Published in final edited form as:
Int J Med Inform. 2020 March ; 135: 104053. doi:10.1016/j.ijmedinf.2019.104053.

Mining clinical phrases from nursing notes to discover risk factors of patient deterioration

Zfania Tom Korach[a,b,*], **Jie Yang**[a,b], **Sarah Collins Rossetti**[c,d], **Kenrick D. Cato**[d], **Min-Jeoung Kang**[a,b], **Christopher Knaplund**[d], **Kumiko O. Schnock**[a,b], **Jose P. Garcia**[a], **Haomiao Jia**[d], **Jessica M. Schwartz**[d], **Li Zhou**[a,b]

[a]Division of General Internal Medicine and Primary Care, Brigham and Women's Hospital, Boston, MA, United States

[b]Harvard Medical School, Boston, MA, United States

[c]Department of Biomedical Informatics, Columbia University, New-York, NY, United States

[d]School of Nursing, Columbia University, New-York, NY, United States

Abstract

Objective: Early identification and treatment of patient deterioration is crucial to improving clinical outcomes. To act, hospital rapid response (RR) teams often rely on nurses' clinical judgement typically documented narratively in the electronic health record (EHR). We developed a data-driven, unsupervised method to discover potential risk factors of RR events from nursing notes.

Methods: We applied multiple natural language processing methods, including language modelling, word embeddings, and two phrase mining methods (TextRank and NC-Value), to identify quality phrases that represent clinical entities from unannotated nursing notes. TextRank was used to determine the important word-sequences in each note. NC-Value was then used to globally rank the locally-important sequences across the whole corpus. We evaluated our method both on its accuracy compared to human judgement and on the ability of the mined phrases to predict a clinical outcome, RR event hazard.

Results: When applied to 61,740 hospital encounters with 1,067 RR events and 778,955 notes, our method achieved an average precision of 0.590 to 0.764 (when excluding numeric tokens).

[*]Corresponding author at: Brigham and Women's Hospital, Department of Medicine, 399 Revolution Dr., Somerville, MA 02145, United States. zkorach@bwh.harvard.edu (Z.T. Korach).

Authors' contributions

Zfania T. Korach designed the study protocol and adapted methods and implemented them. Jie Yang participated in the method adaptation. Sarah Collins Rossetti, Kenrick D. Cato, Christopher Knaplund and Kumiko O. Schnock provided clinical guidance. Min-Jeoung Kang and Jose P. Garcia participated in data collection and validation. Haomiao Jia provided statistical guidance on survival analysis. Jessica M. Schwartz and Min-Jeoung Kang conducted manual annotations. Li Zhou and Sarah Collins Rossetti oversaw the study as principal investigators.

Transparency document

The Transparency document associated with this article can be found in the online version.

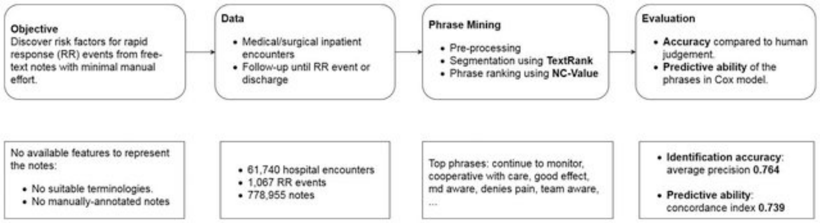
Declaration of Competing Interest

Dr. Korach reports grants from National Institute of Nursing Research, grants from CRICO, during the conduct of the study; personal fees from K Health, ownership of Clirq LLC, outside the submitted work.. Jessica Schwartz is a Pre-doctoral fellow funded by the National Institute of Nursing Research Reducing Health Disparities Through Informatics (RHeaDI) T32NR007969. Dr. Zhou reports grants from Agency for Healthcare Research and Quality, grants from CRICO, during the conduct of the study.

Time-dependent covariates Cox model using the phrases achieved a concordance index of 0.739. Clustering the phrases revealed clinical concepts significantly associated with RR event hazard.

Discussion: Our findings demonstrate that our minimal-annotation, unsurprised method can rapidly mine quality phrases from a large amount of nursing notes, and these identified phrases are useful for downstream tasks, such as clinical outcome predication and risk factor identification.

Graphical Abstract


Objective
Discover risk factors for rapid response (RR) events from free-text notes with minimal manual effort.
Data
Medical/surgical inpatient encounters
Follow-up until RR event or discharge
Phrase Mining
Pre-processing
Segmentation using TextRank
Phrase ranking using NC-Value
Evaluation
Accuracy compared to human judgement.
Predictive ability of the phrases in Cox model.
No available features to represent the notes:
No suitable terminologies.
No manually-annotated notes
61,740 hospital encounters
1,067 RR events
778,955 notes
Top phrases: continue to monitor, cooperative with care, good effect, md aware, denies pain, team aware, ...
Identification accuracy: average precision 0.764
Predictive ability: concordance index 0.739

Keywords

Data mining; Nursing informatics; Hospital Rapid response team; Unsupervised machine learning

1. Introduction

Timely identification of patient deterioration is crucial to patient safety. Rapid response (RR) teams are charged with responding to non–intensive care unit (ICU) patients at risk for rapid deterioration. They function as safety nets with both observational component of deterioration detection and interventional component of providing critical care resources and interventions at the patient’s current location. [1] The identification of imminent clinical deterioration and prompt treatment was demonstrated to reduce mortality [2,3]. Existing approaches to RR detection mainly focus on structured information from flowsheets and measurements. However, in addition to objective measures, the triggers for RR typically include a subjective component such as “Staff member is worried about the patient”, which might be recorded only in the narrative texts in the patient charts [4]. Studies have shown that nurses’ concern is an important indicator that a patient’s condition is likely deteriorating [5]. Therefore, prediction of RR events might benefit from analysis of narrative nursing documentation.

While clinically important, the subjective criterion encompasses a multitude of clinical findings, and the reporting clinician might even lack a clear culprit [6]. Currently, no existing nursing terminology captures diverse expressions documented in their notes, that convey a nurse’s concern about a patient’s conditions that may be associate with clinical outcomes. In addition, free-text notes cannot be used as-is for statistical modelling. Rather, they need to be transformed to numerical values in a process called “feature engineering”. In addition to the required effort, feature hand-crafting (e.g. deciding what signs and symptoms to extract from the notes) poses a challenge to exploratory studies that look to elucidate new associations with potentially unknown factors.

