Foundations of Data Science

Answer Template for CW2: Critical Evaluation of a Data Science Study

Which Study did you choose?

- Real-time tracking of self-reported symptoms to predict potential COVID-19

1. The following questions relate to the scientific paper of your chosen study. Note that the average length of a sentence is 15-20 words, the average length of a paragraph has between 3 to 6 sentences.
    a. [Basic] What is the scientific goal of the study? Write one to three sentences.

       The study aims to investigate whether anosmia is a symptom specific to COVID-19. Using this information, it strives to create a mathematical model combining symptoms most predictive of the virus and apply it to the data of users who were not tested, to identify individuals who are likely to be infected.

    b. [Basic] What is the type of study (e.g., randomized trial, prediction model)? Include a brief explanation why it falls onto that category. Write one to three sentences

       The study is a case-control study. There are two groups divided based on the outcome (positive/negative) of the RT-PCR SARS-CoV-2 test. The two groups are compared based on the suspected causal attribute, anosmia, to determine if it is associated with COVID-19. This fits the definition of observational case-control study (Wikipedia, 2021).

    c. [Advanced] What are the hypotheses of the study? Write one paragraph (3-6 sentences).

       The study addresses two hypotheses. The first is that anosmia is associated with being infected with SARS-CoV-2. It is used to determine whether anosmia should be included in the predictive model the study aims to develop. The second hypothesis is that the association between anosmia and COVID-19 might have been influenced by mainstream media. It is used to question the validity of the model. The hypotheses, especially the second one, are rather implicit.

d. [Advanced] What is the methodology of the study and what is the results (how do the researchers go about testing the hypothesis, and what is the outcome? Write two or three paragraphs.

Data in the form of self-reported information related to COVID-19 were collected from the users of the mobile app *COVID Symptom Study* in the UK and US. To investigate the association between anosmia and COVID-19, the odds ratio (OR) of anosmia was calculated for the UK and US cohort, respectively. The results were then combined, yielding an OR of 6.74 (95% CI = 6.31–7.21). The result is that the effect size of anosmia is approximately 574%, indicating a *strong* association between anosmia and COVID-19. The study does not investigate the underlying mechanism causing anosmia in the presence of infection, instead providing pointers for further investigation.

Having identified anosmia as a dominant predictor for COVID-19, logistic regression was applied to devise the predictive model. The OR for each independent variable is given in Fig. 1a. The resultant sensitivity and specificity of the model on the UK test set was 0.65 (95% CI = 0.62–0.67) and 0.78 (95% CI = 0.76–0.80), respectively, as can be seen from Fig. 1b. Similar values are reported for the US validation set (Fig. 1c).

The study questions the validity of the model by postulating that the association between anosmia and COVID-19 might have been influenced by mainstream media. The researchers identified three date ranges with increasing media coverage of the association between anosmia and COVID-19. The OR of anosmia was calculated for each date range. As a result, an increasing trend for the UK cohort was observed and used to argue that the association between anosmia was influenced by mainstream media in the UK. No such trend was observed for the US cohort, which is used to argue that the association between anosmia and COVID-19 is strong.

e. [Intermediate] What are the statistical methods used in the study and how are they applied to the data? Write one or two paragraphs.

Records where the self-reported characteristics such as age, height, weight, BMI and temperature fell within reasonable ranges were used for analysis. Metrics were reported along with a 95% confidence interval (CI), which was obtained by bootstrapping with replacement. To investigate the association between anosmia and COVID-19, an implicit null hypothesis was posed: “There is *no* association between anosmia and COVID-19.” The method of multivariate logistic regression including age, sex and BMI as predictors was used to obtain the odds ratio of anosmia. This was done for both, UK and US users of the app who underwent the SARS-CoV-2 test, respectively. The associated p-value was strictly less than 0.0001. Similarly, upon

combining the ORs using inverse variance fixed-effects meta-analysis, the resultant p-value was in the same interval, indicating strong evidence against the null hypothesis. Thus, the null hypothesis was rejected, instead asserting the (implicit) alternative hypothesis that there *is* an association between anosmia and COVID-19.

To train the predictive model, the data were filtered to only include UK individuals who had been tested and answered a question about anosmia and at least 9 out of 10 other questions on the symptom report. The dataset was randomly split into training and test sets in an 80:20 ratio. The method of stepwise multivariate logistic regression combining forward and backward algorithms was applied to the training data. The predictors were anosmia, ten other symptoms[1], age, sex and BMI. The performance of the model was evaluated on the UK train set using tenfold cross-validation with additional validation on US data. The model with the lowest Akaike information criterion was selected. Finally, the ROC-AUC, sensitivity, specificity, positive and negative predictive values were reported. This was repeated with stratification for sex and age groups, yielding similar results. Furthermore, another model was trained excluding anosmia as a predictor, leading to a higher specificity at the cost of a large decrease in sensitivity to approximately 0.33. Therefore anosmia was kept as a predictor.

To determine whether the association between anosmia and COVID-19 might have been influenced by mainstream media, the OR of self-reported anosmia for the UK cohort was calculated for each date range using multiple regression as previously mentioned. The resultant ORs were 4.98 (95% CI = 4.47–5.56), 6.64 (95% CI = 5.75–7.68) and 10.40 (95% CI = 9.08–11.91), indicating an increasing trend. The same method was applied to the US cohort, yielding the OR values: 8.13 (95% CI = 5.18–12.78), 12.30 (95% CI = 8.96–16.90) and 9.13 (95% CI = 6.73–12.38), with no observable trend.

f. [Basic] What are the stated conclusions of the study? Write one paragraph.

