STATISTICAL METHODS FOR RISK PREDICTION IN PANDEMIC DATA

Keshav Gandhi and Sophia Torrellas, SPIN Internship Program

National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign

INTRODUCTION

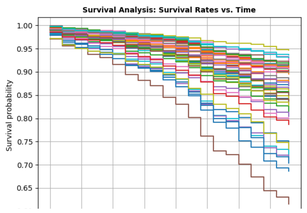
Goals
- Predict a person's probability of testing positive for COVID-19 at a given time
- Determine which features act as significant risk or protective factors for COVID-19
- Optimize modeling by utilizing simulated pandemic data

Dataset
- 1,000 simulated UIUC students
- Analyzed about 150 different covariates
- Features are survey question responses
- Survival analysis of right-censored COVID-19 data, cutoff date at about 100 days

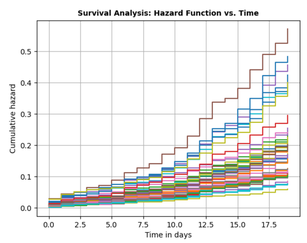
Importance
- Compare with "real" dataset
- Predict who has COVID-19 and allocate limited resources based on risk
- Reduce the spread of COVID-19

METHODS

Random Survival Forest
- Some features are immutable characteristics, but most represent controllable behaviors
- Survival analysis assumes all initial behaviors remain constant through the entire study



Survival curves represent survival probability of not getting COVID-19 for 50 randomly selected students

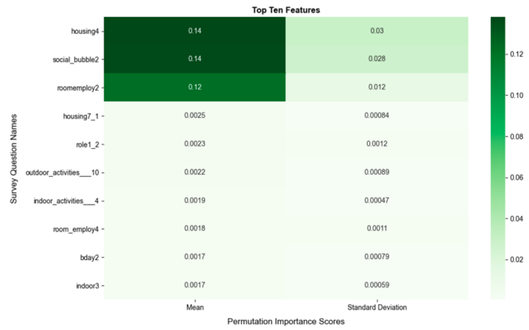


The cumulative hazard function represents the risk of getting a positive COVID-19 test

METHODS CONT.

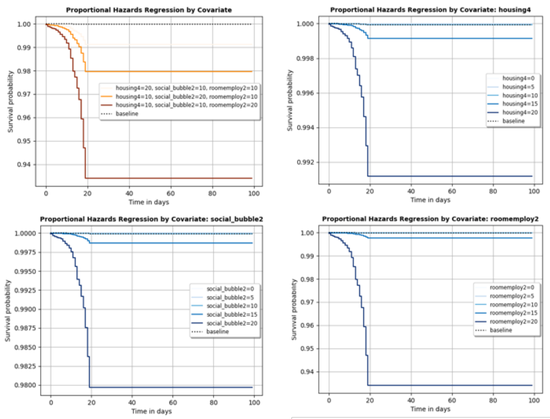
Permutation Feature Importance
- Decrease in model's score when one feature is randomly shuffled and others are controlled
- Determines which covariates had the greatest effect on random survival forest

Important covariates: housing4 (number of roommates), social_bubble2 (size of one's social bubble), roomemploy2 (having a high-risk job)



Proportional Hazards
- Semi-parametric test with more assumptions
- Provided more robust summary statistics
- Observed non-linear interaction effects

Separated effects of three most important covariates in random survival forest model via controlled proportional hazards plots

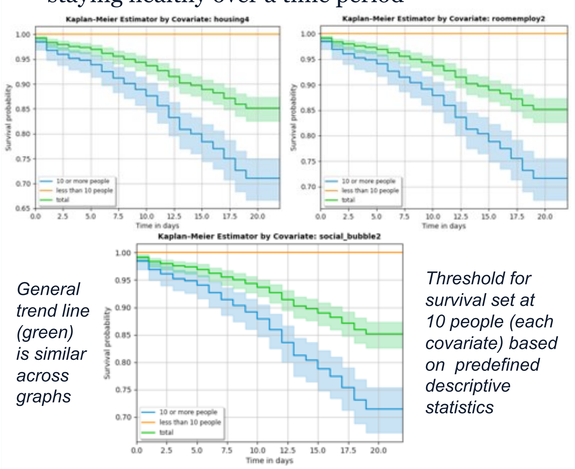


Expected value is the population times the survival probability at a given time, which depends on the percentage tested in real life

Combined Methods
- Permutation feature importance ranks
- Proportional hazards summary statistics
- Served as inclusion criteria for Kaplan–Meier estimator for subsequent modeling using binarized features

APPLICATIONS

Kaplan–Meier Estimator
- Curves show the decreasing probability of staying healthy over a time period



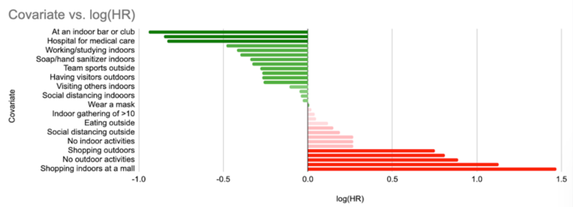
General trend line (green) is similar across graphs

Threshold for survival set at 10 people (each covariate) based on predefined descriptive statistics

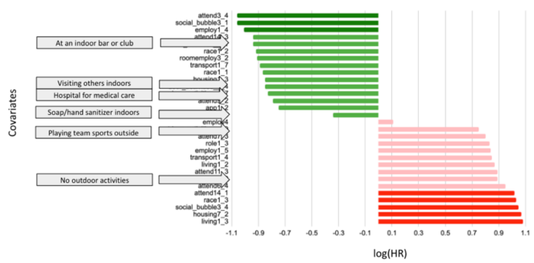
<10 people, no one predicted sick, but ≥ 10 people, people get sick

Hazard Ratios of Everyday Activities
- Calculated overall risk in engaging in a variety of university student's everyday activities

log(hazard ratio) = 0 means that covariate has no significant effect on COVID-related outcomes



Statistically Significant Covariates
- Graphed the hazard ratios of statistically significant variables (p-value < 0.05)
- Most covariates representing everyday activities do not confer statistically significant risk



The data being simulated means that these patterns do not always translate to real life, which may account for such discrepancies

RESULTS AND DISCUSSION

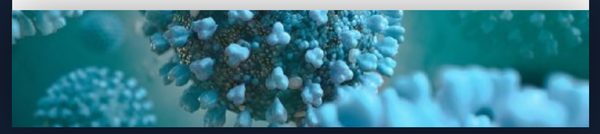
Challenges
- Interval data effects overestimated vs. binary
- Sociological features significant in simulated data only
- Trade-off of inferring direction of risk vs. fewer underlying assumptions (needed to integrate results from both parametric and non-parametric models)

Results
- Over 40 covariates significantly affect risk of a positive COVID-19 test, with most of them being behavioral
- *Risk factors*: having a higher number of roommates, a larger social circle, a high-risk job, etc.
- *Protective factors*: having a lower number of roommates, a smaller social circle, washing hands with soap or using hand sanitizer several times a day, etc.

FUTURE WORK

Simulated vs. Real Data
- Calculate concordance indices, precision/recall

Additional Techniques
- Exponential and Weibull distributions



ACKNOWLEDGEMENTS

Mentors: Dr. Weihao Ge, Dr. Liudmila Mainzer
Students Pushing Innovation: Olena Kindratenko
Dataset Creator: Wayne Wan

ILLINOIS
NCSA | National Center for Supercomputing Applications

ILLINOIS