Predicting Postoperative Acute Respiratory Failure in Critical Care using Nursing Notes and Physiological Signals

Vijay Huddar[1], Vaibhav Rajan[2], Sakyajit Bhattacharya[2] and Shourya Roy[2]

Abstract— **Postoperative Acute Respiratory Failure (ARF) is a serious complication in critical care affecting patient morbidity and mortality. In this paper we investigate a novel approach to predicting ARF in critically ill patients. We study the use of two disparate sources of information – semi-structured text contained in nursing notes and investigative reports that are regularly recorded and the respiration rate, a physiological signal that is continuously monitored during a patient's ICU stay. Unlike previous works that retrospectively analyze complications, we exclude discharge summaries from our analysis envisaging a real time system that predicts ARF during the ICU stay.**

Our experiments, on more than 800 patient records from the MIMIC II database, demonstrate that text sources within the ICU contain strong signals for distinguishing between patients who are at risk for ARF from those who are not at risk. These results suggest that large scale systems using both structured and unstructured data recorded in critical care can be effectively used to predict complications, which in turn can lead to preemptive care with potentially improved outcomes, mortality rates and decreased length of stay and cost.

I. INTRODUCTION

Acute Respiratory Failure (ARF) is a serious postoperative complication occurring in many patients. It occurs when the respiratory system fails in oxygenation and/or CO_2 elimination. Like other postoperative complications it worsens patient outcome and mortality and often prolongs hospital stays leading to increased costs.

Critical care units are data-rich environments where multiple parameters of patients are continuously monitored. Two important sources of patient information include physiological signals (or vital signs) that are measured in bedside monitors and various text sources that are regularly recorded such as nursing notes, reports from radiology, biochemistry and other investigations. While risk factors for many complications, including ARF, have been studied earlier, automated systems that can predict ARF using text sources and/or physiological signals in critical care have not been investigated, to the best of our knowledge. We take the first step in this direction of using machine learning techniques to analyze nursing notes and vital signs to predict ARF. The novelty of our system lies in the fact that we use only nursing notes and investigative reports during the patient's ICU (Intensive Care Unit) stay to predict the risk of ARF.

Discharge summaries[1] are intentionally excluded from our analysis since they are written at the end of the patient's stay and cannot be used in a real-time prediction system within the ICU. This also makes the problem harder since discharge summaries contain comprehensive information of patients' past and current medical history which nursing notes lack. Discharge summaries are formal documents and systems analyzing them (using linguistic techniques) rely on their grammatical structure. In comparison nursing notes are informally written and contain nonstandard and inconsistently used abbreviations.

We also evaluate the predictive performance when features obtained from these textual sources are combined with statistical features from respiration rate. Our experiments show that text sources contain sufficient statistical signal to distinguish between postoperative patients who develop ARF and those who do not develop ARF.

A. Respiratory Failure: Incidence and Risk Factors

Postoperative Acute Respiratory Failure is most commonly defined as the inability to be extubated 48 hours after surgery [3]. It occurs postoperatively in about 3% of all surgical cases and death within 30 days occurs in nearly 26% of the cases [8]. Incidence and mortality rates have been found to be similar in multiple studies across USA and surprisingly, there has been no change in the rates over the last 10 years [8].

Khuri et al. [13] show that ARF is an independent predictor of mortality and Dimick et al. [4] have studied the large cost and length of stay associated with ARF. A predictive model for ARF can hence also be utilized in predictive systems for mortality, cost and length of stay.

Risk factors for postoperative respiratory failure have been studied by several authors and can be divided into patient-specific and operation-specific factors. Patient-specific risk factors include health status (e.g. age, body-mass index), functional status, pulmonary status (including smoking), neurologic status, cardiac status, renal and fluid status and operation-specific risk factors include location of surgery and type of anesthesia used [1]. Various authors have investigated the risk of ARF in specific procedures such as liver transplantation [11], blood transfusion [5], head and neck surgeries [14] and abdominal surgery [2].

[1]International Institute of Information Technology (IIIT), Bangalore, India vijay.huddar at xerox.com

[2]Xerox Research Centre India (XRCI), Bangalore, India {sakyajit.bhattacharya, vaibhav.rajan, shourya.roy} at xerox.com

*This work was done when Vijay Huddar was an intern at XRCI.

