Quantifying Surgical Complexity through Textual Descriptions of Current Procedural Terminology Codes

Alexander Van Esbroeck, BS[1], Ilan Rubinfeld, MD, MBA[2], Zeeshan Syed, PhD[1]
[1]University of Michigan, Ann Arbor, MI; [2]Henry Ford Hospital, Detroit, MI

Abstract

Models for surgical complications are a requirement for evaluating patients by the bedside or for risk-adjusted quality and outcomes assessment of healthcare providers. Developing such models requires quantifying the complexities of surgical procedures. Existing approaches to quantify procedural complexity rely on coding system generalities or factors designed for reimbursement. These approaches measure complexity of surgical procedures through the time taken for the procedures or their correspondence to rough anatomical ranges. We address this limitation through a novel approach that provides a fine-grained estimate of individual procedural complexity by studying textual descriptions of current procedure terminology (CPT) codes associated with these procedures. We show that such an approach can provide superior assessment of procedural complexity when compared to currently used estimates. This text-based score can improve surgical risk adjustment even after accounting for a large array of patient factors, indicating the potential to improve quality assessment of surgical care providers.

Introduction

Quality and outcomes assessment is a burgeoning area in health care, acting as a tool for performance evaluation and a way to measure improvements resulting from changes in practices. At a high level, assessing quality and outcomes involves evaluating the rates of negative outcomes, such as mortality, morbidities, or readmission across healthcare providers. A major challenge in doing such an assessment, however, is that different healthcare providers have very different patient populations and case mixes. As a result, a direct comparison of the rates of mortality and morbidity across providers is inherently biased against providers with higher-risk patient populations, where higher rates of adverse clinical outcomes are largely a function of the kinds of patients being treated. In such a case, a more objective approach for assessing quality and outcomes adjusts for variations in the characteristics of patients and procedures across healthcare providers while comparing their performance.

Risk adjustment aims to factor in a patient's likelihood of negative outcomes independent of the quality of treatment received. A patient coming into the emergency department with a stroke is at a higher baseline risk of mortality than a patient with a broken arm, regardless of the admitting institution. For meaningful performance evaluation, it is necessary to adjust rates of negative outcomes for "level of difficulty", by comparing the occurrence of outcomes with our expectation given the patients treated. Determining this expectation of outcomes requires the development of statistical models that assess the likelihood of the outcome independent of provider. These models can incorporate a variety of patient factors, including demographics, comorbidities, and laboratory tests.

For an accurate expectation of clinical risk, information about patients must be supplemented with information about procedural factors. This is particularly true in surgical departments, where procedures play a critical role in the patient's outcomes. Different procedures carry different levels of risk associated with variations in their complexities: open-heart surgery carries a higher risk of mortality than a bone marrow biopsy regardless of where the operation is performed. Accurately estimating procedural complexity presents a number of challenges. This includes the question of finding a way to compactly characterize a fairly broad range of procedures, e.g., through features that are rich enough to reflect variations across different procedures while also incorporating information that is helpful for assessing complexity. Similarly, another challenge is the use of historical case records to derive empirically motivated estimates of procedural complexity that may be sparsely populated for certain procedures.

Most clinical registries presently retain information about procedures through either current procedural terminology (CPT) codes or their work relative value units (RVUs). The inclusion of CPT codes is largely motivated by reimbursement considerations, but in the absence of other approaches that compactly define the procedures undergoing by patients a variety of approaches have been applied to leverage CPT codes to estimate procedural complexity. These approaches make helpful but overly broad assumptions about the similarity of procedures based on their numbering in the CPT code system. While providing useful measures of complexity, these approaches fall short of the desired level of accuracy. Similarly, RVUs are typically a component of fee schedule, and do not directly take into account the complexity of surgical procedures.

We seek to address this deficiency by exploring a more robust approach to quantifying procedural complexity. Our approach focuses on leveraging short textual descriptions associated with procedures as a way of providing a compact but generalizable set of features that can be used to train models for procedural complexity on historical case records. Instead of relying on general assumptions about the coding system, we relate codes to one another based on their textual descriptions. Codes described with similar terms are expected to have similar complexity: for example biopsies, which vary greatly in their anatomic targets, are generally considered to be low risk. We propose an approach that automatically learns the relationship between words and procedural complexity from a surgical dataset. We hypothesize that the rich set of relationships that can be gleaned from textual CPT code descriptions can be leveraged to provide a more effective approach to complexity estimation.

