## PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

## ARTICLE DETAILS

| | |
|---|---|
| **TITLE (PROVISIONAL)** | Persons Diagnosed with COVID-19 in England in the Clinical Practice Research Datalink (CPRD): A Cohort Description |
| **AUTHORS** | Andersen, Kathleen; McGrath, Leah; Reimbaeva, Maya; Mendes, Diana; Nguyen, Jennifer; Rai, Kiran; Tritton, Theo; Tsang, Carmen; Malhotra, Deepa; Yang, Jingyan |

## VERSION 1 – REVIEW

| | |
|---|---|
| **REVIEWER** | Williams, Richard<br>The University of Manchester |
| **REVIEW RETURNED** | 05-Jul-2023 |

| | |
|---|---|
| **GENERAL COMMENTS** | My main issue with this paper is around the objectives and the findings. The objectives are not all novel. The objective to define case definitions for COVID diagnosis and COVID vaccinations has already been done in the UK by this paper (COVID-19 trajectories among 57 million adults in England: a cohort study using electronic health records), and with a superset of the data in the CPRD. If there is novelty in this paper, then the other paper should be cited and the novelty in this paper clearly defined. Another objective is to create code sets for 3 definitions of "high risk patients". At least one of these has already been done (download from https://digital.nhs.uk/coronavirus/treatments/methodology/coding-classifications-used) so again this is not novel. Also the authors have only provided med and prod codes which are CPRD specific and so not helpful to other EHR researchers - whereas the definitions in the above paper, and other sites such as https://www.opencodelists.org/, https://clinicalcodes.rss.mhs.man.ac.uk/ and https://phenotypes.healthdatagateway.org/ would have code sets in a variety of terminologies including SNOMED and Read. In fact I would recommend uploading any code sets you have to one of these repositories. Researchers often find code sets from repositories like this, rather than speculatively searching for papers with supplementary materials.<br><br>The second objective is better, but it's quite specific, as it is seeing to what extend the cohort of COVID patients in the CPRD is representative of COVID patients in England. This is of use to other CPRD researchers, but not more widely than that, so the objectives, findings and conclusions need toning down to reflect that this is a niche finding rather than a generalisable finding.<br><br>My other comments<br><br>1. Use COVID-19 rather than COVID throughout<br>2. "Given the mandatory reporting of PCR tests, regardless of result, |

| | |
|---|---|
| | back to GP, the CPRD should have nearly complete capture of confirmed COVID cases in this pre-antigen test era" - I don't think it was mandatory. Also your citation to back this up is a link to a press release from EMIS saying they will do it, not that it was mandatory. That's fine, but please point out that CPRD Aurum is just EMIS practices so that you expect most tests to make it to the primary care record. It is also worth noting that the press release also says that "All tests where it is possible to identify the test recipient’s NHS number will be shared". I suspect the fact that you have observed fewer tests in the CPRD than expected is because there were many tests where a nhs number was not provided by the patient.<br>3. Do you know how quickly the U07.1 code was made available to clinical coders, and how quickly it became used routinely?<br>4. "we excluded persons who were admitted to the hospital on or before their primary care recorded date of COVID diagnosis". Can you elaborate? This makes it sound like someone with an admission in 1998 would be excluded but I suspect you're just talking about admitted to hospital with a primary diagnosis of U07.1.<br>5. "Hospitalizations for COVID were defined as persons admitted with a primary diagnosis of COVID (ICD-10 U07.1 “COVID-19”) within 12 weeks of the initial diagnosis recorded in primary care" I don't understand why a person with a primary care record and an admission with the U07.1 code would not be included if they didn't have a COVID diagnosis in their primary care data. I realise you only want people with a primary care record to analyse all the other factors (comorbidities, weight, smoking etc..), but why do they also need the COVID diagnosis. At a minimum it would be informative to know how many people were in this situation.<br>6. There are some places in the implementation of the high risk definitions that could do with more detail e.g. "and asthma requiring at least daily use of preventative and/or reliever medication" - you list the medication codes but what did you do with them? Was it anyone with at least one code - or did you try and work out based on prescription frequency how often they used them?<br>7. "captured 15% of COVID cases in a database that covers 24% of persons in England" I can't see where the 15% has come from. 2,257,907 (primary care cPRD covid cases) / 20,386,786 (all covid cases) = 11%. Also rather than your guess it's to do with CPRD not being representative enough, I think the fact that tests required a nhs number to be shared had a lot more to do with it.<br>8. "Notably, among hospitalized cases, no patients had completed a primary COVID vaccination series" - you do mention that it's possibly because not many people would have time to get two doses - "Therefore, the calendar period under study allowed for many persons to have had a COVID diagnosis in periods at which “full vaccination” was not achievable." - Given the hospital data only went up to 1st April 2021, I'd change that "many persons" to "most persons" or "almost all persons". This is a big limitation for this finding, and could probably just be removed because you don't have enough data.<br>9. I would consider using the RECORD checklist rather than the STROBE checklist as it is more applicable to this research. |

