Original Study

ANNALS OF SURGERY OPEN

OPEN

The Accuracy of the NSQIP Universal Surgical Risk Calculator Compared to Operation-Specific Calculators

Mark E. Cohen, PhD,* Yaoming Liu, PhD,* Bruce L. Hall, MD, PhD, MBA, FACS,*† and Clifford Y. Ko, MD, MS, MSHS, FACS*‡

Objective: To compare the performance of the ACS NSQIP "universal" risk calculator (N-RC) to operation-specific RCs.

Background: Resources have been directed toward building operation-specific RCs because of an implicit belief that they would provide more accurate risk estimates than the N-RC. However, operation-specific calculators may not provide sufficient improvements in accuracy to justify the costs in development, maintenance, and access.

Methods: For the N-RC, a cohort of 5,020,713 NSQIP patient records were randomly divided into 80% for machine learning algorithm training and 20% for validation. Operation-specific risk calculators (OS-RC) and OS-RCs with operation-specific predictors (OSP-RC) were independently developed for each of 6 operative groups (colectomy, whipple pancreatectomy, thyroidectomy, abdominal aortic aneurysm (open), hysterectomy/myomectomy, and total knee arthroplasty) and 14 outcomes using the same 80%/20% rule applied to the appropriate subsets of the 5M records. Predictive accuracy was evaluated using the area under the receiver operating characteristic curve (AUROC), the area under the precision-recall curve (AUPRC), and Hosmer-Lemeshow (H-L) *P* values, for 13 binary outcomes, and mean squared error for the length of stay outcome.

Results: The N-RC was found to have greater AUROC ($P = 0.002$) and greater AUPRC ($P < 0.001$) compared to the OS-RC. No other statistically significant differences in accuracy, across the 3 risk calculator types, were found. There was an inverse relationship between the operation group sample size and magnitude of the difference in AUROC ($r = –0.278$; $P = 0.014$) and in AUPRC ($r = –0.425$; $P < 0.001$) between N-RC and OS-RC. The smaller the sample size, the greater the superiority of the N-RC.

Conclusions: While operation-specific RCs might be assumed to have advantages over a universal RC, their reliance on smaller datasets may reduce their ability to accurately estimate predictor effects. In the present study, this tradeoff between operation specificity and accuracy, in estimating the effects of predictor variables, favors the N-R, though the clinical impact is likely to be negligible.

Keywords: ACS NSQIP, risk calculator, accuracy

INTRODUCTION

NSQIP's universal surgical risk calculator (N-RC) was designed to provide a predictive device to support planning and shared surgical decision-making with informed consent, which both surgeons and patients could use.[1] The N-RC accurately estimates risk for many of the most common surgical procedures using a small, fixed set of accessible predictor variables.[2,3] As the calculator is developed and periodically updated using cases only in the ACS-NSQIP, the calculator does not have information for all types of surgical specialties. However, the overall development strategy has resulted in a widely-used, easily-maintained tool with accuracy sufficient for its intended purposes and relevant to a wide variety of procedure types.

Nevertheless, the N-RC has been criticized for having purportedly poorer performance compared to operation-specific RCs. This inferiority might stem, in theory, from 2 sources. (1) In the context of a (logistic) regression model, optimal parameter values for the model's predictors might be inconsistent across different operation types. An RC that optimizes parameter estimates for specific operations might, therefore, outperform a universal RC that relies on parameter estimates derived from all operations. This is equivalent to arguing that there are operation groups by other-predictor-variable interactions, which are not accounted for by a universal RC. This interaction problem is avoided by operation-specific risk calculators. (2) The N-RC's reliance on a fixed set of predictors, common to all operations, precludes it from taking advantage of potentially potent operation-specific predictors. In that regard, studies of the N-RC have sometimes attributed poor performance to the absence of important operation-specific or other predictors, and comparative studies have sometimes attributed nominal superiority (nominal because of sometimes unresolved issues of research design) of alternative RCs to the inclusion of these predictors.[2,4,5]

These are valid concerns, but there are counterarguments. Regarding point (1) above, there are accuracy costs when building RCs from smaller samples of operation-specific cases. While the parameter estimates might be more appropriate to the specific operation group, the smaller sample from which the parameter values are estimated might make them less reliable

From the ˙Division of Research and Optimal Patient Care, American College of Surgeons, Chicago, IL; †Department of Surgery, Washington University in St. Louis, Center for Health Policy and the Olin Business School at Washington University in St Louis, John Cochran Veterans Affairs Medical Center; and BJC Healthcare, St. Louis, MO; and ‡Department of Surgery, University of California Los Angeles David Geffen School of Medicine and the VA Greater Los Angeles Healthcare System, Los Angeles.

