ORIGINAL CLINICAL RESEARCH REPORT

# A Bayesian Comparison of Frailty Instruments in Noncardiac Surgery: A Cohort Study

Daniel I. McIsaac, MD, MPH, FRCPC,*†‡ Sylvie D. Aucoin, MD, MSc, FRCPC,* and Carl van Walraven, MD, MSc, FRCPC†‡§

**BACKGROUND:** Frailty—a multidimensional syndrome related to age- and disease-related deficits—is a key risk factor for older surgical patients. However, it is unknown which frailty instrument most accurately predicts postoperative outcomes. Our objectives were to quantify the probability of association and relative predictive performance of 2 frailty instruments (ie, the risk analysis index-administrative [RAI-A] and 5-item modified frailty index [mFI-5]) with postoperative outcomes in National Surgical Quality Improvement Program (NSQIP) data.

**METHODS:** Retrospective cohort study using Bayesian analysis of NSQIP hospitals. Adults having inpatient small or large bowel surgery 2010–2015 (derivation cohort) or intermediate to high risk mixed noncardiac surgery in 2016 (validation cohort) had preoperative frailty assigned using 2 unique approaches (RAI-A and mFI-5). Probabilities of association were calculated based on posterior distributions and relative predictive performance using posterior predictive distributions and Bayes factors for 30-day mortality (primary outcome) and serious complications (secondary outcome).

**RESULTS:** Of 50,630 participants, 7630 (14.0%) died and 19,545 (38.6%) had a serious complication. Without adjustment, the RAI-A and mFI-5 had >99% probability being associated with mortality with a ≥2.0 odds ratio (ie, large effect size). After adjustment for NSQIP risk calculator variables, only the RAI-A had ≥95% probability of a nonzero association with mortality. Similar results arose when predicting postoperative complications. The RAI-A provided better predictive accuracy for mortality than the mFI-5 (minimum Bayes factor $3.25 \times 10^{14}$), and only the RAI-A improved predictive accuracy beyond that of the NSQIP risk calculator (minimum Bayes factor = $4.27 \times 10^{13}$). Results were consistent in leave-one-out cross-validation.

**CONCLUSIONS:** Translation of frailty-related findings from research and quality improvement studies to clinical care and surgical planning will be aided by a consistent approach to measuring frailty with a multidimensional instrument like RAI-A, which appears to be superior to the mFI-5 when predicting outcomes for inpatient noncardiac surgery. (Anesth Analg XXX;XXX:00–00)

## KEY POINTS

- **Question:** What frailty instrument that is routinely used in surgical registry, data is most strongly associated with, and predictive of, postoperative morbidity and mortality?
- **Finding:** In this cohort study of noncardiac surgery patients, the risk analysis index had a higher probability of association and was more predictive of outcomes than the 5-item modified frailty index.
- **Meaning:** A multidimensional frailty instrument should be preferred over a comorbidity based instrument when assessing frailty before surgery.

## GLOSSARY

**ASA** = American Society of Anesthesiologists; **ASD** = absolute standardized difference; **BF** = Bayes factor; **CI** = credible interval; **COPD** = chronic obstructive pulmonary disease; **CPT** = current procedural terminology; **mFI-5** = 5-item modified frailty index; **NSQIP** = National Surgical Quality Improvement Program; **OR** = odds ratio; **PUF** = participant use file; **RAI-A** = risk analysis index-administrative; **SD** = standard deviation; **SIRS** = systemic inflammatory response syndrome; **SSI** = surgical site infection; **TRIPOD** = transparent reporting of a multivariable prediction model for individual prognosis or diagnosis

From the *Departments of Anesthesiology & Pain Medicine, University of Ottawa and The Ottawa Hospital, Ottawa, Ontario, Canada; †School of Epidemiology & Public Health, University of Ottawa, Ottawa, Ontario, Canada; ‡Clinical Epidemiology Program, Ottawa Hospital Research Institute, Ottawa, Ontario, Canada; and §Departments of Medicine and General Internal Medicine, University of Ottawa and The Ottawa Hospital, Ottawa, Ontario, Canada.

Accepted for publication October 7, 2020.



DOI: 10.1213/ANE.0000000000005290

Funding: D.I.M. acknowledges salary support received from The Ottawa Hospital Anesthesia Alternate Funds Association.

The authors declare no conflicts of interest.

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's website (www.anesthesia-analgesia.org).

Reprints will not be available from the authors.

Address correspondence to Daniel I. McIsaac, MD, MPH, FRCPC, Department of Anesthesiology & Pain Medicine, The Ottawa Hospital, Room B311, 1053 Carling Ave, Ottawa, ON K1Y 4E9, Canada. Address e-mail to dmcisaac@toh.ca.