| | |
|---|---|
| **REVIEWER** | Radanliev, Petar<br>University of Oxford |
| **REVIEW RETURNED** | 07-Jul-2023 |

| | |
|---|---|
| **GENERAL COMMENTS** | Very interesting and timely article. I think it deserves publication and I am recommending accept with corrections. There are some issues that require your attention. I list these corrections below as feedback |

| | |
|---|---|
| | / comments, and I am looking forward to reading the updated version of this article.<br>-- The article is a bit short, I am not certain on the journal page limit, but if you have space, try to expand, with a focus on contribution. One way how to improve your contributions is to improve your review and comparison of existing literature and knowledge.<br>- I have finished reading the article and I didn't see any mention on the ethics of data privacy risk. You have done a really good job at reviewing so many articles, but not a single article on the ethics and risk. There are recent articles on this topic that reviews recent and relevant literature, for example, on the related topic of 'ethics of shared Covid-19 risks' - see: https://doi.org/10.1007/s12553-021-00565-3 and on the related topic of 'Ethics and Shared Responsibility in Health Policy' - see: https://doi.org/10.3390/su13158355 It would be interested to read your take on this area, maybe just a few sentences review and comparison of your work in relations to these recent studies in related topics.<br>-- You don't have conclusion chapter. You have a paragraph starting with 'in conclusion', and it's the last paragraph of your article, but it's not a conclusion chapter, its part of the discussion. If you think you have covered everything, that's OK, but just to mention that conclusion is the best chapter to outline your key findings and key conclusions. So, you should make use of this chapter to make your article more readable, and since most readers would focus a great deal of their attention on the conclusion, this section should make the key conclusions more visible (and hence more interesting). |

## VERSION 1 – AUTHOR RESPONSE

Response to Reviewer: 1 (Dr. Richard Williams, The University of Manchester)

1. Comments to the Author: My main issue with this paper is around the objectives and the findings. The objectives are not all novel. The objective to define case definitions for COVID diagnosis and COVID vaccinations has already been done in the UK by this paper (COVID-19 trajectories among 57 million adults in England: a cohort study using electronic health records), and with a superset of the data in the CPRD. If there is novelty in this paper, then the other paper should be cited and the novelty in this paper clearly defined.

*We thank the reviewer for highlighting the study by Thygesen et al. The case definitions for COVID diagnosis in "COVID-19 trajectories among 57 million adults in England: a cohort study using electronic health records" by Thygesen et al differ from those in our submission. In brief, there are 19 SNOMED Description ID's reported in Andersen et al that are encapsulated in 7 SNOMED Concept ID's reported in Thygesen et al (SNOMED structure allows for multiple descriptions within a given concept).*

*First, the code list in Andersen et al was built by winnowing the list of codes published by CPRD as related to vaccination, testing, diagnosis, vaccination, advice and possible findings (reference 10 in the manuscript: https://cprd.com/sites/default/files/2022-05/SARS-CoV-*

*2%20counts%20May2022.pdf ). The code list in Thygesen et al is described as "To identify COVID-19 from EHR spanning all health-care settings, we combined all relevant COVID-19 events—ie, diagnosis codes in primary or secondary care, SARS-CoV-2 laboratory testing, disease outcomes, and the provision of ventilatory support (within and outside of the ICU)", without further specification of the process to derive the list.*