As part of a larger research study, CONCERN, which investigates nurses' judgment, in both narrative and structured information, about impending deterioration of acute and critical care patients, we attempted to discover risk factors of RR events from nursing documentation [5]. In this study, we leveraged data-mining methods to overcome the aforementioned challenges to clinical free-text analysis and tested both their accuracy and usefulness for RR event prediction.

1.1. Background and related work

The basic natural language processing (NLP) approaches to document representation for statistical modelling, including bag-of-words (BoW; occurrence of individual words irrespective of order) and bag-of-*N*-grams (BoNG; occurrence of word sequences typically 2–4 words long), suffer from inherent deficiencies. BoW cannot model the word order or multi-word concepts. BoNG suffers from the "curse of dimensionality", where the number of N-grams grows exponentially with length to millions of features, most of which appear only rarely. Such high number of features impairs both the statistical strength of the analysis and the ability to interpret and manage the extracted features. To overcome these problems, machine-learning (ML) methods may use *distributed representations*, low-dimensional (hundreds to thousands of elements) vectors of numerical values that captures the content of words, sentences and documents. While successful in many NLP tasks, these vectors are not interpretable, and their numeric values bear little relationship to conventional clinical concepts [7]. With the increasing adoption of ML in clinical research and practice, so strengthens the demand for transparency and interpretability of the used methods, granting users and the subjects of the data the "right to explanation" of the algorithm's result [8]. Thus, there is a need for a free-text analysis method that is simultaneously low-effort, semantically rich, independent of existing resources, and interpretable.

Quality Phrase Mining (QPM) has been studied previously both by the NLP community, to recognize technical terms [9–11] and by the information-retrieval (IR) community with the goal of identifying the most salient concepts to index. Typically, QPM is divided to two main stages: *first*, due to their sparsity and immense number, the full set of N-grams is filtered to generate a much smaller set of candidate terms. Lexical filters typically use part-of-speech (PoS) tagging and filter in noun phrases while other works employing supervised noun phrase chunking techniques and dependency parsing while others leverage annotated documents to automatically learn these filters [12–15]. In the *second* stage, the candidates are ranked by their *termhood*, defined as "degree that a linguistic unit is related to domain-specific concepts" and *unithood* defined as the "degree of strength or stability of syntagmatic combinations and collocations" [16]. In other words, "termhood" reflects the degree that the word sequence represents a recognized concept (e.g. a disease or a symptom) rather than an ad-hoc description while "unithood" reflects the stability of the sequence of words across its words' occurrences in the corpus. Accordingly, unithood is defined only for multi-word sequences, and is based on occurrence statistics such as mutual-information while termhood applies to both single- and multi-word sequences and stems typically from IR measures such as informativeness. These statistics are sometimes enhanced by comparison to another corpus from a different domain ("reference corpus") or manual annotations to learn the statistics' parameters [11].

Among different QPM methods, C-Value/NC-Value and TextRank have been evaluated in the clinical domain [9]. Liu et al. compared them along with PrefixSpan, another sequential pattern mining algorithm, as the components of a genetic algorithm [17,18]. When assessed individually, C-Value and TextRank outperformed PrefixSpan substantially (F-measure 70.24 % and 69.12 % vs. 9.82 %). When applied to discharge summaries to generate a semantic lexicon, C-Value combined with a machine-learning (conditional-random fields [CRF] named-entity recognition [NER] classifier) filter and custom linguistic normalization rules achieved a precision of 83 % [19]. The generated lexicon was used in a downstream task of concept extraction from the i2b2 dataset, significantly improving the accuracy (F-measure 82.52 % vs 82.04 %) compared to the UMLS-based lexicon, while using 97 % fewer terms and being 100 times faster. However, the dependence on linguistic processing and domain-dependent rules hinders the application of QPM to custom domains less amenable to POS-tagging or rule-based filtering.

As an alternative to lexical rules, methods such as SegPhrase depend on manual annotations to learn the filtering [20]. However, this dependence poses a challenge due to the sample size and expertise needed for manual annotation. To overcome this challenge, AutoPhrase, an extension of SegPhrase, combines unithood (e.g. point-wise mutual information), termhood (e.g. inverse-document frequency) and PoS-tags to partition each sentence to segments with high probability of containing a quality phrase [21]. Leveraging structured knowledge-bases (e.g. Wikipedia), it uses dynamic programming to learn the optimal segmentation (corresponding to the lexical filter in other methods) and phrase mining parameters.

In the current study, we used unsupervised learning to discover narrative factors associated with RR events. Our previous work on RR risk-factor identification used a different unsupervised method, latent-Dirichlet allocation (LDA) topic modelling. The limitations of topic modelling, including challenges in interpretability and the limitation to single words, motivated us to look for methods that capture informative multi-word entities. The lexical filter used by C-Value was found unsuitable for two reasons: First, many of the phrases sought by the subject matter experts (SMEs), e.g. “denies pain”, are not noun phrases. Second, in preliminary testing, the reported PoS-based filter (using the GENIA tagger) missed 55 % of the N-grams deemed quality phrase by two human annotators [22]. While AutoPhrase offers an annotation-free alternative to C-Value’s rule-based filtering, it is highly dependent on PoS tagging, rendering it susceptible to the same issues. Therefore, we sought a data-driven method to mine phrases.

Following Liu et al., who combined C-Value with TextRank in a genetic algorithm, we combined these methods to complement each other: [17] C-Value/NC-Value incorporates global information about the phrase (occurrence patterns across the whole corpus), while TextRank captures local information (importance relative to other sequences in the same document). The proposed method uses TextRank to identify locally important word-sequences (replacing the lexical filter or PoS-segmentation used by C-Value and AutoPhrase, respectively) and then use C-Value/NC-Value to rank them globally against other locally-important sequences.

In summary, our goal was to develop a data-driven, low-manual effort feature engineering method for clinical knowledge discovery from nursing notes. We hypothesize that phrase mining will yield clinically meaningful features with little effort from SME.

2. Material and methods

Our methods consist of 1) data collection; 2) phrase mining; and 3) evaluation, as described in Fig. 1. The study was approved by the institutional review board.

2.1. Data collection

The study population included all inpatients from Partners' Healthcare, a healthcare delivery network in Boston, MA, hospitalized between 2015 and 2018.

Inclusion criterion: Inpatients admitted to any general medical or surgical acute care or critical care unit for > 24 h.