Frailty is a syndrome related to accumulation of age- and disease-related deficits and increases the risk of adverse health outcomes.[1,2] In surgical populations, frailty is associated with a ≥2-fold increase in the risk of morbidity and mortality[3–5] and a ≥5-fold increase in the odds of nonhome discharge.[6,7]

Despite the strong and consistent associations reported between frailty and adverse outcomes, methods used to measure frailty are heterogeneous, which impedes the translation of frailty research in older surgical patients. Much of the surgical epidemiology relating to frailty uses data from the American College of Surgeons National Surgical Quality Improvement Program (NSQIP). Such studies typically quantify frailty using the 5-item modified frailty index (mFI-5),[8] which is based on the 70-item Canadian Study of Health and Ageing Study Frailty Index.[9] However, despite showing a dose-response association with morbidity and mortality,[9] the mFI-5 has significant limitations as a frailty index. Accumulating-deficit frailty indices require at least 30 variables[10] reflecting multiple domains understood to contribute to frailty[11]; in contrast, the mFI-5 has only 5 variables, with 4 of them indicating comorbidity status. As such, the mFI-5 is closer to a reduced comorbidity index rather than a frailty measure.[12]

The risk analysis index-administrative (RAI-A) is another frailty instrument measurable in NSQIP data (adapted from the minimum data set mortality risk index–revised). The RAI-A captures granular and multidimensional constructs reflecting frailty and, compared to the mFI-5, has many more possible values with which to quantify frailty.[13] We, therefore, used NSQIP data to directly compare the mFI-5 and RAI-A to predict the risk of mortality and complications following noncardiac surgery. We used Bayesian methods to directly quantify the probability that (1) the RAI-A has a stronger association with postoperative outcomes than the mFI-5, (2) the RAI-A more accurately predicts outcomes than the mFI-5, and (3) the RAI-A increases the accuracy of outcome prediction beyond the NSQIP risk calculator more than the mFI-5.[14]

## METHODS

### Design and Data Source

Ethical approval was granted (Ottawa Health Sciences Network Research Ethics Board 20160439-01H), including a waiver for the need for written informed consent. We conducted a retrospective cohort study using prospectively collected data from the NSQIP participant use file. These data were collected by trained nurse assessors at participating hospitals using standardized definitions, techniques, and both local and central quality checks.[15] A protocol was prespecified and registered (https://osf.io/n8xmg/), informed by methodological guidelines for prognostic research[16] and Bayesian analysis.[17] Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) guidelines informed reporting.[18]

### Cohorts

Our primary cohort included adults having inpatient small or large bowel surgery (2010–2015). Although we were not developing novel prediction models, we wanted to test whether our findings generalized beyond our initial cohort. Therefore, we undertook an analysis similar to the internal–external validation procedure described by Steyerberg and Harrell[19] as an intermediary between internal validation (ie, split-sample or resampling methods) and true external validation. The validation cohort included all intermediate- and high-risk noncardiac procedures in 2016.[20] Both cohorts were defined using current procedural terminology (CPT) codes (Supplemental Digital Content, Appendix 1, http://links.lww.com/AA/D264).

### Exposures

Preoperative frailty was measured using 2 methods. The mFI-5 was calculated as the sum of binomial indicators for preoperative heart failure, diabetes, hypertension requiring medication, chronic obstructive pulmonary disease or pneumonia, or nonindependent functional status (range 0–5).[8] The RAI-A is the sum of points assigned for a combination of age and malignancy status, sex, weight loss, renal failure, heart failure, shortness of breath, prehospital nursing home residence, and degree of functional dependence (range 0–81; Supplemental Digital Content, Appendix 2, http://links.lww.com/AA/D264).[13] For both instruments, a higher score indicated greater frailty.

The mFI-5 and RAI-A have both been analyzed in many formats (eg, continuous, categorical, and binary). Since we were interested in the performance of the overall frailty construct (as opposed to a specific cutoff), and because of the extensive drawbacks from categorizing continuous predictors,[21–23] primary analyses expressed the instruments in a continuous form.

Following recommendations for Bayesian modeling,[24] all variables were standardized by centering them at the mean and dividing by twice their standard deviation (SD) to give each variable a mean of 0 and an SD of 0.5. Standardizing permitted prior distributions to be appropriately scaled for all predictors. It also allowed measurement of effect sizes on a common scale, as the raw values of the mFI-5 and the RAI-A differ substantially. However, we also expressed each frailty score as categorical variables with 0 as reference for the mFI-5 (6 levels) and the lowest RAI-A quintile as the reference (5 levels).

These transformations tested the impact of frailty instrument parameterization on our results.

### Outcomes

The primary outcome was all-cause mortality within 30 days of surgery. The secondary outcome was the occurrence of a serious complication within 30 days of surgery. Serious complications included cardiac arrest, myocardial infarction, pneumonia, progressive renal insufficiency, acute renal failure, pulmonary embolism, deep vein thrombosis, return to the operating room, deep incisional surgical site infection (SSI), organ space SSI, systemic sepsis, unplanned intubation, urinary tract infection, and wound disruption.

