Automating the Classification of Post-surgery Complication Grades

EJ Jung[*] Robin Costé[†] Ashish Gehani[‡] Samrat Ray[§] Chahal Arora Surajit Nundy[¶]

Abstract

We aim to develop a fast, efficient, and automated way to assess the severity of complications after surgeries, so that medical professionals may compare different surgery and treatment options, as well as tailor the treatment plan for individual patients. The Clavien-Dindo (CD) classification is commonly used to categorize post-surgical complications into different grades. We used the data of 494 patients from Raxa, an Electronic Medical Record (EMR) provider in India, to build a model that automatically performs a binary classification of severity with 84% accuracy. We also report preliminary results of predicting the CD grade using this model.

Keywords: EMR, post-surgery complications, Clavien-Dindo, data mining, predictive model

1 Introduction

Evaluating a patient for surgical procedure requires the surgeon to assess not just the patient's symptoms but also his physical characteristics and underlying conditions. Surgeons take a comprehensive look at all available information to decide on a treatment plan. To make an informed choice, surgeons need to be able to *compare* the likelihood of post-surgical complications, based on pre-existing conditions and different procedural options, such as open versus laparoscopic surgery.

The *Clavien-Dindo (CD)* classification [3] and the *Comprehensive Complication Index (CCI)* [11] are commonly used to describe the severity of post-surgical complications. The CD classification contains 7 grades: I, II, IIIa, IIIb, IVa, IVb, V [4]. These metrics are designed to categorize and generalize the post-surgery complications in an objective and reproducible way. They aim to produce numerical values that doctors can use to compare the outcomes of different surgical procedures, evaluate their effectiveness, and assess the potential risks. However, the manual process of assigning the CD grade and CCI score is time consuming, and the information is not always available.

Our aim is to provide a fast, efficient, and automated way to suggest a CD grade so that the doctors can save time and create a valuable knowledge base.

2 Data Description

Raxa [10] is an *electronic medical record (EMR)* software vendor in India. Their system's data model is derived from the OpenMRS project, an open-source EMR format. Raxa provided the following three kinds of data.

1. **Medical records:** Raxa provided a sanitized database of medical records for 494 patients. This database recorded 2,938 days of patient hospital stays, 17,127 encounters between patients and medical professionals, and 113,897 lines of observations made by medical professionals. The data was anonymized by removing personally identifiable information, such as patients' names, birthdays, and phone numbers. All the data was from gastrointestinal surgery patients.

   The medical records include both structured and unstructured data. The structured data contains quantitative health measurements, such as vital signs and the medication doses. The unstructured data is free-form text, such as after-visit notes written by medical professionals.

2. **Ground truth:** Raxa provided a table of "ground truth" – that is, the CD grade assigned by a medical doctor to a *patientID*, an anonymized identifier for a patient in the database. However, we were not able to use all the patients' CD grades in this table. This was because some patients had multiple surgeries with different CD grades assigned afterwards. In such cases, we were not able to tell which surgery an anonymized database entry corresponded to. We also had patients whose medical records in the database were not complete. For example, if a patient's records include a computed tomography (CT) scan, the expected CD grade would be *IIIa* (indicating a radiological intervention). However, if the CD