Exclusion criteria: Patients less than 18 years of age; hospice or palliative care patients and those without a hospital encounter; special units, such as obstetrics and oncology services.

Nursing notes included the following types: Progress Notes, Consults, Procedures, Discharge Summaries, Assessment & Plan Note, Nursing Note, Code Documentation, Significant Event, Transfer/Sign Off Note, Nursing Summary and Family Meeting documented by Registered Nurse. While Rapid Response Documentation notes are in the scope of the full CONCERN study, they were excluded from this analysis to prevent leakage of outcome information to the phrase mining process. RR events were collected from flowsheets. Data was collected only from the time intervals in included units and censored at the earliest occurrence of discharge, RR event or 1,282 h since admission, the 99th percentile of time from admission to RR event among the CONCERN study population.

2.2. Phrase mining

We made two main changes to the original C-Value/NC-Value methods: [9] a) instead of a lexical filter, each document was segmented using TextRank, filtering in only the selected segments, and b) while NC-Value uses only nouns, adjectives and verbs when considering invidvidual words, we used all words except stop-words. Our phrase mining method consists of three major components (preprocessing, segmentation and term ranking) and nine steps. The process is outlined in Fig. 2.

1) Note Preparation—The notes were tokenized and sentence-segmented. Dates and numbers were collapsed to placeholder tokens. Counts of all N-grams were collected.

1) N-gram Enumeration—For each sentence, all possible N-grams up to length 4 were generated excluding those a) appearing < 5 times in the whole corpus, b) beginning/ending with a stop-token, c) having a non-word token (defined as a token not matching the regular expression *[a-zA-Z][-a-zA-Z_0–9^]*+, or d) containing the conjunction "and".

1) N-gram Representation—N-grams' meaning was represented by FastText embeddings trained on the qualifying notes from the full study cohort, using the default hyperparameters except dimensionality 300, window size 2, minimal count 5 and word N-grams 3 [23]. Phrase embeddings were generated by averaging the embeddings of the phrase's words. While compositional phrase embeddings is an active area of research and more sophisticated composition methods exists, algebraic operations offer reasonable accuracy at a lower computational cost [24].

1) TextRank—TextRank is an adaptation of Google's PageRank algorithm to textual units [25,26]. It determines the importance of an item based on the other items pointing (voting) to it and their importance. The score of an item is the sum of the weight of each vote multiplied by the voting item's score, calculated recursively in the same manner (Equation 1). The weight $w_{j,i}$ between the items is at the core of the TextRank algorithm and the function determining the weights guides the results and differs by use-case. In this work, each document was segmented by TextRank separately, using the document's N-grams as items and the cosine similarity of their embeddings as the vote weight.

1) TextRank Graph Construction—The adjacency matrix describing the graph contains the cartesian product of all the distinct N-gram, so cell *i,j* contains the cosine similarity between the embeddings of N-grams *i* and *j*, yielding an undirected weighted graph. The columns of the matrix were transformed to a probability distribution (range of 0–1 and sum of 1).

1) TextRank Score Calculation—The score for each N-gram was initialized randomly and updated iteratively by multiplying the scores vector by the adjacency matrix using the hyper parameters reported in the original article (damping factor of 0.85 and stopping criteria of a total change in items score of 1E-6 or 200 iterations).

1) Segmentation—After convergence, each sentence was segmented using a greedy algorithm selecting the top-scored N-gram not overlapping with any of the previously selected ones until either exhaustion or absence of any non-overlapping N-gram.

1) Count rectification—The count of each distinct N-gram among the selected segments was collected, yielding a rectified count differing from the raw (all occurrences) counts.

1) C-Value/NC-Value—C-value ranks the unithood according to Equation 2, rewarding longer and frequent phrases and penalizing those who are nested inside other frequent phrases (to capture the longest possible phrase). NC-Value re-ranks the C-value's top-terms using contextual information, rewarding candidates that are accompanied by words that frequently accompany high-scored terms, according to Equation 3. The rectified counts were used as the input for the C-Value/NC-Value algorithm. The NC-value stage requires selection of the number of candidates from the first step to use. Since the rectification diminishes the number and counts of N-grams, we used a lower count threshold of 2 and a higher top-terms proportion of 10 %.

2.3. Evaluation

We first evaluated the accuracy of identified phrases, and then assessed the value of identified phrases for risk predication.

2.3.1. Accuracy of phrase identification—N-grams were adjudicated by two clinicians as being a quality phrase or not, following guidelines listed in Box Box 1. Initially the experts annotated a random sample of 100 randomly selected N-grams. However, low inter-rater reliability (IRR) was observed (Cohen's kappa score = 0.39) despite repeated annotator training sessions, signaling the high subjectivity of the definition of quality phrases. When limiting the comparison to N-grams found verbatim in SNOMED CT, the IRR improved to 0.625, suggesting that subjectivity decreases for high-quality terms. Therefore, we leveraged the pooling approach used in TREC IR competitions: instead of judging a random sample of documents, the human annotation effort is focused on a set of cases deemed relevant by other ranking systems [27]. We used NC-Value (without segmentation) and TextRank segmentation rectified counts alone. The top 150 scored phrases from each method were merged yielding 240 phrases. The phrases were shuffled and annotated by two clinicians, improving IRR to 0.72. Overall, 99 phrases were deemed quality phrases. The evaluation metric was *average precision* (AP), defined as:

$$\mathrm{AP} = \sum_{n=1}^{|N|} \left(Recall_n - Recall_{n-1} \right) \times Precision_n$$

Where N is the number of adjudicated N-grams, $Precision_n$ is the proportion of ranked phrases up to n that are true quality phrase and $Recall_n$ is the proportion of all true quality phrases that are included in the ranked list up to n. Thus, AP equals the mean of precisions achieved at each threshold (ranked item) weighted by the increase in recall from the previous threshold. A perfect ranking method, i.e. one that places all positively-adjudicated phrases at the top of the list and vice versa will achieve an AP of 1.

2.3.2. Predictive value of identified phrases—To judge the phrases' usefulness as features for predictive and explanatory modelling, they were used as time-dependent predictors in an extended Cox model of RR event. The process to transform each document to a numerical feature vector is depicted in Fig. 3. The top 500 min. d phrases were selected. For each document, all qualifying N-grams (i.e. those satisfying the filter described in "Graph construction" above) were collected, and their similarity to each of the selected phrases was calculated. The similarities were averaged for each selected phrase, yielding a fixed-length vector of 500 features for each document. This method was preferred over N-gram counts since the high linguistic variability of nursing notes could result in many documents containing none of the selected 500 phrases, leading to an empty feature vector. In contrast, distributed representations allow meaningful estimation of a phrase's weight in a note even if it is absent verbatim. The documents' features, along with its calendar hour and the patient's age and sex were used as the time-dependent covariates. The model's goodness-of-fit was evaluated by its concordance-index.