The authors declare no conflicts of interest except being employees of the American College of Surgeons, which has a proprietary claim to the ACS NSQIP Risk Calculator.

Reprints: Mark E. Cohen, PhD, American College of Surgeons, 633 N. Saint Clair St, 23nd Floor, Chicago, IL 60611-3211. Email: mcohen@facs.org.

Annals of Surgery Open (2023) 4:e358

Received: 31 August 2023; Accepted 9 October 2023

Published online 15 November 2023

DOI: 10.1097/AS9.0000000000000358

compared to those from a universal RC. It is unclear how this tradeoff between appropriateness and reliability resolves for a particular operation-specific RC. In addition, the recent transition of the N-RC from logistic regression to machine learning (ML), could make the interaction issue (operation-group by other-predictor-variables) moot, as ML can inherently address interactions among variables.[6] Regarding point (2) above, the use of additional variables would be expected (in almost all situations) to improve predictions, but the marginal improvement in accuracy, emanating from those additional variables, may or may not be dramatic and might not justify the business enterprise costs of creating and maintaining operation-specific datasets and large numbers of operation-specific RCs. It might also be the case that reliance on special variables might necessitate that data be collected from a small number of hospitals contributing to the model development effort, which could make the RC less generalizable.

Furthermore, the net result of these competing influences could be influenced by the sample size of the operation group, the event rate for the outcome of interest, and the predictive strength of the operation-specific variables. To evaluate universal versus operation-specific RCs, 6 operation groups, with varied sample sizes and different sets of operation-specific variables, and 14 outcomes with varied event rates, were studied. This empirical investigation is intended to quantify the benefits (or absence of benefits) associated with operation-specific RCs with, and without, operation-specific predictors, compared to the N-RC.

METHODS

Risk predictions for the N-RC were based on the dataset and machine learning (ML) methods described elsewhere.[6] In summary, data from 5,020,713 NSQIP operations, during calendar years 2016 to 2020, were split so that 80% of cases were used for training on an extreme gradient boosting ML learning algorithm and cases that were included in the 6 surgical groups from the remaining 20% of all cases, were used for validation. The N-RC uses the 21 predictors described in Table 1 to create 14 algorithms for the 14 outcomes studied here (death, morbidity, serious morbidity, cardiac complication, pneumonia, venous thromboembolism, renal failure, urinary tract infection, surgical site infection, sepsis, return to OR, unplanned readmission, discharge to nursing or rehabilitation facility, and length of stay [LOS]).

Relevant portions of the same 2016 to 2020 dataset were used for the construction of each of 6 sets of operation-specific risk calculators. Thus, the training sets for the operation-operation specific calculators were much smaller than the single training set for the N-RC dataset. Operations providing a variety of risk levels and sample sizes were selected; they were colectomy, whipple pancreatectomy, thyroidectomy, abdominal aortic aneurysm (open), hysterectomy/myomectomy, and total knee arthroplasty. Sample sizes for each operation group and outcome are shown in Table 2, and Table 3 shows event rates for each operation group and outcome.

The risk calculators were studied as follows. (1) Predictions were made for each outcome using the N-RC's standard 21-predictor set. (2) As just described, except that predictions were made by new operation-group-specific RCs built using the 80%/20% training/validation strategy applied to each operative group—this approach is identified as OS-RC (operation-specific risk calculator). (3) As just described, except that, in addition to the standard 21 predictors, appropriate NSQIP operation-specific preoperative risk variables were also used—this approach is identified as OSP-RC (operation-specific with operation-specific predictors risk calculator). The additional OSP-RC risk predictors are listed in Table 1 (total knee arthroplasty does not have operation-specific variables in NSQIP).

