Automated Knowledge Graph Construction for Healthcare Domain

Markian Jaworsky

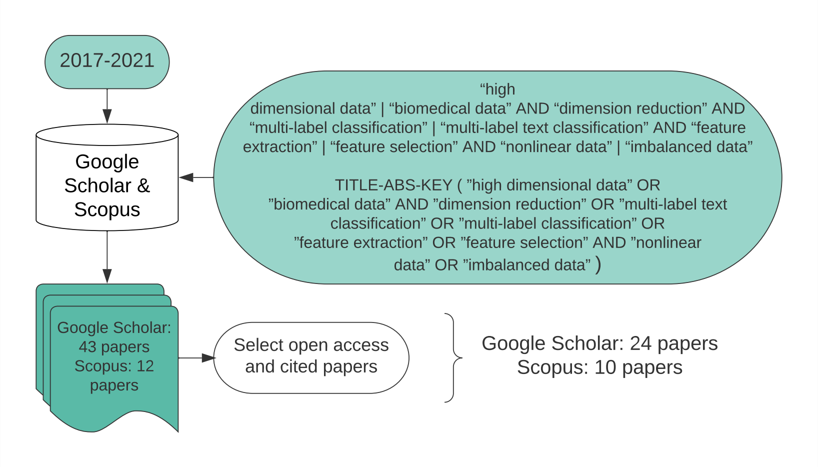
Co-Authors:
(1) Xiaohui Tao, Jianming Yong, Ji Zhang
(2) Lei Pan, Shiva Pokhrel

1: School of Mathematics, Physics, and Computing, University of Southern Queensland, Australia
2: School of Information Technology, Deakin University, Australia

Introduction

- A range of risk factors can be used as predictor variables in the likelihood of developing chronic illness.
- With awareness, patients can adapt their lifestyle in order to improve their chances of longer term survival. [1]
- Risk factors can be categorized as being lifestyle, environmental or biomedical and can change over time. [2]

Literature Review - Scope



Literature Review - Data Challenges

- High Dimensional Datasets - 21 (38%) review citations
- Imbalanced Class Datasets - 13 (24%) review citations
- Missing and Erroneous Data - 12 (22%) review citations
- Multiple and Nonlinear Datasets - 9 (16%) review citations

Research Concept - Knowledge Graphs

- Knowledge graphs promise alternate use to synthetic resampling and upsampling data manipulation methods in the prediction of rare diseases from datasets with highly imbalanced classes. [3]

Research Concept - Knowledge Graphs

- Our review of knowledge graph studies indicate that most research is dependent on knowledge maintenance. [4]
- By identifying a complete source of knowledge we can automate the construction of a complete knowledge graph.
- Using a knowledge-based method to select features gives a predictive model assurance against detecting spurious correlations.

Research Design - Dataset

- The United States CDC make available an anonymized annual Behavioral Risk Factor Surveillance System (BRFSS) survey data, which is free to the public domain and may be copied and distributed without permission.

From: Garvin, William S. (CDC/DDNID/NCCDPHP/DPH) <wsg1@cdc.gov>
Sent: 03 September 2021 23:02
To: Markian Jaworsky <Markian.Jaworsky@usq.edu.au>
Cc: Barrett, Drue H. (CDC/DDPHSS/OS/OSI) <dhb1@cdc.gov>
Subject: RE: Question on statement of CDC BRFSS Annual Questionnaire Data Ethics Approval

Dear Markian,

The Behavioral Risk Factor Surveillance System (BRFSS) is approved by US Office of Management and Budget (OMB) to collect data from the US general population under OMB Control number 0920-1061.

The CDC Human Research Protection Office has determined that this research activity (BRFSS data collection) remains exempt under 45 CFR 46.101(b)(2).

The BRFSS is a state based survey conducted in partnership with the participating state-health departments. A common core questionnaire and standardized optional modules are approved by the states and CDC programs each year. The state health departments implement the survey and oversee the ongoing data collection for their state, whether through a contracted data collector or in-house data collection. The states and data collectors have institutional review boards which review state-specific questionnaire content and determine what is applicable for inclusion in the BRFSS for a given state. The questionnaire and basic BRFSS data collection protocol are covered in the BRFSS Overview documentation released with the public use data set each year on the BRFSS website.

