Assessing Pre-Existing Comorbidities as Predictors of Two-Year Long COVID Mortality Using Machine Learning and N3C Data

Shuntai Zhou, PhD[1]; Alexander Toppo, MD, MPH[2]; Serena S. Dasani, MD, MBA[3]; Prakash R. Ganesh, MD, MPH[4]; Claire Han, PhD, DNP, RN[5]; Nikki Reyes, PhD, MPH[6]; **Mahbubul (Shaon) Hasan, PhD[7]**

[1]University of North Carolina at Chapel Hill, Chapel Hill, NC; [2]Texas Tech University Health Sciences Center, El Paso, TX; [3]University of Texas Health Science Center at Houston, Houston, TX; [4]Case Western Reserve University School of Medicine, Cleveland, OH; [5]The Ohio State University College of Nursing, Columbus, OH; [6]University of Miami, Miami, FL; [7]NCATS & Axle Informatics, Rockville, MD

AIM AHEAD

NIH National Center for Advancing Translational Sciences

Health Data Science Training Program

Scan for more details!

Purpose

Clinical utility: Pinpoints patients with Post-acute sequelae of SARS-CoV-2 infection (**PASC, or Long COVID**) at highest mortality risk to enable precision follow-up and targeted comorbidity management.

Important problem: PASC is common, clinically heterogeneous, and associated with sustained excess mortality, yet robust post-PASC risk stratification tools are lacking.

Public health perspective: Defining high-risk subgroups can sharpen resource allocation and help reduce the long-term population burden of PASC.

- **Routinely available baseline clinical data** contain strong predictive signal for 2-year all-cause mortality after PASC diagnosis.
- **XGBoost** modestly improved mortality risk discrimination over Cox-based benchmarks, with comparable prediction error.
- The recalibrated XGBoost model may enable **individualized, time-specific risk estimation and clinically interpretable risk stratification**.

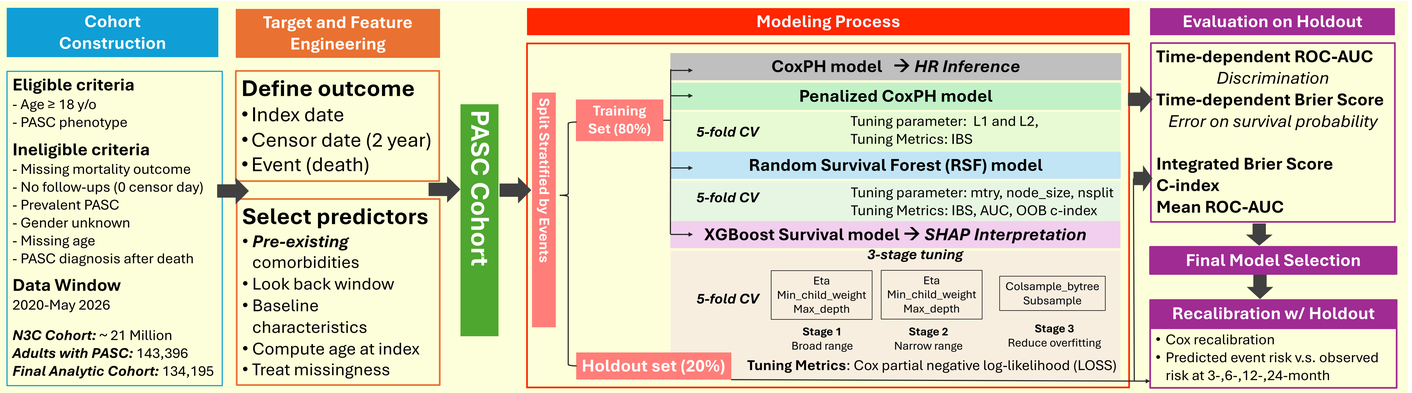


Table 1. Baseline clinical and non-clinical characteristics of the study participants. *Comorbidities with prevalence above 2% are included in this table

Characteristics	Overall (N=134,195)
Outcome	
Event	4319 (3.2%)
Median time to event/censor, days (IQR)	672 (347-731)
Non-clinical characteristics	
Median age at index, y/o (IQR)	55 (42-67)
Sex, male	45128 (33.6%)
Comorbidity/Exposure*	
Systemic corticosteroids	79157 (59.0%)
Prior covid diagnosis	75118 (56.0%)
Obesity	64351 (48.0%)
Hypertension	63738 (47.5%)
Chronic lung disease	47369 (35.3%)
Depression	42465 (31.6%)
Diabetes, uncomplicated	29623 (22.1%)
Substance use disorder	28507 (21.2%)
Diabetes, complicated	20676 (15.4%)
Other immunocompromised	20555 (15.3%)
Tobacco smoker	20016 (14.9%)
Coronary artery disease	17426 (13.0%)
Kidney disease	17276 (12.9%)
Rheumatologic disease	16450 (12.3%)
Heart failure	15759 (11.7%)
Malignant cancer	15137 (11.3%)
Mild liver disease	14634 (10.9%)
Congestive heart failure	10598 (7.9%)
Cerebrovascular disease	9956 (7.4%)
Prior remdesivir exposure	9396 (7.0%)
Peripheral vascular disease	8098 (6.0%)
Myocardial infarction	7840 (5.8%)
Cardiomyopathies	6527 (4.9%)
Pulmonary embolism	5992 (4.5%)

Figure 1. Time dependent Brier Score and Time dependent ROC-AUC Score of models over evaluated time.

