Foundations of Data Science

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

Which Study did you choose?

- Influence of socioeconomic deprivation on interventions and outcomes for patients admitted with COVID-19 to critical care units in Scotland.

Advice:

- For each question, we suggest a length in terms of a number of paragraphs. Note that the average length of a sentence is 15-20 words, the average length of a paragraph has between 3 to 6 sentences.
- When you first read the studies, you will probably find that you don't understand several concepts in the papers. Don't worry; this is normal. Part of the skill of understanding a study is to identify parts you don't understand, and see what you can understand from the rest of the paper.
- "How to read a paper" by Michael Mitzenmacher gives very good advice on reading a paper *critically.*

Answer the following questions:

Section 1 The following questions relate to the scientific paper of your chosen study.

1a [10 points]. Describe the aims of the study and its stated contributions. Write up to one paragraph.

The aims of the study are to describe the impact of socioeconomic deprivation on 30-day mortality following critical care admission for COVID-19, and the impact of COVID-19 on critical care capacity in Scotland. This study claims importance for the planning of critical care services in a future wave to minimize deaths related to COVID-19. It contributes to the effort to optimize critical care units by suggesting a more targeted strategy to allocating additional resources instead of a per capita approach. The paper shows the temporal evolution of hospitality rates and ICU loads stratified by socioeconomic status. It demonstrates (i) higher peak demand for critical care , (ii) increased mortality and (iii) generally poorer outcomes for people living in socioeconomically deprived areas. Similarly, ICUs in those areas spent more time above bed capacity and were forced to expand more severely.

1b [15 points]. Describe the methods used to collect and process the data in the study. Be specific about what is being measured and how. Write two paragraphs.

This was an observational national cohort study using the data of 735 confirmed COVID-19 patients over the age of 16 admitted to critical care units in Scotland from 1/3/20 to 20/6/20. Data was collected and linked from the Scottish Morbidity Record 01, Electronic Communication of Surveillance in Scotland database, National Records of Scotland death records, and the Scottish Intensive Care Society Audit Group database, using Community Health Index numbers. Linking the databases allowed the researchers to determine previous health status, acute illness variables, demographic variables, exposure to socioeconomic deprivation, and whether the patient passed away or not. A patient's previous health status was determined by their number of emergency hospital admissions in the year before, and any comorbidities. The acute illness variables consist of the Acute Physiology Score, the PF ratio, time from hospital admission to critical care admission, and number of organ systems supported on the day of critical care admission. Demographics included age, sex, and ethnicity. A patient's exposure to socioeconomic deprivation was determined using quintiles of the Scottish Index of Multiple Deprivation, which is an area based ranking based on income, employment, health, education, skills and training, housing, geographic access and crime.

As this study was fixed by the number of admissions, sample size calculations were not necessary/feasible. Missing data for ethnicity and Acute Physiology Score were accounted for by using an indicator variable to group missing entries.

1c [15 points]. Identify the statistical methods used in the study and explain how they are applied to the data. Write two paragraphs.

The descriptive statistical methods used in this study were the median and the IQR. These were used to show summary statistics in the data exploration stage, since, unlike the mean and range, the median and IQR are resistant to heavy skewness. They were used in Table 1, to gather baseline characteristics of each variable. The parametric inferential statistical method used in this study was multivariate logistic regression. The logistic regression model was applied to the data to make conclusions about the effect of each variable using the adjusted odds ratios/coefficients.

The authors used sequential modelling to observe the effects of the different groups of variables. First, a univariable model was used. Then a baseline multivariate model was applied, only adjusting for age, sex, and ethnicity. Lastly, another multivariate model was applied, adjusting for age, sex, ethnicity, preexisting health conditions and severity of illness. No survival models were necessary, because the 30-day follow ups were complete.

1d [30 points]. Provide a critical discussion of the paper. Evaluate how strongly the data and analysis support the stated conclusions. Identify limitations of the study. Write three to four paragraphs.

This study reaches conclusions that may only be weakly supported by the data. The researchers claim that those admitted to critical care units from socioeconomically deprived areas have a significantly higher 30-day mortality. If we take a look at Table 4, we see the odds ratios relative to the least deprived group, and their respective 95% confidence intervals and p-values (***α*** = 0.05). As the authors mention, the univariate model sees no effect of deprivation on mortality, as all of the 95% CIs of the odds ratio include 1 and have a p-value much greater than 0.05. The baseline multivariate model adjusting for age, sex, and ethnicity, also has 95 % CIs including 1 and p-values much greater than 0.05 for all deprivation classes except most deprived. The most deprived group has an odds ratio of 1.97, which makes it seem like there is a significant effect in this model. However, when we see the 95 % confidence interval of (1.13, 3.41) with a p-value of 0.016, we realize that this effect disappears when using 99% confidence and does not adjust for confounding health conditions. Now, after adjusting for more potential confounders in the second multivariate model, we see even less of an effect, which supports the conclusion that the previous model was not reliable. The 95 % confidence intervals for all groups except most deprived include 1 and have p-values much greater than 0.05, and additionally, the most deprived group's confidence interval (1.01, 3.15) is very wide and almost includes 1, with a p-value of 0.046, which is just barely under 0.05. Using any confidence level greater than 95 % would have rendered this effect non-significant. The marginally significant p-value of 0.046 leaves open the worry/possibility that the authors tried various percentiles (eg. quartiles, deciles…etc), in which case the p-values would have to be adjusted for multiple testing.

