ClinicoEconomics and Outcomes Research

Dovepress

open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

Surgical Complication Risk Factor Identification Using High-Dimensional Hospital Data: An Illustrative Example in Hemostasis-Related Complications

Stephen Johnston[1], Aakash Jha[2], Sanjoy Roy[3], Esther Pollack[3]

[1]MedTech Epidemiology and Real-World Data Sciences; Johnson & Johnson, New Brunswick, NJ, USA; [2]Decision Science, Mu Sigma, Bangalore, India; [3]Ethicon, Cincinnati, OH, USA

Correspondence: Stephen Johnston, MedTech Epidemiology and Real-World Data Sciences, Johnson & Johnson, 410 George Street, New Brunswick, NJ, 08901, USA, Tel +1-443-254-2222, Email sjohn147@its.jnj.com

Purpose: To describe an approach wherein high-dimensional hospital data can be used to identify generalizable risk factors for surgical complications for which there may be limited prior knowledge, as illustrated in the context of hemostasis-related complications (HRC).

Patients and Methods: This was a retrospective study of the Premier Healthcare Database. Patients included for the study underwent video-assisted thoracoscopic lobectomy (VATL), laparoscopic right colectomy (LRC), or laparoscopic sleeve gastrectomy (LSG) on an inpatient setting between Oct-2015 and Feb-2020 (first = index). The outcome, HRC, comprised hemorrhage, control of bleeding, and acute posthemorrhagic anemia. For each cohort, a high-dimensional dataset (ie, comprising 1000s of candidate risk factors) was constructed using taxonomies from the Clinical Classification Software Refined (CCSR). Candidate risk factors were fed into logistic regression models with a 70%/30% train/test split for each cohort; clinically plausible risk factors that were consistently significant predictors of HRC across the 3 training models were then used in a final parsimonious model including sex, age, race, and payor; finally, the parsimonious model was applied to the test data to compare predicted risk with observed incidence of HRSC.

Results: The study included 11,141 VATL, 20,156 LRC, and 121,547 LSG patients, in whom 7.5%, 7.8%, and 1.2% experienced HRSC, respectively. Ultimately, 6 clinically plausible CCSR categories were identified as being statistically significant predictors across all 3 cohorts (eg, coagulation and hemorrhagic disorders, malnutrition, alcohol-related disorders, among others). In the parsimonious model applied to the test data, the observed incidence of HRSC was substantially higher in the top quintile vs bottom quintile of predicted risk: LSG 2.05% vs 0.53%, LRC 13.30% vs 4.11%, VATS 12.49% vs 5.04%.

Conclusion: High-dimensional real-world data can be useful to identify risk factors for outcomes that generalize across multiple cohorts. The risk factors identified herein should be considered for inclusion in future studies of hemostasis-related complications.

Keywords: hemostasis, high-dimensional, real-world data, surgery, complications

Introduction

Surgical complications are a common outcome of interest in surgical epidemiology, particularly when comparing across alternative medical devices, surgical approaches, or specific procedures. Such complications, however, are often multifactorial in their causes and may be difficult to predict.

Over the past decade, there has been a dramatic increase in the use of real-world data sources – such as electronic health records, hospital billing data, or administrative insurance claims data – to conduct research in the area of surgical epidemiology. Concurrently, advances have been made in the use of "high-dimensional" data techniques, which use the large amounts of clinical data in real-world databases to construct datasets with 1000s to >10,000s of variables. These data have in turn been incorporated into propensity score approaches, such as high-dimensional propensity scores and

Received: 25 June 2022
Accepted: 18 October 2022
Published: 5 November 2022

large-scale propensity matching, as well as disease risk scores.[1–3] These techniques all involve forms of prediction to generate a patient-specific estimated probability of either receiving a given treatment (propensity score) or experiencing a given outcome (disease risk score), typically through the use of logistic regression.

However, when applied in an individual research study, a given propensity or disease risk score's model specification is not typically intended for future application in a patient-level predictive analytic setting for practical clinical decision-making. Accordingly, in the setting of a single study, dimensionality/shrinkage methods are often used to reduce high-dimensional data to a more circumscribed set of covariates that is particularly well suited for the specific cohort under study through methods such as principal component analysis, regularization, and other forms of covariate prioritization.[4,5] However, both traditional (non-high-dimensional) and high-dimensional methods may be susceptible to overfitting, resulting in a list of covariates that may be most appropriate for the internal validity of the study at hand but perhaps with suboptimal generalizability more broadly.

