nature research

Peer Review File

Manuscript Title: OpenSAFELY: factors associated with COVID-19 death in 17 million patients

Reviewer Comments & Author Rebuttals

Reviewer Reports on the Initial Version:

Referee #1 (Remarks to the Author):

This large cohort study provides valuable prognostic information for COVID-19 cases in UK hospitals. Its inclusion of 40% of the population is a major strength of the study. The authors perform a rigorous analysis that includes careful nonlinear modeling of age effect, rigorous sensitivity analyses for missing data, and specific investigation of purported smoking and hypertension effects that comprise the primary novel results. The results mostly agree with other reported studies reported in the literature. Transparent sharing of analysis scripts is commendable. The data quality is high, analysis methodology reasonable, and writing is clear.

Specific comments:
• The post-hoc analysis of smoking and hypertension effects are potentially illuminating, helping provide some insight into results that if simplistically interpreted could lead to invalid conclusions. While the discussion of causal and mediation analysis is helpful, this manuscript only looks at covariate adjustment and interaction, not using any propensity score analysis or other causal approaches to tease out the results in a more rigorous fashion. Such an approach may have strengthened the paper especially given these are the primary novel results obtained in this analysis.
• Is any data available on covid-19 treatments used and/or other drugs currently being taken by the hospitalized patients (e.g. and importantly, blood thinners). If available, these would add further insight and increase the impact of this cohort and study.
• One weakness of the analysis is that no attempt was made to stratify based on ICU or ventilator use, or to assess the risk of progression to these higher levels of care.
• Are there any data sharing plans? It would be nice if there was some way to appropriately share some of the data so that the broader covid research community can benefit.
• Given the fast-moving nature of this crisis and the fast-accruing data, there is a great deal of data that has matured since the 6 April date – coming up on 2 months ago. If there was a way to perform an updated analysis, this would also add to the impact of the paper.

Referee #2 (Remarks to the Author):

The authors should be commended for a fine study. They linked records in the UK NHS to investigate risk factors for SARS Cov 2 in over 17 million participants, with a resultant 5,651 confirmed cases. They were able to confirm some of the findings from smaller studies, and form the early Chinese studies, with some exceptions. And they report an ethnic differential that persists afer adjusting for SES factors. The study is very important of its timeliness and for the power they achieve from such a large study base.

The study is well-designed, and well-executed. The results justify the conclusions. There are a few concerns:

1. Ethnicity - it appears that "Asian" likely combines East Asia and Southwest (subcontinent) populations, who may be quite different in social and genetic factors.

2. Case ascertainment. Especially in the early part of the epidemic, what would be the effect of false negative tests?

3. Existing disease as co-morbid risk factors. The results do seem counterintuitive. Is it possible that some models are overadjusted?

4. Frontline workers is a concept rapidly changing. A paper in PlosONe lastr week showed that hospital workers were at quite low risk compared to others in contact with people- drivers, transit workers, etc. This concept reeds revision here. Liklihood of infection by occupations not usually wearing makes them much higher risk.

Referee #3 (Remarks to the Author):

This was a well-written and large study examining racial disparities in covid-19 in the UK. The strengths include the large dataset, the inclusion of multiple health risk factors, and the prospective study design. The weaknesses include the use of a clinical sample (only most severe cases), lack of contextual data (e.g. occupation and crowding).

Author Rebuttals to Initial Comments:

In response to the comments from the reviewers, we have undertaken a number of additional analyses. In particular, we have updated our analysis using the most recent data available. Our updated analysis includes more than 10,000 COVID-19 deaths in contrast to the previous version which included approximately 5,700 deaths. We have also undertaken a more detailed exploration of the results by ethnic group.

We respond point by point to the reviewer comments below.

Referees' comments:

Referee #1 (Remarks to the Author):

This large cohort study provides valuable prognostic information for COVID-19 cases in UK hospitals. Its inclusion of 40% of the population is a major strength of the study. The authors perform a rigorous analysis that includes careful nonlinear modeling of age effect, rigorous sensitivity analyses for missing data, and specific investigation of purported smoking and hypertension effects that comprise the primary novel results.

