Original Study

ANNALS OF SURGERY OPEN

OPEN

Assessment of Racial Bias within the Risk Analysis Index of Frailty

Michael A. Jacobs, MS,* Leslie R. M. Hausmann, PhD,*†; Robert M. Handzel, MD, MS,*‡ Susanne Schmidt, PhD,§ Carly A. Jacobs, MPH,* and Daniel E. Hall, MD, MDiv, MHSc*‡‖¶

Objective: Our objective was to assess potential racial bias within the Risk Analysis Index (RAI).

Background: Patient risk measures are rarely tested for racial bias. Measures of frailty, like the RAI, need to be evaluated for poor predictive performance among Black patients.

Methods: Retrospective cohort study using April 2010–March 2019 Veterans Affairs Surgical Quality Improvement Program and 2010–2019 National Surgical Quality Improvement Program data. The performance of the RAI and several potential variants were compared between Black and White cases using various metrics to predict mortality (180-day for Veterans Affairs Surgical Quality Improvement Program, 30-day for National Surgical Quality Improvement Program).

Results: Using the current, clinical threshold, the RAI performed as good or better among Black cases across various performance metrics *versus* White. When a higher threshold was used, Black cases had higher true positive rates but lower true negative rates, yielding 2.0% higher balanced accuracy. No RAI variant noticeably eliminated bias, improved parity across both true positives and true negatives, or improved overall model performance.

Conclusions: The RAI tends to predict mortality among Black patients better than it predicts mortality among White patients. As existing bias-reducing techniques were not effective, further research into bias-reducing techniques is needed, especially for clinical risk predictions. We recommend using the RAI for both statistical analysis of surgical cohorts and quality improvement programs, such as the Surgical Pause.

INTRODUCTION

Patient risk measures, such as comorbidity indices, are commonly used to assess and improve healthcare quality, allocate additional resources to patients in need, and adjust medical service payments. While these indices can be used to improve outcomes or reduce cost,[1] they can also manifest unintended racial bias.[2] Racial bias can occur in any predictive algorithm, driven by several possible causes. Training models on a biased outcome can drive racial disparities; for example, healthcare cost models often underestimate the healthcare risks of Black patients because such patients often have reduced access to care that artificially lowers their healthcare costs.[2] In addition, social risk factors can drive race-based differences in relationships between predictor variables and outcomes. The result of this phenomenon is often underappreciated: when any statistical model is calibrated across varying social groups (eg, race), the model will estimate coefficients closer to those of the larger group and make less accurate predictions for subgroups with fewer observations.[3] Worse, these two, distinct sources of bias can co-occur.

Frailty is a syndrome of decreased physiologic reserve characterized by physical weakness, exhaustion, and low or slow activity levels.[4] Although frailty often includes measures of comorbidity, its success in predicting surgical outcomes is largely due to its focus on functional performance assessment.[5,6] One of the most thoroughly validated measures of surgical frailty is the Risk Analysis Index (RAI),[7–12] which outperforms the now obsolete modified Frailty Index in surgical cohorts,[13] and has been shown to improve perioperative outcomes in programs for identifying and mitigating perioperative frailty.[14,15] The RAI comes in 3 versions: (1) a clinical questionnaire,[7] (2) an administrative score using variables from surgical registry data (Veterans Affairs Surgical Quality Improvement Program [VASQIP] and the American College of Surgeons' National Surgical Quality Improvement Program),[7] and (3) an index scoring International Classification of Diseases-10 codes.[16] Based on converging evidence from all 3 versions,[7–12] programs are increasingly using a clinical questionnaire version of the RAI to guide perioperative decision-making. For example, Epic makes the RAI available to any client worldwide as an official clinical program,[17] and the Veterans Health Administration's National Surgery Office adopted the RAI-based Surgical Pause program at the national level in 2023.[18] In the Surgical Pause, all patients being evaluated for surgery are screened for frailty using a

*From the *Center for Health Equity Research and Promotion, VA Pittsburgh Healthcare System, Pittsburgh, PA; †Division of General Internal Medicine, University of Pittsburgh School of Medicine, Pittsburgh, PA; ‡Department of Surgery, University of Pittsburgh School of Medicine, Pittsburgh, PA; §Department of Population Health Sciences, University of Texas Health San Antonio, San Antonio, TX; ‖Geriatric Research Education and Clinical Center, VA Pittsburgh Healthcare System, Pittsburgh, PA; ¶Wolff Center, University of Pittsburgh Medical Center, Pittsburgh, PA.*

Disclosures: This research was supported by a grant support from the VHA Office of Research and Development (HSR&D I01HX003095). The authors disclose other grant funding from the NIH and VHA ORD outside the scope of this work. Dr. Hall discloses a consulting relationship with FutureAssure, LLC.

Disclaimer: The funding sources had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication. The opinions expressed here are those of the authors and do not necessarily reflect the position of the United States government.

SDC *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 Web site (www.annalsofsurgery.com).*

Reprints: Daniel E. Hall, UPMC Presbyterian, Suite F1264, 200 Lothrop Street, Pittsburgh, PA 15213. Email: hallde@upmc.edu

Annals of Surgery Open (2024) 4:e490

Received: 16 July 2024; Accepted 7 August 2024

Published online 25 September 2024

DOI: 10.1097/AS9.0000000000000490

clinical questionnaire RAI, with those in the highest risk decile[7] directed to preoperative goal clarification or prehabilitation.[19]

With more than 1 million patients screened across more than 50 Veterans Affairs (VA) and private sector medical centers, the Joint Commission and the National Quality Forum recognized the Surgical Pause with the 2023 John M. Eisenberg Award for Patient Safety and Quality at a National Level,[20] setting the stage for even wider adoption of the Surgical Pause and the RAI. However, potential algorithmic bias in frailty measures, including the RAI, has not been properly assessed. Given the evidence that Black patients, compared with their White counterparts, have higher illness burdens at younger ages,[21] have increased risk of death[22,23] from disproportionate exposure to chronic stressors (eg, "weathering"), and that such weathering has driven bias in other comorbidity measures,[2,24] the possibility of racial bias within the RAI requires investigation.