Background

We start by discussing the challenges of estimating procedural complexity, and by reviewing existing approaches for this task, focusing especially on the shortcomings of existing methods. We then motivate the use of textual descriptions of procedures to address these issues.

There are several key difficulties in estimating procedural complexity. First, procedures lack a descriptive representation, making them challenging to analyze. Unlike patient factors, which contain a wide variety of continuous and categorical variables, procedural information is limited to the CPT code and its RVUs. These two variables were designed for reimbursement, and are not well-suited to quantifying complexity. For example, RVUs are largely a component of fee schedule and do not directly take into account the complexity of surgical procedures. Moreover, these work units abstract away from the underlying procedure and do not provide enough specificity for the procedure performed. Similarly, a critical challenge with the use of CPT codes for complexity assessment is their overwhelming number. There are well over 10,000 codes in the CPT code set, many of which are uncommon in practice. While it is feasible to directly estimate the complexity of more common procedures, estimating the complexity of less common codes is challenging due to the limited amount of available data.

There are a number of approaches that have been designed for determining procedural complexity using CPT codes[3,4]. These approaches avoid the estimation of complexity on a code by code basis, finding methods to improve the effective sample size and improve accuracy when estimating procedural complexity. Many of these approaches has focused on identifying categories of similar CPT codes to improve complexity estimates. The intuition behind these methods is that it is possible to increase the number of examples used for estimating the complexity of uncommon procedures by identifying related groups of codes. One such approach organizes the codes into 9 groups, based primarily on anatomy (e.g. abdominal procedures, endocrine procedures), allowing for a large number of incidents in each group considered. However, procedures related to an anatomical region may vary greatly in complexity. For example, while pancreaticoduodenectomy and inguinal hernia repair are both gastrointestinal operations, they are vastly different procedures. This coarse organization of codes provides little specificity in assessing complexity.

A more sophisticated system for code grouping by Raval et al. organized CPT codes into 135 different categories, designed specifically for risk-adjustment and providing a finer level of granularity than the anatomic ranges[3]. A downside to this categorization is that it requires manual organization of the codes, a process that is time and resource consuming to apply to alternative sets of codes, or to update as the CPT code standard changes.

Another method, by Syed et al. relied on the intuition that CPT codes, which range from 1 to 99,499, tend to exhibit local similarity[4]. For instance, codes 31621 through 31659 all correspond to variations of a bronchoscopy. Like the grouping methods, this approach (referred to as CPT-SVM) utilized patterns in the code numbering to improve complexity estimates. This fully automated approach learned a support vector machine (SVM) classifier with a radial basis kernel, which allowed the complexity for one code to be propagated to nearby codes, giving better estimates for rare codes by considering the complexity of more common codes nearby. This method permitted even finer-grained distinctions than the categorization methods. However, the method's reliance on the CPT code numbering system, which has no meaning in terms of complexity, does not allow it to factor in what the procedure numbers actually correspond to, limiting the utility of such a method for complexity estimation.

These methods all face a tradeoff between specificity and sample size. The use of larger groups (or a larger range for complexity propagation in the CPT-SVM) increases the amount of data used to estimate complexity, yet decreases the granularity of the resulting estimates. For instance, the method that groups codes into nine anatomic ranges may provide an accurate assessment of the overall complexity of abdominal procedures, but this estimate has little utility in predicting of the complexity of individual procedures.

As an alternative to estimating complexity on a code by code basis, RVUs have also been utilized for complexity assessment. RVUs are used by Medicare to determine reimbursement for physician's services, and were designed to reflect the amount of resources that go into a procedure, including the amount of work required of surgeons and staff. RVUs have been considered as a surrogate measure of procedural complexity because they are a predefined continuous measure available on a code by code basis. Their application to complexity assessment assumes that the amount of work needed and the complexity of a procedure are related. While there is sense behind this intuition, the amount of work required provides a poor estimate of the complexity of an operation, and the relationship between the two is not strong enough for RVUs to provide accurate complexity assessment.

