Mutlu Yuksel (Dalhousie University)

Yigit Aydede (Saint Mary's University) - Presenter

Andrea Giusto (Dalhousie University)

Research Portal on Machine Learning for Social and Health Policies

Halifax, Canada

What can be predicted from a national health survey? Is cancer one of them?

- **BigSurv20**
- **https://www.bigsurv20.org**
- **Big Data Meets Survey Science**
- November 6 – December 4

Motivation

Even after a quarter century of extensive research, researchers are still trying to determine whether cancer is preventable.

Cancer is caused by both internal factors (such as inherited mutations, hormones, and immune conditions) and environmental factors (such as tobacco, diet, radiation, and infectious organisms).

Studies indicate that only 5–10% of all cancer cases are due to genetic defects and that the remaining 90–95% are due to environment and lifestyle.

The main objective of this study is to see whether the life-style and environmental factors can be identified as predictors of cancer with high-dimensional data.

Data- CCHS

Canadian Community Health Survey (CCHS) is the largest national health survey with more than 130 thousand observations in annual files.

It has bout 1,500 variables (in the major confidential files) that contain information ranging from the amount of weekly carrot consumption to the weekly time in minutes that the person spends in Olympic weightlifting workouts.

Even though CCHS is the most extensive health survey in Canada it is almost useless for predictions as well as causal analyses due to its cross-sectional structure.

For example, a naïve work can discover that eating healthy food strongly predicts if the person has a chronic disease, such as cancer.

Objectives

1

To see how tangible improvements in the cross-sectional nature of CCHS increases the predictive power of the data.

2

To see if this practice can also lead to discovering “casual predictors” of a disease by advance machine learning models.

Link to Discharge Abstract Database (DAD)

- We built a panel dataset, by linking DAD and CCHS
- The linked data traces the people in 2001 CCHS for the following 10 years between 2002 and 2011.
- We dropped people who were cancer patients in 2001 or had gone through a cancer treatment prior to 2001.
- Among more than 60 thousand people who have no cancer in 2001 and never had a cancer before 2001, we identified around 6 thousand people who had developed a cancer in the following 10 years.

Predicting cancer…

With this data set, we have an opportunity to use every feature in the survey without having a concern about possible reverse causality problems or model-leaking issues.

With a binary outcome, we first tried to see whether cancer, as a common disease, can be predicted with non-medical data and, if it can, what predictors can be identified (for those between 55-75).

To give an idea about the dimension of the data, with only first-level interactions, our sample can be extended to more than 100 thousands features.

Predicting cancer…

Since more than 80% of people who had cancer in the ten years following 2001 are between 55 and 75, we built predictive models only for that age group.

The DAD files report 11 different major cancer types in medical records. In our first attempt, we use a binary outcome that cover all types of cancer patients.

Although “cancer” is the name given to a collection of related diseases in which some of the body’s cells begin to divide without stopping and spread into surrounding tissues, **cancer is not one disease but more than 100**.

Predicting cancer…

With a binary outcome, we first tried to see whether cancer, as a common disease, can be predicted with non-medical data and, if it can, what predictors can be identified.

With our multi-stage algorithms, we could not exceed 62% of prediction accuracy measured by AUC (after removing the effect of age).

In addition to shrinkage methods, we mostly used nonparametric methods. We also used duration/hazard models as we have time in months.

This shows that, **when the age is controlled for**, there are very few predictors common for all cancer types.

“Causal” predictors …

In the second step, we used two different outcomes: all cancer patients with and without **lung cancer**.

The results show that first-hand and second-hand smoking are **the main predictors** for respiratory cancer patients.

But they are not identified as predictors when we exclude respiratory cancer patients from the sample. We call this practice as **using prediction for causal interpretations**.

These initial results imply that smoking may not be a common predictor for cancer when respiratory cancer types are removed from the data.

Remarks

- This study converts a public survey into a high-dimensional panel dataset that contains more than 60 thousand people and 6 thousand cancer patients observed over 10 years between 2001 and 2011.
- The scope of the information even in the base survey with more than 1500 features is unprecedented.
- At this point, our initial results provide **evidence about so-called lifestyle factors on cancer for the first time using a high-dimensional data** with unpresented number of cancer patients.

Remarks

- The results are robust and tested with more than 15 different and advance machine and deep learning algorithms.
- Although the scope of information in CCHS about the environmental and lifestyle risk factors is enormous, **we cannot verify many major risk factors identified in the medical literature**.
- We are developing a hub at ARDC (Atlantic Research Data Center) where the scripts for data merging will be available.
- This is an ongoing project. The more results and scripts will be posted on my blog and MLPortal

Thank you!