Pressure Ulcer Injury in Unstructured Clinical Notes: Detection and Interpretation

Mani Sotoodeh, MS,[1] Zelalem H. Gero, MS,[1] Wenhui Zhang, PhD, MS, RN,[2] Roy L. Simpson, DNP, RN,[2] Vicki Stover Hertzberg, PhD,[2] Joyce C. Ho, PhD[1]
[1]Department of Computer Science, Emory University, Atlanta, GA, US
[2]Center for Data Science, Nell Hodgson Woodruff School of Nursing, Emory University, Atlanta, GA, US

Abstract

Hospital-acquired pressure ulcer injury (PUI) is a primary nursing quality metric, reflecting the caliber of nursing care within a hospital. Prior studies have used the Braden scale and structured data from the electronic health records to detect/predict PUI while the informative unstructured clinical notes have not been used. We propose automated PUI detection using a novel negation-detection algorithm applied to unstructured clinical notes. Our detection framework is on-demand, requiring minimal cost. In application to the MIMIC-III dataset, the text features produced using our algorithm resulted in improved PUI detection when evaluated using logistic regression, random forests, and neural networks compared to text features without negation detection. Exploratory analysis reveals substantial overlap between key classifier features and leading clinical attributes of PUI, adding interpretability to our solution. Our method could also considerably reduce nurses' evaluations by automatic detection of most cases, leaving only the most uncertain cases for nursing assessment.

Introduction

Pressure injury, the current terminology replacing pressure ulcer in 2016,[1] is a key indicator for patient safety and healthcare quality.[2,3] Pressure ulcer or injury (PUI) is defined as "a localized injury to the skin and/or underlying tissue usually over a bony prominence or related to a medical or other devices, as a result of pressure, or pressure in combination with shear".[1,4] Over 2.5 million Americans develop PUI each year.[3] Hospital admissions due to PUI are 75% higher than for other medical conditions. Prevalence of hospital-acquired PUI is from 5% to 6%.[3,5,6] PUI is associated with pain, reduced quality of life, higher mortality, longer hospital stays, more institutionalization after hospitalization, and economic burdens ($500-7,000 per pressure ulcer) on patients and healthcare system.[2,3,7–10] From 2015, any hospital-acquired PUI, that is, PUI developing after initial evaluation within 24 hours of admission, is considered preventable, and Medicare applies payment penalties to the bottom 25% of the lowest-performing hospitals.[11]

Hospitals have been endeavoring to better evaluate PUI risk to allocate resources such as specialized pressure mattresses and nurses to monitor skin status, especially for intensive and postoperative care units. Risk assessment tools such as the Braden scale[12] have been used for PUI risk evaluation in the past. However such tools have highly variable sensitivity and specificity ,[13] poor prognostic accuracy,[14] and no apparent impact on decreasing PUI incidence.[15] Recently personalized risk algorithms have also been developed with a broader range of patient-level factors than these earlier tools, using structured data from electronic health records (EHRs).[16] Most of these works have focused on classifying PUI versus non-PUI patients to identify risk factors retrospectively rather than predicting PUI development within a prospective time interval,with the exception of Cramer et al[16] who posed a prediction task but only considered structured data and not notes. One Korean study[13] developed a decision support tool based on a Bayesian network risk model, resulting in a significant incidence reduction from 21% to 4%. This finding supported the usefulness of PUI risk alert tools. Unfortunately, this tool used only structured EHR data such as billing codes, which cannot provide real-time and accurate detection for PUI. Moreover, the model performance could be further improved by including unstructured data such as nursing notes, which contain useful patient information.

In practice, responsibilities for PUI prevention and treatment fall to the nurse. Yet nurse-collected data is not actively mined for valuable patient information. For example, nursing notes can contain changes in patient vital signs, symptoms, and care and thus offer more predictive power for PUI detection compared to structured data.

However, unstructured data is often hard to mine and unlikely to be used routinely for decision-making. For example, the National Pressure Ulcer Adversary Panel developed a template with the needed documents to facilitate discovery

of severe PUI development through a review of the timeline of events.[17] In this 18-page general template, most PUI risk factors could only be captured by unstructured data such as skin outlook descriptions, re-positioning, and support surface as documented in nursing notes. Unfortunately, the template only serves as guidance or an educational tool.

Therefore, it would be beneficial if PUI-related unstructured data, especially nursing notes, could be leveraged for PUI prediction prior to its occurrence for early detection, to inform nurses to implement appropriate interventions in time. Our proposed approach aims to address these challenges, by setting up a PUI detection pipeline that utilizes hospital notes after a negation-aware processing step, as an input to a classifier for detection of PUI. Our evaluation of 3,589 hospital-stays suggests the negation-aware processing step improves the predictive performance. Moreover, this automated PUI detection yields a clearer picture of its incidence in real-time, and is also interpretable in a general sense, i.e. we can pinpoint specific keywords that play the most important role in the occurrence of PUI.