The results mostly agree with other reported studies reported in the literature. Transparent sharing of analysis scripts is commendable. The data quality is high, analysis methodology reasonable, and writing is clear.

Response

Thank you for these encouraging comments.

Specific comments:

· The post-hoc analysis of smoking and hypertension effects are potentially illuminating, helping provide some insight into results that if simplistically interpreted could lead to invalid conclusions. While the discussion of causal and mediation analysis is helpful, this manuscript only looks at covariate adjustment and interaction, not using any propensity score analysis or other causal approaches to tease out the results in a more rigorous fashion. Such an approach may have strengthened the paper especially given these are the primary novel results obtained in this analysis.

Response

Our study is the first large-scale population-based study in patients in England to show many of these associations, so we believe that our manuscript contains novel results above and beyond the smoking and hypertension associations. However, these two associations are certainly the most unexpected novel results.

As explained in the manuscript, the estimated hazard ratios from our models will not necessarily reflect causal effects; rather, our aim was to provide a detailed description of independent associations with COVID-19 mortality. For smoking, for example, the

inclusion of variables such as respiratory disease, which are likely on the causal pathway between smoking and COVID-19 death, means that the smoking hazard ratio cannot be interpreted as the causal effect of smoking on COVID-19 death. Despite this, the estimated hazard ratios for smoking and hypertension are somewhat unexpected, and therefore deserve further exploration. Our post-hoc analyses regarding these variables were intended to provide an explanation, evidenced by empirical data, about how and why the direction of these hazard ratios is different to what might be expected.

With regards to the analysis approach undertaken, we do not feel that propensity score analysis or other causal approaches would add to our manuscript. All causal analysis approaches that rely on adjustment to remove confounding, including both multivariable regression and propensity score analysis, depend on the same set of underlying assumptions. Systematic reviews and simulation studies have shown that regression and propensity score approaches give comparable estimates in practice (e.g. Shah, 2005). Further, we think that a multivariable regression approach is better suited to this analysis considering a broad range of risk factors; adopting a propensity score approach would necessitate different propensity score models for each risk factor of interest.

We stress that a full causal analysis of the effects of smoking and hypertension on COVID-19 death is beyond the scope of our manuscript. Each of these questions is deserving of its own focused study, with data and methodology specifically targeted to the exposure of interest. However, we agree with the reviewer that such analyses would be very valuable. We have therefore amended the text in the discussion stating the need for rigorous individual and specific causal analyses to further explore the relationships between smoking, hypertension and COVID-19 death, replacing the original text:

"Where a causal effect is desired, a carefully designed causal analysis should be undertaken focusing on the specific exposure of interest."

with the new text (Discussion, paragraph 2):

"Our study has highlighted a need for carefully designed causal analyses specifically focusing on the causal effect of smoking on COVID-19 death. Similarly, there is a need for analyses exploring the causal relationships underlying the associations observed between hypertension and COVID-19 death"

· Is any data available on covid-19 treatments used and/or other drugs currently being taken by the hospitalized patients (e.g. and importantly, blood thinners). If available, these would add further insight and increase the impact of this cohort and study.

Response

Clarifying effects of in-hospital treatments for infected symptomatic patients is indeed an important priority. Unfortunately, there is no central source of electronic hospital prescribing data in the UK so we were not able to pursue this. Other studies, collecting data directly on small numbers of hospitalised patients (e.g. the ISARIC study of hospitalised COVID-19 patients; Docherty, BMJ, 2020), would be better placed to tackle this question.

One weakness of the analysis is that no attempt was made to stratify based on ICU or ventilator use, or to assess the risk of progression to these higher levels of care.

Response

We thank the reviewer for raising this point. Risk of progression through levels of care is indeed an interesting and important question. We have not attempted to tease out the different patient pathways, in part due to incomplete data on the full patient journey. We are not aware of any databases containing data on all the steps in that journey at present.

