FDS S2 Week 4 Workshop, 2024/25

Model answers for COVID study

These answers are produced by FDS course staff. You may have spotted things we've missed or have queries. Please post any queries or comments on this Piazza post.

All unattributed quotations are from Menni C, Valdes, AM, Freidin MB. et al. (2020). Real-time tracking of self-reported symptoms to predict potential COVID-19 *Nat Med* **26,** 1037–1040. doi: 10.1038/s41591-020-0916-2

Part 1B – Read the paper

1. Question: What question does the paper address?

The main goal is to identify the combinations of symptoms that best predict whether an individual has COVID-19, and to determine if this could be used, at a time when there was limited COVID testing, as a tool to suggest whether individuals showing symptoms should self-isolate. Another question is whether anosmia (loss of taste or smell) can predict COVID-19. A third question is whether logistic regression can be used to predict positive COVID-19 tests.

2. Category: What type of paper is this? A paper collecting survey data, a randomised controlled trial, an analysis of an existing dataset?

It is an observational study, the data from which is being used to build a prediction model. It is observational since it makes use of available data; the subjects are not randomly selected from a population, so there is no little control over the sample. We can also say that the study is longitudinal, since data is over time for the same participants, and it can also be seen as a case-control study.

3. Context: What dataset(s) does it use? How were the datasets collected?

The researchers used a symptom-tracking app developed by their research group (Drew et al. *Science* **368**:1362-1367, 2020) to collect daily reports of symptoms from 2,450,569 UK and 168,293 US users, who also reported if they had had a COVID test, and if so, whether it was negative or positive. Data collected between 24 March 2020 and 21 April 2020 was filtered by age (16-90 years in the UK; 18-90 years in the US), height (110-220cm), weight, BMI and body temperature.

4. Which methods were used to analyse the problem?

A multivariate logistic regression was fit to the 15,638 UK users who had had COVID tests. The independent variables were age, sex, BMI and loss of smell or taste and the dependent variable was whether the test was positive or negative.

Further logistic regression models were trained including the other ten symptoms reported by participants ("fever, persistent cough, fatigue, shortness of breath, diarrhoea, delirium, skipped meals, abdominal pain, chest pain and hoarse voice") as independent variables.

5. Correctness: Do the assumptions appear to be valid?

The assumptions largely have to do with the data and how it was collected. The paper reasonably splits UK and US individuals, but there could still be some issues with bias in the dataset, because the dataset is self-selected, not random – for example the requirement to own a smartphone may induce a bias. Also, as the report states, the symptoms are self-reported rather than assessed by a doctor.

Furthermore, as the report states, the set of individuals tested for COVID-19 is not random, since individuals who are tested are more likely to ones severely infected or who have been in contact with individuals known to have COVID-19. It is also not clear what time window was used to link symptoms and a positive test – e.g. did the symptoms have to be present on the same day as the test, or within a window of 1 or 2 days before the test?

Also the choice of symptoms chosen to include in the model could count as assumptions – perhap there are other symptoms (e.g. aching limbs) that could have been included, and which could have affected the results.

However, the dataset is still large, and some of the above issues may be quite hard to solve in practice (e.g. having a doctor assessing symptoms would likely greatly reduce dataset size and would increase risk of infection). Given the circumstances, it's arguable that the assumptions were valid enough to draw some conclusions from the data.

6. Contributions: What are the paper's main contributions?

The authors state that the "proportion of participants who reported loss of small and taste was higher in those with a positive test result (4,668 of 7,178 individuals; 65.03%) than in those with a negative test result (2,436 of 11,223 participants; 21.71% (odds ratio = 6.74; 95% confidence interval 6.31-7.21)". They conclude that, in addition to other symptoms, the loss of smell is a "potential predictor" of COVID-19. They also conclude that, on the basis of their prediction model, that 17.42% of app users reporting symptoms are likely to have COVID-19.

The above conclusions are their main contributions, but the development of the symptom-tracking app and the data obtained from it are also an important contribution.

7. Clarity: Is the paper well written?

Generally, yes. The goal of the paper is clear, and so is most of the methodology and conclusions. There are some potential issues – some details of the methodology aren't given (e.g. it's not clear what method is used to generate the confidence intervals or the *p*-values), Table 1 is arguably too big/overwhelming, and the paper would be better with sections (though these were likely skipped due to formatting and size restrictions imposed by the publisher). There are likely other issues that you may notice too.
However, it's important to realise that no paper is completely perfect, and this one is one of the more decent ones.

8. Your own understanding: From what you've looked at so far, is there anything you still don't fully understand? It's very unlikely that the answer to this will be "No" - even experienced researchers don't understand every part of every paper they read immediately. Furthermore, there is at least one method - ROC - that hasn't been covered in lectures yet, so it's likely you won't know exactly what it means at this point.

This obviously has no set answer, but it's important to be honest (at least with yourself) about what you don't understand. Some things to consider:

- Do you understand the prediction model equation and what it implies about each variable?
- Do you understand odd and what an odds ratio is (since the paper mentions those frequently)?
- Do you understand what the results are and how they lead to the conclusions?
- Do you understand what each figure is trying to say?

If you don't understand some of the above (or something else), it's important to realise it and then figure out how to learn more about it until you understand it.