Methods

Here we describe the dataset, details of our cohort selection and our proposed reliable proxy for ground truth of PUI incidence in hospital-stays. We then move on to describe the methodology used.

MIMIC-III Dataset

To ensure the replicability of our experiments and results, we used the openly available MIMIC-III dataset.[18] This dataset holds information of patients admitted to intensive care units (ICU) of a populated tertiary care hospital from 2001 to 2012. There are 49,785 unique hospital admissions for patients aged 16 years and older. These records come from 38,597 unique adult individuals.[18] MIMIC-III is one of the only benchmarks datasets in machine learning for healthcare. The data span multiple tables, linking demographics, lab results, diagnosis, notes, and medications associated with hospital-stays, of which only the unstructured notes are used in this paper as features for prediction.

Selection of an Appropriate Cohort

We chose hospital stays as the unit of analysis, to reflect the real-world assessment of all the chart for a stay by nurses. After removing hospital-stays with illogical attributes, e.g. those with a negative length of stay, about 50k unique stays remained. Since in MIMIC-III comparatively very few positive cases of PUI were present in the younger population, hospital-stays of individuals 20 years and younger were removed. PUI happens in younger population with spinal cord injuries, however there were negligible number of such stays in our data. We restricted our analysis to stays longer than 2 days and shorter than 120 days, since shorter stays provide insufficient notes for satisfactory representation, and extremely long stays are most probably erroneous records in this specific dataset. This further cuts down the number of unique hospital-stays to about 26K. The available notes for each stay, therefore, will be its features. For our detection task, we concatenated all the different categories of notes found for each hospital admission into a single document.

Establishing Presence of PUI

Two sources of information for each hospital-stay, ICD-9 diagnosis codes, and notes, are used to determine the presence or absence of PUI. A hospital-stay is indicative of PUI from an ICD-9 perspective if any of the PUI ICD-9 codes in Definition 1 are found in its diagnosis codes. Similarly if any of the explicit PUI keywords in Definition 2 or string versions of ICD-9 codes in Definition 1 appears in the notes, that stay is indicative of PUI from a notes perspective. Hospital-stays that indicate PUI in both sources constitute our positive class and those with no indication of PUI in either, will be labeled as negative. To avoid ambiguity in our dataset, we discard stays that indicate PUI only in notes or only in ICD-9 codes, as reading through notes of a few sample of such positive for notes but negative for ICD-9 codes stays, revealed they aren't always indicative of PUI in that stay, e.g. the PUI keyword found in notes were actually referring to a previous hospital stay. Note that ICD9 codes at dishcharge time are only used for establishing labels and are never used as features in the prediction.

Definition 1 (PUI ICD-9 Codes)
[707, 707.1, 707.2, 707.3, 707.4, 707.5, 707.6, 707.7, 707.9, 707.11, 707.21, 707.22, 707.23, 707.24, 707.25]

Definition 2 (PUI Explicit Keywords)
[Pressure Ulcer Prevention, Skin Surveillance, Decubitus Ulcers, Impaired Tissue Integrity, Impaird Skin Integrtiy, Bed Sores, Pressure Ulcer, Pressure sore]

Determination of PUI Case/Control Samples

Given that the ratio of positive samples (PUI) to negative samples (no PUI) is extremely low (3.5%), heavily inhibiting the learning algorithm's ability in distinguishing the two, we resorted to using a case-control design as a solution. Case-control study designs are commonly used in biomedical research and have also been proposed in Artificial Intelligence for healthcare.[19] A given positive sample is matched with 4 negative samples (stays), closest to it in terms of age, gender, the total length of stay and ICU length of stay. We chose the number of potential negative matches to be 4 to account for some diversity in matched stays' properties while maintaining some consistency in the pool of samples. To perform this matching, a 4 nearest neighbor algorithm was trained with all the negative samples, and then for each stay positive for PUI, the 4 closest negative samples without PUI were added to the pool. Negative samples were not enforced to be unique for each positive sample, and some negative samples might be matched with multiple positive ones, i.e the selection process happened with replacement.

Using the above criteria for positive and negative samples, 856 stays were marked as positive for PUI and using the case-control study, 2733 negative stays for PUI were chosen for our experiments. Our final cohort consists of 3589 hospital samples, with a 31.3% positive to negative sample ratio.

