



DATA NOTE

A comprehensive high cost drugs dataset from the NHS in England - An OpenSAFELY-TPP Short Data Report [version 1; peer review: 3 approved]

Anna Rowan¹, Chris Bates², William Hulme ¹, David Evans ¹, Simon Davy¹, Nicholas A Kennedy ^{3,4}, James Galloway⁵, Kathryn E Mansfield ⁶, Katie Bechman⁵, Julian Matthewman⁶, Mark Yates⁵, Jeremy Brown⁵, Anna Schultze ⁶, Sam Norton², Alex J. Walker¹, Caroline E. Morton¹, Krishnan Bhaskaran⁶, Christopher T. Rentsch ⁶, Elizabeth Williamson ⁶, Richard Croker ¹, Seb Bacon¹, George Hickman ¹, Tom Ward¹, Amelia Green¹, Louis Fisher¹, Helen J Curtis ¹, John Tazare ⁶, Rosalind M. Eggo⁶, Peter Inglesby¹, Jonathan Cockburn², Helen I. McDonald ⁶, Rohini Mathur ⁶, Angel YS Wong ⁶, Harriet Forbes⁶, John Parry ², Frank Hester², Sam Harper², Ian J Douglas⁶, Liam Smeeth ⁶, Laurie A Tomlinson ⁶, Charlie W Lees⁷, Stephen Evans ⁶, Catherine Smith ⁸, Sinéad M. Langan^{6,8}, Amir Mehkar¹, Brian MacKenna ¹, Ben Goldacre¹

¹The DataLab, Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, OX2 6GG, UK

²TPP, 129 Low Lane, Horsforth, Leeds, LS18 5PX, UK

³Department of Gastroenterology, Royal Devon & Exeter NHS Foundation Trust, Exeter, UK

⁴IBD Research Group, University of Exeter, Exeter, UK

⁵Centre of Rheumatic Diseases, King's College London, London, UK

⁶Electronic Health Records Research Group, Faculty of Epidemiology and Population Health, London School of Hygiene & Tropical Medicine, London, WC1E 7HT, UK

⁷Centre for Genomics and Experimental Medicine, University of Edinburgh, Edinburgh, UK

⁸St John's Institute of Dermatology, Guy's and St Thomas' NHS Foundation Trust, London, SE1 9RT, UK




v1 First published: 22 Dec 2021, 6:360
<https://doi.org/10.12688/wellcomeopenres.17360.1>
 Latest published: 22 Dec 2021, 6:360
<https://doi.org/10.12688/wellcomeopenres.17360.1>

Abstract

Background: At the outset of the COVID-19 pandemic, there was no routine comprehensive hospital medicines data from the UK available to researchers. These records can be important for many analyses including the effect of certain medicines on the risk of severe COVID-19 outcomes. With the approval of NHS England, we set out to obtain data on one specific group of medicines, "high-cost drugs" (HCD) which are typically specialist medicines for the management of long-

Open Peer Review

Approval Status 

	1	2	3
version 1			
22 Dec 2021	view	view	view

1. **Richard Day** , St Vincent's Hospital, Sydney, Australia

term conditions, prescribed by hospitals to patients. Additionally, we aimed to make these data available to all approved researchers in OpenSAFELY-TPP. This report is intended to support all studies carried out in OpenSAFELY-TPP, and those elsewhere, working with this dataset or similar data.

Methods: Working with the North East Commissioning Support Unit and NHS Digital, we arranged for collation of a single national HCD dataset to help inform responses to the COVID-19 pandemic. The dataset was developed from payment submissions from hospitals to commissioners.


Results: In the financial year (FY) 2018/19 there were 2.8 million submissions for 1.1 million unique patient IDs recorded in the HCD. The average number of submissions per patient over the year was 2.6. In FY 2019/20 there were 4.0 million submissions for 1.3 million unique patient IDs. The average number of submissions per patient over the year was 3.1. Of the 21 variables in the dataset, three are now available for analysis in OpenSAFELY-TPP: Financial year and month of drug being dispensed; drug name; and a description of the drug dispensed.


Conclusions: We have described the process for sourcing a national HCD dataset, making these data available for COVID-19-related analysis through OpenSAFELY-TPP and provided information on the variables included in the dataset, data coverage and an initial descriptive analysis.

