**Research Brief – DSI 2024**

***Date***: October 20th, 2024

**Team 3**: Marcio Sugar, Lisa Feeney, Serhii Bohoslavets, Hossein Baghdadi, Joao Tiago Foa Binsztajn

**Business Issue/ Challenge/ Opportunity:**

***What is the problem you are trying to solve with this project?***

Physicians are faced with the task of developing a prognosis based on the patient’s symptoms. Is there an opportunity to use Machine Learning to predict the prognosis based on the patient’s symptoms? This would enable the physician to make a faster prognosis and spend more time interacting with the patient to discuss treatment options. It would also reduce physician burnout due to time spent doing administrative work, increase physician satisfaction levels and reduce health care costs.

***Background:***

The dataset has 132 parameters (symptoms) on which 41 different types of diseases can be predicted.

***What do we already know?***

* “The utility of digital symptom assessment apps relies on coverage, accuracy and safety. While no digital tool outperformed GPs, some came close, and the nature of iterative improvements to software offers scalable improvements to care.[[1]](#footnote-1)”
* 5 Best Medical Symptom Checkers [[2]](#footnote-2)
  + Ada: Incredibly well designed and accurate symptom checker. Diagnostic accuracy of 77%.
  + Buoy: Buoy is an AI-powered chat-based symptom checker. Diagnostic accuracy of 52%.
  + Isabel: Relied upon by doctors and nurses around the world. Diagnostic accuracy of 68%.
  + Symptomate: Powered by Infermedica. Expect many branched logic follow-up questions. Diagnostic accuracy of 60%.
  + WebMD Symptom Checker: Most popular online symptom checker with interactive body map. Diagnostic accuracy of 59%.
* Infermedica 860+ conditions, 1700+ symptoms, and 320+ risk factors[[3]](#footnote-3)

**Research Objective (Prediction):**

Use the Kaggle data set to make the task of physicians easier by identifying which symptom features are the most significant predictors of a disease.

***How will it be used? Who will use it?***

It would be used by physicians in a clinical setting as part of patient intake. The symptoms would be input, and the physician would be provided a predicted prognosis to be used as part of the full diagnostic review.

**Dataset**

https://www.kaggle.com/datasets/kaushil268/disease-prediction-using-machine-learning/data

**Potential Revised Data Structure**

If we need to consolidate the dataset to reduce the number of diseases/ symptoms, we can use one of the following data structures in conjunction with/ as validation for our clustering analysis.

***Canadian Chronic Disease Surveillance System (CCDSS)[[4]](#footnote-4)***

**Cardiovascular diseases**

* acute myocardial infarction
* heart failure
* hypertension, excluding gestational hypertension
* ischemic heart disease
* stroke
* hospitalized stroke

**Chronic respiratory diseases**

* asthma
* chronic obstructive pulmonary disease

**Diabetes**

* diabetes mellitus (types combined), excluding gestational diabetes

**Mental illnesses**

* use of health services for mental illness and alcohol/drug induced disorders
* use of health services for mood and anxiety disorders
* use of health services for schizophrenia
* schizophrenia

**Musculoskeletal disorders**

* use of health services for arthritis
* gout and other crystal arthropathies
* juvenile idiopathic arthritis
* osteoarthritis
* rheumatoid arthritis
* osteoporosis
* osteoporosis-related fractures (any fracture, forearm, hip, humerus, pelvic, spine)
* osteoporosis-related fracture care gap (diagnosis, bone mineral density test, prescribed medication)

**Neurological conditions**

* dementia, including Alzheimer disease
* epilepsy
* multiple sclerosis
* parkinsonism, including Parkinson disease

**Multimorbidity (see “Additional items to consider” section below for further review)**

* two or more of the selected CCDSS chronic conditions
* three or more of the selected CCDSS chronic conditions

Two other potential sources to leverage to reduce/ validate the list of symptoms:

***Option 2 - Most commonly experienced symptoms[[5]](#footnote-5)***

* Flu
* Abdominal Pain
* Arm Pain
* Back Pain
* Body Aches
* Breast Pain
* Breathing Difficulty
* Chest Pain
* Congestion
* Cough
* Diarrhea
* Ear Pain
* Excessive Sweating
* Faintness
* Fatigue
* Fever
* Gas
* Headache
* Irregular Periods
* Joint Pain
* Leg Pain
* Mouth Lesions
* Nausea
* Neck Pain
* Rash
* Rectal Bleeding
* Skin Lump
* Sore Throat
* Vaginal Itching
* Vomiting

***Option 3 - Symptoms of Serious Diseases and Health Problems[[6]](#footnote-6)***

* Heart attack
* Stroke
* Reproductive health
* Breast problems
* Lung problems
* Stomach or digestive problems
* Bladder problems
* Skin problems
* Muscle or joint problems
* Emotional problems
* Headache problems (not including everyday tension headaches)
* Eating or weight problems

**Potential Data Limitations**

**Multimorbidity[[7]](#footnote-7):**

* Multimorbidity refers to the concurrent presence of two or more independent chronic medical conditions within an individual. As a chronic condition surveillance indicator, it involves tracking the prevalence and patterns of multiple chronic conditions within a population over time. The list of selected CCDSS chronic conditions in the multimorbidity estimates includes asthma, chronic obstructive pulmonary disease, dementia (including Alzheimer disease), diabetes (types combined, excluding gestational diabetes), epilepsy, gout and other crystal arthropathies, heart failure, hypertension (excluding gestational hypertension), ischemic heart disease, multiple sclerosis, osteoarthritis, osteoporosis, parkinsonism (including Parkinson disease), rheumatoid arthritis, schizophrenia, and stroke.
* 12.9% of Canadians report 2+ chronic diseases and 3.9% report 3+ chronic diseases[[8]](#footnote-8).

**Multicollinearity**

* Multicollinearity arises when at least two highly correlated predictors are assessed simultaneously in a regression model.[[9]](#footnote-9)

**Lack of Demographic and Social Determinants of Health (SDOH) variables**

* Article “Identifying diseases symptoms and general rules using supervised and unsupervised machine learning”[[10]](#footnote-10) references a Kaggle dataset “Disease Symptoms and Patient Profile Dataset[[11]](#footnote-11)” that was leveraged for their findings. This specific Kaggle dataset includes Age and Gender variables which are important for a robust analysis.

1. Gilbert S, Mehl A, Baluch A, et alHow accurate are digital symptom assessment apps for suggesting conditions and urgency advice? A clinical vignettes comparison to GPsBMJ Open 2020;10:e040269. doi: 10.1136/bmjopen-2020-040269 [↑](#footnote-ref-1)
2. https://openmd.com/directory/symptoms [↑](#footnote-ref-2)
3. https://infermedica.com/medical-knowledge-base [↑](#footnote-ref-3)
4. https://health-infobase.canada.ca/ccdss/data-tool/ [↑](#footnote-ref-4)
5. https://www.everydayhealth.com/symptom-checker/ [↑](#footnote-ref-5)
6. https://www.medicinenet.com/symptoms\_of\_serious\_diseases\_and\_health\_problems/article.htm [↑](#footnote-ref-6)
7. https://health-infobase.canada.ca/ccdss/data-tool/ [↑](#footnote-ref-7)
8. https://www.canada.ca/en/public-health/services/reports-publications/health-promotion-chronic-disease-prevention-canada-research-policy-practice/vol-35-no-6-2015/prevalence-patterns-chronic-disease-multimorbidity-associated-determinants-canada.html [↑](#footnote-ref-8)
9. https://pmc.ncbi.nlm.nih.gov/articles/PMC4888898/ [↑](#footnote-ref-9)
10. https://www.nature.com/articles/s41598-024-69029-8 [↑](#footnote-ref-10)
11. https://www.kaggle.com/datasets/uom190346a/disease-symptoms-and-patient-profile-dataset [↑](#footnote-ref-11)