Accuracy metrics were compared when the N-RC, the OS-RC, and the OSP-RC all operated on the same 20% validation datasets. The accuracy metrics studied were the area under the receiver operating characteristic curve (AUROC), the area under the precision-recall curve (AUPRC), and the Hosmer-Lemeshow (H-L) *P*, for binary outcomes, and the mean squared error for the LOS outcome. As described elsewhere, the AUROC is a commonly used as a measure of event/nonevent discrimination which can range from 0.5 (chance discrimination) to 1.0 (perfect discrimination), the AUPRC is a measure of discrimination that is appropriate when data are imbalanced (many nonevent cases) and interest is centered on correctly predicting events (the chance AUPRC is equal to the event rate), and the H-L *P* is a measure of calibration ranging from 0.0 to 1.0, where larger values indicate less evidence of miscalibration ($P \leq 0.05$ indicates statistically significant miscalibration, which might not be of practical importance for large data sets).[6]

RESULTS

This study is not concerned with evaluating accuracy metrics observed for different operation groups and outcomes. Rather, interest is directed towards differences in those metrics associated with N-RC, OS-RC, and OSP-RC methods for deriving risk estimates. In this context, the different operation groups and outcomes serve as samples from the population of potential operations and outcomes for which RCs could be developed. Therefore, scientific interest is directed at comparing the accuracy of the different methods averaged across the "sampled" types of operations and outcomes.

Table 4 shows that the N-RC is superior to the OS-RC with respect to both AUROC (mean = 0.698 and 0.679, respectively; $P = 0.002$) and AUPRC (mean = 0.122 and 0.114, respectively; $P < 0.001$). No other statistically significant difference in accuracy metrics was observed between the N-RC and either the OS-RC or the OSP-RC. Table 5 provides detailed information about risk calculator performance for the mortality and morbidity outcomes. While interest has been directed towards differences averaged across surgical groups and outcomes, this Table provides some insight into the variability observed for individual surgical group and outcome components.

The magnitude of the difference in AUROC, between N-RC and the OS-RC methods, was shown to be inversely related to the operation sample size (taken from Table 2) on which the OS-RC was trained and validated (Fig. 1A; N = 78; r = –0.278; $P = 0.014$) but not to the event rate (taken from Table 3) for the outcome that the RC was directed toward (N = 78; r = –0.180; $P = 0.116$). The magnitude of the differences in AUPRC, between the N-RC and the OS-RC methods, was also shown to be inversely related to sample size (Fig. 1B; N = 78; r = –0.425; $P < 0.001$), but not to the event rate (N = 78; r = –0.095; $P = 0.407$). For both AUROC and AUPRC, the smaller the operation group sample size used for OS-RC algorithm construction and validation, the greater the difference between N-RC and OS-RC, favoring the N-RC where the algorithm was trained using 80% of the 5M-case dataset.

DISCUSSION

The N-RC was shown to provide significantly greater AUROC and AUPRC than the OS-RC, although the clinical impact of these statistically significant improvements is likely to be negligible. The magnitude of this superiority was shown to be inversely proportional to the sample size of cases used to develop the OS-RC. This suggests that the performance of the N-RC is related to a more accurate assessment of predictor effects due to much larger sample sizes compared to OS-RCs and OSP-RCs. The OS-RC's assumed better ability, compared to the N-RC, to estimate predictor effects unique to the operation group (which

TABLE 1.

Predictors Used in the Standard N-RC and Additional Predictors Used for Each of the OSP-RC Operation Groups Identified by CPT Code

