Predicting Surgical Risk: How Much Data is Enough?

Ilan Rubinfeld, MD[1], Maria Farooq, MBBS[1], Vic Velanovich, MD[1], Zeeshan Syed, PhD[2]
[1] Henry Ford Health System, Detroit, MI; [2] University of Michigan, Ann Arbor, MI

Abstract

As medicine becomes increasingly data driven, caregivers are required to collect and analyze an increasingly copious volume of patient data. Although methods for studying these data have recently evolved, the collection of clinically validated data remains cumbersome. We explored how to reduce the amount of data needed to risk stratify patients. We focused our investigation on patient data from the National Surgical Quality Improvement Program (NSQIP) to study how the accuracy of predictive models may be affected by changing the number of variables, the categories of variables, and the times at which these variables were collected. By examining the implications of creating predictive models based on the entire variable set in NSQIP and smaller selected variable groups, our results show that using far fewer variables than traditionally done can lead to similar predictive accuracy.

Introduction

The sheer volume of information that must be collected and analyzed for each patient poses a serious challenge to health care professionals. Caregivers often have to keep track of a variety of patient data, including demographics, comorbidities, laboratory values, imaging results, parameters for continuous monitoring, and a record of interventions. While recent advances allow these variables to be combined, for many clinical applications, into models with excellent risk adjustment and prediction, the collection processes for these variables are associated with increased health care costs as well as impose demands on both caregivers and patients.

We investigated the effect of reducing the amount of data that needs to be collected to build accurate models for risk stratification. We focused our investigation on surgical patients, seeking answers to three questions using data from the National Surgical Quality Improvement Program (NSQIP)[1]:

- To what extent can the number of risk variables be decreased without compromising on predictive accuracy?
- How much information do laboratory results add to predictive models constructed from detailed demographics and comorbidity data?
- How does the performance of predictive models change when they are developed on data collected from earlier years?

Our work is motivated by the observation of Birkmeyer et al.[2] that previous studies on NSQIP typically flatten in terms of accuracy with roughly 10 variables. This is primarily because the variables are too rare, internally correlated, or not sufficiently explanatory to contribute meaningfully for risk adjustment or prediction.

While we address a slightly different problem, i.e., predicting surgical risk (where models trained on historical data are used to predict new data) as opposed to outcomes analysis (where models trained on historical data are used to identify high or low outliers in those data), we examine the observation of Birkmeyer et al more formally, and explore how the amount of data needed to model surgical risk can be decreased.

The Clinical Context of Data

The ability to translate clinical data into acuity models is relevant for both the goal of predicting patient risk and outcomes analysis. However, this process depends critically on the presence of validated data that can be used to train models through inductive inference and to apply these models to individual patients in the future.

We note, for example, that the collection of audited and verified data is both one of the main strengths of NSQIP and one of its major challenges. NSQIP has developed into the leading standard of acuity-based research for surgery. Each hospital participating in NSQIP is required to undergo audit visits and meet stringent guidelines that represent some of the highest in the industry; over 90% for inter-rater reliability (IRR)[3]. Data is gathered by nurse data-coordinators and submitted centrally. This verified data is then utilized to construct risk models. NSQIP creates observed-to-expected ratios (O/E) for adverse events for each institution to allow comparison between institutions and to assess progress.

While the kinds of data collected by NSQIP are a valuable resource to develop acuity models with high predictive accuracy, and can be generalized to the set of variables that can potentially be used as a basis for patient care, they are resource-intensive to maintain. These data generally require nursing abstractors and

custom information technology links to assist with abstracting data for verification from administrative sources different at each institution. The data can be categorized by source and type. Each category of data requires a different interface and method for collection, with variable impact on nursing or registrar abstractor time. The NSQIP sampling methodology utilizes laboratory data, which at most institutions are gathered manually and entered into the computerized system. Some institutions have developed direct interfaces at great expense. In the past four years of NSQIP acuity adjusted data (2005-2008), only serum albumin has been utilized consistently as a continuous variable in the acuity adjustment models[4]. Other laboratory results appear as categorized variables intermittently (for example, creatinine >1.2, or serum glutamic oxaloacetic transaminase [SGOT] >40). In the setting of scarce resources with intense demands for efficiency, it is often not worthwhile to collect all categories of data. Understanding the value of each category may lead to improved efficiency while preserving accuracy.

