Performance, Revision, and Extension of the National Nosocomial Infections Surveillance System's Risk Index in Brazilian Hospitals

Fernando Martín Biscione, MD, PhD;[1] Renato Camargos Couto, MD, PhD;[1] Tânia M. G. Pedrosa, MD, PhD[1]

OBJECTIVE. To assess the benefit of using procedure-specific alternative cutoff points for National Nosocomial Infections Surveillance (NNIS) risk index variables and of extending surgical site infection (SSI) risk prediction models with a postdischarge surveillance indicator.

DESIGN. Open, retrospective, validation cohort study.

SETTING. Five private, nonuniversity Brazilian hospitals.

PATIENTS. Consecutive inpatients operated on between January 1993 and May 2006 (other operations of the genitourinary system [n = 20,723], integumentary system [n = 12,408], or musculoskeletal system [n = 15,714] and abdominal hysterectomy [n = 11,847]).

METHODS. For each procedure category, development and validation samples were defined nonrandomly. In the development samples, alternative SSI prognostic scores were constructed using logistic regression: (i) alternative NNIS scores used NNIS risk index covariates and cutoff points but locally derived SSI risk strata and rates, (ii) revised scores used procedure-specific alternative cutoff points, and (iii) extended scores expanded revised scores with a postdischarge surveillance indicator. Performances were compared in the validation samples using calibration, discrimination, and overall performance measures.

RESULTS. The NNIS risk index showed low discrimination, inadequate calibration, and predictions with high variability. The most consistent advantage of alternative NNIS scores was regarding calibration (prevalence and dispersion components). Revised scores performed slightly better than the NNIS risk index for most procedures and measures, mainly in calibration. Extended scores clearly performed better than the NNIS risk index, irrespective of the measure or operative procedure.

CONCLUSIONS. Locally derived SSI risk strata and rates improved the NNIS risk index's calibration. Alternative cutoff points further improved the specification of the intrinsic SSI risk component. Controlling for incomplete postdischarge SSI surveillance provided consistently more accurate SSI risk adjustment.

Infect Control Hosp Epidemiol 2012;33(2):124-134

The National Nosocomial Infections Surveillance (NNIS) system's risk index is the most widely used methodology worldwide for adjusting the risk of surgical site infection (SSI) in daily surveillance.[1] Unlike its immediate precursor, the Study on the Efficacy of Nosocomial Infection Control (SENIC) project risk index, the NNIS risk index was not developed with a multivariate modeling technique.[2] Indeed, the NNIS risk index is best viewed as a convenience adaptation of the SENIC risk index, one aimed at making the basic conclusions already drawn by the SENIC risk index more applicable in routine practice.[1]

The value of the NNIS risk index as a benchmarking tool has been criticized in many settings.[3-5] One of the main purported advantages of the NNIS risk index over the SENIC risk index is that it is procedure specific—that is, the risk of SSI is adjusted within predefined operative procedure categories.[1] However, despite being procedure specific, the categorization of each variable that composes the NNIS risk index is the same irrespective of the operative procedure considered and was arbitrarily defined (see "Methods").[1] In addition, the failure of the index to account for incomplete postdischarge follow-up has been a major concern.[6-8] As many as 12% to 84% of SSIs are detected after patients are discharged from the hospital, and an ever-increasing proportion of SSIs manifests after discharge, driven by the progressively shorter length of postoperative hospital stay of surgical patients.[9] The Centers for Disease Control and Prevention currently recommends the use of postdischarge surveillance to detect SSIs after operative procedures.[10] The NNIS risk index, however, does not explicitly recognize the problem of incomplete postdischarge surveillance.[11,12]

In this study, we sought to explore whether using alternative cutoff points for the variables in the NNIS risk index would improve its predictive accuracy. We also aimed at investigating whether using a postdischarge surveillance indicator in SSI risk prediction scores would provide any benefit in terms of predictive ability.

Affiliation: 1. Health Sciences and Tropical Medicine Postgraduate Course, Minas Gerais Federal University School of Medicine, Belo Horizonte, Minas Gerais, Brazil.

Received April 21, 2011; accepted October 17, 2011; electronically published December 23, 2011.

© 2011 by The Society for Healthcare Epidemiology of America. All rights reserved. 0899-823X/2012/3302-0005$15.00. DOI: 10.1086/663702

https://doi.org/10.1086/663702 Published online by Cambridge University Press

Box 1: Measures of Model Performance

Area under the receiver operating characteristic curve (A_{ROC}). A_{ROC} provides an indication of how well a model can discriminate between patients who develop SSI and patients who do not.[25,26] The A_{ROC} is the probability of concordance between outcomes and predictions. For binary outcomes (eg, SSI), this is the probability that a randomly chosen individual with an SSI will have a higher predicted probability than a randomly chosen individual without an SSI. An A_{ROC} of 0.5 indicates random predictions, whereas values higher and significantly different from 0.5 indicate discriminatory power, with 1 indicating perfect discrimination. Comparisons between A_{ROC} values were made using DeLong's test, and significance was adjusted for multiple comparisons using Sidak's correction.