Data Analysis

Medical documents often contain terms that only when considered in the context of a sentence, can be interpreted as a sign for the presence/absence of a condition.[20] Therefore, we propose a negation-aware processing step on hospital-stays' aggregated notes. Our process identifies the negative mention of conditions based on the sentence context, putting a negation prefix right before its mention in the processed notes (e.g. no_edema) to denote a different word. After the negation, we transforms each hospital-stay's aggregated nursing notes (both before and after negation-aware processing step) into a vectorized feature representation. The vectorized feature representation is then used for predicting the presence of PUI. We explored three different classifiers to investigate the following two questions:

- What is the impact of our proposed negation-aware framework on the performance of PUI detection?
- What are the most salient text features and do the features overlap with known medical factors of PUI?

Figure 1 provides an overall illustration of our proposed methodology.

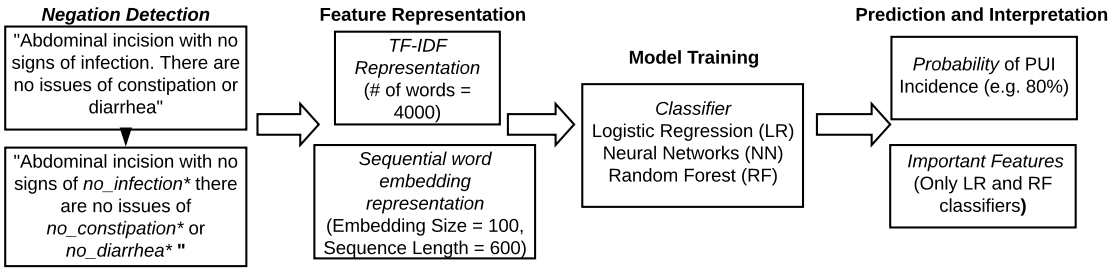

Figure 1: Overview of PUI Detection: Negation Detection, Model Training, Interpretation

Negation Detection in Text Data

The mere existence of a term in clinical notes does not necessarily indicate the presence of the clinical condition suggested by the term. Terms and phrases can be used to rule out the existence of a particular disease in a patient. In fact, it is common for healthcare providers to describe the absence of specific patient findings in clinical notes. For example, the sentence "showed no evidence of congestive heart failure or pneumonia" rejects the existence of congestive heart failure and pneumonia. Unfortunately, standard text processing fails to handle negations. As a result, in many natural language processing pipelines, these can easily be mistaken for the presence of the conditions.

Before applying our negation detection scheme we clean the notes syntactically. All punctuation signs are removed except for full stop and colon which are used as a boundary for sentence segmentation. Standard stop-words that do not indicate the existence of a negation tone (words such as not and doesn't) are removed. We also remove all kinds of tab characters, end of line carriage return characters, consecutive multiple white space characters, and all special characters used to anonymize personal data such as [****]. Finally all the charachters in the text are lower-cased.

To determine whether diseases or symptoms mentioned in the clinical notes were negated by the dictating physician, we combined two clinical text processing methods, Scispacy[21] and NegEX.[20] Scispacy extracts all mentions of named entities including disease, medication, symptoms, and chemicals. NegEX uses defined regular expressions that cover several phrases indicating negation, filters out sentences with phrases that falsely appear to be negation phrases, and limits the scope of the negation phrase. First, our text processing determines the sentence boundaries in each text using the full stop and colon as the boundary to limit the scope of the negation. The mentions of named entities in each sentence are identified using Scispacy. NegEx is then run on the sentence to determine which named entities should be negated. As an example of our preprocessing step, the above example will be replaced by "showed evidence of no_congestive_heart_failure or no_pneumonia". Thus, our negation-aware processing step helps with the recognition of positive and negative mentions of a condition, which we expect to enhance the predictive ability of our algorithm.

Transforming Text into Vectorized Features

For classification, we utilize two common standard text vector representations. A vectorized form of a hospital-stay's aggregated notes was created using either the term frequency-inverse document frequency (TF-IDF) representation or a sequential word embedding representation.

Term Frequency–Inverse Document Frequency (TF-IDF representation)

A commonly used representation is the TF-IDF representation, which assigns weights to each unique word in the document. The weight accounts for the number of times a word appears in a document (or hospital-stay) and also adjusts for the frequency of the words in the overall corpus (across all hospital-stays). Thus, words that occur more frequently in a document will have a higher impact while rare words will have little influence. However, words that appear frequently across many documents will be less important. Under this representation, the sequence of words in the notes is not modeled as each document is viewed as a collection of words.

Sequential Word Embedding Representation

A major limitation of TF-IDF is the inability to preserve the word order in the text. To address this, the sequential word embedding representation assigns each word token a unique token-number that preserves the word order in the text. A maximum number of words are specified for each document, and shorter documents are padded with zeros while longer ones are trimmed to the maximum number of words. These vectorized text representations were later passed to an embedding input layer where a dense representation of the word is learned.