*Second, a side-by-side comparison of Thygesen et al and Andersen et al lists reveal important differences in case definition. The Lancet paper includes immunoglobulin A, G and M, as well as total immunoglobulin, findings. We did not consider these a current and confirmed infection, as immunoglobulin assays (also referred to as "antibody tests" on page 6 in the manuscript) measure resolution of a prior infection not a current infection. Further, we did not include codes that represent sequelae of infection, such as "Pneumonia caused by severe acute respiratory syndrome" or "Myocarditis caused by acute severe respiratory syndrome coronavirus 2" as these may occur on a later date than the index diagnosis. Finally, we did not include a code that indicated a test had been performed without a result, such as "Coronavirus ribonucleic aciddetection assay (observable entity)", as observable entities in SNOMED taxonomy indicate the test was performed, not necessarily the presence of coronavirus ribonucleic acid in the performed assay.*

*In methods, under the subheading COVID Case Definition, the first sentence is "With each monthly data release, CPRD publishes feasibility counts for SARS-CoV-2 related codes in CPRD primary care data with corresponding code lists", and the end of the paragraph states "We defined a current and confirmed COVID episode as a diagnosis code, positive PCR, or antigen test. We did not include COVID vaccination, antibody tests, possible cases, exposure to COVID or post-COVID clinic referral codes in the COVID case definition." We believe this is a sufficient description of the differences between the code lists.*

*We have added the following discussion of the differences between Andersen and Thygesen in the discussion section: "This manuscript reports results from a case definition of confirmed and current infection. We did not include codes for immunoglobulin titers, as measurable antibodies indicate a resolved infection rather than date of onset. We did not include codes indicating a sequela of prior infection, as these most often occur on a later date than index diagnosis. We did not include codes indicating a test without a result, as people with a negative test result should not be included in a COVID-19 case definition. Our results therefore identified fewer cases, although with greater specificity, than other studies in published literature that allow for such heterogeneity".*

2. Another objective is to create code sets for 3 definitions of "high risk patients". At least one of these has already been done (download from https://digital.nhs.uk/coronavirus/treatments/methodology/coding-classifications-used) so again this is not novel. Also the authors have only provided med and prod codes which are CPRD specific and so not helpful to other EHR researchers - whereas the definitions in the above paper, and other sites such as https://www.opencodelists.org/, https://clinicalcodes.rss.mhs.man.ac.uk/and https://phenotypes.healthdatagateway.org/ would have code sets in aariety of terminologies including SNOMED and Read. In fact I would recommend uploading any code sets you have to one of these repositories. Researchers often find code sets from repositories like this, rather than speculatively searching for papers with supplementary materials.

*We agree this is similar to our effort to define NHS Highest Risk Conditions, and have revised the manuscript as such: "After the completion of this work, NHS Digital published a code list for "Targeted Conditions", which includes each element in the NHS Highest Risk category. Among these, the NHS code list can be repurposed for 4 of the 14 conditions in PANORAMIC criteria and 3 of the 11 conditions in UKHSA*

*Clinical Risk. To our knowledge, we offer the first publication of code lists for capture of all elements in the PANORAMIC criteria as well as UKHSA Clinical Risk criteria for these high-risk definitions, which can now be readily used in datasets that contain CPRD medical and product, ICD-10 and OPCS Classification of Interventions and Procedures codes."*

*We are amenable to updating the code list files to include SNOMED Description ID codes pending acceptance and editor input. CPRD Aurum does not use Read codes.*

*Our organization's legal and data compliance teams are reviewing these code list repository sites for data privacy considerations and we are unable to provide a final answer as of the date of this submission.*

3. The second objective is better, but it's quite specific, as it is seeing to what extend the cohort of COVID patients in the CPRD is representative of COVID patients in England. This is of use to other CPRD researchers, but not more widely than that, so the objectives, findings and conclusions need toning down to reflect that this is a niche finding rather than a generalisable finding.