Herein, we describe an approach wherein high-dimensional hospital data can be used to identify generalizable risk factors for surgical complications that may be considered for multiple future studies. Taking an approach in which we favor clinical face validity of risk factors over "black box" predictive analytics, we specifically search for risk factors for hemostasis-related complications, a difficult-to-predict outcome for which there is limited prior knowledge.

Materials and Methods

Data Source

We extracted the study data from the Premier Healthcare Database® (PHD), which is a population-based hospital database that contains clinical and administrative data routinely contributed by over 700 US hospitals that are members of the Premier healthcare performance improvement alliance, representing approximately 25% of annual US inpatient discharges.[6] This database includes discharge-level information on patient demographics, diagnoses, procedures, medical supplies, costs, and hospital and provider characteristics. This information is provided in the form of standardized administrative fields, hospital charge master data, and International Classification of Diseases, 10th Revision, Clinical Modification and Procedure Classification System (ICD-10-CM/ICD-10-PCS) diagnosis and procedure codes. The PHD has been widely used for epidemiologic research, forming the basis of over 600 peer-reviewed publications since 2006.

We conducted this study under an exemption from Institutional Review Board oversight for US-based studies using de-identified healthcare records, as dictated by Title 45 Code of Federal Regulations (45 CFR 46.101(b)(4)).

Patient Selection

Using the PHD, we selected patients undergoing one of the three minimally invasive surgeries in which hemostasis-related complications have been studied as an outcome of particular interest: Video-Assisted Thoracoscopic Lobectomy (VATL), Laparoscopic Right Colectomy (LRC), or Laparoscopic Sleeve Gastrectomy (LSG). Eligible patients had an elective inpatient encounter carrying an ICD-10-PCS code for one of these procedures in the primary procedure position between Oct-2015 and Feb-2020 (first = index). To increase the homogeneity of the groups with respect to surgical approach, patients were excluded if they had a hospital charge master record, ICD-10-PCS, or Current Procedural Terminology code indicating the use of robotic assistance. Patients undergoing LSG were also required to have an ICD-10-CM diagnosis code related to obesity in the primary diagnosis position or a Medicare Severity Diagnosis Related Group related to obesity.

Measurement of High-Dimensional Dataset of Candidate Risk Factors

We focused our search for candidate risk factors for hemostasis-related complications solely on ICD-10-CM diagnosis codes that were designated as "present on admission" during index. That is, any diagnoses that were designated as "not present on admission" would represent acute circumstances that developed during index and would be inappropriate for inclusion in most, though not all, propensity or disease risk score applications; furthermore, such diagnoses could represent sequelae of hemostasis-related or other complications and therefore simply be the subject of reverse causation.

For each of the three surgical groups, we began with a high-dimensional dataset of candidate risk factors comprising all eligible ICD-10-CM diagnosis codes recorded during index; this resulted in >5000 unique diagnosis codes across the surgical groups, many that were conceptually similar (eg, in the ICD-10-CM, type 2 diabetes [E11] has over 80 sub-codes). Therefore, to reduce the dimensionality of the data and improve the clinical interpretability and relevance of the individual features, we employed the Agency for Healthcare Research and Quality/Healthcare Cost and Utilization Project's Clinical Classifications Software Refined (CCSR). The CCSR aggregates more than 70,000 ICD-10-CM diagnosis codes into over 530 clinically meaningful categories.[7] Additional covariates included patient age, sex, race, and payer type (a correlate with socioeconomic status). Laboratory result, biometric, imaging, and other clinical data beyond diagnoses were either not available or only available for a subset of patients and therefore not used for this study.

Measurement of Hemostasis-Related Complications

The outcome of interest was intra- or post-operative hemostasis-related complications, defined as a composite of the following: intraoperative hemorrhage, intraoperative hematoma, postprocedural hemorrhage, postprocedural hematoma, postprocedural seroma, melena (LSG and LRC only), gastrointestinal hemorrhage (LSG and LRC only), hemorrhage not elsewhere classified, acute posthemorrhagic anemia, and control of bleeding. Control of bleeding was based on ICD-10-PCS procedure codes recorded during the index admission; all other components were identified via ICD-10-CM diagnosis codes that were not designated as "present on admission" during index recorded during the index admission.

