*Journal of the American Medical Informatics Association*, 28(3), 2021, 541–548
doi: 10.1093/jamia/ocaa263
Advance Access Publication Date: 17 November 2020
Research and Applications

Research and Applications

# Natural language processing to measure the frequency and mode of communication between healthcare professionals and family members of critically ill patients

**Filipe R. Lucini ,[1,2] Karla D. Krewulak,[1] Kirsten M. Fiest,[1,3,4] Sean M. Bagshaw,[5,6] Danny J. Zuege,[1,6] Joon Lee,[2,3,7] and Henry T. Stelfox[1,3]**

[1]Department of Critical Care Medicine, Cumming School of Medicine, University of Calgary and Alberta Health Services, Calgary, Alberta, Canada, [2]Data Intelligence for Health Lab, Cumming School of Medicine, University of Calgary, Calgary, Alberta, Canada, [3]Department of Community Health Sciences and O'Brien Institute for Public Health, Cumming School of Medicine, University of Calgary, Calgary, Alberta, Canada, [4]Department of Psychiatry & Hotchkiss Brain Institute, Cumming School of Medicine, University of Calgary, Calgary, Alberta, Canada, [5]Department of Critical Care Medicine, Faculty of Medicine and Dentistry, University of Alberta, and Alberta Health Services, Edmonton, Alberta, Canada, [6]Critical Care Strategic Clinical Network, Alberta Health Services, Alberta, Canada and [7]Department of Cardiac Sciences, Cumming School of Medicine, University of Calgary, Calgary, Alberta, Canada

Corresponding Author: Filipe R. Lucini, PhD, Centre for Health Informatics (CHI), 5E-04 TRW building, 3280 Hospital Drive NW, Calgary, Alberta T2N 4Z6, Canada (filipe.lucini@ucalgary.ca)

Received 30 June 2020; Editorial Decision 30 September 2020; Accepted 9 October 2020

## ABSTRACT

**Objective:** To apply natural language processing (NLP) techniques to identify individual events and modes of communication between healthcare professionals and families of critically ill patients from electronic medical records (EMR).

**Materials and Methods:** Retrospective cohort study of 280 randomly selected adult patients admitted to 1 of 15 intensive care units (ICU) in Alberta, Canada from June 19, 2012 to June 11, 2018. Individual events and modes of communication were independently abstracted using NLP and manual chart review (reference standard). Preprocessing techniques and 2 NLP approaches (rule-based and machine learning) were evaluated using sensitivity, specificity, and area under the receiver operating characteristic curves (AUROC).

**Results:** Over 2700 combinations of NLP methods and hyperparameters were evaluated for each mode of communication using a holdout subset. The rule-based approach had the highest AUROC in 65 datasets compared to the machine learning approach in 21 datasets. Both approaches had similar performance in 17 datasets. The rule-based AUROC for the grouped categories of patient documented to have family or friends (0.972, 95% CI 0.934–1.000), visit by family/friend (0.882 95% CI 0.820–0.943) and phone call with family/friend (0.975, 95% CI: 0.952–0.998) were high.

**Discussion:** We report an automated method to quantify communication between healthcare professionals and family members of adult patients from free-text EMRs. A rule-based NLP approach had better overall operating characteristics than a machine learning approach.

**Conclusion:** NLP can automatically and accurately measure frequency and mode of documented family visitation and communication from unstructured free-text EMRs, to support patient- and family-centered care initiatives.

**Key words:** natural language processing, electronic medical records, intensive care units, family, communication

## INTRODUCTION

Patient- and family-centered care (PFCC) is an approach to improving healthcare through the inclusion of patients and families as partners throughout the healthcare process.[1] PFCC is especially important in an intensive care unit (ICU) where family members often assume the role of surrogate decision-makers for critically ill patients.[2] The Society of Critical Care Medicine recommends regular communication with family members of critically ill patients,[3] and studies have identified that communication occurs in multiple ways including during dedicated family meetings, patient care rounds, informal updates, and telephone calls.[4] Measures of frequency of family communication are important to evaluate the adherence of healthcare settings to these recommendations, as well as establish benchmarks and improve PFCC.

The increasing availability of electronic medical records (EMRs) provides an opportunity to better understand communication with patient families. However, family communication is often recorded as free-text in EMRs. Manual abstraction and classification of this free-text data is time-consuming, prone to human error, and difficult to scale. One solution would be to standardize the structure of information capture in EMRs, but this is likely to be deficient for documenting complex social constructs like family communication. Another strategy would be to automate abstraction and classification of free-text data. Natural language processing (NLP) presents promising computational techniques to generate structured information from free-text EMRs.[5] We sought to apply NLP techniques to automatically identify individual events and modes of family communication from EMRs to test the ability of NLP to generate useful information related to family–care provider interactions in the ICU.