[*] Department of Computer Science, University of San Francisco

[†] École Polytechnique, while visiting SRI International

[‡] SRI International

[§] Sir Ganga Ram Hospital

[¶] Raxa

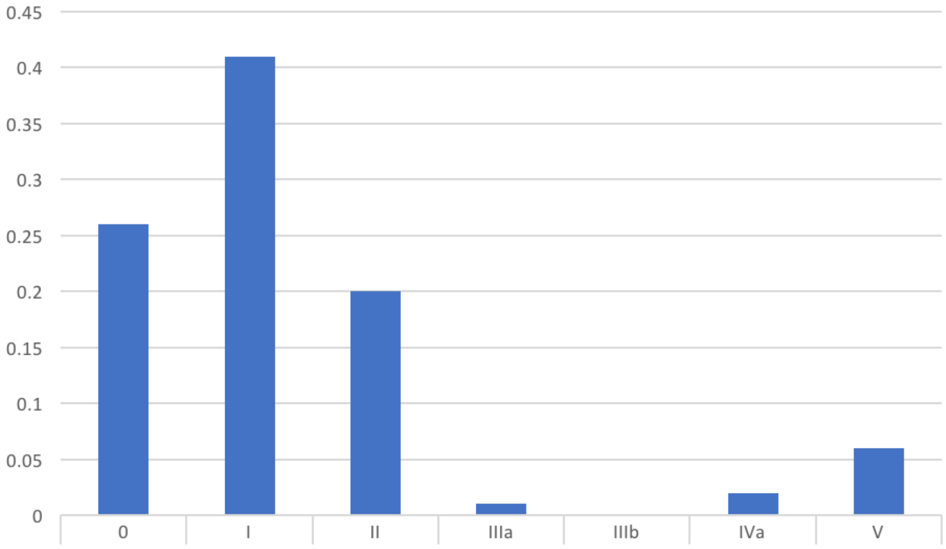

Figure 1: Fraction of all data that was contained in each CD grade.

grade reported was *II*, the CT scan might have been performed as part of a normal post-surgery procedure. Such outliers were manually removed from the dataset.

3. **Medical ontology:** Even though the free-form text was entered in English, in many cases medical shorthand was used. For example, when "BP 110/70" was recorded in the unstructured data portion of a patient's medical record, the intended semantics are that the "systolic blood pressure = 110" and that the "diastolic blood pressure = 70". To help with extracting and interpreting such text features, we utilized a custom ontology provided by medical doctors.

CD Grade Distribution: The original Clavien-Dindo scale consists of a system of letter grades [0, I, II, IIIa, IIIb, IVa, IVb, V], for which 0 means no complications occurred after surgery. To support the development of a classifier, we use the Clavien-Dindo Numeric (CDN) model instead. In this model, the grades are represented by numerical variables [0, 1, 2, 3, 4, 5, 6, 7]. The distribution of CD grades in our data is shown in Figure 1. The majority of the patients had CDN 0–2, which represents the range from no complications to relatively mild ones.

A CD grade of *IIIa* or higher indicates that the patient had more serious complications. For example, they may have needed to undergo surgical, endoscopic, or radiological intervention. This motivated us to split the distribution into two categories – CD $<$ *IIIa*, and CD $\geq$ *IIIa*. The result is shown in Figure 1. Only 8% of the patients were assigned a CD grade of *IIIa* or higher – that is, our dataset is highly skewed.

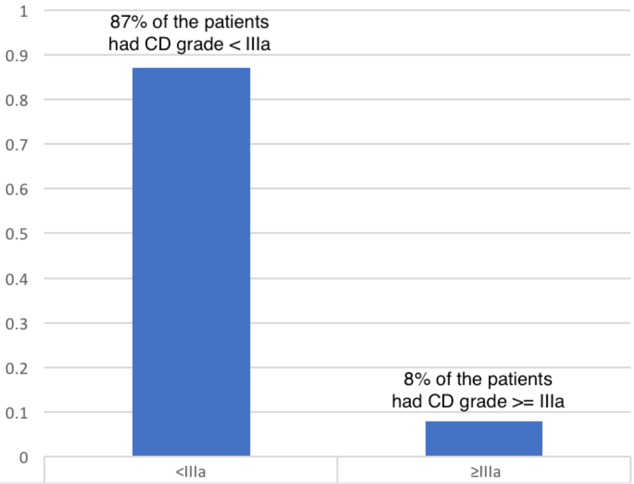

Figure 2: Fraction of CD grades in two categories. This split became the basis of the classification model.

3 Constructing Classifier Features

The raw data had to be transformed into input that could be used with classification algorithms. Next, we describe our approach.

3.1 Extracting Terms: *cTakes* [6] is an open-source natural language processing tool. It is designed to extract information from free-form text found in EMRs, such as doctors' notes following a patient' hospital visit. We use cTakes to extract a consistent set of terms from various descriptions of the same symptom. For example, an observation may contain "abundant bleeding from the arm" or "The arm was bleeding heavily". In both cases, the output of cTakes contains "arm", "bleeding" and "arm bleeding." Similarly, the output will contain "fever", regardless of whether a doctor's note indicates "fever" or "febrile". We used cTakes to extract 2,232 terms in the following categories (that were defined by the tool): *ProcedureMention*, *SignSymptomMention*, *DiseaseDisorderMention*, *AnatomicalSiteMention*, *MedicationMention*.