Keywords

Medications, healthcare administration, biosimilars, OpenSAFELY

University of New South Wales, Sydney,
Australia

2. **Ruth H. Jack** , University of Nottingham,
Nottingham, UK

3. **Katie Page** , University of Technology
Sydney, Sydney, Australia

Lutfun Hossain, University of Technology
Sydney, Sydney, Australia

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

Corresponding authors: Brian MacKenna (brian.mackenna@phc.oc.ac.uk), Ben Goldacre (ben.goldacre@phc.ox.ac.uk)

Author roles: **Rowan A:** Data Curation, Formal Analysis, Methodology, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Bates C:** Data Curation, Software, Writing – Review & Editing; **Hulme W:** Software, Writing – Review & Editing; **Evans D:** Software, Writing – Review & Editing; **Davy S:** Software, Writing – Review & Editing; **A Kennedy N:** Data Curation, Writing – Original Draft Preparation; **Galloway J:** Data Curation, Writing – Review & Editing; **E Mansfield K:** Writing – Review & Editing; **Bechman K:** Writing – Review & Editing; **Matthewman J:** Writing – Review & Editing; **Yates M:** Writing – Review & Editing; **Brown J:** Writing – Review & Editing; **Schultze A:** Writing – Review & Editing; **Norton S:** Software, Writing – Review & Editing; **J. Walker A:** Data Curation, Software, Writing – Review & Editing; **E. Morton C:** Software, Writing – Review & Editing; **Bhaskaran K:** Writing – Review & Editing; **T. Rentsch C:** Writing – Review & Editing; **Williamson E:** Writing – Review & Editing; **Croker R:** Writing – Review & Editing; **Bacon S:** Data Curation, Software, Supervision, Writing – Review & Editing; **Hickman G:** Supervision, Writing – Review & Editing; **Ward T:** Software, Writing – Review & Editing; **Green A:** Writing – Review & Editing; **Fisher L:** Writing – Review & Editing; **J Curtis H:** Data Curation, Project Administration, Writing – Review & Editing; **Tazare J:** Writing – Review & Editing; **M. Eggo R:** Writing – Review & Editing; **Inglesby P:** Software, Writing – Review & Editing; **Cockburn J:** Data Curation, Software, Writing – Review & Editing; **I. McDonald H:** Writing – Review & Editing; **Mathur R:** Writing – Review & Editing; **YS Wong A:** Writing – Review & Editing; **Forbes H:** Writing – Review & Editing; **Parry J:** Writing – Review & Editing; **Hester F:** Resources, Writing – Review & Editing; **Harper S:** Software, Writing – Review & Editing; **J Douglas I:** Writing – Review & Editing; **Smeeth L:** Resources, Supervision, Writing – Review & Editing; **A Tomlinson L:** Writing – Review & Editing; **W Lees C:** Writing – Review & Editing; **Evans S:** Writing – Review & Editing; **Smith C:** Writing – Review & Editing; **M. Langan S:** Writing – Review & Editing; **Mehkar A:** Data Curation, Methodology, Project Administration, Writing – Review & Editing; **MacKenna B:** Conceptualization, Data Curation, Investigation, Methodology, Project Administration, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing; **Goldacre B:** Conceptualization, Methodology, Resources, Supervision, Writing – Review & Editing

Competing interests: All authors have completed the ICMJE uniform disclosure form and declare the following: BG has received research funding from Health Data Research UK (HDRUK), the Laura and John Arnold Foundation, Wellcome, the NIHR Oxford Biomedical Research Centre, the NHS National Institute for Health Research School of Primary Care Research, the Mohn-Westlake Foundation, the Good Thinking Foundation, the Health Foundation, and the World Health Organisation; he also receives personal income from speaking and writing for lay audiences on the misuse of science. IJD has received unrestricted research grants and holds shares in GlaxoSmithKline (GSK). CS received departmental research funding from AbbVie, Boehringer Ingelheim, GlaxoSmithKline, Leo, Pfizer, Novartis, Regeneron, SwedishOrphan Biovitrum AB, and Roche and is an investigator within consortia that have industry partners (see biomap.eu and psort.org.uk). JG has received honoraria from Abbvie, Amgen, Celgene, Chugai, Galapagos, Gilead, Janssen, Lilly, Novartis, Pfizer, Roche, Sobi, and UCB, and has research funding from Amgen, Aztra-Zeneca, Gilead, Janssen, Medicago, Novovax and Pfizer. MY has received honoraria from AbbVie and UCB. CWL has received honoraria from Abbvie, Celltrion, Ferring, Galapagos, Gilead, GSK, Iterative Scopes, Janssen, Pfizer and Takeda. TPP and NECS CSU provided technical expertise and data infrastructure centre pro bono in the context of a national emergency. The views expressed are those of the authors and not necessarily those of the NIHR, NHS England, Public Health England or the Department of Health and Social Care.

