DATA NOTE

REVISED Therapeutics Dataset from COVID-19 Medicine Delivery Units in England: an OpenSAFELY Data Report

[version 2; peer review: 2 approved with reservations]

Linda Nab[1], Amelia Green[1], Rose Higgins[1], Bang Zheng[2], Anna Schultze[2], John Tazare[2], Viyaasan Mahalingasivam[2], Peter Inglesby[1], Simon Davy[1], Rebecca Smith[1], Amir Mehrkar[1], Christopher Bates[3], Jonathan Cockburn[3], Michael Marks[4], Michael Brown[5], Milan Wiedemann[1], Alex Walker[1], Ian Douglas[2], Ben Goldacre[1], Brian MacKenna[1], Laurie Tomlinson[2], Helen Curtis[1]

[1]Bennett Institute for Applied Data Science, Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, England, OX2 6GG, UK
[2]London School of Hygiene and Tropical Medicine Faculty of Epidemiology and Population Health, London, England, UK
[3]TPP House, TPP, 129 Low Lane, Horsforth, Leeds, LS18 5PX, UK
[4]Division of Infection, University College London Hospitals NHS Foundation Trust, London, England, UK
[5]University College London Division of Infection and Immunity, London, England, UK

v2 **First published:** 07 Aug 2024, **9**:425
https://doi.org/10.12688/wellcomeopenres.22721.1
Latest published: 19 Aug 2025, **9**:425
https://doi.org/10.12688/wellcomeopenres.22721.2

Abstract

Background

Between December 2021 and June 2023, COVID-19 medicine delivery units (CMDUs) in England offered antiviral medicines and neutralising monoclonal antibodies (paxlovid, sotrovimab, molnupiravir, remdesivir, and casirivimab/imdevimab) to non-hospitalised individuals with COVID-19, identified at high risk of developing severe outcomes. In order to prescribe and supply medicines CMDUs were required to notify NHS England of every prescription via an electronic form. This data was supplied to OpenSAFELY, a secure analytics platform for electronic patient records, as the COVID-19 "Therapeutics" dataset. We aimed to explore the analytic potential of the dataset for research into the use and effectiveness of these therapeutics offered by CMDUs.

Methods

Working on behalf of NHS England, we assessed the content and data

Open Peer Review

Approval Status ? ?

	1	2
version 2 (revision) 19 Aug 2025		
version 1 07 Aug 2024	? view	? view

1. **Tanja Mueller**, University of Strathclyde, Glasgow, UK

2. **Mahmoud Shams**, Kafrelsheikh University, Kafr El-Shaikh, Egypt

Any reports and responses or comments on the article can be found at the end of the article.

quality of the COVID-19 Therapeutics dataset within OpenSAFELY. We focused on therapeutics provided in outpatient settings by CMDUs. We described for each field the: data format, completeness and summarised its content.

Results

The COVID-19 Therapeutics dataset contained 18 columns and 58,590 rows of data, for 54,435 distinct patient IDs (92.9%) treated in outpatient settings. The dataset was well-structured, with completeness of almost all fields of 100%. The dataset included details on the specific treatment received, date administered, high-risk group(s) to which the patient belonged and the region in which they were assessed.

Conclusion

The COVID-19 Therapeutics dataset is well-structured, complete, and is suitable for research. It is linked to other data sources in OpenSAFELY (e.g., primary care), enabling important research on the impact of treatment and health disparities.

Keywords
SARS-CoV-2, COVID-19, Covid Medicine Delivery Unit, antivirals, antibodies, electronic health records

Corresponding author: Helen Curtis (helen.curtis@phc.ox.ac.uk)

Author roles: Nab L: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Writing – Original Draft Preparation, Writing – Review & Editing; **Green A**: Conceptualization, Data Curation, Formal Analysis, Methodology, Validation, Writing – Original Draft Preparation, Writing – Review & Editing; **Higgins R**: Writing – Review & Editing; **Zheng B**: Writing – Review & Editing; **Schultze A**: Writing – Review & Editing; **Tazare J**: Writing – Review & Editing; **Mahalingasivam V**: Writing – Review & Editing; **Inglesby P**: Conceptualization, Data Curation, Resources, Software; **Davy S**: Conceptualization, Data Curation, Resources, Software; **Smith R**: Conceptualization, Data Curation, Formal Analysis, Resources, Software; **Mehrkar A**: Conceptualization, Data Curation, Funding Acquisition, Investigation, Methodology, Project Administration, Resources, Supervision; **Bates C**: Data Curation, Investigation, Resources, Software; **Cockburn J**: Conceptualization, Investigation, Software; **Marks M**: Writing – Review & Editing; **Brown M**: Writing – Review & Editing; **Wiedemann M**: Writing – Review & Editing; **Walker A**: Conceptualization, Data Curation, Funding Acquisition, Methodology, Writing – Review & Editing; **Douglas I**: Writing – Review & Editing; **Goldacre B**: Conceptualization, Funding Acquisition, Methodology, Resources, Writing – Review & Editing; **MacKenna B**: Conceptualization, Data Curation, Funding Acquisition, Methodology, Writing – Review & Editing; **Tomlinson L**: Writing – Review & Editing; **Curtis H**: Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project Administration, Resources, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: BG has received research funding from the Bennett Foundation, the Laura and John Arnold Foundation, the NHS National Institute for Health Research (NIHR), the NIHR School of Primary Care Research, NHS England, the NIHR Oxford Biomedical Research Centre, the Mohn-Westlake Foundation, NIHR Applied Research Collaboration Oxford and Thames Valley, the Wellcome Trust, the Good Thinking Foundation, Health Data Research UK, the Health Foundation, the World Health Organisation, UKRI MRC, Asthma UK, the British Lung Foundation, and the Longitudinal Health and Wellbeing strand of the National Core Studies programme; he also receives personal income from speaking and writing for lay audiences on the misuse of science. BMK is also employed by NHS England working on medicines policy and clinical lead for primary care medicines data. IJD has received unrestricted research grants and holds shares in GlaxoSmithKline (GSK). LAT is funded by an NIHR Research Professorship NIHR302405. JT is employed by LSHTM on a fellowship sponsored by an unrestricted GSK grant.