Statistical Analyses

With an initial list of >500 CCSR categories, we further reduced the dimensionality of the dataset by retaining only those candidate risk factors that were present in at least 0.5% of each surgical group, resulting in the following numbers of candidate risk factors for the surgical groups: 122 for VATL, 121 for LRC, and 84 for LSG. Next, separately for each surgical group, we entered the candidate risk factors into a multivariable logistic regression model with a 70%/30% train/test split. Risk factors that were consistently statistically significant predictors of hemostasis-related complications across the 3 surgical groups in the test data at a P-value ≤ 0.05 were evaluated by a group of surgeons external to the study team to determine clinical face validity/plausibility, to ensure the selection of clinically relevant predictors. Finally, the risk factors that passed screening for clinical face validity were included in a parsimonious model including patients' sex, age, race, and payor type, which was used to quantify the collective ability of the identified risk factors to stratify patients by observed risk of hemostasis-related complications. As the purpose of this exercise was not purely to develop a predictive analytic model per-se, but rather to identify specific candidate risk factors, we did not use model-oriented measures of discriminative accuracy such as the area under the receiver operating curve for the parsimonious model.

Results

The study included 11,141 VATL, 20,156 LRC, and 121,547 LSG patients, in whom 7.5%, 7.8%, and 1.2% experienced HRSC, respectively. Ultimately, 6 clinically plausible CCSR categories were identified as being statistically significant ($P<0.05$) risk factors across all 3 surgical groups: coagulation/hemorrhagic disorders (CCSR code BLD006), cardiac dysrhythmias (CCSR code CIR017), gastrointestinal disorders (CCSR code DIG025), malnutrition (CCSR code END008), fluid/electrolyte disorders (CCSR code END011), and alcohol-related disorders (CCSR code MBD017). The gastrointestinal disorders grouping was driven primarily by diagnoses for peritoneal adhesions. Table 1 shows patient demographics and the prevalence of the 6 selected risk factors, stratified by those with vs without hemostasis-related complications during index.

Figure 1 shows the observed risk of hemostasis-related complications in each group by quintile of predicted risk based on the parsimonious model, thereby visually depicting the collective ability of the 6 selected risk factors (plus demographics) to stratify patients by risk. The observed risk of hemostasis-related complications was 2.2 to 3 times higher in the top quintile vs bottom quintile of predicted risk, illustrating the merit of these risk factors for consideration in future propensity or disease risk score models. Notably, the observed risk of hemostasis-related complications increased monotonically with increasing quintiles of predicted risk in the LRC and LSG surgical groups; however, it was uniform in quintiles 1–4 with a strong spike in quintile 5 for VATL.

Table 1 Patient Characteristics and Final Selected Risk Factors for Hemostasis-Related Complications

	VATL		LRC		LSG	
	Yes HRC	No HRC	Yes HRC	No HRC	Yes HRC	No HRC
	834	10,307	1576	18,580	1410	120,137
Female, %	59.1%	56.9%	57.3%	54.5%	78.7%	79.6%
Age, mean (SD)	67.0 (9.6)	67.8 (10.0)	65.5 (13.2)	69.0 (12.7)	47.3 (11.8)	43.9 (12.1)
Race, %						
Black	7.3%	7.4%	10.4%	11.2%	23.9%	18.4%
White	85.2%	84.5%	80.4%	79.1%	62.6%	66.9%
Other	5.6%	6.7%	7.9%	7.9%	11.8%	12.4%
Unknown	1.8%	1.4%	1.3%	1.8%	1.7%	2.3%
Payer, %						
Commercial	21.7%	25.2%	22.8%	34.9%	52.6%	59.6%
Medicare	69.5%	65.1%	69.7%	56.4%	22.7%	13.5%
Medicaid	5.2%	6.6%	4.4%	5.0%	19.4%	18.9%
Other	3.6%	3.1%	3.0%	3.7%	5.3%	8.0%
Final selected risk factors for HRC, %						
Coagulation/hemorrhagic disorders (BLD006)	3.5%	1.8%	3.0%	1.4%	2.1%	0.9%
Cardiac dysrhythmias (CIR017)	15.2%	9.7%	16.1%	8.7%	4.6%	2.4%
Gastrointestinal disorders (DIG025)	6.5%	4.5%	17.8%	13.5%	13.0%	8.7%
Malnutrition (END008)*	5.4%	1.3%	3.7%	1.9%	0.3%	0.02%
Fluid/electrolyte disorders (END011)	7.1%	3.7%	8.4%	4.2%	5.2%	1.2%
Alcohol-related disorders (MBD017)	4.0%	1.8%	1.8%	0.9%	0.3%	0.1%

