and HADM_ID were used as join keys across ADMISSIONS, DIAGNOSES_ICD, PROCEDURES_ICD, and NOTEEVENTS. Outlined columns were brought into the joined table. For the final output, admission_type was converted to a binary column, UHA.

consisted of over 3 million rows, where each row corresponds to a unique patient, hospital stay, and clinical note. In order to optimize the dataset for model training, we began by selecting a subset of the data containing only the rows with primary diagnosis (dx) codes that rank among the top 10 most frequent dx codes in the dataset (Table 1). We performed this subselection using the MIMIC-III link to Google BigQuery, resulting in an initial raw dataset of 182,425 rows and 6 columns: `subject_id`, `hadm_id`, `admission_type`, `primary_px_ICD9`, `primary_dx_ICD9`, and `note_text`. We also created the binary column, `UHA`, which outputs a 1 for rows with an admissions type of "EMERGENCY" or "URGENT" and 0 for rows with an admissions type of "ELECTIVE".

To preprocess the clinical notes text, we designed two functions. The first function employs list comprehension to remove alphanumeric and punctuation characters, while the second function leverages the NLTK library to eliminate stop words from the input text. To ensure accurate concatenation and proper processing of our input features within our models, it's essential to maintain alignment in the second dimension. This alignment was achieved by padding diagnosis (dx) codes and procedure (px) codes to the target length of 100. We employed the Word2Vec algorithm to generate embeddings for both px and dx codes. Separate Word2Vec models were trained for dx and px codes. To encode sequences using previously generated Word2Vec embeddings, two

Table 1: Most frequent ICD9 diagnosis codes in the MIMIC-III dataset.

ICD9 Code	Description
41401	Coronary atherosclerosis of native coronary artery
51881	Acute respiratory failure
4241	Aortic valve disorders
4280	Congestive heart failure, unspecified
5849	Acute kidney failure, unspecified
42731	Atrial fibrillation
5990	Urinary tract infection, site not specified
9971	Cardiac complications, not elsewhere classified
53081	Esophageal reflux
4019	Unspecified essential hypertension

functions were created. These functions compute the mean vector of word embeddings within each sequence. As a result, encoded representations of dx and px code are generated using the Word2Vec embeddings (`encoded_dx_code` and `encoded_px_code`).

3.3 Metrics

In order to compare and evaluate different models for selection and optimization, we will compare validation accuracy scores as well as F1 scores. Since our data contains a large class imbalance, the F1 score is important to calculate in addition to accuracy, as it provides more insight into the accuracy of each class, rather than the overall accuracy. A successful model should have an accuracy over 80% and an F1 score over 70%.

4 Results and Discussion

4.1 Baseline

To establish an initial baseline, we developed a logistic regression model using only dx codes and px codes as features with the binary UHA classification as the output. A primary drawback of this model type is its ineffectiveness in handling class imbalances. In our processed dataset, the positive class comprised of 142,572 rows, while the negative class was represented by only 39,853 rows. To tackle the class imbalance, we applied NearMiss undersampling of the positive class, which resulted in approximately ∼8,000 rows in both classes. Despite implementing the NearMiss undersampling technique to balance the classes, our baseline still demonstrated mediocre performance, with an accuracy of 57% and a macro average F1-score of 56%.

4.2 Models

4.2.1 Model Selection

Next, we tested 3 different strategies for model building: Bidirectional Encoder Representations from Transformers (BERT), Convolutional Neural Network (CNN), and a combined BERT + CNN model. All models used the clinical notes alone as the input and binary classification of UHAs as the output. For the BERT and BERT + CNN models, we tokenized the notes using the pretrained "bert-base-cased" BERT tokenizer from Hugging Face. For the CNN model, we tokenized the notes using TensorFlow's WhitespaceTokenizer.

For direct comparison, all models used the same hyperparameters, number of hidden layers, and accuracy measurements. All models were trained on the full processed dataset (with no need to balance the classes here), with training allowed on all layers over 1 epoch. The resulting validation accuracies can be found in Table 2.

Interestingly, we noticed that adding an attention layer to the CNN model decreased the validation accuracy to mirror those of the BERT and BERT + CNN models (Table 2). Upon analysis into the encoded text input, we noticed that a large portion of the tokens were encoded as "UNK" for unknown. This is likely the reason for the decrease in accuracy upon adding attention, as trying to contextualize unknown words will convolute the output, rather than add relevant information.