Grant information: The OpenSAFELY platform is principally funded by grants from NHS England [2023-2025], The Wellcome Trust (222097/Z/20/Z) [2020-2024], MRC (MR/V015737/1) [2020-2021]. Additional contributions to OpenSAFELY have been funded by grants from MRC via the National Core Study programme, Longitudinal Health and Wellbeing strand (MC_PC_20030, MC_PC_20059) [2020-2022] and the Data and Connectivity strand (MC_PC_20058) [2021-2022], NIHR and MRC via the CONVALESCENCE programme (COV-LT-0009, MC_PC_20051) [2021-2024], and NHS England via the Primary Care Medicines Analytics Unit [2021-2024]. The views expressed are those of the authors and not necessarily those of the NIHR, NHS England, UK Health Security Agency (UKHSA), the Department of Health and Social Care, or other funders. Funders had no role in the study design, collection, analysis, and interpretation of data; in the writing of the report; and in the decision to submit the article for publication.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Nab L, Green A, Higgins R *et al.* **Therapeutics Dataset from COVID-19 Medicine Delivery Units in England: an OpenSAFELY Data Report [version 2; peer review: 2 approved with reservations]** Wellcome Open Research 2025, **9**:425 https://doi.org/10.12688/wellcomeopenres.22721.2

First published: 07 Aug 2024, **9**:425 https://doi.org/10.12688/wellcomeopenres.22721.1

REVISED **Amendments from Version 1**

In response to reviewer feedback, we have made several updates to our manuscript, including:

- **Abstract:** We added the names of the COVID-19 treatments and suggested potential future uses of the data. We also shortened the description of OpenSAFELY and removed the sentence "The values were largely plausible."
- **Introduction:** We expanded the background by adding the names of the treatments delivered through COVID-19 Medicine Delivery Units (CMDUs) and clarified that "Each CMDU followed the national specification for selecting patients for treatment."
- **Methods:** We clarified the patient coverage of the dataset by specifying that "We access and analyse only the data for patients registered with TPP practices, i.e. approximately 40% of England's registered patients." We also added details about the availability of the "Date of symptom onset" and "risk cohort" information in the data.
- **Discussion:** We revised the terminology used for patient treatment status, changing "Treatment complete" to "Approved," to better reflect the data source.
- **Figures and Tables:** We updated Table 2 to include additional footnotes clarifying the date of symptom onset and risk cohort fields.

Any further responses from the reviewers can be found at the end of the article

Introduction

From December 16th 2021 to June 26th 2023, regional COVID-19 medicine delivery units (CMDUs) offered antiviral medicines, including oral antivirals and neutralising monoclonal antibodies (nMABs), to non-hospitalised people with COVID-19 across England. Available treatments were paxlovid, sotrovimab, molnupiravir, remdesivir, and casirivimab/imdevimab. Each CMDU followed the national specification for selecting patients for treatment[1]. Eligibility criteria varied slightly for each individual treatment, but broadly, those eligible were patients aged 12 and over with SARS-CoV-2 who were believed to be at increased risk of severe COVID-19 outcomes (e.g., solid organ transplant recipients and people with chronic kidney disease)[1]. CMDUs were required to notify NHS England, the national body for managing the NHS in England, of every prescription via an electronic form.

OpenSAFELY is a new secure analytics platform for electronic patient records built by our group on behalf of NHS England to deliver urgent academic and operational research during the COVID-19 pandemic[2]. In OpenSAFELY, pseudonymised primary care records are analysed in situ within the secure data centres of electronic health record software providers. TPP is one such provider, and OpenSAFELY-TPP covers about 40% of English general practices. Analyses run across all patients' full raw pseudonymised primary care records, with patient-level linkage to various other data sources.

To enable important research into the use, effectiveness and safety of antivirals and nMABs offered in outpatient settings, collated data from NHS England was linked in OpenSAFELY-TPP and made available to researchers as the COVID-19 Therapeutics dataset.

Here we set out to systematically assess the content and data quality of the COVID-19 Therapeutics dataset within OpenSAFELY-TPP, focussing on the subset of therapeutics offered in outpatient settings by CMDUs. We described for each field: data format, completeness, and summarised its content. This paper is intended to support all studies using the COVID-19 Therapeutics dataset in OpenSAFELY-TPP, to inform and increase transparency of research into the use and effectiveness of COVID-19 therapeutics.

Methods

Data source

The COVID-19 Therapeutics dataset was supplied by NHS England. Most medicines in the NHS are paid for through overall hospital contracts or "tariffs''. However, certain "high-cost" or specialised medicines are excluded from tariffs, including COVID-19 therapeutics. Detailed information about such treatments is entered in the CMDU through systems such as BlueTeq, and passed to the responsible commissioner to enable payment, producing a detailed dataset including patient details[3]. OpenSAFELY-TPP obtained the national COVID-19 Therapeutics dataset directly from NHSE who received it from BlueTeq. Data was supplied weekly from 27th Jan 2022 to 28th June 2023, in alignment with the end of the national "COVID-19 treatment services" (26th June)[4], after which, COVID-19 therapeutics were funded and managed as part of routine, decentralised NHS services. Forms held in the BlueTeq system but not approved, e.g., those not yet submitted to NHS England or where treatment was decided against, were not supplied to OpenSAFELY; included statuses are shown in Table 1.