Notes: The codes in parentheses next to each risk factor are CCSR codes; *Though END008 was present in <0.5% of the LSG sample, it was present in >0.5% of the total sample and thus not eliminated via heuristic feature selection.

Abbreviations: CCSR, Clinical Classification Software-Refined; HRSC, Hemostasis-Related Complications; LRC, Laparoscopic Right Colectomy; LSG, Laparoscopic Sleeve Gastrectomy; VATL, Video-Assisted Thoracoscopic Lobectomy.

Discussion

We used multivariable analysis applied to high-dimensional data to search for risk factors for hemostasis-related complications, a difficult-to-predict outcome for which there is limited prior literature on risk factors. We identified 6 risk factors with clinical face validity that we will consider in our future propensity and disease risk score models for studies in which hemostasis-related complications are an outcome.

With respect to the risk factors, coagulation/hemorrhagic disorders and alcohol-related disorders make intuitive sense as having an empirical association with increased risk of hemostasis-related complications. We hypothesize that the primary underlying mechanism by which cardiac dysrhythmias were associated with hemostasis-related complications is pharmacotherapy with long-term oral anticoagulants to decrease the risk of stroke. As the Premier Healthcare Database does not have information on chronic medications, rather only those that are administered within the hospital, future research in other databases is needed to confirm our hypothesis. It is possible that malnutrition and fluid/electrolyte disorders may be associated with compromised tissue quality, leading to an increased risk of hemorrhage, but more so these conditions are very likely to increase the risk of acute post-hemorrhagic anemia, a component of our definition of hemostasis-related