Standard N-RC predictors	CPT code (which yields an NSQIP-proprietary linearized CPT-specific risk score and access to an RVU value), Age, gender, functional status, emergent status, ASA class, steroid use, ascites, sepsis category, ventilator dependent, disseminated cancer, diabetes, hypertension, congestive heart failure, dyspnea, smoker, COPD, dialysis, acute renal failure, and BMI
Colectomy (44140, 44141, 44143, 44144, 44145, 44146, 44147, 44150, 44151, 44160, 44204, 44205, 44206, 44207, 44208, 44210)	Colon steroid/immunosuppressant use, colon preoperative, mechanical bowel prep, colon preoperative oral antibiotic, colon chemotherapy within 90 days, colon primary indication for surgery Colon indication for surgery if emergent, colon operative approach Colon pathologic T stage, colon pathologic n stage, colon pathologic M stage
Whipple pancreatectomy (48150, 48152, 48153, 48154, 48155)	Preoperative obstructive jaundice, preoperative antibiotics, preoperative biliary stent, chemotherapy within 90 days, radiation therapy within 90 days, operative approach, pancreatic duct size, pancreatic gland texture, pancreatic reconstruction WHIPPLE, drains Vascular resection
Thyroidectomy (60200, 60210, 60212, 60220, 60225, 60240, 60252, 60254, 60260, 60270, 60271)	Primary indication for surgery, if nodule goiter or graves- clinical toxicity, prior neck surgery, preoperative needle biopsy result, operative approach, central neck dissection performed, use of harmonic scalpel or ligasure or other vessel sealant device, intra-operative electrophysiologic or electromyographic RLN monitoring, drain usage, neoplasm, if cancer tumor T classification, multifocal cancer, if cancer lymph node N classification, if cancer distant metastasis M classification, postoperative calcium level checked, postoperative parathyroid (PTH) level checked, postoperative calcium and vitamin D replacement
AAA (open) (34830, 34831, 34832, 35081, 35082, 35091, 35092, 35102, 35103)	Indication for surgery, aneurysm diameter category, prior abdominal aortic surgery, surgical approach, proximal clamp location, proximal aneurysm extent, distal extent, management of inferior mesenteric artery, renal revascularization, visceral revascularization, lower extremity revascularization, abdominal nonarterial repair, or excision
Hysterectomy/myomectomy (58140, 58145, 58146, 58150, 58152, 58180, 58200, 58210, 58240, 58260, 58262, 58263, 58267, 58270, 58275, 58280, 58285, 58290, 58291, 58292, 58294, 58541, 58542, 58543, 58544, 58545, 58546, 58548, 58550, 58552, 58553, 58554, 58570, 58571, 58572, 58573, 58575, 58940, 58943, 58950, 58951, 58952, 58953, 58954, 58956)	Prior abdominal operations, prior pelvic operations, pelvic inflammatory disease, gynecologic cancer case, presence of gross abdominal disease, uterine weight
TKA (27447, 27486, 27487)	N/A

AAA indicates abdominal aortic aneurysm; ASA, American Society of Anesthesiology; BMI, body mass index; COPD, chronic obstructive pulmonary disease; CPT, current procedural terminology; PTH, postoperative parathyroid; RLN, recurrent laryngeal nerve; RVU, relative value unit; TKA, total knee arthroplasty.

TABLE 2.

OS-RC and OSP-RC Dataset Sample Sizes

	N (5 Years of Data With 80% Used for Training and 20% for Validation)					
	Colectomy	Whipple Pancreatectomy	Thyroidectomy	AAA (Open)	Hysterectomy/Myomectomy	TKA
Mortality	203,001	22,978	30,535	2686	183,975	331,523
Morbidity	203,001	22,978	30,535	2686	183,975	331,523
Serious Morbidity	203,001	22,978	30,535	2686	183,975	331,523
Cardiac complication	203,001	22,978	30,535	2686	183,975	331,523
Pneumonia	201,872	22,955	30,523	2672	183,951	331,486
VTE	203,001	22,978	30,535	2686	183,975	331,523
Renal failure	202,698	22,974	30,534	2681	183,970	331,511
UTI	202,424	22,949	30,528	2682	183,767	331,346
SSI	197,374	22,426	30,533	2673	183,798	331,048
Sepsis	189,259	22,283	30,528	2625	183,784	331,181
Return OR	203001	22,978	30,535	2686	183,975	331,523
Unplanned readmission	194,734	21,874	30,473	2264	183,648	331,028
Discharge Destination (to nursing/rehab)	196,049	22,225	30,464	2313	183,670	331,209
LOS (days)	199,526	22,171	30,497	2562	183,772	331,221

The differences in sample sizes, within each operation group, result from eligibility criteria specific to certain outcomes. For example, preoperative UTI patients were ineligible for modeling postoperative UTI. Eighty percent of each sample was used for training, and 20% for validation for each of 154 calculators (6 operation groups X 14 outcomes for OS-RC, and 5 operations X 14 outcomes for OSP-RC). The N-RCs 14 calculators (for 14 outcomes) were previously constructed using a sample of size exceeded 5M records and were not operation specific. The 20% validation sample, for each operation group shown below, was used to estimate the accuracy of all 3 risk calculators.

AAA indicates abdominal aortic aneurysm; SSI, surgical site infection; TKA, total knee arthroplasty; UTI, urinary tract infection; VTE, venous thromboembolism.

TABLE 3.

