Surgeons: Buyer beware—does “universal” risk prediction model apply to patients universally?

Makoto Mori, MD,^{a,b} David M. Shahian, MD,^{c,d} Chenxi Huang, PhD,^b Shu-Xia Li, PhD,^b Sharon-Lise T. Normand, PhD,^{e,f} Arnar Geirsson, MD,^a and Harlan M. Krumholz, MD, SM^{b,g,h}

Supplemental material is available online.

Although many surgical risk models are specific for particular operations,¹ some commonly used models are designed more generally to encompass many types of operations. Such general models include the National Surgical Quality Improvement Program surgical risk model² and EuroSCORE II (European System for Cardiac Operative Risk Evaluation).³ They leverage the observation that certain risk factors are common to various types of surgery and offer the simplicity of having one size for many clinical situations, and they may increase the power when estimating the risk of rare outcomes in a low-volume subgroup (Figure 1). However, there is a question of the consistency of model performance across all included procedure types.

Such general models are sometimes called “universal models.” Universal models are derived from data of a variety of procedure types,²⁻⁴ and developers report the model performance of the universal model to demonstrate its credibility. However, in order for a surgeon to determine whether the universal model’s prediction is accurate for the particular type of operation he/she performs, performance of the model over each constituent case type must be reported. This is often not the case for existing universal models. Reporting only the performance values for the universal model without that of clinically relevant constituent procedure types can mislead the users when

Universal Surgical Risk Model	
Strengths • “One-size fits all” • Avoids impracticality of developing each model for all possible case types • Allows modeling of case types with small case volume or rare events • As long as a strong performance over each case type is ensured, predicted risk may be reliable	Weaknesses • Global model performance depends on included case types • Poor model performance for low-volume case types can be masked
Proposed Solution • Reporting performance of the universal model over each case type	

Summary of strengths, weaknesses, and proposal to improve the use of universal risk model.

CENTRAL MESSAGE

Universal risk model’s global model performance may be misleading. Critical evaluation of such models should entail evaluating the model performance over each constituent case type.

See Commentaries on pages 180, 181, 183, and 184.

there is a discordance between the performance of the universal model and the model applied over a constituent procedure type.

We performed modeling experiments of various procedure types using a claims-based dataset to test how the performances of universal models fair when tested on individual constituent procedure types. We developed universal models consisting of case types of various event incidences and case volumes to best demonstrate how model performance may change depending on various combinations of case compositions. The resulting combinations of case types are unrealistic in terms of what a single surgeon may encounter. However, the illustrated limitations of universal models are applicable to subspecialty-specific universal risk models, such as EuroSCORE II, because the event incidence and relative case volume of the case, and not the actual case type, are the parameters of interest. We discuss the importance of reporting model performance over constituent procedure types to encourage surgeons to be cognizant of potential pitfalls in using such universal models in clinical settings.

From the ^aSections of Cardiac Surgery and ^bCardiovascular Medicine, Department of Internal Medicine, Yale School of Medicine, New Haven, Conn; ^cCenter for Outcomes Research and Evaluation, Yale-New Haven Hospital, New Haven, Conn; ^dDepartment of Surgery and ^eCenter for Quality and Safety, Massachusetts General Hospital, Harvard Medical School, Boston, Mass; ^fDepartment of Health Care Policy, Harvard Medical School, Boston, Mass; ^gDepartment of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, Mass; and ^hDepartment of Health Policy and Management, Yale School of Public Health, New Haven, Conn.

Received for publication Oct 22, 2019; revisions received Nov 18, 2019; accepted for publication Nov 21, 2019; available ahead of print Feb 19, 2020.

Address for reprints: Harlan M. Krumholz, MD, SM, Center for Outcomes Research and Evaluation (CORE), 1 Church St, Suite 200, New Haven, CT 06510 (E-mail: harlan.krumholz@yale.edu).