We found that there was high variance in the number of observations per patient. Similarly, medical professionals exhibit varying degrees of verbosity. Consequently, we opted to use the output of cTakes to indicate whether each term was present or absent, rather than counting the frequency of occurrence of each term.

3.2 Value Selection: The OpenMRS data model contains fields for numerical data. However, we noticed that in practice most medical professionals did not utilize them. Instead, they recorded quantitative information, such as vital signs, in fields defined for unstructured data, such as observation notes. As a result, we extracted 8 vital signs from the free-form text: *pulse*, (liquid) *output*, (liquid) *intake*, *temp*(erature), *stoma*, *RT* (respiratory therapy), and *SPO2* (oxygen saturation).

For many of these features, a patient has multiple measurements, even on a single day. We tested several ways of grouping multiple values: computing the average, the median, or the difference between the minimum and maximum. Finally, we opted to simply use the last value reported. The motivation is that it has the most up-to-date information. In addition, it is also the most likely to capture an abnormal value if a complication was in the process of developing.

3.3 Custom Dictionary: We built a custom dictionary to map certain events to a corresponding CD grade. The utility of this can be seen through an example: Per the CD grade definitions, single organ failure would result in grade *IVa* and multiple organ failure would be grade *IVb*. When we see "renal failure" in a patient's data, their CD grade is *IVa* if there was no other organ failure. We capture this as an entry in the dictionary.

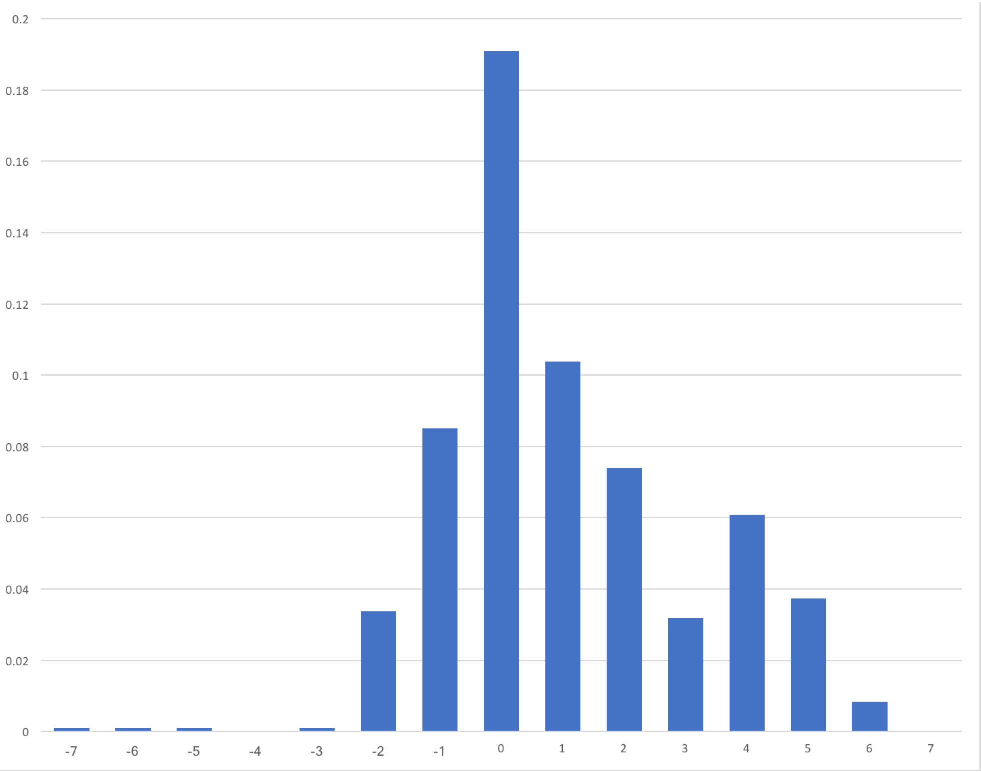

Figure 3: Errors using the custom dictionary: $CDN_{dictionary} - CDN_{true}$.

Using the custom dictionary, 80% of the CD grades were correctly assigned. The false positives and false negatives were then manually investigated to remove outliers from the data set. We found our dictionary tends to overestimate the CD grade, as shown in Figure 3. $CD_{dictionary}$ is the CD grade assigned based on our dictionary. CD_{true} is the CD grade assigned by medical doctors. 50.7% of the patients have $CD_{dictionary} \geq CD_{true}$, of which 31.6% of the patients have $CD_{dictionary} > CD_{true}$.