Event Rates for the Samples (Training and Validation Combined) Described in Table 2

	Event Rate (%) or Mean LOS Based on the Full 5-Year Sample					
	Colectomy	Whipple Pancreatectomy	Thyroidectomy	AAA (Open)	Hysterectomy/Myomectomy	TKA
Mortality	3.08	1.74	0.09	11.84	0.11	0.11
Morbidity	18.10	32.04	2.95	35.41	7.32	4.13
Serious Morbidity	14.72	27.79	2.38	32.95	5.82	3.42
Cardiac complication	1.59	2.38	0.16	11.80	0.19	0.27
Pneumonia	2.36	3.82	0.22	7.86	0.29	0.28
VTE	2.03	3.96	0.18	2.08	0.56	1.00
Renal failure	1.53	1.70	0.04	11.19	0.14	0.15
UTI	1.61	2.32	0.28	1.60	2.44	0.60
SSI	7.47	21.36	0.74	3.44	3.24	1.14
Sepsis	3.05	9.36	0.12	3.96	0.62	0.23
Return OR	4.81	5.41	1.29	11.17	1.30	1.06
Unplanned readmission	10.08	17.98	2.19	6.27	3.31	3.01
Discharge	10.26	10.22	0.53	22.65	0.92	13.06
Destination (to nursing/rehab)						
LOS (days)	6.00	9.40	1.10	8.40	1.50	2.10

AAA indicates abdominal aortic aneurysm; SSI, surgical site infection; TKA, total knee arthroplasty; UTI, urinary tract infection; VTE, venous thromboembolism.

TABLE 4.

Mean AUROC, Mean AUPRC, Mean H-L *P*, and Mean MSE for LOS in Days for N-RC, OS-RC, and OSP-RC Prediction Methods

	Mean AUROC	Mean AUPRC	Mean H-L *P*	Mean LOS MSE
N-RC	0.698	0.122	0.293	14.607
OS-RC	0.679	0.114	0.286	14.027
N-RC–OS-RC	0.018	0.008	0.007	0.280
N, *P*	78, 0.002	78, <0.001	78, 0.861	6, 0.099
N-RC	0.700	0.136	0.277	17.004
OSP-RC	0.700	0.138	0.313	15.716
N-RC–OSP-RC	0.000	−0.002	−0.037	1.288
N, *P*	65, 0.979	65, 0.374	65, 0.438	5, 0.072

The N of 78 represents paired (for N-RC and OS-R) data points for 6 operation groups X 13 binary outcomes. The N of 65 represents paired (for N-RC and OSP-RC) data points for 5 operation groups (TKA does not have operation-specific predictors necessary for OSP-RC) X 13 binary outcomes. The N of 6 is for LOS MSE and represents paired data points for 6 operations X the 1 continuous outcome, while N of 5 is the result of dropping TKA. *P* for paired *t* tests on the 78, 65, 6, or 5 pairs of data points. Values for N-RC appear twice, once computed using all operation groups for comparison to the OS-RC, and once where the TKA operation group is dropped for comparison to the OSP-RC (TKA does not have operation-specific variables).
MSE indicates mean squared error.

might have been the case when using logistic regression but not the case using ML) could have been insufficient to compensate for this.

However, accuracy did not significantly differ in any respect between the N-RC and the OSP-RC. It is generally assumed that additional variables will improve prediction but apparently this benefit was only able to compensate for limitations associated with estimates derived from smaller datasets, and no more. In addition, the similarity in accuracy for N-RC and OSP-RC might be due to the strength of the N-RC's 21 predictors, leaving little opportunity for the operation-specific predictors to improve OSP-RC performance.

These findings suggest that the assumption that operation-specific surgical risk calculators will provide better estimates than a universal risk calculator should be carefully evaluated, at least in real-world settings where there are limitations in available sample sizes. One can speculate that future operation-specific calculators might tend to focus on clinically interesting (higher risk), but less frequently undertaken, operations. If this happens, reduced sample sizes for modeling or algorithm training might continue to challenge OS or OSP risk calculator performance.

The need for operation-specific calculators, to replace the N-RC, has been raised in the literature many times, but the arguments supporting this proposition can be problematic in several ways. (1) While some studies have reported that the N-RC did not perform well when applied to specific groups of operations, many of these studies had design flaws related to sample size, case-mix heterogeneity, and generalizability (evaluations done on a small number of hospitals) which could lead to unjustified conclusions about the N-RC.[2] (2) It has been observed that when the N-RC is applied to progressively narrower groups of operations (e.g., All Cases, General, Colorectal, Colectomy), the AUROC systematically declines. While some suggest that this reduced discrimination demonstrates weakness of the N-RC, which could be remedied by using operation-specific RCs, this might not be the case. The reduction in AUROC could equally be due to greater case homogeneity (which reduces the range of predictor variable values on which the RC operates), which would not be remediated by using an operation-specific RC, unless it uses important new operation-specific predictors.[7] (3) To the extent that the N-RC might not have performed well when applied to groups of specific operations in the past, this might be due to the N-RC's prior reliance on logistic regression. With the N-RC's transition to ML, which has a greater capacity to account for hypothesized operation groups by other-predictor-variable interactions, this problem is likely moderated. (4) When there are direct comparisons between the N-RC and an

Table 5.