In line with OpenSAFELY standards on privacy and security, the national COVID-19 Therapeutics dataset was linked to primary care records in the secure data warehouse of TPP, who make the Electronic Health Record software for >40% of GPs in England. OpenSAFELY is deployed inside their data infrastructure for secure analysis. This linkage was done using hashed NHS numbers. Linked data related to patients' clinical risk, treatment types and timings. The region of the CMDU was also included, but all fields pertaining to Trust, Trust Organisation Data Service (ODS) code and Clinical Commissioning Group (CCG) name were not linked. We access and analyse only the data only for patients registered with TPP practices, i.e. approximately 40% of England's registered patients.

The dataset included information for those who received treatment in either outpatient settings (i.e., those who were non-hospitalised and received treatment by CMDUs) or inpatient settings (i.e., while hospitalised). We focussed our investigation on outpatient settings, as our aim was to explore the analytic potential of the dataset for research into the use and effectiveness of these therapeutics offered by CMDUs.

Dataset validation

We assessed the dataset schema, data formats, completeness and range of values. We compared the date values in different

Table 1. Overview of BlueTeq Form statuses and corresponding use and meaning that were included in the COVID-19 Therapeutics dataset supplied to OpenSAFELY-TPP by NHS England.

BlueTeq Form statuses	Use and meaning
Approved	The request for treatment has been submitted by a COVID-19 medicine delivery unit (CMDU) and meets all the validation criteria in the form.
Treatment Not Started	The request for treatment has been most likely "Approved" after which a user at a CMDU has gone to the record and used the "End of Treatment" function and selected "Treatment not Started". Using this function, the date at which the treatment was not started is recorded. The user at the CMDU also enters a supporting text explaining the reason for not starting treatment (e.g. because the patient died).
Treatment Stopped	As with above. The user at the CMDU used the "End of Treatment" function and selected "Treatment Stopped". The date at which the treatment was stopped is recorded with a supporting text.
Treatment Complete	As with above. The user at the CMDU used the "End of treatment" function and selected "Treatment Complete". Of note, the "End of Treatment" function is rarely used when a treatment is complete. The status of a request for treatment is automatically moved to "Treatment Complete" if a "Continuation" of treatment is requested, typically used for chronic conditions where the previous request covered a given time period. This is not applicable for COVID-19 therapeutics currently.

date columns, and assessed how many patients had more than one record, and which columns differed in such cases. To minimise disclosure risks, row/patient counts were rounded to the nearest 5 and numbers below 7 suppressed[5]. Percentages were calculated after rounding, which may result in mismatches with totals.

Analysis was initially carried out in January 2022 when the first data was supplied. Here, we repeated the analysis to provide a complete overview of the final dataset.

Ethics approval

This study was approved by the Health Research Authority (Research Ethics Committee reference 20/LO/0651, 02/04/2020) and by the London School of Hygiene and Tropical Medicine Ethics Board (reference 21863, 02/04/2020).

Dataset validation

Full list of fields

When linked to OpenSAFELY-TPP, the last import of the COVID-19 Therapeutics dataset in June 2023 contained 18 columns and 110,140 rows. 53% of rows were for therapeutics prescribed in an outpatient setting (58,590/110,140), with the remaining 47% prescribed to inpatients, not analysed here. Table 2 provides a complete list of each column, with a brief description of the field type and specification, distinct values and missing values.

Key fields

The field CurrentStatus reflects the status of the BlueTeq form (Box 1) and was in 99.9% of rows 'Approved' (58,535/58,590). FormName contained 40 different values, reflecting the name and version of the patient registration form used to register the treatment. As we used the field COVID_indication to filter the dataset on non-hospitalised patients, the field holds the singular value 'non_hospitalised'. Diagnosis was equal to 'Covid-19' in all rows. Intervention was key for identifying the type of treatment. The five different values and number of occurrences in the dataset in descending order were: 'Paxlovid' (40.0%; 23,455/58,590), 'Sotrovimab' (36.3%; 21,290/58,590), 'Molnupiravir' (23.2%; 13,565/58,590), 'Remdesivir' (0.6%; 325/58,590), 'Casirivimab and imdevimab' (0.1%; 55/58,590). The interventions molnupiravir, sotrovimab and casirivimab/imdevimab had corresponding columns for the high-risk groups and date of symptom onset (columns starting 'MOL1', 'SOT02', and CASIM05' respectively). Similar fields were later created for paxlovid and remdesivir but not supplied to OpenSAFELY.

Date fields

There were two fields in datetime format related to when treatment took place: Received and TreatmentStartDate. The Received date was generated when the form was submitted, and ranged between 16/12/2021 and 26/6/2023 while TreatmentStartDate was entered by the clinician and could represent either a future planned start date or a past date at the time of form submission. For TreatmentStartDate, 0.07% of values (40/58,590) were before the date at which CMDUs were launched across England (16/12/2023) and 0.02% of values (10/58,590) after the last date the data was loaded (28/6/2023). The Received date was the same as the TreatmentStartDate in 45.7% of rows (26,750/55,590) and later than the TreatmentStartDate in 44.6% of rows (26,130/55,590, with a median difference of 14 days [Q1-Q3:5-40]) (Table 3).

A further three fields were supplied in date-like format but supplied in an unstructured format: MOL1_onset_of_symptoms, SOT02_onset_of_symptoms, and CASIM05_date_of_symptom_onset. Each of these reflect the date of COVID-19 symptom onset reported by the patient, corresponding to the particular intervention under consideration i.e., MOL1_onset_of_symptoms for molnupiravir, etcetera. We found that the dates were inconsistently formatted (e.g., dd.mm.yy or d/m/yy).

Table 2. Summary of the COVID-19 Therapeutics dataset in OpenSAFELY-TPP filtered on outpatients.

