ORIGINAL ARTICLE

CrossMark

Prospective Validation of the Iowa Rectal Surgery Risk Calculator

Scott K. Sherman[1] · Jennifer E. Hrabe[2] · Emily Huang[1] · John W. Cromwell[2] · John C. Byrn[3]

Abstract

Background The Iowa Rectal Surgery Risk Calculator estimates risk for proctectomy procedures. The Iowa Calculator performed well on NSQIP 2010–2011 training and 2005–2009 validation datasets, but was not prospectively validated and did not include low anterior resections. This study sought to demonstrate validity on new independent data, to update the calculator to include low anterior resection, and to compare performance to other risk assessment tools.

Methods Non-emergent ACS-NSQIP proctectomy and low anterior resection data from 2010 to 2015 ($n = 65{,}683$) were included. The Iowa Calculator generated risk estimates for 30-day morbidity using 2012–2015 data. An Updated Calculator used 2010–2011 training data to include low anterior resection, with validation on 2012–2015 data. NSQIP data provided NSQIP Morbidity Model predictions and a custom web-script collected ACS-NSQIP Online Surgical Risk Calculator predictions for all patients.

Results Proctectomy morbidity (not including low anterior resection) decreased from 40.4% in 2010–2011 to 37.0% in 2012–2015. Low anterior resection had lower morbidity (22.4% in 2012–15). The Iowa Calculator demonstrated good discrimination and calibration using 2012–2015 data (C-statistic 0.676, deviance + 9.2%). After including low anterior resection, the Updated Iowa Calculator performed well during training (c-statistic 0.696, deviance 0%) and validation (C-statistic 0.706, deviance + 7.9%). The Updated Iowa Calculator had significantly better discrimination and calibration than morbidity predictions from the ACS Online Calculator (C-statistic 0.693, $P < 0.001$, deviance − 28.1%) and NSQIP General/Vascular Surgery Model (C-statistic 0.703, $P < 0.05$, deviance − 40.8%).

Conclusion When applied to new independent data, the Iowa Calculator supplies accurate risk estimates. The Updated Iowa Calculator includes low anterior resection, and both are prospectively validated. Risk estimation by the Iowa Calculators was superior to ACS-provided risk tools.

Keywords Rectal surgery · Morbidity · Complications · Risk calculator

Introduction

Complex pelvic anatomy, difficult disease processes including cancer and inflammatory bowel disease, cases that are clean-

✉ Scott K. Sherman
Scott.Sherman@uchospitals.edu

1 Department of Surgery, University of Chicago, 5841 S. Maryland Ave. S214, Chicago, IL, USA

2 Department of Surgery, University of Iowa Carver College of Medicine, Iowa City, IA, USA

3 Department of Surgery, University of Michigan, Ann Arbor, MI, USA

contaminated at best, and anastomoses within unfavorable geometry all contribute to high complication rates in rectal surgery. Single institution series and our own previous work with the National Surgical Quality Improvement Project (NSQIP) database showed that the rate of 30-day complications after rectal surgery approaches 40%.[1–5] Despite these high observed complication rates, earlier risk adjustment models supplied with NSQIP data predicted a complication rate of just 23% for rectal surgery.[4] The large discrepancy between NSQIP-predicted and actual complication rates spurred development of the Iowa Rectal Surgery Risk Calculator, which provided more accurate risk estimates for these operations.[4,5]

To develop the Original Iowa Rectal Surgery Risk Calculator (OIRC), NSQIP data were analyzed to identify factors associated with 30-day complications in proctectomy. From these, 17 factors available prior to operation and strongly associated with complications were included in a

multivariable logistic regression model. The model was trained using 2010–2011 data and back-validated on 2005–2009 data. The OIRC performed significantly better than the NSQIP-supplied General/Vascular Surgery Morbidity Model (NSQIP-Model) in terms of both discrimination and calibration using 2005–2009 back-validation data (C-statistic of 0.637 and percent deviance of − 1.7%, compared to 0.621 and − 48.4% for the NSQIP-Model).

Despite this strong performance, deficiencies remained in the OIRC. At the time, no new data were available for the prospective validation recommended for clinical risk estimation tools.[6] This forced a back-validation with older data. In the older dataset, a problem with recording bleeding complications necessitated imputation for this data point, potentially introducing error.[4] Additionally, the OIRC did not provide risk estimates for low anterior resection (LAR), which represents a significant proportion of rectal surgery in the USA. Finally, it was not possible to compare the OIRC's performance to the then-newly-released American College of Surgeons Online Surgical Risk Calculator (ACS-Calculator).[7] Therefore, the present study sought to expand applicability of the Iowa Calculator to LAR, to provide a full prospective validation, and to determine whether the Iowa Calculator provides superior risk estimation compared to the ACS-Calculator.

Methods

Patients and Variables Data were obtained from ACS-NSQIP Participant-Use-Data-Files for 2010–2015.[8] Inclusion and exclusion criteria, variables, and procedure categorization matched those described for the OIRC with LAR procedures added ($n = 65,683$, Table 1).[4,5] To ensure a fair comparison between the ACS-Calculator and the UIRC, records missing data such that a risk score could not be calculated by both models were excluded ($n = 282$ in 2010–2011 dataset). Pre- and post-operative variables included in the original study but no longer recorded in NSQIP were coded as not present in more recent patients when missing. Morbidity was defined as in the original model as death, reoperation, or any NSQIP-recorded complication within 30 days of surgery, matching definitions in the ACS-Calculator.[4,8,9] Significance in comparisons of continuous variables was assessed by Welch's t test and with Wilcoxon-rank-sum test for length of stay. Chi-squared or Fisher-exact tests assessed categorical variables. The Benjamini-Hochberg false-discovery-rate adjustment was used to correct P values for multiple comparisons when appropriate.[10] This study was Institutional Review Board-exempt.

Models and Predictions The OIRC is based on a multivariable logistic regression model described previously and was not altered for validation.[4] After adding LAR procedures to the 2010–2011 training dataset, the model was refitted using the same predictor variables to create the UIRC. LAR procedures were coded as laparoscopic or open without considering whether a stoma was created (Supplement). Risk predictions for the 2012–2015 validation dataset were generated using both Original and Updated Calculators. NSQIP-Model predictions were obtained from NSQIP (MORBPROB in Participant-Use-Data-Files). To obtain ACS-Calculator any-morbidity predictions, a custom web-script was developed which submitted data from each NSQIP entry for all 2012–2015 patients and recorded the ACS-Calculator prediction.[7,9] ACS-Calculator predictions were obtained in February–March, 2017.