Cox's calibration regression. Calibration refers to the accuracy of predictions compared with the observed data. To quantify the degree of miscalibration between observed and predicted probabilities, Cox[27] proposed fitting an ordinary logistic regression model in the validation sample (ie, with the observed outcomes as the dependent variable) and the predicted risk as the only independent variable. This model has the general form observed log odds = $\alpha + \beta \times$ predicted log odds, so that α and β reflect the degree of agreement between observed and predicted risk. For a model with perfect calibration, $\alpha = 0$ and $\beta = 1$. The value of β represents the degree of variability in the predicted probabilities. If $0 < \beta < 1$, predicted probabilities are too extreme (ie, predicted probabilities are lower for low-risk patients, higher for high-risk patients, or both). If $\beta > 1$, predicted probabilities show the right general pattern of variation but do not vary enough. In general, $\alpha > 0$ when predicted probabilities are globally low and $\alpha < 0$ when they are globally high, compared with the observed risk. Likelihood ratio statistics were used to test each type of unreliability, with P values less than .05 indicating a significant lack of calibration:

- Significant overall unreliability (H_0: $\alpha = 0, \beta = 1$);
- Significant unreliability due to overall prevalence error (H_0: $\alpha = 0 | \beta = 1$); and
- Significant unreliability due to incorrect degree of variation, given prevalence correction (H_0: $\beta = 1 | \alpha$).

Goodman-Kruskal statistic (G). G is a nonparametric correlation coefficient for ordinal data that measures the strength of the association between 2 cross-classified ordered polytomies, in our case the SSI risk stratum as defined by the scores and the actual SSI status.[28] G ranges from -1 (perfect negative correlation) to $+1$ (perfect positive correlation), with 0 indicating no significant correlation.

Yates's decomposition of Brier's score. Yates[29] demonstrated that Brier's score, which is the mean square error between outcomes and predictions, can be decomposed into informative components:

- The excess variance of predictions (V_{exc}/V_{min}), which represents the degree of unnecessary variation in the predictions. It is calculated by decomposing the total variance of the predictions, $V(p)$, as $V_{min} + V_{exc}$, where V_{min} represents the minimum variance possible for predictions that would be just as good as the actual predictions, and V_{exc} represents the excess variance of the predictions above this minimum. For perfect predictions, $V_{exc} = V_{min}$, so that $V_{exc}/V_{min} = 0$.
- The covariance of outcome and prediction ($\mathrm{Cov}(Y, p)$), which is a measure of how accurately the predictions correspond to the outcomes and is closely related to discrimination. For perfect predictions, $\mathrm{Cov}(Y, p)$ equals the variance of the observed outcome ($V(Y)$), so that $\mathrm{Cov}(Y, p)/V(Y) = 1$.

Trend across ordered groups (Cuzick's test). This is a nonparametric test that measures the trend across ordered groups, in our case the risk of SSI across risk scores strata.[30] Test statistics above $|1.96|$ and associated P values less than .05 indicate a significant trend. The higher the test statistic (and the lower the associated P values), the higher the trend.

Model χ^2. This statistic is a measure of overall model performance.[31] It is based on a likelihood ratio test and measures how better it is to use the probabilities predicted by the risk scores than simply forecasting the mean risk of SSI for every patient. Test statistics above 3.84 and associated P values less than .05 indicate that using the model is better than predicting the mean SSI risk. The higher the test statistic (and the lower the associated P values), the better the model performance.

METHODS

Patients and Setting

This study was performed using data collected from 60,692 operative procedures belonging to 1 of 4 selected NNIS operative procedures categories:[13] other operations of the genitourinary system (OGU; $n = 20{,}723$), abdominal hysterectomy (HYS; $n = 11{,}847$), other operations of the integumentary system (OSK; $n = 12{,}408$), and other operations of the musculoskeletal system (OMS; $n = 15{,}714$). Data were prospectively collected from January 1993 to May 2006 at 5 private, nonuniversity, secondary or tertiary care healthcare facilities (range, 49–230 beds) located in Belo Horizonte, Brazil. These institutions are general acute care hospitals or institutions devoted to the care of women's health.

Data Collection and Definitions

The participating institutions implemented prospective SSI surveillance programs based on the NNIS system's protocols and definitions[9,13] in early 1992. For each patient who underwent an operation, the NNIS system's risk index variables (ie, duration of surgery, American Society of Anesthesiologists' physical status [ASA-PS] score, and wound class) were prospectively recorded shortly after completion of the surgery. All variables were recorded in their original form (ie, in minutes for duration of surgery, 1–5 for ASA-PS score, and

https://doi.org/10.1086/663702 Published online by Cambridge University Press

TABLE 1. Characteristics of Development and Validation Samples and Description of Alternative Risk Indexes and Weights