J Thorac Cardiovasc Surg 2020;160:176-9

0022-5223/\$36.00

Copyright © 2020 by The American Association for Thoracic Surgery

<https://doi.org/10.1016/j.jtcvs.2019.11.144>

FIGURE 1. Advantages and limitations of universal risk models. The figure summarizes advantages and limitations of universal risk model. Limitations mainly pertain to the potentially misleading nature of global model performance, which may deviate from the model performance over the case type of specific interest. Reporting model performance over relevant case type allows for surgeons critically assess whether the model can be trusted.

MODELING APPROACH

We fitted a logistic regression model on the National Inpatient Sample dataset⁵ to predict in-hospital mortality based on age, sex, case types, and Elixhauser comorbidity index⁶ following laparoscopic cholecystectomy (CCY), esophagectomy, or coronary artery bypass grafting (CABG). We chose each procedure type to represent different levels of in-hospital mortality (risk surrogate) and procedure volume, with CCY representing high-volume and low-risk procedure, esophagectomy representing low-volume and high-risk, and CABG representing high-volume and high-risk procedure type (Table 1).

We defined universal model as the model developed and tested in the combined cohort of CCY, CABG, and

esophagectomy. The performance of the universal model was compared with the performance of the universal model tested on individual constituent procedure types. The dataset was randomly split into 30:70 testing and training samples and the process was iterated 20 times. We evaluated model performance using c-statistics, which characterizes the discriminatory ability of the model, and Brier score,⁷ which is the mean-squared error of prediction and lower values indicate higher accuracy. Although many complementary metrics exist,⁴ we chose them for their familiarity and convenience in assessing the concordance of model performance between universal and constituent procedure models. Odds ratios, confidence intervals, and *P* values for the included covariates for the model including all 3 case types are summarized in Tables E1-E3.

DISCREPANCIES BETWEEN UNIVERSAL AND CONSTITUENT CASE MODEL PERFORMANCES

We first demonstrated that the models developed in the universal cohort and tested in the cohort of each constituent procedure type (CCY, CABG, and esophagectomy) differed in their c-statistics and Brier scores from that of the universal model (Table 1: Model 1 vs 2, 3, 4). The performances for models tested on each constituent procedure type are the metrics that should arguably be regarded as the gold standard. The performance of the universal model is what has been reported in existing universal models, and its deviations from c-statistics and Brier scores of the model tested on constituent procedure types argues for the importance of reporting performance of the universal model over each constituent procedure type.

Although the risk estimated by such universal model could still be accurate for selected procedure types, a thoracic surgeon performing esophagectomy may be misled to think the model performance for this procedure had a c-statistic of 0.864, the global model performance when tested

TABLE 1. Summary of model performances by combinations of case types

Model	Training cohort	Testing cohort	C-statistic Brier score	SD	Testing event rate	Testing N
1*	CCY + CABG + Esoph	CCY + CABG + Esoph	0.864 0.0115	0.004 0.0002	1.2%	437,901
2	CCY + CABG + Esoph	CCY	0.889 0.0037	0.009 0.0001	0.4%	268,572
3	CCY + CABG + Esoph	CABG	0.769 0.0234	0.007 0.0005	2.5%	165,583
4	CCY + CABG + Esoph	Esoph	0.708 0.0389	0.032 0.0034	4.1%	3,746
5	CCY + CABG	CCY + CABG	0.864 0.0112	0.003 0.0002	1.2%	434,155
6	Esoph + CABG	Esoph + CABG	0.771 0.0233	0.006 0.0004	2.5%	169,329

SD, Standard deviation; CCY, cholecystectomy; CABG, coronary artery bypass graft; Esoph, esophagectomy. *Model 1 is the universal model.

on all 3 procedures, when in fact the performance of the model in the esophagectomy cohort was 0.708. Brier scores showed similar discrepancies. The explanation for this observation is likely that the c-statistic can be increased when the population is quite heterogeneous, which facilitates discrimination.⁸ This discrepancy cannot be assessed when only the performance of the universal model is reported, and is particularly problematic when inferior performances in a certain procedure type is masked by an acceptable performance of the universal model.