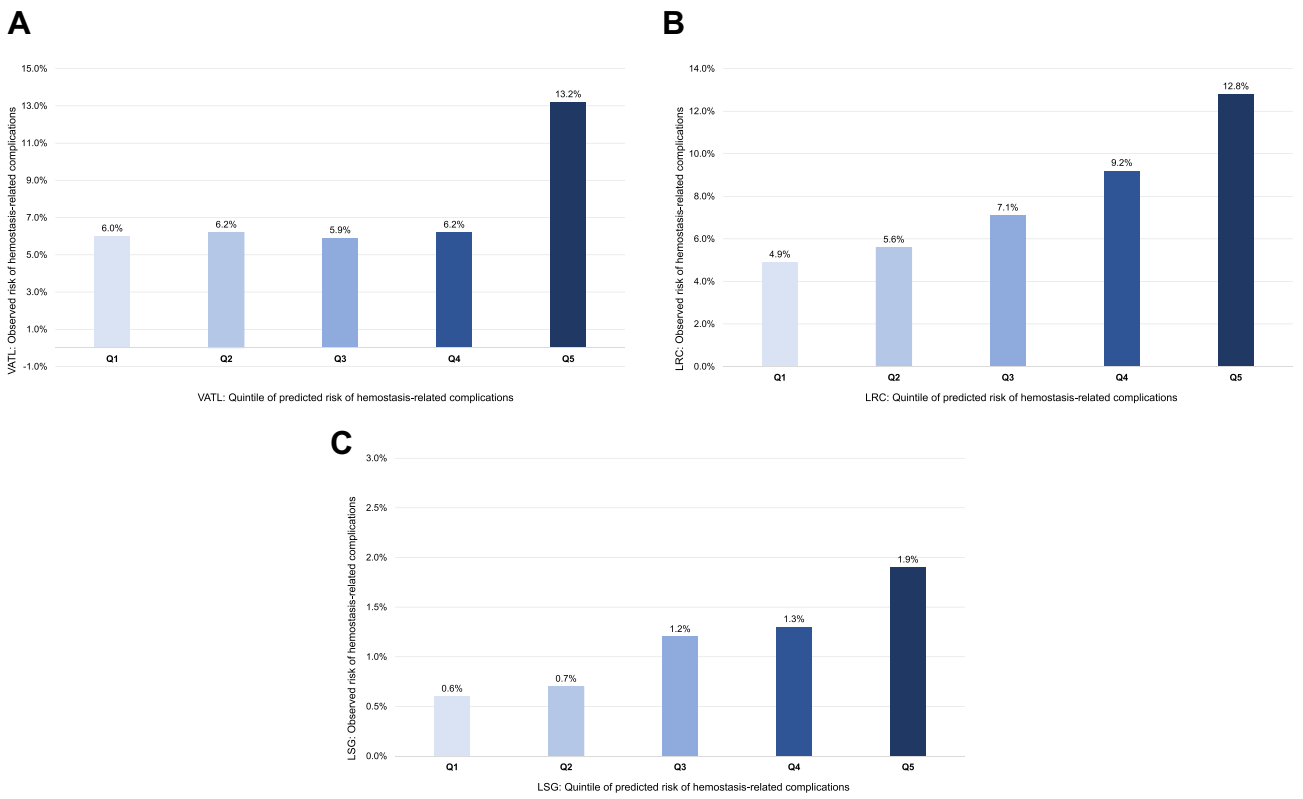

Figure 1 Observed risk of hemostasis-related complications by quintile of predicted risk. <figures to be inserted stacked vertically as follows: VATL.pdf, LRC.pdf, LSG.pdf >. (**A**) VATL. (**B**) LRC. (**C**) LSG.
Abbreviations: LRC, Laparoscopic Right Colectomy; LSG, Laparoscopic Sleeve Gastrectomy; VATL; Q1-Q5, quintiles 1–5; Video-Assisted Thoracoscopic Lobectomy.

complications. Finally, the gastrointestinal disorders category was largely driven by diagnoses related to peritoneal adhesions, particularly in the LRC and LSG cohorts, which can complicate surgical resection in the gastrointestinal tract.

Prior research on risk factors for hemostasis-related complications in the areas of VATL, LRC, and LSG has primarily focused on post-operative bleeding. Using a variety of data sources, predictors, and research designs, some similarities in identified risk factors between the previous literature and the present study include anemia/hemophilia (LRC), nutritional disorders (LRC), general comorbidity (VATL), and anticoagulation (LSG).[8–10]

This analysis and approach are subject to limitations. First, although we applied the CCSR taxonomy to the initial high-dimensional dataset in order to reduce the dimensionality of the data and improve the clinical interpretability and relevance of the individual features, in doing so it is possible that one or only a few of the individual diagnoses that a specific CCSR category comprises could be the predominant driver of an association between a candidate risk factor and a given outcome. To mitigate this in the future, it would be possible to deconstruct a given CCSR category and test its constituent diagnosis codes; however, this would be done at the cost of time and efficiency. Second, we used the Premier Healthcare Database in this effort because it is the database that we most frequently use in our applied research within our research team; from a clinical risk factor standpoint one could argue that an electronic health record database with rich data on laboratory results, biometrics, and/or imaging could provide much deeper insights into risk factors; however, such features are unavailable within the Premier Healthcare Database and therefore nonfungible to research conducted in that database. Third, in our example, we relied on consistency in risk factor findings across the surgical groups as a form of external cross-validation; such an approach prioritizes risk factors that generalize across surgeries but could miss ones that are idiosyncratic to an individual procedure. Fourth, diagnoses recorded in real-world databases are subject to measurement error; for example, alcohol disorders may be underdiagnosed relative to their true prevalence – thus, cross-database differences in the sensitivity and positive predictive value of data capture for predictors may influence their database-specific magnitude of influence on risk.