3. Results

Overall, 61,740 encounters of 45,817 patients (48.9 % male, average age 61.6, standard deviation [SD] 17.4 years) with 1,067 events (1.7 % of all encounters) were found. RR events occurred at a median of 82 h after the admission. The time the patients were included in the study averaged 131 (SD = 162) hours with a median of 75 and interquartile range of 48–86 hours. The types of the 778,955 qualifying notes containing 10,699,976 sentences and 125,809,359 tokens is in Table 1.

Out of the 45,513,425 distinct N-grams (up to length of 4) found in the notes, 2,171,428 N-grams qualified. The segmentation yielded an average of 7.78 (SD = 9.07) and a median of 6 segments per sentence. Segmentation examples are shown in Table 2. The top-20 ranked phrases are listed in Box 2.

3.1. Accuracy of phrase identification

Our method achieved an AP of 0.590. In manual inspection, many of the mislabeled phrases contained number placeholders. Nursing notes frequently quote the values of vital signs. Since often these are continuous variables, they can amount to huge number of distinct tokens, increasing sparsity. Therefore, during pre-processing numbers are collapsed to placeholder tokens, reducing N-gram sparsity and increasing the termhood of N-grams such as “hr NUMBER” and “bp NUMBER”. On phrases without number placeholders, AP increased to 0.764. The precision-recall curves with and without number-placeholders can be found in Fig. 4.

3.2. Predictive value of identified phrases

Since variance-inflation factor analysis revealed substantial collinearity between the top-500 phrases, principal-component analysis was performed. Using a cumulative explained variance cutoff of 99 %, the first 96 components, along with age, sex and calendar hour as the covariates of another extended Cox model, achieving a concordance index of 0.739. Tests of proportionality based on correlation of Schoenfeld residuals with time were found not significant, supporting the proportionality assumption. To explore the entities associated with RR event, we clustered the phrase variables using Spearman’s correlation measure to 50 clusters, based on our previous experience about the entities associated with clinical nursing outcomes. In manual inspection of the clusters, 30 corresponded to a clinical concept and they were used to fit another extended Cox model. Table 3 shows the phrase clusters associated with RR events and their hazard ratio (HR).

4. Discussion

To the best of our knowledge, this is the first work evaluating automatic term recognition methods on nursing notes and on clinical outcome prediction. In the current study, we used QPM to transform obscure unsupervised features (distributed representations) to interpretable ones (phrases) while minimizing the required manual effort. As nurses’ concern is an important indicator that a patient’s condition is likely deteriorating, information from nursing notes can be leveraged to capture nurses’ general concern about the patient [5]. After the mining stage, the phrases can be rapidly applied to newly written notes as features to allow prediction of RR event risk in real time. While the mined phrases did not achieve an absolutely high concordance-index (0.739), they can offer a reasonable baseline requiring no SME effort and enable exploratory or ad-hoc analysis of narrative clinical data. With a relatively small manual effort (review of the 50 phrase clusters was the only manual step in the process) the mined phrases could be further interpreted and grouped to higher level concepts. The fitted Cox model revealed multiple clinically plausible risk/protective factors such as “patient decision” or “emotional support” (hazard-increasing) and “bowel movement” (hazard-decreasing). Less clear findings (e.g. the hazard-decreasing effect of “pain” cluster) call for further investigation. Such results could also stem from deficiencies in the application of the mined phrases to represent each document, such as negation and hedging information.

In the CONCERN study, we needed to develop custom terminologies due to the lack of suitable ones targeting nursing and rapid response domains. The mined phrases can be used to facilitate terminology development since adjudication of automatically generated results, in our experience, is easier and faster than curation of terms de-novo by SMEs.

While focusing on the nursing domain and RR event outcome, our method is not inherently limited to them. It has no dependence on domain-specific terminologies and all of its components are unsupervised requiring only raw text, typically abundantly available from the EHR.

In addition to the development and evaluation of the phrase mining method, the current study demonstrated the validity and applicability of the phrase-adjudication guidelines. These findings can support future phrase-mining projects in the clinical domain. Comments on the adjudicated phrases revealed the subjectivity and indefinite nature of quality phrase definition, particularly regarding completeness and consistency.

In the course of this work, other methods were explored. Modelling the phrase mining as a NER task, we explored the feasibility of distant supervision (using SNOMED CT) to train a bidirectional long-short term memory-CRF (BiLSTM-CRF) sequence labeler. However, on the concept detection i2b2 dataset it achieved a precision of 47 % and recall of 29 %, suggesting the limits of distant supervision alone to capture the semantics of clinical entities and the gaps between terminologies, even interface ones such as SNOMED CT, and narrative clinical documentation [28].

Our study is affected by several limitations. The low IRR precluded the manual annotation of randomly selected N-grams, precluding the estimation of the true recall and precision. The increase in IRR on the SNOMED-matching and pooled N-grams suggests that clarification of the criteria for definite non-quality phrases could improve IRR. As evident from the improvement in identification accuracy by exclusion of certain token classes (number placeholders), QPM is sensitive to the pre-processing logic. A possible solution could be to tune the QPM to prefer high recall and use clustering, active-learning or feature selection (e.g. LASSO regression) to prune the excessive list of phrases. Due to lack of working reference implementation, no direct comparison was made to other established methods hindering conclusion about the optimal phrase-mining method. Our data also came from a single institute, affecting the generalizability of our findings.

Future steps for our work include validation of the identified risk factors in other settings and incorporation of structured information such as flowsheet and medication data to enhance the accuracy of the Cox model. Recently developed context-aware representation methods such as BERT may enhance QPM by incorporating deeper context information into the segmentation process to address contextual issues such as negation, timing and experiencer [29]

5. Conclusion

Automatic term recognition can generate useful and interpretable textual features for a clinical outcome prediction from nursing narrative notes with minimal manual effort.