We suspect some of the false positives are due to incomplete data and illustrate this with an example. A patient may have complications after surgery that necessitate admission to the ICU (intensive care unit). This qualifies for a CD grade of IVa. However, it is also possible for a patient to be sent to the ICU as part of a normal recovery. This may be necessary for a number of reasons, such as the patient being in a chemically-induced coma or requiring assisted breathing facilities. Though this is termed POICU (post-operative ICU), some doctors may simply mention ICU. As a result, the assignment of CD grade IVa becomes an overestimate. This can be seen in the bottom right corner of Figure 4 – in the cases where $CD_{dictionary}$ is IVa or higher, and CD_{true} is 1 or 2.

It is not sufficient to just use a dictionary of terms to assign CD grades. This can be seen in the frequency distribution of dictionary keywords over CD grades in the ground-truth table. Figure 5 shows the distribution for the word "expire" provides a clear signal for grade V. In contrast, the distribution for "ICU" is spread over almost all CD grades, including 0 (indicating no complications), even though an ICU stay counts as a grade IV event.

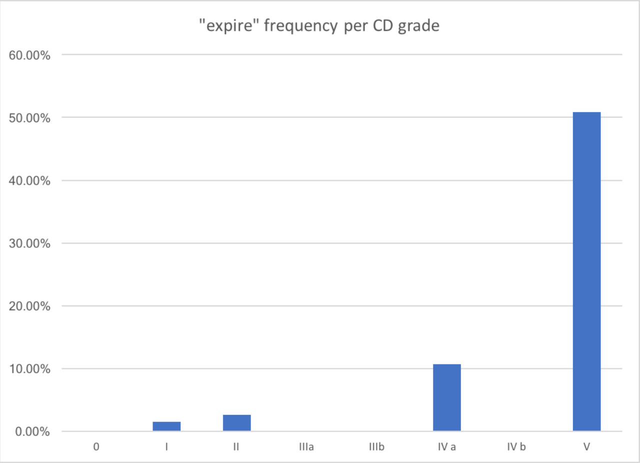

Figure 5: Frequency of 'expire' per CD grade.

As shown in Figure 6, "expire" is a clear signal for grade V whereas "ICU" is used in almost all grades, including 0 (no complication), even though an ICU stay counts as a grade IV event. In other words, the dictionary alone does not provide clear enough signals to achieve a highly accurate CD classification system.