POOR REPRESENTATION OF LOW-VOLUME CASE TYPE

A related problem is that the performance of low-volume procedure types may be poorly represented in the performance of universal model. When using the model trained and tested on the universal cohort including CABG ($n = 165,583$), CCY ($n = 268,572$), and esophagectomy ($n = 3,746$), the c-statistic of the universal model was highly skewed toward that of CCY model (universal = 0.864, CCY = 0.889), the highest-volume procedure type (Table 1: Model 1 vs 2). In addition, the performances of models did not differ between the model consisting of 2 high-volume procedure types (CCY + CABG) and the model that also included low-volume case type (CCY + CABG + esophagectomy), although c-statistics and Brier scores differed significantly between the universal model and the model trained on universal cohort and tested on esophagectomy cohort (c-statistics: 0.864 vs 0.708 and Brier scores: 0.011 vs 0.039) (Table 1: Model 1 vs 4). These observations indicate that the relative contribution of low-volume procedure type toward the performance of universal model is small, and reporting the performance of universal model alone obscures potentially poor performance of low-volume procedure types.

In the case of EuroSCORE II, the model performance over aortic operation may suffer from this issue. EuroSCORE II is derived from a cohort that is predominantly CABG and valve operations, and aortic operation comprises the small fraction of the modeled cohort. Therefore, model performance driven primarily by CABG and valve operations may mask the actual model performance over aortic operations.

UNIVERSAL MODEL PERFORMANCE CHANGES BY THE INCLUDED CASE TYPES

In addition, the performance of universal model depends on the composition of procedure types used to develop and test the model, which complicates interpretation of such performance values. For example, for a cardiac surgeon deciding whether a universal model is useful in estimating a patient's risk of CABG in a situation where only the global performance of the universal model is known, a model derived from the universal cohort of CABG + CCY +

esophagectomy would have a c-statistic of 0.864 and Brier score of 0.011, and may appear as the best option. In contrast, although using the same exact CABG cohort, a universal model developed from CABG + esophagectomy cohort would have a c-statistic of 0.771 and Brier score of 0.023 (Table 1: Model 1 vs 6). Because the performance of the universal model is sensitive to the procedure types included in deriving the model, it does not serve as a reliable metric to decide whether to use a model in a surgeon's specialized practice.

This finding is partly explained by previous observation that the degree of risk heterogeneity within the cohort impacts the model performance.^{8,9} For example, it is intuitively easier to distinguish which patients are more likely to experience operative death among the heterogeneous cohort of CABG and CCY patients than among the homogeneous cohort of CABG patients only. This is because the procedure type (CABG or CCY) can serve as a strong predictor of event in the heterogeneous cohort, while this predictor is not available in the cohort solely consisting of CABG. The problem lies in the fact that such procedure type (CABG or CCY) is a contextually limited predictor, in that a practicing surgeon rarely performs both of such distinct operations in his/her practice.

VARIABLE EFFECT SIZES OF UNOBSERVED COVARIATES BETWEEN CASE TYPES

Although unexamined in our modeling exercise, another potential limitation of universal risk models is related to the effect size of unobserved covariates that may vary between different case types. For example, a given level of frailty may have a more pronounced effect toward mortality risk in case types that predominantly consist of older individuals, such as CABG and transcatheter aortic valve replacement, compared with cases performed for younger individuals with more reserve, such as mitral valve repair for degenerative diseases. Such differences in the effect size can be accounted for by treatment of interactions, if the variable is observed and included in the model. However, when such a variable that is selectively important to a certain case type is not observed, those case types suffer from inferior model performance more than the case types in which the unobserved variable is not as important would. A remedy may require collection of such variable pertinent to particular case type for all cases to allow for adjustment in the universal model. However, in the current convention of manual data collection based on case report forms, such approach may require prohibitive amount of resource.