	OGU		OSK		OMS		HYS	
	Development samples							
Period	1993–2002		1993–2000		1993–2003		1993–2002	
Size	14,506		5,682		11,000		8,293	
No. of SSIs	415		166		259		426	
Measured SSI risk, %	2.86		2.92		2.35		5.14	
Effective PDS, %	62.59		56.44		36.77		69.60	
	Validation samples							
Period	2002–2006		2000–2006		2003–2006		2002–2006	
Size	6,217		6,726		4,714		3,554	
No. of SSI	144		100		109		122	
Measured SSI risk, %	2.32		1.49		2.31		3.43	
Effective PDS, %	68.97		39.21		35.87		71.58	
	Selected cutoff points and weighting for alternative risk indexes							
	Cutoff	Weight[a]	Cutoff	Weight[a]	Cutoff	Weight[a]	Cutoff	Weight[a]
Alternative NNIS risk index								
Surgery length, minutes	>*T* (120)	NA	>*T* (120)	1	>*T* (180)	1	>*T* (120)	1
Wound class	Co/I	NA	Co/I	1	Co/I	1	Co/I	7
ASA-PS score	NS	...	≥3	2	≥3	3	NS	...
Revised risk index								
Surgery length, minutes	60–120	2	61–230	1	>120	NA	>120	1
	>120	4	>230	3				
Wound class	Cl/Co/I	1	CC/Co/I	1	Co/I	NA	Cl/Co/I	3
ASA-PS score	≥2	1	≥2	2	≥2	NA	≥2	1
Extended risk index								
Surgery length, minutes	60–120	1	61–230	1	>120	1	>120	1
	>120	3	>230	2				
Wound class	Cl/Co/I	1	CC/Co/I	1	Co/I	2	Cl/Co/I	3
ASA-PS score	≥2	1	≥2	1	≥2	1	≥2	2
Effective PDS	Present	5	Present	2[b]	Present	3	Present	3

NOTE. ASA-PS, American Society of Anesthesiologist's physical status; CC, clean-contaminated; Cl, clean; Co, contaminated; HYS, abdominal hysterectomy; I, infected; NA, not applicable (logistic coefficients of covariates were very similar, so no weighting was applicable); NS, not selected (variable was not selected in the logistic regression models; see "Risk Score Development"); OGU, other operations of the genitourinary system; OMS, other operations of the musculoskeletal system; OSK, other operations of the integumentary system; PDS, postdischarge surveillance; SSI, surgical site infection.

[a] For unweighted risk indexes, nonreference categories count 1 point. All reference categories count 0 points.

[b] As suggested by the bias-corrected maximum likelihood method (see "Risk Score Development").

clean, clean-contaminated, contaminated, and infected for wound class).

The outcome variable for the NNIS risk index is the occurrence of SSI within 30 days after surgery, and this was also the time span used in this study. The 1992 Centers for Disease Control and Prevention's surveillance criteria for SSI were used as case definitions throughout the study.[9] Case finding included in-hospital and postdischarge surveillance. A detailed description of the surveillance methods used at our institutions can be found elsewhere.[8,14]

Study Design

An open, retrospective, validation cohort study was conducted. Within each operative procedure category, development and validation samples were defined nonrandomly so that they would differ from each other by a systematic characteristic. Because we wanted to test the external validity of our risk models, surgeries performed in more recent years were allocated to the validation samples, which amounts to challenging the historical component (and, to a lesser extent, the methodological component) of external validity.[15,16] The size of the validation samples was defined so that at least 30% of the original sample and at least 100 SSIs would be retained in the validation samples[17,18] and at least 20 SSIs per candidate degree of freedom would be guaranteed for the logistic regression models in the development samples[19] (see "Risk Score Development").

https://doi.org/10.1086/663702 Published online by Cambridge University Press

TABLE 2. National Nosocomial Infections Surveillance (NNIS) and Alternative Risk Indexes in the Validation Samples

	No.[a]	SSI risk, %	
		Observed[b]	Predicted[c]
		Other operations of the genitourinary system	
NNIS risk index			
0	5,725	2.17	0.36
1	467	3.85	0.85
2–3	25	8.00	2.92
Alternative NNIS risk index			
Unweighted			
0	5,788	2.19	2.64
1	420	3.81	4.98
2	9	11.11	15.38
Revised risk index			
Unweighted			
0	1,221	0.98	1.65
1	3,294	2.13	2.58
2	1,450	3.38	3.71
3–4	252	5.16	6.36
Weighted			
0–1	1,677	1.49	1.69
2	2,895	1.97	2.67
3	1,211	3.63	3.42
4	277	3.97	4.88
5–6	157	4.46	6.63
Extended risk index			
Unweighted			
0	483	0	0.46
1	1,728	0.75	1.03
2	2,678	2.61	3.32
3	1,151	4.43	5.09
4–5	177	5.65	7.29
Weighted			
0–1	1,473	0.07	0.48
2–3	399	0.5	1.56
4–5	795	1.89	2.82
6–7	3,246	3.48	3.99
≥8	304	4.28	6.67
		Other operations of the integumentary system	
NNIS risk index			
0–3	6,726	1.49	1.29
Alternative NNIS risk index			
Unweighted			
0	5,028	1.03	2.39
1	1,519	2.37	3.61
2–3	179	6.70	10.68
Weighted			
0	5,028	1.03	2.39
1	1,400	2.43	3.56
2	141	3.55	5.47
3–4	157	5.73	13.56
Revised risk index			
Unweighted			
0	2,642	0.57	1.59
1	2,726	1.87	2.44
2	1,098	2.46	4.46
3–4	260	2.69	10.97
Weighted			
0	2,642	0.57	1.59
1	2,253	1.95	2.39
2–3	1,571	2.16	4.28
≥4	260	2.69	10.97

https://doi.org/10.1086/663702 Published online by Cambridge University Press

TABLE 2 (*Continued*)