Part 2B – Advanced questions

1. What is the methodology of the study and what are the results? (In other words, how do the researchers go about testing the hypothesis, and what is the outcome?)

Note: This and the next question are essentially more detailed answers to question 4 from Part 1B.

After collecting the data from the app, a multivariate logistic regression was fit to the 15,638 UK users who had had COVID tests. The independent variables were age, sex, BMI and loss of smell or taste and the dependent variable was whether the test was positive or negative; it is not clear how many days before the test the symptoms had to occur to be counted. In the UK participants, the odds ratio for the loss of smell arising from this model was 6.40, with a 95% confidence interval of 5.96-6.87. A separate model trained on the US participants' data gave an odds ratio of 10.01 (95% CI: 8.23-12.16).

Further logistic regression models were trained including the other ten symptoms reported by participants ("fever, persistent cough, fatigue, shortness of breath, diarrhoea, delirium, skipped meals, abdominal pain, chest pain and hoarse voice") as independent variables. All symptoms were found to be associated with a positive test in the UK data, though not the US data, where only loss of smell or taste, fatigue and skipped meals were found to be significant. To identify the symptoms most strongly associated with COVID in the UK data models with varying combinations of parameters were trained on 80% of the data and tested on 20% of the data. The best predictive model was determined by a measure called the Akaike Information criterion, and included age, sex, loss of small or taste, severe or persistent cough, severe fatigue and skipped meals. The prediction model had sensitivity of 0.65 (CI 0.62-0.67) and selectivity of 0.78 (CI 0.76-0.80).

The predictive model was applied to the 805,753 UK and US app-users who had not been tested. The model predicted around 17.42% of those reporting some symptoms (14.45-20.39%) were positive, which was 5.36% of the total app users.

2. Identify the statistical methods used in the study and explain how they are applied to the data.

Logistic regression is used to compute odds ratios with confidence intervals for the factors by which each symptom increased the odds of having COVID. *P*-values were reported, implicitly for the null hypothesis that the odds ratio was one. It is not clear what method is used to generate the confidence intervals or the *p*-values. A correction for multiple testing is reported, but it is not clear exactly how this has been done.

Splitting the data into training and testing data was used when training and reporting the error on the predictive model. In addition, cross-validation was used in connection with computing the Area Under the Curve (AUC) metric – perhaps to obtain the confidence intervals. The sensitivity and selectivity were reported for the model, as well as the area under the curve, an overall measure of classifier performance. The bootstrap was used to obtain the 95% confidence intervals around the estimated number of infections. A measure called the Akaike information criterion was used to estimate the prediction error on unseen samples – essentially it is trying to give a measure of the goodness of fit, but penalised for larger number of parameters, which will tend to cause the model to over-fit (Wikipeda page on Akaike Information Criterion).

3. Does the paper use confidence intervals? If so, where, and do they make sense? If not, where would be good to use them?

Yes, the paper frequently uses confidence intervals (dubbed "CI"), primarily for odds rations, but also for some other metrics (AUC, sensitivity (SE), specificity (SP), positive predictive value (PPV) and negative predictive value (NPV)) in Figure 1. As stated in the answer to the last question, it's not clear how most of these were obtained, but the values themselves do make sense – most of the CI's are fairly tight, which is reasonable with the amount of data involved, and the CI for the US data (8.23–12.16) is wider than that for the UK data (5.96–6.87), which is again reasonable given that there were more than 10 times more UK participants than US participants.

4. Provide a critical discussion of the paper. Evaluate how strongly the data and analysis support the stated conclusions. Identify limitations of the study.

The paper (when it was published) was extremely important to global public health, as it suggested a way of improving detection of COVID-19, especially when testing facilities are limited. It also helped to identify that anosmia is a key symptom that suggests a COVID-19 infection, a finding which has influenced global public health advice. The paper is based on a large sample size, which will tend to lead to smaller confidence intervals than with a smaller dataset.

All the concerns raised in question 5 from part 1B are also valid criticisms.

5. Identify and explain the ethical issues investigated in the study, or connected with how the study was conducted. How well do the authors discuss these issues?

In the UK, the institutional ethics committee (King's College London) approved the app. Users provided consent for non-commercial use. In the US the "Partners Human Research Committee" approved the research. An informal consultation with TwinsUK members (a consortium in which the authors are involved) is also mentioned. The potential benefits of the research in terms of finding COVID cases and potentially saving lives are large. Privacy is an issue: there are potential harms to app users if their data

is leaked or compromised, so there is an ethical obligation on the researchers to maintain data integrity. The authors mention their competing interests, I.e. their role as consultants to the firm who developed the app.

6. Try to figure out what the ROC is (there is information about the ROC in Chapter 22 of the Lecture Notes) and what its corresponding results mean. In papers and articles, you may often come across terms or methods that you're not familiar with - trying to figure them out is an important skill.

The purpose of this question here is to give a specific example of something to look up during the process of reading the paper. ROC is discussed in chapter 22.3 of the lecture notes (go to the "Predictive accuracy" paragraph), but you may find it easier to find external explanations of it as well. The description of ROC in the lecture notes relies on sensitivity and selectivity, so you would have to know what those are as well (the paper mentions sensitivity, and Chapter 12.3 of the lecture notes define both).