PUI Classifiers

After curating the 2 versions of nursing notes (raw notes and negation-detected notes) and transforming the text to the appropriate vectorized representation, we predicted PUI incidence within each stay using only these features. We explored three different classifiers: logistic regression, random forest, and neural networks. Below, an overall view of each classifier is provided.

Logistic Regression: Interpretable and Intuitive

The logistic regression (LR) classifier assumes a Bernoulli distribution between the features and target classes and is parameterized by a vector of weights W. Given the feature matrix X, and their corresponding target classes Y, the optimal values for weights W are found to minimize the mismatch between predicted labels $\hat{Y}$ and Y. LR is chosen since it is a straightforward and highly interpretable model. There is a one-to-one mapping from weights to features, which can be interpreted as the relative contribution (importance) of that feature to LR's decision toward the positive class. The TF-IDF feature representation is used as input for the LR classifier.

Random Forest: Trading less Interpretability for better Performance

The random forest (RF) classifier combines multiple learners for more accurate predictions. More specifically, RF combines multiple decision trees trained across different subsets of features and samples. For the final classification, the weighted average probability of belonging to the positive class across all the decision trees is used. Due to its ensemble nature, it often achieves better predictive performance. RF also calculates feature importance based on the number and level of splits made with each feature across all the trees, however, the exact contribution is not readily apparent. The TF-IDF feature representation is also used as input for the RF classifier. In summary, applying RF on our PUI detection task can attain better performance at the expense of losing some interpretability.

Neural Networks: Many Parameters, Data Hungry

A neural network-based (NN) classifier attempts to find a non-linear decision boundary by connecting multiple layers of units connecting the features to the target classes. Each unit's output is computed as a weighted sum of its input that is passed through a local activation function. NN classifiers have been increasingly adopted for text data due to their stellar success.[22] We note that some NN architectures are more suitable than others for certain application domains and kinds of data. We fed the sequence of notes' words embedding vectors to a sequential model consisting of input word embedding, global max pooling layer, and several dense layers, with the output layer yielding the probability of PUI incidence. We also tried the long short-term memory (*LSTM*) network, given the sequential nature of the text. However, we excluded it from the results due to its poor performance.

Experimental Setup

Here we outline our experiments along with measures taken to ensure the robustness of our results, provide detailed parameters of classifiers, and describe our metric for evaluation. Finally, we report how feature importance pertaining to each classifier is leveraged to reveal keywords specific to its criteria for distinguishing PUI vs no PUI, where applicable. These extracted keywords for classifiers pave the way for more intuitive comparison across classifiers, and more exploration of these keywords' compatibility with clinical characteristics of PUI. The implementation and the Python code for all our experiments is available on this repository[1].

Experiments Overview and Data Split

Our classification task is predicting the incidence of PUI in our test data given the training set. We used stratified sampling (i.e., prevalence of PUI is maintained) to obtain three splits: 68% training, 12% validation and 20% test data. To assure our comparisons across the various predictive models and input data were generalizable, we repeated each experimental setting 30 times (i.e., over 30 different splits of training/validation/test). For a given split, all three classifiers were trained on the same training data, their hyperparameters were tuned on the same validation set, and made decisions about the labels of the same test data. Once the optimal hyperparameters for each classifier are found using the validation set, the classifier is retrained with these parameters on the training and validation data. For the final comparison, we report the average and variance of the predictive performance on the test data across all 30 splits.

Given our constructed versions of collapsed notes of each stay, untouched notes as they're found in MIMIC-III, and processed notes using negation detection, and the three chosen classifiers Logistic Regression (LR), Random Forest (RF), and Neural Network (NN), we will have a total of 6 experimental settings.

Evaluation Metric

Our chosen cohort is unbalanced in terms of positive and negative cases of PUI. Thus, to truly capture the performance of our classifiers in different settings, we report the Area Under the Receiver Operating Characteristic Curve (AUC ROC) and F1 score on the test set. AUC ROC is based on the area of True Positive Rate (TPR) and False Positive Rate (FPR) curve across different classification thresholds in the range 0 to 1.The F1 score is computed as the harmonic mean of precision and recall (TPR) using a 0.5 classification threshold.

[1] https://github.com/manisci/PU_Detection_Notes

Classifier-Specific Hyperparameter Tuning

The optimal hyperparameters for the classifiers in each experimental setting are determined based on the best AUC ROC on the validation dataset.