Acknowledgement

This study was funded by the National Institute of Nursing Research (NINR): 1R01NR016941-01, Communicating Narrative Concerns Entered by RNs (CONCERN): Clinical Decision Support Communication for Risky Patient States. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Abbreviations:

ICU	intensive care unit
RR	rapid response
EHR	electronic health record
NLP	natural language processing
RR	Rapid response
BoW	bag-of-words
BoNG	bag-of-*N*-grams
ML	machine-learning
QPM	Quality Phrase Mining
IR	information-retrieval
PoS	part-of-speech
CRF	conditional-random fields
NER	named-entity recognition
SMEs	subject matter experts
IRR	inter-rater reliability

AP	average precision
SD	standard deviation
HR	hazard ratio

References

[1]. Lyons PG, Edelson DP, Churpek MM, Rapid response systems, Resuscitation 128 (2018) 191–197, 10.1016/j.resuscitation.2018.05.013. [PubMed: 29777740]

[2]. Winters BD, Weaver SJ, Pfoh ER, Yang T, Pham JC, Dy SM, Rapid-response systems as a patient safety strategy: a systematic review, Ann. Intern. Med 158 (2013) 417–425, 10.7326/0003-4819-158-5-201303051-00009. [PubMed: 23460099]

[3]. Solomon RS, Corwin GS, Barclay DC, Quddusi SF, Dannenberg MD, Effectiveness of rapid response teams on rates of in-hospital cardiopulmonary arrest and mortality: a systematic review and meta-analysis, J. Hosp. Med 11 (2016) 438–445, 10.1002/jhm.2554. [PubMed: 26828644]

[4]. Bellomo R, Goldsmith D, Uchino S, Buckmaster J, Hart G, Opdam H, Silvester W, Doolan L, Gutteridge G, Prospective controlled trial of effect of medical emergency team on postoperative morbidity and mortality rates, Crit. Care Med 32 (2004) 916–921. [PubMed: 15071378]

[5]. Collins SA, Cato K, Albers D, Scott K, Stetson PD, Bakken S, Vawdrey DK, Relationship between nursing documentation and patients' mortality, Am. J. Crit. Care 22 (2013) 306–313, 10.4037/ajcc2013426. [PubMed: 23817819]

[6]. Parr MJ, Hadfield JH, Flabouris A, Bishop G, Hillman K, The Medical Emergency Team: 12 month analysis of reasons for activation, immediate outcome and not-for-resuscitation orders, Resuscitation. 50 (2001) 39–44. [PubMed: 11719127]

[7]. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J, Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ (Eds.), Distributed Representations of Words and Phrases and Their Compositionality, Curran Associates, Inc., 2013, pp. 3111–3119 (Accessed September 1, 2017), http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf.

[8]. Goodman B, Flaxman S, European Union regulations on algorithmic decision-making and a "right to explanation", AI Mag. 38 (2017) 50, 10.1609/aimag.v38i3.2741.

[9]. Frantzi K, Ananiadou S, Mima H, Automatic recognition of multi-word terms:. the C-value/NC-value method, Int J Digit Libr. 3 (2000) 115–130, 10.1007/s007999900023.

[10]. Park Y, Byrd RJ, Boguraev BK, Automatic glossary extraction: beyond terminology identification, Proceedings of the 19th International Conference on Computational Linguistics - Volume 1, Association for Computational Linguistics, Stroudsburg, PA, USA, 2002, pp. 1–7, 10.3115/1072228.1072370.

[11]. Zhang Z, Iria J, Brewster C, Ciravegna F, A Comparative Evaluation of Term Recognition Algorithms, LREC, 2008.

[12]. Chen K, Chen H-H, Extracting noun phrases from large-scale texts: a hybrid approach and its automatic evaluation, Proceedings of the 32Nd Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, Stroudsburg, PA, USA, 1994, pp. 234–241, 10.3115/981732.981764.

[13]. Punyakanok V, Roth D, The use of classifiers in sequential inference, in: Leen TK, Dietterich TG, Tresp V (Eds.), Advances in Neural Information Processing Systems 13, MIT Press, 2001, pp. 995–1001 http://papers.nips.cc/paper/1817-the-use-of-classifiers-in-sequential-inference.pdf.

[14]. Xun E, Huang C, Zhou M, A unified statistical model for the identification of English baseNP, Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, Stroudsburg, PA, USA, 2000, pp. 109–116, 10.3115/1075218.1075233.

[15]. Koo TK, Carreras X, Collins M, Simple Semi-supervised Dependency Parsing, ACL, 2008.

[16]. Kageura K, Umino B, Methods of Automatic Term Recognition - a Review, (1996).

[17]. Liu W, Chung BC, Wang R, Ng J, Morlet N, A genetic algorithm enabled ensemble for unsupervised medical term extraction from clinical letters, Health Inf. Sci. Syst 3 (2015) 5, 10.1186/s13755-015-0013-y. [PubMed: 26664724]

[18]. Mortazavi-Asl H Pinto UD, Mining sequential patterns by pattern-growth: the PrefixSpan approach, IEEE Trans. Knowl. Data Eng 16 (2004) 1424–1440, 10.1109/TKDE.2004.77.

[19]. Jiang M, Denny JC, Tang B, Cao H, Xu H, Extracting semantic lexicons from discharge summaries using machine learning and the C-Value method, AMIA Annu. Symp. Proc 2012 (2012) 409–416. [PubMed: 23304311]

[20]. Liu J, Shang J, Wang C, Ren X, Han J, Mining quality phrases from massive text corpora, Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, ACM, New York, NY, USA (2015) 1729–1744, 10.1145/2723372.2751523.

[21]. Liu J, Shang J, Han J, Phrase mining from massive text and its applications, Synth. Lect. Data Min. Knowl. Discov 9 (2017) 1–89, 10.2200/S00759ED1V01Y201702DMK013.

[22]. Tsuruoka Y, Tateishi Y, Kim J-D, Ohta T, McNaught J, Ananiadou S, Tsujii J, Developing a robust part-of-speech tagger for biomedical text, in: Bozanis P, Houstis EN (Eds.), Advances in Informatics, Springer, Berlin Heidelberg, 2005, pp. 382–392.

[23]. Bojanowski P, Grave E, Joulin A, Mikolov T, Enriching Word Vectors With Subword Information, CoRR. abs/1607.04606, (2016) http://arxiv.org/abs/1607.04606.

[24]. Scheepers T, Kanoulas E, Gavves E, Improving word embedding compositionality using lexicographic definitions, Proceedings of the 2018 World Wide Web Conference, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland (2018) 1083–1093, 10.1145/3178876.3186007.