Using high-quality VASQIP data,[25] we designed this study to quantify and compare model performance between the administrative RAI and 3 modified versions intended to reduce racial bias between Black and White patients. We sought to quantify any Black-White racial bias present within the existing RAI and assess whether any commonly used technique was useful to alleviate such bias.

METHODS

Study Design and Cohort

The study was determined to be exempt by the VA Pittsburgh Healthcare System's Institutional Review Board and reported according to Strengthening the Reporting of Observational Studies in Epidemiology guidelines.[26] The cohort included VASQIP cases performed between April 2010 and March 2019 and was a subset of the cohort used to calibrate the RAI.[7] VASQIP is a quality assessment database containing nurse-abstracted data on VA surgeries, using methods described elsewhere.[25] VASQIP provides superior capture of patient risk factors and outcomes[27] and tracks 180-day mortality. We excluded cases that had missing race/ethnicity (15.3%), were not Black or White (6.2%), or were missing variables needed to calculate the RAI (0.003%, 78.4% of cases remaining). Because surgical mortality is a rare outcome, the Asian/Pacific Islander and Native American groups lacked adequate cases with 180-day mortality to adequately assess the RAI's performance (32 and 80 cases, respectively).

Outcome

The VASQIP variable DWIN6MO was used for 180-day mortality. DWIN6MO is derived from various sources, including VA sources, Medicare, and Social Security Death Master File, to capture patient mortality more completely.

Exposure

The administrative RAI[7] was calculated with preoperative VASQIP variables. Prior work demonstrates that the RAI provides a composite measure of patient-related risk that avoids problems with model fit associated with models that include each of the underlying VASQIP variables.[8,10]

Bias Mitigation Strategies

To assess possible solutions to any detected racial bias, we recalibrated the base RAI scoring system (RAI_{Base}) using 3 approaches used to minimize bias in other algorithms: (1) balancing the sample by reweighting racial strata such that Black and White patients were equally represented (ie, 50/50),[28] (2) matching the mean age of Black patients to the mean age of White patients (ie, adding 3.5 years to each Black patient's age),[28] and (3) creating a new RAI scoring system using a penalized form of logistic regression designed to balance false positive and false negative rates between the 2 groups.[29] For each approach, we used the dataset from the original RAI recalibration analysis, re-solving the regression equation for new parameter estimates, and generating predicted probabilities of 180-day mortality for the 3 RAI variants: the (1) reweighted balanced-sample RAI ($RAI_{Balanced}$), (2) age-adjusted RAI ($RAI_{Age\text{-}adjusted}$), and (3) Fair RAI ($RAI_{Fair}$).

Bias Quantification and Comparison

Logistic regression models between each version of the RAI and 180-day mortality, with fixed effects adjustment for the VA site, were performed on the entire cohort. C-statistics were calculated for each model to evaluate statistical discrimination. To generate dichotomized predictions of mortality, several thresholds were assessed: (1) an RAI value of 30 with a 6.7% predicted probability of 180-day mortality (used by the Surgical Pause program), (2) an RAI value of 36 with a 15% predicted probability (phenotypically obvious frailty), and (3) an RAI value of 26 with a 3.5% predicted probability (cutoff for optimal sensitivity and specificity[30]). After ensuring that the optimal threshold (RAI = 26) was similar in the models generated by each of the 3 bias mitigation strategies, the cohort was then split into Black and White patients, with the sensitivity, specificity, false positive and false negative rates, accuracy, balanced accuracy,[31] Matthew's correlation coefficient,[32] and F1 score[33] calculated separately for each race group. We compared the performance metrics listed above, stratified by Black and White subgroups. In addition, and following methods described elsewhere,[34] we calculated metrics of equal opportunity (the ratio of true positives in Black and White samples) and predictive equality (the ratio of true negatives in Black and White samples). The fairness criterion of equalized odds is defined as having both equal opportunity and predictive equality. Finally, we examined separate receiver operating characteristic (ROC) curves for Black and White patients in the RAI_{Base} to characterize its predictive performance between the 2 groups. Analyses were performed using R 4.3.1, with fair regression implemented using a Python script[29] shared by O.M. and B.Z. in the acknowledgments and adapted by authors R.M.H. and M.A.J.

Sensitivity Analysis

Sensitivity analyses replicating the approach described above were conducted in a sample of National Surgical Quality Improvement Program (NSQIP) cases to examine if findings were similar in a national-level, private sector sample of surgical cases. We examined the same years (2010–2019), analyzed only the RAI_{Base}, and selected a threshold of 2.3% predicted probability of 30-day mortality (RAI = 30). We examined performance metrics between Black and White cases, and separate ROC curves, as described above.

RESULTS

Patient Demographics

Analyses included 377,107 surgical cases (Table 1), 92.3% male with mean (SD) age of 60.8 (12.9). Black patients represented 18.6% of the sample. Most White cases had normal frailty (RAI 21–29, 56.0%), followed by robust (RAI ≤20, 33.5%), frail (RAI 30–39, 8.2%), and very frail (RAI ≥40, 2.3%), while most Black cases had robust frailty (50.2%), followed by normal (38.7%), frail (7.8%), and very frail (3.2%). RAI components with noticeable racial differences were renal failure (1.2% in White cases vs. 3.7% in Black cases, $P < 0.001$), cognitive deterioration (3.4% in White cases vs. 4.6% in Black cases, $P < 0.001$), and functional status (7.6% partially or totally dependent White cases vs. 9.0% Black cases, $P < 0.001$).