Finally, a key design decision that characterizes our approach was ensuring that all candidate risk factors based on diagnosis codes were designated as present on admission, to mitigate the risk of reverse causation. This "POA" indicator therefore is an important feature of the Premier Healthcare Database, but it is important to note that the POA designation is sometimes not available and sometimes not comprehensively populated in other real-world datasets. Despite these limitations, the consistency and face validity of the 6 identified risk factors suggests that there is merit in the approach we have described herein. Although the present study was based on findings across three different surgical procedures, future research in which hemostasis-related complications are studied in the context of other surgical procedures would be useful to corroborate the present study's findings; furthermore, evaluating the utility of the CCSR relative to other diagnosis taxonomies in the ability to efficiently aggregate highly granular diagnosis data is needed.

Conclusion

High-dimensional real-world data can be useful to identify generalizable risk factors for surgical complications that may be considered for multiple future studies. Within the context of high-dimensional data reduction, we find that the CCSR software is a useful clinical tool to organize diagnoses into clinically meaningful candidate risk factors. The risk factors identified herein should be considered for inclusion in future studies of hemostasis-related complications.

Acknowledgments

The abstract for this paper was presented as a poster at ICPE 2022, the 38th International Conference on Pharmacoepidemiology and Therapeutic Risk Management (ICPE), Copenhagen, Denmark, 26–28 August, 2022.

Funding

This work was supported by Johnson & Johnson; however, it received no specific grant but rather was conducted as routine methodological research work.

Disclosure

Stephen Johnston, Sanjoy Roy, and Esther Pollack are employees and stockholders of Johnson & Johnson. Aakash Jha is an employee of Mu Sigma, which was paid by Johnson & Johnson to provide data analytics support. The authors report no other conflicts of interest in this work.

References

1. Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology*. 2009;20(4):512–522. PMID: 19487948; PMCID: PMC3077219. doi:10.1097/EDE.0b013e3181a663cc
2. Tian Y, Schuemie MJ, Suchard MA. Evaluating large-scale propensity score performance through real-world and synthetic data experiments. *Int J Epidemiol*. 2018;47(6):2005–2014. PMID: 29939268; PMCID: PMC6280944. doi:10.1093/ije/dyy120
3. Arbogast PG, Ray WA. Performance of disease risk scores, propensity scores, and traditional multivariable outcome regression in the presence of multiple confounders. *Am J Epidemiol*. 2011;174(5):613–620. PMID: 21749976. doi:10.1093/aje/kwr143
4. Kumamaru H, Schneeweiss S, Glynn RJ, Setoguchi S, Gagne JJ. Dimension reduction and shrinkage methods for high dimensional disease risk scores in historical data. *Emerg Themes Epidemiol*. 2016;13:5. doi:10.1186/s12982-016-0047-x
5. Kumamaru H, Gagne JJ, Glynn RJ, Setoguchi S, Schneeweiss S. Comparison of high-dimensional confounder summary scores in comparative studies of newly marketed medications. *J Clin Epidemiol*. 2016;76:200–208. PMID: 26931292. doi:10.1016/j.jclinepi.2016.02.011
6. Premier Applied Sciences®, Premier Inc. Premier healthcare database white paper: data that informs and performs; 2019. Available from: https://learn.premierinc.com/white-papers/premier-healthcaredatabase-whitepaper. Accessed February 25, 2020.
7. Clinical Classifications Software Refined (CCSR) for ICD-10-CM Diagnoses. Healthcare Cost and Utilization Project (HCUP). Rockville, MD: Agency for Healthcare Research and Quality; 2022. Available from: www.hcup-us.ahrq.gov/toolssoftware/ccsr/dxccsr.jsp. Accessed October 21, 2022.
8. Uramoto H, Shimokawa H, Tanaka F. Postoperative bleeding after surgery in patients with lung cancer. *Anticancer Res*. 2014;34(2):981–984. PMID: 24511043.
9. Chen D, Afzal N, Sohn S, et al. Postoperative bleeding risk prediction for patients undergoing colorectal surgery. *Surgery*. 2018;164(6):1209–1216. doi:10.1016/j.surg.2018.05.043
10. Mocanu V, Dang J, Ladak F, Switzer N, Birch DW, Karmali S. Predictors and outcomes of bleed after sleeve gastrectomy: an analysis of the MBSAQIP data registry. *Surg Obes Relat Dis*. 2019;15:1675–1681. PMID: 31590999. doi:10.1016/j.soard.2019.07.017