Column Name	Description	Type	Length	Nullable	Distinct Values	Missing Values
Patient_ID	Pseudonymised patient ID	bigint	8	No	54,435**	0 (0%)
AgeAtReceivedDate	Age	int	4	Yes	98	0 (0%)
Received	Date form submitted	datetime	8	Yes	558	0 (0%)
Intervention	Intervention/therapeutic name	varchar	1,000	Yes	5	0 (0%)
Diagnosis	Diagnosis	varchar	1,000	Yes	1	0 (0%)
CurrentStatus	Status of form/application	varchar	1,000	Yes	4	0 (0%)
FormName	Name of form	varchar	1,000	Yes	40	0 (0%)
TreatmentStartDate	Treatment start date (actual/planned)	datetime	8	Yes	607	1-7 (0%)
Region	Region of CMDU	varchar	1,000	Yes	7	0 (0%)
MOL1_onset_of_symptoms	Date of symptom onset (Molnupiravir)	varchar	1,000	Yes	946	45,030 (76.9%)
MOL1_high_risk_cohort	Risk cohort (Molnupiravir)	varchar	1,000	Yes	70	45,115 (77.0%)
SOT02_onset_of_symptoms	Date of symptom onset (Sotrovimab)	varchar	1,000	Yes	2,413	37,400 (63.8%)
SOT02_risk_cohorts	Risk cohort (Sotrovimab)	varchar	1,000	Yes	87	37,505 (64.0%)
CASIM05_date_of_symptom_onset	Date of symptom onset (Casirivimab/ imdevimab)	varchar	1,000	Yes	22	58,535 (99.9%)
CASIM05_risk_cohort	Risk cohort (Casirivimab/ imdevimab)	varchar	1,000	Yes	12	58,535 (99.9%)
COVID_indication	Treatment setting/ indication	varchar	1,000	Yes	1	0 (0%)
Count	Number of forms	int	4	Yes	4	0 (0%)
Der_LoadDate	Data load date	varchar	1,000	Yes	1	0 (0%)

The number of distinct values and missing values is shown for each column of data. Counts of missing values are rounded to the nearest 5 and small numbers shown as "1-7". * For a description of these dataset characteristics we refer to the open OpenSAFELY documentation and references available therein[6]. ** Number of distinct Patient_IDs have been redacted to the nearest five. Date of symptom onset and risk cohort information was only supplied to OpenSAFELY-TPP for sotrovimab, molnupiravir, and casirivimab/imdevimab; and these fields were only completed when they corresponded with the Intervention.

Table 3. Comparison of the values of 'Received' and 'TreatmentStartDate' fields of the COVID-19 Therapeutics dataset in OpenSAFELY-TPP.

Comparison	Number of Occurrences (%)	Median Difference in Days [ICR]
Received < TreatmentStartDate	5,710 (9.7%)	1 [1-2]
Received = TreatmentStartDate	26,750 (45.7%)	-
Received > TreatmentStartDate	26,130 (44.6%)	14 [5-40]
TreatmentStartDate is missing	1-7 (0%)	-

ICR = Interquartile Range [Quartile 1 - Quartile 3].

Other fields

Region appeared to be an automatically generated field based upon the location of the CMDU submitting the form, and corresponding to the seven NHS England regions: 'East of England', 'London', 'Midlands', 'North East and Yorkshire', 'North West', 'South East', and 'South West'. AgeAtReceivedDate ranged between 0-100 and was never missing: 0 may have been supplied when age was unknown. Both region and age would normally be derived from linked GP records in EHR analysis so we did not explore these fields further. Der_LoadDate was '28/6/2023' for all rows in the final import of the dataset in OpenSAFELY-TPP.

Three text fields (derived from tick-boxes) represented the high-risk group(s) to which the patient was considered to belong, corresponding to the particular intervention under consideration: MOL1_high_risk_cohort (molnupiravir), SOT02_risk_cohorts (sotrovimab) and CASIM05_risk_cohort (Casirivimab/imdevimab). Date of symptom onset and risk cohort information was only supplied to OpenSAFELY-TPP for these three drugs, as the corresponding information for the other available treatments was collected later. In total there were 16 distinct risk groups available for selection (Table 4). When more than one box was checked, risk groups were joined by the word 'and'. There were 70, 87 and 12 distinct combinations of risk groups, for molnupiravir, sotrovimab and casirivimab/imdevimab, respectively. Because these were selected from checkboxes, they were highly consistent, although the options and what they represented changed over time[1,7]. For example, there were two variations on "rare neurological conditions/diseases" (Table 4).

Table 4. Distinct risk groups in the COVID-19 Therapeutics dataset in OpenSAFELY-TPP.

Risk Group
Downs syndrome
HIV or AIDS
IMID*
haematologic malignancy
haematological diseases
immune deficiencies
liver disease
primary immune deficiencies
rare neurological conditions
rare neurological diseases
renal disease
sickle cell disease
solid cancer
solid organ recipients
stem cell transplant recipients
None

Note some terms represent wider eligible groups, e.g. "Downs syndrome" also includes other chromosomal disorders, and some criteria changed over time[1,6]. * *IMID = Immune-Mediated Inflammatory Disorder.*

Completeness

Completeness of most fields was 100% (Table 2). Exceptions were fields relating to symptom onset date and risk groups for each treatment (Molnupiravir, Sotrovimab and Casirivimab/imdevimab). By filtering the data by treatment (using the field Intervention), we verified that the corresponding symptom onset fields were always complete, except for 1-7/13,565 rows for molnupiravir; and the high-risk group fields were only occasionally missing for molnupiravir and sotrovimab (0.66% [90/13,565] and 0.49% [105/21,290] rows respectively) and always complete for Casirivimab/imdevimab.

