STEM Fellowship Big Data Challenge 2020

An Investigation into the Relationship Between Malarious Countries and COVID-19 Cases

Anne Jing, Vimal Raj, and Charles Liu

Faculty of Applied Science and Engineering, University of Toronto

June 1, 2020

Abstract

Studies have shown that a negative correlation may exist between the number of COVID-19 cases and the incidence of malaria. This relationship is important, as it can provide valuable insight into potential treatments of the virus. To further investigate the correlation, 66 other factors (such as country demographic, GDP per capita, Health Index etc.) were examined. Open source data-sets were collected online from institutions such as the John Hopkins Coronavirus Resource Center and analyzed using various Python libraries. By calculating various correlation coefficients, it was determined that the number of tests conducted, air traffic and other several other factors correlated with COVID-19 cases, and may be responsible for the lower number of COVID-19 cases instead. To further strengthen the results, a machine learning model was constructed to evaluate the most relevant factors contributing to the transmission of COVID-19. Here, malaria was not deemed as an important feature. Instead, factors such as climate and country wealth were most important. A mixture of statistics and machine learning helped show that various factors unrelated to malaria may be responsible for the lower occurrence of COVID-19 in malarious countries, and while this is not enough to prove causation, it makes it likely that the correlation between COVID-19 and the Plasmodium parasite is spurious.

Keywords
COVID-19, malaria, machine learning, cases

1 Introduction

At the end of December 2019, the Wuhan Municipal Health Commission in China reported an abundance of flu-like cases in Wuhan[1]. 14 days later, the first case of this mysterious novel virus was found outside of China. This virus, known as COVID-19, is a pandemic that has infected over 5.7 million people and has killed over 300,000 people, as of May 27, 2020, worldwide[2].

Statistics about the virus including fatality rate and R0 value are influenced by numerous factors. One of the factors being explored is the presence of malaria in a country.

Past studies have shown that malarious countries have a significantly lower number of COVID-19 cases as compared to the global average. In the paper "Global Spread of Coronavirus Disease 2019 and Malaria: An Epidemiological Paradox in the Early Stage of A Pandemic“ from the University of Cagliari, a negative correlation was found between COVID-19 cases and malaria, suggesting that malaria "might play a role in limiting the spread of COVID-19" [3]. While this may be related to anti-malarial drugs or a biological resistance, there are various other factors that may be responsible. This study examines these other factors, and aims to find an explanation for the link between COVID-19 and malaria.

2 Methods

Open-source data regarding the total number of COVID-19 cases was collected from the John Hopkin Coronavirus Resource Center, while data regarding the estimated number of cases of malaria was collected from WHO's data repository. All countries were first divided into 2 categories: malarious and non-malarious, dependent on whether or not the country had at least one known case of malaria. 83 malarious and 103 non-malarious countries, were observed. To account for the vast difference of population between countries, several factors were adjusted for by looking at counts per 100,000 people.

To find links between the spread of COVID-19 and other environmental factors, climate, population, and air travel datasets were downloaded from Worldbank. Moreover, datasets related to the Quality of Life index and the Human Development Index were acquired from Numbeo and Kaggle.

Using the pandas package for Python3, the datasets were cleaned and merged to generate one overarching dataset, containing 66 different factors for each country including discrete but critical factors such as human development, GDP per capita, and wind speed.

A preliminary examination of the factors was conducted by calculating the Pearson Correlation Coefficient[1], denoted as r, between all the factors. These coefficients demonstrate the correlation strength between factors to give a preliminary understanding of the importance of the factors to COVID-19 cases. A +/-1 indicates a perfect correlation while a 0 indicates no correlation. All factors with a correlation coefficient over 0.5 were plotted versus COVID-19 cases to further examine their linear dependencies.