[1]A discharge summary is a report written at the end of a patient's stay in the hospital. It typically includes details of the patient, the healthcare professionals involved during the stay, diagnoses, investigations and complications during the stay, past medical conditions as well as present and future treatment plans.

978-1-4244-7929-0/14/$26.00 ©2014 IEEE

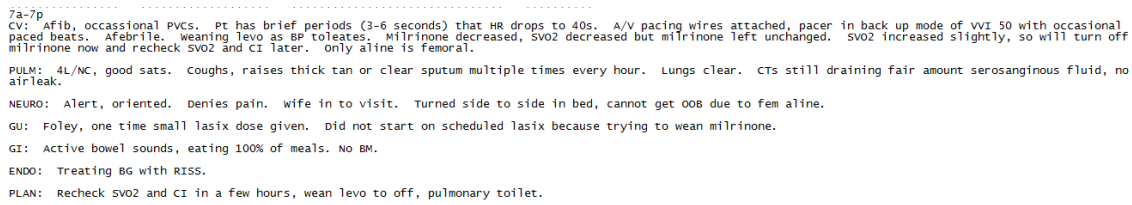

Fig. 1. Sample de-identified nursing notes from an ICU

B. Predictive Text Analytics in Healthcare

Many companies have developed experimental systems that can use heterogeneous electronic sources of data available in hospitals and provide predictive analytics services such as identifying patients at risk for diseases, treatment planning, and hospital resource management. Some well known examples include IBM's ICDA, a platform for intelligent care delivery analytics [7] and MatrixFlow, for analysis of disease progression using clinical event sequences [17]. These generic platforms have not been evaluated for complication prediction within ICUs.

Studies have illustrated the advantages of using (predictive) text analytics in healthcare. Hripcsak et al. [10] demonstrate that text analytics can detect clinical conditions in chest X-rays with a consistency that is indistinguishable from that of physicians reviewing the same reports. A recent study by Murff et al. [9] explores the use of text analytics to predict several postoperative complications. The predictive accuracy ranges from 64% for pneumonia to 91% for myocardial infarction. Source documents for the study include nursing notes, reports and discharge summaries.

None of these studies specifically investigate respiratory failure. Further, most of these studies use discharge notes in the text corpus for retrospective studies. Discharge summaries are written at the end of the patient's stay and cannot be used in a real-time predictive system. Predictive models for postoperative ARF have been designed based on preoperative clinical and demographic factors by several authors [8], [12], [15]. But no previous work has investigated the use of nursing notes and investigative reports during the patient's stay to predict ARF and we take the first step in this direction. We are also unaware of any previous work that combines text sources and physiological signals for complication prediction in ICUs.

II. Predicting Acute Respiratory Failure

The task of predicting ARF can be framed as a supervised classification task: using past cases from two classes A (those who develop ARF during their ICU stay) and B (those who do not), train a classifier that can distinguish between the two classes. The classifier can be used to predict the class label of a new patient based on his/her data. If the predicted label is A, the patient is considered to be at risk for ARF.

III. Data

The source of our data is MIMIC II [18], a publicly available database, part of Physionet [6], containing physiological signals and clinical data of more than 2300 patients in Critical Care. We restrict our study to those patients in the database for whom vital signs (physiological signals such as Respiration Rate, Blood Pressure etc. recorded by bedside monitors) have been provided. From this set, we extract the data of patients with postoperative respiratory failure, identified by ICD9 code 518.5.

We construct two datasets: in the first dataset we do not exclude any surgeries, thus the data contains samples from different kinds of surgeries. This dataset contains 122 patients from class A (those with ARF) and 684 patients from class B (those without ARF): a total of 806 patients. In the second dataset, we include only patients who have undergone coronary bypass surgery (procedure codes 3611–3614). This dataset contains 22 patients from class A (those with ARF) and 30 patients from class B (those without ARF): a total of 52 patients. We will refer to these datasets as dataset I and II respectively.