[25]. Page L, Brin S, Motwani R, Winograd T, The PageRank Citation Ranking: Bringing Order to the Web, (1999) (Accessed February 25, 2019), http://ilpubs.stanford.edu:8090/422/.

[26]. Mihalcea R, Tarau P, TextRank: Bringing Order Into Text, (2004).

[27]. Voorhees EM, Harman D, Overview of the Eighth Text REtrieval Conference (TREC-8), Trec, 1999, p. 103.

[28]. Uzuner Ö, South BR, Shen S, DuVall SL, 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text, J. Am. Med. Inform. Assoc 18 (2011) 552–556, 10.1136/amiajnl-2011-000203. [PubMed: 21685143]

[29]. Devlin J, Chang M-W, Lee K, Toutanova K, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, ArXiv:1810.04805 [Cs], (2018) (Accessed December 19, 2018), http://arxiv.org/abs/1810.04805.

Box 1

N-gram adjudication guidelines

Term is defined as “a *consistent*, *complete* and *pure* semantic unit that *represents* a clinical concept”:

1. “Consistent”: the word sequence is repeatedly used by the domain practitioners to represent the clinical concept. E.g. “head feels bigger” describes the typical symptom of tension-type headache. However, it is not consistent because it is an ad-hoc description (and many other wordings can be used to describe the same thing). In contrast, “tension-headache” is consistent because it is the conventional/typical way to communicate that clinical concept between clinicians.

2. “Complete”: the word sequence includes all the parts of the concept. E.g. “community acquired” is not a complete semantic unit because the full concept is “community-acquired pneumonia”. E.g. “rapid rate” is incomplete because it can be assigned to multiple distinct clinical entities (e.g. “rapid heart rate” and “infusion rate”), and the assignment requires additional words to be present. However, if a word sequence is a complete meaning but of multiple concepts, it can be considered a term. E.g. “PE” is a term because while it has multiple meanings (“pleural effusion” vs “pulmonary embolism”) for each of them, PE is complete.

3. “Pure”: the word sequence does not include any word that is not part of the clinical concept. This requirement applies also to modifiers (laterality, timing etc.) E.g. “left hip fracture” is not pure because the word “left” is not part of the clinical concept “hip fracture” (hip fracture is a distinct clinical concept with a specific diagnostic algorithm, treatment etc. while “left” is a fact that describes an individual instance of hip fracture, and does not have any specific clinical implications or knowledge). On the other hand, “left heart failure” is a term because it has different clinical implications from “right heart failure”.

4. “Clinical concept”: some words might also have a general-world meaning. In this exercise, we look for clinical concepts only.

When judging a word sequence, please do not consider its association with specific clinical outcomes (RR, mortality etc.) The association with clinical outcome will be investigated in subsequent steps (predictive modelling). In this step, the purpose is to find terms that nurses use to represent clinical entities in their notes and the word sequence should be judged based on that criterion solely.

Box 2

Top-20 phrases by descending order

Continue to monitor, cooperative with care, good effect, md aware, denies pain, team aware, case management progress, case management, available for consultation, steady gait, cont to monitor, tolerating regular diet, continue to follow, risk screening completed, high risk criteria, bed alarm, regular diet, bm this shift, per report, abd soft.

Summary table

What was already known on the topic

a

a. Rapid-response activation criteria depend on subjective and unstructured criteria.

b. Data-mining methods can discover factors from free-text but do not suit clinical notes.

What this study added to our knowledge

a

a. A validated data-mining method to discover clinical phrases from nursing notes with no manual annotation.

b. The discovered phrases are significantly associated with rapid-response event hazard.

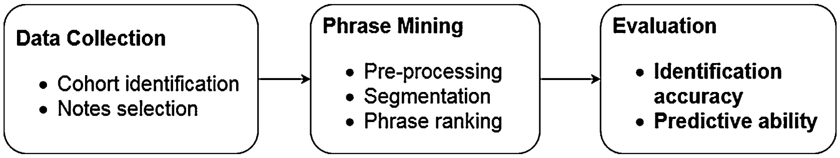
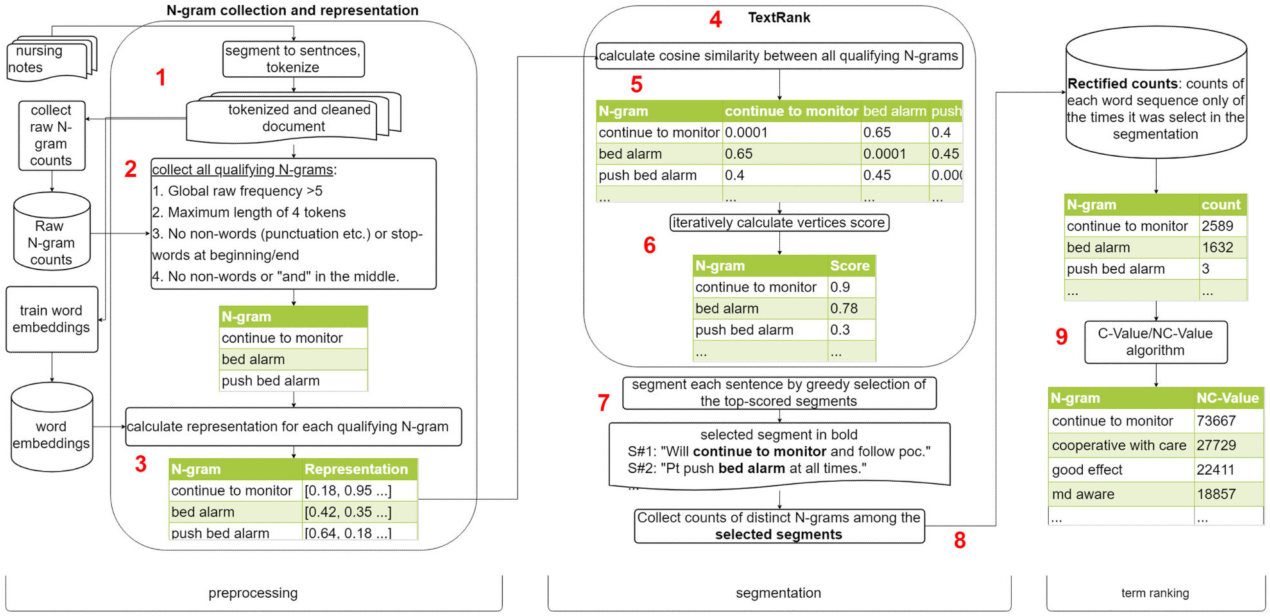


Fig. 1.