Grant information: This work was supported by Wellcome [205039; 222097 to BG; 107731 to KB; 201375 to RM]; the Medical Research Council [MR/V015737/1]. BG's work on better use of data in healthcare more broadly is currently funded in part by: NIHR Oxford Biomedical Research Centre, NIHR Applied Research Collaboration Oxford and Thames Valley, the Mohn-Westlake Foundation, NHS England, and the Health Foundation; all DataLab staff are supported by BG's grants on this work. LS reports grants from Wellcome, MRC, NIHR, UKRI, British Council, GSK, British Heart Foundation, and Diabetes UK outside this work. JB is funded by a studentship from GSK. AS is employed by LSHTM on a fellowship sponsored by GSK. KB holds a Sir Henry Dale fellowship jointly funded by Wellcome and the Royal Society. HIM is funded by the National Institute for Health Research (NIHR) Health Protection Research Unit in Immunisation, a partnership between Public Health England and LSHTM. AYSW holds a fellowship from BHF. EW holds grants from MRC. ID holds grants from NIHR and GSK. RM holds a Sir Henry Wellcome Fellowship funded by Wellcome. HF holds a UKRI fellowship. RME is funded by HDR-UK and the MRC. SML was supported by a Wellcome Senior Research Fellowship in Clinical Science (205039/Z/16/Z). SML was also supported by Health Data Research UK (Grant number: LOND1), which is funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation and Wellcome. SML is an investigator on the European Union Horizon 2020-funded BIOMAP Consortium. CS acknowledges support for this research from the NIHR Biomedical Research Centre (BRC) at King's College London and Guy's and St Thomas' NHS Foundation Trust and the Psoriasis Association. CWL is funded by a UKRI Future Leaders Fellowship.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2021 Rowan A *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: 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 Research 2021, 6:360 <https://doi.org/10.12688/wellcomeopenres.17360.1>

First published: 22 Dec 2021, 6:360 <https://doi.org/10.12688/wellcomeopenres.17360.1>

Introduction

Medicines data can help answer important questions about the patterns of use of medications, associated costs and possible risks and benefits of pharmaceutical treatments on patient outcomes, such as death from COVID-19. In 2019–20 the NHS spent £20.9bn on medicines in England¹. Of this total, medicines issued in hospitals accounted for 55.9% (£11.7bn) and medicines issued in primary care accounted for 43.5% (£9.1bn). Detailed records and data exist for medicines used in hospitals; however, at the outset of the pandemic this information was not made routinely available by the NHS^{2,3}.

Although the NHS is a single-payer healthcare system, it includes different internal payment mechanisms, and consequently different sources of medicines data. Briefly, medicines in the NHS in England are funded either centrally, by NHS England (NHSE) specialised commissioning, or locally by Clinical Commissioning Groups (CCGs). There are approximately 135 CCGs who fund all medicines prescribed in primary care, and the majority of medicines in secondary care such as those funded through overall hospital contracts or “tariffs”. However, a specific group of “high-cost” medicines, typically specialist medicines⁴, are excluded from tariffs and funding is provided directly either from NHSE specialised commissioning or a CCG, depending on the medicine, condition or service it is used for. As a consequence, detailed payment information, including patient details, is passed between a hospital and the responsible commissioner, resulting in detailed data that is held locally by each commissioner. To our knowledge this has never been collated in a single place and made available for routine analysis at national level.

OpenSAFELY-TPP is a new secure analytics platform for electronic patient records built by our group on behalf of NHSE to deliver urgent academic and operational research during the pandemic^{5,6}. Analyses run across all patients’ full raw pseudonymised primary care records in 40% of English general practices where TPP electronic health record (EHR) software is deployed, with patient-level linkage to various sources of secondary care data. Code and analysis are shared openly for inspection and re-use.

With the approval of NHS England, we set out to: obtain a source of hospital high-cost drug (HCD) data; make these data available in OpenSAFELY-TPP to support analysis of important questions related to COVID-19; better understand the information collected and available for analysis and generate descriptive outputs. This report is intended to support all researchers and studies carried out in OpenSAFELY-TPP, and those elsewhere working with the present dataset or similar data, to help inform the response to the COVID-19 pandemic.