REPORTING THE MODEL PERFORMANCE OVER THE COHORT OF CLINICAL INTEREST

Because universal model performances vary depending on the included procedure types and case volumes, performances of the universal model over clinically relevant

procedure types should be reported to allow for evaluation of the model credence in the surgeon's specialized practice. As we demonstrated, the clinical significance of "good" universal model performance is unclear, because the global performance of the universal model depends on the composition of cases included in the derivation cohort. Universal model's risk estimate over some procedure types may indeed be accurate, especially by adjusting for procedure types or using tree-based modeling approach to account for higher-order interactions between covariates and each procedure type.¹⁰ However, the problem lies in the potential discrepancy between the global performance value reported for the universal model and the actual performance of constituent case-specific models, latter of which is often not reported but of clinical interest. Therefore, we recommend reporting the universal model's performance for clinically relevant procedure types and comparing with procedure-specific models such as those used in the Society of Thoracic Surgeons' Adult Cardiac Surgery Database.¹

CONCLUSIONS

Current enthusiasm for innovative modeling techniques applied over big data containing a variety of procedure types must be critically examined. Although universal models developed from heterogeneous procedure cohorts offer simplicity, their c-statistics may be overoptimistic because of the substantial heterogeneity of their population of procedures. Depending on the combinations of case types, event rates, and procedure volumes included in the training set, they may not perform as well as procedure-specific models. These performance characteristics should be explicitly tested before applying universal models to specific procedures.

Conflict of Interest Statement

Dr Normand: Patient long-term prognosis quantitative prediction and intervention system and method; National

Center for Cardiovascular Diseases, China. Dr Krumholz: Medtronic, US Food and Drug Administration, Johnson & Johnson (Janssen), Centers for Medicare & Medicaid Services; Shenzhen Center, National Center for Cardiovascular Diseases in Beijing, for Health Information, Arnold & Porter Law Firm, Martin Baughman Law Firm, Siegfried and Jensen Law Firm, UnitedHealth, IBM Watson Health Life Sciences Board, Element Science, Facebook, Aetna, HugoHealth, Refactor Health. All other authors have nothing to disclose with regard to commercial support.

References

1. Shahian DM, O'Brien SM, Filardo G, Ferraris VA, Haan CK, Rich JB, et al. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: part 1—coronary artery bypass grafting surgery. *Ann Thorac Surg.* 2009;88:S2-22.
2. Bilimoria KY, Liu Y, Paruch JL, Zhou L, Kmieciak TE, Ko CY, et al. Development and evaluation of the universal ACS NSQIP surgical risk calculator: a decision aid and informed consent tool for patients and surgeons. *J Am Coll Surg.* 2013;217:833-42. e1-3.
3. Nashef SA, Roques F, Sharples LD, Nilsson J, Smith C, Goldstone AR, et al. EuroSCORE II. *Eur J Cardiothorac Surg.* 2012;41:734-44; discussion 744-5.
4. Chiew CJ, Liu N, Wong TH, Sim YE, Abdullah HR. Utilizing machine learning methods for preoperative prediction of postsurgical mortality and intensive care unit admission. *Ann Surg.* 2019;270:e138.
5. Healthcare Cost and Utilization Project (HCUP). HCUP Nationwide Inpatient Sample (NIS). 2004-2011. Agency for Healthcare Research and Quality, Rockville, MD. Available at: <http://www.hcup-us.ahrq.gov/nisoverview.jsp>. Accessed March 20, 2019.
6. Elixhauser A, Steiner C, Harris DR, Coffey RM. Comorbidity measures for use with administrative data. *Med Care.* 1998;36:8-27.
7. Murphy AH. A new vector partition of the probability score. *J Appl Meteorol.* 1973;12:595-600.
8. Austin PC, Reeves MJ. The relationship between the C-statistic of a risk-adjustment model and the accuracy of hospital report cards: a Monte Carlo Study. *Med Care.* 2013;51:275-84.
9. Merkow RP, Hall BL, Cohen ME, Dimick JB, Wang E, Chow WB, et al. Relevance of the c-statistic when evaluating risk-adjustment models in surgery. *J Am Coll Surg.* 2012;214:822-30.
10. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Presented at: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016; San Francisco, Calif.