Model Analysis Models were compared in terms of discrimination using the C-statistic with DeLong's test for significance.[11,12] Percent deviance of observed to expected rates of complications assessed calibration, and the Brier Skill Score was used to measure combined discrimination and calibration.[13,14] Statistical analyses used R v.3.4.0 (Vienna, Austria).

Results

Patient Characteristics Included were 65,683 patients in 2010–2015. The predominant procedure in 2012–2015 was laparoscopic LAR (41.6% of operations), followed by LAR (25.8%), and abdominoperineal resection (APR, 7.8%) (Table 1). Preoperative patient characteristics and risk factors were first compared in the original, non-LAR proctectomy cohort (Table 2, left columns). This revealed several differences, which, while statistically significant, were small in absolute terms. Median age increased by 1 year to 57.0, median body mass index (BMI) increased by 0.3 kg/m^2, and hematocrit increased by 0.4% ($P < 0.01$ for all). Larger and potentially more clinically meaningful differences between 2010–2011 and 2012–2015 time periods were observed in the proportions of patients treated laparoscopically (27.3 to 35.0%) and those with disseminated cancer (4.8 to 8.1%, $P < 0.001$ for both). Notably, NSQIP stopped recording a history of cerebrovascular accident or radiation therapy, and these declined sharply such that no patient was reported to have either after 2012. Proportions of female sex, diabetes, smoking, COPD, antihypertensive use, large weight loss, and preoperative leukocyte count did not significantly change between the two time periods among non-LAR proctectomy patients.

Compared to the non-LAR proctectomy cohort, patients undergoing LAR in 2012–2015 were significantly older (median 61.0 vs. 57.0 years), more likely to be female (50.5 vs. 43.6%), less likely to be on preoperative steroids or immunosuppressants (3.9 vs. 14.3%), and more frequently had benign postoperative diagnoses (51.7 vs. 13.7%), and laparoscopic

Table 1 Included CPT codes, procedure groups, sample sizes, and percentages from 2010–2011 to 2012–2015 (n = 65,683). Note: included proctectomy case numbers from 2010 to 2011 differ slightly from those previously reported due to exclusion of records for which an ACS Online Risk Calculator prediction could not be calculated because of missing data.[5]

		2010–2011		2012–2015	
CPT code	Procedure type	n	%	n	%
44207, 44208	Laparoscopic low anterior resection	4501	31.4	21,373	41.6
44145, 44146	Low anterior resection	4550	31.7	13,249	25.8
45110	Abdominoperineal resection	1292	9.0	4029	7.8
44155, 44156, 45121	Total proctocolectomy with ileostomy	499	3.5	1247	2.4
45111, 45123	Proctectomy	477	3.3	1563	3.0
44211	Laparoscopic total proctocolectomy with ileal-pouch-anal-anastomosis	476	3.3	1179	2.3
45395	Laparoscopic abdominoperineal resection	426	3.0	2121	4.1
44158	Total proctocolectomy with ileal-pouch-anal-anastomosis	332	2.3	603	1.2
45112, 45114, 45116, 45120	Proctectomy with anastomosis	333	2.3	867	1.7
45119	Proctectomy with j-pouch and coloanal anastomosis	316	2.2	767	1.5
45113	Proctectomy with ileal-pouch-anal-anastomosis	306	2.1	914	7.8
45397	Laparoscopic proctectomy with j-pouch and coloanal anastomosis	279	2.0	1542	3.0
44212	Laparoscopic total proctocolectomy with ileostomy	228	1.6	879	1.7
44157	Total proctocolectomy with ileal-anal anastomosis	140	1.0	341	0.7
45126	Pelvic exenteration	77	0.5	371	0.7
45550	Proctopexy with sigmoid resection	40	0.3	99	0.2
45402	Laparoscopic proctopexy with sigmoid resection	37	0.3	125	0.2
45160	Transsacral proctotomy	30	0.2	75	0.2
Total		14,339		51,344	

CPT current procedural terminology, *ACS* American College of Surgeons

operations (61.7 vs. 35.0%, P < 0.001 for all, Table S1). Due to these differences and the large number of LAR patients influencing the composition of the overall patient population, both the original non-LAR proctectomy cohort and the group with LAR included are presented separately (Table 2, right columns).

Observed Morbidity Patients undergoing non-LAR proctectomies had fewer complications in 2012–2015 compared to 2010–2011 (overall morbidity rate 37.0 vs. 40.4%, P < 0.001, Table 3). Contributing to lower morbidity were small decreases in most types of postoperative complications and significant decreases in rates of superficial wound infections (9.2 to 7.4%), urinary tract infections (5.7 to 4.4%), and prolonged ventilation (1.6 to 1.1%), while deep wound infections increased (2.6 to 3.2%, P < 0.05 for all). Rates of reoperation and readmission were similar between the time periods, while median length of stay decreased to 6.0 from 7.0 days (P < 0.001).

Low anterior resection was associated with significantly fewer adverse 30-day surgical outcomes as compared to other proctectomy procedures (Table S2). The rate of having any complication after LAR was 22.4 versus 37.0% with other proctectomies in 2012–2015. The individual complication demonstrating the largest difference was postoperative

bleeding (7.2 vs. 16.4%), and significantly lower rates of most individual complications were observed for LAR. Readmissions following LAR occurred in only 10.1% of patients, versus 17.2% after non-LAR proctectomy, and median length of stay was 5.0 days, a full day shorter than other proctectomies. The postoperative mortality rate was similar in LAR and other proctectomies at about 1%. Thus, after inclusion of LAR (Table 3, right columns), overall morbidity in 2012–2015 was only 27.2%. In this complete cohort, significant decreases were also observed between the time periods in postoperative bleeding, reintubation, sepsis, stroke, and death (P < 0.05 for all). Readmission and length of stay likewise decreased, and the increase in deep wound infections seen in non-LAR proctectomies was not observed. As noted during development of the Original Iowa Calculator, morbidity varied widely between different procedures,[4] from 79.0% in pelvic exenteration to only 16.0% in laparoscopic LAR (Table S3).