### Statistical Analyses

Descriptive statistics were performed using SAS 9.4 for Windows (SAS Institute, Cary NC). They compared characteristics between people with a dichotomized frailty status (0–15 vs ≥16 for the RAI-A, 0–1 vs 2–5 for the mFI-5). Absolute standardized differences were computed for each variable with values exceeding 0.10, indicating a substantial difference.[25]

Bayesian analyses were conducted using the R package "brms" (R Foundation for Statistical Computing, Vienna, Austria).[26] A Bayesian analysis was used instead of a frequentist approach because it more closely aligned with our objectives. Frequentist analyses return *P* values, which quantify the likelihood that the study data (or more extreme data) would be observed over multiple study repetitions if the true difference between frailty instruments was exactly 0. In contrast, Bayesian analyses produce a posterior probability distribution reflecting the combination of prior beliefs and the measured data to directly estimate the probability of observed associations given the data collected. These methods also allow the calculation of 95% credible intervals (CI), which describe an interval where there is 95% probability that the true parameter would be found, given the data and prior knowledge.

Bayesian logistic regression models were used. As recommended, prior distributions for the primary analyses were weakly informative to decrease the likelihood of estimating unrealistically large or small effects without having a substantive effect on regression parameters.[27] Sensitivity analyses were conducted using prior distributions that were noninformative (ie, flat) or strongly informative (centered at an effect size equal to an odds ratio (OR) of 1.5; Supplemental Digital Content, Appendix 3, http://links.lww.com/AA/D264). Adequate mixing of chains and autocorrelation were evaluated using visual plots and Geweke diagnostics for chain convergence.[17,28] Specifically, we wanted to see that plots of the posterior distributions overlapped, that results within- and between-chains were similar, and that autocorrelation plots were relatively flat after several time lags. These ensured that the different chains estimating our models were estimating similar quantities but that the individual chains were not overly similar to chains originating at a similar time in the estimation process.

To estimate the probability that each frailty instrument was associated with each outcome, we calculated the probability that its exponentiated regression coefficient (ie, its OR) exceeded 1 (the null value), 1.5 (a moderate effect size), or 2 (a large effect size). Because Bayesian analyses produce a posterior distribution (ie, a distribution of values for each parameter that are plausible based on the prior distribution and the data analyzed), we were able to calculate the proportion of the posterior distribution that exceeded each of these threshold effect sizes. This reflected the probability that the effect was as large or larger than the threshold. For each outcome, we constructed models that included only the frailty instrument (unadjusted model), as well as the frailty instrument plus the NSQIP risk calculator variables (adjusted model; Supplemental Digital Content, Appendix 4, http://links.lww.com/AA/D264). These models adjusted for procedure using a random intercept for each CPT code.

In addition to a posterior distribution for each model parameter, Bayesian models also generate posterior predictive distributions, which function as predictions of possible values for future data.[24] To compare the relative predictive performance of each instrument, we used posterior predictive distributions to calculate Bayes factors (BF), which quantify the ratio of data probability given 1 model to data probability given a second model:

$$\mathrm{BF} = \frac{P(M1 \mid D)}{P(M2 \mid D)}$$

where $P$ = probability, $M1$ = comparator model, $M2$ = reference model, and $D$ = data.[29] Therefore, BFs measure the relative predictive accuracy of 2 models based on how well the posterior predictive distribution matched the observed data. A BF = 1 suggests no difference between models, <1 suggests better predictive performance in model 2 (the reference model), and >1 suggests better predictive performance in model 1 (the comparator model). Supplemental Digital Content, Appendix 5, http://links.lww.com/AA/D264, provides a guide for BF interpretation.

BF were calculated for models predicting the outcome based exclusively on the RAI-A versus exclusively on the mFI-5. Marginal likelihoods across posterior predictive distributions were used to compute BF using the bridge sampling method. These calculations were repeated after adjusting for NSQIP risk calculator variables to compare whether frailty

instruments improved the performance of existing approaches to risk stratification.

### Sensitivity Analyses

Additional analyses were conducted to test the impact of varying the (1) frailty instrument parameterization (ie, rerunning primary analyses using categorical frailty scores); (2) prior distributions (ie, rerunning primary analyses with noninformative and strongly informative prior distributions); (3) surgical populations and years (ie, validating primary analyses in a mixed intermediate to high cardiac risk procedures); and (4) model comparison framework (ie, using BFs versus Pareto-smoothed importance sampling with leave-one-out cross-validation).[30] Post hoc, we performed a fractional polynomial analysis to determine the best fitting form of each instrument using the "mfp" package in R.

### Sample Size and Missing Data

Our population included all available cases and was not based on a prespecified power calculation. No adjustment for multiplicity was required as a prespecified Bayesian analysis is not subject to the concerns of multiplicity encountered with frequentist analysis. All exposure and outcome data were complete. We used single mean imputation in 3700 observations (7.3%) that were missing body mass index values to calculate the RAI-A (the mean value inputed was 30).

## RESULTS

In the NSQIP dataset, we identified 50,630 individuals undergoing bowel surgery between 2010 and 2016. Mean (SD) values for the mFI-5 and the RAI-A were 1.0 (1.0) and 8.7 (7.5), respectively; the Pearson correlation coefficient between instruments was 0.387. Patient characteristics as a function of frailty varied by the instrument (Table 1).

