Study of Incontinence Patients Data and Pelvic Cancer Influence on Medical Therapy Recommendations

Capstone Project Summary

Agne Vaivodaite

MSc in Business Analytics

Central European University

In the age of technology, medical decision making should be more automated. By doing so, doctors would free up time for an actual human to human interaction which may get lost when the burden of decision making is on them. It should also limit false-negative diagnoses which are crucial for early diagnosis and survival rate.

This study is based on an incontinence related questionnaire for patients that have already been to a urologist. It has around 25 incontinence related questions, including a column marking if a patient was diagnosed with pelvic cancer. The main goals of this study are the following:

- Analyse incontinence related dataset (over 9000 entries) and create an analytic model based on health risk factors and the occurrence of cancer
- Explore if the cancer diagnosis leads to different medical treatment
- Establish a new dataset structure based on findings of the analysis
- Discuss potential new questions to be included in the new dataset

Completing these tasks will help to kick off the client's primary goal of creating a decision-making platform that will help doctors to offer treatment and a correct referral to further examination for their patients.

This study is broken down into several tasks:

- Data cleaning and exploration
- Analytic model creation
- Dissemination of results
- Comparison of treatments prescribed to patients with and without diagnosed cancer
- Comparison of female and male analytic models
- Questionnaire improvement suggestions

- Online decision-making platform examples

Methods used for the analysis were the following:

- Zero imputation: This method was used to impute around 2000 missing values
- Pearson's correlation: This measure helps to identify whether a relationship exists between two variables, the direction and strength of the correlation they have
- Probit: Cancer diagnosis variable is binary – it is either true or false. Binary variables can be studied by models targeted at latent variables, such as probit
- LASSO: this model is used for feature selection and regularization of data models. LASSO feature selection works by shrinking regression coefficients towards zero aiming to reduce if they are not necessary
- ROC, precision-recall curves, AUC: The visual inspection of the two curves allows judging if the model is doing better than random guessing (ROC close to 45 degrees line) and if classified outcomes have a reasonable amount of false-negatives (PR). ROC curve can be summarized with a single number – AUC which is considered to be good between 0.8-0.9 range

The analysis started with data cleaning during which the questionnaire was transformed, and missing data were imputed. All answers were converted to binary and categorical values for easier analysis. The dataset coverage was focused on the female side while the number of males was smaller – just around 20% of all data. The positive cancer cases were also different for men and female – females had almost twice as much diagnosed cancer cases than men (around 8.3%).

Overall, the majority of diagnosed cancer cases in the data belong to the 60+ age group. It is important to have in mind that the majority of the entries in the dataset belong to 60+ group which, naturally, has a higher number of cancer cases, compared 80+ group, where 20 times less data exist. Positive cancer diagnoses for women increase with age, and there is a material shift past 70+. Men, on the other hand, are mostly diagnosed in their 60s, while they are less likely to be diagnosed at a younger age. The decrease of positive diagnosis can be seen in men at around age 80 which could be caused by a lack of observations

Correlation analysis helped to uncover patterns in the data. Certain variables in the dataset were more related to cancer than others. Mobility constraints, UTI and questions related to incontinence severity proved to be the most correlated with a pelvic cancer diagnosis. Questions related to prevention, obesity, and diabetes were least related to pelvic cancer, according to our data. Correlation results showed several questions to be highly co-moving. These, therefore, might be worth reformulating or deleting altogether in future surveys.

A combination of probit and LASSO models was used for analytic modelling. Firstly, a variety of variable groups was created, and interactions were taken into account. Six probit and one LASSO models were created. LASSO model had the lowest holdout RMSE. Non-zero variables from LASSO results were selected and applied to the final probit model. The model was evaluated using ROC curve, AUC, as well as the precision-recall curve. The final model had an AUC of 0.8 which can be considered good. The suggested threshold after which all cases could be considered as positive is around 0.12.

The overall sample model results revealed that fecal incontinence, doctor’s evaluation of incontinence complaints, asymptomatic treatment recommendations, mobility issues and recurring UTI have a positive effect on cancer diagnosis. According to the findings in this study, diabetes, neurological issues, obesity and experienced stroke have a negative effect on cancer diagnosis. These findings proved to be more or less as expected based on initial correlation analysis. When comparing genders, female patients have a slightly higher risk of being diagnosed with cancer if they have recurring UTI, mobility issues and higher doctor’s concerns, similar as for the overall sample. Male patients show different patterns: doctor’s opinion about incontinence complaints and need for asymptomatic treatment as well as incontinence during sneezing or coughing have a positive effect on positive cancer diagnosis in our data. Contrary to females and overall sample, men have a lower risk of being diagnosed with cancer when they experience recurring UTI or have mobility issues.

A similar approach was used to study the difference in doctor-recommended treatments for patients with and without cancer diagnosis. Each treatment was predicted using the probit model. It was possible to compare how different or similar treatments were for cancer and no diagnosed cancer patients using average marginal effects. People without diagnoses tend to receive a wider range of treatments while patients with cancer, who most probably already receive treatments prescribed by different doctors, mostly receive treatments in case if fecal incontinence occurs or if they suffer from neurological problems.

Finally, future survey changes and other suggestions were made based on discoveries from the data exploration and analytic model. An example of decision making/medical prediction platform was also introduced. Some questions in the client's questionnaire need reformatting or being replaced altogether. Some questions proved to be too similar or too long, thus introducing the risk of answer fatigue in patients. Questions related to diabetes or questions with vague given answers are not specific enough, which leads to the loss of important information. Such questions should be specified: diabetes type 1 and type 2 have different roles in the possibility of pelvic cancer; answers similar to "rarely" and "occasionally" are not specific enough and should have more defined time frames.

The findings of this study mostly align with the reality of incontinence and cancer: men are less willing to answer uncomfortable incontinence related questions. Therefore doctor's input can tell more when predicting pelvic cancer; pelvic cancer in women is related to slightly different drivers than men. The suggested treatment for patients with diagnosed cancer slightly different compared to those with cancer.