We note that the issue of collecting data is relevant both for forming the initial models (where a large number of variables need to be collected from many patients) and the subsequent application of these models to patients. In this context, the demands of data collection are both resource-intensive and persistent. We explore how these demands may be reduced.

Methods

Data: Inputs, Endpoints and Patient Population

Our study used NSQIP Participant Use File (PUF) data from 2005-2008. This data was sampled at over 200 hospital sites and contains 240 HIPAA-compliant variables, including demographics, surgical profile, preoperative risk factors, intraoperative and postoperative variables, and 30-day mortality and morbidity outcomes for patients undergoing major surgical procedures in both the inpatient and outpatient settings. Data was obtained under the NSQIP data use agreement. This work was approved by the Institutional Review Board of the Henry Ford Health System.

We focused on two specific outcomes for our study: death and morbidity (i.e., one or more morbidity outcomes) within 30 days following surgery. In developing models to predict these adverse outcomes, we focused on 86 variables collected before the start of surgery[5].

Table 1 presents the number of patients in the NSQIP PUF for each year. To maintain consistency in our experiments, we restricted our analysis to patients who did not have any missing values for any of the variables. This decision ensured that the patient population did not change when comparing the accuracy of risk models trained using different variables. It also prevented risk variables suffering from the bias introduced by methods to fill missing values. We also monitored the missing data based on category and detailed the types of missing data.

Year	Total Number of Patients	Patients without Missing Data
2005	33,930	0
2006	118,560	11,323
2007	211,407	33,259
2008	271,368	42,752

Table 1: Number of patients in NSQIP from 2005-2008. Also shown are the numbers of patients without missing values for any of the variables.

Developing Risk Models

For both mortality and morbidity, we used logistic regression to develop models to predict 30-day risk.

We developed two initial mortality and morbidity models using the 86 variables collected before the start of the surgery. These models were constructed on 2007 NSQIP PUF patients who did not have any missing values. The models were compared to: (1) models constructed on the same patients using the first 1, 5 and 10 variables found using forward selection based on the Wald statistic; (2) models constructed on the same patients using the 53 variables excluding laboratory results; and (3) models constructed on patients without any missing values from the 2005-2006 NSQIP PUF using all 86 variables.

We compared models based on how well they predicted outcomes on patients in the 2008 NSQIP PUF who did not have any missing values. The comparisons were carried out by measuring the area under the receiver operating characteristic curve (AUROC) and the associated standard error for each model, and comparing the AUROC using an unpaired *t*-test.

To analyze category and type, we focused on the 2007 PUF data. We generated logistic regression models similar to NSQIP annual report models. These utilized a mixture of laboratory, demographics and comorbidity related data points.

Results

Models with Fewer Variables

Table 2 presents the results of comparing predictive models constructed on 2007 NSQIP PUF using the first 1, 5 and 10 variables found using forward selection based on the Wald statistic to models constructed using all 86 variables collected before the

Model	Mortality AUROC	P-Value	Morbidity AUROC	P-Value
All Variables	0.907	Ref	0.767	Ref
10 Variables	0.902	0.889	0.759	0.524
5 Variable	0.889	0.642	0.750	0.184
1 Variable	0.689	<0.001	0.654	<0.001

Table 2: Comparison of AUROC for models constructed with 1, 5 and 10 variables with models constructed using all variables.

Model	Mortality AUROC	P-Value	Morbidity AUROC	P-Value
All Variables	0.907	Ref	0.767	Ref
No Laboratory Variables	0.896	0.769	0.761	0.636

Table 3: Comparison of AUROC for models constructed without laboratory results with models constructed using all variables.