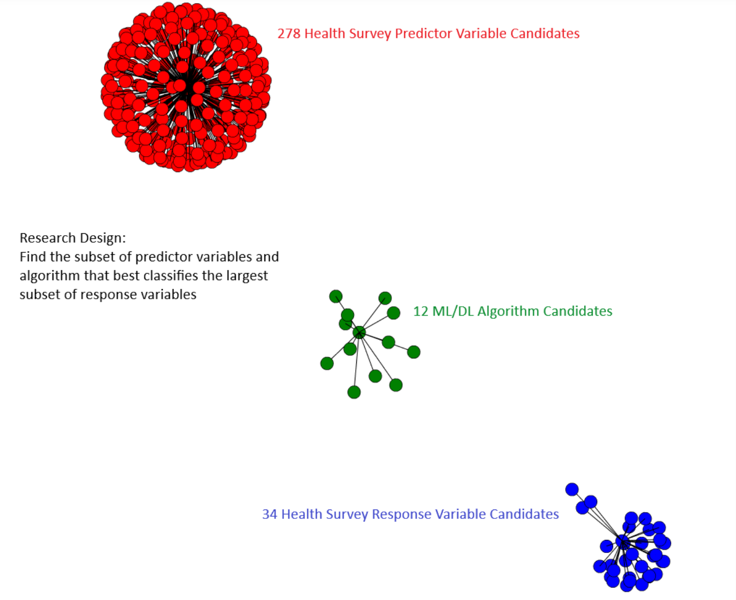
Hopefully this provides the information needed for explanation of the approval process of the BRFSS.

If you have additional questions or concerns please let us know.

Thank you,

Bill Garvin
Survey Operations
Population Health Surveillance Branch
Division of Population Health
National Center for Chronic Disease Prevention and Health Promotion
Centers for Disease Control and Prevention
(770) 488-4621

Research Design



Knowledge Graph - Automation

- R programs on GitHub convert WHO ICD codes and health survey question text vectors into a Knowledge Graph CSV file.
- WHO ICD Code chapters can be pruned to focus on specific human organ systems.

https://github.com/mjaworsky/KnowledgeGraph

Knowledge Graph - Nodes

- The knowledge graph nodes represent a feature matched with another feature that was significantly correlated by Wilcox Rank.

Knowledge Graph - Edges

- The knowledge graph edges are scored by determining the number of text words of a feature was matched with another feature significantly correlated by Wilcox Rank.

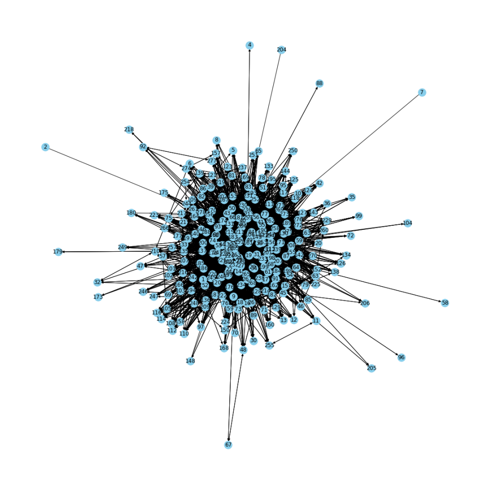
Research Evaluation - Metrics

- Data Set 1 - Baseline dataset with all features containing no less than 50% of values valid (not missing).
- Data Set 2 - Top 25% of baseline features significantly Wilcox Rank Correlated.
- Data Set 3 - Top 10% of baseline features significantly Wilcox Rank Correlated.
- Macro Average F1-Score - 35 (65%) review citations.

Knowledge Graph-Based - Feature Selection Results

Survey	Classifier	Data Set	Macro Avg F1
BRFSS 2020	Rusboost	DataSet3	0.56
BRFSS 2019	Rusboost	DataSet3	0.498
BRFSS 2018	Rusboost	DataSet3	0.66
BRFSS 2017	Rusboost	DataSet3	0.7

- We can automate the construction of a knowledge graph by identifying significantly related health survey questions using text frequency vs chapter rankings of the WHO ICD.



Conclusion 2

- In this study, we demonstrate that constructing a knowledge graph can improve feature selection by directly seeking the relationships between features and then prioritizing the aggregate of the features with the most relationship edge values.

USQ Human Research Ethics Approval

- USQ HREC ID: H21REA222
- Approval date: 15/10/2021
- Review date: 13/10/2022
- Expiry date: 15/10/2024
- USQ HREC status: Approved

References

[1] Hyuna Sung et al. "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries". In: *CA: a cancer journal for clinicians* 71.3 (2021), pp. 209–249.

[2] Australian Institute of Health. *LungScreen Australia monitoring report 2011*. 64. AIHW, 2011.

[3] Xuedong Li et al. "Improving rare disease classification using imperfect knowledge graph". In: *BMC medical informatics and decision making* 19.5 (2019), pp. 1–10. DOI: 10.1186/s12911-019-0938-1. URL: https://doi.org/10.1186/s12911-019-0938-1.

[4] Xiangxiang Zeng et al. "Toward better drug discovery with knowledge graph". In: *Current opinion in structural biology* 72 (2022), pp. 114–126.

Thanks

Questions?

Markian Jaworsky
Markian.Jaworsky@usq.edu.au