STEM Fellowship Big Data and AI Challenge 2023

Eliminating Randomness in selecting International Passengers for RT-PCR

Umid Ghimire[1], Dharmik Limbani [1], Erwin Mongui[1], and Sairam Kilari[1]

[1]Lambton College

June 11, 2023

Abstract

Travel across borders has been severely impacted by the COVID-19 pandemic, prompting the deployment of strict precautions to stop the virus' cross-border spread. Pre-travel testing using reports from RT–PCR tests (polymerase chain reaction) is one such measure. However, it still needs to be determined whether or not these reports can be relied upon to identify people who require retesting before travelling. As the new variations start to emerge and spread more quickly, concerns about whether they may make the recently authorized vaccinations less effective are unavoidable. This study intends to determine how well COVID-19 RT–PCR reports perform in differentiating those needing a retest before travelling. By thoroughly analyzing data from the World Health Organisation (WHO) and Govt of Canada Covid sources, the study aims to offer insights that can help the Immigration unit at Toronto Pearson International Airport identify particular travellers who are more likely to contract COVID-19 and may need special attention.

Keywords
covid-19, travel, pcr, retest

1 Introduction

The WHO first recognized Pneumonia of Unusual Cause, also known as COVID-19, on 31 December 2019 [1]. Years later, as of June 2023, we still find traces of COVID-19 even after vaccination, booster doses, and proper quarantine methods. The World Health Organisation has been interested in a novel COVID-19 sub-variant. Patients with this COVID-19 omicron variant, also known as "Arcturus", present with many of the well-known symptoms of earlier variants as well as a brand-new symptom: conjunctivitis [2].

International travel is becoming more accessible with lesser restrictions all around the world. Furthermore, RT–PCR tests are the most widely used method to identify the coronavirus [3]. According to WHO, on 7 June 2023, globally, there were 219,183 new cases last 7 days, 767,750,853 cumulative cases and 6,941,095 cumulative deaths [4]. Although we are taking a slightly relaxed approach to handling the COVID-19 situation, particularly for Ontario, while remaining mindful of new variants and more potent mutations, the number of tourists inbound to Ontario is 1,700,135 only in the first quarter of 2023 [5].

At the moment, the Immigration unit at Toronto Pearson International Airport randomly chooses travellers for RT–PCR reports. However, a more trustworthy strategy can be used because cluster data is readily available. The effectiveness of the screening procedure is increased by being able to choose individuals based on cluster data for repeat testing of their RT–PCR findings. The efficiency and accuracy of identifying probable COVID-19 cases among travellers can be improved by leveraging the available datasets and selection criteria, which will aid in managing and mitigating the virus's transmission overall.

2 Materials & Methods

2.1 Source of Data

The WHO records covid related cases from all over the world and makes the dataset publicly available [6]. We can inspect the new number of cases per thousand within seven days in the dataset across several nations. Moreover, the Canada health dataset for the epidemiology of COVID-19 consists of cases(hospitalizations,

ICU and deaths), as well as age and gender data over time [7]. A collection of RT–PCR reports from various travellers is also available from multiple countries. These reports include thorough details about the travellers, including the test date, age, gender, and pcr status (positive/yes or negative/no). Some of the reports were created with the sample available on the internet.

2.2 Statistical Analysis

2.2.1 Govt of Canada Covid Data

The Canada covid-19 data was carefully analyzed, and different visualization plots were created in order to compare the data with age group, sex, rate per 100000, and the status of the patient (ICU, deaths, hospitalization). Furthermore, a few more plots were observed, which were a combination of age, sex, and status, with the total count of the dataset.

2.2.2 WHO dataset

Based on the country, the WHO dataset was observed to have countries, cases - newly reported in the last 7 days, and cases - newly reported in the last 7 days per 100000 population. Using these attributes, two additional attributes were created: the percentage of cases newly reported in the last 7 days per 100000 population and the level of risk.

$$percentage\% = \frac{x}{sum(x)} * 100$$

where, x = cases newly reported in last 7 days per 100000 population.
The three values were assigned to the level of risk according to the percentage of cases where,

$$risklevel = \begin{cases} 2, & \text{if } percentage > 2 \\ 1, & \text{if } 1 <= \text{percentage} < 2 \\ 0, & otherwise \end{cases} \quad (1)$$

2.3 Report Extraction

The collected reports were extracted from PDF format to a readable CSV using Python operations. The CSV file includes information such as the date, RT-PCR status (positive/yes or negative/no), gender, age, and location. Additionally, the country attribute is added, extracted from the location, along with the RT-PCR validity, which indicates the number of days after the report date.

$$validity = Today() - date$$

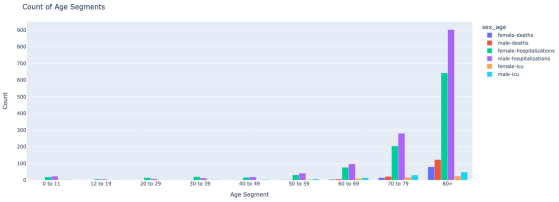

Figure 1: Age segmentation according to sex and case status

3 Results & Discussion

Overall, from the last six months' data, it can be observed that the number of individuals in the age segment 70-79 and 80+ holds the major proportion of the overall count while 50-69 holds a comparable number however, the rest of the categories display a minimal count. So for further observation, the data has been divided into 3 categories of age groups where first is 70+, the second is 50-69, and the third contains data of 0-49 aged people.

