Integrating household consumer data utilizing a multi-language pipeline may improve prediction models for coronary heart disease

Consumer data and risk stratification for coronary heart disease

Iben M. Ricket MPH[1]; Inas Khayal, PhD[2]; Jeremiah R. Brown, PhD, MS[2,3]

[1] Quantitative Biomedical Sciences, School of Graduate and Advanced Studies, Dartmouth College, Lebanon, NH
[2] The Dartmouth Institute for Health Policy and Clinical Practice, Lebanon, NH
[3] Department of Epidemiology , Dartmouth Geisel School of Medicine, Lebanon, NH

Email: Iben.Ricket.gr@dartmouth.edu, | **GitHub**: https://github.com/iben-mickrick | **Twitter**: @imickrick

INTRODUCTION

Many risk models utilize co-morbidities to predict adverse health outcomes.

Co-morbidities alone fail to include information on factors that create or enable risk for adverse health outcomes themselves.

Consumer patterns may reflect or be influenced by social determinants of disease, which may offer unique information about risk for adverse health outcomes beyond information offered by presence of co-morbidities alone.

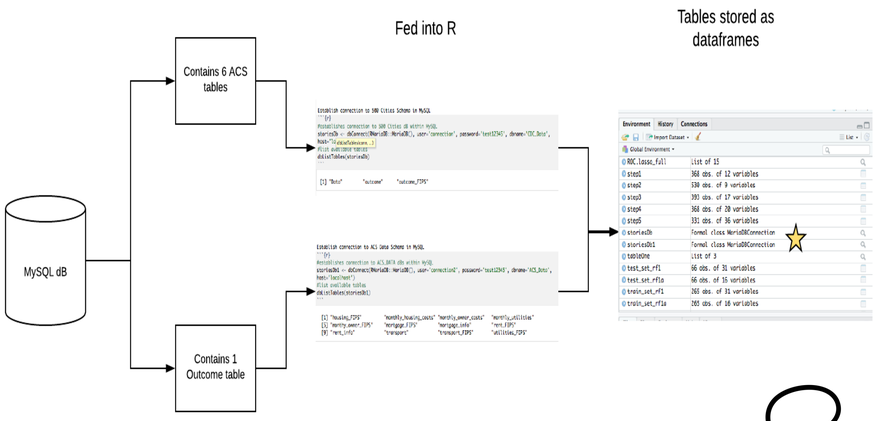
METHODS

DATA SOURCES:

2015 American Community Survey

2015 Centers for Disease Control & Prevention 500 Cities

DATA PIPELINE:



ANALYSIS:

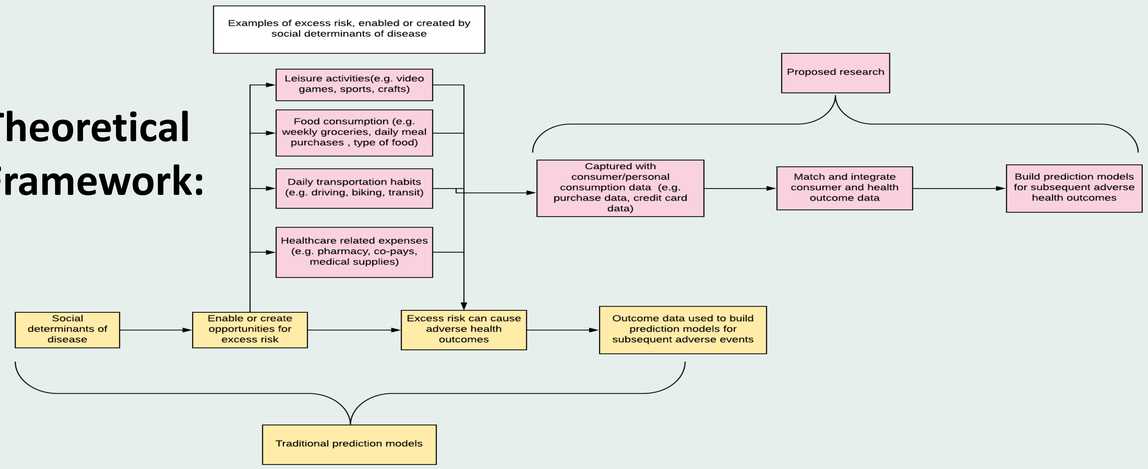
Logistic regression with L1 regularization

RESULTS

Model	AUROC	Std. error	95% CI
Base	0.893	0.039	0.817, 0.968
Full	0.939	0.029	0.881, 0.998

Consumer variables selected by full model:

Theoretical Framework:



Scan code to access GitHub repository with abstract, code, and references

Disclaimer: This is a new style of poster #betterposter

Data manipulations: Missing data were imputted with mean values, unique FIPS codes were collapsed into geographic four regions, continuous skewed variables were log-transformed, and the following variables were removed: extra geographic (e.g. geo location, city name), gender specific outcomes (e.g. pap smear, mammography), non-disease related prevalence (e.g. preventative screens), non-consumer related information, and crude estimates.

Outcome candidate variables:

- Blood pressure, blood pressure medication, smoking, diabetes, high cholesterol, chronic kidney disease, poor mental health, poor physical health, obesity
- Health insurance, arthritis, cancer, asthma*

Consumer candidate variables:

- Adjusted median household income, number of vehicles available at the household, housing costs
- Median housing costs, second mortgage prevalence, equity loan, public transit for work, use of vehicle for work, households without cars, geographic region, using the following as fuel: electricity, oil, wood, solar, or "other"*

*Collapsed into binary variables based on median

Variables selected from full LASSO model

Variable	Coefficient
Adjusted median income	-1.17
Arthritis prevalence low	-1.01
Blood pressure prevalence	0.009
Cancer prevalence low	-0.288
Asthma prevalence low	1.67
COPD prevalence	2.40
Diabetes prevalence	0.540
Poor physical health prevalence	0.338
Health insurance coverage low	-0.666
Households with cars	0.140

No statistically significant difference between models (p=0.338)

Additional model performance metrics

Metric	Base	Full
Accuracy	0.894	0.939
Sensitivity	0.844	0.938
Specificity	0.941	0.941
PPV	0.931	0.938
NPV	0.865	0.941
Precision	0.931	0.931
Recall	0.843	0.844
F1	0.885	0.938
Brier Score	0.058	0.043