	No.[a]	SSI risk, %	
		Observed[b]	Predicted[c]
Extended risk index			
Unweighted			
0	1,607	0.19	0.28
1	2,633	0.72	1.53
2	1,845	2.71	3.15
3	534	4.68	7.60
4–5	107	2.80	12.78
Weighted			
0–1	3,205	0.31	0.49
2	1,752	1.03	2.07
3	1,281	3.75	4.20
4	395	5.57	7.48
≥5	93	2.15	12.80
		Other operations of the musculoskeletal system	
NNIS risk index			
0	4,083	1.37	0.63
1	541	7.95	0.94
2–3	90	11.11	1.78
Alternative NNIS risk index			
Unweighted			
0	4,083	1.37	2.06
1	541	7.95	4.50
2–3	90	11.11	4.85
Weighted			
0	4,083	1.37	2.06
1	429	8.16	3.73
≥2	202	8.91	7.14
Revised risk index			
Unweighted			
0	2,620	1.07	1.88
1	1,633	3.06	2.64
2–3	461	6.72	5.88
Extended risk index			
Unweighted			
0	1,675	0.36	0.38
1	1,991	2.11	3.16
≥2	1,048	5.82	5.06
Weighted			
0	1,675	0.36	0.38
1–2	1,255	2.95	2.33
3–4	1,592	3.33	4.36
≥5	192	6.77	5.40
		Abdominal hysterectomy	
NNIS risk index			
0	3,039	3.39	1.36
1	498	3.21	2.32
2–3	17	17.65	5.17
Alternative NNIS risk index			
Unweighted			
0	3,063	3.36	4.86
1–2	491	3.87	6.28
Weighted			
0	3,063	3.36	4.86
1	478	3.77	6.05
>1	13	7.69	25.00
Revised risk index			
Unweighted			
0	2,464	3.00	4.52
1	889	4.39	5.54
2–3	201	4.48	9.55

https://doi.org/10.1086/663702 Published online by Cambridge University Press

TABLE 2 (*Continued*)

	No.[a]	SSI risk, %	
		Observed[b]	Predicted[c]
Weighted			
0	2,464	3.00	4.52
1	738	3.93	5.39
2–3	277	6.14	6.92
4–5	75	2.67	16.77
Extended risk index			
Unweighted			
0	660	0.45	2.51
1	2,079	3.99	4.99
2	682	4.11	6.24
3–4	133	6.02	12.21
Weighted			
0–1	727	0.55	2.57
2–3	2,052	4.04	5.10
4–5	559	4.11	6.13
6	176	5.68	8.97
7–9	40	5.00	20.45

NOTE. NNIS, National Nosocomial Infections Surveillance; SSI, surgical site infection.

[a] Size of the validation samples.

[b] Risk of SSI as observed in the validation samples.

[c] For the alternative risk indexes, this is the risk of SSI as predicted by the development samples. For the NNIS risk index, predicted values for 2004 are shown,[11] but annually updated values in the NNIS system's reports were used for performance assessment (see "Risk Score Performance and Validation").

Risk Score Development

For each operative category, three alternative SSI risk indexes were developed in the development samples, beginning with binary logistic regression models for selection and weighting of covariates. The alternative risk models differed by the candidate explanatory variables considered for inclusion in these models, as follows:

1. The alternative NNIS risk models used the original cutoff points of the NNIS risk index—that is, an ASA-PS score of more than or equal to 3 (vs less than 3), a surgical wound classified as contaminated or infected (vs clean or clean-contaminated), and an operation lasting more than T hours (vs less than T hours), with T representing the approximate 75th percentile of operation length and depending on the operative procedure performed.

2. For the revised risk models, the NNIS risk index variables were recategorized using procedure-specific alternative cutoff points, which were defined using visual inspection of density histograms, contingency table analysis, and decision tree analysis with the exhaustive CHAID (χ^2 automatic interaction detector) algorithm as the growing method.[20] Cutoff points defined in the ordinal sense of the variables were given priority. We constrained up to 2 categories for ASA-PS and wound class and up to 3 categories for surgery length.

3. The extended risk models expanded the revised models to account for the proportion of procedures with incomplete 30-day postdischarge surveillance. To accomplish this task, a postdischarge surveillance indicator was created, which was assigned the value +1 whenever the patient did not develop an SSI during hospitalization and was reached by the postdischarge surveillance efforts (ie, effective postdischarge surveillance) and 0 otherwise. In this way, the model recognizes that patients reached by postdischarge surveillance will obviously be more likely to have an SSI detected and will have a higher measured SSI risk than patients not reached. More details about this strategy can be found elsewhere.[8]

Because the occurrence of an SSI is a rare event, which may bias logistic regression coefficients (β coefficients), we conducted sensitivity analysis by comparing the results of three methods of inference for the β coefficients: the (asymptotic) unconditional maximum likelihood method,[21] the conditional exact method,[22] and the rare-event, finite-sample, bias-corrected maximum likelihood method described in King and Zeng.[23] Manual backward elimination was used for covariate selection, starting with the models with all candidate predictors and retaining in the final models those with a P value less than or equal to .15 in at least 1 exact hypothesis test (ie, conditional probability, conditional score, or exact likelihood ratio).