Prospective Validation of the OIRC Hypothesizing that valid risk estimation tools should accommodate differences in preoperative risk factors while still returning accurate predictions for postoperative outcomes, the OIRC was used to generate risk estimates for all 2012–2015 non-LAR proctectomy patients. Risk estimates were also obtained from the NSQIP-

Table 2 Patient characteristics and preoperative risk factors using original (without LAR) inclusion criteria, and with LAR included. *P* values represent comparisons between time periods

Preoperative characteristic	Without LAR			LAR Included		
	2010–2011 *n* = 5288	2012–2015 *n* = 16,722	*P* value	2010–2011 *n* = 14,339	2012–2015 *n* = 51,344	*P* value
Age, median (IQR)	56.0 (45.0–67.0)	57.0 (46.0–68.0)	*0.003*	59.0 (50.0–69.0)	60.0 (50.0–69.0)	*0.04*
Sex (female %)	45.2	43.6	0.06	49.2	48.2	0.05
Race, overall[a]			*< 0.001*			*< 0.001*
Asian (%)	3.0	3.2	0.4	2.5	3.1	*< 0.001*
Black (%)	7.5	7.2	0.4	6.9	7.1	0.5
Hispanic (%)	4.1	5.0	*0.008*	3.9	5.0	*< 0.001*
Native American (%)	0.9	0.4	*< 0.001*	0.5	0.3	*0.01*
Unknown/other (%)	8.2	9.6	*0.003*	8.2	9.6	*< 0.001*
White (%)	76.4	74.5	*0.012*	78.0	74.9	*< 0.001*
Body mass index, median kg/m^2 (IQR)	26.3 (22.9–30.5)	26.6 (23.1–30.7)	*0.004*	27.2 (23.6–31.4)	27.5 (24.0–31.7)	*< 0.001*
Diabetes (%)	11.8	12.3	0.3	12.7	13.1	0.2
Smoking (%)	17.6	16.8	0.2	18.1	17.6	0.2
Steroid or immunosuppressive use (%)	12.8	14.3	*0.008*	6.5	7.3	*0.002*
Dyspnea (%)	5.7	5.0	0.07	7.0	5.3	*< 0.001*
COPD (%)	3.1	3.5	0.2	4.0	3.9	0.7
History of stroke (%)	1.0	0.1	*< 0.001*	1.0	0.0	*< 0.001*
Poor functional status (%)	3.0	1.7	*< 0.001*	2.5	1.5	*< 0.001*
Use of antihypertensives (%)	37.4	36.4	0.2	44.1	43.5	0.2
Disseminated cancer (%)	4.8	8.1	*< 0.001*	4.3	6.3	*< 0.001*
Recent large weight loss[b] (%)	6.2	5.7	0.2	4.7	4.3	0.07
Recent radiotherapy (%)	14.1	1.1	*< 0.001*	8.6	0.7	*< 0.001*
Hematocrit, median (IQR)	38.3 (34.9–41.3)	38.7 (34.1–41.8)	*< 0.001*	39.1 (35.7–42.1)	39.6 (36.0–42.6)	*< 0.001*
Leukocyte count, median 1000/dL (IQR)	6.5 (5.1–8.4)	6.5 (5.1–8.3)	0.07	6.7 (5.3–8.4)	6.7 (5.3–8.4)	0.3
Laparoscopic procedure (%)	27.3	35.0	*< 0.001*	41.5	53.0	*< 0.001*
Post-operative diagnosis, overall[a]			*< 0.001*			*< 0.001*
Benign (%)	15.8	13.7	*< 0.001*	38.8	39.9	0.3
Cancer (%)	56.2	59.6	*< 0.001*	50.2	51.3	*0.04*
Inflammatory bowel disease (%)	28.0	26.7	0.1	10.9	9.4	*< 0.001*

[a] By Chi-squared test for heterogeneity

[b] > 10% of body weight

LAR low anterior resection, *IQR* interquartile range, *COPD* history of chronic obstructive pulmonary disease. *P* values are false-discovery-rate-adjusted due to multiple comparisons

Model and the ACS-Calculator (Table 4). Model calibration was first examined using the percent deviance of predicted to observed morbidity. Of the risk models examined, the OIRC's prediction of a 40.4% overall complication rate was closest to the observed rate of 37.0%, with a percent deviance of predicted-to-actual of + 9.2%. As it had for 2005–2011 data,[4] the NSQIP-Model predicted a morbidity rate lower than was observed (20.0% predicted, − 45.9% deviance). The ACS-Calculator performed better, predicting a 25.2% complication rate for − 31.9% deviance. Each model provided reasonable discrimination of higher versus lower risk patients. The ACS-Calculator returned a C-statistic of 0.653 (95% confidence interval, CI 0.645–0.662), while the NSQIP-model reached 0.671 (CI 0.663–0.680). The OIRC provided the best performance of the three with a C-statistic of 0.676 (CI 0.667–0.684). This exceeded its C-statistic of 0.660 in the 2010–2011 training dataset and 0.637 in the 2005–2009 back-validation.[4]

Brier Skill Scores (BSS) for each model were calculated to combine measures of discrimination and calibration. Using this metric, a score of 0 indicates predictive skill equal to assigning the average complication rate to all patients. Numbers greater than 0 indicate better performance, while scores less than zero correspond to predictive skill worse than guessing the average. By this measure, the OIRC performed

Table 3 Selected 30-day postoperative outcomes in cohorts with and without LAR. *P* values represent comparisons between time periods