Performance Across the Surgical Categories for the Mortality and Morbidity Outcomes

	Mortality			Morbidity		
	N-RC	OS-RC	OSP-RC	N-RC	OS-RC	OSP-RC
AUROC						
Colectomy	0.921	0.920	0.924	0.714	0.713	0.721
Pancreatectomy	0.699	0.662	0.667	0.578	0.576	0.623
Thyroidectomy	0.751	0.531	0.796	0.667	0.661	0.682
AAA (open)	0.785	0.791	0.812	0.657	0.649	0.676
Hysterectomy/myomectomy	0.908	0.910	0.914	0.626	0.630	0.636
TKA	0.802	0.792		0.631	0.631	
AUPRC						
Colectomy	0.359	0.359	0.370	0.387	0.386	0.391
Pancreatectomy	0.040	0.025	0.026	0.390	0.381	0.408
Thyroidectomy	0.013	0.005	0.008	0.094	0.087	0.102
AAA (open)	0.413	0.371	0.478	0.508	0.504	0.527
Hyster/myomect	0.043	0.022	0.024	0.133	0.133	0.138
TKA	0.013	0.009		0.078	0.079	
H-L *P*						
Colectomy	0.421	0.131	0.505	0.116	0.466	0.618
Pancreatectomy	0.162	0.048	0.300	0.322	0.093	0.003
Thyroidectomy	0.296	0.007	0.369	0.279	0.222	0.486
AAA (open)	0.008	0.170	0.183	0.244	0.216	0.894
Hyster/myomect	0.152	0.731	0.831	0.142	0.293	0.470
TKA	0.656	0.684		0.170	0.006	
LOS MSE						
Colectomy	19.032	18.318	17.998			
Pancreatectomy	28.204	26.635	25.845			
Thyroidectomy	2.454	2.530	2.150			
AAA (open)	33.217	31.993	30.540			
Hyster/myomect	2.110	2.076	2.045			
TKA	2.627	2.612				

Higher values of AUROC and AUPRC indicate better discrimination and higher Hosmer-Lemeshow *P* values indicate better calibration.
AAA indicates abdominal aortic aneurysm; MSE, mean squared error; TKA, total knee arthroplasty.