The authors conclude that there is a *strong* association between COVID-19 and anosmia. This association remains strong despite possible bias introduced by mainstream media reports in the UK. The authors therefore suggest that anosmia become part of routine screening for COVID-19 and be added to the WHO symptom list for COVID-19. Finally, it was predicted using the devised model that 17.42% (95% CI = 14.45%–20.39%) of users who reported symptoms are likely to have COVID-19.

[1] *Fever, persistent cough, fatigue, shortness of breath, diarrhoea, delirium, skipped meals, abdominal pain, chest pain and hoarse voice*

g. [Advanced] Provide a critical discussion of the paper. How important or impactful is it? Are there any obvious errors or potential flaws? Write one or two paragraphs.

The study is impactful because it has implications for public health. Eight months after its publication, the pandemic is still ongoing. Identifying anosmia as a strong indicator and devising a predictive model of infection with COVID-19 enables organisations such as the WHO and local governments to make more effective measures, such as recommending self-isolation, ultimately protecting public health. Nonetheless, the study has its shortcomings. It reports the following. Firstly, data were self-reported. The lack of formal evaluation of symptoms and the possibility of human error when entering the data could have biased the data. Secondly, the study uses the outcome of RT-PCR SARS-CoV-2 tests as ground truth in training the model. The tests have limited sensitivity and specificity, which may have introduced false positives and false negatives into the datasets, making the model less accurate. Thirdly, it is not known whether anosmia develops prior to other symptoms and whether it accompanies the illness or comes after it. Such information would have implications for the role of anosmia in predicting COVID-19 infections. Furthermore, the sample used to train the model is likely not representative of the general population. It consists of those who underwent an RT-PCR SARS-CoV-2 test. These people are more likely to have the virus in the first place. This could have biased the data and resulted in the model overestimating the number of positive cases. Similarly, the users of the app are a self-selected group who may not be representative of the general population, indicating that the model might not scale well. Lastly, there may be bias in the data due to media reports. As the study indicates, media reports on the association of anosmia with COVID-19 may have biased data in the UK.

However, there are other possible flaws that the study fails to mention. Firstly, multicollinearity was not investigated. For example, the predictors “anosmia” and “skipped meals” might be correlated, as loss of smell and taste likely reduces one’s appetite. If the correlation is strong, it should be adjusted for (e.g. by performing PCA), to prevent unstable estimates of regression coefficients. Secondly, the model might not be perfectly applicable to US individuals. The model was trained on UK data only and then applied to untested individuals from both, UK and US, to estimate the number of infections. However, the study itself reported that although all ten symptoms were associated with testing positive in the UK cohort, only three were associated with a positive test result in the US cohort. This implies that the model trained on UK data might not scale perfectly to US data. Thirdly, there may be lurking variables. Anosmia, age, sex, BMI and ten other symptoms reported by the WHO were used as predictors in the model. However, adjusting for different medical conditions (e.g. asthma) or lifestyles (e.g. smoker) could have improved the performance of the

model. Finally, the method for choosing the date ranges used to study the association between media reports and anosmia was not explained. Neither the means of quantifying the effect of media reports, nor the method used to determine the date ranges were explained. This makes the study less transparent.

h. [Intermediate] Are there any ethical implications of the study? If none, please state so; if yes, how well do the authors relate to these implications? Write one paragraph.

The study complies with ethical standards. Competing interests were clearly outlined. Consultation of and approval by the King's College London Ethics Committee (UK) and Partners Human Research Committee (US) is clearly indicated. Furthermore, the source code for the application was made publicly available. Any form of data collection was preceded by user consent, and the collected data were stored in a protected manner. Collected data were anonymised and made available to other researchers.

2. The following questions relate to the media report of your chosen study:
Chosen report: Coronavirus: research reveals way to predict infection – without a test
   a. [Basic] Provide a brief summary of the report. Write up to three sentences.

The report explains the motivation for developing the application, how it is used, and it provides a simple run-down of gathered data and the accuracy of the mathematical model. It also reports on ongoing activities and calls out authorities for failing to update lists of symptoms on their websites. Finally, it encourages readers to self-isolate if they experience anosmia and invites them to download the app and contribute with personal data in an aim to contain the virus.

   b. [Advanced] How accurately did the report summarize the study? Write one or two paragraphs.

The report contains information not included in the study, such as the ongoing developments, for example validating the model on more incoming data and producing data on risk factors. In what concerns the study, the report provides an informal overview of the gathered data, but it focuses mostly on the mathematical model and the prominence of anosmia as a predictor therein. However, the report fails to mention the limitations of the study, such as the possibility of bias in the data which were used to train the mathematical model. Furthermore, the report uses a headline ("Research reveals way to predict infection – without a test") which is moot

since the model indeed is not 100% accurate. The report chooses accuracy as the (only) metric of choice when presenting the model, claiming it is 80% accurate. However, accuracy alone is not a sufficient measure of the model's performance, especially for imbalanced problems such as the one at hand, where the number of infected may be significantly lower than the number of not infected. In fact, the study itself does not mention the accuracy of the model, instead stating metrics such as specificity and sensitivity.

To conclude, whether intentionally or not, the report leaves out some crucial details. Therefore, it cannot be said that the report is an accurate summary of the study.

Bibliography

Wikipedia, 2021. *Case-control study.* [Online]
Available at: https://en.wikipedia.org/wiki/Case-control_study
[Accessed January 2021].