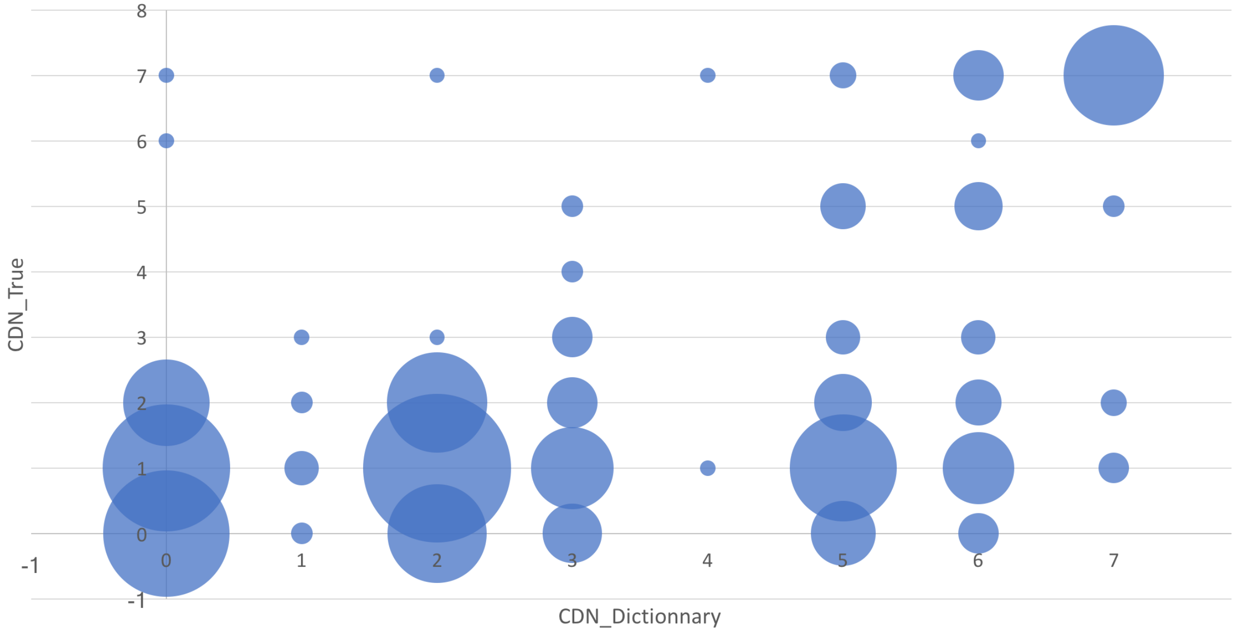

Figure 4: Comparing $CDN_{dictionary}$ and CDN_{true}.

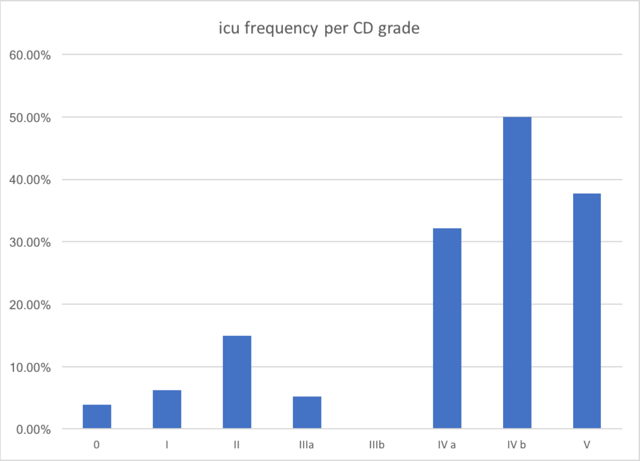

Figure 6: Frequency of 'icu' per CD grade.

4 Model Building

We aimed to develop an automated CD grade classifier with the following properties:

- **Robust:** Each medical professional has a unique style of writing that manifests as diversity in the text describing patients' symptoms and treatments. Our feature extraction normalizes the differences in style, using cTakes and our custom dictionary. This makes the classifier robust in the face of writing idiosyncrasies.

- **Efficient:** The system must be fast enough to handle input describing each patient discharged in a single day at a large hospital. At the same time, it should be able to operate with the limited computing resources of a small hospital. Our current implementation takes about an hour on a laptop to process 500 patients' records. This includes (i) extracting features using cTakes, (ii) selecting a subset of high-utility key-value pairs, and (iii) identifying whether terms in the custom dictionary are present.

- **Comprehensible:** We anticipate medical professionals will trust the system's output if its design choices are clear. Consequently, we solicited and utilized input from doctors when selecting features from the medical records. In particular, we aimed to ensure they could understand which symptoms, observations, and other data influenced the classification.

With these goals in mind, we tried several classification algorithms, including *decisions trees*, *random forests*, and *naive Bayes*. We did not investigate the use of neural networks because the dataset size was small (putting it at high risk of overfitting).

Data Skew: As described in Section 2, a CD grade of *IIIa* or higher indicates serious complications. Based on this, we aimed to classify patients into two groups – those with CD grades below *IIIa* and those with CD grades of *IIIa* or higher. Initially, we used 10-fold cross-validation with an 80-20 training-test data split to build our model. The skew in the data, seen in Figure 1, posed a challenge. Even a zero-classifier, which randomly assigns low or high CD grades to patients, achieves an accuracy of 81%. Of the three algorithms tried, *random forests* performed best. However, its 83% accuracy is not significantly higher than the zero-classifier.

To address the skew in the data, we constructed a balanced dataset. The number of patient records with low CD grades was limited to match the number of records of those with high CD grades. By construction, the zero-classifier's accuracy was reduced to 50%. Using this dataset, we built another model using the earlier approach (of applying 10-fold cross-validation with an 80-20 training-test data split). Table 1 shows the confusion matrix obtained with the updated dataset. The model has an accuracy of 79.3%.