Our proposed approach circumvents some challenges of code-by-code complexity estimation by using textual descriptions of the CPT codes to infer similarity between procedures. Descriptive text characterizes procedures, providing relevant features that can be used for data-driven modeling of procedural complexity. Instead of associating complexity with numerical ranges, the method estimates complexity related to individual words. The complexity of a procedure can be determined by considering the complexity related to each of the words in its description. This allows for an association between codes based on qualitative similarities of the actual operations performed, giving better estimates of complexity. These similarities in the textual descriptions of can associate rare procedures with common ones in an analogous way to how existing methods leverage CPT code numbering. In a basic case, it allows association among CPT codes that differ only in minor ways, such as the large number of CPT codes assigned to bronchoscopies. However it also allows for more complex associations, such as grouping CPT codes relating to biopsies, which are generally low complexity, and have a wide numerical range in the CPT code system (ex. 20200 denotes a muscle biopsy, while 85102 denotes a bone marrow biopsy). This provides the potential for much more sensitive and specific identification of procedural similarity than approaches relying on numerical ordering.

Estimates of procedural complexity contribute substantially to quality assessment for surgical departments. The American College of Surgeons (ACS) National Surgical Quality Improvement Program (NSQIP) was designed to address the need for outcome-based, risk-adjusted evaluation of surgical care quality[1,2]. NSQIP develops quality scores based on outcomes, adjusting for patient factors as well as procedural information. An improved method for quantifying procedural complexity would have a direct application in the improving quality assessments provided by NSQIP.

Methods

Data from multiple years of the NSQIP was used to train and evaluate models to estimate procedural complexity. NSQIP collects data on hundreds of thousands of patients from hundreds of institutions across the country, with records consisting of an array of patient factors, procedural information, and 30-day outcomes. The NSQIP data for 2005-2006, 2007, and 2008 was obtained under the data use agreement of the ACS and with the approval of the Institutional Review Board at the Henry Ford Hospital. The data sets consisted of patients undergoing general and vascular surgery: the 2005-2006 data consisted of 152,290 patients from 121 sites, the 2007 data of 211,407 patients from 186 sites, and the 2008 data of 271,368 cases from 211 sites. Each patient record included a primary CPT code and the corresponding RVUs, 30-day mortality, and a variety of preoperative variables, including demographics (age, sex, race), comorbidities, and laboratory tests. More details on the variables, outcomes, and data acquisition procedures have been described previously[5].

We associated each CPT code in the data with a short piece of text. These standard descriptions ranged in length from 1 to 5 words, denoting procedure names rather than a detailed description of the procedure.

Each CPT code was associated with a feature vector corresponding to the word frequencies for its short description. For example, code 20200 (muscle biopsy) would be represented with a 1 for muscle, a 1 for biopsy, and a 0 for all other words. We associate each patient with the word frequencies associated with their primary CPT code. We then treat the word frequencies for each patient as a set of predictive features, and use these to train a mortality prediction model. This model estimates the risk or complexity associated with individual terms, and the complexity associated with a code is defined as the sum of the complexities associated with the terms describing it. For a particular patient, the model takes as input the word frequencies associated with their primary CPT code, and outputs a complexity score reflecting the likelihood of negative outcomes for that individual.

Due to the large number of words in the descriptions, we employ support vector machines (SVM). SVMs learn a separating boundary between two classes of examples, in a way that tries to maximize the distance between the data points and the boundary. This property makes SVMs well-suited to dealing with high-dimensional data. With a large

number of features, there are countless boundaries capable of classifying the data points correctly. By maximizing the distance between the data and the boundary, an SVM can identify an optimal separation among this large range of possibilities. We used LIBSVM[6], a software package for learning SVM parameters, for the training of the model, using a linear kernel and default parameters. The model was trained only on data from 2005-2006.

The text-based model (referred to as Text SVM) was compared to several existing methods both on the 2005-2006 data used for training, as well as on data from 2007 and 2008 (which represents a clearly separated test set). In addition to considering different alternate approaches that use CPT variables, we also compared our work to the use of RVU to estimate procedural complexity. In particular, the RVU for a procedure was used as a predictive variable, with larger values corresponding to higher-complexity procedures.