IV. Text Preprocessing and Feature Extraction

Observing the text data, we notice that the data is not completely unstructured but is structured into various headings such as "NEURO", "PULM" etc. See figure 1 for an example. These headings are neither consistent nor unique; for example, "CARDIO" is also written as "CV" and "CARD" in some notes. The key idea of our preprocessing method lies in realizing that the importance of a word or phrase in the text, in the context of a complication, is relative to the heading within which it resides. The significance of the same word differs when it is under the heading "PULM" than when it is under "NEURO". Hence, we extract features for each heading separately and assign an importance value to each word based on its frequency of occurrence in the training data. Only a fixed percentage of the extracted words are used in further preprocessing of the text. The complete sequence of steps performed is listed below. We denote by *text observation* all the text data for a single patient concatenated together which includes nursing notes and investigative reports of a patient but excludes the discharge summary.

- Extract all the headings from all text observations using predefined rules that identify headings. For example, a word in the beginning of a sentence, followed by colon is considered a heading.
- Eliminate headings and regroup data. Since headings are not consistently provided in the text, several different headings could in reality refer to the same word (example 'CARDIO', 'CV' and 'CARD' all refer to the same

heading). A dictionary of such "synonyms" are created to map synonymous headings together.

- Words within the same heading are processed together for each text observation. Stemming, stop word removal and punctuation removal are performed to obtain a list of stem words under each heading (for each text observation).
- Let $n_w(C, H)$ be the number of text observations from class C wherein the word w occurs under heading H. The importance of a word is computed as $I_w(H) = n_w(A, H) - n_w(B, H)$ for classes A and B. Thus words that are more frequent in class A are positive and those for class B are negative and the importance value is an approximate measure of the word's discriminatory power.
- For each heading H, we sort the words with respect to their importance values $I_w(H)$, select the top and bottom 5% (thus selecting from both the most negative and most positive values), and discard the rest. Within each heading, each of these words forms a feature and the number of occurrences of the word within a text observation is the feature value. A patient's data consists of a feature vector containing all the feature values (for all the headings).

With these preprocessing steps we obtain 6413 features in dataset I and 4662 features in dataset II.

V. Experimental Results

Three standard classifiers are used in our experiments: Support Vector Machine (SVM), Logistic Regression (LR) and Random Forest (RF). All implementations are in python using the scikit package [16]. Default parameter settings are used except for the option that adjusts for class imbalance. To evaluate the classification performance, we perform leave-one-out cross validation: for a dataset of n samples, n runs are executed for n distinct choices of test samples – the classifier is trained on the remaining $n-1$ samples and tested on the single chosen test sample. We use three performance metrics: accuracy, sensitivity and specificity. Accuracy is defined as the proportion of the total test samples in which a classifier accurately predicted the class. Sensitivity is defined as the proportion of the total samples from class A that are correctly identified as belonging to class A. Specificity is defined as the proportion of the total samples from class B that are correctly identified as belonging to class B. We test two settings. First, only features from text data are used. In the second setting, from the respiration rate (RR) we extract the mean and coefficient of variation for each patient and concatenate these two features with the text features (by adding extra columns to the text feature value matrix).

Figure 2 shows the accuracy, sensitivity and specificity values obtained (the values are averaged over the 806 cross validation runs) using text data alone and when RR is concatenated to the text data in dataset I. We note that both Logistic Regression and Random Forest obtain above 80% accuracy in classification. The addition of RR does not affect the performance of the classifiers significantly.