### Thirty-Day Mortality

All models converged had adequate, effective sample size (ie, there were adequate data available from the Markov chains to estimate the distributions), and did not suffer from substantive autocorrelation (Supplemental Digital Content, Appendix 6, http://links.lww.com/AA/D264). Within 30 days of surgery, 7066 (14.0%) of the cohort died. Both the mFI-5 and RAI-A were associated with mortality. As sole predictors in the model, the probability of a nonzero association exceeded 99% for both frailty instruments.

**Table 1. Demographics Grouped by High and Low Frailty Scores**

| Characteristic | mFI-5 <2 n = 37,311 | mFI-5 ≥2 n = 13,319 | ASD | RAI-A <16 n = 42,605 | RAI-A ≥16 n = 8025 | ASD |
|---|---|---|---|---|---|---|
| Female | 53.9 | 55 | 0.02 | 55.9 | 44.9 | 0.22 |
| Elective surgery | 14.7 | 19.3 | 0.12 | 14.9 | 21.3 | 0.17 |
| Large bowel surgery versus small bowel | 72.6 | 73.3 | 0.02 | 71.9 | 77.8 | 0.14 |
| Diabetes mellitus | 2.9 | 53.7 | 1.37 | 15.1 | 22.3 | 0.19 |
| Hypertension | 40.2 | 94.1 | 1.40 | 52.8 | 62.4 | 0.20 |
| Heart failure | 0.3 | 12 | 0.50 | 2.1 | 10.4 | 0.35 |
| Dyspnea at rest | 1.6 | 8.9 | 0.33 | 1.2 | 15.7 | 0.54 |
| Moderate dyspnea | 4.3 | 14.2 | 0.35 | 6.3 | 10.1 | 0.14 |
| Smoker | 20.6 | 20.2 | 0.01 | 21.1 | 17.6 | 0.09 |
| COPD | 2.6 | 34 | 0.89 | 9.6 | 25.8 | 0.43 |
| Dialysis | 2.2 | 8.3 | 0.28 | 2.6 | 9.9 | 0.31 |
| Acute kidney injury | 2.5 | 7.8 | 0.24 | 2.5 | 11.1 | 0.35 |
| Metastatic cancer | 7.4 | 6.3 | 0.04 | 0.4 | 43 | 1.21 |
| Preoperative ventilation | 4.3 | 13.25 | 0.32 | 4.6 | 17.5 | 0.42 |
| SIRS | 12.3 | 11.7 | 0.02 | 12.3 | 11.2 | 0.03 |
| Sepsis | 26.8 | 28.5 | 0.04 | 27.6 | 26.6 | 0.02 |
| Septic shock | 9.6 | 22.8 | 0.36 | 10.4 | 27.2 | 0.44 |
| Ascites | 3.9 | 5.6 | 0.08 | 3.5 | 9.5 | 0.25 |
| Steroid | 9.3 | 14.6 | 0.16 | 9.8 | 15 | 0.16 |
| Independent functional status | 95.5 | 62.6 | 0.88 | 94 | 48.8 | 1.16 |
| Partially dependent | 2.3 | 24.9 | 0.70 | 4.9 | 25.9 | 0.61 |
| Totally dependent | 1.2 | 11.3 | 0.43 | 0 | 24.4 | 0.80 |
| ASA physical status ≤II | 31.8 | 4 | 0.78 | 28.1 | 5.9 | 0.62 |
| ASA physical status III | 43.8 | 38.8 | 0.10 | 43.8 | 35.4 | 0.17 |
| ASA physical status IV | 21.9 | 50.1 | 0.61 | 25.3 | 50.4 | 0.54 |
| ASA physical status V | 2.5 | 6.7 | 0.20 | 2.8 | 8.3 | 0.24 |
| Age <65 | 56.5 | 32.3 | 0.50 | 52.3 | 38.9 | 0.27 |
| Age 65–74 | 20.4 | 28.4 | 0.19 | 22.4 | 23.3 | 0.02 |
| Age 75–84 | 16.8 | 29.1 | 0.30 | 18.6 | 28 | 0.22 |
| Age 85+ | 6.2 | 10.2 | 0.15 | 6.7 | 10.1 | 0.12 |

All values represent % with characteristic.

Abbreviations: ASA, American Society of Anesthesiologists; ASD, absolute standardized difference; COPD, chronic obstructive pulmonary disease; mFI-5, 5-item modified frailty index; RAI-A, risk analysis index-administrative; SIRS, systemic inflammatory response syndrome.

**Table 2. Mortality Effect Sizes and Probabilities of Association**

| | | | Probability of effect size | | |
|---|---|---|---|---|---|
| Model | OR[a] | 95% CI | OR >1.0 | OR >1.5 | OR >2.0 |
| Weakly informative prior | | | | | |
| RAI-A | 2.61 | 2.53-2.69 | >.99 | >.99 | >.99 |
| mFI-5 | 3.00 | 2.89-3.16 | >.99 | >.99 | >.99 |
| RAI-A + NSQIP | 1.63 | 1.46-1.82 | >.99 | .93 | 0 |
| mFI-5 + NSQIP | 1.14 | 0.73-1.79 | .73 | .11 | .01 |
| Strongly informative prior | | | | | |
| RAI-A | 2.61 | 2.53-2.72 | >.99 | >.99 | >.99 |
| mFI-5 | 3.00 | 2.89-3.13 | >.99 | >.99 | >.99 |
| RAI-A + NSQIP | 1.63 | 1.46-1.84 | >.99 | .93 | 0 |
| mFI-5 + NSQIP | 1.15 | 0.74-1.82 | .72 | .11 | .01 |
| Noninformative prior | | | | | |
| RAI-A | 2.61 | 2.53-2.69 | >.99 | >.99 | >.99 |
| mFI-5 | 3.00 | 2.89-3.16 | >.99 | >.99 | >.99 |
| RAI-A + NSQIP | 1.63 | 1.46-1.84 | >.99 | .94 | 0 |
| mFI-5 + NSQIP | 1.13 | 0.75-1.75 | .69 | .1 | 0 |