For comparison to our Text SVM method, we considered two methods for translating CPTs into procedural complexity. The first (referred to as CPT Ranges) divided CPT codes into nine groups of anatomically related categories: integumentary and musculoskeletal procedures (CPT range 10000 to 29999), respiratory/hemic/lymphatic procedures (CPT range 30000 to 32999 and 38000 to 39999), cardiovascular procedures (CPT range 33001 to 34900), vascular procedures (CPT range 35001 to 37799), upper digestive tract procedures (CPT range 40000 to 43499), other digestive tract/abdominal procedures (CPT range 43500 to 49429 and 49650 to 49999), hernia repair procedures (CPT range 49491 to 49611), endocrine procedures (CPT range 60000 to 60999), and other including urinary and nervous system procedures (CPT range 50000 to 59999 and 61000 to 99999). The second approach (referred to as CPT-SVM) trained a radial basis kernel SVM on CPT codes, which related the complexities of CPT codes with similar numbers. The selection of parameters was set to be consistent with the work presented in Syed et al[4].

Evaluation of risk adjustment effectiveness was done using the area under the receiving operating characteristic curve (AUC). The AUC measures the discriminative ability of a continuous measure with respect to an outcome. We compared the AUC of the text-based model with the other approaches to assess whether the resulting model achieved better estimation. The method of DeLong was used to compute a p values to compare the ROC curves of the different methods[7].

We also used net reclassification improvement (NRI) and integrated discriminative improvement (IDI) to assess whether the addition of the text-based model improved prediction performance[8]. NRI and IDI both estimate a measure's predictive improvement compared to a baseline measure. NRI assesses the number of patients with events whose risk increases in the new model, as well as the number of patients without events whose risk decreases. IDI is similar, but considers the magnitude of the risk increases and decreases in its evaluations. These measures are growing in popularity for the evaluation of clinical predictors because they reflect the practical improvement of using one predictor over another, in terms of its effect on patient decision-making, better than the comparison of AUCs.

In addition to assessing the relative performance of the text-based model for procedural complexity, we evaluated the ability of this measure to improve risk adjustment when incorporating patient factors. This experiment demonstrates whether the additional predictive power of the text-based descriptions is related to preoperative variables, or if the new information is complementary to that of other factors. A multivariate logistic regression model was trained using stepwise backward elimination with the Wald statistic. The preoperative variables considered were: age, sex, race, height, weight, American Society of Anesthesiologists class, emergency status, preoperative functional status, diabetes, renal disease, dyspnea, ascites, chronic obstructive pulmonary disease, current pneumonia, ventilator dependence, chronic steroid use, bleeding disorders, heart failure, hypertension, coronary artery disease, peripheral vascular disease, disseminated cancer, weight loss, current chemotherapy or radiotherapy, neurologic deficit, preoperative transfusion, wound class, preoperative sepsis/septic shock/systemic inflammatory response syndrome, alcohol use, smoking, sodium, creatinine, albumin, bilirubin, aspartateaminotransferase, alkaline phosphatase, white blood cell count, hematocrit, platelet count, partial thromboplastin time, and international normalized ratio These variables were chosen to be consistent to prior work in the area[3,4].

Several versions of the model were trained. The first included all preoperative variables present in NSQIP, including laboratory tests, RVUs, and CPT Ranges. A second version included all features from the first model, in addition to the text-based model scores. This two-model comparison was repeated with the exclusion of laboratory tests, as a large percentage of patients in NSQIP had missing values for these tests.

All evaluation was conducted using SPSS and MATLAB.

Results

To compute procedural complexity estimates, only patients with recorded values for CPT code and RVUs were used in evaluating AUCs. Table 1 shows the resulting sample sizes and the AUCs of the methods evaluated. As expected, the text-based approach performs substantially better than the use of RVUs or CPT ranges. Additionally, the AUC values for the 2005-2006 data on which the text-based (Text SVM) and CPT-SVM models were fit, indicate significantly higher performance for the Text SVM when compared to the CPT-SVM ($p < 0.001$). This trend continued for the two testing data years, where again the Text SVM approach significantly outperformed the CPT-SVM ($p < 0.001$).