TABLE 1.

Cohort Demographics

	Overall	Black	White	*P* value
Number (%)*	377,107	70,301 (18.6)	306,806 (81.4)	
Age mean [SD]	60.8 [12.9]	58.0 [12.1]	61.5 [13.0]	<0.001
Male	347,912 (92.3)	61,745 (87.8)	286,167 (93.3)	<0.001
Cancer	7,755 (2.1)	1,535 (2.2)	6,220 (2.0)	0.009
Weight loss	10,141 (2.7)	2,270 (3.2)	7,871 (2.6)	<0.001
Renal failure	6,185 (1.6)	2,617 (3.7)	3,568 (1.2)	<0.001
Congestive heart failure	2,645 (0.7)	565 (0.8)	2,080 (0.7)	<0.001
Dyspnea				<0.001
None	330,746 (87.7)	63,106 (89.8)	267,640 (87.2)	
At minimal exertion	42,194 (11.2)	6,445 (9.2)	35,749 (11.7)	
At rest	4,167 (1.1)	750 (1.1)	3,417 (1.1)	
Admission from				<0.001
Admitted from home	366,018 (97.1)	67,817 (96.5)	298,201 (97.2)	
Acute care hospital	4,282 (1.1)	891 (1.3)	3,391 (1.1)	
Nursing home	5,635 (1.5)	1,256 (1.8)	4,379 (1.4)	
Other	1,172 (0.3)	337 (0.5)	835 (0.3)	
Cognitive deterioration	13,657 (3.6)	3,258 (4.6)	10,399 (3.4)	<0.001
Functional status				<0.001
Independent	348,744 (92.5)	63,972 (91.0)	284,722 (92.8)	
Partially dependent	22,100 (5.9)	4,713 (6.7)	17,387 (5.7)	
Totally dependent	6,263 (1.7)	1,616 (2.3)	4,647 (1.5)	
RAI mean [SD]	21.7 [7.4]	20.8 [8.0]	21.9 [7.3]	<0.001
RAI				<0.001
Robust (≤20)	138,068 (36.6)	35,312 (50.2)	102,756 (33.5)	
Normal (21–29)	199,125 (52.8)	27,235 (38.7)	171,890 (56.0)	
Frail (30–39)	30,669 (8.1)	5,484 (7.8)	25,185 (8.2)	
Very frail (≥40)	9,245 (2.5)	2,270 (3.2)	6,975 (2.3)	
30-day mortality	4,195 (1.1)	806 (1.1)	3,389 (1.1)	0.350
6-month mortality	13,870 (3.7)	2,617 (3.7)	11,253 (3.7)	0.493
1-year mortality	22,212 (5.9)	3,959 (5.6)	18,253 (5.9)	0.001

*Percentages are by row, the rest are by column.

RAI indicates Risk Analysis Index; SD, standard deviation.

Bias Assessment of the RAI_{Base}

Race-stratified model performance metrics are tabulated for the RAI_{Base} and 3 RAI variants at the RAI 30 threshold (Fig. 1) and RAI 36 threshold (Fig. 3). Bias is quantified as the difference in each model performance metric between Black and White cases. Green and red shading represent bias in favor of Black or White cases, respectively, whereas gray shading represents the absence of bias. At the RAI 30 threshold, the RAI_{Base} performed as good or better among Black cases across all performance metrics versus White, with a 2.8% higher true positive rate, equivalent true negative rate, and 1.4% greater balanced accuracy (Fig. 1). At the RAI 36 threshold, the patterns of bias are more pronounced: Black cases had 4.5% higher true positive rates but 0.5% lower true negative rates, yielding 2.0% higher balanced accuracy. The differences between the 2 thresholds are further illustrated by the ROC curves, which favor Black cases across most RAI thresholds (Fig. 2A), but converge and cross at higher RAI values (Fig. 2B).

Comparison of Model Performance and Bias Across the RAI Variants

Differences in model performance between the RAI variants are quantified for both thresholds in Figure 1 and Figure 3, where green and orange shading represent increased and decreased performance, respectively, in each metric compared to the RAI_{Base}. At the RAI 30 threshold (Fig. 1), the $RAI_{Balanced}$ demonstrated discrimination equivalent to the RAI_{Base} (c = 0.837) with a slightly higher balanced accuracy among both Black (Δ_{Acc} = 0.1%) and White (Δ_{Acc} = 0.2%) cases, producing a simultaneous increase in the true positive rate and decrease in true negative rate. However, these differences were small. Changes in model performance for the $RAI_{Age\text{-}adjusted}$ were more complex: Overall discrimination increased (c = 0.878), but the balanced accuracy among White cases decreased 0.3%, representing increased disparity between Black and White cases. Finally, the RAI_{Fair} demonstrated the lowest overall discrimination (c = 0.765) with large reductions in balanced accuracy for both White (Δ_{BalAcc} = −8.1%) and Black (Δ_{BalAcc} = −10.0%) cases, effectively reversing the direction of bias—Black cases had worse detection of true positives and better detection of true negatives.

At the threshold of RAI 36 (Fig. 3), the magnitude of bias between Black and White cases was both larger and more heterogeneous. For example, Black cases now had better detection of true positives (Δ_{Sens} = 4.5%) and worse detection of true negatives (ΔSpec = −0.5%). The bias between Black and White cases changed direction and magnitude across the 3 RAI variants, with RAI_{Fair} demonstrating the least biased equal opportunity ratio (0.90), but with better true positive detection in White cases than Black.

Results at the threshold of RAI 26 are provided in the Supplement (Supplemental Table 1, http://links.lww.com/AOSO/A401).