Model	Mortality AUROC	P-Value	Morbidity AUROC	P-Value
All Variables (2007)	0.907	Ref	0.767	Ref
All Variables (2006)	0.895	0.759	0.763	0.780

Table 4: Comparison of AUROC for models constructed using 2007 NSQIP PUF and models constructed using 2006 NSQIP PUF.

start of surgery. The first 10 variables found using this approach for mortality corresponded to (in order): functional health status prior to surgery, ASA classification, preoperative serum albumin, age, presence of disseminated cancer, preoperative BUN, DNR status, emergent vs. non-emergent case, work relative value unit, and presence of ascites. The first 10 variables for morbidity were: ASA classification, work relative value unit, preoperative albumin, emergent vs. non-emergent case, functional status prior to surgery, inpatient vs. outpatient case, preoperative systemic sepsis, age, steroid use for chronic condition, and weight.

Our results show while the predictive accuracy of the models decreased as fewer variables were used, the change in the AUROC values for mortality and morbidity was not statistically significant for both the 5 and 10 variable models (Table 2).

Models without Laboratory Results

Table 3 compares models for mortality and morbidity on 2007 NSQIP PUF constructed using all variables with models that were constructed on the same patient population without using laboratory results. While the AUROC for both mortality and morbidity prediction decreased slightly when laboratory data were excluded, these changes were not statistically significant in this patient population.

Models Trained on Older Data

Table 4 shows the change in AUROC when older data (i.e., from the 2005-2006 NSQIP PUF) was used for training predictive models instead of the data from the 2007 NSQIP PUF. The models were consistently evaluated on patient outcomes in the 2008 NSQIP PUF. Since all of the 2005 patients were missing data for one or more variables, this corresponded to exclusively constructing a model from the 2006 patients. As was the case for the experiment with laboratory results, using data from previous years resulted in a marginally lower AUROC. This change was not significant for either mortality or morbidity.

Missing Data

The NSQIP patient sample size grew over the years covered by our study (i.e., 2005 to 2008) as did the number of hospitals contributing data. Only 13.7% of patients over all four years had zero missing data (Table 1). When categorizing type of data missing for the year 2007, which was used to develop most of the models in our study, over 74% of the patients were

Laboratory Result	# Patients Missing Data	% Patients Missing Data
Hematocrit	29,278	14%
White Blood Cells	33,823	16%
Platelets	33,834	16%
Creatinine	40,321	19%
Sodium	41,318	20%
Blood Urea Nitrogen	44,124	21%
Serum Glutamic Oxaloacetic Transaminase	91,895	44%
Alkaline Phosphate	92,708	44%
Bilirubin	94,188	45%
Albumin	96,260	46%
Prothrombin Time	90,578	57%
Partial Thromboplastin Time	86,311	59%

Table 5: Breakdown of missing laboratory results in the 2007 NSQIP PUF.

missing laboratory results. In contrast, less than 22% of the patients were missing comorbidity or demographic data.

Table 5 gives the detail by individual type of laboratory data for the year 2007. The average missing rate was 35%. Albumin, generally considered one of the most useful laboratory values to predict adverse outcomes, was missing over 46% of the time.

Discussion

The practice of medicine is becoming an increasingly data-driven process. Caregivers are required to collect and analyze a large number of variables across many different categories for each patient. Even though methods for studying this data have evolved in recent years, collecting clinical validated data is a cumbersome process. This has resulted in models for acuity adjustment and risk stratification requiring a resource-intensive methodology to create and maintain.

As we study the kinds of data we collect presently, and compare the relative predictive powers of different risk variables in the context of outcomes and risk stratification, opportunities for efficiency and simplification may emerge. In this work we have discussed the issue of how related categories of data only incrementally increase the accuracy of existing risk stratification models. A long-held dictum of the management community has been to seek efficiencies through the 80:20 rule or Pareto principle[6]. For example, it is typical that 80% of the commissions at a brokerage are generated by 20% of the brokers. This principle may be true in health care settings as well, where 80% of the risk related to the patient may be modeled using 20% of the data, or maybe 20% of the effort to gather the data. In this sense the incremental benefit of each new variable or new category of variables, each with its own gathering and verification system, can be viewed critically.

