Predictive Performance and Algorithmic Equity in a Comprehensive Primary Care Plus Two-Step Risk Stratification Program

Gary E. Weissman,[1,2,3,4] Andrew Crane-Droesch,[1,5] Asaf Hanish,[5] Corey Chivers,[5] J.T. Howell,[6] Maryanne Peifer,[6] Sebastian Haines,[7] Matthew J. Press,[3,7,8] Anna U. Morgan[3,8]

[1] Palliative and Advanced Illness Research (PAIR) Center, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania [2] Pulmonary, Allergy, and Critical Care Division, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania [3] Leonard Davis Institute of Health Economics, University of Pennsylvania, Philadelphia, Pennsylvania [4] Penn Institute for Biomedical Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania [5] Penn Predictive Healthcare, Penn Medicine, Philadelphia, Pennsylvania [6] Department of Family Medicine and Community Health, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania [7] Primary Care Service Line, Penn Medicine, Philadelphia, Pennsylvania [8] Division of General Internal Medicine, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania

pair
Palliative and Advanced Illness Research Center

Background

- Practices participating in Track 2 of the Comprehensive Primary Care Plus (CPC+) program are required to risk stratify patients in two steps: first, with a computer-based algorithm, and second, with a clinician adjustment based on their intuition and knowledge of the patient

- However, it is unknown whether a widely implemented and proprietary risk stratification algorithm from EPIC would perform and, secondarily, how it would perform by patient race

Objective

- Evaluate the performance and algorithmic equity of EPIC's proprietary general risk score among Medicare Fee-for-Service (FFS) beneficiaries seen at clinics participating in Track 2 of CPC+ at a large academic medical center
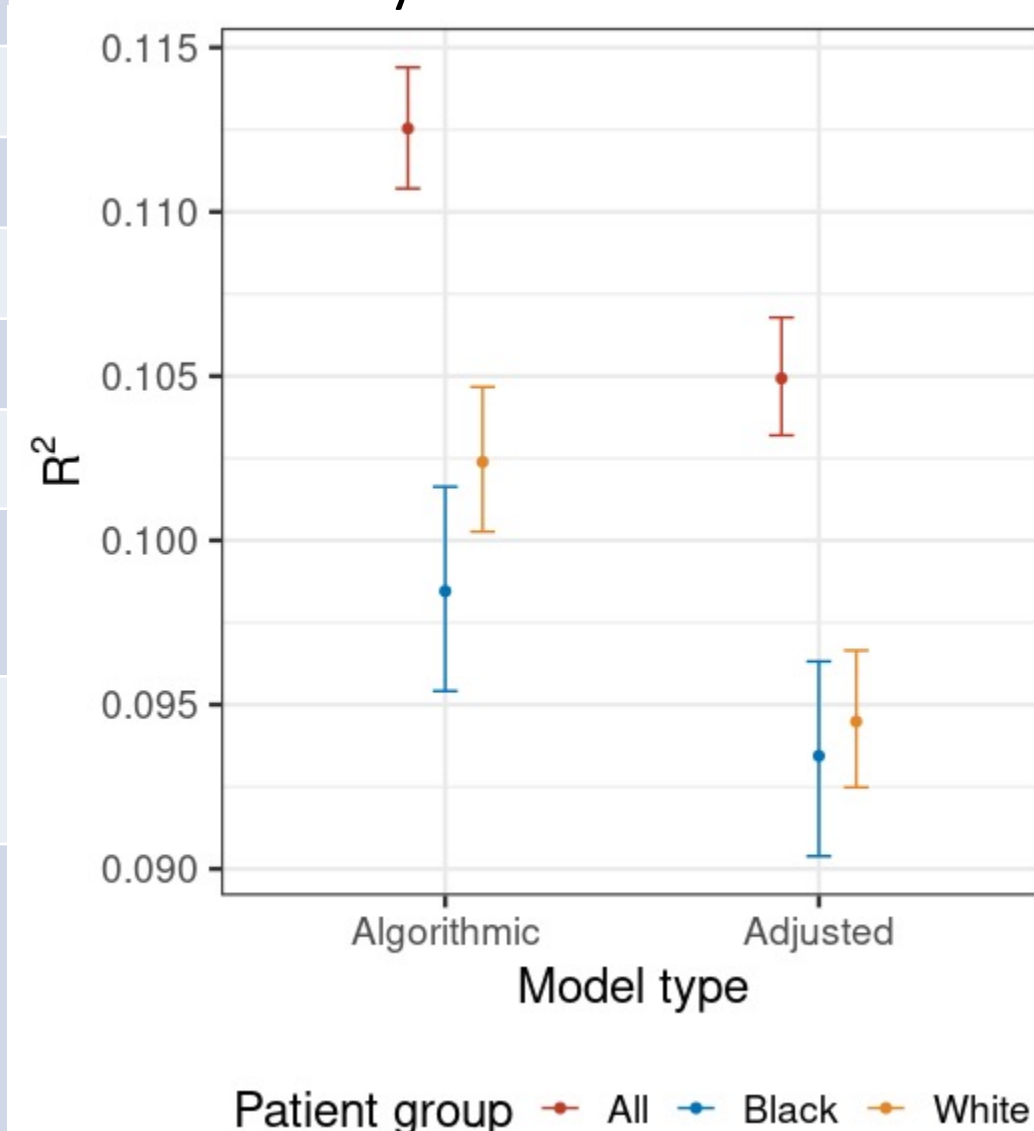
Methods