The count for different statuses: hospitalization, deaths, and icu also shows the same trend as the count of aged groups where people falling under the age group 70+ are having significant proportion in each category. Almost 85% of people falling under this age group need hospitalization, making it a crucial factor for further analysis.

People under the age group 70+ and 0-11 have a higher proportion of males however, for the rest of the age group number of females surpasses the males. Hospitalization and deaths exhibit a similar trend in their number while ICU status remains fluctuated through the last 6 months time period, although, it also demonstrates a downfall trend.

After visualizing the data, international travellers can be segmented into three possible segments - Immediate Priority, High Priority & Medium Priority and Low Priority.
Immediate Priority- The Passengers who need immediate attention, i.e. should be selected for RT-PCR testing.
High Priority and Medium Priority - Passengers who can be randomly selected for RT-PCR testing but do not require immediate attention.
Lower Priority - Passengers not fulfilling the Immediate or High Priority criteria.

The Definition of the Priority is decided by Canada's Covid Cases Data visualizations from the last 7 days based on the factors of Age and Gender. Based on the data visualizations, the definitions of the segments are,
Immediate Priority Level 3- Travellers of age

70 and above, with RT-PCR positive reports more than 14 days (2 weeks) old. Both Males and Females travel from highly affected countries (Risk Level 2).
Immediate Priority Level 2A- Travellers of age 70 and above, with RT-PCR positive reports less than 14 days (2 weeks) old. Both Males and Females travel from highly affected countries (Risk Level 2).
Immediate Priority Level 2B- Travellers of age 70 and above, with RT-PCR positive reports greater than 14 days (2 weeks) old. Both Males and Females travel from low-affected countries (Risk Level 1 or less).
High Priority Level 3- Travellers of age 50 to 69, RT-PCR positive reports greater than 14 days old. Both Males and Females travel from highly affected countries (Risk Level 2).
High Priority Level 2A- Travellers of age 50 to 69, RT-PCR positive reports less than 14 days (2 weeks) old. Both Males and Females travel from highly affected countries (Risk Level 2).
High Priority Level 2B- Travellers of age 50 to 69, RT-PCR positive reports greater than 14 days old. Both Males and Females travel from low-affected countries (Risk Level 1 or less).
Medium Priority Level 2A: Travellers of age 50 and above with RT-PCR negative reports.
Medium Priority Level 2B: Travellers of age 0 to 49 with RT-PCR positive reports.
Low Priority Level: Travellers of age 0 to 49 with RT-PCR negative reports.

In the data extracted from reports, we have assigned different segments to group the travellers. It was observed that two travellers fall under Immediate Level 2B, eight on High Level 2B and nineteen on Medium Level 2A.

Positive	Gender	Age	Country	Validity
yes	female	72	india	93
yes	male	72	mali	647

Table 1: Travellers on segment Immediate level 2B

Positive	Gender	Age	Country	Validity
yes	male	58	India	576
yes	male	67	India	648
yes	female	67	India	658
yes	female	58	India	576
yes	male	56	India	938
yes	male	64	Albania	681
yes	male	64	Albania	994
yes	male	64	Albania	994

Table 2: Travellers on segment High level 2B

Positive	Gender	Age	Country	Validity
no	female	60	India	784
no	male	60	India	784
no	female	55	India	797
...	...	...	...	...
no	female	55	India	797
no	female	72	Greenland	573
no	male	61	Greenland	573

Table 3: Travellers on segment Medium 2A

Conclusions

The numbers of new COVID-19 strains are emerging and coinciding with the relaxation of travel measures. The peak of the pandemic was a dire situation, and it is uncertain whether vaccinations provided are adequate protection. Being cautious is most important, as we can observe from the data collected over time. Also, randomly selecting and checking the traveller is not an optimal strategy for dealing with travel protocols. With this study, we could segment international travellers according to age, sex, location and RT-PCR reports. This method facilitates more efficient arrangements for screening travelling individuals. By leveraging the information based on demographic groups, we can identify the severity and make good decisions. Consequently, proactive measure enables us to be prepared for potential surges in COVID-19 spread and enhance our ability to provide adequate care and support.

Acknowledgements

We would like to thank our Prof. Mr. Ashish Gupta for his mentor ship and support through out the challenge. We would also like to thank STEM Fellowship for organising the challenge and providing us the opportunity to participate.

References

[1] WHO. Pneumonia of unknown cause – china, . URL https://www.who.int/emergencies/disease-outbreak-news/item/2020-DON229.

[2] Meagan Drillinger. What to know about the new covid-19 strain 'arcturus'. URL https://www.healthline.com/health-news/new-covid-19-strain-arcturus-what-to-know.

[3] Stephanie Ward, Andrew Lindsley, Josh Courter, Amal Assa'ad. Clinical testing for COVID-19. 146(1):23–34, 2020. doi:

https://www.sciencedirect.com/science/article/pii/S0091674920307296.

[4] WHO. Global situation, . URL https://covid19.who.int/.

[5] Ontario. Tourism research statistics. URL https://www.ontario.ca/page/tourism-research-statistics.

[6] WHO. Data information, . URL https://covid19.who.int/data.

[7] Government of Canada. Covid-19 epidemiology update: Current situation. URL https://health-infobase.canada.ca/covid-19/current-situation.html.