Outcome	Without LAR			LAR Included		
	2010–2011 $n = 5288$	2012–2015 $n = 16,722$	*P* value	2010–2011 $n = 14,339$	2012–2015 $n = 51,344$	*P* value
Morbidity (overall %)	40.4	37.0	< 0.001	31.2	27.2	< 0.001
Bleeding (%)	16.5	16.4	0.9	11.5	10.2	< 0.001
Superficial wound infection (%)	9.2	7.4	< 0.001	8.3	5.9	< 0.001
Deep wound infection (%)	2.6	3.2	0.049	1.7	1.7	0.9
Organ space wound infection (%)	7.0	7.0	1.0	5.1	5.3	0.4
Urinary tract infection (%)	5.7	4.4	< 0.001	4.3	3.1	< 0.001
Prolonged ventilation (%)	1.6	1.1	0.02	1.3	1.0	0.01
Reintubation (%)	1.8	1.4	0.2	1.5	1.2	0.01
Pneumonia (%)	2.0	2.0	1.0	1.6	1.6	0.9
Sepsis (%)	6.0	5.1	0.06	4.6	3.8	< 0.001
Septic shock (%)	1.4	1.1	0.1	1.1	1.1	0.9
Stroke (%)	0.3	0.2	0.3	0.3	0.2	0.004
Deep vein thrombosis (%)	1.8	1.8	1.0	1.3	1.2	0.4
MI or cardiac arrest (%)	0.9	1.0	0.9	1.0	0.8	0.2
Reoperation (%)	6.8	6.9	0.9	5.4	5.5	0.9
Death (%)	1.2	1.0	0.3	1.1	0.9	0.007
Readmission within 30 days (%)	16.6[a]	17.2	0.7	13.6[a]	12.4	0.01
Length of stay, median days (IQR)	7.0 (5.0–10.0)	6.0 (5.0–10.0)	< 0.001	6.0 (4.0–9.0)	5.0 (4.0–7.2)	< 0.001

[a] Not recorded in 2010

LAR low anterior resection, *MI* myocardial infarction, *IQR* interquartile range

P values are false-discovery-rate adjusted

well with BSS of + 0.089, as compared to − 0.047 for the NSQIP-Model and + 0.005 for the ACS-Calculator. When plotted as a density function, the OIRC predictions show a wide range of risk estimates roughly centered around the true complication rate, whereas most predictions of the other two tools cluster below the observed complication rate (Fig. 1). From this, we conclude that the OIRC is valid for non-LAR proctectomy risk estimation.

**Inclusion of LAR** We next hypothesized that the same predictors of complications in the OIRC would accurately predict risk in LAR. To test this, LAR procedures from 2010 to 2011 ($n = 9,051$) were included with the original training dataset, and the model was regenerated to produce the Updated Iowa Rectal Surgery Risk Calculator. The UIRC performed well in the training dataset, predicting a mean complication rate equal to that observed (31.2%, 0% deviance), achieved a C-statistic of 0.696 (CI 0.687–0.705), and demonstrated a Brier Skill Score of + 0.107.

It remained possible that inclusion of a large number of LAR procedures might cause the lower morbidity profile of LAR to overpower other proctectomy procedures for which

Table 4 Model performance in the full 2012–2015 validation cohort ($n = 51,344$)

	UIRC	NSQIP-Model	ACS-Calculator
Actual morbidity (%)	27.2	27.2	27.2
Predicted morbidity (%)	29.3	16.1	19.5
Deviance, mean predicted-to-actual morbidity (%)	+ 7.9	− 40.8	− 28.1
C-statistic[a]	0.706	0.703	0.693
95% confidence interval	0.701–0.711	0.698–0.708	0.688–0.698
Brier Skill Score	0.114	0.026	0.056

[a] All between-group comparisons statistically significant at false discovery rate adjusted *P* < 0.05 by DeLong test

UIRC Updated Iowa Rectal Surgery Risk Calculator, *NSQIP* National Surgical Quality Improvement Project, *ACS* American College of Surgeons

Distribution of Morbidity Predictions in Non–LAR Proctectomy 2012–15
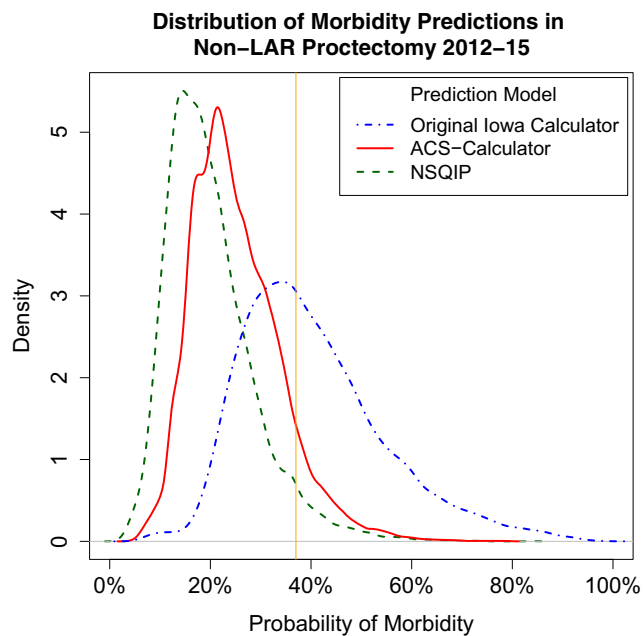
Fig. 1 Density plot showing the distribution of 2012–2015 predicted morbidity probabilities (*x-axis*) in 16,722 non-LAR proctectomy patients. The Original Iowa Calculator (*dashed and dotted blue line*) includes a wider range of predictions and centers closer to the true 37.0% morbidity rate (*vertical orange line*), than the predictions of the ACS-Calculator (*solid red line*) or NSQIP morbidity model (*dashed green line*)

the calculator was initially developed. Examination of the Original and Updated Calculators' Odds Ratios (OR) revealed that incorporating LAR resulted in only small changes in ORs for many predictors, but larger changes for some (Table S4). The OR for preoperative hematocrit remained 0.94, as it was in the OIRC, for example, but the OR for history of stroke decreased from 2.01 in the OIRC to 1.58 in the Updated Calculator. Odds ratios of the procedure term for LAR and laparoscopic LAR were 0.29 and 0.52 (compared to the reference of 1.00 for APR). Much of the different risk profile of LAR might therefore be accounted for by this procedure term, but to verify that changes in the UIRC's odds ratios for other risk factors would not adversely affect its performance in non-LAR proctectomies, predictions from the Original and Updated Calculators in 2010–2011 non-LAR proctectomies were compared. This revealed similar performance by C-statistic (0.657 with UIRC vs. 0.659 by OIRC, $P = 0.09$) and percent deviance (0% for both), indicating that the UIRC performed well in non-LAR proctectomy patients.