We are separately exploring factors associated with the risk of being admitted to an intensive care unit with COVID-19 in our study population. We decided not to cover this in the present paper because admission to ICU, unlike death, is a subjective decision made partly due to expectations about subsequent survival, and affected by other external factors including the availability of ICU beds. For example, it is likely that very elderly or frail patients will have a relatively low "risk" of admission to ICU, where this is thought unlikely to provide substantial benefit. Such effects, reflecting decisions around care and service availability as well as underlying risk, would not be readily generalisable internationally, and would complicate interpretation. We therefore suggest that adding these results to our current manuscript would detract from the clear message in the current presentation, with our single outcome of COVID-19 death.

· Are there any data sharing plans? It would be nice if there was some way to appropriately share some of the data so that the broader covid research community can benefit.

Response

We have shared all underlying codelists and code for data management and analysis, to support detailed review of our work, and to facilitate re-analysis and reproducibility by any other groups with access to similar data. Our analysis runs against the full primary care health record of 17 million patients. Although the name, address, and birth date are removed, these records are extremely detailed. They therefore contain a very large volume of confidential medical information, and present a very substantial risk of re-identification. They therefore cannot be shared. While the OpenSAFELY platform was produced on a very short timeline to deliver urgent analyses during the global COVID-19 emergency we are also working towards a model of expanded access for the wider research community, with appropriate permissions.

· Given the fast-moving nature of this crisis and the fast-accruing data, there is a great deal of data that has matured since the 6 April date – coming up on 2 months ago. If there was a way to perform an updated analysis, this would also add to the impact of the paper.

Response

We agree with the reviewer that, given the rapid changes occurring, an updated analysis is warranted. We have therefore updated our analysis with the most recent data available to us. Our revised results include more than 10,000 COVID-19 deaths in contrast to the previous version which included ~5,700 deaths.

The most up-to-date data available to us on deaths now come from Office of National Statistics (ONS) death registrations, rather than inpatient hospital death notifications data (CPNS), so we have used the ONS data in the revised version. This improves our ascertainment of COVID-19 deaths as out-of-hospital deaths are also now captured (see also response to referee #3). We have updated the manuscript accordingly; patterns of results remain very similar, though the trend by age is even stronger, likely reflecting a relatively high proportion of out-of-hospital deaths in the elderly that we did not capture in the original analysis.

We have removed the original results on inpatient hospital deaths (from CPNS data) from the revised manuscript as we do not think they add to the paper in the light of our updated analysis with the more complete ONS death registration data. However, we are open to adding these to the extended data if the Editors wish.

Referee #2 (Remarks to the Author):

The authors should be commended for a fine study. They linked records in the UK NHS to investigate risk factors for SARS Cov 2 in over 17 million participants, with a resultant 5,651 confirmed cases. They were able to confirm some of the findings from smaller studies, and form the early Chinese studies, with some exceptions. And they report an ethnic differential that persists after adjusting for SES factors. The study is very important of its timeliness and for the power they achieve from such a large study base.

The study is well-designed, and well-executed. The results justify the conclusions.

Response

We thank the reviewer for these comments.

There are a few concerns:

1. Ethnicity - it appears that "Asian" likely combines East Asia and Southwest (subcontinent) populations, who may be quite different in social and genetic factors.

Response

Our "Asian" category excludes East Asian populations (these are included in the "Other" category due to the comparatively small number of East Asians in the UK). We thank the reviewer for raising this issue. We have renamed the Asian group to "South Asian" in our revised manuscript for clarity.

In addition, we have undertaken an additional secondary analysis exploring a more granular grouping of ethnicity. We found that all groups except Irish and Other White had higher risks of COVID-19 death compared with White British. These positive associations were statistically significant in all cases except Chinese, which was a particularly small group. The strongest associations were seen in the Bangladeshi and African groups. We have added these results to the extended data (Table A1).