## MATERIALS AND METHODS

### Design, setting, and population

We conducted a retrospective multicenter population-based cohort study. A random sample of adult patients (aged $\geq$ 18 years) admitted to 1 of 15 ICUs in Alberta, Canada (14 medical/surgical and 1 neuroscience providing critical care services to 4.3 million residents) from June 19, 2012 to June 11, 2018 with an ICU stay greater than 24 hours were included in this study. The primary data source was eCritical, an EMR for multidisciplinary clinical documentation and automated capture of device and laboratory data in use in all ICUs across Alberta.[6,7] eCritical is subjected to regular audits and rigorous quality assurance.[8] The study was reviewed and approved by the Conjoint Health Research Ethics Board at the University of Calgary (Reference number: REB17-1842). The need for informed consent was waived.

### Measures of family communication

The measures of family communication investigated in this study were identified by a team comprising ICU clinicians and researchers. Three categories of information were included. The first category (Documented Family or Friends) was the documented existence of patients' family members or friends. It allowed the identification of persons that are closely related to the patient and included 22 subcategories of kinship and gender (eg, child-boy/man). The second category (Visits) referred to documented visits of family members or friends and included 37 subcategories of kinship, gender (girl/woman or boy/man) and visitation (eg, child-boy/man-visit). The third category (Phone Calls) captured documented phone calls from or to family members or friends and included 30 subcategories of kinship, gender and telephone communication (eg, child-boy/man-phone call). Gender (girl/woman or boy/man)[9] of the family member or visitor was included because, in many societies, patient family caregivers are disproportionally women,[10] and this may influence communication. A total of 89 measures of family communication were examined (Supplementary Appendix A).

### Data and variables

Data were abstracted from patient admission to discharge from the ICU including admission demographic variables (age, gender, ICU admission/discharge date, ICU site and admission class [medical, surgical, neuroscience, or trauma]), outcomes (survival and length of stay), and time-stamped free-text notes. The latter was the main input to our study and included all textual data from 59 note parameters, related to 4 categories: psychosocial/family/social work (eg, family orientation)[n = 31, 52.5%], history/admission (eg, emergency contact)[n = 21, 35.6%], admission/discharge/transfer (eg, discharge location)[n = 6, 10.2%] and continuous quality improvement (eg, patient or family teaching)[n = 1, 1.7%]. These free-text notes were entered by members of the ICU care team including bedside registered nurses (RN), charge RNs, allied health professionals (registered respiratory therapists, physiotherapist, social workers, spiritual care), attending physicians, residents, or nurse practitioners. Free-text notes were available in all patients charts and ranged from 2 to 3566 characters (median 51, IQR 20–142).

To train, validate, and test the proposed NLP methods, a reference standard for family communication was created by an independent and blinded researcher who manually reviewed all EMRs within a random sample of 280 patients. All available free-text notes were individually analyzed, and a table was created to abstract measures of family communication. Table rows corresponded to notes (ie, 1 row for each evaluated note; eg, "son in to visit") and columns to measures (ie, 1 column for each measure of family communication; eg, child-boy/man-visit). Table cells, referring to the intersection of notes and measures, were filled out using ones and zeros, where ones represented the presence of the measure in the text and zeros represented its absence. As a result, all free-text notes were evaluated in relation to all measures of family communication. The characteristics of the patients included in the reference standard are presented in Table 1. Interrater agreement for the reference standard was evaluated by having a second blinded researcher classify a random sample of 10% (n = 28) of the 280 patients included in the reference standard assessment. The Cohen's Kappa estimate was 0.923.

### Natural language processing

The ability of NLP techniques to identify EMR documented family visitation and communication in ICUs was evaluated following a multistep framework. Figure 1 provides an overview of the framework, which is divided into 3 modules: (i) Preprocessing and dataset construction, (ii) NLP training, and (iii) performance evaluation. All computations were performed in Python 3.7.[11] Codes referring to the proposed framework are available on the Data Intelligence for Health Lab GitHub repository.[12]

## Preprocessing and datasets construction

### Preprocessing

The analysis of free-text notes is difficult for 4 reasons. First, texts do not follow a standard structure, and typographical errors might be present. Second, sometimes the relationship between people and the patient is indirectly registered, demanding text interpretation to correctly categorize (eg, her friend's wife). Third, in the English language, some family relationships do not explicitly identify gender (eg, "child"), which makes tasks that require such information difficult (eg, tasks within the subcategory "child-girl/woman"). Finally, some notes refer only to names (eg, "John called"), demanding a previous knowledge of the patient's network of family and friends. The main steps applied to overcome these difficulties and prepare records for the NLP training were:

1) Spaces between words and between words and punctuation were adjusted to be 1 blank character (eg, "brother –in– law" became "brother—in—law"). 2) Acronyms and informal names of relationships were identified by randomly inspecting notes within note parameters "Family Quick View Summary" and "Contact Information Family." Once common terms were discovered, their respective conventional references were assigned, resulting in a dictionary (eg, "sis" referring to "sister"). Occurrences of terms in the sampled data were manually inspected and there was no need for term disambiguation. This dictionary was applied to all notes to substitute terms (ie, whenever a known acronym or informal name was found in text, it was substituted by its conventional reference according to the dictionary). A separate dictionary was created to identify and adjust compound nouns of relationships into single terms (eg, "brother – in – law" referring to "brother_in_law"). This dictionary was also applied to all notes. 3) The Microsoft Excel spell checker was used to correct typographical errors. 4) Indirect registers of family members were substituted by a direct version (eg, "patient's sister's husband" became "patient's brother_in_law"). Whenever it was not possible to relate the mentioned person to the patient, the direct version assumed 1 of 3 options: "other_girl/woman," "other_boy/man" or "other_unknown" (eg, "her friend's wife" became "other_girl/woman"). 5) Words were reduced to their basic form, excluding inflectional endings (ie, lemmatization[13]) 6) A patients' network table was built. It considered textual information available in notes related to the parameters "Contact Information Family" and "Family Quick View Summary" and summarized known relationships and names associated to each patient. People's names were identified by either using the Stanford named entity recognition tagger[14] or by searching a list of names. The list of names, as well as their respective genders, was obtained after web crawling 2 online guides used to choose baby names.[15,16] Relationships were identified by searching for specific terms in the text. Once names and relationships were identified, the relation between them was estimated. Most free-text in these notes followed a similar structure (eg, "Name Surname [relationship]. Phone xxx-xxx-xxxx"), with differences in the order that information was presented. Therefore, it was assumed that a name or relationship would be related to the subsequent relationship or name, respectively. Whenever 2 subsequent names or relationships were found, it was assumed that the complementary information of the first occurrence was missing. 7) For each patient, known names were searched in all free-text notes and substituted by their corresponding relationships. Information about gender was included in ambiguous relationships based on their corresponding names. Whenever a name was common in both genders or no information was available regarding a specific name, the gender was considered "unknown." 8) Numbers and punctuation were excluded and all capital letters were lowercased.

**Table 1.** Characteristics of the patients included in the reference standard

| Characteristic | Patients Included in the Reference Standard (n = 280) |
|---|---|
| Age (years), mean ± SD (range) | 59.2 ± 15.0 (20–91) |
| Gender | |
| Man, n (%) | 149 (53.2) |
| Length of ICU stay (days), median (IQR) | 4.3 (5.6) |
| Mortality, n(%) | |
| ICU | 23 (8.2) |
| Hospital | 37 (14.4) |
| Admission class, n (%) | |
| Medical | 155 (55.4) |
| Surgical | 78 (27.9) |
| Neuroscience | 22 (7.9) |
| Trauma | 17 (6.1) |
| No admission category assigned | 8 (2.9) |

Abbreviation: IQR, interquartile range; SD, standard deviation.

### Dataset construction

The reference standard was organized into 156 datasets (Supplementary Appendix A). Each dataset was related to 1 measure of family communication (eg, child-boy/man-visit) and was structured according to a specific granularity of data. Each dataset included all available notes (ie, the entire corpus) and contained 2 fields of information. The first (target variable) was a binary variable representing the occurrence of the measure of family communication (ie, the annotation of zeros and ones in the reference standard). The sec-