*We agree this is of value to CPRD researchers. This dataset has been used in over 3,000 peer-reviewed publications and represents one of the largest real world data assets in the United Kingdom. The intent of aim 2 is to empirically evaluate whether the sample of persons in CPRD can be used to draw conclusions about the population of England. We have revised the statement of the objective to: "Second, we aimed to evaluate these definitions in a sample of persons with COVID, using the CPRD, to assess whether this cohort's sociodemographic and clinical characteristics were generalizable to population-level COVID epidemiology in England." We added qualifiers of "persons with COVID in CPRD" in the results section.* However, we disagree with the *reviewer's conclusion "this is a niche finding rather than a generalisable finding", as the definition of generalizability is whether the findings in the study population can be used in a different sample or population. This study presents evidence of where the CPRD COVID cohort does and does not allow for generalizability of findings to the population of England using national data.*

4. Use COVID-19 rather than COVID throughout

*Revised throughout.*

5. "Given the mandatory reporting of PCR tests, regardless of result, back to GP, the CPRD should have nearly complete capture of confirmed COVID cases in this pre-antigen test era" - I don't think it was mandatory. Also your citation to back this up is a link to a press release from EMIS saying they will do it, not that it was mandatory. That's fine, but please point out that CPRD Aurum is just EMIS practices so that you expect most tests to make it to the primary care record. It is also worth noting that the press release also says that "All tests where it is possible to identify the test recipient's NHS number will be shared". I suspect the fact that you have observed fewer tests in the CPRD than expected is because there were many tests where a nhs number was not provided by the patient.

*Revised to "From August 2020 through March 2022, all tests booked via the National Health Service (NHS) website for polymerase chain reaction (PCR) tests for SARS-CoV-2, regardless of result, were reported to GP offices that use the EMIS electronic health record software". In the methods section, first sentence of "Study Setting and Population", there is mention that CPRD Aurum uses EMIS.*

6. Do you know how quickly the U07.1 code was made available to clinical coders, and how quickly it became used routinely?

   *We see HES records of hospitalization in our data with a U07.1 code as early as March 1, 2020. The Office of National Statistics has published death counts due to COVID-19 and begin their enumeration on March 1, 2020 as well: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/bulletins/deathsinvolvingcovid19englandandwales/deathsoccurringinjune2020 .*

7. "we excluded persons who were admitted to the hospital on or before their primary care recorded date of COVID diagnosis". Can you elaborate? This makes it sound like someone with an admission in 1998 would be excluded but I suspect you're just talking about admitted to hospital with a primary diagnosis of U07.1.

   *For example, if someone was admitted to the hospital on May 1 and their GP recorded date of COVID diagnosis as May 5, we do not know their true date of clinical diagnosis. Given potential lags in reporting to the GP, the person might have tested positive on May 1 but their GP was advised several days later. We have revised the methods section to say "Fourth, we excluded persons who were admitted to the hospital with a primary diagnosis of U07.1 on or before their primary care recorded date of COVID-19 diagnosis"*

8. "Hospitalizations for COVID were defined as persons admitted with a primary diagnosis of COVID (ICD-10 U07.1 "COVID-19") within 12 weeks of the initial diagnosis recorded in primary care" I don't understand why a person with a primary care record and an admission with the U07.1 code would not be included if they didn't have a COVID diagnosis in their primary care data. I realise you only want people with a primary care record to analyse all the other factors (comorbidities, weight, smoking etc..), but why do they also need the COVID diagnosis. At a minimum it would be informative to know how many people were in this situation.

   *We do not have an estimate for this number. The process as per our data license agreement with CPRD to receive an HES dataset is to first pull all persons with a condition of interest (here, GP recorded COVID) and then request corresponding hospitalization records for that set of people.*

9. There are some places in the implementation of the high risk definitions that could do with more detail e.g. "and asthma requiring at least daily use of preventative and/or reliever medication" - you list the medication codes but what did you do with them? Was it anyone with at least one code - or did you try and work out based on prescription frequency how often they used them?

   *CPRD Aurum contains prescriptions written, not necessarily dispensed. Further, the majority of medication records are missing dose or quantity. While an important point, we were unable to ensure daily sure of medications. The full set of codes used were reviewed by practicing physicians in the UK, as well as pharmacoepidemiologists trained in medication measurement using secondary data sources.*

10. "captured 15% of COVID cases in a database that covers 24% of persons in England" I can't see where the 15% has come from. 2,257,907 (primary care cPRD covid cases) / 20,386,786

(all covid cases) = 11%. Also rather than your guess it's to do with CPRD not being representative enough, I think the fact that tests required a nhs number to be shared had a lot more to do with it.