Year	*n*	*Text SVM*	*CPT-SVM*	*RVU*	*CPT Ranges*	*p value*
2005-2006	152418	0.871	0.822	0.657	0.636	<0.001
2007	211365	0.835	0.798	0.687	0.646	<0.001
2008	271250	0.838	0.802	0.683	0.661	<0.001

Table 1 AUC values for CPT code, CPT description, and RVU based predictors of mortality. The SVM models were trained on data from 2005-2006. The p value reflects the largest p-value when comparing the Text SVM to the other three models.

Table 2 compares the AUCs for multivariate models with (+Text) and without (-Text) the Text SVM scores for procedures included alongside patient variables. We note that the stepwise backward elimination process always retained the Text SVM scores, and therefore the -Text model was derived by starting the backward elimination process without including the Text SVM scores. The inclusion of preoperative patient variables, particularly the inclusion of laboratory tests, greatly reduced the number of records available for the analysis(since many patients did not have complete records for these additional preoperative variables). While the AUCs and R^2 values showed a marginal (but consistent) improvement with the addition of procedural complexity information, both the NRI and IDI showed substantial improvements in model discrimination and reclassification when the Text SVM information was factored into multivariate models.

Year	*n*	*AUC (-Text)*	*R^2*	*AUC (+Text)*	*R^2*	*NRI (p)*	*IDI (p)*
2005-2006	39597	0.915	0.380	0.923	0.400	0.599 (<0.001)	0.011 (<0.001)
2007	54230	0.919	0.374	0.920	0.378	0.317 (<0.001)	0.001 (0.070)
2008	68953	0.919	0.393	0.921	0.398	0.528 (<0.001)	0.003 (<0.001)

Table 2 Performance of stepwise backward elimination models including preoperative variables (including laboratory tests), RVUs, CPT ranges, and CPT-SVM scores, compared to a model that additionally included the text-based scores.

When laboratory tests were excluded from the set of patient factors, the sample sizes greatly increased (Table 3). In these cases the text-based score was again always retained in the backwards elimination process. Table 3 shows the performance of the models both with and without the text-based score. The text-based score again resulted in small but consistent improvements in the AUCs and R^2 values, as well as substantial improvements in the NRI and IDI.

Year	*n*	*AUC (-Text)*	*R^2*	*AUC (+Text)*	*R^2*	*NRI (p)*	*IDI (p)*
2005-2006	151377	0.937	0.402	0.943	0.419	0.630 (<0.001)	0.005 (<0.001)
2007	209554	0.938	0.392	0.939	0.400	0.482 (<0.001)	0.003 (<0.001)
2008	271240	0.942	0.407	0.945	0.415	0.534 (<0.001)	0.004 (<0.001)

Table 3 Performance of stepwise backward elimination models including preoperative variables (excluding laboratory tests), RVUs, CPT ranges, and CPT-SVM scores, compared to a model that additionally included the text-based scores.

Figure 1 Visualization of descriptor words displaying the complexity associated with each term in the data. The size of a code reflects the complexity estimated by the model in the dataset.

To better understand the derived risk model, we visualized the risks associated with individual words. Figure 1 depicts the most commonly occurring words in the dataset, with higher risk reflected by larger font size. The variation in color is for readability, and does not reflect properties of the words. Many of the low complexity terms are consistent with expectation, including a number of imaging and medication related terms. The larger terms reflect more complex procedures, like duodenotomy and tracheoplasty.

Term	Number of Codes	Example Code Descriptions
Repair	165	nerve repair, repair blood vessels, hernia repair
Excision	99	excision of tumor, excision of cyst, excision of lesion
Surgery	83	incisional surgery, excisional surgery, resectional surgery
Graft	70	excision of graft, skin graft, nerve graft
Bypass	65	gastric bypass, arterial bypass, bypass graft
Treatment	63	fracture treatment, treatment of cyst
Fracture	60	fracture treatment
Tumor	51	resection of tumor, excision of tumor, destruction tumor
Arthroscopy	39	hip arthroscopy, arthroscopy knee, ankle arthroscopy
Drainage	37	incision and drainage, drainage of abscess, drainage of cyst

Table 4 List of the 10 words occurring in the largest number of different CPT codes, with examples of the descriptive text for the matching codes.