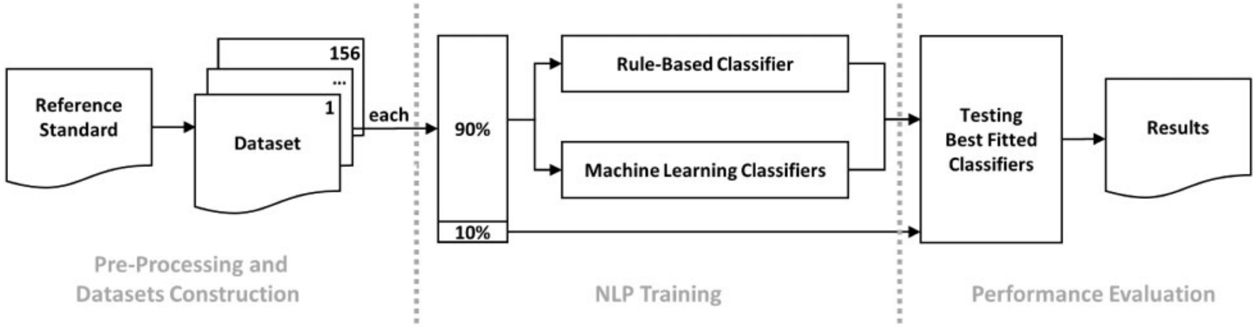

**Figure 1.** Framework overview.

ond (input variable) contained the full content of all free-text notes available in the reference standard. Two levels of data granularity were examined: micro and macro. Micro granularity considered data in their most granular form available (ie, individual notes in the reference standard), allowing temporal analysis of measures (ie, timestamps of notes were used as temporal references to measures). Macro granularity, on the other hand, considered data in a patient-based format (ie, each record of the dataset referred to all available data for a specific patient) and was used to identify the occurrence of measures during the whole ICU stay. Patient-based textual records were formed by appending all free-text notes related to each patient. The target variable of each patient-based record was determined using the "OR" operator (ie, if the target variable of any of the free-text notes related to the patient was equal to 1, the patient-based target variable assumed the value of 1; otherwise, it assumed the value of 0). Macro granularity was explored for all categories of family communication measures. Micro granularity, differently, was only explored for "Visits" and "Phone Calls," as the category "Documented Family or Friends" do not demand temporal analysis.

## NLP training

Two different approaches were tested during NLP training: a rule-based classifier (RBC) and machine learning classifiers (MLC). The information retrieval task was characterized as a binary classification problem, where 1 class was related to the presence of the information in the text (class "True") and the other class was related to its absence (class "False"). There is a tradeoff when choosing the proportion between training-validation and testing (also known as a holdout) subsets. Having larger training-validation subsets means models see more training examples, which might result in improved adjustment of methods. On the other hand, larger testing subsets are desirable to evaluate how well the model generalizes to unseen data. As such, data were randomly split by patients in stratified (ie, same ratio between classes as in the original data) training-validation (90%) and testing (10%) subsets. Only datasets that had at least 11 patients in each class were trained and evaluated. This was necessary to guarantee at least 1 patient of each class within each of the 10 folds used during the cross-validation grid search method (MLC training). Consequently, 53 datasets were excluded from training and evaluation. Most of the excluded cases related to the "Phone Calls" category (n = 26, 49.1%), followed by "Visits" (n = 20, 37.7%) and "Documented Family or Friends" (n = 7, 13.2%). The number of records in each class and dataset, as well as the relation of included and excluded datasets is available in Supplementary Appendix A.

### Rule-based classifier

Rule-based classifiers refer to any classification scheme that uses IF-THEN rules for class prediction. Rules are typically formed by a set of conditions (also called antecedent) that must be met to derive a conclusion (also called consequent).[17] The general rule of the proposed classifier states that if a record contains information related to the category (condition 1) and subcategory (condition 2) being analyzed, then it will be classified as "True"; otherwise, it will be classified as "False." Inclusion and exclusion criteria, based on the occurrence of specific words, expressions, and parameters, were used to evaluate the fulfillment of conditions. When at least 1 inclusion and no exclusion criteria were met, the analyzed condition was satisfied.

The train-validation subset was randomly split by patients in stratified training (90%) and validation (10%) subsets. The training subset was then manually analyzed, and several terms, expressions, and parameters were selected to compose candidate inclusion and exclusion criteria, which were evaluated according to their coverage (ie, number of records affected by the criterion) and to their class frequency. Candidate criteria that covered at least 2 records and that were present in only 1 class were kept. In total, 254 inclusion criteria and 25 exclusion criteria were considered in this study. For example, for the category "Phone Calls," the use of the code parameter "Comment Family Phone Call" was used as an inclusion criterion. However, notes from other code parameters also registered phone calls and, in those cases, specific terms were used as inclusion criteria, such as "called" or "telephoned." Some notes were specific to in-person visits (eg, code parameter "Comment Family In"), and their occurrence was used as an exclusion criterion. Some notes contained information regarding phone calls that were not related to communication with family or friends. For this reason, several terms and expressions were used as exclusion criteria, such as "ems called." In relation to subcategories, inclusion and exclusion criteria were formed by terms and expressions. For instance, the subcriteria "child-boy/man" considered only 1 term as inclusion criterion ("son"). Synonyms of terms and expressions were also included to expand the list of criteria. The inclusion and exclusion criteria for all categories and subcategories are available in Supplementary Appendix B. The validation subset was used to validate the criteria built on the training subset.

### Machine learning classifiers

Machine learning classifiers were tested as a second approach for NLP training. Tested classifiers were based on logistic regression, support vector machine, random forest, adaptive boosting, and neural networks architectures. Several options of text vectorization (ie, process to convert textual documents into numeric vectors) were also tested, including the use of uni-, bi-, and trigrams (ie, sequences of 1, 2, or 3 adjacent words), minimal and maximal frequency of variables (ie, *n*-grams) among documents (ie, notes), variable occurrence (binary, term-frequency and term frequency—inverse document frequency), and variable normalization (l1, l2, and no normalization).