Abbreviations: CI, credible interval; mFI-5; 5-item modified frailty index; NSQIP, National Surgical Quality Improvement; OR, odds ratio; RAI-A, risk analysis index-administrative.

[a]For each change of 2 standard deviations (15 points for RAI-A, 2 points for mFI-5).

**Table 3. Bayes Factors for Mortality**

| Comparator model | Reference model | Bayes factor[a] |
|---|---|---|
| Weakly informative prior | | |
| RAI-A | mFI-5 | $1.85 \times 10^{102}$ |
| RAI-A + NSQIP | NSQIP | $4.27 \times 10^{13}$ |
| mFI-5 + NSQIP | NSQIP | 0.1 |
| RAI-A + NSQIP | mFI-5 + NSQIP | $4.65 \times 10^{14}$ |
| Strongly informative priors | | |
| RAI-A | mFI-5 | $1.82 \times 10^{102}$ |
| RAI-A + NSQIP | NSQIP | $3.79 \times 10^{37}$ |
| mFI-5 + NSQIP | NSQIP | 0.1 |
| RAI-A + NSQIP | mFI-5 + NSQIP | $3.70 \times 10^{38}$ |
| Noninformative priors | | |
| RAI-A | mFI-5 | $1.82 \times 10^{102}$ |
| RAI-A + NSQIP | NSQIP | $1.96 \times 10^{14}$ |
| mFI-5 + NSQIP | NSQIP | 0.66 |
| RAI-A + NSQIP | mFI-5 + NSQIP | $3.25 \times 10^{14}$ |

Abbreviations: mFI-5, 5-item modified frailty index; NSQIP, National Surgical Quality Improvement Program; RAI-A, risk analysis index-administrative.

[a]A Bayes factor is the ratio of the probability of the data for the comparator model to the probability of the data given in the reference model; therefore, a result >1 signifies that the comparator model provides a better fit, results <1 signify that the reference model provides a better fit.

For the RAI-A, an increase of 2 SDs (eg, from 0 to 15 points) was associated with a 2.61-fold increase in the odds of mortality (95% CI, 2.53-2.69). For the mFI-5, an increase of 2 SDs (eg, from 0 to 2 points) was associated with a 3.00-fold increase in the odds of mortality (95% CI, 2.89-3.16).

Table 2 presents the probability of different strengths of association for both frailty instruments with 30-day death risk as a function of adjustment for NSQIP risk calculator and prior distribution. The probability that both mFI-5 and RAI-A (estimated separately) were associated with an OR exceeding 2.0 (ie, a large effect size) was more than 99% regardless of whether a weakly informative, strongly informative, or noninformative prior distribution was used. After adjusting for the NSQIP risk calculator, however, there was a higher probability of association with 30-day death for the RAI-A than the mFI-5. The BF comparing unadjusted models having RAI-A versus mFI-5 (Table 3) was $1.85 \times 10^{102}$. After adjusting for NSQIP risk calculator variables, the BF comparing models with RAI-A versus mFI-5 was $4.65 \times 10^{14}$.

Only the RAI-A improved mortality risk prediction when added to the NSQIP risk calculation (Table 2). Compared to a model solely containing the NSQIP risk calculation, the NSQIP plus RAI-A greatly improved mortality risk prediction (BF = $4.27 \times 10^{13}$). In contrast, a model with NSQIP + mFI-5 actually worsened mortality risk prediction (BF = 0.10). Sensitivity analyses returned similar results, although categorized mFI-5 provided weak evidence of improved fit compared to the NSQIP alone (BF = 2.29). The categorical RAI-A showed a superior fit than the mFI-5 categorical (BF = $7.64 \times 10^5$). Changes in prior distributions made no substantive difference in point or probability estimates (Tables 2 and 3); the proportion who died by category of each instrument are provided in Supplemental Digital Content, Appendix 7, http://links.lww.com/AA/D264. Results were entirely consistent when using leave-one-out cross-validation (Supplemental Digital Content, Appendix 8, http://links.lww.com/AA/D264). For both frailty instruments, a linear form was identified as the best fitting polynomial.