On inspection of the associations between words and codes, it is evident that the use of text can link multiple procedures in the model by relevant characteristics. Table 4 lists the 10 words appearing in the largest number of

distinct CPT codes, and some examples of the code descriptions. These words show significant breadth in the variety of related procedures, while still quantifying a common property among the set.

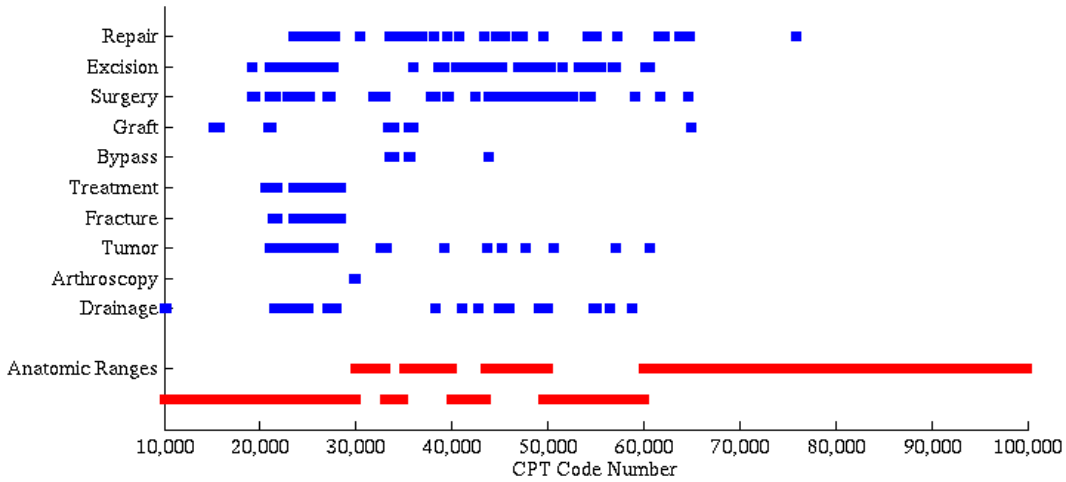

Figure 2 CPT code numbers associated with the 10 words spanning the most codes. A square for a word indicates a code in that numerical range. The bottom displays one of the groupings of CPT codes, which organize contiguous sets of codes based on anatomy.

To compare the connections made by the descriptive words with those made by numbering-based approaches, we also looked at the distribution of codes for these 10 words across the full CPT range. Figure 2 shows the placement of the CPT codes containing these 10 words across the CPT numbering spectrum, and compares these with the set of anatomic ranges used in the CPT ranges approach. The codes associated with the words span a broad numerical range, with many of the words represented across a large number of anatomic ranges. This demonstrates the ability of text to go beyond existing approaches and make novel connections between procedures.

Discussion

This study explores the use of descriptive text to improve procedural complexity estimates. By relying on textual descriptions rather than the raw numbering of surgical procedures, the relationships and similarities between surgical procedures can be established and leveraged to provide an accurate quantitative measurement of procedural complexity. In particular, textual descriptions provide a useful feature set for capturing variations between surgical procedures, while simultaneously providing a way to relate procedures and to jointly use information between procedures with similar text when estimating the complexity of the large number of uncommon procedures.

A comparison of the predictive power of a number of methods for estimating procedural complexity, using the AUC, showed a significant improvement in complexity estimation when using the text-based approach compared to existing alternatives. More importantly, a comparison of models supplementing preoperative patient variables with and without the inclusion of the text-based procedural complexity score indicated that our approach also improves overall risk adjustment. Our measured AUC values improved in all the years within ACS NSQIP that we considered, and the NRI and IDI values were highly significant. The ability to improve the overall model shows the immediate utility of this approach in quality assessment: this significant improvement to risk adjustment can benefit the evaluation of healthcare institution performance, for instance in the ACS NSQIP, while also potentially providing better patient care by the beside.