The neural network classifier referred to a fully connected neural network comprised of 2 layers. The first layer presented the same number of neurons as the number of variables in the training subset (limited to 2048 neurons). The ReLU activation function was used and the dropout technique (dropout = 0.2) was applied to mitigate overfitting.[18] The second layer was comprised of 1 neuron and the sigmoid activation function was used to produce an output between 0 and 1. Several thresholds were tested to classify notes, ranging from 0.01 to 0.99 with a stride of 0.01. Computations were done using keras 2.2.4.[11] (neural network architecture) and sklearn 0.22.[11] (remaining architectures). Table 2 presents the tested methods, functions, and parameters. Complementary hyperparameters were the default options available in keras and sklearn.

A stratified 10-fold cross-validation grid search was applied to the training-validation set. It was used to identify the best combination of text vectorization, architecture, and hyperparameters. For each combination, the training-validation subset was divided into 10 mutually exclusive subsets of equal size and class distribution, such that 1 subset was used for validation and the remaining 9 sub-

**Table 2.** Tested methods, functions, and parameters

| Method | Tested functions and parameters |
|---|---|
| Text vectorization | Function: TfidfVectorizer() |
| | ngram_range: [(1,1), (1,2), (1,3)] |
| | max_df: [0.70, 0.80, 0.90, 0.95, 1.0] |
| | min_df: [2, 10, 50] |
| | binary: [False, True] |
| | use_idf: [False, True] |
| | norm: ['l1', 'l2', None] |
| Logistic regression | Function: LogisticRegression() |
| | penalty: 'none' |
| | class_weight: 'balanced' |
| | max_iter: 1e4 |
| | solver: 'saga' |
| Support vector machine | Function: SVC() |
| | kernel: 'linear' |
| | class_weight: 'balanced' |
| | max_iter: 1e4 |
| Random forest | Function: RandomForestClassifier() |
| | class_weight: 'balanced' |
| Adaptive boosting | Function: AdaBoostClassifier() |
| Neural networks | Function: Sequential() |
| | Layers: |
| | Dense(units = number of variables, activation = 'relu') |
| | Dropout(dropout = 0.2) |
| | Dense(units = 1, activation = 'sigmoid') |
| | optimizer: 'adam' |
| | loss: 'binary_crossentropy' |
| | metrics: 'binary_accuracy' |
| | epochs: 1000 |
| | callbacks: EarlyStopping(monitor = 'val_loss', min_delta = 0.01) |
| | output threshold: [0.01, 0.02, …, 0.98, 0.99] |

sets were used for training. For each combination, this process was carried out 10 times alternating the validation subset. Performance statistics were calculated from results, and the best fit of the grid search was the one that maximized the mean AUROC of the 10 predicted validation subsets.

## Performance evaluation

For each dataset, the testing subset was used to evaluate and compare the performance of the different NLP approaches with the reference standard; 2 classifiers were evaluated (RBC and the best fit for MLC). However, MLC were first retrained using the whole training-validation subset and their respective best hyperparameters (discovered during NLP training). Performance was measured using AUROC, sensitivity, and specificity for each dataset individually. Reference mean values (including 95% CI, based on mean and standard deviation) were estimated for different combinations of category and data granularity. All computations were performed in Python 3.7.[11]

# RESULTS

## NLP training

### Rule-based classifier

RBC's performance in identifying measures of family communication from EMRs for training and validation subsets was measured for each dataset individually (Supplementary Appendix C). The mean training sensitivity, specificity and AUROC, considering all datasets, were 0.870 (95% CI 0.847–0.894), 0.951 (95% CI 0.939–0.963), and 0.911 (95% CI 0.898–0.924), respectively. The mean validation sensitivity, specificity, and AUROC, considering all datasets, were 0.854 (95% CI 0.807–0.902), 0.960 (95% CI 0.949–0.972), and 0.907 (95% CI 0.883–0.931), respectively. Considering datasets individually, the mean difference between training and validation subsets for sensitivity, specificity, and AUROC were 0.144 (95% CI 0.110–0.178), 0.027 (95% CI 0.022–0.032), and 0.078 (95% CI 0.061–0.095), respectively.

### Machine learning classifiers

To identify the best combination of text vectorization, architecture and hyperparameters, a stratified 10-fold cross-validation grid search was used for each dataset individually. Performance results (Supplementary Appendix D) showed that the best fitted models were support vector machine (n = 41, 39.8%), followed by adaptive boosting (n = 31, 30.1%), neural networks (n = 27, 26.2%), logistic regression (n = 3, 2.9%), and random forest (n = 1, 1.0%). The most common hyperparameters were "ngram_range = (1, 1)" (n = 55, 53.4%), "max_df = 0.70" (n = 27, 26.2%), "min_df = 2" (n = 50, 48.5%), "binary = False" (n = 60, 58.3%), "use_idf = False" (n = 53, 51.5%), and "norm = l2" (n = 52, 50.5%). The mean training sensitivity, specificity, and AUROC, considering all datasets, were 0.979 (95% CI 0.973–0.985), 0.928 (95% CI 0.907–0.948), and 0.953 (95% CI 0.942–0.964), respectively. The mean validation sensitivity, specificity, and AUROC, considering all datasets, were 0.849 (95% CI 0.826–0.873), 0.902 (95% CI 0.883–0.921), and 0.876 (95% CI 0.861–0.891), respectively. Considering datasets individually, training and validation subsets overlapped for sensitivity, specificity, and AUROC in 49 (47.6%), 70 (68.0%) and 48 (46.6%) datasets, respectively.