Prospective Validation of UIRC As models are expected to perform well when tested against the same data used to develop them, we next performed a prospective validation of the Updated Iowa Risk Calculator using independent 2012–2015 proctectomy data including LARs ($n = 51,344$). In the complete prospective validation dataset, the UIRC was well-calibrated, predicting an average

complication risk of 29.3%, a + 7.9% deviance from the actual complication rate of 27.2%. Discrimination of higher and lower risk patients was better in the validation than training dataset, with a C-statistic of 0.706 (CI 0.701–0.711). The Brier Skill Score was + 0.114.

In the validation cohort, the Original and Updated Iowa Calculators' performance was examined in LAR and non-LAR subgroups. In the subgroup of 2012–2015 non-LAR proctectomies, the UIRC performed better than the OIRC in terms of discrimination (C-statistic 0.679 vs. 0.676, $P < 0.001$) and calibration (+ 8.5 vs. + 9.2% deviance). The distribution of predicted risks was similar between the two models and centered around the observed non-LAR proctectomy morbidity risk of 37.0% (Figure S1). When considering only LARs in 2012–2015, the Updated Calculator returned a C-statistic and percent deviance of 0.689 and + 7.5%. These results show that broadening the scope of the model to include LAR did not compromise performance for non-LAR proctectomies.

Comparison to Other Prediction Tools In the complete 2012–2015 cohort, the NSQIP General/Vascular Surgery Morbidity Model underestimated risk, predicting an average morbidity of 16.1%, with deviance of − 40.8% (Table 4). The ACS-Calculator was closer, predicting 19.5% morbidity, which deviated from the true complication rate by − 28.1%. Both performed well in discriminating patients at greater and lesser risk with C-statistics of 0.703 and 0.693, respectively. Despite small absolute differences, each C-statistic was significantly different from the others by DeLong's test (adjusted $P < 0.05$ for all between-group comparisons). The NSQIP-Model and ACS-Calculator both recorded Brier Skill Scores above zero. In each of these metrics, the UIRC's performance exceeded the other risk prediction tools. Examination of the probability predictions' density plots again showed most predictions from the NSQIP-Model and ACS-Calculator fell to the left of the true 27.2% morbidity rate, while the UIRC predictions distributed more evenly above and below this level (Fig. 2).

To determine how calibration varied across procedures, predictions by the NSQIP-Model, ACS-Calculator, and UIRC were compared by procedure group. Of the 18 included procedure types, the UIRC returned average predictions within 20% of actual morbidity for 13 of 18 procedure types, representing 93% of total patients (Figure S2A-D). ACS-Calculator predictions fell within 20% of actual morbidity for 3 of 18 procedure categories, representing 44% of patients. The NSQIP-model's predictions were below the − 20% bound for all procedure groups.

Calibration of each set of predictions was examined graphically by ordering the predicted morbidity probabilities and comparing the true and predicted morbidity rates of 10 approximately equally sized groups (Fig. 3). By this analysis, the UIRC showed excellent calibration with mean predicted
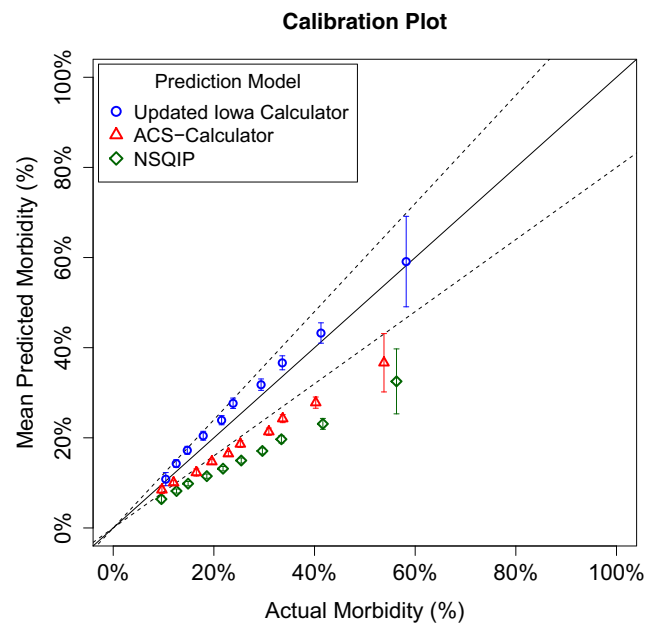
Distribution of Morbidity Predictions in 2012–15 for Proctectomy Including LAR
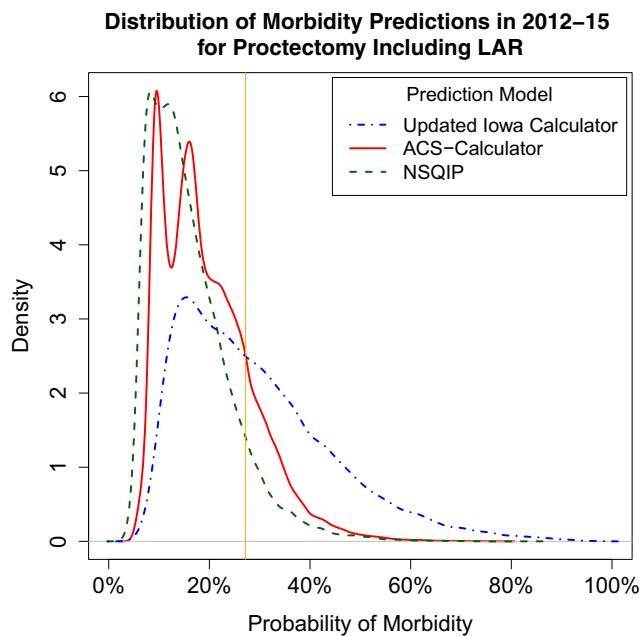
Fig. 2 Density plot showing the distribution of all 2012–2015 (*n* = 51,344 with LAR included) proctectomy morbidity predictions (*x-axis*) by the Updated Iowa Calculator (*dashed and dotted blue line*), ACS-Calculator (*solid red line*), and NSQIP morbidity model (*dashed green line*). The Updated Iowa Risk Calculator predictions more closely approximate the true 27.2% morbidity rate (*vertical orange line*), while the other models' predictions skew to the left of the actual complication probability

