Name of the Researcher: Rakesh Kumar Muraleedharan

Course Name: Master of Science – Data Analytics

College Name: CCT College, Westmoreland St, Dublin 2

Thesis Title:

Predict the screening colonoscopy numbers across Ireland in order to assist in the capacity planning of the colonoscopy units, by using an appropriate regression Machine Learning model.

Objectives:

1. Use descriptive statistics, to identify the variations in colonoscopies across genders and age groups, identify the variations in the colonoscopy numbers across different months of a year in order to better plan the colonoscopy capacity in different units, and further do a hypothesis testing to test if the colonoscopy counts for male population are less compared to females in order to assist the colonoscopy units better.
2. Identify the features that influence the colonoscopy counts, by undertaking appropriate co-relation studies between the features. Use visualisation charts to understand how the colonoscopy counts vary for each gender, identify if there are any seasonal factors affecting the colonoscopy numbers.
3. Taking the age range extension proposal into consideration and, in order to assist in predicting the increase in colonoscopy numbers across Ireland, make use of the historical colonoscopy records as the secondary data source, and the eligible population in the extended age group from the register as the primary data source, predict the future colonoscopies by applying a suitable regression machine learning model, the best model should be identified after applying hyper parameter tunings and cross validations.

SAMPLING STRATEGY:

It is proposed to use three different sampling strategies as part of the research, this includes both probabilistic (simple and stratified) sampling, this is primarily used while attaining the objective stated above. Non-probabilistic sampling (judgemental) is done to support the research and get inputs on the factors influencing the research.

This research project proposes to use two separate populations for achieving the above objectives. The first population are the males who have undergone colonoscopy across different age groups. The second population are the females who have undergone colonoscopy across different age groups. Three different samples are used to achieve the above objectives, and also get some insights in addition to the literature review.

In the first sample, each population(males) is divided into different age groups, statistical analysis is done on these age groups. Since the population selected is first grouped together for each age group, and a sample will be collected for further analysis, *stratified sampling* strategy is applied here. This is because the population is divided and grouped to different ages and then random selection is done. Since every record in the sample have undergone colonoscopy, simple sampling is applied on the selected sample. This is a probabilistic sampling, since every record in the sample have undergone colonoscopy, which means every unit has a chance. Also, in the sample the population is further grouped for each age group. The age group that has undergone colonoscopy as part of the colorectal is between 61-69 years. We will divide and group the data as between 61-63, 64-67 and 68-69 for the

sake of our analysis. Sample records (30%) across the years, are then selected using the simple random sampling approach. Learnings from the above sampling is understood and applied on the second population for Females. Descriptive statistics are then applied to this selected samples so that the objective of identifying the max, min, average colonoscopies for the population is identified.

The third population that is proposed to be used in the research is the number of eligible clients based on the census 2022 data. *Probabilistic* sampling since is done on this population, since all the clients are equally eligible for colonoscopy. We will club the census data with the actual colonoscopy numbers and then apply suitable machine learning models on the sample records, we will use 70% of the data to train the model and the remaining 30% to test the model. *Simple* random sampling strategy is proposed to be used here since any random records can be selected for the research. This is because in this sample random sampling is done on the population without any age grouping.

Further, in order to better understand the dependencies of the research topic a mixture of judgement and convenience sampling is applied. Depth interview will be done as detailed in the primary research section below. There are several points in the research that is beyond the literature review and needs input from the experts. The experts with regards to the topic are readily available and convenient hence the strategy applied is convenience strategy. The experts selected would represent the population and hence the judgemental sampling strategy.

Ethical considerations to be aware of are the probability sampling are done considering three populations like male, female, based on the literature reviews. The data is from the clinical database and the issue with regards to the bias are taken care.

PRIMARY RESEARCH METHODOLOGY

This research proposes to use depth interview as a primary research methodology. Number of patients to undergo colonoscopy is directly proportional to the number of patients with a positive FIT result which is in turn proportional to the eligible population. However, the literature reviews have suggested statistically men participate less in the Bowel screening programme, also there are impacts of deprived areas, age group, certain months having high/less colonoscopy counts. These answers should be answered by the health care experts in the field of cancer screening. This is the reason I choose depth interview as my primary research method. Data privacy and security is one of the biggest concerns to be taken care of. Ethical consideration is taken care by interviewing the data security officer within this is based on judgement and convenient sampling.