Unweighted alternative scores were constructed simply by adding up the number of factors retained in the final logistic regression models (and present in the patient at the time of surgery). This summation determined different SSI risk strata in the development samples. As in the NNIS risk index, when SSI rates for adjacent risk strata were not significantly different or returned low absolute frequencies, they were combined into a single risk category. Weighted versions of the scores were similarly constructed by adding up the number of weighted predictors retained in the logistic models. Weighting was accomplished by assigning 1 point to the predictor

https://doi.org/10.1086/663702 Published online by Cambridge University Press

with the lowest β coefficient (β_{min}), dividing the β coefficients of the other predictors by β_{min}, and then rounding to the nearest integer.[24]

Risk Score Performance and Validation

In the validation samples, we calculated the original NNIS risk index and the newly developed alternative risk indexes and compared predicted with observed SSI risk within each risk category. For the original NNIS index, the predicted risk of SSI was that periodically updated in the annual NNIS system's reports. For the alternative risk indexes, the predicted risk of SSI was that observed in the development samples for each risk stratum. Risk scores were evaluated for discriminatory ability, calibration, and overall performance (see Box 1). All analyses were performed using Stata version 9, SPSS version 15.0, and LogXact version 8.0.0. Institutional review board approval was obtained.

RESULTS

Table 1 shows the size, period, and risk of SSI in development and validation samples. The proportion of successful postdischarge surveillance contact in the study period was nearly 55% but varied across procedures. The table also shows the selected cutoff points and weighting of covariates for the alternative risk indexes. No alternative cutoff points were found for wound class in OMS and surgery length in HYS. In the logistic regression models, all three methods of estimation (see "Risk Score Development") suggested the same weighting for the explanatory variables, except for effective postdischarge in OSK. The ASA-PS score was not admitted in the logistic regression models for the alternative NNIS risk index for OGU and HYS, and weighting was not possible for the alternative NNIS index for OGU and the revised index for OSK because the logistic regression coefficients of their covariates were very similar.

Table 2 shows the NNIS index and the alternative indexes in the validation samples and compares observed and predicted risk of SSI. In general, the alternative risk indexes arranged cases across risk categories more evenly than the NNIS index.

Table 3 shows the NNIS risk index and the best-performing alternative indexes in the validation samples. The weighted version of the extended score for OSK showed better performance than the unweighted version for all measures. For the other procedures and scores, however, no consistent benefit of weighting was evident. Unweighted extended scores showed consistently better performance than the NNIS risk index for all procedures and for almost all measures. Revised scores for OGU and OSK also performed better than the NNIS risk index for all measures; for OMS and HYS, revised scores significantly improved calibration. The most evident benefit of the alternative NNIS scores over the NNIS index was in terms of calibration, although it also performed better for most other measures in OSK and OMS. No alternative NNIS or revised risk index performed better than the corresponding extended index.

DISCUSSION

In this study, we sought to explore whether using alternative cutoff points for the variables of the NNIS risk index would improve its predictive accuracy. We also aimed at investigating whether using a postdischarge surveillance indicator would provide any benefit in terms of predictive ability. We are aware of few studies that have attempted to improve the NNIS risk index performance by selecting alternative cutoff points for their explanatory variables.[5,32] Most studies have just focused on defining a locally derived T for surgery length.[32] Others explored more structural changes to the cutoff points,[5] but all failed to demonstrate any significant benefit over the predictive power of the NNIS risk index.

Although a first inspection of the performance measures reported in Table 3 would suggest poor overall performance of NNIS and alternative risk indexes, interpretation of these measures should be approached with much caution. Most previous reports in the literature have judged the performance of NNIS or alternative risk indexes in relation to "perfect performance" values (see Box 1). Because all of these indexes aim at adjusting only for patient- and procedure-related risk factors, extrinsic (ie, quality of care) factors are deliberately excluded from the models, so they will obviously never reach perfect performance.[4] Accordingly, it is more correct to judge a risk index by comparing its performance against that of another risk index applied to the same sample.

An unexpected finding of this study was that weighted scores did not show consistently better performance than their unweighted versions. In the development samples, on the contrary, weighted scores performed better (slightly, but consistently) for all procedures and for most measures (data not shown). The most likely explanation for this phenomenon is the overfitting of the weighted scores to the development samples. Overfitting is the degree to which a prognostic model fits random (ie, nonreproducible) noise rather than real patterns in the data and is more likely to occur when model construction relies heavily on data-driven decisions,[33] of which weighting is an example. Data-driven decisions imply a better fit to the data under study because model construction will make the greatest possible use of any and all idiosyncrasies of those particular data, sometimes producing spurious associations.[33] However, this does not necessarily mean that we learned more about the underlying population. As a consequence, when significant overfitting occurs during model construction, the model will produce overoptimistic predictions in the model-building sample but may not generalize well in a validation sample.[33]

