Prediction of extended length of stay with the PHTS

# METHODS

## PHTS

## Internal validation

## Missing data

The count and percent of missing values was tabulated overall and by length of stay status. Missing values were imputed after data were split into training and testing sets during each replicate of cross-validation. All information from the testing data was withheld during this process. Specifically, missing values were imputed in both the training and testing data using an imputation model that only leveraged the training data. Outcome values were not used to impute missing values, as this would imply knowing the outcome value before predicting it. A single imputed dataset was formed by applying a nearest neighbor imputation model to each predictor variable with missing values, separately. For each patient with a missing value for the current predictor variable, nearest neighbor imputation identifies patients in the training data who are most similar to the current patient and have an observed value for the current predictor variable. An imputed value is created by aggregating the observed values for these most similar patients. In the current study, 10 nearest neighbors were identified using Gower’s distance and imputations were formed using the median for continuous and mode for categorical variables.(1)

## Statistical analysis

All analyses were conducted using SAS version 9.4 or R version 4.0.3.

*Missing values*

The number and proportion of missing values for each candidate predictor were tabulated. Missing patterns

# RESULTS

1. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2020. Available at: <https://www.R-project.org/>.