Sensitivity Analyses at the Threshold that Optimizes Sensitivity and Specificity

At the threshold jointly maximizing sensitivity and specificity, model metrics showed similar patterns to those at the RAI 30 threshold (Fig. 1) where the RAI_{Base} performed as well or better among Black cases versus White across all performance metrics—consistent with the ROC curves in Figure 2. At this threshold, the equal opportunity and predictive equality ratios were 1.01 and 1.05, respectively, and none of the RAI variants effectively reduced the bias favoring Black cases.

		Bias Assessment				Comparative Model Performance		
		RAI_{Base}	$RAI_{Balanced}$	$RAI_{Age-adj}$	RAI_{Fair}	Δ from RAI_{Base}		
		Statistic	Statistic	Statistic	Statistic	$\Delta RAI_{Balanced}$	$\Delta RAI_{Age-adj}$	ΔRAI_{Fair}
Accuracy	Black	0.905	0.903	0.890	0.923	-0.2%	-1.5%	1.8%
	White	0.904	0.901	0.907	0.900	-0.3%	0.3%	-0.4%
	Bias	0.1%	0.2%	-1.7%	2.3%	0.1%	-1.8%	2.2%
Balanced Accuracy	Black	0.751	0.752	0.760	0.651	0.1%	0.9%	-10.0%
	White	0.737	0.739	0.734	0.656	0.2%	-0.3%	-8.1%
	Bias	1.4%	1.3%	2.6%	-0.5%	-0.1%	1.2%	-1.9%
True Positive Rate (Sensitivity)	Black	0.584	0.588	0.619	0.358	0.4%	3.5%	-22.6%
	White	0.556	0.564	0.548	0.392	0.8%	-0.8%	-16.4%
	Bias	2.8%	2.4%	7.1%	-3.4%	-0.4%	4.3%	-6.2%
True Negative Rate (Specificity)	Black	0.917	0.915	0.900	0.945	-0.2%	-1.7%	2.8%
	White	0.917	0.913	0.921	0.920	-0.4%	0.4%	0.3%
	Bias	0.0%	0.2%	-2.1%	2.5%	0.2%	-2.1%	2.5%
False Positive Rate	Black	0.083	0.085	0.099	0.055	0.2%	1.6%	-2.8%
	White	0.083	0.087	0.079	0.080	0.4%	-0.4%	-0.3%
	Bias	0.0%	0.2%	-2.0%	2.5%	0.2%	-2.0%	2.5%
False Negative Rate	Black	0.416	0.411	0.381	0.642	-0.5%	-3.5%	22.6%
	White	0.444	0.436	0.452	0.608	-0.8%	0.8%	16.4%
	Bias	2.8%	2.5%	7.1%	-3.4%	-0.3%	4.3%	-6.2%
F1 score	Black	0.314	0.311	0.296	0.257	-0.3%	-1.8%	-5.7%
	White	0.298	0.294	0.302	0.224	-0.4%	0.4%	-7.4%
	Bias	1.6%	1.7%	-0.6%	3.3%	0.1%	-2.2%	1.7%
MCC	Black	0.315	0.313	0.304	0.230	-0.2%	-1.1%	-8.5%
	White	0.296	0.294	0.298	0.203	-0.2%	0.2%	-9.3%
	Bias	1.9%	1.9%	0.6%	2.7%	0.0%	-1.3%	0.8%
Brier score	Black	0.0309	0.0308	0.0309	0.0339	0.0001	0.0000	0.0030
	White	0.0306	0.0306	0.0306	0.0335	0.0000	0.0000	0.0029
	Bias	-0.0003	-0.0002	-0.0003	-0.0004	0.0001	0.0000	-0.0001
Harrel's C	Black	0.8546	0.833	0.8334	0.7614	-0.0216	-0.0212	-0.0932
	White	0.8335	0.8545	0.8546	0.7867	0.0210	0.0211	-0.0468
	Bias	0.0211	-0.0215	-0.0212	-0.0253	0.0426	0.0423	0.0464
Equal Opportunity[a]		1.05	1.04	1.13	0.91	-1.0%	8.0%	-14.0%
Predictive Equality[b]		1.00	1.00	0.98	1.03	0.0%	-2.0%	3.0%

For comparisons of racial bias, indicates metrics with better performance among Black patients than White, while indicates metrics with better performance among White patients than Black. For comparisons of various RAI versions, indicates improvements in predictive performance, while indicates reductions in predictive performance.

[a]Ratio of Sensitivity between Black and White cases – ratio higher than 1 suggests better sensitivity in Black patients than White, ratio less than 1 suggests better sensitivity in White patients than Black.

[b]Ratio of Specificity between Black and White cases – ratio higher than 1 suggest better specificity in Black patients than White, ratio less than 1 suggests better specificity in White patients than Black.

Abbreviations: RAI, Risk Analysis Index; MCC, Matthew's correlation coefficient

Figure 1. Measures of predictive performance between Black and White cases for the threshold of detectable frailty (RAI 30, 6.7% predicted 180-day mortality).

Sensitivity analysis of NSQIP data

Examining an RAI 30 threshold in NSQIP data, Black cases had better specificity (1.5%) but worse sensitivity (–1.3%), yielding similar balanced accuracies (73.3% in Black cases vs. 73.0% in White) and better accuracy in Black cases (1.5%, Supplemental Table 2, http://links.lww.com/AOSO/A401). At this threshold, the equal opportunity and predictive equality ratios were 0.96 and 1.02, respectively. The ROC curve showed better predictive performance in Black cases across most RAI thresholds (Supplemental Figure 1, http://links.lww.com/AOSO/A401). However, the ROC curves for Black and White cases crossed at the upper range of RAI values.