It is remarkable that by entering the NNIS risk index variables in an unconstrained form the exhaustive CHAID growing method spontaneously selected several cutoff points not present in the NNIS risk index (Table 1). This reinforces the

https://doi.org/10.1086/663702 Published online by Cambridge University Press

https://doi.org/10.1086/663702 Published online by Cambridge University Press

TABLE 3. Performance of National Nosocomial Infections Surveillance (NNIS) Risk Index and Selected Extended Alternative Risk Indexes in the Validation Samples by Operative Procedure Category

	OGU risk indexes		OSK risk indexes		OMS risk indexes		HYS risk indexes	
Performance measure	NNIS system	Unweighted extended	NNIS system	Weighted extended	NNIS system	Unweighted extended	NNIS system	Unweighted extended
Model χ^2								
Statistic	6.96	70.17	0.07	92.80	63.77	77.92	1.61	21.10
P	.008	<.001	.785	<.001	<.001	<.001	.201	<.001
A_{ROC}								
Point estimate	0.53	0.68[a]	0.50	0.77[a]	0.68	0.73[b]	0.51	0.59[a]
95% CI	0.52–0.54	0.67–0.69	0.50–0.50	0.76–0.78	0.67–0.70	0.72–0.74	0.49–0.52	0.58–0.61
P	.003	<.001	1.0	<.001	<.001	<.001	.332	<.001
Goodman-Kruskal correlation (*G*)								
Statistic	0.314	0.518	NA	0.665	0.718	0.662	0.055	0.340
P	.037	<.001		<.001	<.001	<.001	.678	<.001
Trend across ordered groups (Cuzick's test)								
Statistic	2.91	7.82	NA	10.09	10.96	9.00	0.93	3.97
P	.004	<.001		<.001	<.001	<.001	.352	<.001
Yates's decomposition of Brier's score								
$V(Y)/\mathrm{Cov}(Y,p)$	1,879	88	−227,182	46	568	66	1,351	149
V_{exc}/V_{min}	706	99	92,646	76	46	62	952	225
Cox's calibration regression								
β	0.660	1.165	−2.116	0.992	2.522	1.200	0.524	1.142
α	−0.108	0.309	−13.397	−0.429	8.706	0.539	−1.149	0.005
H_0: $\alpha = 0, \beta = 1$								
Statistic	264.67	9.23	38.41	17.87	140.95	2.71	65.07	21.92
P	<.001	.010	<.001	<.001	<.001	.258	<.001	<.001
H_0: $\alpha = 0 \mid \beta = 1$								
Statistic	262.14	8.18	2.25	17.86	114.41	1.56	63.49	21.59
P	<.001	.004	.134	.001	<.001	.212	<.001	<.001
H_0: $\beta = 1 \mid \alpha$								
Statistic	2.53	1.04	36.16	0.01	26.54	1.15	1.57	0.33
P	.111	.307	<.001	.945	<.001	.284	.210	.564

NOTE. A_{ROC}, area under the receiver operating characteristic curve; CI, confidence interval; $\mathrm{Cov}(Y,p)$, covariance of outcome and prediction; HYS, abdominal hysterectomy; NA, not applicable; OGU, other operations of the genitourinary system; OMS, other operations of the musculoskeletal system; OSK, other operations of the integumentary system; V_{exc}, excess variance of predictions; V_{min}, minimum variance of predictions; $V(Y)$, variance of the observed outcome.

[a] $P < .001$ versus NNIS risk index.

[b] $P = .068$ versus NNIS risk index.

previous impression that accounting for the specificities of each procedure category is important. To keep the alternative indexes as clinically credible as possible, we gave priority to cutoff points defined in the ordinal sense of the covariates. For OGU and HYS, however, clean-contaminated sites had lower SSI rates than the other wound classes (Table 1). A nonordinal increase in the risk of SSI with each increment of the wound class has been reported in many studies; in particular, many authors have observed the lowest risk of SSI in clean-contaminated procedures.[34-36] That this association was observed in the development and validation samples (not shown) and for clinically related procedures (ie, OGU and HYS) suggests a real rather than a spurious association. The differential use of prophylactic and therapeutic antibiotics according to wound class is the most likely explanation for the nonordinal increase in the risk of SSI.[37]

To the best of our knowledge, this is one of the very few studies that have attempted to validate an SSI prognostic model in a sample other than that used for development. Two previous studies that externally tested alternative SSI risk models suggested an impairment in their performance compared with the performance in the development sample.[38,39] This impairment is a common observation in validation studies and was observed in our own data (not shown), reinforcing the need for externally validating risk models before they are used in practice. Simply testing the performance of a model in the model-building sample is known to give an overoptimistic picture of performance.[16,33] This is because for a model to perform well in a new setting all factors that influence outcome (including patient factors and quality of care) must either be included in the model or have the same distribution in the new setting as in the sample used to develop the model. Differences between countries and over time make this second condition unlikely.