Descriptive text is readily available for CPT codes, and the approach leverages an existing resource to garner advances in complexity estimation. This descriptive text can be leveraged to associate complexity with certain aspects of a procedure, leading to a richer model that can identify the particular properties of procedures that contribute to their complexity. These benefits are not exclusive to CPT codes, and could be extended to other kinds of medical codes, such as ICD-9 diagnostic codes.

Categorization approaches are limited by their ability to associate each code with a single group. Different choices of categorization could relate based on anatomy or the type of procedure (e.g. hip arthroscopy), but because of the

discrete groupings a choice must be made as to which scheme is more useful. With a text-based approach it is possible to leverage multiple kinds of connections between procedures at once, providing greater associative power.

A key benefit of this machine learning approach is its automatic, data-driven nature. Methods that rely on careful grouping of procedures require much work on the part of experts to engineer a categorization system. Because CPT codes are added, removed, and modified on a yearly basis, these categories must be updated regularly. Alternative coding systems require this effort to be replicated again for each new standard used. By using an automatic approach like the one outlined in this paper, it is possible to achieve a surgical stratification model that is both accurate and easily maintained and extended.

There are a number of directions for future improvement of the presented methods. The use of longer, more detailed descriptions of procedures may provide more useful information to relate procedures and estimate complexity. Identifying prefixes and suffixes shared across terms may give greater associative power than using individual word frequencies. The use of a cleaner set of descriptive text, which includes more consistent abbreviation and avoids concatenation of words would allow the use of more inter-code relationships.

There have been concerns raised about the significance of the AUC in terms of its ability to calibrate clinical predictive models, and the difficulty of evaluating the significance of an improvement[9]. The use of NRI and IDI helps address these concerns by evaluating the metrics in terms of potential real-world benefits, however the improvement of methods for evaluating clinical risk factors is an active area of research[10].

Conclusions

We explored an approach that leveraged descriptive text for procedures in order to provide better estimates of procedural complexity. A model generated using a complexity score based on this descriptive text achieved higher accuracy in risk assessment than existing methods. The model significantly improved overall risk assessment according to a variety of metrics, even after accounting for a wide variety of preoperative variables. Text-based models can lead to better estimates of quality for healthcare institutions.

Acknowledgments

This material is based upon work supported by the National Science Foundation.

References

Khuri SF, Daley J, Henderson W, Hur K, Demakis J, Aust JB, Chong V, Fabri PJ, Gibbs JO, Grover F, Hammermeister K, Irvin G, McDonald G, Passaro E, Phillips L, Scamman F, Spencer J, Stremple JF. The Department of Veterans Affairs' NSQIP: the first national, validated, outcome-based, risk-adjusted, and peer-controlled program for the measurement and enhancement of the quality of surgical care. Ann Surg 1998;228:491–507.

Khuri SF. The NSQIP: a new frontier in surgery. Surgery 2005;138:837–843.

Raval MV, Cohen ME, Ingraham AM, Dimick JB, Osborne NH, Hamilton BH, Ko CY, Hall BL. Improving American College of Surgeons National Surgical Quality Improvement Program risk adjustment: incorporation of a novel procedure risk score. J Am Coll Surg 2010;211:715–723.

Syed Z, Rubinfeld I, Patton P, Ritz J, Jordan J, Doud A, Velanovich V. Using diagnostic codes for risk adjustment: A non-parametric learning approach. J Am Coll Surg 2010;211:S99--S100

American College of Surgeons. American College of Surgeons National Surgical Quality Improvement Program participant user guide. Available at: http://acsnsqip.org. Accessed March 10, 2012.

Chang CC, Lin CJ. LIBSVM: a library for support vector machines. ACM Trans Intel Sys Tech 2011;2:1--27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44:837–845.

Pencina MJ, D'AgostinoSr RB, D'AgostinoJr RB, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Statistics in Medicine 2008;27(2):157-172.

Merkow RP, Bilimoria KY, Hall BL. Interpretation of the c-statistic in the context of ACS-NSQIP models. Ann Surg Oncol 2010;18:1-1.

McGeechan K, Macaskill P, Irwig L, Liew G, Wong TY. Assessing new biomarkers and predictive models for use in clinical practice: a clinician's guide. Arch Intern Med. 2008;168(21):2304-2310.