Methods

Data source - obtaining a national high-cost drugs dataset

Hospitals in England supply medicines to patients either directly or through “homecare” providers who deliver medicines to a patient’s home. The majority of medicines are funded through overall hospital contracts, included in tariffs; however, for certain HCDs, hospitals are required to provide a submission for

each patient to the relevant commissioner, either NHSE or one of 135 local CCGs¹, in order to receive payment. The majority of submissions relate to a prescription of a HCD, although some submissions relate to associated services (i.e. home care delivery charges). There is a national list of the medicines that are funded by NHSE⁷ and locally agreed lists for each CCG. These patient-level submissions are processed by intermediate organisations, Commissioning Support Units (CSUs), to support financial payments and associated activities like summary reporting. To our knowledge there has been no single collation of the submissions data across NHSE and all 135 CCGs for these HCDs. To address this gap, we arranged for a single CSU, the North East Commissioning Support Unit (NECS), to collate all the data flows from their partner CSUs into a single comprehensive dataset. The data was collated by NECS in May 2020. To use the dataset in OpenSAFELY-TPP it was determined that NHS Digital must approve access, via the NHS Digital Data Access Request Service (DARs), which was granted in September 2020 and made available shortly thereafter. In line with OpenSAFELY-TPP standards on² privacy and security the HCD dataset was linked at individual patient level to primary care records in the secure data warehouse of TPP.³

Variable overview

The HCD dataset is a patient-level dataset and includes variables on patient characteristics, clinical indications and medicine prescribed (**Table 1**). NHSE have a standard data collection specification for each individual submission, the Drugs Patient Level Contract Monitoring Data Set; however each CCG is independent and has local versions of the collection.

The national specification for submissions is published on the NHS Data Model and Dictionary [website](#)⁸. A full list of the variables collected via submissions and the specification for each variable can be found on the website. Further information on the NECS data collation and standardisation processes can be found in the documentation on [Github](#)⁹.

A description of each variable in the OpenSAFELY-TPP HCD dataset is provided below. **Table 1** provides a brief overview of each variable and **Table 2** provides information on the completeness of the data collection.

Currently, three variables from the HCD dataset can be queried in an OpenSAFELY-TPP study: FinancialMonth, FinancialYear and DrugName. These variables are the ones most relevant to current research questions and add new information not available in other OpenSAFELY-TPP datasets. These are also some of the most complete variables in the dataset.

Analysing the high-cost drugs dataset within OpenSAFELY-TPP

The HCD dataset covering submissions from April 2018 to March 2020 was made available to researchers within the OpenSAFELY-TPP software framework, to inform responses to the COVID-19 pandemic.

¹ 135 CCGs in FY 2019/20

Table 1. Full list of variables included in the OpenSAFELY-TPP high-cost drugs dataset.

Variable name	Variable type	Specification details	Variable description
Patient_Id	n10	Mandatory where relevant	Pseudonymised patient identification, used to match dataset to other datasets within OpenSAFELY-TPP.
Financial Month	Max an2	Mandatory	Currently able to query in OpenSAFELY-TPP study definition Financial month the prescribed item was administered to patient. 1 = April; 12 = March
Financial Year	an6	Mandatory	Currently able to query in OpenSAFELY-TPP study definition Financial year the prescribed item was administered to patient. FY 2018/19 = 201819
PersonAge	n	Derived	Age of patient when prescribed item was administered to patient. Some submissions included age at intervention. Where missing this variable was derived using clinical intervention date and date of birth.
Person Gender	an1	Mandatory where relevant	Gender as stated by the patient. 1 = Male 2 = Female 9 = Indeterminate (unable to be classified as either male or female)
Activity Treatment Function Code	an3	Mandatory where relevant	Code to describe the clinical area that prescribing is taking place in, based on main speciality. Full list of codes found online ¹⁰ .
Therapeutic Indication Code	Min an6 Max an20	Mandatory where relevant	Should be a SNOMED CT Code but looks like input varies based on organisation collecting the data. Code used to identify the reason for administering drug to the patient.
HighCost TariffExcluded DrugCode	Min an6 Max an20	Optional	Should be a SNOMED CT Code but looks like input varies based on organisation collecting the data. This should be the dm+d description of medicine administered to patient. Only populated when the provider has a dm+d enabled system.
DrugName	Max an255	Mandatory where relevant	Currently able to query in OpenSAFELY-TPP study definition Input standardisation is at a provider level rather than a national level - non-standardised text input. The name of the prescribed item. Where possible this should be the SNOMED CT name. For drugs not listed in dm+d, this must be the valid name in UPPER CASE.
RouteOf Administration	Min an6 Max an20	Mandatory where relevant	Should be a SNOMED CT code but looks like input varies based on organisation collecting the data. To be populated by providers with an e-prescribing system.
DrugStrength	Max an100	Mandatory where relevant	The amount of ingredient substance in the prescribed item.
DrugVolume	Max an100	Mandatory where relevant	The volume of the drug administered to a patient when given in liquid form.
DrugPackSize	Max an100	Optional	The amount of product in a pack or container.
DrugQuantity OrWeight Proportion*	Max n4.	Mandatory where relevant	The quantity prescribed in terms of either the packsize or number of doses. * To note, the variable name is misspelled.
UnitOf Measurement		Mandatory where relevant	Should be a SNOMED CT code but looks like input varies based on organisation collecting the data. Describes what the DrugQuantity Or WeightProportion variable is measuring.
Dispensing Route	an1	Mandatory where relevant	Describes where the prescription item was dispensed to the patient. 1 = Inpatient (via internal pharmacy) 2 = Outpatient (via internal pharmacy) 3 = Outsourced pharmacy 4 = Homecare delivery 5 = Community pharmacy (FP10) 6 = Other (not listed)
HomeDelivery Charge	Max n18. Max n8	Mandatory	The amount charged for delivery of items to the patient's home.