Since both models only see semi-significant effects of deprivation when only comparing least deprived with most deprived, the data do not fully support the authors' claims. It would have been informative if instead of using dummy variables to measure deprivation, they had simply used the SIMD ranking itself. Then we would at least be able to better see the linear effect of exposure to deprivation on outcome, instead of only focusing on the two extremes (most and least deprived).

One idea that is not discussed in the paper is that ICU capacity could play a role in the outcome. It seems plausible that if an ICU is at 165% capacity versus one that is at 136%, the first may simply have less resources for each patient (oxygen, doctors…). Since more deprived areas in Scotland clearly had higher peak capacity and over longer duration, this could possibly explain the seemingly insignificant results of the two models. The authors could have added a model which factors in ICU capacity, or included a scatterplot of number of days over capacity vs. deprivation. Another interesting observation which is not discussed by the authors is that health is a factor in the SIMD deprivation index. I assume the authors looked into this, because this would mean that higher deprivation would be correlated with health of a patient, and clearly the health of a patient plays a major role in their 30-day mortality.

The limitations acknowledged by the authors include the very small sample size, small outcome rate (14-18%), the unavailability of less severe comorbidities data(eg. respiratory diseases), and the fact that the SIMD measure is not an individual index but rather one based on small areas. These are all true, but the authors never mention the limitations of the confidence level they used, or discuss the very weak support for the conclusion. Additionally, any conclusion regarding absolute numbers, (number of deprived ICU admissions) assumes that the population size of each area is either close to constant or is uncorrelated with socioeconomic deprivation, which was not discussed at all.

1e [10 points]. Identify and explain the ethical issues connected with the study. Evaluate how well the authors discuss these issues. Write one or two paragraphs.

One ethical issue with this study is that it involved sensitive data of the deceased. The problem with this is that the dead had no opportunity to decline. The researchers are in no position to decide for the patients whether or not they take part in the study. Although the data is not publicly accessible, it "involved data on unconsented participants", deceased or not, which poses an ethical issue. The authors only mentioned that they have legal access to the data, but did not discuss the possibility that patients might not want to be involved in the study. It was also not mentioned if the data was anonymized, or if the authors could identify real people based on their sensitive data. Thus, although the authors state the non-consent and legality of the study, they do not consider the ethical issues with it well enough.

Another ethical issue is the suggestion that "a more targeted approach to additional resource should be considered" in the future. Although it sounds reasonable to suggest that we should consider a different approach to resource allocation, this is an opinion. It suggests that we can further minimize the number of deaths due to covid by allocating more resources to those with a higher death rate. This is a controversial subject, and the authors should either make clear that this is their opinion or simply state the facts. This was not discussed at all, and the authors made it seem like considering a targeted approach was something everyone would agree with.

Section 2. The following questions relate to the media report of your chosen study:

2a [5 points]. Summarise the report in your own words. Write up to three sentences.

Researchers from Glasgow and Edinburgh found that more patients from deprived areas in Scotland were much more likely to be admitted to critical care and die. Hospitals in more deprived areas were also more likely to have higher peak capacity, which could be explained by poor housing, public transport, and financial pressure. The hospitals and patients in deprived areas will need extra support as the pandemic continues, in order to mitigate socioeconomic inequality, as the mortality rates of the most deprived are "significantly higher".

2b [15 points]. Evaluate how accurately the report summarized the study. You could identify aspects of the study that were not included in the report and discuss how important it would be to include them to give a fair impression of the research. Write two paragraphs.

I think the report summarized the study terribly. There is so much detail that is left out, and the conclusion of the paper is not discussed at all. The report simply says that death rates were "significantly higher" in patients from the most deprived area, without mentioning any numbers. This wording is dangerous, as statistical significance pertains to the trustworthiness of an effect, rather than its magnitude. The report should discuss how the effect would be considered insignificant had the authors used any higher confidence level, and that after adjusting for confounding, the effect is almost negligible. The report also fails to mention any limitations of the study and does not discuss the conclusion in context of the data or potential confounding variables, which it should.

Another interesting remark is that the report mentions that the study used anonymized data. This was never said in the paper, and unless the journalist received the information externally, it may not be true. It is very important for this report to give a fair impression of the research, as the claims it makes may not fully be supported by the data. The media often ignores or inflates certain parts of studies to get more clicks. This comes at the cost of spreading false or exaggerated information which is believable because it is coming from trusted experts.