Lastly, a multivariate linear regression model[2] was constructed to estimate the number of COVID-19 cases in a country using all the factors collected as features. The 'Extreme Gradient Boosting' algorithm was used, present in Python's xgboost library, to improve prediction accuracy. The algorithm utilizes decision tree ensembles[3] to reduce misclassification rate. After training the model on the entirety of the dataset, the importance of each feature using the F score was examined[4].

3 Results

The choropleth maps in Figure 1 show the spread of COVID-19 and malaria around the world. It can be observed that countries with more cases of malaria generally tend to have lower cases of COVID-19. However, there are outliers such as northern parts of South America and South Asia. This demonstrates that other factors influence the number of confirmed cases of COVID-19.

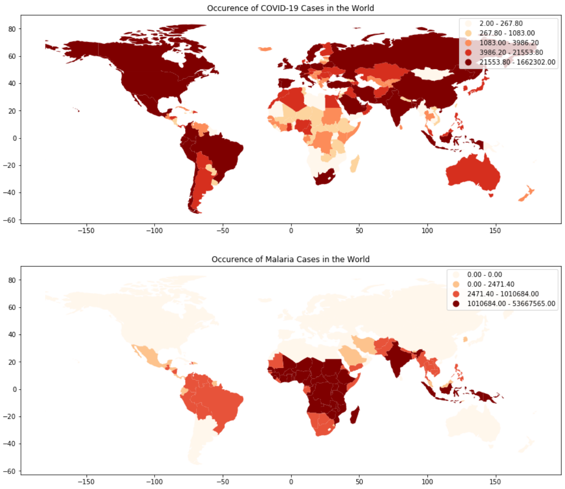

Figure 1: Choropleth maps of COVID-19 and malaria occurrences.

Figure 2 illustrates this relationship even further. There is a clear difference between the average number of COVID-19 cases in non-malarious and malarious countries, proving that there is a link that is to be explored.

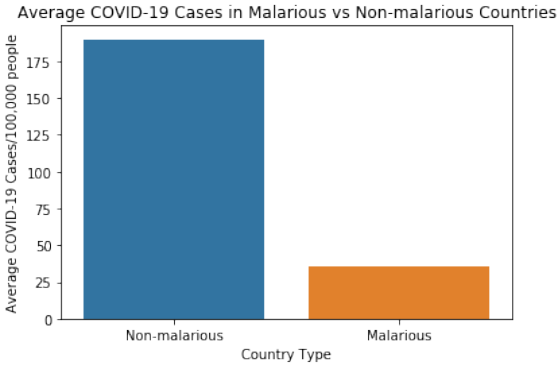

Figure 2: The number of average COVID-19 cases reported in malarious and non-malarious countries.

Figure 3 shows a positive correlation ($r = 0.93$) between the total number of COVID-19 per 100,000 people and tests conducted per 100,000 people. Interestingly, it can be seen that non-malarious tend to have more tests conducted, and a higher number of cases.

[1] A measure of the linear relationship between two variables

[2] multiple correlated dependent variables are predicted using independent variables

[3] A graphical representation of tests conducted and their respective outcome

[4] Measures a test's accuracy

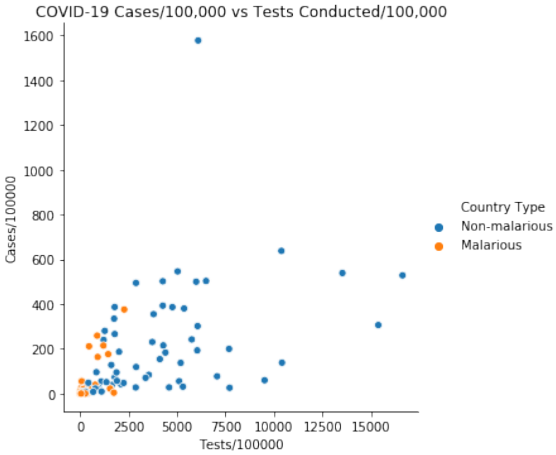

Figure 3: A scatter graph of the total COVID-19 cases against the total number of tests conducted. Countries on the graph are divided by malarious or non-malarious.