TABLE E1. ORs and CIs for the model including CABG, cholecystectomy, and esophagectomy

Variables	OR (95% CI)	P value
Cholecystectomy (ref. CABG)	0.18 (0.16-0.19)	<.0001
Esophagectomy (ref. CABG)	1.23 (1.02-1.47)	.031
Age (per 1-y increase)	1.04 (1.04-1.05)	<.0001
Female	1.37 (1.29-1.46)	<.0001
Elixhauser comorbidity		
AIDS	1.92 (0.92-4.01)	.082
Alcohol abuse	0.72 (0.58-0.91)	.0045
Anemia (non-blood loss)	0.64 (0.58-0.70)	<.0001
Rheumatoid arthritis	0.77 (0.61-0.97)	.027
Blood loss anemia	0.50 (0.34-0.72)	.0003
Congestive heart failure	3.16 (2.85-3.50)	<.0001
Chronic lung disease	0.87 (0.80-0.94)	.0002
Coagulopathy	1.47 (1.37-1.57)	<.0001
Depression	0.55 (0.46-0.65)	<.0001
Diabetes without complications	0.71 (0.66-0.76)	<.0001
Diabetes with complications	0.59 (0.52-0.67)	<.0001
Drug abuse	1.07 (0.78-1.47)	.66
Hypertension	0.42 (0.40-0.45)	<.0001
Hypothyroidism	0.62 (0.55-0.70)	<.0001
Liver disease	1.53 (1.29-1.81)	<.0001
Lymphoma	1.41 (1.04-1.92)	.030
Fluid and electrolyte disorders	2.32 (2.19-2.46)	<.0001
Metastatic cancer	1.60 (1.24-2.07)	.0004
Other neurologic disorders	1.13 (0.99-1.30)	.077
Obesity	0.85 (0.77-0.93)	.0002
Paralysis	1.68 (1.43-1.99)	<.0001
Peripheral vascular disease	1.57 (1.45-1.69)	<.0001
Psychosis	0.90 (0.72-1.11)	.32
Pulmonary circulation disorder	2.40 (1.99-2.88)	<.0001
Renal failure	1.75 (1.63-1.88)	<.0001
Solid tumor without metastasis	1.11 (0.87-1.41)	.40
Peptic ulcer disease	1.21 (0.27-5.43)	.806
Valvular disease	1.06 (0.88-1.27)	.54
Weight loss	1.99 (1.83-2.16)	<.0001

Bold values represent P value <.05. OR, Odds ratio; CI, confidence interval; CABG, coronary artery bypass graft; AIDS, acquired immunodeficiency syndrome.

