EXECUTIVE SUMMARY

Over-capacity hospitals with strained resources and beds in hallways became one of the most alarming markers of the COVID-19 pandemic as the disease swept the world in 2020. Anticipating these hospital surges continues to be key in managing the crisis for communities. Our analysis used Google trends search data to predict hospital surge at the metro area level.

DATA

Our work relied on three data sources: the department of Health and Human Services weekly hospital-level reports[1], internet search trends retrieved from Google Trends, and county-level U.S. Census Bureau data. To obtain the Google Trends dataset, we pulled search data for ten keywords that are common COVID-19 symptoms according to the CDC[2]. Specifically, we obtained the weekly search frequency for "breath", "throat", "smell", "taste", "fever", "cough", "vomit", "fatigue", "chills", and "diarrhea", aggregated by metro area (as defined by Google[3]). Then, we merged the three data sources together using publicly available crosswalk files[4, 5]. Among the 210 metropolitan cities in Google Trends, we successfully mapped 108 cities to the hospital-level data. We used these data to build a variety of predictive models to predict hospital surges.

ANALYSIS

Our data spanned a 6-month time frame, during which 27% of hospital observations qualified as experiencing a "surge". To investigate if keywords can predict hospital surges, we constructed predictive models using all cities with complete data (n=108).

K-nearest neighbors (kNN), random forest, support vector machine, and decision tree models were our most successful predictive models. We evaluated the results based on AUC, precision, recall and F1-score, the latter of which combines precision and recall. Four of our models achieved accuracy above 80 percent, and our best performing model, kNN, had 83 percent accuracy as measured by the F1-score.

Recognizing that predictive models that perform well overall may underperform for specific metro areas, we used county-level census data to examine performance by metro demographics. We grouped metro areas in our test dataset by whether they had at least one correctly-predicted surge and compared the two groups using t-tests. The size of our dataset limits our ability to detect these differences and we therefore did not find any significant demographic differences between the groups. We hoped to conduct a cluster analysis to better understand how demographics varied across well predicted cities but were unable to do so.

---

[1] https://healthdata.gov/Hospital/COVID-19-Reported-Patient-Impact-and-Hospital-Capa/anag-cw7u
[2] https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html
[3] https://support.google.com/trends/answer/4355212
[4] https://sites.google.com/view/jacob-schneider/home?authuser=0
[5] https://www.unitedstateszipcodes.org/zip-code-database/

CONCLUSION

Anticipating hospital surges is key to managing the ongoing COVID-19 pandemic. To our knowledge, others have not yet leveraged Google Trends data to predict these surges, although case and death prediction at state levels has been done. Using ten search terms, we were able to build a variety of predictive models that offer moderate insight into forthcoming hospital capacity issues. While we note several limitations of using Google Trends to predict surges, namely, limited data coverage, we demonstrate that Trends may offer a novel, timely data source that could be leveraged by governmental bodies, health departments, hospital systems, and others to get an earlier signal of forthcoming hospital surges.

PRESENTATION

https://www.youtube.com/watch?v=QTa4BAw8ETw