Figure 4 shows another positive correlation, this time between the total number of COVID-19 cases and the number of air passengers carried to a country. It also seems that non-malarious countries tend to have higher air traffic.

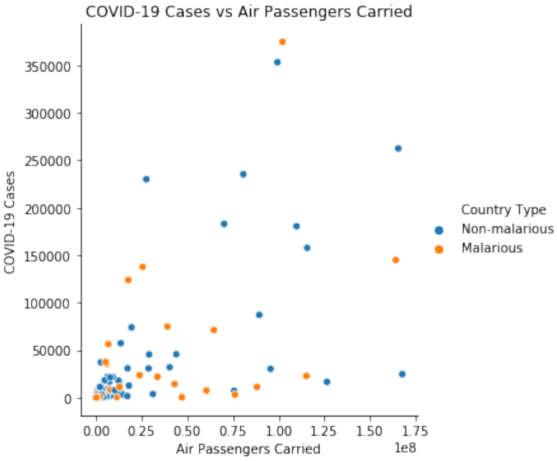

Figure 4: A scatter graph of the total COVID-19 cases against the number of air passengers to said country. Countries on the graph are divided by malarious or non-malarious.

This is further demonstrated by Figure 5, which shows that the number of air passengers travelling to non-malarious countries is significantly greater than to malarious countries.

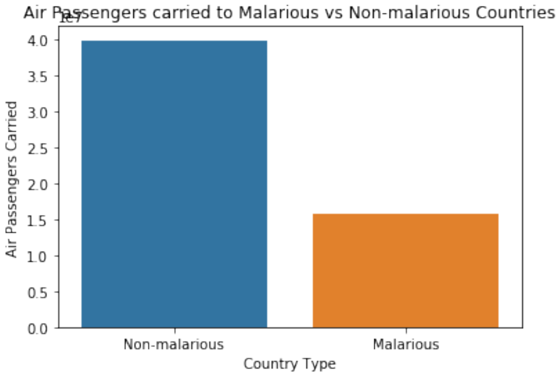

Figure 5: A bar graph portraying the average number of air passengers carried to malarious and non-malarious countries.

In Figure 6, a positive correlation between the number of COVID-19 and the net number of migrants was observed. Here, a negative number of migrants indicates more people have left the country, while a positive number indicates more people have migrated to the country.

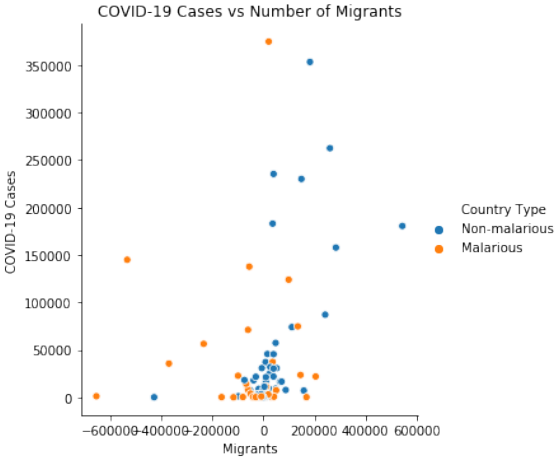

Figure 6: A scatter graph of the total COVID-19 cases against the total number of migrants to said country. Countries on the graph are divided by malarious or non-malarious.

Moreover, Figure 7 shows that people tend to migrate away from malarious countries and into non-malarious countries.

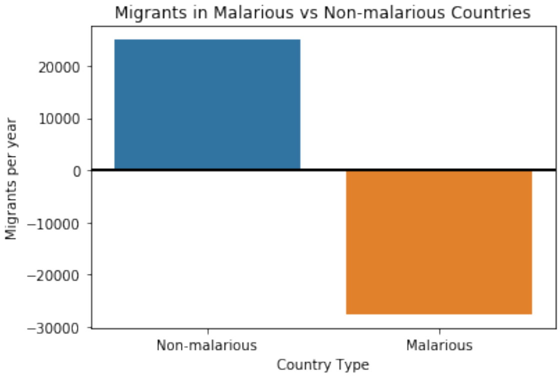

Figure 7: A bar graph plotting the migrants in malarious and non-malarious countries.