- LR: We used the *sklearn*[23] *LogisticRegression* library, with *class_weight = "balanced"*, *solver = "lbfgs"* and *max_iter=10000*. We performed cross validation over 6 different values of regularization strength c, in the range 0.01 to 0.5.
- RF: We used the *RandomForestClassifier* module of *sklearn*[23] with *class_weight="balanced"*. We explored 6 settings of the *max_depth* of the trees in the range of 2 to 7.
- NN: We used a sequential NN model with the following layers: an embedding layer with *input_dim=4000*, output dimension of 100, and *input_length=600*; a 1-dimensional global max pooling layer; a dense layer with 32 units using with the rectified linear unit (*relu*) activation function; and an output layer using the *sigmoid* activation function. We train the model with the *keras* [24] library over 10 epochs using *Adam* optimizer, *binary_crossentropy* loss function and accuracy as a metric. We performed a grid search for the best *batch_size* using 5 values in range 8 to 128.

Inferring Word Significance from Feature Importance

We assess the feature importance of specific words using the final trained model. Note that this is only applicable to LR and RF models, as NN requires additional methods to extract feature importance. For LR, we use the learned weights as indicators of feature importance. By sorting the weights in descending order, the features positively correlated with a PUI stay appear in the beginning, while the ones associated with the absence of PUI appear at the end. For RF, we report the feature importance that is calculated using the number and level of splits made on each feature across all decision trees. In contrast to LR however, we only know the relative importance of each feature in the model's classification but do not know the direction of correlation.

Results and Discussion

We provide our evaluation results for the 2 versions of data (raw notes from MIMIC-III and the negation-aware notes) and the 3 chosen classifiers (LR, RF, and NN) and discuss the trends observed. We first assess the merit of the negation detection step for PUI detection in multiple experimental settings and compare the performance of classifiers in predicting PUI. A comparison of extracted significant words with and without applying negation detection is presented to shed more light on the reasons for better performance of the negation detection step. Lastly, we discuss how relevant some of the significant words are to known medical contributors and/or certain comorbidities of PUI.

Impact of Negation Detection on AUC ROC and F1 Score for PUI Detction

Table 1 presents the average AUC and F1 score of LR, NN, and RF along with their standard deviation across the 30 splits. From the results, the negation detection step leads to AUC and F1 score improvements in all three classifiers. The greatest gain belongs to LR (around 3% and 5% for AUC and F1 respectively), followed by NN and RF. We performed a one-sided paired t-test for each classifier to determine whether the improvement had a p-value below 0.05. The p-value for LR was less than machine precision for both metrics, while for NN and RF it was 0.2309 (0.0802) and 0.3826 (0.2599) (p-value for F1 is reported in parentheses) . This further portrays the greatest utility of negation detection for improving LR performance relative to improvement for NN and RF.

We posit the small predictive improvement in RF is due to the fact that the negation words are minority features. Since RF performs random subsampling for each tree, these minority negative words are even less likely to be included in individual trees. However, in the NN and LR models, there is no random feature subsampling, thus these same minority negative words have a higher chance of inclusion in the final model. This is also further supported by comparing the number of negative words among the important features when negation detection is used for these three classifiers.

Table 1: Average AUC and F1 score of classifiers with and without negation detection over 30 runs. * denotes a p-value < 0.05 under a one-sided paired t-test.

Classifier	**Average Test AUC *(SD)***		**Average Test F1 *(SD)***	
	Negation-Aware	w/o Negation Detection	Negation-Aware	w/o Negation Detection
Neural Networks (NN)	0.8462 *(0.0169)*	0.8440 *(0.0161)*	0.6252 *(0.0291)*	0.6189 *(0.0302)*
Logistic Regression (LR)*	0.9022 *(0.0120)*	0.8720 *(0.0155)*	0.6905 *(0.0188)*	0.6455 *(0.0248)*
Random Forest (RF)	0.9533 *(0.0086)*	0.9530 *(0.0071)*	0.7887 *(0.0226)*	0.7862 *(0.0219)*

Comparison of Predictive Performance across Classifiers

A comparison of the test AUC and F1 scores in Table 1 reveals that NN performs the worst. This is likely due to over-fitting as NN has an exponential number of parameters with respect to the number of layers and units. Given our small cohort, we anticipated NN to not perform well, as confirmed by our results. Also, NN is not easily interpretable due to the enormous number of parameters, making it unappealing for exploratory analysis of factors leading to PUI. Another drawback is that NN's run time is considerably higher than both LR and RF (by a factor of 2). Moreover, it is unlikely to have many positive PUI cases available for training, due to PUI's low incidence rate. This combined with privacy concerns for data collaboration across hospitals suggests that the predictive model should be robust even when trained on a smaller number of data samples. In summary, LR's decent performance combined with its greatest interpretability across the 3, makes it a truly appealing choice for PUI detection. However, if we are willing to forego some interpretability, ensemble models such as RF may have better predictive performance as evidenced by our results.