Calibration Plot

Fig. 3 Calibration plot of 2012–2015 (*n* = 51,344) actual versus mean predicted morbidity for the Updated Iowa Risk Calculator (*blue circles*), the ACS-calculator (*red triangles*), and the NSQIP-Model (*green diamonds*). For each risk tool, predictions were ordered and then divided into 9 groups of 5000 morbidity predictions (with the 10th containing 6344). For each group, the actual complication rate (*x-axis*) is plotted against the mean predicted morbidity (*y-axis*). Bars show standard deviation of predicted morbidity for each group. Perfect prediction is indicated by the *45° black line*, while the *dashed black lines* show ± 20%. All 10 UIRC groups fall within this 20% error bound. The ACS-Calculator placed 2/10 groups within the bound, while all 10 NSQIP-Model prediction groups fell outside of this limit

morbidity in all 10 subgroups falling within 20% of those groups' true complication rate. For the ACS-Calculator, 2/10 groups were within this limit and 0/10 NSQIP-Model prediction groups were well-calibrated. The Original and Updated Iowa Risk Calculators, including their complete risk prediction equations are available online at http://apps.cromwell.lab.uiowa.edu/calculator/rectalcalc.html.

Discussion

Despite the known tendency of clinical risk models to perform better when tested against the same data used to develop them, systematic reviews of published risk calculators conclude that for most, external validation is reported inappropriately or not at all.[6,15,16] Serious limitations in the development and testing of clinical risk prediction models have led expert consensus panels to conclude that most reported risk calculators "should not be used," and to advance standards for new prediction tools.[6,17] While adhering to these guidelines, the present study found that the OIRC returned valid morbidity estimates for patients undergoing proctectomy procedures in 2012–2015. Using the same predictor variables for 2010–2011 patients, an Updated Calculator also performed well on prospective validation with independent 2012–2015 data, providing better risk discrimination than in the training dataset with good calibration. Risk estimates from both the OIRC and UIRC were

more accurate for their respective populations than the estimates given by the NSQIP-Model and the ACS Online Risk Calculator. The Iowa Rectal Surgery Risk Calculators therefore provide clinically meaningful risk estimates to inform actual clinical practice.

The UIRC extends the scope of the OIRC to include LAR patients. Leaving out LAR from the OIRC despite it representing the most frequently performed proctectomy procedure was not specifically intended. Rather, OIRC-included procedures were those NSQIP placed in its "proctectomy basket," whereas NSQIP grouped LAR with colectomies.[4] The lower LAR complication rate (25.9% in 2010–2011), which more closely matches that of colectomy (24.4% reported in 2009–2012),[7] as compared to non-LAR proctectomy (40.4% in 2010–11), seems to justify such a grouping and exclusion of LAR from a higher-risk proctectomy calculator. Further justification comes from the nature of operations constituting LAR. NSQIP does not record how much rectum is removed in an LAR. Rather, application of an LAR CPT code requires only that a coloproctostomy be performed. Therefore, procedures coded as LAR range from definite proctectomies such as total mesorectal excision with ultralow anastomosis, to procedures more closely resembling colectomies, such as

sigmoidectomy for diverticulitis with high coloproctostomy. That 33% of LARs in the present 2012–2015 cohort had "diverticulitis" as the primary diagnosis suggests that many NSQIP LARs are the latter. An LAR case-mix that skews towards such operations, with less-extensive pelvic dissection and more favorable anastomotic geometry, could explain the overall lower LAR complication rate and more colectomy-like risk profile. Still, because many LARs involve significant rectal resection and because including LAR in the model did not degrade its performance in non-LAR proctectomy, we believe that inclusion of LAR in the UIRC represents an improvement over the Original Calculator.

Differences existed in preoperative characteristics and patient outcomes between the 2010–2011 and 2012–2015 time periods. Many of these were statistically significant due more to large cohort size than actual clinically meaningful differences. Larger preoperative demographic changes worth noting include higher rates of disseminated cancer (from 4.8 to 8.1% of patients in non-LAR proctectomies, a 69% relative increase). This could reflect either greater efforts to ensure charting of a strong predictor of higher complication rates for risk adjustment purposes, or greater willingness of surgeons to operate on patients with metastatic rectal cancer. In a review of 1600 hepatectomy patients with colorectal liver metastases, including 561 with rectal primaries, recent median disease-specific survival improved to 64 months from 43 months in the era prior to modern chemotherapy.[18] Indeed, more effective chemotherapy and strong survival results for hepatic metastastectomy support that selected patients with metastatic rectal cancer should receive operations.

Another important change in preoperative factors was the decrease in recorded history of stroke or radiation due to NSQIP no longer recording these variables. A valid model accounts for variation in patient characteristics, but when predictors are missing entirely, factors distinguishing higher and lower risk individuals are lost. Stroke strongly predicted complications in both the OIRC and UIRC, and omission of this data point, which required treating it as "not present" in later patients, decreases the calculators' predictive abilities. Few patients (1% in 2010–2011) had a history of stroke, mitigating this impact. Potentially more important was loss of information about radiation treatment, which correlated with complication risk in the OIRC,[4] and which many rectal cancer patients receive (14.1% of all non-LAR proctectomy patients in 2010–2011). Still, despite these changes in data collection, the OIRC and UIRC still provided excellent discrimination and calibration in morbidity prediction.

Overall complications showed a substantial relative decrease of approximately 9% in both non-LAR and LAR-included cohorts between 2010–2011 and 2012–2015. Lower rates of superficial wound infections (SSI), urinary tract infection (UTI), prolonged ventilation, and sepsis led this trend. The influence of colorectal surgery subspecialty training

and surgeon case volumes on these outcomes cannot be determined from these data as NSQIP does not record these factors (Supplement). While NSQIP data do not prove exact causes for this change, large-scale efforts targeting exactly these complications have been implemented during the years included in this study, as payors increasingly refuse payment for hospital-acquired infections such as SSI, catheter-associated UTI, and pneumonia.[19–21] Increased adherence to recommendations for preventing SSI, UTI, and DVT in colorectal operations has been shown to produce detectable reductions in these complications.[22] Improved administration of preoperative antibiotics, resurgence in the use of bowel preparation, and early urinary catheter removal have each been associated with lower morbidity and potentially positively impacted these complications during the study period.[20,23–25] Relative contributions of different aspects of prevention to falling complication rates deserve further study, but irrespective of precise causes, that SSI decreased from 11.0% in 2005–2009 non-LAR proctectomy,[4] to 9.2% in 2010–2011, to 7.4% in 2012–2015 suggests real progress in reducing proctectomy morbidity.