TABLE E2. ORs and CIs for the model including CABG and cholecystectomy

Variables	OR (95% CI)	P value
Cholecystectomy (ref. CABG)	0.17 (0.16-0.19)	<.0001
Age (per 1-y increase)	1.04 (1.04-1.05)	<.0001
Female	1.37 (1.29-1.45)	<.0001
Elixhauser comorbidity		
AIDS/HIV	1.77 (0.81-3.89)	.15
Alcohol abuse	0.72 (0.58-0.91)	.0057
Anemia (non-blood loss)	0.64 (0.58-0.70)	<.0001
Rheumatoid arthritis	0.79 (0.63-0.99)	.044
Blood loss anemia	0.50 (0.34-0.73)	.0004
Congestive heart failure	3.25 (2.93-3.60)	<.0001
Chronic lung disease	0.86 (0.80-0.93)	.0001
Coagulopathy	1.45 (1.36-1.56)	<.0001
Depression	0.55 (0.46-0.65)	<.0001
Diabetes without complications	0.71 (0.66-0.76)	<.0001
Diabetes with complications	0.59 (0.52-0.68)	<.0001
Drug abuse	1.09 (0.79-1.50)	.60
Hypertension	0.43 (0.40-0.45)	<.0001
Hypothyroidism	0.62 (0.55-0.70)	<.0001
Liver disease	1.50 (1.26-1.78)	<.0001
Lymphoma	1.45 (1.07-1.98)	.0177
Fluid and electrolyte disorders	2.31 (2.18-2.45)	<.0001
Metastatic cancer	2.91 (2.21-3.83)	<.0001
Other neurologic disorders	1.13 (0.99-1.30)	.081
Obesity	0.86 (0.78-0.94)	.0007
Paralysis	1.68 (1.42-1.98)	<.0001
Peripheral vascular disease	1.57 (1.45-1.70)	<.0001
Psychosis	0.90 (0.72-1.12)	.35
Pulmonary circulation disorder	2.30 (1.89-2.79)	<.0001
Renal failure	1.75 (1.62-1.88)	<.0001
Solid tumor without metastasis	1.06 (0.82-1.37)	.66
Peptic ulcer disease	1.27 (0.28-5.77)	.76
Valvular disease	1.07 (0.88-1.29)	.50
Weight loss	2.04 (1.87-2.23)	<.0001

OR, Odds ratio; CI, confidence interval; CABG, coronary artery bypass graft; AIDS, acquired immunodeficiency syndrome.

TABLE E3. ORs and CIs for the model including CABG and esophagectomy

Variables	OR (95% CI)	P value
Esophagectomy (ref. CABG)	1.40 (1.16-1.67)	.0003
Age (per 1-y increase)	1.03 (1.03-1.03)	<.0001
Female	1.64 (1.54-1.76)	<.0001
Elixhauser comorbidity		
AIDS/HIV	1.59 (0.68-3.72)	.28
Alcohol abuse	0.61 (0.47-0.79)	.0002
Anemia (non-blood loss)	0.62 (0.56-0.69)	<.0001
Rheumatoid arthritis	0.70 (0.54-0.92)	.012
Blood loss anemia	0.39 (0.25-0.62)	<.0001
Congestive heart failure	2.82 (2.45-3.25)	<.0001
Chronic lung disease	0.81 (0.75-0.88)	<.0001
Coagulopathy	1.34 (1.25-1.44)	<.0001
Depression	0.55 (0.45-0.67)	<.0001
Diabetes without complications	0.67 (0.62-0.73)	<.0001
Diabetes with complications	0.54 (0.46-0.62)	<.0001
Drug abuse	0.99 (0.69-1.42)	.95
Hypertension	0.40 (0.38-0.43)	<.0001
Hypothyroidism	0.60 (0.52-0.68)	<.0001
Liver disease	1.89 (1.52-2.35)	<.0001
Lymphoma	1.06 (0.71-1.57)	.78
Fluid and electrolyte disorders	2.12 (1.99-2.26)	<.0001
Metastatic cancer	0.62 (0.40-0.95)	.026
Other neurologic disorders	1.08 (0.92-1.27)	.36
Obesity	0.83 (0.76-0.92)	.0003
Paralysis	1.53 (1.27-1.84)	<.0001
Peripheral vascular disease	1.55 (1.43-1.69)	<.0001
Psychosis	0.76 (0.58-1.00)	.050
Pulmonary circulation disorder	2.55 (1.94-3.34)	<.0001
Renal failure	1.71 (1.58-1.85)	<.0001
Solid tumor without metastasis	1.06 (0.80-1.40)	.69
Peptic ulcer disease	1.91 (0.38-9.54)	.43
Valvular disease	1.62 (1.26-2.07)	.0002
Weight loss	1.66 (1.50-1.83)	<.0001

OR, Odds ratio; CI, confidence interval; CABG, coronary artery bypass graft; AIDS, acquired immunodeficiency syndrome.