- Analyze data from Black and white FFS beneficiaries from September 2017 through December 2020 seen at Penn Medicine clinics

- The EPIC General Risk Score (range 0 to 15) and all clinician alterations to that score were extracted from the electronic health record to determine the algorithmic and adjusted scores, respectively

- Scores were evaluated as categorical inputs to a logistic regression model to calculate a composite outcome of hospitalization or emergency room visit within 6 months of the score

- Model performance was evaluated using the R^2 overall and by patient-reported race and inspection of calibration curves
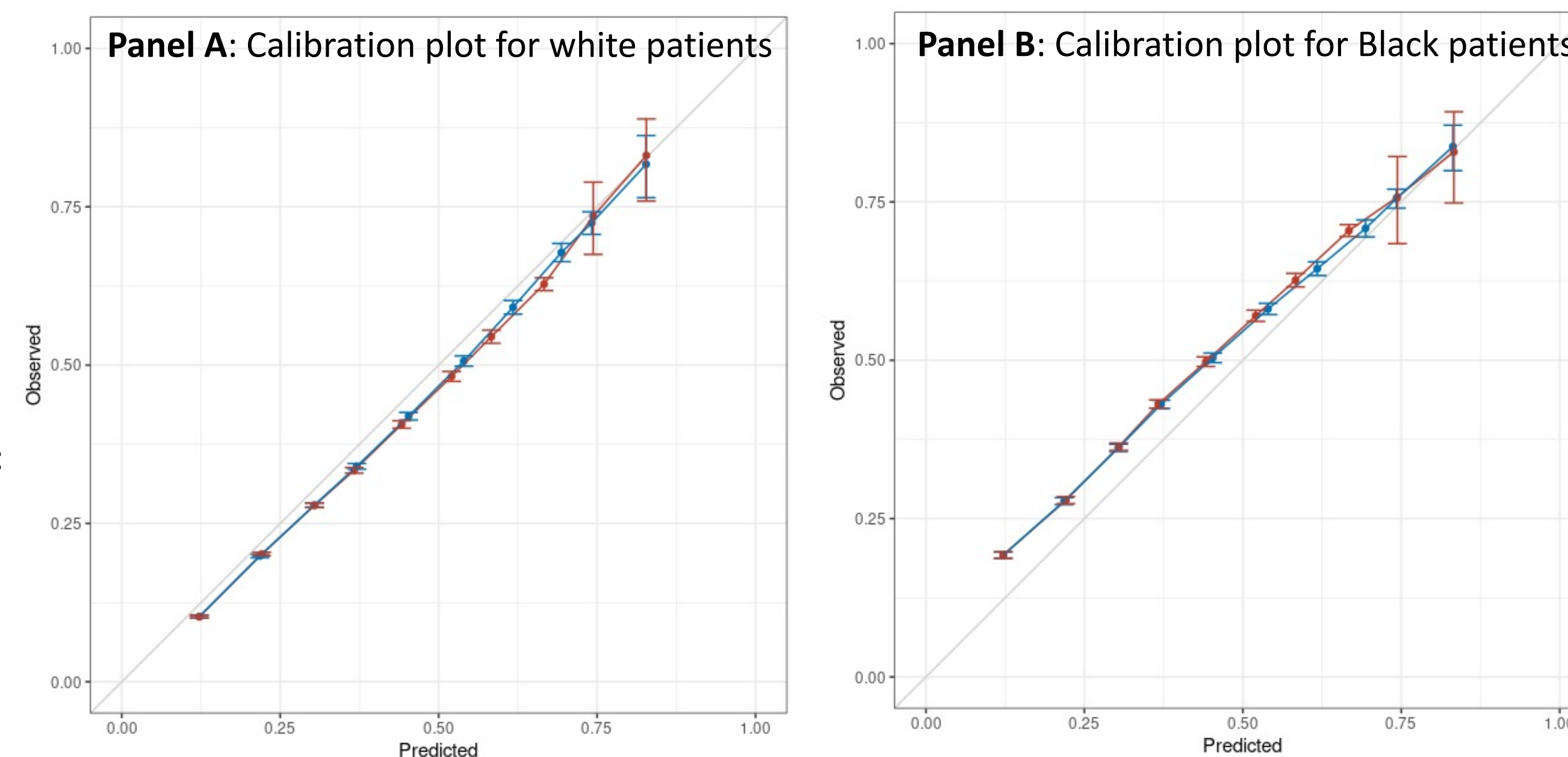
Results