*From Results, sentences 1 and 2: "From 1 August 2020 through 31 January 2022, the UK Government's Coronavirus dashboard reported 14,744,991 COVID-19 cases in England.[24] The final CPRD cohort contained 2,271,072 persons diagnosed with COVID-19, regardless of care setting, in England in the same time period* ***(eTable 6)."*** *(2,271,072 / 14,744,991 = 15%).*

*The count of 20,286,786 is described in the footnote of the table where presented as "For all COVID cases, coronavirus.data.gov.uk no longer publishes case counts by region at the day level. It was not possible to limit to the study period of August 1, 2020 – January 31, 2022, and therefore the results above reflect the figures as accessed on January 31, 2023 of the total number of cases since the start of the pandemic."*

*We have added "The requirement for an NHS number in order for results to be shared may explain some of this attrition as well" to the discussion section.*

11. "Notably, among hospitalized cases, no patients had completed a primary COVID vaccination series" - you do mention that it's possibly because not many people would have time to get two doses - "Therefore, the calendar period under study allowed for many persons to have had a COVID diagnosis in periods at which "full vaccination" was not achievable." - Given the hospital data only went up to 1st April 2021, I'd change that "many persons" to "most persons" or "almost all persons". This is a big limitation for this finding, and could probably just be removed because you don't have enough data.

    *Revised to "most persons".*

12. I would consider using the RECORD checklist rather than the STROBE checklist as it is more applicable to this research.

    *Thank you for the suggestion. BMJ Open requests STROBE for reporting of observational studies in epidemiology: https://bmjopen.bmj.com/pages/authors.*

Reviewer: 2 (Dr. Petar Radanliev, University of Oxford)

1. Comments to the Author: Very interesting and timely article. I think it deserves publication and I am recommending accept with corrections. There are some issues that require your attention. I list these corrections below as feedback / comments, and I am looking forward to reading the updated version of this article.

    *Thank you for your time in reviewing, and for your kind words.*

2. The article is a bit short, I am not certain on the journal page limit, but if you have space, try to expand, with a focus on contribution. One way how to improve your contributions is to improve your review and comparison of existing literature and knowledge.

*The editor commented that the piece was too long. We are willing to revise, pending editor's input.*

3. I have finished reading the article and I didn't see any mention on the ethics of data privacy risk. You have done a really good job at reviewing so many articles, but not a single article on the ethics and risk. There are recent articles on this topic that reviews recent and relevant literature, for example, on the related topic of 'ethics of shared Covid-19 risks' - see: https://doi.org/10.1007/s12553-021-00565-3 and on the related topic of 'Ethics and Shared Responsibility in Health Policy' - see: https://doi.org/10.3390/su13158355 It would be interested to read your take on this area, maybe just a few sentences review and comparison of your work in relations to these recent studies in related topics.

*This is a very interesting topic and we appreciate the reviewers viewpoint. CPRD conducts a rigorous data privacy review process. We've included this in the Methods section (last paragraph of methods section). As we've followed standard practice for using de-identified data, we feel further discussion of data privacy ethics to be out of scope for this paper.*

4. You don't have conclusion chapter. You have a paragraph starting with 'in conclusion', and it's the last paragraph of your article, but it's not a conclusion chapter, its part of the discussion. If you think you have covered everything, that's OK, but just to mention that conclusion is the best chapter to outline your key findings and key conclusions. So, you should make use of this chapter to make your article more readable, and since most readers would focus a great deal of their attention on the conclusion, this section should make the key conclusions more visible (and hence more interesting).

*We created a heading of "Conclusion", thank you.*

## VERSION 2 – REVIEW

| | |
|---|---|
| **REVIEWER** | Williams, Richard<br>The University of Manchester |
| **REVIEW RETURNED** | 15-Nov-2023 |

| | |
|---|---|
| **GENERAL COMMENTS** | Thank you for addressing my comments. |