Threshold Selection: When the model is applied to a patient's medical record, it computes a quantitative estimate of complication severity. If this is above a threshold, the patient is classified in the category of patients with CD grades of *IIIa* or higher. The selection of this threshold therefore determines a tradeoff between false positives and false negatives. If the patient should not have been classified as having CD grade *IIIa* or higher but was, this constitutes a false positive. A false negative occurs when the patient should have

Classified / Actual	CD < IIIa	CD ≥ IIIa	Total
CD < IIIa	423 (74.7%)	143 (25.3%)	566 (50%)
CD ≥ IIIa	91 (16.0%)	475 (84.0%)	566 (50%)
Total	514 (45%)	618 (55%)	1132

Table 1: Confusion matrix using the balanced dataset.

been so classified, but was not. In borderline cases, the choice of threshold can change the category the patient is classified into. When selecting a threshold, we reasoned that it is preferable for the model to produce an overestimate rather than an underestimate because a doctor can manually override such false positives.

5 Predicting CD Grades

We explored whether the system could predict the CD grade of a patient before a qualifying event. For example, assume a patient experiences renal failure on the tenth day after surgery. We studied whether the system could predict (before the tenth day) this patient was likely to experience a severe complication – in this case, one with CD grade *IVa*.

Each patient's data was split into sets of records from individual post-operative days (PODs). Using this, a list of recording dates and values was constructed for each feature. In total, the dataset contained elements spanning 2,938 days and 494 patients.

Validating the model's predictive power proved challenging due to the nature of the dataset. To explain the types of issues encountered, we describe three patients' cases. Their data is plotted in Figure 8. One patient had been assigned CD grade *I*, while the other two had received CD grade *IVa*. The x axis represents the number of days elapsed since the patient's surgery. The y axis corresponds to the complication severity. The green line shows the classifier's threshold. Values above it indicate a CD grade of at least *IIIa*.

In each case, it was not possible to make a prediction before the threshold was crossed. However, the reasons are different for each patient. In the first case, the patient's score remains below the threshold for the entire hospital stay. This means there was no point in time when the threshold was crossed. The second patient developed severe complications after POD 5. However, there were no records between POD 5 and 9. As a result, the prediction does not cross the threshold until POD 9. The third patient developed severe complications immediately after the surgery. It was therefore not possible to make a prediction before the threshold was crossed.

If we had CD grades for every POD of all patients, we could have validated the predictive power of the model. In practice, we had a single CD grade for each patient. Recall that the CD grades are manually assigned by doctors. As a result, we did not have ground truth for the daily CD grade of each patient. Instead, we studied the CD grade predicted by the model on each day of each patient's stay in the hospital. Each predicted grade was compared the actual CD grade assigned at the end by a doctor. The results are summarized in Figure 7. For each predicted CD grade, the confusion matrix shows the distribution of the actual CD grades.