## Complications

A serious complication occurred in 19,545 individuals (38.6%). Models for complications required 4000 additional iterations to achieve adequate convergence and effective sample size. Before adjustment for NSQIP calculator variables, the RAI-A was associated with complications (OR, 1.55, 95% CI, 1.52-1.60), as was the mFI-5 (OR, 1.79, 95% CI, 1.73-1.84). The BF strongly suggested superior prediction in the unadjusted models by mFI-5 (Supplemental Digital Content, Appendix 7 and 8, http://links.lww.com/AA/D264). After adjustment, however, effect sizes for both instruments were notably attenuated. Only the RAI-A had <95% probability of a nonzero association with postoperative complications (Supplemental Digital Content, Appendix 9 and 10, http://links.lww.com/AA/D264). While the BF for the RAI-A plus NSQIP decisively favored addition of the RAI-A, addition of the mFI-5 worsened predictions compared to the NSQIP alone.

## Validation

In the validation cohort (Table 4, Supplemental Digital Content, Appendix 11, http://links.lww.com/AA/D264), effect sizes were increased for the RAI-A and mFI-5, although the probability of association was unchanged for the RAI-A. BF favored the NSQIP plus

**Table 4. Validation Effect Sizes and Probabilities of Association**

| Model | OR[a] | 95% CI | Probability of effect size OR >1.0 | OR >1.5 | OR >2.0 |
|---|---|---|---|---|---|
| 30-d mortality | | | | | |
| RAI-A | 3.63 | 3.39-3.90 | >.99 | >.99 | >.99 |
| mFI-5 | 3.06 | 2.77-3.39 | >.99 | >.99 | >.99 |
| RAI-A + NSQIP | 1.85 | 1.42-2.44 | >.99 | .94 | .3 |
| mFI-5 + NSQIP | 2.1 | 0.14-31.19 | .72 | .61 | .52 |
| Serious complications | | | | | |
| RAI-A | 1.68 | 1.63-1.73 | >.99 | >.99 | 0 |
| mFI-5 | 1.63 | 1.58-1.68 | >.99 | >.99 | 0 |
| RAI-A + NSQIP | 1.32 | 1.19-1.48 | >.99 | .01 | 0 |
| mFI-5 + NSQIP | 1.39 | 0.10-21.98 | .6 | .47 | .39 |

Abbreviations: CI, credible interval; mFI-5, 5-item modified frailty index; NSQIP, National Surgical Quality Improvement Program; OR, odds ratio; RAI-A, risk analysis index-administrative.

[a]For each change of 2 standard deviations (15 points for RAI-A, 2 points for mfi-5).

RAI-A over the NSQIP alone, but the NSQIP alone over the NSQIP plus mFI-5.

## DISCUSSION

In this retrospective cohort study using prospectively collected surgical data, we found convincing evidence that the RAI-A is superior to the mFI-5 for predicting mortality and serious complications in bowel and mixed noncardiac surgeries. Using Bayesian methods, we consistently found that with or without adjustment for a robust set of preoperative variables, the probability that models containing the RAI-A provided a better fit to the data than the mFI-5 was at least thousands of times higher. Furthermore, the RAI-A, but not the mFI-5, consistently improved outcome prediction when added to the NSQIP calculator. Therefore, together with the fact that the construct underlying the RAI-A is more consistent with consensus definitions of frailty,[31] we strongly suggest that future research and quality improvement initiatives using the NSQIP employ the RAI-A and not the mFI-5 when analyses regarding frailty are conducted.

Frailty is a key risk factor when considering surgery for older people, with most studies reporting associations in excess of a 1.5-fold increase in the risk of morbidity and mortality,[4,32,33] and 5-fold increases in the odds of losing independence.[6,7] Data from our study are consistent with these findings; we estimated a >99% probability that on their own, the RAI-A and mFI-5 are associated with an increased risk of mortality and complications. However, our data suggest that the RAI-A is more likely to provide unique information for risk prediction. Independent of other variables known to be strong predictors of mortality, we found a 99% probability of a nonzero association between the RAI-A and 30-day death risk. In contrast, the probability that the mFI-5 was associated with mortality was only 73% after adjustment for other predictors. The fact that the components of the RAI-I reflect the multidimensional nature of frailty,[31] in contrast to the comorbidity-oriented mFI-5, may explain this finding.

In addition to evidence that the RAI-A was more strongly associated with mortality and complications, we also found that the RAI-A more accurately predicted outcomes than the mFI-5. This is a novel finding that addresses an important knowledge gap. Unfortunately, few studies directly compare different frailty instruments.[34] Most pertinent to the current study, Hall et al[13] compared the mFI-11 to the RAI-A and found minimal differences in discrimination (ie, area under the curve [the probability that someone with a higher predicted risk actually experiences the outcome]) when predicting death or complications. However, among other limitations, as a measure of predictive accuracy, discrimination is insensitive to changes in model performance.[35,36] Furthermore, calibration (agreement between observed and predicted outcome rates) and gold-standard model fit comparisons for frequentist analysis (such as the log-likelihood test) were not computed.[35,37,38] In our study, using a Bayesian approach to model assessment (which compares the posterior predictive distribution [ie, the expected outcomes given the model and associated uncertainty] to the actual outcomes [and therefore has similarities to calibration or a proper scoring rule]),[24,37,39] strong and consistent evidence suggests that the RAI-A has thousands of times higher probability of accurately predicting outcomes than the mFI-5. This finding was consistent across procedure groups and after adjustment for known predictors.