From Table 3 it is evident that, although alternative indexes significantly improved calibration in relation to the NNIS risk index, most of them were still miscalibrated. However, when a scoring system derived from patients in one country or healthcare system during a given time period is applied to patients admitted for care in other settings or during other time periods, the interpretation of the lack of calibration is not straightforward. Moreover, the interpretation of the lack of calibration is further complicated for risk models that adjust only for intrinsic SSI risk factors because if large variations occur in the quality of care between the development and validation samples then prediction of SSI solely on the basis of intrinsic patient characteristics will be less or more efficient depending on the relative contribution of the extrinsic component to the overall SSI risk. In fact, in our own hospitals fundamental changes in quality of care (eg, antibiotic prophylaxis and sterilization protocols) occurred during the development years, which would make it unlikely that alternative risk indexes would calibrate well in the validation samples. However, the better calibration of the alternative indexes compared with the NNIS risk index still suggests a better specification of the intrinsic risk component.

Almost all previous studies reporting alternative SSI risk indexes in the literature relied heavily on single performance measures, most notably the *G* statistic. From Table 3 we see that, although the unweighted extended risk index for OMS had a lower value for the *G* statistic than the NNIS risk index, it was clearly superior in other measures, so the alternative index would still be judged to be superior to the NNIS index if all performance measures were jointly considered. This raises major concerns about what we have learned from previously published articles and strengthens the need for multidimensional evaluations of SSI risk indexes.

Incomplete postdischarge SSI surveillance is a formidable methodological challenge for hospital epidemiologists. Some authors have used an indicator variable to adjust for incomplete postdischarge surveillance in SSI prognostic regression models.[4,40] We refer to a previous study for a thorough discussion about the construct validity of this approach.[8] All of these studies have shown the importance of incorporating such an adjustment. However, no attempts have been made to transfer this adjustment into simple risk scores that are suitable for use in routine surveillance. The simple risk adjustment introduced in the extended indexes moves the patient one or more strata upward whenever postdischarge surveillance was accomplished, thus acknowledging that the measured SSI risk will obviously be higher than if no postdischarge surveillance was conducted. Accordingly, that the extended risk indexes had the best performance is not surprising, since a substantial proportion of SSIs are diagnosed after patient discharge (in our study, approximately 70%). This improvement in performance forces us to acknowledge that, without some adjustment for incomplete postdischarge surveillance, a substantial proportion of the measured SSI risk will always remain unexplained. Factors leading to more or fewer SSIs being detected after discharge may influence the performance of the extended indexes in different settings and for different procedures.

The use of alternative procedure-specific cutoff points for the NNIS risk index covariates can improve the specification of the intrinsic SSI risk component, and controlling for incomplete postdischarge SSI surveillance can provide more accurate SSI risk adjustment. Further work is warranted to evaluate both approaches.

ACKNOWLEDGMENTS

Financial support. F.M.B. was partially funded by a grant from Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG).

Potential conflicts of interest. All authors report no conflicts of interest relevant to this article. All authors submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest, and the conflicts that the editors consider relevant to this article are disclosed here.

Address correspondence to Fernando Martín Biscione, MD, PhD,

https://doi.org/10.1086/663702 Published online by Cambridge University Press

Health Sciences and Tropical Medicine Postgraduate Course, Minas Gerais Federal University School of Medicine, 190 Alfredo Balena Avenue, Room 533, Santa Efigênia, Belo Horizonte, Minas Gerais, Brazil 30-130-100 (fernandobiscione@yahoo.com.ar or fernandobiscione@med-trop.dout.ufmg.br).