Figure 8 plots the ten most important features, as determined by the regression model during training. Factors such as air passengers carried and tests conducted are important, matching with the statistical results. However, other factors such as population density, latitude and urban population are of equal importance. While the number of malaria cases is used by model to make predictions, it appears to be less important than the others.

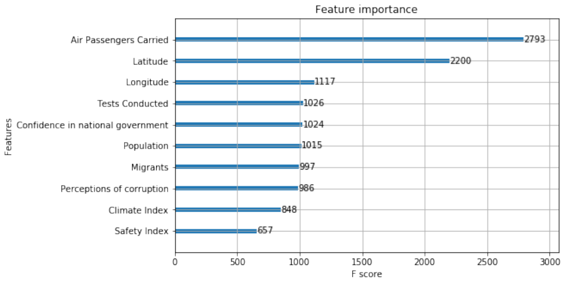

Figure 8: A bar graph showing the feature importance of the machine learning model.

4 Discussion

Figure 1 and 2 prove the existence of an inverse correlation between COVID-19 and malaria. They are two many hypotheses that can explain this effect. Firstly, an evolutionary adaptation caused by the malarial parasite may provide a resistance against COVID-19. One example is a change in the ACE2 receptor, which would prevent the virus from infecting human cells [4]. More studies are required to provide substance to this hypothesis.

Additionally, there could be a direct link based on availability of anti-malarial drugs and virtual resistance. According to the CDC, malarious countries all treat and have direct access to chloroquine, which akin to hydroxychloroquine, is derived from the quinoline molecule [5, 6]. However, there are multiple findings indicating that antimalarial drugs are ineffective against COVID-19 [7].

Figure 3 shows an obvious and intuitive correlation between the number of tests conducted and the number of COVID-19. It can also be observed that malarious countries have conducted fewer tests. This can be explained by the lack of access to testing facilities, as malarious countries have lower GDP per capita, higher corruption levels and unsatisfactory healthcare systems [8]. It is important to keep in mind that this does not result in a lower number of COVID-19 cases, but rather indicates an underestimation of the real number of cases.

Air traffic was another significant factor in the spread of COVID-19. As seen in Figure 4, the relationship between them is one of high correlation ($r = 0.81$). Moreover, past studies have shown the same link. [9]. Figure 5 shows that air travel is rarer in malarious countries. Together, these graphs can be used to show that air traffic may be responsible for the lower number of COVID-19 cases, rather than malaria.

In Figure 7, it can be seen how migrants generally move from malarious countries to non-malarious countries. Meanwhile, Figure 6 shows the number of migrants is positively correlated to the number of COVID-19 cases. While not a strong correlation ($r = 0.59$), there may be potential reasons for this. Countries with more migrants tend to be more diverse, and in turn have more citizens travelling back and forth. However, it is also possible that this correlation is spurious and therefore should not be considered highly.

These statistical results have proven the existence of a link between various biological and economical factors unrelated to malaria, and the lower number of COVID-19 cases observed in malarious countries. While this is not enough to prove a causation, it provides valid reasons for why the relationship between malaria and COVID-19 may be fictitious.

These correlations provide an overview of the relationship between two distinct factors (such as COVID-19 cases and the number of air passengers), relationships in real life tend to be more complex. In particular, instead of multiple binary situations, factors that may affect the transmission of COVID-19 form a complex web in which all the factors are interwoven.