That the RAI-A substantially increased the probability of a better model fit when added to NSQIP risk calculator, whereas the mFI-5 provided a worse fit than the NSQIP calculator alone, further supports the assertion that the construct of the RAI-A more accurately reflects multidimensional frailty than does the mFI-5. As the computation of the posterior predictive probabilities used to calculate BFs penalizes more complex models that do not provide information more valuable than the complexity incurred, this suggests that the mFI-5 simply did not provide important new data to support accurate outcome prediction. In other words, making a model more complex through adding variables largely already incorporated in the baseline model (as most components of the mFI-5 are in the NSQIP risk calculator) is unlikely to improve performance. Translated into the clinical context, these data suggest that added value from guideline-recommended frailty assessment is likely only to be realized using instruments that capture the multidimensional nature of frailty. Familiarity and ease of computation may currently explain the preferential use of the mFI-5 over the RAI-A in studies of NSQIP data (75% greater use according to our structured review[40]); therefore, knowledge translation strategies

may include active dissemination and confirmation of our findings, as well as consideration of calculation and inclusion of the RAI-A in future releases of NSQIP participant use file data.

### Strengths and Limitations

The strengths and limitations of our study should be considered. Our analyses were prespecified in a registered protocol that specified outcomes, analyses, and choice of prior distributions; all of which can both influence BF estimation and are often said to make Bayesian analyses too subjective. Furthermore, we conducted extensive sensitivity analyses, including varied prior distributions, and found consistency in our results. This included an "internal–external" validation step,[19] where analyses were rerun in a temporally and procedurally distinct cohort to confirm generalizability.

However, limitations also exist. We cannot estimate the generalizability of our results to non-NSQIP hospitals, and the RAI-A and mFI-5 were specifically derived for NSQIP data. Our analyses were limited to inpatient noncardiac surgeries; generalization to cardiac and ambulatory surgical populations is required. Although we prespecified parameterizing both frailty instruments continuously, in keeping with recommendations for reducing bias and overfitting in prediction models, our data suggest that the mFI-5 provides a better model fit as a categorical variable. That said, even categorically expressed, the RAI-A had a 300,000-fold higher probability of fitting our data than did the categorical mFI-5. Furthermore, our analyses could compare only the relative performance of the RAI-A and mFI-5; we are unable to provide further insights into comparisons with other frailty instruments. As the RAI-A may not be easily used in clinical practice, users could consider other clinically oriented tools shown to be as, or more accurate than, complex instruments (eg, the clinical frailty scale).[34]

## CONCLUSIONS

Frailty is a key prognostic factor in older surgical patients that must be addressed to improve outcomes in our aging surgical population. When operationalizing frailty in research and quality improvement settings, our data suggest that the RAI-A substantially out-performs the mFI-5 and provides unique prognostic data beyond that captured by the NSQIP universal risk calculator.

### DISCLOSURES

**Name:** Daniel I. McIsaac, MD, MPH, FRCPC.
**Contribution:** This author helped conceive, design, acquire the data for, analyze, interpret and draft, revise, and approve the final manuscript, and is the guarantor.
**Name:** Sylvie D. Aucoin, MD, MSc, FRCPC.
**Contribution:** This author helped conceive, design, interpret, draft, revise, and approve the final manuscript.
**Name:** Carl van Walraven, MD, MSc, FRCPC.
**Contribution:** This author helped conceive, design, interpret, draft, revise, and approve the final manuscript.
**This manuscript was handled by:** Robert Whittington, MD.