DISCUSSION

Given the wide and expanding application of the RAI in clinical practice, this study was designed to probe for any algorithmic bias according to race. We found that, at the clinical cutoff used by the Surgical Pause program (RAI 30), the RAI_{Base} demonstrated a slight algorithmic bias in favor of Black cases. None of the other RAI variants noticeably eliminated bias, improved parity across both true positives and true negatives, or improved overall model performance. The pattern favoring Black cases remained similar at various thresholds. We consider the minor differences found to be reassuring, and our data suggest that the RAI can continue to be used for preoperative decision-making. However, the ROC curves for Black and White cases crossed at the upper range of RAI values, suggesting that certain clinical thresholds could yield predictions that disadvantage Black patients. We recommend the existing threshold of 30 for the VASQIP-derived RAI, which corresponds to a threshold of 37 for the clinical questionnaire version of the RAI currently used in Surgical Pause screening.[7] Although our data limited authoritative comparisons for non-Black minority groups, absent compelling evidence of poor RAI performance in these groups, we consider the benefits of the RAI to outweigh the potential risks (especially considering that RAI-based frailty screening can produce an almost threefold reduction in postoperative mortality).[15]

Ideally, bias mitigation strategies should reduce the advantage observed in 1 group, not degrade overall model performance for the disadvantaged group, and performance across groups should not be markedly worse. Our data demonstrate that this is not always possible. For example, the RAI_{Fair} had markedly

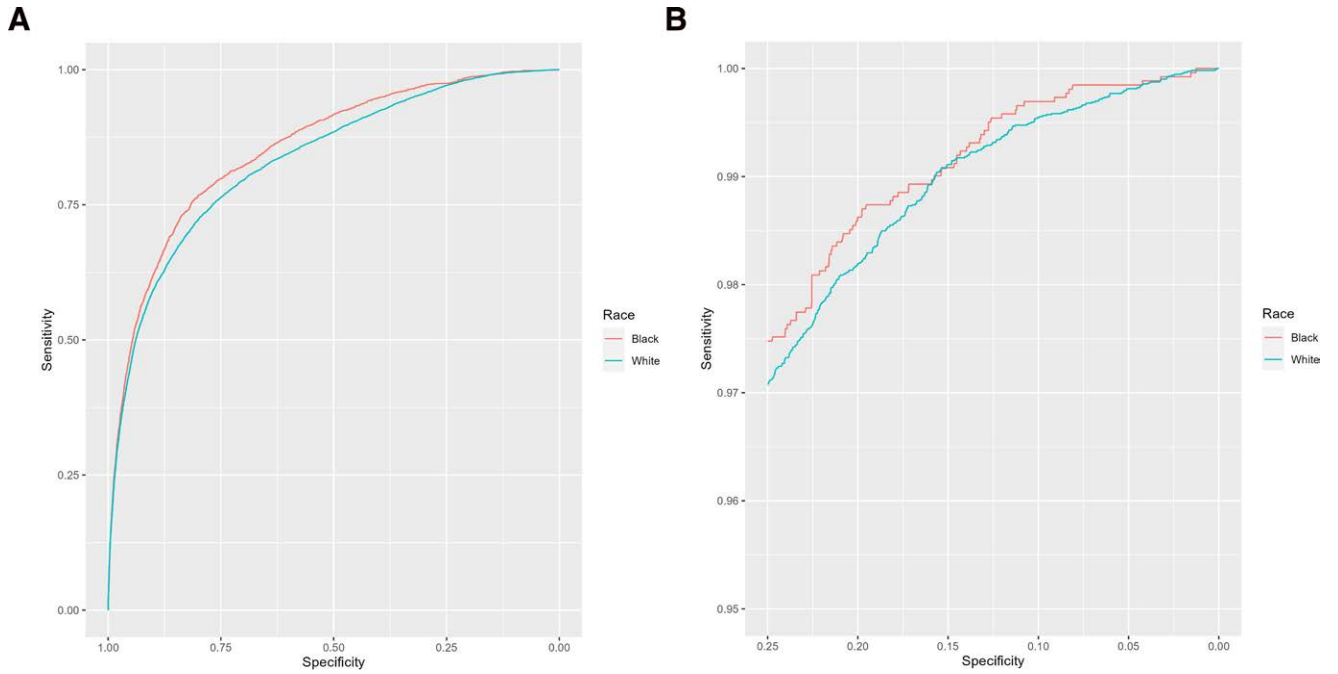

Figure 2. Receiver operating curves for RAI base among White and Black cases. RAI indicates Risk Analysis Index.

worse true positive rates than the RAI_{Base} among both White and Black cases, making this model worse for both races. Although the $RAI_{Balanced}$ had slightly improved true positive rates among Black and White cases, the true negative rates were lower in both racial groups, and a small bias emerged in favor of Black patients. Although an argument could be made to use the $RAI_{Balanced}$ because of its slightly improved equal opportunity ratio, we contend that the change is too small to warrant revising the scoring systems currently in use. These findings suggest that RAI_{Base} remains acceptable for use in both statistical analysis of surgical cohorts and quality improvement programs, such as the Surgical Pause—at least for patients who are White or Black.

To the extent that the RAI_{Base} had better predictive performance, particularly in sensitivity, among Black patients versus White, the RAI increases the likelihood that Black patients receive appropriate prehabilitation and accurate counseling about surgical risk—which may help alleviate racial disparities in surgical outcomes. Algorithms should include race/ethnicity as a predictor when it increases the chance of appropriate care for an at-risk group but not when it reduces their chance of appropriate care.[35] In our case, while race or ethnicity are not included in the RAI, Black patients have an increased opportunity for appropriate treatment when the RAI is used.