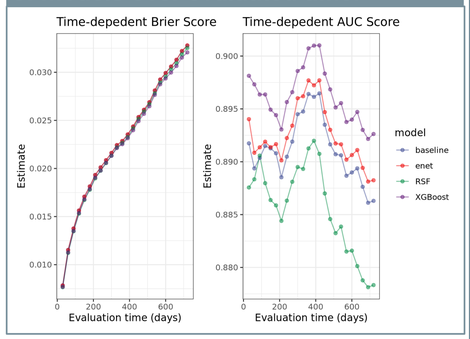


Figure 2. Time-specific calibration of recalibrated XGBoost Cox model for survival probability. Test set are divided into 20 groups by risk levels.

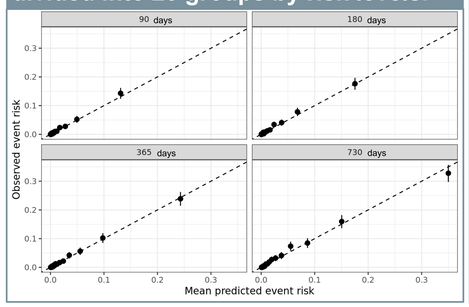


Figure 3. SHAP beeswarm summary plot for the final XGBoost model.

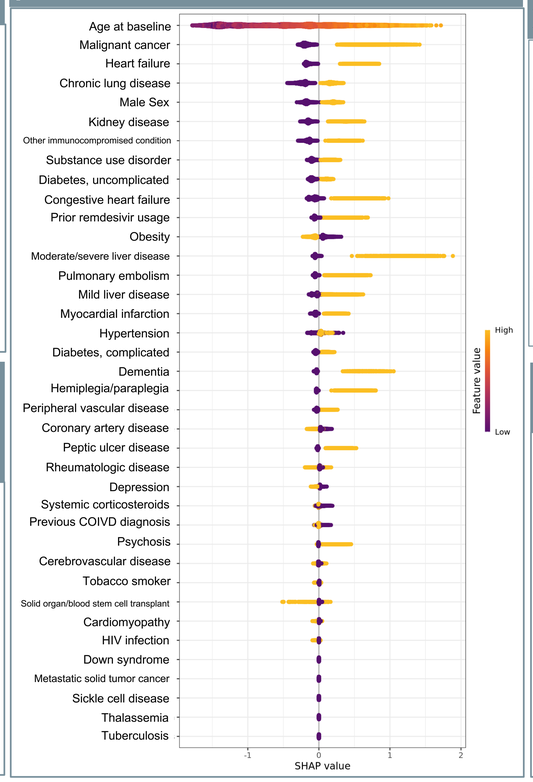


Figure 4. SHAP dependence plot for selected predictors in the final XGBoost model

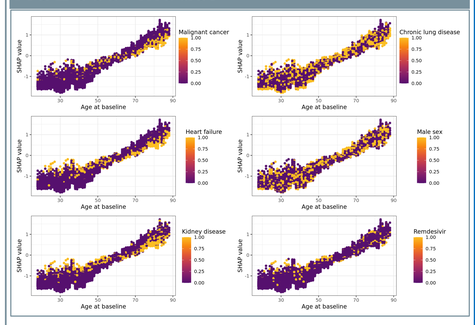
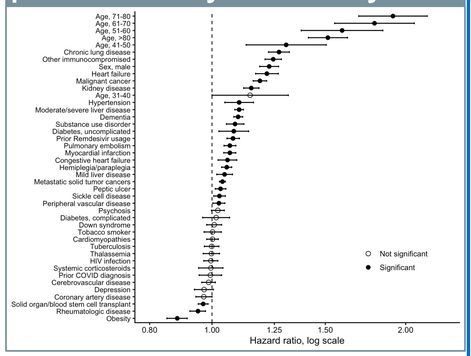


Figure 5. Baseline Cox model predictors of 2-year mortality



Clinical Implication

- **2-year mortality risk after PASC can be stratified** using routinely available EHR-based clinical and comorbidity data.
- **Risk prediction at diagnosis may support clinical decision-making**, including follow-up intensity, chronic disease optimization, and risk-informed care planning.
- **Younger and older adults with PASC may have distinct risk profiles**, and future work should evaluate whether different comorbidity patterns drive mortality across age groups
- **The model is intended to support, not replace, clinical judgment** and requires further external validation before clinical implementation.

Next Steps

- **Conduct age-stratified analyses** of younger versus older adults with PASC.
- **Incorporate additional predictors,** including biomarkers, longitudinal clinical data, and social determinants of health.
- **Validate model performance** in external datasets and prospective PASC cohorts.
- **Evaluate whether risk-stratified care** improves clinical follow-up and chronic disease management.

Acknowledgements: The research reported in this poster was supported by the AIM-AHEAD Coordinating Center at the University of North Texas Health Science Center at Fort Worth. This research was, in part, funded by the National Institutes of Health (NIH) Agreement No. 1OT2OD032581. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the NIH. The analyses described in this poster were conducted with data or tools accessed through the NCATS N3C Data Enclave https://n3c.ncats.nih.gov/clinical-cohort and N3C Attribution & Publication Policy v 1.2-2020-08-25b supported by NCATS Contract No. 75N95023D00001. This research was possible because of the patients whose information is included within the data and the organizations (https://ncats.nih.gov/n3c/resources/data-contribution/data-transfer-agreement-signatories) and scientists who have contributed to the ongoing development of this community resource https://doi.org/10.1093/jamia/ocaa196.