Duplication

There were 54,435 distinct patient IDs present in the data (92.9% of the 58,590 rows in total). Of these, 92.9% appeared once (50,550/54,435), 6.7% twice (3,630/54,435), 0.4% three times (240/54,435), and <0.1% four times or more (15/54,435). Patients with multiple records most commonly had different values for Received, TreatmentStartDate, AgeAtReceivedDate, Intervention, or both Intervention and one of the treatment date fields (Table 5). The field Count, representing the number of forms submitted for a patient, was in 99.8% of rows 1 (58,490/58,590) and 2 or more in the remaining 100 rows.

Analysing the COVID-19 Therapeutics dataset within OpenSAFELY-TPP

The COVID-19 Therapeutics dataset covering antiviral and nMABs prescriptions from December 2021 to June 2023 was made available to researchers within the OpenSAFELY-TPP software framework, to inform responses to the COVID-19 pandemic. The provision of this data for researchers incorporated findings from the present work, for example, the date of symptom onset fields were excluded due to their unstructured format. We also removed identical duplicates, and processed the risk groups for ease of use (e.g., allowing all three columns to be queried together and splitting out those previously joined by "and" to more standard comma separation).

The guidance to query the COVID-19 Therapeutics dataset using ehrQL (Electronic Health Records Query Language) via OpenSAFELY-TPP is published online and available to all[6]. ehrQL is a query language custom built to retrieve records from the OpenSAFELY database. It was designed by the OpenSAFELY team specifically for use with EHR data but is portable to other settings[8]. Table 6 provides an example of the ehrQL code used to include information on COVID-19 Therapeutics prescriptions within an OpenSAFELY "dataset definition"; this code is used to define a cohort.

Strengths and limitations

The provision of the COVID-19 Therapeutics dataset to OpenSAFELY-TPP by NHS England allows researchers to conduct important research into these therapeutics' real-world

Table 5. Duplicate patients. Patients appearing more than once in the COVID-19 Therapeutics dataset in OpenSAFELY-TPP and fields that differed in each appearance.

Field	Patient Count	Patients with multiple records (N = 3,885)	Total distinct patients (N = 54,435)
Received	3705	95%	6.8%
TreatmentStartDate	3550	91%	6.5%
AgeAtReceivedDate	2165	56%	4.0%
Intervention	1815	47%	3.3%
Intervention_AND_Received	1710	44%	3.1%
Intervention_AND_ TreatmentStartDate	1670	43%	3.1%
Region	150	4%	0.3%
SOT02_risk_cohorts	145	4%	0.3%
MOL1_high_risk_cohort	90	2%	0.2%
none_of_these	40	1%	0.1%
CurrentStatus	35	1%	0.1%

Table 6. Example of ehrQL code used in an OpenSAFELY-TPP dataset definition to query the Therapeutics dataset.

```
prescribed_paxlovid = (
  covid_therapeutics.where(covid_therapeutics.intervention.is_in(["Paxlovid"]))
   .where(covid_therapeutics.covid_indication.is_in(["non_hospitalised"]))
  .where(
       covid_therapeutics.treatment_start_date.is_on_or_between(
         "2021-12-16", "2023-06-26"
       )
    )
  .sort_by(covid_therapeutics.treatment_start_date)
  .first_for_patient()
)
```

The example ehrQL (Electronic Health Records Query Language) code above flags the first prescriptions of all patients who were prescribed paxlovid in an outpatient setting between 16 December 2021 and 26 June 2023, in the COVID-19 Therapeutics dataset. Further guidance on querying the COVID-19 therapeutics dataset via an OpenSAFELY-TPP dataset definition can be found online.

use and effectiveness. We efficiently re-used a dataset collected for administrative purposes, not adding any further burden on CMDUs (or hospitals). The availability of the COVID-19 Therapeutics dataset in OpenSAFELY-TPP means that this information can be readily combined with data from primary and secondary care (e.g., hospital admissions).

The dataset was well-structured, with completeness of almost all fields of 100% and the data values had a high level of plausibility. The dataset included details on the specific treatment received, date administered, high-risk group(s) to which the patient belonged and the region in which they were assessed. The dataset is made available for all researchers in OpenSAFELY and linked to other data sources, enabling important research. To date, the incorporation of the COVID-19 Therapeutics' dataset in OpenSAFELY-TPP has been important for monitoring who has received these therapeutics[9], and assessing the comparative effectiveness[10–12] and effectiveness[13] of these therapeutics. This report can support future research of these therapeutics in OpenSAFELY-TPP, or elsewhere using the present dataset or similar data.

Some of our findings were suggestive of minor data quality issues. In 99.9% of records the treatment status was "Approved", and as such "Treatment complete" was very rarely used. This indicates that forms were completed at the time of treatment

decision and not updated (or resubmitted) when treatment was complete. In our investigation of the difference between date treatment was received and date of treatment start, we found that there is often a delay between a patient being assessed for treatment and the form being submitted. As new updates of the data were provided continuously to OpenSAFELY between January 2022 and June 2023, to allow for timely analyses, there may have been some delay in finding the correct number of people treated over time. While the nature of how the data was collected (i.e., using forms with pre specified values to select from) there were very few fields which were able to contain implausible values, some patients appeared in the data multiple times and it is not known whether multiple treatments were given (if different) or other reasons for multiple forms being submitted. We did have access to a field for a variable called Count which represents how many times patients have appeared in the data, but this field appeared to represent a count of duplicates which were aggregated prior to data being shared with OpenSAFELY. The fields representing the date of symptom onset were supplied in inconsistent format and could not be automatically converted to dates in the same format; these fields were therefore not made available for researchers to use, in line with OpenSAFELY policies. Symptom onset and high-risk group fields were not supplied to OpenSAFELY for a subset of interventions (Paxlovid, remdesivir). Validation was only possible for two fields (age and intervention) and that the validity of one of these (intervention) fields has already been discussed in detail in another paper and we did not seek to explore this any further within this paper[9].