Our findings highlight issues with "bias-reducing" statistical methods. Namely, in the pursuit of equity, these methods can worsen a model's predictive ability.[36,37] In our case, the RAI_{Fair} achieved higher specificity in Black patients, but drastically reduced sensitivity for both Black and White patients. Such changes are in direct contrast with the express purpose of the RAI, to accurately capture frailty in patients considering surgery and enable either preoperative intervention or alternative, nonsurgical treatment.[15] The impact of bias mitigation strategies also varies across model performance metrics, which highlights the importance of defining the most clinically relevant metric. For example, if priority is placed on identifying frailty and mitigating its associated risks, efforts should be focused on reducing sensitivity bias, even at the expense of specificity. Alternatively, if the priority is placed on reducing the inconvenience, expense, and potential harm of falsely categorizing patients as frail, then efforts should focus on maximizing specificity. Global measures of performance, such as balanced accuracy, F1, or MCC, may not always reflect what is of greatest clinical importance.

Concerns about racial bias in predictive algorithms are well-founded and healthcare algorithms need to be screened for potential racial disparity. To our knowledge, the RAI is the first frailty measure to be assessed for racial bias, and various non-frailty comorbidity indices, such as the Charlson,[38] Elixhauser,[39] and Gagne[40] scores, have also not yet been tested for racial bias. Until such testing occurs, we advise caution when adjusting for comorbidity indices to estimate the causal effects of racial discrimination, social risk factors, or any nonclinical risk factor heavily related to race. Without definitive evidence of racial bias or its absence, researchers should perform sensitivity analyses with and without comorbidity adjustment when investigating such factors.

More broadly, meaningful differences in medical conditions persist across certain groups. The risk of breast cancer is not the same between men and women.[41] The risk of sickle-cell anemia is not the same between Black and White patients.[42] As such, it may not be possible to expect a model to both (1) make the same predictions of risk across both groups and (2) have equal false positive and false negative rates between both groups.[43,44] Decisions about trade-offs in such situations require input from physicians, statisticians, and patients. While using race in clinical algorithms can reinforce disparities or racial stereotypes, "excluding factors that have proven to be predictive, albeit highly imperfect… can sometimes carry great human costs".[37] Patients expect predictions of their risk to be as accurate as possible, given the information we have. Reducing accuracy in the name of equity accomplishes neither.

Finally, predictive algorithms can have ROC curves that cross,[45] so the choice of threshold is critical for many algorithmic fairness measures. A model can be racially fair or more predictive in the minority group across a range of thresholds but be racially biased when other thresholds are chosen. Examining and comparing different thresholds is a critical component of developing fair and equitable predictive algorithms and, in some situations, might be more effective than "bias-reducing" statistical methods. Further research into bias-reducing techniques is needed, especially for clinical risk predictions.

		Bias Assessment				Comparative Model Performance		
		RAI_{Base}	$RAI_{Balanced}$	$RAI_{Age\text{-}adj}$	RAI_{Fair}	Δ from RAI_{Base}		
		Statistic	Statistic	Statistic	Statistic	$\Delta RAI_{Balanced}$	$\Delta RAI_{Age\text{-}adj}$	ΔRAI_{Fair}
Accuracy	Black	0.944	0.944	0.938	0.957	0.0%	-0.6%	1.3%
	White	0.947	0.947	0.948	0.956	0.0%	0.1%	0.9%
	Bias	-0.3%	-0.3%	-1.0%	0.1%	0.0%	-0.7%	0.4%
Balanced Accuracy	Black	0.688	0.687	0.704	0.556	-0.1%	1.6%	-13.2%
	White	0.668	0.665	0.664	0.562	-0.3%	-0.4%	-10.6%
	Bias	2.0%	2.2%	4.0%	-0.6%	0.2%	2.0%	-2.6%
True Positive Rate (Sensitivity)	Black	0.412	0.408	0.451	0.123	-0.4%	3.9%	-28.9%
	White	0.367	0.36	0.359	0.136	-0.7%	-0.8%	-23.1%
	Bias	4.5%	4.8%	9.2%	-1.3%	0.3%	4.7%	-5.8%
True Negative Rate (Specificity)	Black	0.964	0.965	0.957	0.989	0.1%	-0.7%	2.5%
	White	0.969	0.97	0.97	0.988	0.1%	0.1%	1.9%
	Bias	-0.5%	-0.5%	-1.3%	0.1%	0.0%	-0.8%	0.6%
False Positive Rate	Black	0.036	0.035	0.043	0.011	-0.1%	0.7%	-2.5%
	White	0.031	0.03	0.03	0.012	-0.1%	-0.1%	-1.9%
	Bias	-0.5%	-0.5%	-1.3%	0.1%	0.0%	-0.8%	0.6%
False Negative Rate	Black	0.588	0.592	0.549	0.877	0.4%	-3.9%	28.9%
	White	0.633	0.64	0.641	0.864	0.7%	0.8%	23.1%
	Bias	4.5%	4.8%	9.2%	-1.3%	0.3%	4.7%	-5.8%
F1 score	Black	0.353	0.353	0.353	0.175	0.0%	0.0%	-17.8%
	White	0.335	0.333	0.334	0.186	-0.2%	-0.1%	-14.9%
	Bias	1.8%	2.0%	1.9%	-1.1%	0.2%	0.1%	-2.9%
MCC	Black	0.328	0.327	0.331	0.174	-0.1%	0.3%	-15.4%
	White	0.309	0.307	0.308	0.180	-0.2%	-0.1%	-12.9%
	Bias	1.9%	2.0%	2.3%	-0.6%	0.1%	0.4%	-2.5%
Brier score	Black	0.0309	0.0308	0.0309	0.0339	0.0001	0.0000	0.0030
	White	0.0306	0.0306	0.0306	0.0335	0.0000	0.0000	0.0029
	Bias	-0.0003	-0.0002	-0.0003	-0.0004	0.0001	0.0000	-0.0001
Harrel's C	Black	0.8546	0.833	0.8334	0.7614	-0.0216	-0.0212	-0.0932
	White	0.8335	0.8545	0.8546	0.7867	0.0210	0.0211	-0.0468
	Bias	0.0211	-0.0215	-0.0212	-0.0253	0.0426	0.0423	0.0464
Equal Opportunity[a]		1.12	0.86	1.26	0.90	-26.0%	14.0%	-22.0%
Predictive Equality[b]		1.00	1.00	0.99	1.00	0.0%	-1.0%	0.0%

For comparisons of racial bias, indicates metrics with better performance among Black patients than White, while indicates metrics with better performance among White patients than Black. For comparisons of various RAI versions, indicates improvements in predictive performance, while indicates reductions in predictive performance.