## Performance evaluation

The testing subset was used to evaluate and compare the performance of the 2 tested approaches with the reference standard. RBC had the highest AUROC in 65 datasets compared to MLC in 21 datasets. RBC and MLC had the same AUROC in 17 datasets. Supplementary Appendix E presents performance results for both approaches and all datasets. As a reference, Table 3 shows results for different combinations of category and data granularity.

# DISCUSSION

In this retrospective cohort study, we developed and tested 2 NLP approaches that automatically identify individual events and modes of family communication from free-text contained in EMRs of ICU patients. The RBC approach had the highest AUROC in 65 datasets compared to MLC in 21 datasets. However, it is noteworthy that the implementation and maintenance of an RBC model requires substantial manual work. RBC is an effective and accurate approach to extract the frequency and mode of communication between healthcare professionals and family members of ICU patients from free-text EMRs.

There are several examples of NLP techniques to extract family history information in noncritically ill patients in the literature.[19–22] Similar to the current study, these extractions include recognition of family member mentions and definitions of the family member's relationship to the patient. For example, a rule-based NLP technique accurately categorized relatives, with a sensitivity of 0.93,[19] from

**Table 3.** Summary of results according to categories, data granularities, and approaches for the testing subset

| Category<br>(data granularity) | Measurement | Rule-based classifier<br>[mean (95% CI)] | Machine learning classifiers<br>[mean (95% CI)] |
|---|---|---|---|
| Documented Family or Friends (macro) | Sensitivity | 0.954 (0.882–1.000) | 0.875 (0.758–0.992) |
| | Specificity | 0.990 (0.980–1.000) | 0.971 (0.954–0.988) |
| | AUROC | 0.972 (0.934–1.000) | 0.923 (0.862–0.984) |
| Visits (micro) | Sensitivity | 0.761 (0.644–0.878) | 0.801 (0.740–0.861) |
| | Specificity | 0.958 (0.936–0.980) | 0.940 (0.916–0.964) |
| | AUROC | 0.860 (0.800–0.919) | 0.871 (0.839–0.902) |
| Visits (macro) | Sensitivity | 0.856 (0.745–0.967) | 0.674 (0.572–0.776) |
| | Specificity | 0.908 (0.873–0.942) | 0.871 (0.831–0.910) |
| | AUROC | 0.882 (0.820–0.943) | 0.772 (0.726–0.819) |
| Phone Calls (micro) | Sensitivity | 0.915 (0.861–0.970) | 0.800 (0.702–0.899) |
| | Specificity | 0.970 (0.948–0.993) | 0.780 (0.674–0.886) |
| | AUROC | 0.943 (0.916–0.970) | 0.790 (0.711–0.869) |
| Phone Calls (macro) | Sensitivity | 0.980 (0.939–1.000) | 0.689 (0.588–0.791) |
| | Specificity | 0.969 (0.952–0.987) | 0.776 (0.699–0.853) |
| | AUROC | 0.975 (0.952–0.998) | 0.733 (0.669–0.796) |

Note: bold text refers to the best mean AUROC among the tested NLP approaches.

admission notes of a multicenter primary care general hospital, which is similar to the performance of the RBC approach described in the current study. However, it is noteworthy that writing styles used in ICU notes may be different from non-ICU notes. For example, ICU documentation may be rushed and scanty with both critically ill patients requiring urgent intervention as well as larger quantities of data to document.

Measures of frequency and mode of communication between healthcare professionals and family members of patients are important to benchmark PFCC in critical care medicine, including the application of open visitation policies ,[23] family participation on ICU rounds ,[24] discussion of goals of care,[25] and shared decision-making.[10] A qualitative study on family participation in ICU rounds included suggestions to increase family member participation in educational initiatives and training.[26] Interrupted time series analyses using the NLP described in the current study could potentially evaluate the effects of education initiatives to increase the frequency of family participation in rounds using less resource-intensive methods compared to direct chart reviews. A recent study on the impact of adopting a family-centered ICU policy on nurses' ability to deliver PFCC included structured observations and on-site discussions with nursing staff over a 2-month period.[27] Though this study was able to record the frequencies and content of nurse–family interactions, it required labor-intensive data collection and may have missed unobserved interactions (when the study team was not present).[27] The NLP technique developed in our study could capture the frequency and mode of nurse–family interactions with the opportunity to scale to multiple ICUs, larger numbers of patients, and over longer periods of time.