Impact of Negation Detection on Extracted Important Features (Words) from Models

Next, we compare the extracted significant words' lists for the original notes and negation-aware notes. In particular, we focus on the significance of words for the negative class in the two versions of data, and especially those containing the prefix *no_* in the negation-aware version. These are of special interest since they highlight the efficacy of our proposed negation detection scheme. The top 10 most influential words by feature importance alluding to the absence or presence of PUI are presented for both untouched notes and negation-aware notes. Table 2 summarizes the words for both LR and RF.

We first observed that among the 10 most important words indicating no PUI in LR and RF, 5 and 4 words respectively are the direct product of negation detection step, proving its utility. Moreover, the words from the negation-aware version have comparatively higher weights than their untouched notes counterparts (-0.2566 vs -0.1955 for LR, and 0.0067 vs 0.00038 for RF). This means the 10 features have a higher correlation (e.g., for logistic regression higher log-odds) with the outcome. Furthermore, some of the words appearing in the top 10 for the model trained on untouched notes are solely non-specific general descriptors, which is rare in negation-aware versions (e.g. all words in RF untouched notes case except "ganx" and "fio2" are non-specific).

Inferred Salient Features and their Overlap with Leading Medical Factors of PUI

We investigate how closely the most salient contributing words resemble known medical covariates of PUI. After review by our diverse team of computer, nursing, biostatistics, and health informatics scientists, we present the high importance keywords that are aligned with the established evidence on PUI guidelines including the definition, staging, Braden Scale, personalized algorithms, and root cause analysis template. Table 3 shows these keywords extracted from the model for different experimental settings of note version, classifier, and the direction of the words contribution. For example, the keyword *no_erythema* is consistent with the updated definition of "Stage 1 Pressure Injury: Non-blanchable erythema of intact skin".[1] The keywords *swangaz* (abbreviation for "Swan-Ganz catheterization") and *no_tube* indicate PUI related to medical devices.[1] Keywords such as *sedate* or *PACU* (post anesthesia care unit); *no_secretions* or *no_stool*; *independent*; *no_obstruction* and *multipodus* are related to the Braden risk categories of sensory perception, moisture, mobility or activity, nutrition as well as friction and shear respectively.[12] Also, Key-

Table 2: Top 10 most important features (words) in different experimental settings

Classifier	Found only in	Indicating PUI	Words in the Set *(importance)*
LR	Negation-aware notes	Absence	{mso *(-0.3182)*, groundglass *(-0.2953)*, swanganz *(-0.2915)*, preoperative *(-0.2730)*, no_ectopy *(-0.2632)*, no_edema *(-0.2560)*, independent *(-0.2531)*, no_sob *(-0.2137)*, no_pneumothorax *(-0.2107)*, number *(-0.1918)*, no_pulmonary *(-0.1881)*}
LR	Untouched notes	Absence	{ganz *(-0.3563)*, mso4 *(-0.3431)*, lat *(-0.3431)*, ward *(-0.2286)*, lima *(-0.1443)*, hyperthermia *(-0.1169)*, pepcid *(-0.1156)*, neoplasm *(-0.1043)*, Sao2 *(-0.1023)*, pyrexia *(-0.1006)*}
RF	Negation-aware notes	Presence or Absence	{no_wound *(0.0016)*, apply *(0.0011)*, multipodus *(0.0011)*, swanganz *(0.0008)*, sch *(0.0005)*, clip *(0.0004)*, [no_skin, no_pneumothorax, unit, no_infection] *(all 0.0003)*}
RF	Untouched notes	Presence or Absence	{lat *(0.0007)*, ptitle *(0.0006)*, ganz *(0.0005)*, name *(0.0004)*, [followup, numeric, lastname, identifier] *(all 0.0003)*, [fi02, defined] *(all 0.0002)*}

words *diuresis* and *multipodus* are consistent with recently identified predictive features for PUI in the intensive care unit.[16] Diabetes glycemic control indicated by *no_insulin*, *no_hypotension* and *no_infection* were also related to the risky comorbidities in the root cause analysis template.[17] Although some keywords with relatively high importance did not stand out in the past evidence, they could inform future PUI research directions. For example, *no_her* could indicate gender difference.