It would be useful to have data on individuals assessed at CMDUs for whom treatment was not approved, e.g., those found to be ineligible, to better facilitate future analyses of comparative effects. Here we focused our investigation on therapeutics supplied in outpatient settings. It is also of interest to explore the potential use of the dataset in hospital settings, and we encourage and invite other research teams to explore this.

Ethics and consent

This study was approved by the Health Research Authority (Research Ethics Committee reference 20/LO/0651) and by the London School of Hygiene and Tropical Medicine Ethics Board (reference 21863).

Data sharing

Primary care records managed by the GP software provider, TPP were linked to the COVID-19 Therapeutics dataset through OpenSAFELY and were linked, stored and analysed securely using the OpenSAFELY platform, https://www.opensafely.org/, as part of the NHS England OpenSAFELY COVID-19 service. Data in OpenSAFELY include pseudonymised data such as coded diagnoses, medications and physiological parameters. No free text data are included. All code is shared openly for review and re-use under MIT open license at https://github.com/opensafely/covid-therapeutics-notebook. Detailed pseudonymised patient data is potentially re-identifiable and therefore not shared.

Information governance

NHS England is the data controller for OpenSAFELY-TPP; TPP is the data processor; all study authors using OpenSAFELY have the approval of NHS England[14]. This implementation of OpenSAFELY is hosted within the TPP environment which is accredited to the ISO 27001 information security standard and is NHS IG Toolkit compliant[15].

Patient data has been pseudonymised for analysis and linkage using industry standard cryptographic hashing techniques; all pseudonymised datasets transmitted for linkage onto OpenSAFELY are encrypted; access to the platform is via a virtual private network (VPN) connection, restricted to a small group of researchers; the researchers hold contracts with NHS England and only access the platform to initiate database queries and statistical models; all database activity is logged; only aggregate statistical outputs leave the platform environment following best practice for anonymisation of results such as statistical disclosure control for low cell counts[16].

The service adheres to the obligations of the UK General Data Protection Regulation (UK GDPR) and the Data Protection Act 2018. The service previously operated under notices initially issued in February 2020 by the Secretary of State under Regulation 3(4) of the Health Service (Control of Patient Information) Regulations 2002 (COPI Regulations), which required organisations to process confidential patient information for COVID-19 purposes; this set aside the requirement for patient consent[17]. As of 1 July 2023, the Secretary of State has requested that NHS England continue to operate the Service under the COVID-19 Directions 2020[18]. In some cases of data sharing, the common law duty of confidence is met using, for example, patient consent or support from the Health Research Authority Confidentiality Advisory Group[19].

Taken together, these provide the legal bases to link patient datasets using the service. GP practices, which provide access to the primary care data, are required to share relevant health information to support the public health response to the pandemic, and have been informed of how the service operates.

Data availability

Access to the underlying identifiable and potentially re-identifiable pseudonymised electronic health record data is tightly governed by various legislative and regulatory frameworks, and restricted by best practice. The data in OpenSAFELY is drawn from General Practice (GP) data across England where TPP is the data processor. TPP developers initiate an automated process to create pseudonymised records in the core OpenSAFELY database, which are copies of key structured data tables in the identifiable records. These pseudonymised records are linked onto key external data resources that have also been pseudonymised via SHA-512 one-way hashing of NHS numbers using a shared salt. Bennett Institute for Applied Data Science developers and PIs holding contracts with NHS England have access to the

OpenSAFELY pseudonymised data tables as needed to develop the OpenSAFELY tools. These tools in turn enable researchers with OpenSAFELY data access agreements to write and execute code for data management and data analysis without direct access to the underlying raw pseudonymised patient data, and to review the outputs of this code. All code for the full data management pipeline, from raw data to completed results for this analysis, and for the OpenSAFELY platform as a whole is available for review at https://github.com/OpenSAFELY.

The data management and analysis code for this paper was led by HC and contributed to by LN, PI, SD and RS.

Software availability statement

Source code available from: https://github.com/opensafely/covid-therapeutics-notebook.

License: Data was extracted using SQL and analysed in Python in line with the relevant OpenSAFELY access policy[20]

Administrative

Acknowledgements

We are very grateful for all the support received from the TPP Technical Operations team throughout this work, and for generous assistance from the information governance and database teams at NHS England and the NHS England Transformation Directorate.

Transparency statement

The lead authors affirm that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

Copyright / license for publication

References

1. **Highest-risk patients eligible for new COVID-19 treatments: a guide for patients.** In: GOV.UK. [cited 15 Jan 2024].
Reference Source

2. Williamson EJ, Walker AJ, Bhaskaran K, *et al.*: **Factors associated with COVID-19-related death using OpenSAFELY.** *Nature.* 2020; **584**(7821): 430–436.
PubMed Abstract | Publisher Full Text | Free Full Text

3. Rowan A, Bates C, Hulme W, *et al.*: **A comprehensive high cost drugs dataset from the NHS in England - an OpenSAFELY-TPP short data report [version 1; peer review: 3 approved].** *Wellcome Open Res.* 2021; **6**: 360.
PubMed Abstract | Publisher Full Text | Free Full Text

4. **Treatments for coronavirus (COVID-19).** In: *NHS Digital.* [cited 31 Jan 2024].
Reference Source

5. **Safe outputs and requesting release of files from the Level 4 server - OpenSAFELY documentation.** [cited 26 Jan 2024].
Reference Source

6. **tpp schema for COVID therapeutics - OpenSAFELY documentation.** [cited 21 Jun 2024].
Reference Source

7. **Defining the highest risk clinical subgroups upon community infection with SARS-CoV-2 when considering the use of neutralising monoclonal antibodies (nMABs) and antiviral drugs (updated March 2023).** In: GOV.UK. 2023; [cited 21 Jun 2024].
Reference Source