[a]Ratio of Sensitivity between Black and White cases – ratio higher than 1 suggests better sensitivity in Black patients than White, ratio less than 1 suggests better sensitivity in White patients than Black.

[b]Ratio of Specificity between Black and White cases – ratio higher than 1 suggest better specificity in Black patients than White, ratio less than 1 suggests better specificity in White patients than Black.

Abbreviations: RAI, Risk Analysis Index; MCC, Matthew's correlation coefficient

Figure 3. Measures of predictive performance between Black and White cases for the threshold of phenotypically obvious frailty (RAI 36, 15% predicted 180-day mortality).

Limitations

We limited our investigation to Black and White patients for 2 main reasons: (1) preexisting evidence of surgical disparities between Black and White patients and (2) insufficient cases of 180-day mortality for other race/ethnicity groups in our data. Because surgical registries are designed to be systematic samples of procedures,[25,46] they will necessarily reflect the small number of racial minorities undergoing surgery. Furthermore, for Asian patients, our data lacked the necessary granularity. Meaningful differences between subgroups of Asian Americans exist, and combining these subgroups into one "Asian" group can dramatically skew results and mask meaningful health disparities.[47] Future research should examine racial disparities in risk prediction for Hispanic, Asian, Pacific Islander, and Native American populations, and should use larger cohorts with more refined racial data to do so. Further explorations of racial disparities might include data on social risk factors or genetic ancestry.

CONCLUSIONS

We examined racial bias within the RAI, a commonly used measure of frailty. We found that, across various thresholds, the RAI tends to predict mortality among Black patients better than it predicts mortality among White patients. Attempts to reduce bias with the RAI failed to (1) noticeably reduce bias, (2) improve parity across true positives and true negatives, or (3) improve overall model performance. As existing bias-reducing techniques were not effective, further research into bias-reducing techniques is needed, especially for clinical risk predictions. We recommend using the RAI for both statistical analysis of surgical cohorts and quality improvement programs, such as the Surgical Pause.

ACKNOWLEDGMENTS

We thank Brian Ziebart, PhD and Omid Memarrast, PhD for both sharing Python code from their previous work and assisting the authors in implementing it.

REFERENCES

1. Hsu J, Price M, Vogeli C, et al. Bending the spending curve by altering care delivery patterns: the role of care management within a pioneer ACO. *Health Aff (Millwood)*. 2017;36:876–884.
2. Obermeyer Z, Powers B, Vogeli C, et al. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. 2019;366:447–453.