Table 3: Most medically meaningful keywords for different experimental settings

Classifier	Type of Words (or notes)	Indicating PUI	Most Medically Meaningful Keywords in the Set *(importance)*
LR	Only no_ .... words	Presence	{no_wound *(0.2951)*, no_erythema *(0.2303)*, no_skin *(0.1629)*, no_infection *(0.1420)*, no_obstruction *(0.1403)*, no_ct *(0.1291)*, no_secretions *(0.0850)*, no_lesions *(0.0728)*}
LR	Only no_ .... words	Absence	{no_edema *(-0.2560)*, no_stool *(-0.1241)*, no_pain *(-0.0923)*, no_diuresis *(-0.0744)*, no_hemorrhage *(-0.0677)*, no_bleed *(-0.0589)*, no_bleeding *(-0.0523)*}
LR	Only Negation-aware	Absence	{swanganz *(-0.2915)*, no_edema *(-0.2560)*, independent *(-0.2531)*, pacu *(-0.1765)*, bloodtinged *(-0.1314)*, no_stool *(-0.1241)*}
RF	Only no_ .... words	Presence or Absence	{no_wound *(0.0016)*, [no_skin, no_tube] *(0.0003)*, [no_infection, no_blood, no_insulin, no_erythema, no_hypotension] *(all 0.0002)*, [no_diuresis, no_abscess, no_pain, no_stool, no_edema, no_gtt] *(all 0.0001)*}
RF	Only Negation-aware	Presence or Absence	{no_wound *(0.0016)*, multipodus *(0.0010)*, [no_skin , no_infection , no_tube, no_insulin] *(all 0.0003)*, sedate *(0.0002)*}

Implications for Nursing Practice

Given our promising results, from the perspective of efficient nursing quality assurance, the predicted incidence of PUI is a proxy for its incidence in reality. To achieve more reliable PUI detection, we can present nurses with only the most uncertain cases (for instance, those with around 50% probability to develop PUI) for confirmation, rather than overwhelming nurses with all hospital-stays. Using such an approach has the potential to exponentially improve efficiency and outcomes for nurse-driven plans of care. Once PUI is detected through machine learning, the RN has the knowledge to direct caregivers in the tactical care of the patient for eliminating the impact of PUI on the patient.

Conclusion and Future Work

Detection of PUI in the clinical setting for quality assurance purposes and improving care by timely interventions, is a critically valuable tool in nursing care. The current practice of quarterly assessments of one hospital in a single day by supervisor nurses has many disadvantages. High cost, subjectivity, substantial disagreement among nurses, missed opportunities to instantly alter practices leading to inadequate care are all areas needing improvement in PUI detection. Thus, a real-time and accurate PUI detection tool is necessary to inform nurses about required appropriate interventions.

Our proposed method addresses many of these hurdles by leveraging collective notes acquired from the entire care team in one hospital-stay. Despite their informativeness, clinical notes often do not get sufficient attention in clinical decision-making due to their multiple sources and unstructured form. We first proposed a negation detection step for notes that moves the presence or absence of a medical conditions closer to their mention in a sentence. Through experimental results with three representative classifiers for the PUI detection task, we showed the efficacy of the negation detection method in improving models' predictive performance. We further teased apart the keywords in notes, and analyzed the keywords contributing the most to the detection of PUI and their promising overlap with medical knowledge on PUI. We also showcased that many of the negated condition keywords were actually among the most important words for PUI detection. In summary, we depicted how our approach can be applied to hospital notes for detection of PUI as a tool that is accurate, fast, on-demand, and highly reflective of current medical knowledge.

There are potential limitations to our work. Since our focus was not solely on attaining the most accurate model, even though the prevalence of PUI in our data is not the true prevalence, we believe our exploratory studies and its insight are still relevant. Also, the detection performance can be improved further by additional hyperparameter tuning and utilizing more complex machine learning models. In particular, due to the large search space and the sheer number of parameters in NN, extensive hyperparameter tuning may have yielded a better predictive performance. Since our goal was to explore the model's features and its interpretability as well, we limited our search space and settled on a reasonably well-performing model.

In terms of future directions for research in this area, we plan to use an extension of the tool we used for more accurate identification of medical terms called ConText in the next steps and potentially tailor it to a particular datset as well. Moreover, our model was not designed to predict PUI prior to the occurrence where effective nursing and medical interventions might prevent PUI emergence.[13] We also did not attempt to detect the PUI's stage, a topic of high interest among clinicians.[25] Braden scale was not incorporated into our modelling, as either a baseline or additional features, since a recent study[16] showed poorer performance in either case on the same dataset, however, investigating its reason may be insightful. These are all open areas of research, some of which we hope to tackle in our future work.

Acknowledgement.

This research was supported by Natinal Library Of Medicine of the National Institutes of Health under award number R01LM013323-01.