Table 2: Comparison of BERT, CNN, and BERT + CNN models against the baseline for model selection.

Model	Accuracy	F1 Score
Baseline	57%	56%
BERT	78.1%	86.3%
CNN	83.3%	88.6%
CNN + Attention	78.1%	86.3%
BERT + CNN	78.1%	86.3%

4.2.2 Model Optimization

Expanding on our initial CNN model, which incorporates a primary input layer (clinical notes), we reintroduced the dx and px features from the baseline model. Two additional input layers `input_px` and `input_dx` were defined with the same shape as the primary input. We concatenated the output from the global max pooling layer with the two new layers that were created. We applied the same number of hidden layers and ReLU activation function to the output of the convolutional layer as previous models. The model was compiled using adam optimizer, binary cross-entropy loss and the accuracy metric for training. This achieved an accuracy of 84.1%, marking it as our most successful model outcome.

Despite our high accuracy and F1 scores, we noticed an increase in the loss function over 5 epochs in the validation set, indicative of potential overfitting (Figure 2a). In an attempt to correct the overfitting, we first reduced the complexity of the CNN by decreasing the number of filters from 32 to 16 (Figure 2b). While this strategy resulted in a small improvement, we still observed an increase

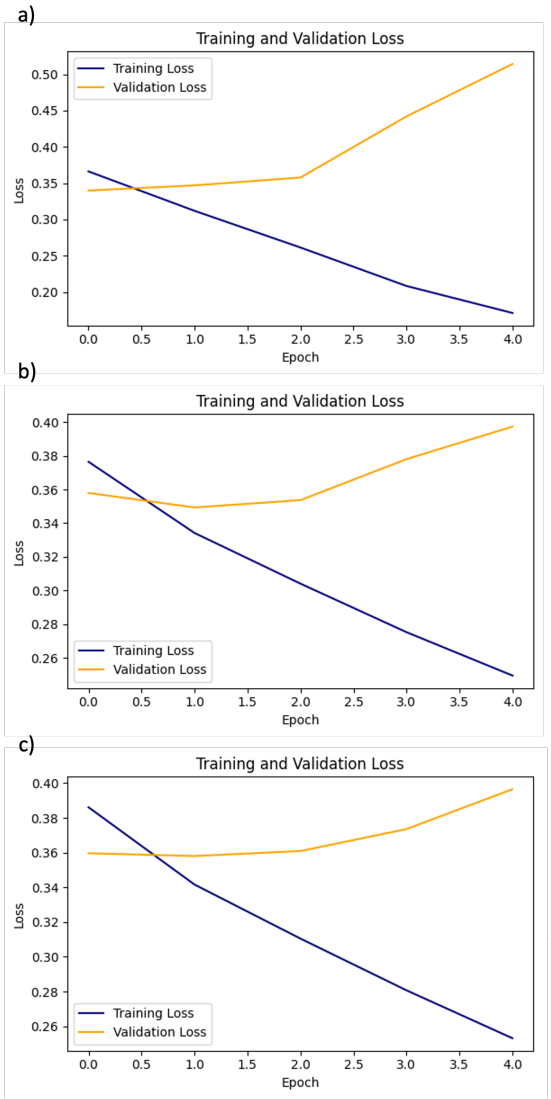

Figure 2: Loss function outputs for optimized model. a) Loss function output over 5 epochs for optimized model using 32 filters, b) 16 filters, and c) 16 filters with regularization.

in the loss function over time for the validation set. As a final attempt, we added regularization of the dense layer to penalize complex models (Figure 2c). Again, we saw a small improvement in the loss function, but further work is needed to combat overfitting.

5 Conclusion

In this work, we have shown the power of clinical notes to directly predict unplanned hospital admissions from electronic health records, as well as the potential for improving industry-standard structured data only models by incorporating unstructured clinical notes. Our optimized model, integrating clinical notes with dx and px codes, outperformed our initial baseline model by nearly 30%.

Future optimization of these models should make use of medical-specific pretraining models to limit the number of unknown words in the training set. This would 1) allow for more context-specific learning using attentions and 2) decrease overfitting. Examples of these models include BioBERT (Lee et al., 2020), PubMedBERT (Gu et al., 2021), and GatorTron (Yang et al., 2022), which are trained on a combination of clinical notes and medical texts.