Outline of the study architecture.


N-gram collection and representation
nursing notes
segment to sentnces, tokenize
1
tokenized and cleaned document
collect raw N-gram counts
Raw N-gram counts
2
collect all qualifying N-grams:
1. Global raw frequency >5
2. Maximum length of 4 tokens
3. No non-words (punctuation etc.) or stop-words at beginning/end
4. No non-words or "and" in the middle.
train word embeddings
word embeddings
calculate representation for each qualifying N-gram
3
4
TextRank
calculate cosine similarity between all qualifying N-grams
5
iteratively calculate vertices score
6
7
segment each sentence by greedy selection of the top-scored segments
Collect counts of distinct N-grams among the selected segments
8
Rectified counts: counts of each word sequence only of the times it was select in the segmentation
9
C-Value/NC-Value algorithm
preprocessing
segmentation
term ranking

N-gram
continue to monitor
bed alarm
push bed alarm

N-gram	Representation
continue to monitor	[0.18, 0.95 ...]
bed alarm	[0.42, 0.35 ...]
push bed alarm	[0.64, 0.18 ...]

N-gram	continue to monitor	bed alarm	push
continue to monitor	0.0001	0.65	0.4
bed alarm	0.65	0.0001	0.45
push bed alarm	0.4	0.45	0.00(
...	...	...	...

N-gram	Score
continue to monitor	0.9
bed alarm	0.78
push bed alarm	0.3
...	...

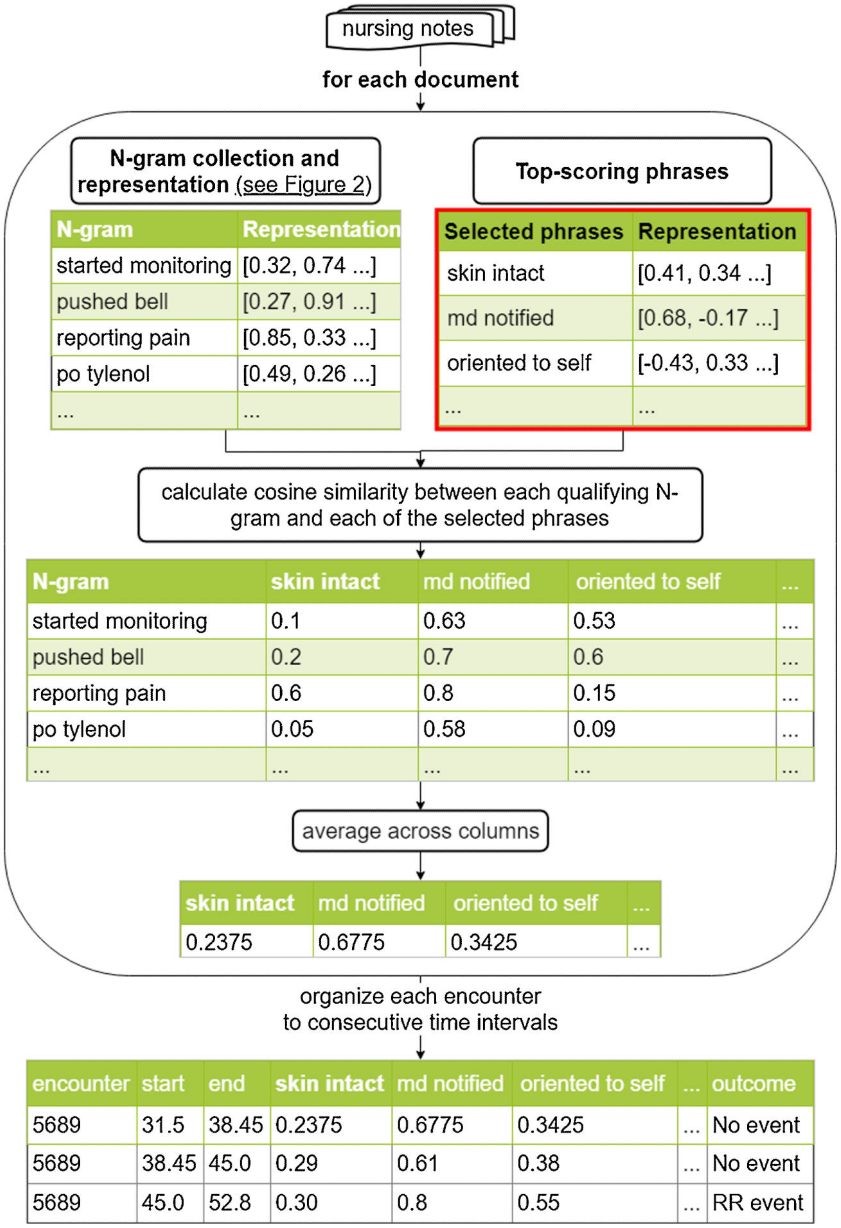
selected segment in bold
S#1: "Will **continue to monitor** and follow poc."
S#2: "Pt push **bed alarm** at all times."
...

N-gram	count
continue to monitor	2589
bed alarm	1632
push bed alarm	3
...	...

N-gram	NC-Value
continue to monitor	73667
cooperative with care	27729
good effect	22411
md aware	18857
...	...

Fig. 2.

Phrase mining process.


nursing notes
for each document
N-gram collection and representation (see Figure 2)
Top-scoring phrases
calculate cosine similarity between each qualifying N-gram and each of the selected phrases
average across columns
organize each encounter to consecutive time intervals

N-gram	Representation
started monitoring	[0.32, 0.74 ...]
pushed bell	[0.27, 0.91 ...]
reporting pain	[0.85, 0.33 ...]
po tylenol	[0.49, 0.26 ...]
...	...

Selected phrases	Representation
skin intact	[0.41, 0.34 ...]
md notified	[0.68, -0.17 ...]
oriented to self	[-0.43, 0.33 ...]
...	...

N-gram	skin intact	md notified	oriented to self	...
started monitoring	0.1	0.63	0.53	...
pushed bell	0.2	0.7	0.6	...
reporting pain	0.6	0.8	0.15	...
po tylenol	0.05	0.58	0.09	...
...	...	...	...	...

skin intact	md notified	oriented to self	...
0.2375	0.6775	0.3425	...

encounter	start	end	skin intact	md notified	oriented to self	...	outcome
5689	31.5	38.45	0.2375	0.6775	0.3425	...	No event
5689	38.45	45.0	0.29	0.61	0.38	...	No event
5689	45.0	52.8	0.30	0.8	0.55	...	RR event

Fig. 3.