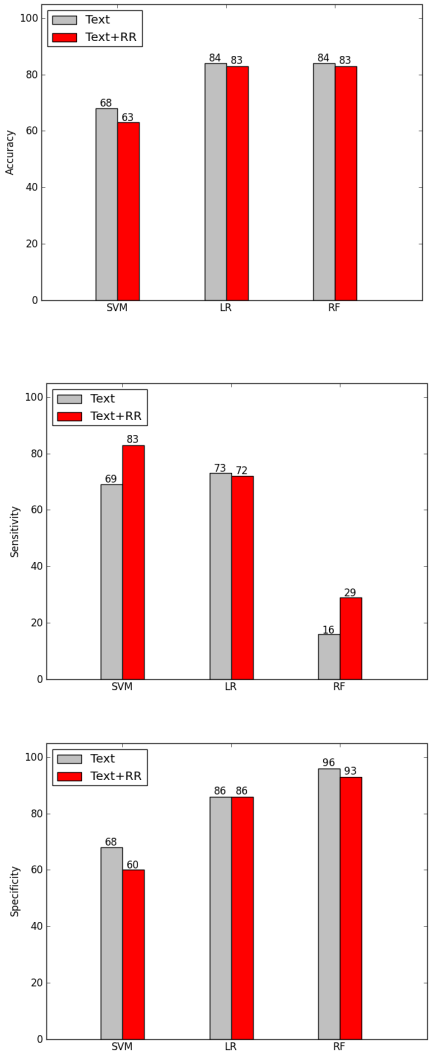

Fig. 2. Classifier performance: accuracy (above), sensitivity (middle) and specificity (below) in two scenarios, using text features alone and using text features with mean respiration rate (RR) in dataset I. Values are averages obtained over 806 leave-one-out cross-validation runs. 3 classifiers used – RF: Random Forest, LR: Logistic Regression, SVM: Support Vector Machine.

The specificity values are higher than sensitivity due to the unbalanced classes used in training the classifier. With more training samples from class B, we expect the sensitivity values as well as the total accuracy to increase. The accuracy remains close to 85% when only 25 features are used after transforming the dataset using PCA.

Figure 3 shows the results for dataset II. We observe that using text features alone the highest accuracy obtained is 92% with 81% sensitivity and 100% specificity (using SVM). On a dataset of 25 features, obtained through PCA, the performance is of the same order: 96% accuracy, 90% sensitivity and 100% specificity. The performance of any of the classifiers does not deteriorate on combining the features from text data and respiration rate. These experiments also suggest that it may be more valuable to build models for specific complications that are restricted to specific patients groups (in this case, patients who underwent cardiac surgery).

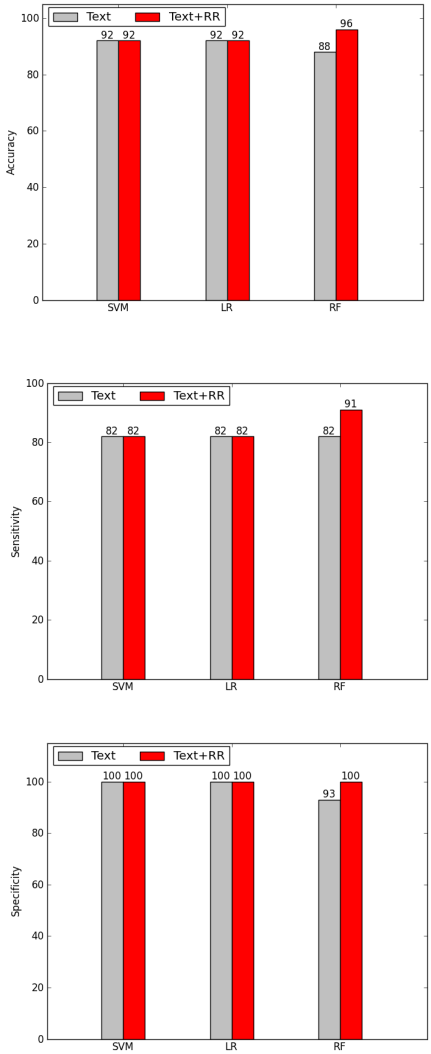

Fig. 3. Classifier performance: accuracy (above), sensitivity (middle) and specificity (below) in two scenarios, using text features alone and using text features with mean respiration rate (RR) in dataset II. Values are averages obtained over 52 leave-one-out cross-validation runs. 3 classifiers used – RF: Random Forest, LR: Logistic Regression, SVM: Support Vector Machine.

VI. CONCLUSION

We present a novel approach to predict postoperative acute respiratory failure (ARF) using text sources of information commonly available in ICUs. We also evaluate the performance when these are combined with features derived from vital signs such as respiration rate. Our experiments strongly suggest that nursing notes contain sufficient statistical signal to distinguish between postoperative patients who develop ARF and those who do not develop ARF. In the future we would like to explore dimensionality reduction techniques and data fusion techniques for combining heterogeneous sources of data – text and physiological signals, to improve upon our preliminary results. We believe that text sources within ICU can be powerful sources for building analytics tools for predicting complications like Respiratory Failure.

REFERENCES

[1] Ahsan M Arozullah, Jennifer Daley, William G Henderson, Shukri F Khuri, National Veterans Administration Surgical Quality Improvement Program, et al. Multifactorial risk index for predicting postoperative respiratory failure in men after major noncardiac surgery. *Annals of surgery*, 232(2):242, 2000.