There is a tension between documenting clinical information in coded formats versus recording it as free-text. Although coded data may be too limiting and may not allow the writer to express the full clinical picture, free-text is variable and includes shorthand, misspelled words, and acronyms. As such, using an NLP technique to extract meaning from this unstructured data also has limitations. Reasons for false positive associations include storytelling cases (eg, when a family member was reporting that someone visited or called) and incorrect identification of family members (eg, the method was unable to correctly relate names and relations due to different patterns in relation to the training data). Reasons for false negative associations include family members that were not identified (eg, method did not recognize a name) and the occurrence of notes registered using unusual parameters (eg, a phone call registered using the note parameter "Comment Family In").

## Study strengths

The current study has several strengths: first, family communication variables (ie, parameters and common terms appearing in notes) were identified by a team comprising ICU clinicians and researchers, which ensured we sampled the relevant sources of family communication in our EMR. Second, based on such variables, we manually built a reference standard using data sampled from 15 ICUs that provided a wide variety of note-writing styles and was comprehensive and reliable (excellent agreement between graders). Third, for each mode of communication and data granularity, we explored over 2700 different combinations of NLP methods and hyperparameters, giving a reliable base to support the methods employed. Finally, the methods identified in our study could be employed to evaluate the frequency and mode of communication between healthcare professionals and patient families in other clinical settings (eg, outpatient clinics) or patient populations (eg, pediatrics) where free-text documentation is used in EMRs.

## Study limitations

Our study has limitations. First, the NLP techniques were developed using data from a single large population using a single EMR; the operating characteristics with other EMRs and other jurisdictions and patient populations is unknown. However, in our study we have included details that would make it possible for other healthcare systems to adopt the same NLP technique to identify the frequency and modes of communication between healthcare professionals and patient families. Second, our methods are dependent on the quality of documentation in the EMR. Discrepancies likely exist between documented and actual communication (eg, undocumented communication). Moreover, the proposed NLP techniques capture the frequency and mode of communication between family and healthcare professionals, but provide no information on the quality of communication (ie, whether a meaningful communication occurred). The quality of communication is just as important as the frequency of communication, and future work is needed to develop NLP extraction methods in this sense. Third, manually extracting data from an EMR is labor intensive and, accordingly, our sample size was limited to 280 patients. The split proportion (90/10 for train-validation and test subsets) aimed to facilitate a better adjustment of the training methods. However, it may have led to significant differences between training and validation and between validation and test performance measurements. Further validation in a larger dataset and in additional jurisdictions would provide more precise estimates of the operating characteristics of these NLP techniques. Fourth, we investigated a limited number of measures of family communication, and additional measures may provide complementary elements to better target PFCC initiatives. Fifth, term disambiguation was not applied during the preprocessing phase, as it was not needed for the study dataset. It is unclear whether term disambiguation should be used when applying the reported methods in other datasets. Sixth, we investigated specific preprocessing methods. The operating characteristics for other preprocessing options are unknown. Seventh, inclusion and exclusion criteria of the RBC were mostly based on terms and expressions identified using the training subset. Despite having added synonyms to the inclusion and exclusion criteria, additional relevant terms and expressions may have been missed which could have impacted operating characteristics.

## CONCLUSION

Patient and family centered care is central to the care of critically ill patients. Multiple communication strategies with family members of critically ill patients have been observed including participation in rounds, person meetings, and telephone calls. Measures of the frequency and mode of family communication are important first steps to measure and benchmark PFCC.[28] This study identified RBC as the NLP technique with the best operating characteristics to automatically extract the frequency and mode of documented family visitation and communication from free-text EMRs. NLP provides a reliable, valid, and efficient tool to quantify communication between healthcare professionals and patient family members.

## FUNDING

This work was supported by University of Calgary Eyes High Postdoctoral Scholar Program (Filipe Lucini recipient).

## AUTHOR CONTRIBUTIONS

All authors contributed to the study by conceptualizing the study question (HTS) and design (all), developing the technical framework (FRL, JL), extracting and managing data (KDK, FRL), developing and testing the algorithms (FRL, JL), interpreting results (all), and drafting (FRL, KDK), revising (all), and approving (all) the manuscript.

## SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

## CONFLICT OF INTEREST STATEMENT

None declared.