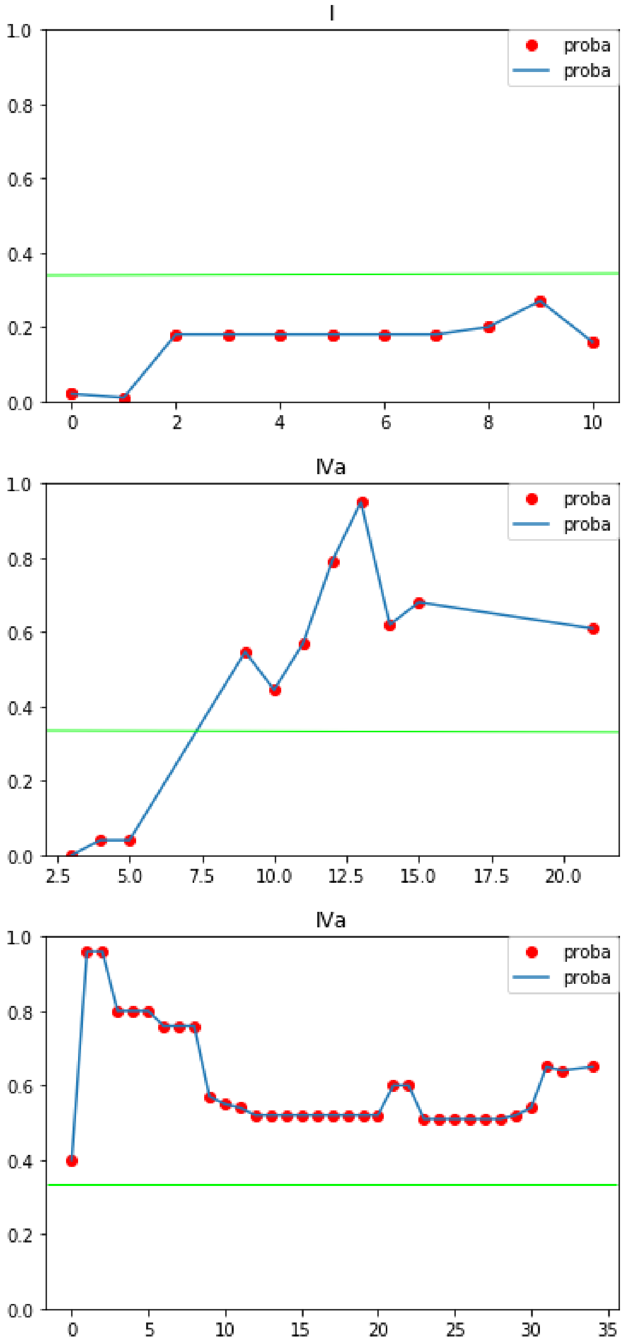

Figure 8: Predicted severity as a function of post-operative days elapsed.

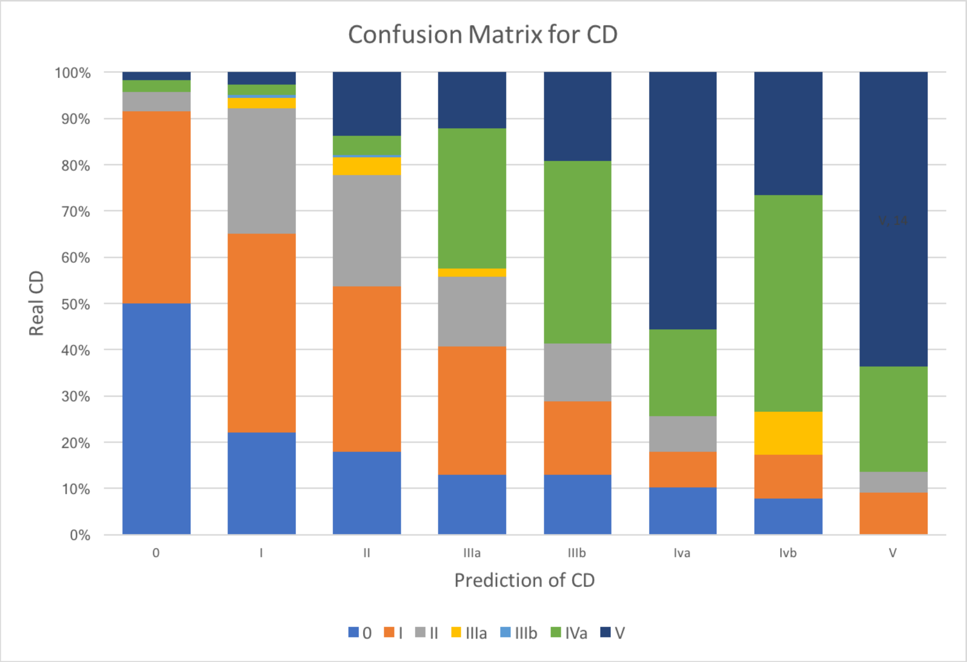

Figure 7: Actual versus predicted CD grades.

6 Related Work

Researchers have been applying machine learning to medical data for over four decades [7]. Advances in computing have facilitated significant improvements in machine learning over the years. In 1986, the U.S. National Library of Medicine created the Unified Medical Language System (UMLS) [2]. It codifies over a billion concepts, helping standardize medical expressions. In 2002, Liu et al. [8] used machine learning techniques to reduce ambiguity in the UMLS. This has facilitated automated processing of free-form medical text.

Rajkomar et al. [9] used over 46 million data points from 216,221 patients to predict in-hospital mortality (equivalent to CD grade V in our study) with 93–94% accuracy. They utilized deep learning to analyze electronic health records. To compare our system's accuracy to theirs, we would need to test perform tests limited to CD grade V classification. However, the small size of our dataset precluded this.