References

Julian Abel, Helen Kingston, Andrew Scally, Jenny Hartnoll, Gareth Hannam, Alexandra Thomson-Moore, and Allan Kellehear. 2018. Reducing emergency hospital admissions: a population health complex intervention of an enhanced model of primary care and compassionate communities. *British Journal of General Practice*, 68(676):e803–e810.

Pei Fu Chen, Ssu Ming Wang, Wei Chih Liao, Lu Cheng Kuo, Kuan Chih Chen, Yu Cheng Lin, Chi Yu Yang, Chi Hao Chiu, Shu Chih Chang, and Feipei Lai. 2021. Automatic ICD-10 coding and training system: Deep neural network based on supervised learning. *JMIR Medical Informatics*, 9(8).

Caitlin Dreisbach, Theresa A. Koleck, Philip E. Bourne, and Suzanne Bakken. 2019. A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data.

Rubèn González-Colom, Carmen Herranz, Emili Vela, David Monterde, Joan Carles Contel, Antoni Sisó-Almirall, Jordi Piera-Jiménez, Josep Roca, and Isaac Cano. 2023. Prevention of Unplanned Hospital Admissions in Multimorbid Patients Using Computational Modeling: Observational Retrospective Cohort Study. *J Med Internet Res 2023;25:e40846 https://www.jmir.org/2023/1/e40846*, 25(1):e40846.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1).

Elias Hossain, Rajib Rana, Niall Higgins, Jeffrey Soar, Prabal Datta Barua, Anthony R. Pisani, and Kathryn Turner. 2023. Natural Language Processing in Electronic Health Records in relation to healthcare decision-making: A systematic review.

Jia Lien Hsu, Teng Jie Hsu, Chung Ho Hsieh, and Anandakumar Singaravelan. 2020. Applying convolutional neural networks to predict the ICD-9 codes of medical records. *Sensors (Switzerland)*, 20(24).

Alistair E W Johnson, Tom J Pollard, Lu Shen, Li-Wei H Lehman, 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(1):1–9.

Jet H. Klunder, Sofie L. Panneman, Emma Wallace, Ralph de Vries, Karlijn J. Joling, Otto R. Maarsingh, and Hein P.J. van Hout. 2022. Prediction models for the prediction of unplanned hospital admissions in community-dwelling older adults: A systematic review. *PLOS ONE*, 17(9):e0275116.

Theresa A. Koleck, Caitlin Dreisbach, Philip E. Bourne, and Suzanne Bakken. 2019. Natural language processing of symptoms documented in free-text narratives of electronic health records: A systematic review.

Lola Jean Kozak, Margaret J. Hall, and Maria F. Owings. 2017. Trends In Avoidable Hospitalizations, 1980–1998. *https://doi.org/10.1377/hlthaff.20.2.225*, 20(2):225–232.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Irene Li, Jessica Pan, Jeremy Goldwasser, Neha Verma, Wai Pan Wong, Muhammed Yavuz Nuzumlali, Benjamin Rosand, Yixin Li, Matthew Zhang, David Chang, R. Andrew Taylor, Harlan M. Krumholz, and Dragomir Radev. 2022. Neural Natural Language Processing for unstructured data in electronic health records: A review.

Hong Liang, Xiao Sun, Yunlei Sun, and Yuan Gao. 2017. Text feature extraction based on deep learning: a review. *Eurasip Journal on Wireless Communications and Networking*, 2017(1):1–12.

Lucila Ohno-Machado. 2011. Realizing the full potential of electronic health records: the role of natural language processing.

Barbara Starfield, Leiyu Shi, and James Macinko. 2005. Contribution of Primary Care to Health Systems and Health. *The Milbank Quarterly*, 83(3):457–502.

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E. Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B. Costa, Mona G. Flores, Ying Zhang, Tanja Magoc, Christopher A. Harle, Gloria Lipori, Duane A. Mitchell, William R. Hogan, Elizabeth A. Shenkman, Jiang Bian, and Yonghui Wu. 2022. A large language model for electronic health records. *npj Digital Medicine 2022 5:1*, 5(1):1–9.