8. Nab L, Schaffer AL, Hulme W, *et al.*: **OpenSAFELY: a platform for analysing Electronic Health Records designed for reproducible research.** *Pharmacoepidemiol Drug Saf.* 2024; **33**(6): e5815.
PubMed Abstract | Publisher Full Text | Free Full Text

9. Green ACA, Curtis HJ, Higgins R, *et al.*: **Trends, variation, and clinical characteristics of recipients of antiviral drugs and neutralising monoclonal antibodies for Covid-19 in community settings: retrospective, descriptive cohort study of 23.4 million people in OpenSAFELY.** *BMJ Med.* 2023; **2**(1): e000276.
PubMed Abstract | Publisher Full Text | Free Full Text

10. Zheng B, Green ACA, Tazare J, *et al.*: **Comparative effectiveness of sotrovimab and molnupiravir for prevention of severe covid-19 outcomes in patients in the community: observational cohort study with the OpenSAFELY platform.** *BMJ.* 2022; **379**: e071932.
PubMed Abstract | Publisher Full Text | Free Full Text

11. Zheng B, Tazare J, Nab L, *et al.*: **Comparative effectiveness of nirmatrelvir/ritonavir versus sotrovimab and molnupiravir for preventing severe COVID-19 outcomes in non-hospitalised high-risk patients during Omicron waves: observational cohort study using the OpenSAFELY platform.** *Lancet Reg Health Eur.* 2023; **34**: 100741.
PubMed Abstract | Publisher Full Text | Free Full Text

12. Zheng B, Campbell J, Carr EJ, *et al.*: **Comparative effectiveness of sotrovimab and molnupiravir for preventing severe COVID-19 outcomes in patients on kidney replacement therapy: observational study using the OpenSAFELY-UKRR and SRR databases.** *Clin Kidney J.* 2023; **16**(11): 2048–2058.
PubMed Abstract | Publisher Full Text | Free Full Text

13. The OpenSAFELY collaborative, Tazare J, Nab L, *et al.*: **Effectiveness of sotrovimab and molnupiravir in community settings in England across the Omicron BA.1 and BA.2 sublineages: emulated target trials using the OpenSAFELY platform.** *medRxiv.* 2023; 2023.05.12.23289914.
Publisher Full Text

14. **The NHS England OpenSAFELY COVID-19 service - privacy notice.** In: *NHS Digital.* [cited 17 Jan 2024].
Reference Source

15. **Data security and protection toolkit.** In: *NHS Digital.* [cited 17 Jan 2024].
Reference Source

16. **ISB1523: anonymisation standard for publishing health and social care data.** In: *NHS Digital.* [cited 17 Jan 2024].
Reference Source

17. **[withdrawn] Coronavirus (COVID-19): notice under regulation 3(4) of the health service (Control of Patient Information), Regulations 2002 – general.** In: GOV.UK. [cited 17 Jan 2024].
Reference Source

18. **COVID-19 public Health directions 2020.** In: *NHS Digital.* [cited 17 Jan 2024].
Reference Source

19. **Confidentiality advisory group.** In: *Health Research Authority.* [cited 17 Jan 2024].
Reference Source

20. **Policy for OpenSAFELY access by platform developers - OpenSAFELY documentation.** [cited 26 Jan 2024].
https://docs.opensafely.org/developer-access-policy/

Open Peer Review

Current Peer Review Status: ? ?

Version 1

Reviewer Report 21 September 2024

https://doi.org/10.21956/wellcomeopenres.25023.r97037

?

Mahmoud Shams
[1] Kafrelsheikh University, Kafr El-Shaikh, Gharbia Governorate, Egypt
[2] Kafrelsheikh University, Kafr El-Shaikh, Gharbia Governorate, Egypt

Strengths:

1. **Clarity and Structure:** The abstract is well-organized and easy to follow. It clearly states the purpose, methods, results, and conclusion, making it straightforward for the reader to understand the research.

2. **Relevance of Topic:** The research is highly relevant given the ongoing interest in COVID-19 therapeutics. Exploring the potential of the dataset for further research adds value to public health efforts.

3. **Data Quality and Completeness:** The authors effectively communicate the high quality and completeness of the dataset, which is a significant strength. This is crucial for inspiring confidence in future research that may use this dataset.

4. **Data Availability:** The availability of the dataset to all researchers through OpenSAFELY is an important point that enhances transparency and reproducibility in research.

Areas for Improvement:

1. **Background:**

 - The introduction could briefly mention the specific types of antiviral medicines and monoclonal antibodies offered by the CMDUs. This would give a clearer picture of the scope of treatments involved.
 - The phrase “non-hospitalised individuals with COVID-19, identified at high risk” could benefit from elaboration. What defines “high risk”? Are there specific conditions or factors that categorize these patients?

2. **Methods:**

- The abstract mentions that the data quality and content of the dataset were assessed, but it would be helpful to provide more detail on how this was done. Were specific data quality metrics (e.g., missingness, validity checks) used? Even a brief mention of key methods could enhance transparency.
- If possible, the methodology could specify whether any additional analyses (e.g., basic statistical summaries, trends over time) were conducted on the dataset.

3. **Results:**

- The abstract indicates that 92.9% of patient IDs were treated in outpatient settings, but it would be useful to clarify whether the remaining 7.1% were excluded from the analysis or if they had different characteristics.
- The term “values were largely plausible” is somewhat vague. What specifically was checked for plausibility (e.g., consistency with clinical guidelines)? Providing a bit more insight here would strengthen the credibility of the results.

4. **Conclusion:**

- While the dataset is said to be "well-structured" and "complete," the conclusion could briefly mention what kinds of research questions or areas this dataset would be most suitable for (e.g., efficacy of specific treatments, health disparities).
- The availability of the dataset for researchers is a key strength, but it would be useful to mention any known limitations or areas where further data collection might be needed. For example, are there any biases or gaps (e.g., underrepresentation of certain demographics)?