3. Chen I, Johansson FD, Sontag D. Why Is My Classifier Discriminatory? In: Bengio S, Wallach H, Larochelle H, et al. editors. *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2018. Available at: https://proceedings.neurips.cc/paper_files/paper/2018/file/1f1baa5b8edac74eb4eaa329f14a0361-Paper.pdf
4. Fried TR, Bradley EH, Towle VR, et al. Understanding the treatment preferences of seriously ill patients. *N Engl J Med*. 2002;346:1061–1066.
5. Panayi AC, Orkaby AR, Sakthivel D, et al. Impact of frailty on outcomes in surgical patients: a systematic review and meta-analysis. *Am J Surg*. 2019;218:393–400.
6. McIsaac DI, Wong CA, Huang A, et al. Derivation and validation of a generalizable preoperative frailty index using population-based health administrative data. *Ann Surg*. 2019;270:102–108.
7. Arya S, Varley P, Youk A, et al. Recalibration and external validation of the risk analysis index: a surgical frailty assessment tool. *Ann Surg*. 2020;272:996–1005.
8. Yan Q, Kim J, Hall DE, et al. Association of frailty and the expanded operative stress score with preoperative acute serious conditions, complications and mortality in males compared to females: a retrospective observational study. *Ann Surg*. 2023;277:e294–e304.
9. George EL, Hall DE, Youk A, et al. Association between patient frailty and postoperative mortality across multiple noncardiac surgical specialties. *JAMA Surg*. 2020;156:e205152.
10. Shinall MC, Youk A, Massarweh NN, et al. Association of preoperative frailty and operative stress with mortality after elective vs emergency surgery. *JAMA Netw Open*. 2020;3:e2010358.
11. Shinall MC, Arya S, Youk A, et al. Association of preoperative patient frailty and operative stress with postoperative mortality. *JAMA Surg*. 2020;155:e194620.
12. Yan Q, Hall DE, Shinall MC, et al. Association of patient frailty and operative stress with postoperative mortality: no such thing as low-risk operations in frail adults. *J Am Coll Surg*. 2020;231:S134–S135.
13. McIsaac DI, Aucoin SD, van Walraven C. A Bayesian comparison of frailty instruments in noncardiac surgery: a cohort study. *Anesth Analg*. 2021;133:366–373.
14. Varley PR, Buchanan D, Bilderback A, et al. Association of routine preoperative frailty assessment with 1-year postoperative mortality. *JAMA Surg*. 2023;158:475–483.
15. Hall DE, Arya S, Schmid KK, et al. Association of a frailty screening initiative with postoperative survival at 30, 180, and 365 days. *JAMA Surg*. 2017;152:233–240.
16. Dicpinigaitis AJ, Khamzina Y, Hall DE, et al. Adaptation of the risk analysis index for frailty assessment using diagnostic codes. *JAMA Netw Open*. 2024;7:e2413166.
17. Walker P. *Risk Score Helps Patients and Doctors Make Informed Decisions About Whether to Go Ahead with Surgery*. EpicShare, 2021 [cited Jan 16, 2024]; Available at: https://www.epicshare.org/share-and-learn/risk-score-helps-patients-and-doctors-make-informed-decisions-about-whether-to-go-ahead-with-surgery.
18. *The Surgical Pause Practice Adopted as National Program*, 2023 [cited Feb 16, 2024]; Available at: https://www.hsrd.research.va.gov/impacts/surgical-pause.cfm.
19. U. S. Department of Veterans Affairs. *SAGE QUERI What Matters -- Surgical Pause*. VA Healthc. - VISN 4. [cited Jun 27, 2024]; Available at: https://www.visn4.va.gov/VISN4/SAGE/matters.asp.
20. John M, Joint Commission. *Eisenberg Patient Safety and Quality Awards*. Trust. Partn. Patient Care Jt. Com, 2024 [cited Jul 3, 2024]; Available at: https://www.jointcommission.org/resources/awards/john-m-eisenberg-patient-safety-and-quality-award/.
21. Geronimus AT, Hicken M, Keene D, et al. "Weathering" and age patterns of allostatic load scores among blacks and whites in the United States. *Am J Public Health*. 2006;96:826–833.
22. Astone NM, Ensminger M, Juon HS. Early adult characteristics and mortality among inner-city African American women. *Am J Public Health*. 2002;92:640–645.
23. Geronimus AT, Bound J, Waidmann TA, et al. Excess mortality among blacks and whites in the United States. *N Engl J Med*. 1996;335:1552–1558.
24. Navathe AS, Park SH, Hearn CM, et al. *(In)equity in Risk Prediction: Examining and Mitigating Racial Bias in the Veterans Affairs Care Assessment Needs (CAN) Risk Model*, 2023 [cited Feb 17, 2023]; Available at: https://www.hsrd.research.va.gov/meetings/2023/abstract-display.cfm?AbsNum=1004.
25. Massarweh NN, Kaji AH, Itani KMF. Practical guide to surgical data sets: Veterans Affairs Surgical Quality Improvement Program (VASQIP). *JAMA Surg*. 2018;153:768–769.
26. Ghaferi AA, Schwartz TA, Pawlik TM. STROBE reporting guidelines for observational studies. *JAMA Surg*. 2021;156:577–578.
27. Best WR, Khuri SF, Phelan M, et al. Identifying patient preoperative risk factors and postoperative adverse events in administrative databases: results from the Department of Veterans Affairs National Surgical Quality Improvement Program. *J Am Coll Surg*. 2002;194:257–266.
28. Parikh RB, Linn KA, Park SH, et al. Age as a mechanism of racial bias in mortality prediction models. [unpublished manuscript]
29. Rezaei A, Fathony R, Memarrast O, et al. Fairness for robust log loss classification. *Proc AAAI Conf Artif Intell*. 2020;34:5511–5518.
30. Perkins NJ, Schisterman EF. The inconsistency of "Optimal" cutpoints obtained using two criteria based on the receiver operating characteristic curve. *Am J Epidemiol*. 2006;163:670–675.
31. Mower JP. PREP-Mt: predictive RNA editor for plant mitochondrial genes. *BMC Bioinf*. 2005;6:96.
32. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta*. 1975;405:442–451.
33. Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Med Imaging*. 2015;15:29.
34. Yuan C, Linn KA, Hubbard RA. Algorithmic fairness of machine learning models for Alzheimer disease progression. *JAMA Netw Open*. 2023;6:e2342203.
35. Vaughan Sarrazin MS. Balancing statistical precision with societal goals to reduce health disparities using clinical support tools. *JAMA Netw Open*. 2023;6:e2331140.
36. Pfohl SR, Foryciarz A, Shah NH. An empirical characterization of fair machine learning for clinical risk prediction. *J Biomed Inform*. 2021;113:103621.
37. Pierson E. Accuracy and equity in clinical risk prediction. *N Engl J Med*. 2024;390:100–102.
38. Charlson ME, Pompei P, Ales KL, et al. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J Chronic Dis*. 1987;40:373–383.
39. Elixhauser A, Steiner C, Harris DR, et al. Comorbidity measures for use with administrative data. *Med Care*. 1998;36:8–27.
40. Gagne JJ, Glynn RJ, Avorn J, et al. A combined comorbidity score predicted mortality in elderly patients better than existing scores. *J Clin Epidemiol*. 2011;64:749–759.
41. Ly D, Forman D, Ferlay J, et al. An international comparison of male and female breast cancer incidence rates. *Int J Cancer*. 2013;132:1918–1926.
42. Pokhrel A, Olayemi A, Ogbonda S, et al. Racial and ethnic differences in sickle cell disease within the United States: from demographics to outcomes. *Eur J Haematol*. 2023;110:554–563.
43. Kleinberg J. Inherent Trade-Offs in Algorithmic Fairness. In: *Abstracts of the 2018 ACM International Conference on Measurement and Modeling of Computer Systems*. Irvine, CA, USA: ACM, 2018 [cited 2023 Dec 26]. p. 40.Available at: https://dl.acm.org/doi/10.1145/3219617.3219634
44. Chouldechova A. Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. *Big Data*. 2017;5:153–163.
45. Gigliarano C, Figini S, Muliere P. Making classifier performance comparisons when ROC curves intersect. *Comput Stat Data Anal*. 2014;77:300–312.
46. Raval MV, Pawlik TM. Practical guide to surgical data sets: national surgical quality improvement program (NSQIP) and pediatric NSQIP. *JAMA Surg*. 2018;153:764–765.
47. Madhusoodanan J. Researchers are working to disaggregate Asian American health data—here's why it's long overdue. *JAMA*. 2024;331:1350–1353.