To assess the strength of different prediction models for estimating proctectomy risk, this study evaluated discrimination and calibration. Discrimination reflects each model's ability to accurately rank individuals at higher and lower risk. Calibration describes agreement between the level of risk predicted to that observed.[4,6] As commented upon previously,[4] the C-statistic accurately evaluates models' discrimination, but its dependence on the magnitude and dispersion of underlying risk in the evaluated cohort complicates interpretation.[17,26,27] The UIRC illustrates this. In the complete 2012–2015 cohort, the UIRC recorded a much higher C-statistic than in either the LAR or non-LAR subpopulations (0.706 vs. 0.689 and 0.676). Due to its lower inherent morbidity, adding LAR to non-LAR proctectomies broadens the cohort's risk distribution. Most LARs will then be correctly assigned lower risk than non-LARs, and the C-statistic improves compared to when considering the more closely clustered subpopulation predictions. The true event rate also affects the maximum achievable C-statistic, and calculators for rare events, such as mortality, therefore routinely show near-perfect C-statistics, whereas more common events pose greater difficulties for discrimination.[4,7,26] In light of the observed 2012–2015 morbidity of 27.7%, discrimination by all evaluated models in the 2012–2015 complete cohort was strong. The ACS-Calculator C-statistic for 2012–2015 non-LAR proctectomy of 0.653 closely matched the 0.656 reported for the ACS-Calculator in 2010–2014 proctectomy data.[28]

Of the calculators, the UIRC performed best with a C-statistic of 0.706.

With good discrimination across the board, there were greater differences in calibration, with the Iowa Calculators standing out against the other models. As in our previous study, we elected to forgo Hosmer-Lemeshow (HL) testing

as this test is designed for smaller sample sizes, and is difficult to interpret in a cohort of > 50,000 records.[4,17,29,30] We instead present prediction density diagrams (Figs. 1 and 2) and calibration plots showing the agreement of predicted to actual results in risk subgroups (Fig. 3). These graphics provide more information and support easy comparison of the different models.[27] While the other risk tools showed good discrimination and linear increases in risk across calibration subgroups, deviance in predicted to observed morbidity for these groups was at best − 12.5% and more than − 25% for most.

Assessment of a risk calculator's ability depends on its intended application, especially in light of trade-offs between discrimination and calibration.[26] A calculator intended for risk adjustment across large patient groups needs not be well-calibrated. If it correctly discriminates higher from lower risk patients, risk adjustment may be performed based on patients' risk relative to others without worrying about the magnitude of predicted risk. However, if a risk calculator is to support informed consent and influence treatment decisions for individuals, such as whether to have surgery at all, whether to undergo a simpler or more complex operation, or whether to undergo neoadjuvant as opposed to adjuvant chemotherapy based on the likelihood of postoperative complications,[31] it is essential that the calculator be well-calibrated. In this application, the UIRC provides the most accurate morbidity estimates of tested tools, which could help support patient and surgeon decisions.

Another advantage of the UIRC for proctectomy concerns accessibility. For both the OIRC and UIRC, the available online calculators completely specify the model using open-source code. The model can therefore be examined, tested, and applied to other data, and could be incorporated into electronic health records to provide automatic risk estimates for preoperative patients. Additionally, its conclusions can be tested by others in populations distinct from NSQIP to determine its external validity. The ACS-Calculator's models remain proprietary and inaccessible for external validation or research.[32,33]

Applying the ACS-Calculator to large numbers of patients requires submission of each record through the ACS-Calculator website. As a result, external assessments of the ACS-Calculator's performance have small sample sizes relative to the entire NSQIP population and often represent single-institution data ($n = 95$–1480 in 15 studies listed in Supplement). Many of these have questioned the ACS-Calculator's morbidity predictions,[27] and together suggest that individual risk calculators tailored to specific procedure types could outperform an omnibus calculator. These studies' limited samples, however, make it difficult to determine whether the ACS-Calculator predictions were poorly-calibrated, or the included patients simply differed from the overall NSQIP risk profile.[27,33]

The present study overcame this limitation by evaluating NSQIP records representing the same inputs and endpoints as the ACS-Calculator. A web-script automated NSQIP record submission, allowing tens of thousands of ACS-Calculator predictions to be obtained. With this large and robust sample, our results likewise found poor ACS-Calculator calibration in a complex and high-risk class of procedures. A limitation of this conclusion is that since obtaining ACS-Calculator estimates, the ACS-Calculator has been updated with specific intent to improve calibration.[28] How the current version differs from the one tested here remains unknown. Due to the addition of a feature on the website preventing automated record submission, we are unable to update our analysis. These results support further efforts to validate ACS-Calculator predictions for additional endpoints and surgical subgroups.

Conclusion

The OIRC and UIRC returned accurate proctectomy morbidity predictions when prospectively validated against independent 2012–2015 data. Discrimination and calibration of these predictions exceeded that of the NSQIP-Model and ACS-Calculator, which had good discrimination but significantly underestimated actual morbidity rates. The Iowa Calculators and underlying models are available at http://apps.cromwell.lab.uiowa.edu/calculator/rectalcalc.html and can be used to inform decisions about patient care.

Author's Contributions Sherman: study design, data analysis, data interpretation, writing, critical revision; Hrabe: study design, data interpretation, writing, critical revision; Huang: data interpretation and critical revision; Cromwell: study design, data interpretation, critical revision; Byrn: study design, data interpretation, critical revision.

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflicts of interest.

References

Bolded Names Indicate Co-First Authorship.

1. Balentine, C. J. *et al.* Waist circumference predicts increased complications in rectal cancer surgery. *J. Gastrointest. Surg.* 2010 Nov; **14**(11):1669–1679.
2. Ballian, N. et al. Body mass index does not affect postoperative morbidity and oncologic outcomes of total mesorectal excision for rectal adenocarcinoma. *Ann. Surg. Oncol.* 2010 Jun;**17**(6): 1606–1613.

