Generalizability of prediction models

- How well does a prediction model perform in a target population that differs from the population originally used for model development and/or evaluation.

Datasources

- Framingham study data
- Data from NHANES

Objectives

1. Evaluate performance of a cardiovascular risk prediction model in a target population underlying NHANES (the focus should not be on the model building or getting the optimal model).
2. Conduct the same analysis when the target population is simulated.

Transportability analysis

- Useful references for i) Brier score (MSE for binary outcome) and ii) AUC

JOURNAL ARTICLE

Transporting a Prediction Model for Use in a New Target Population Get access >

Jon A Steingrimsson ✉, Constantine Gatsonis, Bing Li, Issa J Dahabreh

American Journal of Epidemiology, Volume 192, Issue 2, February 2023, Pages 296–304,
https://doi.org/10.1093/aje/kwac128

Published: 22 July 2022 **Article history** ▾

BIOMETRIC METHODOLOGY

Estimating the area under the ROC curve when transporting a prediction model to a target population

Bing Li ✉, Constantine Gatsonis, Issa J. Dahabreh, Jon A. Steingrimsson

Process

- Create a combined dataset from Framingham study data and NHANES data (with the outcome only being available in the Framingham study).
  - Find the variables that are common in both studies
  - Find the subset of the NHANES data that meets the eligibility of the Framingham study (to the extent possible).
- Either use an already existing model or build and evaluate a model (test and training).
- Estimate how that model performs in the NHANES target population .

Formula for transportability analysis

Notation:

- Y outcome
- X covariates
- S population indicator (S=1 if in Framingham study and S=0 if in NHANES)
- g(X) a prediction model for Pr[Y=1|X, D=0].
- D indicator if in test set (if you are using a split into train and test set)
- Estimator for Brier risk in
- target populati

$$\widehat{\psi}_{\widehat{\beta}} = \frac{\sum_{i=1}^{n} I(S_i = 1, D_{\text{test},i} = 1)\widehat{o}(X_i)\left(Y_i - g_{\widehat{\beta}}(X_i)\right)^2}{\sum_{i=1}^{n} I(S_i = 0, D_{\text{test},i} = 1)},$$

where $\widehat{o}(X)$ is an estimator for the inverse-odds weights in the test set, $\frac{\Pr[S=0|X, D_{\text{test}}=1]}{\Pr[S=1|X, D_{\text{test}}=1]}$.

Simulations

- Now assume that individual level data is not available from the target population and only summary statistics are available. Example from a different dataset below.

Table S2: Summary of the variables used in the transportability analysis of lung cancer prediction models in Section 5 of the main paper. Continuous variables are summarized by their weighted mean (weighted standard deviation) and categorical variables are summarized by weighted percentage in each category.

Variable	NLST	NHANES
Age	61.4 (5.0)	62.6(5.4)
BMI	27.9 (5.1)	29.1 (6.5)
Race or ethnic group		
Black	4.4%	8.2%
White	89.9%	81.6%
Hispanic	1.8%	6.2%
Other	3.8%	4.0%
Education level		
less than high school	6.3%	36.2%
high school graduate	38.3%	30.3%
AA degree/some college	23.7%	21.6%
college graduate	31.7%	11.9%
Personal history of cancer	4.1%	22.3%
Smoking status		
Current	48.4%	62.4%
Former	51.6%	37.6%
Smoking intensity [1]	28.5 (11.5)	27.2(12.1)
Duration of smoking (year)	39.8 (7.3)	42.1 (6.7)
Smoking quit time (year) [2]	7.3 (4.8)	6.4 (3.7)

[1] Smoking intensity (the average number of cigarettes smoked per day) has a nonlinear association with lung cancer and was transformed by dividing by 10, exponentiating by the power -1 in the prediction model for lung cancer.
[2] Smoking quit time in former smokers (years).

Simulations

- Use such summary level data to simulate individual level data from the target population and conduct the transportability analysis using this simulated dataset. This will give one estimator of model performance in the target population.
- Across simulations, compare the simulated transportability analysis to the non-simulated transportability analysis.