Style and Readability:

- The abstract is clear, but it could benefit from a slightly more engaging tone, particularly in the introduction and conclusion, to capture readers’ attention.
- The phrase “values were largely plausible” could be replaced with more precise language, such as “data values were consistent with expected clinical outcomes.”

Overall Suggestions:

- Expand on how the CMDUs identified high-risk patients and the specific treatments offered.
- Add more detail on the methods used to assess data quality.
- Clarify whether the 7.1% of patient IDs were excluded or treated differently.
- Specify what kinds of research this dataset can support in the conclusion.

References

1. Hassan E, Shams M, Hikal N, Elmougy S: Detecting COVID-19 in chest CT images based on several pre-trained models. *Multimedia Tools and Applications*. 2024; **83** (24): 65267-65287 Publisher Full Text
2. Tarek Z, Shams MY, Towfek SK, Alkahtani HK, et al.: An Optimized Model Based on Deep Learning and Gated Recurrent Unit for COVID-19 Death Prediction.*Biomimetics (Basel)*. 2023; **8** (7). PubMed Abstract | Publisher Full Text
3. Hassan E, Shams M, Hikal N, Elmougy S: COVID-19 Diagnosis-Based Deep Learning Approaches for COVIDx Dataset: A Preliminary Survey. 2023. 107-122 Publisher Full Text
4. ElAraby ME, Elzeki OM, Shams MY, Mahmoud A, et al.: A novel Gray-Scale spatial exploitation learning Net for COVID-19 by crawling Internet resources.*Biomed Signal Process Control*. 2022; **73**:

103441 PubMed Abstract | Publisher Full Text
5. Abdel Samee N, M. El-Kenawy E, Atteia G, M. Jamjoom M, et al.: Metaheuristic Optimization Through Deep Learning Classification of燙OVID-19 in Chest X-Ray Images. *Computers, Materials & Continua*. 2022; **73** (2): 4193-4210 Publisher Full Text
6. Butler IAE, Butterfield T, Janda M, Gordon DM: Colony life history of the tropical arboreal ant, Cephalotes goniodontus De Andrade, 1999.*Insectes Soc*. 2024; **71** (3): 271-281 PubMed Abstract | Publisher Full Text
7. Shams MY, Elzeki OM, Abouelmagd LM, Hassanien AE, et al.: HANA: A Healthy Artificial Nutrition Analysis model during COVID-19 pandemic.*Comput Biol Med*. 2021; **135**: 104606 PubMed Abstract | Publisher Full Text

Is the rationale for creating the dataset(s) clearly described?
Partly

Are the protocols appropriate and is the work technically sound?
Partly

Are sufficient details of methods and materials provided to allow replication by others?
Partly

Are the datasets clearly presented in a useable and accessible format?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Computer Science, AI.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 11 September 2024

https://doi.org/10.21956/wellcomeopenres.25023.r92857

?

Tanja Mueller
[1] University of Strathclyde, Glasgow, Scotland, UK
[2] University of Strathclyde, Glasgow, Scotland, UK

The presented “data note” is a concise summary of the OpenSAFELY Therapeutics dataset, describing its content as well as pertinent aspects of data quality. Briefly, the therapeutics dataset comprises prescribing records on a number of medications – antiviral drugs and neutralising

monoclonal antibodies – used in an outpatient setting to prevent hospitalisations in vulnerable patients with COVID-19. These treatments were initiated in COVID-19 medicines delivery units (CMDUs) across England between December 2021 and June 2023.

The rationale for creating the therapeutics dataset has been highlighted, and the variables within the dataset are clearly described; the work underpinning the development of the dataset appears technically sound, and methods and processes in place to provide access to data, conduct analyses, and disseminate findings adhere to current standards, both technically and ethically. Data availability and quality seem acceptable for research purposes.

While the manuscript is informative, well-written, and easy to read overall, there are a few details that could be clarified to facilitate understanding of the broader context and avoid misinterpretations of data and/or outputs.

1. Did all CMDUs provide the same service, following the same protocols/guidelines, in a similar manner? Or were there any differences between service providers that may need to be taken into consideration when analysing data? And were these units available to all patients, across England (or were these only accessible to specific patients, or with any regional differences)?

2. If the OpenSAFELY data provider (TPP) only covers about 40% of English GP practices, what does this mean with regards to data availability and population coverage of the therapeutics dataset? Were the CMDUs attached to or independent of a GP practice? Were COVID-19 therapeutics prescription also collected through several data provides but only TPP data was collated within the OpenSAFELY data environment?

3. It would be helpful to clarify how exactly data was collected; it seems this was done through an electronic system – possibly at the time of prescribing – but not all data was collected in a standardised form, and also not always in a very timely manner? In addition, is there any information available on whether treatment was, indeed, given and/or completed, considering that the “end of treatment” field was not mandatory? This is likely not an issue for one-off treatments given at the clinic (sc, iv) but what about multi-day oral treatments, were there take-home options?

4. Was there a particular reason to not supply symptom onset/risk group fields for Paxlovid/remdesivir patients?

Additional minor comments:

- It would be helpful to include the names of the included medicines in both the abstract and the introduction.

- Presenting the share of unique IDs as a percentage of the overall number of records (54,435/58,590, 92.9%) is not necessarily the most useful of summaries; perhaps better to state what the other records indicate (e.g., subsequent treatment episodes?).

- A footnote in table 2, indicating the reason for the high level of missing values in the onset of symptoms/high risk group variables, would be beneficial (in addition to the explanation provided in the text).

Is the rationale for creating the dataset(s) clearly described?

Yes

Are the protocols appropriate and is the work technically sound?
Yes

Are sufficient details of methods and materials provided to allow replication by others?
Partly

Are the datasets clearly presented in a useable and accessible format?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Pharmacoepidemiology, Drug Utilisation Research, Health Services Research, real-world data

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.