Feature extraction process for the selected phrases.

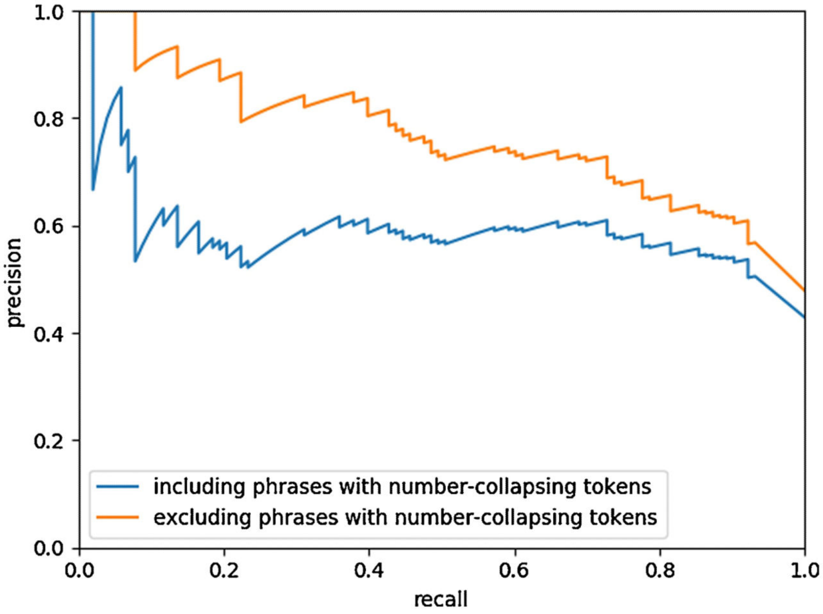


Fig. 4.

Precision-recall curve for the modified NC-Value method on rectified counts.

Table 1

Note type distribution.

Note type	Count
Progress Notes	729249
Nursing Summary	36945
Nursing Note	5420
Procedures	3110
Significant Event	1746
Transfer / Sign Off Note	1679
Code Documentation	272
Family Meeting	39
Total	**778955**

Table 2

Examples of sentence segmentations.

Original sentence	Segments
TTE, worsening MR, BIV failure, and RV dysfunction.	BIV failure, TTE, worsening MR, RV dysfunction
+ edema to BLE, L > R, MD aware.	+ edema to BLE, MD aware
She has a (+) CXR for pna, and a dirty urine.	CXR for pna, dirty urine

Table 3

Phrase clusters and their association with the risk of rapid-response event. A positive coefficient increases the hazard for RR event and vice versa.

Effect	Cluster name	Hazard ratio (95 % confidence interval)	p-value	adjusted p-value	Example phrases
Increases hazard	patient decision **	1.68 (1.28–2.20)	< .001	< .001	pt aware, pt refused
	auxiliary tests **	1.63 (1.38–1.93)	< .001	< .001	am labs, ekg obtained
	cough **	1.47 (1.32–1.65)	< .001	< .001	non productive cough, productive cough
	emotional support **	1.44 (1.27–1.63)	< .001	< .001	emotional support given
	abdominal examination *	1.39 (1.04–1.85)	0.025	0.8	abd softly distended, non tender
	nurse’s shift transfer *	1.38 (1.06–1.78)	0.015	0.49	shift assessment, shift summary
	patient verbal communication	1.24 (0.93–1.64)	0.14	> .99	patient reports, patient states
	MD awareness *	1.21 (1.02–1.43)	0.026	0.83	md made aware, md notified, md paged
	care management	1.21 (0.62–2.33)	0.58	> .99	review homecare, home care services
	fasting *	1.19 (1.01–1.39)	0.038	> .99	npo since midnight, remains npo, ivf infusion
	patient alarm *	1.15 (1.04–1.27)	< .001	0.27	bed alarm, call bell
	Intravenous medications	1.09 (0.91–1.29)	0.35	> .99	iv antibiotics, iv dilaudid
	aspiration	1.07 (0.93–1.23)	0.37	> .99	aspiration precautions maintained, infection afebrile
	Pre/post operation	1.05 (0.97–1.15)	0.22	> .99	post op, pre op
	dressing	1.03 (0.88–1.21)	0.72	> .99	primary dressing, wound vac
	assessment	1.03 (0.77–1.37)	0.86	> .99	team made aware, continue to monitor
	sex (female)	1.02 (0.91–1.16)	0.7	> .99	N/A
	age **	1.02 (1.02 to 1.02)	< .001	< .001	N/A
	skin barrier	1.01 (0.86–1.20)	0.87	> .99	barrier cream, open to air
No effect	Note’s calendar hour	1 (0.99–1.01)	0.65	> .99	N/A
Decreases hazard	Intravenous catheter	0.95 (0.86–1.04)	0.28	> .99	blood return, picc line
	following commands	0.91 (0.82–1.01)	0.071	> .99	follows commands
	diet tolerance	0.86 (0.72–1.03)	0.1	> .99	tolerating house diet, takes pills whole
	completed tests *	0.86 (0.76 to 0.97)	0.014	0.43	cxr completed, echo completed
	vital signs *	0.85 (0.74 to 0.98)	0.024	0.76	hemodynamically stable, neuro exam stable
	risk criteria *	0.79 (0.65 to 0.97)	0.023	0.73	high risk criteria, initial assessment

Effect	Cluster name	Hazard ratio (95 % confidence interval)	p-value	adjusted p-value	Example phrases
	wound care	0.76 (0.56–1.03)	0.08	> .99	ace wrap, dressing cdi
	Pain [*]	0.75 (0.63 to 0.91)	< .001	0.091	abd pain, atc Tylenol, oxycodone prn
	Ambulation [*]	0.75 (0.59 to 0.95)	0.016	0.52	able to ambulate, oob to chair, stand by assist
	communication	0.73 (0.53–1.01)	0.054	> .99	family at bedside, responsible person
	plan of care	0.45 (0.18–1.18)	0.1	> .99	reassessment chart, pt consulted
	bowel movement [**]	0.29 (0.20 to 0.41)	< .001	< .001	bm overnight, bm this shift

N/A not applicable.

[*] Significant at 0.05 level.

[**] Significant after Bonferroni correction of the significance level to 0.05/34 = 1.56E-03.