Table 1

Table 1	N (%)
Patients	94,546
Race	
White	73,403 (78)
Black	21,143 (22)
Women	54,661 (58)
Age (years; median [IQR])	71 (66 to 78)
Algorithmic score, median (IQR)	2 (1.4 to 3)
Adjusted score, median (IQR)	2 (1.4 to 2)
Number of calculated scores per patient, median (IQR)	3 (2 to 7)

Figure 1: Comparison of R^2 values by model and race

Figure 2: Calibration plots of adjusted and algorithmic models by patient race (A: white patients, B: Black patients)

Panel A: Calibration plot for white patients

Panel B: Calibration plot for Black patients

- 463,985 scores were evaluated over the study period for 94,546 patients (**Table 1**)

- Clinician adjustment worsened predictive performance in all groups and slightly reduced the performance gap between white and Black patients, among whom all models performed worse (**Figure 1**)

- Both algorithmic and adjusted models were generally well calibrated but both model approaches tended to overestimate risk for white patients and underestimate risk for Black patients (**Figure 2**)

Limitations

- The study is limited to the Medicare Fee-for-Service population while the two-step risk prediction was applied to all patients at participating practices

- Patients who received care fragmented across health systems, who are more likely to be from minority groups, may have less information available for assessing performance of prediction models

- The study design does not account for interventions such as population health management programs, guided by the risk scores, that may have affected the outcome

- Small numbers of patients who did not identify as Black or white were not included in the analysis

- The original EPIC General Risk Score is not publicly available and so these results represent a proxy assessment

Conclusions and Policy Impact

- The score exhibits moderate performance on par with existing risk scores like the CMS-HCC

- The score exhibits better performance among white compared to Black patients and clinician adjustment worsens performance for all groups

- Further validation and increased transparency around such scores are needed before they should be mandated by CPC+ regulations

Penn LDI LEONARD DAVIS INSTITUTE of HEALTH ECONOMICS

50 YEARS 1967 2017 FIRST THEN. FIRST NOW.

Penn Institute for Biomedical Informatics