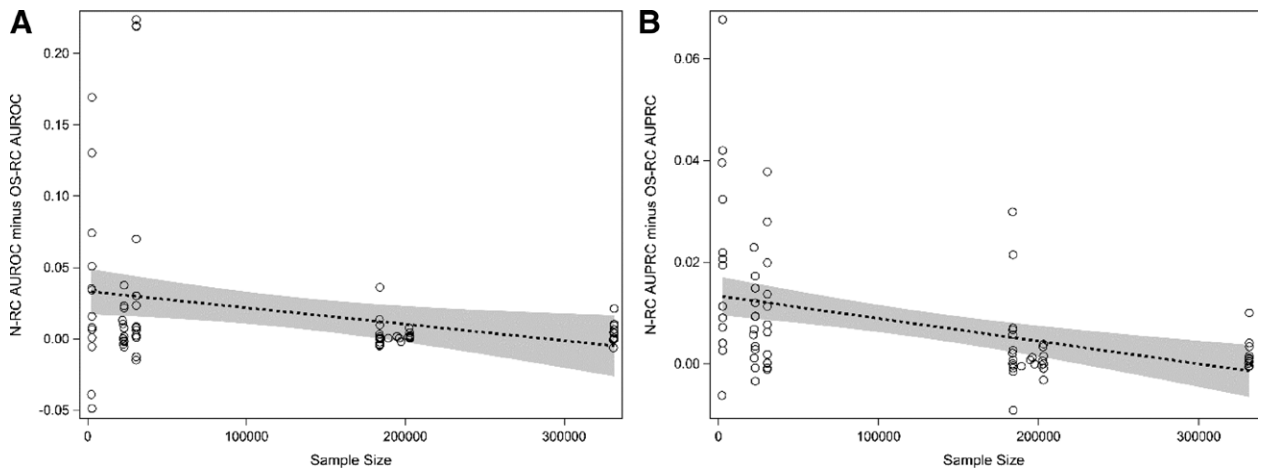

FIGURE 1. A, N-RC AUROC minus OS-RC AUROC as a function of surgical group sample size. B, N-RC AUPRC minus OS-RC AUPRC as a function of surgical group sample size. In both figures, areas under the curve are greater for the N-RC than for the OS-RC, the magnitude of that difference decreases with increasing sample size (both trends are statistically significant), and variability in differences is also observed to decrease with increasing sample size.

OS-RC or an OSP-RC, it is possible that those results might be subject to the same design flaws as noted in item 1 above. With respect to the generalizability issue, some researchers are likely to build operation-specific calculators using a local dataset, particularly when they are augmenting their model with operation-specific predictors. However, when comparing this new calculator's accuracy to that of the N-RC, the observed differences would not be due solely to the calculator. Observed differences would be influenced by relationships between the calculator and the data on which the calculator was built. Applying a new OS-RC or OSP-RC to data from the same hospital(s) on which it was developed (even if to a separate validation dataset) would give it an unfair advantage over the N-RC, which is trained on a national dataset. The problem lies, in part, in the fact that different hospitals have different quality profiles, which makes them different from the average NSQIP hospital. The new OS or OSP calculator implicitly incorporates that local quality effect (which improves predictive performance), but the N-RC cannot. However, any superiority of an OS-RC or an OSP-RC over the N-RC may not exist when applied to patients from other

hospitals. The present study avoids these enumerated problems and provides a fair comparison of RC approaches.

There are several reasons why the observation that the N-RC is superior in some areas, and noninferior in others, compared to operation-specific risk calculators, should be tempered. First, while we studied a variety of operations, with higher and lower risk levels, this still represents only a sample of operations that could have been studied. Results might be different for other operations. Second, we evaluated only the operation-specific predictors collected by NSQIP. It might well be that case that other operation-specific variables exist with such profound predictive value that an OSP-RC would be superior. Another limitation is that the sample size for LOS mean squared error was exceedingly small, with little power to distinguish differences, if they existed. Finally, the inability of the N-RC to address operation-specific outcomes (outcomes that exist only for certain operations) has not been addressed but is clearly a limitation for a universal RC (The N-RC does, in fact, present some "special" outcomes for colorectal surgery and geriatric patients, but there is not a continuing effort in this direction). To the extent that operation-specific outcomes might be of pressing clinical interest, this will, and should, continue to motivate the development of operation-specific RCs.

Despite these limitations, the present evidence supports the continued use of the N-RC for the general purposes of clinical risk assessment for planning and decision-making and specifically shared informed consent. While better operation-specific risk calculators might exist, or could be developed, their superiority over the N-RC cannot be assumed simply because they are operation-specific. In certain situations, the N-RC might have accuracy equivalent, or superior, to operation-specific calculators.

REFERENCES

1. Bilimoria KY, Liu Y, Paruch JL, 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.
2. Cohen ME, Liu Y, Ko CY, et al. An examination of american college of surgeons NSQIP surgical risk calculator accuracy. *J Am Coll Surg*. 2017;224:787–795.e1.
3. Liu Y, Cohen ME, Hall BL, et al. Evaluation and enhancement of calibration in the american college of surgeons NSQIP surgical risk calculator. *J Am Coll Surg*. 2016;223:231–239.
4. Willoughby JE, Baker JF. Utility of surgical risk calculators in spine surgery in patients aged over 80 years: analysis of SpineSage and ACS NSQIP. *Global Spine J*. 2023;13:2168–2175.
5. Sherman SK, Hrabe JE, Huang E, et al. Prospective validation of the iowa rectal surgery risk calculator. *J Gastrointest Surg*. 2018;22:1258–1267.
6. Liu Y, Ko CY, Hall BL, et al. American college of surgeons NSQIP risk calculator accuracy using a machine learning algorithm compared with regression. *J Am Coll Surg*. 2023;236:1024–1030.
7. Merkow RP, Hall BL, Cohen ME, et al. Relevance of the c-statistic when evaluating risk-adjustment models in surgery. *J Am Coll Surg*. 2012;214:822–830.