### REFERENCES


1. Fried LP, Tangen CM, Walston J, et al; Cardiovascular Health Study Collaborative Research Group. Frailty in older adults: evidence for a phenotype. *J Gerontol A Biol Sci Med Sci*. 2001;56:M146–M156.
2. Rockwood K, Song X, MacKnight C, et al. A global clinical measure of fitness and frailty in elderly people. *CMAJ*. 2005;173:489–495.
3. Watt J, Tricco AC, Talbot-Hamon C, et al. Identifying older adults at risk of delirium following elective surgery: a systematic review and meta-analysis. *J Gen Intern Med*. 2018;33:500–509.
4. Lin HS, Watts JN, Peel NM, Hubbard RE. Frailty and post-operative outcomes in older surgical patients: a systematic review. *BMC Geriatr*. 2016;16:157.
5. McIsaac DI, Bryson GL, van Walraven C. Association of frailty and 1-year postoperative mortality following major elective noncardiac surgery: a population-based cohort study. *JAMA Surg*. 2016;151:538–545.
6. McIsaac DI, Taljaard M, Bryson GL, et al. Frailty as a predictor of death or new disability after surgery: a prospective cohort study. *Ann Surg*. 2020;271:283–289.
7. McIsaac DI, Beaule PE, Bryson GL, van Walraven C. The impact of frailty on outcomes and healthcare resource utilization after total joint arthroplasty: a population-based cohort study. *Bone Jt J*. 2016;98:799–805.
8. Subramaniam S, Aalberg JJ, Soriano RP, Divino CM. New 5-factor modified frailty index using American College of Surgeons NSQIP data. *J Am Coll Surg*. 2018;226:173–181.e8.
9. Mitnitski AB, Song X, Rockwood K. The estimation of relative fitness and frailty in community-dwelling older adults using self-report data. *J Gerontol A Biol Sci Med Sci*. 2004;59:M627–M632.
10. Searle SD, Mitnitski A, Gahbauer EA, Gill TM, Rockwood K. A standard procedure for creating a frailty index. *BMC Geriatr*. 2008;8:24.
11. Mitnitski a B, Mogilner a J, Rockwood K. Accumulation of deficits as a proxy measure of aging. *Sci World J*. 2001;1:323–336.
12. Fried LP, Ferrucci L, Darer J, Williamson JD, Anderson G. Untangling the concepts of disability, frailty, and comorbidity: implications for improved targeting and care. *J Gerontol A Biol Sci Med Sci*. 2004;59:255–263.
13. Hall DE, Arya S, Schmid KK, et al. Development and initial validation of the risk analysis index for measuring frailty in surgical populations. *JAMA Surg*. 2017;152:175–182.
14. 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.
15. *User Guide for the ACS NSQIP Participant Use Data File*; 2014. Available at: https://www.facs.org/quality-programs/acs-nsqip/participant-use. Accessed November 9, 2020.
16. Hayden JA, van der Windt DA, Cartwright JL, Côté P, Bombardier C. Assessing bias in studies of prognostic factors. *Ann Intern Med*. 2013;158:280–286.
17. Depaoli S, van de Schoot R. Improving transparency and replication in Bayesian statistics: the WAMBS-checklist. *Psychol Methods*. 2017;22:240–261.
18. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med*. 2015;162:55.

19. Steyerberg EW, Harrell FE Jr. Prediction models need appropriate internal, internal-external, and external validation. *J Clin Epidemiol*. 2016;69:245–247.
20. Liu JB, Liu Y, Cohen ME, Ko CY, Sweitzer BJ. Defining the intrinsic cardiac risks of operations to improve preoperative cardiac risk assessments. *Anesthesiology*. 2018;128:283–292.
21. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. *Stat Med*. 2006;25:127–141.
22. van Walraven C, Hart RG. Leave 'em alone – why continuous variables should be analyzed as such. *Neuroepidemiology*. 2008;30:138–139.
23. Bennette C, Vickers A. Against quantiles: categorization of continuous variables in epidemiologic research, and its discontents. *BMC Med Res Methodol*. 2012;12:21.
24. Gelman A, Carlin JB, Stern H. *Bayesian Data Analysis*. 2nd. Chapman & Hall/CRC; 2003.
25. Austin PC. Using the standardized difference to compare the prevalence of a binary variable between two groups in observational research. *Commun Stat - Simul Comput*. 2009;38:1228–1234.
26. Bürkner P-C. brms : an R package for Bayesian multilevel models using stan. *J Stat Softw*. 2017;80:1–27.
27. Gelman A, Jakulin A, Pittau MG, Su Y-S. A weakly informative default prior distribution for logistic and other regression models. *Ann Appl Stat*. 2008;2:1360–1383.
28. McElreath R. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. CRC Press; 2015.
29. Kass RE, Raftery AE. Bayes factors. *J Am Stat Assoc*. 1995;90:773–795.
30. Vehtari A, Gelman A, Gabry J. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat Comput*. 2017;27:1413–1432.
31. Rodríguez-Mañas L, Féart C, Mann G, et al. Searching for an operational definition of frailty: a Delphi method based consensus statement. The frailty operative definition-consensus conference project. *J Gerontol A Biol Sci Med Sci*. 2013;68:62–67.
32. Watt J, Tricco AC, Talbot-Hamon C, et al. Identifying older adults at risk of harm following elective surgery: a systematic review and meta-analysis. *BMC Med*. 2018;16:2.
33. Kim DH, Kim CA, Placide S, Lipsitz LA, Marcantonio ER. Preoperative frailty assessment and outcomes at 6 months or later in older adults undergoing cardiac surgical procedures: a systematic review. *Ann Intern Med*. 2016;165:650–660.
34. Aucoin SD, Hao M, Sohi R, et al. Accuracy and feasibility of clinically applied frailty instruments before surgery. *Anesthesiology*. 2020:133:78–95.
35. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation*. 2007;115: 928–935.
36. Pepe MS. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *Am J Epidemiol*. 2004;159:882–890.
37. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*. 2010;21:128–138.
38. Cook NR. Quantifying the added value of new biomarkers: how and how not. *Diagn Progn Res*. 2018;2:14.
39. Carpenter B. Bayesian Posteriors are Calibrated by Definition. Statistical Modeling, Causal Inference, and Social Science. 2017. Available at: https://statmodeling.stat.columbia.edu/2017/04/12/bayesian-posteriors-calibrated/. Accessed May 9, 2020.
40. McIsaac D, Alkadri J. A systematic review and meta-analysis of frailty instruments in perioperative electronic health data. *Open Science Foundation*. 2020. Available at: https://osf.io/rvc8k/. Accessed November 09, 2020.