A number of other researchers have performed similar studies in other areas of medicine. For example, Blamey et al. [1] used machine learning to predict risk in biliary surgery; Cruz and Wishart applied machine learning techniques for cancer prediction and prognosis [5]. Similarly, artificial intelligence has been used to augment image recognition in medicine – Xu et al. [13] used such an approach to find cells, while Wolberg et al. [12] applied it to identify cancerous regions in images.

7 Conclusion

We studied whether the assignment of post-surgery complication severity grades could be automated. To build a model, we extracted a set of features from the electronic medical records of 494 patients, a custom dictionary, and specific schema. Using the patient data and a medical text processing tool, we trained a classifier that was then able to achieve its goal with 84% accuracy.

We also attempted to build and use a model to predict when a patient's complications would cross a pre-defined threshold. The lack of ground truth prevented a complete validation. However, we were still able to compare the model's daily predictions to the final ground-truth value for each patient. The results of this were summarized.

The results could be improved in a number of ways. Separate models could be developed for different types of surgeries. The increased specificity may improve the precision of each model. Pre-surgery data, such as family history, could be included. This would allow risk factors to be incorporated into the classifier's model. Finally, an interactive version of the tool could allow doctors to select between multiple options they may be considering, based on predicted outcomes.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant ACI-1547467. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] Stephen Blamey, Kenneth Fearon, Harper Gilmour, Denis Osborne, and Derek Carter, **Prediction of risk in biliary surgery**, *British Journal of Surgery*, Vol. 70(9), 1983.

[2] Olivier Bodenreider, **The Unified Medical Language System (UMLS): Integrating biomedical terminology**, *Nucleic Acids Research*, Vol. 32, 2004.

[3] Pierre Clavien, Jeffrey Barkun, Michelle de Oliveira, Jean Nicolas Vauthey, Daniel Dindo, Richard Schulick, Eduardo de Santibanes, Juan Pekolj, Ksenija Slankamenac, Claudio Bassi, Rolf Graf, Rene Vonlanthen, Robert Padbury, John Cameron, and Masatoshi Makuuchi, **The Clavien-Dindo classification of surgical complications**, *Annals of Surgery*, Vol. 250, 2009.

[4] Clavien-Dindo grades, `http://www.assessurgery.com/clavien-dindo-classification`

[5] Joseph Cruz and David Wishart, **Applications of machine learning in cancer prediction and prognosis**, *Cancer Informatics*, Vol. 2, 2006.

[6] cTakes, `http://ctakes.apache.org/`

[7] Rahul Deo, **Machine learning in medicine**, *Circulation*, Vol. 132(20), 2015.

[14]

[8] Hongfang Liu, Stephen Johnson, and Carol Friedman, **Automatic resolution of ambiguous terms based on machine learning and conceptual relations in the UMLS**, *Journal of the American Medical Informatics Association*, Vol. 9(6), 2002.

[9] Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew Dai, Nissan Hajaj, Peter Liu, Xiaobing Liu, Mimi Sun, Patrik Sundberg, Hector Yee, Kun Zhang, Gavin Duggan, Gerardo Flores, Michaela Hardt, Jamie Irvine, Quoc Le, Kurt Litsch, Jake Marcus, Alexander Mossin, Justin Tansuwan, De Wang, James Wexler, Jimbo Wilson, Dana Ludwig, Samuel Volchenboum, Katherine Chou, Michael Pearson, Srinivasan Madabushi, Nigam Shah, Atul Butte, Michael Howell, Claire Cui, Greg Corrado, and Jeff Dean, **Scalable and accurate deep learning for electronic health records**, *arXiv:1801.07860*, 2018.

[10] Raxa, `https://www.raxa.com/`

[11] Ksenija Slankamenac, Rolf Graf, Jeffrey Barkun, Milo Puhan, and Pierre-Alain Clavien, **The Comprehensive Complication Index (CCI): A novel continuous scale to measure surgical morbidity**, *Annals of Surgery*, Vol. 258(1), 2013.

[12] William Wolberg, Nick Street, and Olvi Mangasarian, **Machine learning techniques to diagnose breast cancer from image-processed nuclear features of fine needle aspirates**, *Cancer Letters*, Vol. 77(2), 1994.

[13] Yan Xu, Tao Mo, Qiwei Feng, Peilin Zhong, Maode Lai, and Eric Chang, **Deep learning of feature representation with multiple instance learning for medical image analysis**, *39th IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014.