References

1. Edsberg LE, Black JM, Goldberg M, McNichol L, Moore L, Sieggreen M. Revised National Pressure Ulcer Advisory Panel Pressure Injury Staging System: Revised Pressure Injury Staging System. Journal of wound, ostomy, and continence nursing : official publication of The Wound, Ostomy and Continence Nurses Society. 2016;43(6):585–597.
2. National Quality Forum. National Voluntary Consensus Standards for Developing a Framework for Measuring Quality for Prevention and Management of Pressure Ulcers; 2011.
3. Agency for Healthcare Research and Quality. Preventing Pressure Ulcers in Hospitals. Content last reviewed October 2014;. Available from: `https://www.ahrq.gov/patient-safety/settings/hospital/resource/pressureulcer/tool/index.html`.
4. National Pressure Ulcer Advisory Panel; European Pressure Ulcer Advisory Panel. Prevention and Treatment of Pressure Ulcers: Clinical Practice Guideline; 2012.
5. VanGilder C, Amlung S, Harrison P, Meyer S. Results of the 2008-2009 International Pressure Ulcer Prevalence Survey and a 3-year, acute care, unit-specific analysis. Ostomy/wound management. 2009 nov;55(11):39–45.
6. Black JM, Cuddigan JE, Walko MA, Didier LA, Lander MJ, Kelpe MR. Medical device related pressure ulcers in hospitalized patients. International wound journal. 2010 oct;7(5):358–365.
7. Reddy M, Gill SS, Rochon PA. Preventing pressure ulcers: a systematic review. JAMA. 2006 Aug;296(8):974–984.

8. Whittington KT, Briones R. National Prevalence and Incidence Study: 6-year sequential acute care data. Advances in skin & wound care. 2004;17(9):490–494.

9. Brem H, Maggi J, Nierman D, Rolnitzky L, Bell D, Rennert R, et al. High cost of stage IV pressure ulcers. American journal of surgery. 2010 oct;200(4):473–477.

10. Russo C, Steiner C, Spector W. Hospitalizations Related to Pressure Ulcers Among Adults 18 Years and Older, 2006: Statistical Brief #64. 2008 Dec. In: Healthcare Cost and Utilization Project (HCUP) Statistical Briefs. Rockville (MD): Agency for Healthcare Research and Quality;. Available from: https://www.ncbi.nlm.nih.gov/books/NBK54557/.

11. Centers for Medicare and Medicaid Services. Readmissions Reduction Program.; 2015.

12. Bergstrom N, Braden BJ, Laguzza A, Holman V. The Braden Scale for Predicting Pressure Sore Risk. Nursing research. 1987;36(4):205–210.

13. Cho I, Park I, Kim E, Lee E, Bates DW. Using EHR data to predict hospital-acquired pressure ulcers: a prospective study of a Bayesian Network model. International journal of medical informatics. 2013 nov;82(11):1059–1067.

14. Ranzani OT, Simpson ES, Japiassú AM, Noritomi DT. The Challenge of Predicting Pressure Ulcers in Critically Ill Patients. A Multicenter Cohort Study. Annals of the American Thoracic Society. 2016;13(10):1775–1783. Available from: https://doi.org/10.1513/AnnalsATS.201603-154OC.

15. Moore ZEH, Patton D. Risk assessment tools for the prevention of pressure ulcers. Cochrane Database of Systematic Reviews. 2019;(1). Available from: https://doi.org//10.1002/14651858.CD006471.pub4.

16. Cramer E, Seneviratne M, Sharifi H, Ozturk A, Hernandez-Boussard T. Predicting the Incidence of Pressure Ulcers in the Intensive Care Unit Using Machine Learning. eGEMs (Generating Evidence & Methods to improve patient outcomes). 2019 sep;7.

17. National Pressure Ulcer Advisory Panel. NPUAP Pressure Ulcer Root Cause Analysis Template;. Available from: http://www.hret-hiin.org/resources/display/npuap-pressure-ulcer-root-cause-analysis-template.

18. Johnson AE, Pollard TJ, Shen L, Li-wei HL, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Scientific data. 2016;3:160035.

19. Gyöngyösi M, Ploner M, Porenta G, Sperker W, Wexberg P, Strehblow C, et al. Case-based distance measurements for the selection of controls in case-matched studies - application in coronary interventions. Artif Intell Medicine. 2002;26(3):237–253.

20. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. Journal of biomedical informatics. 2001;34(5):301–310.

21. Neumann M, King D, Beltagy I, Ammar W. Scispacy: Fast and robust models for biomedical natural language processing. arXiv preprint arXiv:190207669. 2019;.

22. Young T, Hazarika D, Poria S, Cambria E. Recent trends in deep learning based natural language processing. ieee Computational intelligenCe magazine. 2018;13(3):55–75.

23. Buitinck L, Louppe G, Blondel M, Pedregosa F, Mueller A, Grisel O, et al. API design for machine learning software: experiences from the scikit-learn project. In: ECML PKDD Workshop: Languages for Data Mining and Machine Learning; 2013. p. 108–122.

24. Chollet F. keras. GitHub; 2015. https://github.com/fchollet/keras.

25. Coomer NM, McCall NT. Examination of the accuracy of coding hospital-acquired pressure ulcer stages. Medicare & medicaid research review. 2013;3(4).