The machine learning model was constructed to simulate the complex relationships between all factors and provide deeper insights to what may be affecting the transmission of COVID-

19. Traditional multivariate regression was used in conjunction with an extreme gradient boosting algorithm. Here, many independent models that are weak learners come together to form one strong model by utilizing tree ensembles. The models are made sequentially, with new models learning from previous ones, leading to an overall decrease in the number of iterations required to make a good prediction[10].

To prevent multicollinearity, [5] certain features that were strongly correlated with other features were removed. For example, the Quality of Life Index and Life Ladder Rating were very linearly dependent, and so the latter was removed due to it's inaccurate nature. Moreover, a lower number of features reduced the chances of overfitting. While the model learned how to predict COVID-19 cases, the main objective was to see which features it deemed as most important, as shown in Figure 9.

As seen, malaria does not seem to be an important feature, and was not used by the model to predict cases. From the features that were important, the effect of air passengers, tests conducted and migrants has already been investigated into using statistics. However, climate has not yet been considered. The model indicated that the latitude and longitude, factors primarily affecting the weather, along with the climate index are important factors. Studies show that cold, dry, unventilated air may contribute to influenza transmissions[11]and malarious countries are generally situated near the equator and have higher humidity than the global average[12, 13]. This might be contributing to the lower number of COVID-19 cases in these regions.

The other factors presented by the model, such as safety index and confidence in government are more difficult to investigate into. However, they can be connected to a country's overall wealth, and thus their access to testing facilities. This link remains to be explored in future studies.

As the model does not succumb to the same problems faced by a statistical approach, it provides more rigorous evidence against the effect of malaria on the transmission of COVID-19. Together, these approaches provide various different factors that may be responsible instead.

5 Conclusions

The data explored in this study has shown that the inverse correlation between the incidence of malaria and the number of COVID-19 cases is likely to be spurious. The significantly lower amount of COVID-19 cases in malarious countries can instead be explained by factors just as tests conducted, air traffic, and climate. While research into anti-malaria drugs as a treatment for COVID-19, or a biological malarial resistance to the virus may still prove to be useful, it should not be a priority due to the lack of a statistical ground.

Acknowledgements

We would like to thank the committee of STEM Fellowship for hosting this event. We'd also like to thank our mentor, Daiva Nielsen, and Anish Verma for their expertise.

References

[1] WHO. Who timeline - covid-19.

[2] WorldOMeter. Coronavirus cases.

[3] Nioi Napoli. Global spread of coronavirus disease 2019 and malaria: An epidemiological paradox in the early stage of a pandemic. *Journal of Clinical Medicine*, 9(9), 2020.

[4] Ellington Molineux Bull Paff, Nuismer. Virus wars: Using one virus to block the spread of another. *PubMed*.

[5] CDC. Malaria information and prophylaxis, by country.

[6] Amboss. Chloroquine and hydroxychloroquine.

[7] Lawton. Us study indicates ineffectiveness of antimalarial drug in covid-19 patients. *Euractiv*.

[8] Dobromirov Lučić, Radišić. Covid-19 and the cardiovascular system. *Economic Research-Ekonomska Istraživanja*, 1(29):360–379, 2016.

[9] Chad R Wells, Pratha Sah, Seyed M Moghadas, Abhishek Pandey, Affan Shoukat, Yaning Wang, Zheng Wang, Lauren A Meyers, Burton H Singer, and Alison P Galvani. Impact of international travel and border control measures on the global spread of the novel 2019 coronavirus outbreak. *Proceedings of the National Academy of Sciences*, 117(13):7504–7509, 2020.

[5] a situation in which two or more explanatory variables in a multiple regression model

[10] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[11] Ana Sandoiu. How humidity may affect covid-19 outcome. *Medical News Today*.

[12] Bakhtiari Tabatabai Mohammadkhani Mohammadkhani, Khanjani. The relation between climatic factors and malaria incidence in sistan and baluchestan, iran. *Sage Journals*, 9(5), 2019.

[13] Open Learn Create. Communicable diseases module: 6. factors that affect malaria transmission. *Open Learn Create*.