REFERENCES

1. Culver DH, Horan TC, Gaynes RP, et al. Surgical wound infection rates by wound class, operative procedure, and patient risk index. *Am J Med* 1991;91(suppl 3B):S152–S157.
2. Haley RW, Culver DH, Morgan WM, White JW, Emori TG, Hooton TM. Identifying patients at high risk of surgical wound infection: a simple multivariate index of patient susceptibility and wound contamination. *Am J Epidemiol* 1985;121:206–215.
3. Vandenbroucke-Grauls C, Schultsz C. Surveillance in infection control: are we making progress? *Curr Opin Infect Dis* 2002;15: 415–419.
4. Geubbels EL, Grobbee DE, Vandenbroucke-Grauls CM, Wille JC, de Boer AS. Improved risk adjustment for comparison of surgical site infection rates. *Infect Control Hosp Epidemiol* 2006; 27:1330–1339.
5. Brandt C, Hansen S, Sohr D, Daschner F, Rüden H, Gastmeier P. Finding a method for optimizing risk adjustment when comparing surgical-site infection rates. *Infect Control Hosp Epidemiol* 2004;25:313–318.
6. Barnes S, Salemi C, Fithian D, et al. An enhanced benchmark for prosthetic joint replacement infection rates. *Am J Infect Control* 2006;34:669–672.
7. Moro ML, Morsillo F, Tangenti M, et al. Rates of surgical-site infection: an international comparison. *Infect Control Hosp Epidemiol* 2005;26:442–448.
8. Biscione FM, Couto RC, Pedrosa TM. Accounting for incomplete postdischarge follow-up during surveillance of surgical site infection by use of the National Nosocomial Infections Surveillance system's risk index. *Infect Control Hosp Epidemiol* 2009; 30:433–439.
9. Mangram AJ, Horan TC, Pearson ML, et al; Hospital Infection Control Practices Advisory Committee. Guideline for prevention of surgical site infection, 1999. *Infect Control Hosp Epidemiol* 1999;20:250–278.
10. National Healthcare Safety Network (NHSN). *The NHSN Manual: Patient Safety Component Protocol.* Atlanta: Division of Healthcare Quality Promotion, National Center for Infectious Diseases, 2008.
11. National Nosocomial Infections Surveillance system. National Nosocomial Infections Surveillance (NNIS) system report, data summary from January 1992 through June 2004, issued October 2004. *Am J Infect Control* 2004;32:470–485.
12. Edwards JR, Peterson KD, Mu Y, et al. National Healthcare Safety Network (NHSN) report: data summary for 2006 through 2008, issued December 2009. *Am J Infect Control* 2009;37: 783–805.
13. Horan TC, Emori TG. Definitions of key terms used in the NNIS system. *Am J Infect Control* 1997;25:112–116.
14. Biscione FM, Couto RC, Pedrosa TM, Neto MC. Factors influencing the risk of surgical site infection following diagnostic exploration of the abdominal cavity. *J Infect* 2007;55:317–323.
15. Altman DG, Royston P. What do we mean by validating a prognostic model? *Stat Med* 2000;19:453–473.
16. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Ann Intern Med* 1999;130:515–524.
17. Cox DR. A note on data-splitting for the evaluation of significance levels. *Biometrika* 1975;62:441–444.
18. Vergouwe Y, Steyerberg EW, Eijkemans MJC, Habbema JDF. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. *J Clin Epidemiol* 2005;58:475–483.
19. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol* 1996;49:1373–1379.
20. Kass GV. An exploratory technique for investigating large quantities of categorical data. *J R Stat Soc C* 1980;29:119–127.
21. Hosmer DW, Lemeshow S. *Applied Logistic Regression.* New York: Wiley, 1989.
22. Mehta CR, Patel NR. Exact logistic regression: theory and examples. *Stat Med* 1995;14:2143–2160.
23. King G, Zeng L. Logistic regression in rare events data. *Polit Anal* 2001;9:137–163.
24. Moons KG, Harrell FE, Steyerberg EW. Should scoring rules be based on odds ratios or regression coefficients? *J Clin Epidemiol* 2002;55:1054–1055.
25. Hanley J, McNeil B. The meaning and use of the area under a receiver-operating-characteristic curve. *Radiology* 1982;143: 29–36.
26. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988;44: 837–845.
27. Cox DR. Two further applications of a model for binary regression. *Biometrika* 1958;45:562–565.
28. Goodman LA, Kruskal WH. Measures of association for cross classifications. *J Am Stat Assoc* 1954;49:732–764.
29. Yates JF. External correspondence: decomposition of the mean probability score. *Organ Behav Hum Perform* 1982;30:132–156.
30. Cuzick J. A Wilcoxon-type test for trend. *Stat Med* 1985;4:87–90.
31. van Houwelingen JC, le Cessie S. Predictive value of statistical models. *Stat Med* 1990;9:1303–1325.
32. Campos ML, Cipriano ZM, Freitas PF. Suitability of the NNIS index for estimating surgical-site infection risk at a small university hospital in Brazil. *Infect Control Hosp Epidemiol* 2001; 22:268–272.
33. Braitman LE, Davidoff F. Predicting clinical states in individual patients. *Ann Intern Med* 1996;125:406–412.
34. Geubbels EL, Mintjes-de Groot AJ, van den Berg JM, de Boer AS. An operating surveillance system of surgical site infections in the Netherlands: results of the PREZIES national surveillance network. *Infect Control Hosp Epidemiol* 2000;21:311–318.
35. Weiss CA, Statz CL, Dahms RA, Remucal MJ, Dunn DL, Beilman GJ. Six years of surgical wound infection surveillance at a tertiary care center: review of the microbiologic and epidemiological aspects of 20,007 wounds. *Arch Surg* 1999;134:1041–1048.
36. Nguyen D, MacLeod WB, Phung DC, et al. Incidence and predictors of surgical-site infections in Vietnam. *Infect Control Hosp Epidemiol* 2001;22:485–492.
37. Di Leo A, Piffer S, Ricci F, et al. Surgical site infections in an Italian surgical ward: a prospective study. *Surg Infect (Larchmt)* 2009;10:533–538.
38. Chen LF, Anderson DJ, Kaye KS, Sexton DJ. Validating a 3-point prediction rule for surgical site infection after coronary

https://doi.org/10.1086/663702 Published online by Cambridge University Press

artery bypass surgery. *Infect Control Hosp Epidemiol* 2010;31: 64–68.
39. Batista R, Kaye K, Yokoe DS. Admission-specific chronic disease scores as alternative predictors of surgical site infection for patients undergoing coronary artery bypass graft surgery. *Infect Control Hosp Epidemiol* 2006;27:802–808.
40. Rioux C, Grandbastien B, Astagneau P. The standardized incidence ratio as a reliable tool for surgical site infection surveillance. *Infect Control Hosp Epidemiol* 2006;27:817–824.

Note added in proof. The authors note that the article by Mu et al published in the October 2011 issue of the journal concurs with the findings in this report.

https://doi.org/10.1086/663702 Published online by Cambridge University Press