In our investigation, we explored questions related to how the accuracy of predictive models is affected by changing the number of variables, the categories of variables, and the times at which these variables were collected. Our results on the NSQIP dataset show that models to predict adverse surgical outcomes can be constructed using fewer variables, with reduced dependence on laboratory results, and potentially using data that is not recorded in the period immediately preceding model training, while still achieving accuracy similar to a more data-intensive approach.

Our findings motivate the creation of acuity models that can be constructed and applied in an affordable and time-efficient manner with low complexity. For example, we note that laboratory results such as albumin levels have been consistently important in the NSQIP dataset while creating models of patient

risk. Yet to obtain this one laboratory value at the typical institution would require either a laboratory interface or a separate method for lookup clinically. While albumin has been proven to be valuable in risk stratification, it may be possible to construct predictive models without it that have similar accuracy yet eliminate a level of complexity in the pursuit of quality data that is easier to obtain. In particular, our results on the NSQIP dataset show that patient demographics and clinical characteristics, which can be easily obtained from patient histories and physical exams, contain a wealth of information that can be exploited to reduce dependence on variables that are more invasive and expensive to measure.

Constructing data from easier to collect variables can make the use of acuity models more widespread. Our findings regarding missing data were initially presented as methodologic logistics in an attempt to be fully transparent with data limitations and the characteristics of our study. These findings evolved into further evidence of the difficulties in obtaining reliable data. Even in the NSQIP context with outside audit, dedicated nursing data abstraction and an IRR method, the number of complete datasets is limited. The majority of analytic methods falter in the face of missing data, and methods to extrapolate and recreate these missing values fall short of expectations.

Health care institutions, regulatory and reporting agencies, payers, and even patients increasingly require transparency and reliable outcomes data, preferably with acuity adjustment in a validated way. Health care expenditure is rising rapidly and quality projects struggle to justify expenditure in data gathering and capture. Each variable and additional type or category of variable comes at a greater expense stressing ever-tightening budgets and resources. Methods to improve efficiency without sacrificing accuracy are essential to the continued growth of the quality and outcomes movement in health care.

We conclude by noting that while our study focused exclusively on the NSQIP dataset, the results of our work may extend more broadly to other datasets and clinical disciplines. We also believe that our research on addressing the challenges of collecting data for risk stratification (i.e., time and financial costs, need for invasive tests to measure some parameters) may have further relevance in addressing the burden of cognitive overload. Reducing the number of variables needed to predict patient risk creates the opportunity to identify core data elements that can be compactly presented to caregivers for decision-making. Further research is needed to study both the human factors associated with successfully implementing this approach, and to evaluate the potential impact of such work.

Acknowledgment

The authors would like to thank Sarah Whitehouse for helping improve the presentation of this work.

References

1. Khuri SF. The NSQIP: a new frontier in surgery. *Surgery*. 2005;138:837–843.
2. Birkmeyer JD, Shahian DM, Dimick JB, Rinlayson SRG, Flum DR, Ko CY, and Hall BL. Blueprint for a new American College of Surgeons: National Surgical Quality Improvement Program. *American College of Surgeons*. 2008;207:777–782.
3. Shiloach M, Frencher SK, Steeger JE, Rowell KS, Bartzokis K, Tomeh MG, Richards KE, Ko CY, Hall BL. Towards robust information: data quality and inter-rater reliability in the American College of Surgeons National Surgical Quality Improvement Program. *JACS*. 2010;210:6–16.
4. NSQIP semi-annual reports, 2006-2008.
5. ACS NSQIP. *User Guide for the 2008 Participant Use Data File*, July 2009 report.
6. Koch R. The 80/20 Principle, *Nicholas Brealy Publishers,* London, 1997.

NSQIP Disclosure

The American College of Surgeons National Surgical Quality Improvement Program and its participating hospitals are the source of the data used herein; they have not verified and are not responsible for the statistical validity of the data analysis or the conclusions derived by the authors.