3. Canedo, J. A., Pinto, R. A., McLemore, E. C., Rosen, L. & Wexner, S. D. Restorative proctectomy with ileal pouch-anal anastomosis in obese patients. *Dis. Colon rectum* 2010 Jul; **53**(7):1030–1034.

4. **Hrabe, J. E., Sherman, S. K.**, Charlton, M. E., Cromwell, J. W. & Byrn, J. C. Effect of BMI on outcomes in proctectomy. *Dis. Colon rectum* 2014 May; **57**(5):608–615.

5. Sherman, S. K., Hrabe, J. E., Charlton, M. E., Cromwell, J. W. & Byrn, J. C. Development of an improved risk calculator for complications in proctectomy. *J. Gastrointest. Surg.* 2014 May; **18**(5): 986–994.

6. Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. *Ann. Intern. Med.* 2015 Jan; **162**(1):55–63.

7. Bilimoria, K. Y. *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 Nov; **217**(5):833–842.

8. User Guide for the 2015 ACS-NSQIP Participant Use Data File (PUF). American College of Surgeons. https://www.facs.org/~/media/files/quality%20programs/nsqip/nsqip_puf_user_guide_2015.ashx (Oct. 2016).

9. ACS-NSQIP Surgical Risk Calculator. American College of Surgeons. https://riskcalculator.facs.org/RiskCalculator/, Accessed February–March, 2017.

10. Hochberg, Y. & Benjamini, Y. More powerful procedures for multiple significance testing. *Stat. Med.* 1990 Jul; **9**(7):811–818.

11. Robin, X. *et al*. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 2011 Mar; **12**:77.

12. Hanley, J. A. & McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982 Apr; **143**(1):29–36.

13. Brier, G. Verification of forecasts expressed in terms of probability. *Monthly Weather Review.* 1950 Jan; **78**(1):1–3.

14. Wilks, D. S. Statistical Methods in the Atmospheric Sciences. Academic Press, San Diego, CA, Mar. 1995.

15. Bouwmeester, W. *et al*. Reporting and methods in clinical prediction research: a systematic review. *PLoS Med.* 2012; **9**(5):1–12.

16. Collins, G. S. et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC Med. Res. Methodol.* 2014 Mar; **14**:40. 17.

17. Moons, K. G. *et al*. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann. Intern. Med*. 2015 Jan; **162**(1): 1–73.

18. House, M. G. *et al*. Survival after hepatic resection for metastatic colorectal cancer: trends in outcomes for 1,600 patients during two decades at a single institution. *J. Am. Coll. Surg.* 2010 May; **210**(5): 744–752.

19. Molena, D., Mungo, B., Stem, M., Feinberg, R. L. & Lidor, A. O. Prevalence, impact, and risk factors for hospital-acquired conditions after major surgical resection for cancer: a NSQIP analysis. *J. Gastrointest. Surg*. 2015 Jan; **19**(1):142–151.

20. Nygren, J. *et al*. Guidelines for perioperative care in elective rectal/pelvic surgery: Enhanced Recovery After Surgery (ERAS) Society recommendations. *World J. Surg.* 2013 Feb; **37**(2):285–305.

21. Rosenberger, L. H., Politano, A. D. & Sawyer, R. G. The surgical care improvement project and prevention of post-operative infection, including surgical site infection. *Surg. Infect.* (Larchmt) 2011 Jun; **12**(3):163–168.

22. Arriaga, A. F. *et al*. The better colectomy project: association of evidence-based best-practice adherence rates to outcomes in colorectal surgery. *Ann. Surg.* 2009 Oct; **250**(4):507–513.

23. Ohman, K. A. *et al*. Combination of Oral Antibiotics and Mechanical Bowel Preparation Reduces Surgical Site Infection in Colorectal Surgery. *J. Am. Coll. Surg.* 2017 Oct; **225**(4):465–471.

24. Cannon, J. A. *et al*. Preoperative oral antibiotics reduce surgical site infection following elective colorectal resections. *Dis. Colon rectum* 2012 Nov; **55**(11):1160–1166.

25. Connolly, T. M., Foppa, C., Kazi, E., Denoya, P. I. & Bergamaschi, R. Impact of a surgical site infection reduction strategy after colorectal resection. *Colorectal Dis.* 2016 Sep; **18**(9):910–918.

26. Cook, N. R. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation* 2007 Feb; **115**(7), 928–935.

27. Cohen, M. E., Liu, Y., Ko, C. Y. & Hall, B. L. An Examination of American College of Surgeons NSQIP Surgical Risk Calculator Accuracy. *J. Am. Coll. Surg.* 2017 May; **224**(5):787–795.

28. Liu, Y., Cohen, M. E., Hall, B. L., Ko, C. Y. & Bilimoria, K. Y. Evaluation and Enhancement of Calibration in the American College of Surgeons NSQIP Surgical Risk Calculator. *J. Am. Coll. Surg*. 2016 Aug; **223**(2):231–239.

29. Kramer, A. A. & Zimmerman, J. E. Assessing the calibration of mortality benchmarks in critical care: The Hosmer-Lemeshow test revisited. *Crit. Care Med.* 2007 Sep; **35**(9):2052–2056. 30.

30. Paul, P., Pennell, M. L. & Lemeshow, S. Standardizing the power of the Hosmer-Lemeshow goodness of fit test in large data sets. *Stat. Med.* 2013 Jan; **32**(1):67–80.

31. Paruch, J. L., Ko, C. Y. & Bilimoria, K. Y. An opportunity to improve informed consent and shared decision making: the role of the ACS NSQIP Surgical Risk Calculator in oncology. *Ann. Surg. Oncol.* 2014 Jan; **21**(1):5–7.

32. Wanderer, J. P. & Ehrenfeld, J. M. Toward External Validation and Routine Clinical Use of the American College of Surgeons NSQIP Surgical Risk Calculator. *J. Am. Coll. Surg.* 2016 Oct; **223**(4):674.

33. Liu, Y., Cohen, M. E., Ko, C. Y., Bilimoria, K. Y. & Hall, B. L. Considerations in Releasing Equations for the American College of Surgeons NSQIP Surgical Risk Calculator: In Reply to Wanderer and Ehrenfeld. *J. Am. Coll. Surg.* 2016 Oct; **223**(4):674–675.