*COMP4710 Data Mining Project*

Connor Bean

Claire Chen

Eddy He

James Froese

# Introduction

The novel corona virus of 2019 dubbed Covid-19 has profoundly altered life on earth in 2020. In addition to the human sickness and death associated with a pandemic, widespread economic shutdowns have been implemented across the globe in an effort to limit the spread of the virus. As students, our lives have been directly impacted as international travel has been vastly reduced, classes moved to online delivery, and personal connections limited. This unique situation provides us with an importunity to study a problem that is hugely relevant to our day-to-day lives, and where results may become immediately useful in the fight to control this disease.

As a result of these unique circumstances, we have decided to mine Canadian Covid-19 patient data in an effort to determine which demographics are most at risk of hospitalization and death and if these risk demographics change depending on geographic location. Utilizing data provided by federal and provincial government we intend to use a variation of the FP-growth algorithm to determine the most at-risk demographics around the country. Patient data generally includes age, sex, health status, and exposure type which we will mine to determine frequent attributes and hopefully draw some conclusion with regards to high-risk demographics. In addition, in comparing our results province-by-province we will determine if risk has any region-dependency.

Our contributions will be two-fold. First, we will examine high-risk demographics and how they change depending on province. While risk demographics associated with Covid-19 have been studied regionally in Canada already [see source in related work about BC, ON, QC], due to the rapidly evolving nature of an on-going pandemic experiencing a second wave of infections, our project will provide more results on a larger set of cases over more provinces. As British Columbia, Ontario, and Quebec contained the vast majority of the initial infections in spring of 2020, most research has been conducted on these specific provinces. At the time of writing (November, 2020) Alberta, Saskatchewan, and Manitoba are all experiencing significant outbreaks which will add a much larger sample to our mining. Second, We will alter the classic FP-growth algorithm to suit our particular needs efficiently. To achieve this, we propose removing the second database scan by representing the full database as a tree in memory, reusing available tree nodes where subsequent transactions have the same prefix. Each path of the tree will then represent a transaction from the database to which the FP-Tree can be constructed. Removing a second pass from the algorithm will reduce computation time and allow for quick data mining as the Covid-19 patient data available continues to grow throughout this pandemic.

# Related Work

The related and previous work considered in completing this project falls broadly under two categories. The first is the related work that characterizes how demographics are differently affected by Covid-19 infection and how that research educates the direction of our project. The second is how our selected data mining algorithm has been applied previously and how and if our selected changes have been implemented in similar cases.

While the study of Covid-19 only began in earnest this year for obvious reasons, there is significant research already completed that we may observe to direct our own goals for this project. One such preliminary study briefly compares the mortality rate between Chinese and Italian Covid-19 patients and reveals significant disparity in fatality rate between equivalent populations [1]. The fact that distributions of similar people in different geographic locations have different resilience to Covid-19 prompts inspection of this question on a finer regional level. While two countries as geographically and ethnically distinct as China and Italy may have considerable differences it may be worth examining if such regional changes in Covid-19 recovery rates exist regionally in Canada.

Examining how demographics react to Covid-19 across Canada has already been attempted in a paper entitled “Demographic Profile of COVID-19 Cases, Fatalities, Hospitalizations and Recoveries Across Canadian Provinces” [2]. This study compares the rates of hospitalization, fatality, and recovery of Canadians across the country by several different demographic characteristics. The authors break risk populations up by sex and age in a given region and found significant differences in the rate of hospitalization and death between provinces. However, this study was published in May of 2020 and is therefore extremely limited to a few select provinces that had substantial outbreaks at the time, namely British Columbia, Ontario, and Quebec. As of the time of writing (November 22, 2020), most provinces have substantial outbreaks ongoing in addition to significant daily hospitalizations and fatalities. With more up-to-date data and a larger sample size, it’s likely that we can find more similar results across the country.

A broader study done in Sweden examines finer demographics utilizing data about a subject’s age, net income, education, civil status, and country of birth [3]. The additional personal information allows for a finer examination of exactly what factors might influence a patient’s chances of surviving Covid-19. So while a simple examination of age and gender yields valuable information, a greater knowledge of an individual’s health and socio-economic position will provide much more detailed information on what risk factors are most important to patient death.

Finally, while examining how Covid-19 affects demographics differently is not a novel idea, it may provide novel results given that new data is being received on a daily basis. Moving into our method of mining we can take the lessons learned in previous studies to drive the direction of our data mining efforts.

##### References

Note: references are not yet properly formatted, or complete.

1. <https://onlinelibrary.wiley.com/doi/pdfdirect/10.1002/jmv.25860>
2. <https://www-deslibris-ca.uml.idm.oclc.org/ID/10103979>
3. https://www.nature.com/articles/s41467-020-18926-3