[2] Jo Ann Brooks-Brunn. Predictors of postoperative pulmonary complications following abdominal surgery. *CHEST Journal*, 111(3):564–571, 1997.

[3] E Stanley Crawford, Lars G Svensson, Kenneth R Hess, Salwa S Shenaq, Joseph S Coselli, Hazim J Safi, Prita K Mohindra, and Victor Rivera. A prospective randomized study of cerebrospinal fluid drainage to prevent paraplegia after high-risk surgery on the thoracoabdominal aorta. *Journal of vascular surgery*, 13(1):36–46, 1991.

[4] Justin B Dimick, Steven L Chen, Paul A Taheri, William G Henderson, Shukri F Khuri, and Darrell A Campbell Jr. Hospital costs associated with surgical complications: a report from the private-sector national surgical quality improvement program. *Journal of the American College of Surgeons*, 199(4):531–537, 2004.

[5] John P Ebert, Ben Grimes, and Kurt MW Niemann. Respiratory failure secondary to homologous blood transfusion. *Anesthesiology*, 63(1):104–106, 1985.

[6] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. Ch. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000 (June 13).

[7] David Gotz, Harry Stavropoulos, Jimeng Sun, and Fei Wang. ICDA: A platform for intelligent care delivery analytics. In *AMIA Annual Symposium Proceedings*, volume 2012, page 264. American Medical Informatics Association, 2012.

[8] Himani Gupta, Prateek K Gupta, Xiang Fang, Weldon J Miller, Samuel Cemaj, R Armour Forse, and Lee E Morrow. Development and validation of a risk calculator predicting postoperative respiratory failurerisk calculator predicting respiratory failure. *CHEST Journal*, 140(5):1207–1215, 2011.

[9] Murff HJ, FitzHenry F, Matheny ME, and et al. Automated identification of postoperative complications within an electronic medical record using natural language processing. *JAMA*, 306(8):848–855, 2011.

[10] George Hripcsak, Carol Friedman, Philip O Alderson, William DuMouchel, Stephen B Johnson, and Paul D Clayton. Unlocking clinical data from narrative reports: a study of natural language processing. *Annals of internal medicine*, 122(9):681–688, 1995.

[11] Ching-Tzu Huang, Horng-Chyuan Lin, Shi-Chuan Chang, and Wei-Chen Lee. Pre-operative risk factors predict post-operative respiratory failure after liver transplantation. *PloS one*, 6(8):e22689, 2011.

[12] Robert Johnson, Ahsan Arozullah, Neumayer Leigh, William G. Henderson, Patrick Hosokawa, and Shukri F. Khuri. Multivariable predictors of postoperative respiratory failure after general and vascular surgery: Results from the patient safety in surgery study. *Journal of the American College of Surgeons*, 204(6):1188–1198, 2007.

[13] Shukri F Khuri, William G Henderson, Ralph G DePalma, Cecilia Mosca, Nancy A Healey, Dharam J Kumbhani, et al. Determinants of long-term survival after major surgery and the adverse effect of postoperative complications. *Annals of surgery*, 242(3):326, 2005.

[14] Timothy M McCulloch, Niels F Jensen, Douglas A Girod, Terance T Tsue, and Ernest A Weymuller. Risk factors for pulmonary complications in the postoperative head and neck surgery patient. *Head & neck*, 19(5):372–377, 1997.

[15] Kazuyo Nakahara, Kiyoshi Ohno, Junpei Hashimoto, shinichiro Miyoshi, Hajime Maeda, Akihide Matsumura, Takatoshi Mizuta, Akinori Akashi, katuhiro Nakagawa, and Yasunaru Kawashima. Prediction of postoperative respiratory failure in patients undergoing lung resection for lung cancer. *The Annals of Thoracic Surgery*, 46(5):549–552, 1988.

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

[17] Adam Perer and Jimeng Sun. Matrixflow: Temporal network visual analytics to track symptom evolution during disease progression. In *AMIA annual symposium proceedings*, volume 2012, page 716. American Medical Informatics Association, 2012.

[18] Mohammed Saeed, C Lieu, G Raber, and RG Mark. Mimic II: a massive temporal ICU patient database to support research in intelligent patient monitoring. In *Computers in Cardiology, 2002*, pages 641–644. IEEE, 2002.