2. Case ascertainment. Especially in the early part of the epidemic, what would be the effect of false negative tests?

Response

It is possible that some deaths in people with COVID-19 were not identified as such, particularly given limited testing for COVID-19 in the UK at that time. Therefore, while we believe our outcome variable is likely to be highly specific (it is unlikely that large numbers

of deaths that are unrelated to COVID-19 were attributed to COVID-19), the sensitivity may be lower (i.e. there may be COVID-19 deaths that were not identified in our data). This may have led to slight underascertainment of outcomes but the effects on our results are likely to be minimal. Underascertainment of outcomes that is non-differential (i.e. unrelated to other covariates in the analysis) does not typically bias relative measures of effect (e.g. Ebrahim & Bowling, 2005). Furthermore, since the outcome (at this point in the pandemic) is relatively rare, the primary driver of associations will be the profile of risk factors in the identified cases; the risk profile of the small number of unidentified cases will have little statistical influence among the very large population of non-cases.

We have added a sentence to the manuscript acknowledging this limitation in the outcome measurement (Discussion, paragraph 4):

> "There may be people who died with COVID-19 who were misclassified as non-COVID-19 deaths in our analysis, particularly in the early stages of the pandemic. COVID-19 was rapidly recognised as an emerging cause of death in the UK so any such misclassification is likely to have reduced quickly as deaths accumulated. Furthermore, a degree of outcome underascertainment, providing unrelated to patient characteristics, should not have biased our hazard ratios. Due to the rarity of the outcome, the associations observed will be driven primarily by the profile of risk factors in the included cases."

3. Existing disease as co-morbid risk factors. The results do seem counterintuitive. Is it possible that some models are overadjusted?

Response

We assume that this comment relates to exposures such as smoking, where our model simultaneously includes variables that are likely to lie in the causal pathway from the exposures to the outcome. As we explain in our manuscript, our aim is not to undertake causal analyses for each exposure in our multivariable models, but rather to describe factors independently associated with COVID-19 mortality. As we have stated in the manuscript, the estimated associations, in particular for smoking, should not be interpreted as causal effects.

In order to provide some intuition about how and why the "overadjusted" associations arise, we have also provided post-hoc analyses exploring which aspects of the adjustment contribute most to the unexpected associations.

It is beyond the scope of our manuscript to undertake causal analyses for each variable in the model, which we think can only be adequately done in detailed papers focussing on individual risk factors with extensive consideration of variable-specific causal pathways, and employment of appropriate causal modelling methodology bespoke for each individual risk factor. We have added a comment to the manuscript about the need for such analyses (Discussion, paragraph 2):

> "In particular, our study has highlighted a need for carefully designed causal analyses specifically focusing on the causal effect of smoking on COVID-19 death. Similarly, there is a need for analyses exploring the causal relationships underlying the associations observed between hypertension and COVID-19 death."

4. Frontline workers is a concept rapidly changing. A paper in PlosONe last week showed that hospital workers were at quite low risk compared to others in contact with people-drivers, transit workers, etc. This concept reeds revision here. Likelihood of infection by occupations not usually wearing makes them much higher risk.

Response

We agree with the reviewer that "frontline worker" is an ambiguous phrase. We have rephrased the sentence that previously used this term, from

> "Other possible explanations for increased risk among BME groups relate to higher infection risk, including occupation with over-representation in "front-line" jobs with higher exposure to infection, or higher household density."

to the current wording (Discussion, Findings in Context):

"Other possible explanations for increased risk among BME groups relate to higher infection risk, including occupation with over-representation in jobs with higher exposure to infection and less access to personal protective equipment, or higher household density. "

Referee #3 (Remarks to the Author):

This was a well-written and large study examining racial disparities in covid-19 in the UK. The strengths include the large dataset, the inclusion of multiple health risk factors, and the prospective study design. The weaknesses include the use of a clinical sample (only most severe cases), lack of contextual data (e.g. occupation and crowding).