## REFERENCES

1. Fix GM, VanDeusen Lukas C, Bolton RE, *et al.* Patient-centred care is a way of doing things: How healthcare employees conceptualize patient-centred care. *Health Expect* 2018; 21 (1): 300–7. doi: 10.1111/hex.12615
2. Fiest KM, McIntosh CJ, Demiantschuk D, *et al.* Translating evidence to patient care through caregivers: a systematic review of caregiver-mediated interventions. *BMC Med* 2018; 16 (1): 1–10. doi: 10.1186/s12916-018-1097-4
3. Davidson JE, Aslakson RA, Long AC, *et al.* Guidelines for family-centered care in the neonatal, pediatric, and adult ICU. *Crit Care Med* 2017; 45 (1): 103–28.
4. Au SS, Roze Des Ordons AL, Amir Ali A, *et al.* Communication with patients' families in the intensive care unit: a point prevalence study. *J Crit Care* 2019; 54: 235–8.
5. Kreimeyer K, Foster M, Pandey A, *et al.* Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. *J Biomed Inform* 2017; 73: 14–29.
6. Chiasson TC, Manns BJ, Stelfox HT. An economic evaluation of venous thromboembolism prophylaxis strategies in critically ill trauma patients at risk of bleeding. *PLoS Med* 2009; 6 (6): e1000098.
7. Stelfox HT, Brundin-Mather R, Soo A, *et al.* A multicentre controlled pre–post trial of an implementation science intervention to improve venous thromboembolism prophylaxis in critically ill patients. *Intensive Care Med* 2019; 45 (2): 211–22.
8. Brundin-Mather R, Soo A, Zuege DJ, *et al.* Secondary EMR data for quality improvement and research: a comparison of manual and electronic data collection from an integrated critical care electronic medical record system. *J Crit Care* 2018; 47: 295–301.
9. Clayton JA, Tannenbaum C. Reporting sex, gender, or both in clinical research? *JAMA* 2016; 316 (18): 1863–4.
10. Fiest KM, Krewulak KD, Ely EW, *et al.* Partnering with family members to detect delirium in critically ill patients. *Crit Care Med* 2020; 48 (7): 954–61.
11. Raschka S, Mirjalili V. *Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow* 2. Birmingham, United Kingdom: Packt Publishing; 2019.
12. Data Intelligence for Health Lab. GitHub Repository. 2020.https://github.com/data-intelligence-for-health-lab/NLP–modes_of_communication–families-healthcare_professionals Accessed July 7, 2020.
13. Lucini FR, Tonetto LM, Fogliatto FS, *et al.* Text mining approach to explore dimensions of airline customer satisfaction using online customer reviews. *J Air Transp Manag* 2020; 83: 101760.doi: 10.1016/j.jairtraman.2019.101760
14. Finkel JR, Grenager T, Manning C. Incorporating non-local information into information extraction systems by Gibbs sampling. In: *43rd Annual Meeting of the Association for Computational Linguistics;* June 25–30, 2005; Ann Arbor, MI.
15. Belly Ballot. https://babynames.net/Accessed July 19, 2019
16. Name Berry. https://nameberry.com/Accessed Jul 17, 2019
17. Tung AKH. Rule-based classification. In: *Encyclopedia of Database Systems*. Boston, MA: Springer, 2009: 2459–62.
18. Srivastava N, Hinton G, Krizhevsky A, *et al.* Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res* 2014; 15: 1929–58.
19. Friedlin J, McDonald CJ. Using a natural language processing system to extract and code family history data from admission reports. *AMIA Annu Symp Proc* 2006; 2006: 925.
20. Mowery DL, Kawamoto K, Bradshaw R, *et al.* Determining onset for familial breast and colorectal cancer from family history comments in the electronic health record. *AMIA Jt Summits Transl Sci Proc* 2019; 2019: 173–81.

21. Dai HJ. Family member information extraction via neural sequence labeling models with different tag schemes. *BMC Med Inform Decis Mak* 2019; 19 (S10): 1–12.
22. Shi X, Jiang D, Huang Y, *et al.* Family history information extraction via deep joint learning. *BMC Med Inform Decis Mak* 2019; 19 (S10): 1–6.
23. Ning J, Cope V. Open visiting in adult intensive care units—a structured literature review. *Intensive Crit Care Nurs* 2020; 56: 1–8.
24. Au SS, Roze Des Ordons AL, Leigh JP, *et al.* A multicenter observational study of family participation in ICU rounds. *Crit Care Med* 2018; 46 (8): 1255–62.
25. Boulton R, Boaz A. The emotional labour of quality improvement work in end of life care: a qualitative study of Patient and Family Centred Care (PFCC) in England. *BMC Health Serv Res* 2019; 19 (1): 1–9.
26. Roze Des Ordons AL, Au S, Blades K, *et al.* Family participation in ICU rounds—working toward improvement. *J Eval Clin Pract* 2020; 1–9. doi: 10.1111/jep.13345.
27. Rippin AS, Zimring C, Samuels O, *et al.* Finding a middle ground: Exploring the impact of patient- and family-centered design on nurse–family interactions in the neuro ICU. *HERD* 2015; 9 (1): 80–98.
28. Farrier CE, Stelfox HT, Fiest KM. In the pursuit of partnership: Patient and family engagement in critical care medicine. *Curr Opin Crit Care* 2019; 25 (5): 505–10.