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 study aims to use demographic and clinical characteristics stratified by **socioeconomic deprivation** to describe their impact on 30-day mortality following admitting to critical care for COVID-19, and the capacity of critical care units in Scotland due to the COVID-19 pandemic. The study has provided support to previous literature on the association between socioeconomic status and the critical care outcomes for COVID-19 patients. It has given important insights to the planning of critical care services in future waves.

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.

The data is sourced from multiple databases and included acute hospital activity, all virology testing in Scotland, death records, all adult general intensive care (ICU), and high dependency unit (HDU) activity within Scotland. The researchers used a cohort study design which comprised all Scottish residents "aged ≥16 years admitted to general ICUs and HDUs in

Scotland" who are tested positive for SRAS-CoV-2 before or during admission. Only the first admission was included for patients with multiple admissions.

Data collected between March 1st, 2020 to June 20th, 2020 were processed to produce data including 30-day mortality, critical care discharge, length of critical care stay, and type/duration of organ support during critical care stay. The processed data are exposed primarily to socioeconomic deprivation defined using quintiles of the Scottish Index of Multiple Deprivation. Other demographic and clinical variables including age, sex, ethnicity, pre-existing health status before admission, and comorbidities, and acute illness variables comprised the Acute Physiology Score were also exposed to the data for the logistic regression model of multivariable association.

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

This study mainly used logistic regression to assess the univariable and multivariable associations between socioeconomic deprivation, several other factors, and 30-day mortality. Metrics includes generated Odds Ratios with 95% confidence interval and p-values are given to reject the null hypothesis that the odds ratio was one. For the univariable association, the overall survival rate exposed to socioeconomic deprivation, age, and ventilation status was presented via Kaplan-Meier plots. The multivariable regression was done two separate times with a sequential approach where age, sex, and ethnicity were added in the base model as confounders. Pre-existing health status and severity of illness on admission were included in the second regression model in addition to the first regression.

The exact method used to compute Odds Ratios and confidence intervals were not given. It is likely the odds ratio is generated via a logistic regression model fitting using the principle of maximum likelihood. Sample size calculation wasn't performed as the total admissions are fixed. The 95% confidence interval can be obtained by using bootstrapping the logistic coefficients. For multivariable logistic regression, sequential regression simultaneously fits multiple covariates at each stage. Considering the number of covariates is relatively small, naïve sequential regression may be used in the study.

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 is an important reflection on the current status of public health in Scotland, it pointed out socioeconomic deprivation has a significant impact on the survival rate of critical care patients and demand for critical care units. It had demonstrated the change required in critical care capacity to meet demand and provided insights to the planning of critical care service in future waves.

In this study, a complete national cohort study of critical care patients with COVID-19 was analyzed, the study is confident all patients receiving mechanical ventilation ins Scotland is captured. The use of national, linked databases allows for an accurate record of pre-existing health status of patients, and outcomes for all patients with no loss to follow up. A comprehensive collection of variables are considered while the data is sourced, both demographic and clinical variables are exposed to socioeconomic deprivation and included in the multivariable logistic regression. Additionally, the analysis of the data also produced convincing results: the univariable regression suggested the 30-day mortality was "non-significantly higher" in the most socioeconomically deprived quintile relative to the least deprived (42.6% vs 34.0%; OR 1.44, 95%CI 0.87, 2.39, p=0.157) but mortality was "significantly higher" in the first multiple regression (OR 1.97, 95%CI 1.13, 3.41, p=0.016) and sustained in the second multiple regression (OR 1.78, 95%CI 1.01, 3.15, p=0.046) after fully adjusting for other covariables. The analysis for the subgroup of mechanically ventilated patients had produced a similar result when fully adjusted (OR 2.23, 95%CI 1.10, 4.51, p=0.010). In regards to the demand for critical care units, ICU capacity data suggested while baseline ICU capacity was exceeded for 25 days nationally, ICUs serving more deprived areas spent 34 days in overcapacity and 14 days in less deprived areas. The consecutive overcapacity was also prolonged for longer in more deprived areas (33 days) than less deprived areas with a maximum of 7 consecutive days. The more deprived areas also experienced higher overcapacity peaks relative to baseline capacity at 136%, 123%, and 165% respectively for low, medium, and high deprivation groups.