Response

We thank the reviewer for the positive appraisal, and for raising these points. Our original submission was restricted to hospitalised deaths for pragmatic reasons, as data on deaths among hospitalised patients were more up-to-date at the time of our initial analysis. It is true that there are likely to be differences in the COVID-19 positive patients

who end up in hospital and those who do not, and therefore differences in the profile of those who die in hospital versus outside of hospital. In this revision we have updated the analysis using national death registration data from the Office of National Statistics, which captures all COVID-19 related deaths rather than just those among hospitalised patients. In addition to addressing a less clinically-focused outcome, this also enables us to use more recent data, as mentioned in response to reviewer 1.

Associations between risk factors and the updated outcome, all COVID-19 deaths, show very similar results to our original submitted version, though the trend by age is even stronger, reflecting the fact that deaths among older people in the UK often occur outside of hospital.

In regards to occupation and crowding this is indeed a limitation of our data. It would be hugely informative to have information on occupation. Unfortunately, no population-based study in the UK with sufficient numbers to explore COVID-19 outcomes has occupational data. We have added some text in the discussion to highlight this limitation (Discussion, Findings in Context):

> "Other possible explanations for increased risk among BME groups relate to higher infection risk, including occupation with over-representation in jobs with higher exposure to infection and less access to personal protective equipment, or higher household density. Addressing these questions will likely entail bespoke data collection among cases and controls**,** since central linkable sources of electronic data on these variables are not available."

References:

Docherty Annemarie B, Harrison Ewen M, Green Christopher A, Hardwick Hayley E, Pius Riinu, Norman Lisa et al. Features of 20 133 UK patients in hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: prospective observational cohort study BMJ 2020; 369 :m1985

Ebrahim S, Bowling A (eds). Handbook of health research methods: investigation, measurement and analysis. Open University Press. New York. 2005

Shah BR, Laupacis A, Hux JE, Austin PC. Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review. J Clin Epidemiol. 2005;58:550–559.

Reviewer Reports on the First Revision:

Referee #1 (Remarks to the Author):

Thank you for your detailed responses to the issues and questions I raised.

I agree that a detailed causal analysis is beyond the scope of this study, and am happy with the revised discussion that I believe conveys these thoughts.

Updated results indeed strengthen the paper.

I understand why the data cannot of course be shared, but in light of the recent Surgisphere controversy, anything the authors can provide as documentation of the data, with how many hospitals were included, the scope, and verifications, will be helpful to assure the scientific community of the reliability of the data set, and overcome any backlash or skepticism in the broader public about results from large observational cohorts like this.

Referee #2 (Remarks to the Author):

The authors have responded satisfactorily to my concerns. Thank you.

Referee #3 (Remarks to the Author):

The authors have rigorously addressed some of the critiques. However, the manuscript would greatly benefit from a more thorough discussion of the significant disparities that were identified and how this unequal burden of disease can be prevented and mitigated in the UK.

Marmot M, Friel S, Bell R, Houweling TA, Taylor S. Closing the gap in a generation: health equity through action on the social determinants of health. Lancet. 2008

Williams DR, Collins C. Racial residential segregation: a fundamental cause of racial disparities in health. Public Health Rep. 2001

Link BG, Phelan J. Social conditions as fundamental causes of disease. J Health Soc Behav. 1995

Author Rebuttals to First Revision:

1-Reviewer #1 highlights the need for more transparency about the dataset. While we understand that pseudomyzed data cannot be publicly shared, we would like you to include more detail on how the data is acquired, managed and checked for accuracy, both routinely during collection as well as for the study. In addition, a description of the access request process and any validation done to ensure the reliability of the data should be included.

We have added substantial extra detail to the Methods section and an additional Figure in the Extended data to explain how the data are acquired, processed and checked for accuracy in the course of routine direct clinical care. We have additionally added text to introduce the motivation for the design of our new data analysis platform in the Introduction section as this sheds light on the unprecedented scale, the privacy engineering, and the reliability of the underlying data.

Regarding access requests: we rapidly delivered the OpenSAFELY data analysis platform without prior funding to deliver timely analyses on urgent research questions in the context of a global health emergency. Now that the platform is established we are developing a formal data access request process in collaboration with NHS England, details of which will be published shortly on OpenSAFELY.org. We are moving swiftly to make this an accessible resource for additional research by other teams, alongside evaluation of our own work. We have added this information to the Data Availability statement.

Revised text, Introduction: "Patient care is typically managed through electronic health records (EHR) which are commonly used in research. However traditional approaches to EHR analysis rely on intermittent extracts of small samples of historic data. Evaluating a rapidly arising novel cause of death requires a new approach. We therefore set out to deliver a secure analytics platform inside the data centre of major electronic health records vendors, running across the full live linked pseudonymised electronic health records of a very large population of NHS patients, to determine factors associated with COVID-19 related death in England (referred to as "death" in text that follows)."

Revised text, Methods: "The dataset analysed with OpenSAFELY is based on 24 million currently registered patients (approximately 40% of the English population) from GP surgeries using the TPP SystmOne electronic health record system. SystmOne is a secure centralised EHR used in English clinical practice since 1998; it records data entered (in real time) by GPs and practice staff during routine primary care. The system is accredited under the NHS approved systems framework for General Practice. Data extracted from TPP SystmOne have previously been used in medical research, as part of the ResearchOne dataset.[29,30] From this EHR a pseudonymised dataset was created for OpenSAFELY consisting of 20 billion rows of structured data including characterising for example pseudonymised patients' diagnoses, medications, physiological parameters, and prior investigations [Extended Data Figure 2, Level 1]. All OpenSAFELY data processing took place on TPP's servers; external data providers securely transferred pseudoymised data (such as COVID-19 related death from ONS) for linkage to

OpenSAFELY [Extended Data Figure 2, Level 2]; study definitions developed in Python on GitHub were pulled into the OpenSAFELY infrastructure, and used to create a study dataset of one row per patient [Extended Data Figure 2, Level 3]. Statistical code was developed using synthetic data and used to analyse the study dataset; this included code to check data ranges, to check consistency of data columns, and to produce descriptive statistics for comparison with expected disease prevalences to ensure validity, as well as code to fit our analysis models. Only two authors (KB/AJW) accessed OpenSAFELY to run code; no pseudonymised patient-level data were ever removed from TPP infrastructure; only aggregated, anonymous, manually checked study results were released for publication [Extended Data Figure 2, Level 4], All code for data management and analysis is archived online (see Code Availability, below)"

Revised text, Data Availability: "All data were linked, stored and analysed securely within the OpenSAFELY platform https://opensafely.org/. All code is shared openly for review and re-use under MIT open license. Detailed pseudonymised patient data is potentially re-identifiable and therefore not shared. We rapidly delivered the OpenSAFELY data analysis platform without prior funding to deliver timely analyses on urgent research questions in the context of the global Covid-19 health emergency: now that the platform is established we are developing a formal process for external users to request access in collaboration with NHS England; details of this process will be published shortly on OpenSAFELY.org."

2-Reviewer #3 has stated that it would be interesting to include some discussion on how the unequal distribution of disease burden could be addressed.

We observed unequal burden of mortality by deprivation and ethnicity. We do not know any underlying biological mechanisms that could explain these observations. Social factors which we do not have available (such as occupation, access to protective equipment for workers) could influence infection and hence mortality may be an explanation but without further evidence we can only speculate on these factors and potential mitigation. For this reason and keeping in mind word limit constraints, we did not add extended discussion on this topic, but have added some text suggesting factors and policies that could be the subject of further work on addressing inequalities.

Revised text, Discussion: "We have demonstrated - for the first time - that only a small part of the substantially increased risks of COVID-19 related death among non-white groups and among people living in more deprived areas can be attributed to existing disease. Improved strategies to protect people in these groups are urgently needed.[26] These might include specific consideration of BME groups in shielding guidelines and work-place policies. Subsequent studies are needed to investigate the interplay of additional factors we were unable to explore, including employment, access to personal protective equipment and related risk of exposure to infection and household density."