The comprehensive datasets and results produced in the analysis both strongly support the stated conclusion of the study. The author concludes that there is a strong association between 30-day mortality and socioeconomic deprivation, and between critical care unit overcapacity, peak overcapacity, consecutive overcapacity, and socioeconomic deprivation. The authors therefore suggest "A per capita approach to expanding health care services may not be the best strategy to meet future demand. Given that critical care units serving socioeconomically deprived areas experienced a higher peak of demand for critical care and for a more prolonged period of time, a more targeted approach to additional resource should be considered."

However, there are also limitations to the findings of the study. The precision of estimates was affected by relatively small sample size (n=735) and outcome frequency. SIMD, the definition used for socioeconomic deprivation is an area-based instead of an individual-based indicator, which suggests the statistic inference may be limited on an individual level. There could also be selection bias of possibly representing residents of more deprived areas "'sicker' population" had they been subjected to better hospitals or ICU admissions, which could cause some differences between outcomes. Less severe comorbidities related to COVID-19 were not recorded in the study and this may affect the association between socioeconomic deprivation and mortality. Some of the patients were provided non-invasive respiratory support outside of their designated critical care area. A sizable proportion of ethnicity and APS data were missing and were included as missing variables in multiple logistic regressions, which may cause bias in the result association.

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.

The study has mentioned twice both in "2.5 Approval" and "Ethical/information governance approval" that the Scottish Intensive Care Society Audit Group (SICSAG) has approved the use of additional datasets in this study for work undertaken within Public Health Scotland. A Public

Health Scotland information governance review of related internal datasets has approved the access and use of the data for the study's purpose. For the access to patient data, only analysts working in Public Health Scotland can access linked data via a secured NHS network. Data of individuals were not presented in the report nor shared since it involved unconsented participants and it's clearly stated in the study's data sharing statement. These statements ensure the security of highly private data used in the study and its use is ethical and approved.

A transparency statement is included to proclaim there are no important aspects of the study omitted and all discrepancies have been explained. The study's declaration of competing interests also states there are no financial interests that could have skewed or manipulated this work. These statements show the legitimacy of the study and its results and distinguish this study from other studies with academic misconducts, for example, the false claim of MMR vaccine causing autism or the controversial Syngenta funded research on the effect of the herbicide atrazine on amphibians.

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

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

A study has suggested that COVID-19 patients in more socioeconomically deprived areas are more likely to be critically ill and die than less socioeconomically deprived areas. Researchers looked at the critical care admissions in Scotland between March and June 2020 and found more disadvantaged areas also had more ICU admissions, higher 30-day mortality, and were more likely to be over capacity, although these outcomes may also be associated with other factors like age, sex, poor housing, increased use of public transport and the financial pressure to continue working. This study has shown the most deprived communities and their hospitals will need extra support in both the short term and long term in future pandemic waves.

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.

The report summarized the study at a high level with no obvious errors. It correctly included the time period the data is sourced, the total population, and how socioeconomic deprivation is defined, but it missed multiple important factors of the association between socioeconomic deprivation and 30-day mortality. Over a third of participants had at least one comorbidity, and it has a strong correlation with 30-day mortality in the univariable regression model with one (OR 1.88, 95%CI 1.27, 2.77, p=0.001) and two or more comorbidities (OR 1.73, 95%CI 1.10, 2.74, p=0.018), as well as multivariable regression models with one comorbidity (OR 1.51, 95%CI 0.97, 2.33, p=0.065). This implies comorbidities also have a relatively strong association with 30-day mortality, yet pre-existing health status was not mentioned in the report.

Although a few other potential factors are mentioned, the impact of these factors on the association is not clarified, the drastic increase in mortality after adjusting for additional factors was not addressed. The limitations of the study were not reported either, socioeconomic deprivation is defined via SIMD which is an area-based indicator instead of individual-based thus the statistic inference is limited at an individual level. A greater proportion of ethnicity and APS data are missing for the multivariable logistic regression (91.1% of the population in the model are white), which may cause bias to the association.