Variable name	Variable type	Specification details	Variable description
TotalCost	Max n18. Max n8	Mandatory	The total cost of the activity that includes any agreed adjustments.
Derived SNOMED FromName	Max an255	Derived by NECS	dm+d code Over 90% NULL values (see Table 2).
DerivedVTM	Max an255	Derived by NECS	dm+d code - virtual therapeutic moiety Around one third of values are NULL (see Table 2).
DerivedVTM Name	Max an255	Derived by NECS	dm+d name - virtual therapeutic moiety Around one third of values are NULL (see Table 2).

The guidance to analyse the HCD dataset via OpenSAFELY-TPP is published online and available to all¹¹. **Box 1** below provides an example of the code used to include information on HCD prescriptions within an OpenSAFELY “study definition”; this code is used to define a cohort. As part of this process, users also need to create specific codelists, which cover the medications of interest. Due to the nature of the “Drug-Name” variable in the HCD dataset (discussed in more detail in the Results section) the codelists used to query HCD data do not follow an existing naming convention such as the British National Formulary (BNF) or the mandated NHS standard dictionary of medicines and devices (dm+d).

- A list of the existing HCD codelists can be found on the OpenSAFELY codelists website¹².
- If a codelist does not already exist, then the user will need to create one. These codelists will need to be based on the unique values from the DrugName variable. A list of all the unique values for the drug name variable can be found in the analysis code under *Extended data*⁹.

Box 1. Example of code used in an OpenSAFELY-TPP study definition to query the HCD dataset

The example code below flags all patients who were prescribed adalimumab between October 2019 and March 2020, in the HCD dataset. The adalimumab filter is based on the adalimumab codelist, found on OpenSAFELY codelists¹³.

```
prescribed_adalimumab=patients.with_high_
  cost_drugs(
    drug_name_matches= adalimumab_
    codes,
    between = ["2019-10-01", "2020-
03-31"],
    find_first_match_in_period=True,
    returning="binary_flag",
    return_expectations={"incidence"
: 0.05,},
  )
```

Further guidance on querying HCD dataset via an OpenSAFELY-TPP study definition can be found online¹¹.

Full variable list. The national HCD dataset in OpenSAFELY-TPP covers 21 variables, three of which are currently available to query via OpenSAFELY-TPP study definitions. [Table 1](#) provides a complete list of the variables in the dataset with a brief description of the variable type and specification.

Variable completeness. Variable completeness is shown in [Table 2](#). The completeness (the percentage of records with non-missing values) differs across the variables: some variables have very few or no missing values (DrugName, PersonAge, TotalCost) whilst others are much less complete (DrugStrength, DrugVolume, TherapeuticIndicationCode). Variables that have high levels of missing data may not be suitable for inclusion in analysis and could be a target for improving the coverage of the data collection.

The number of unique values recorded in each variable is an indication of whether the variable uses nationally standardised inputs, following a codelist with restricted input at the data collection stage, or uses locally compiled lists, which will vary across providers. The DrugName variable is an example of a variable which uses locally compiled lists rather than national standardised input on collection and therefore has many unique values at a national level (>20,000). By contrast the ActivityTreatmentFunctionCode variable only has 143 unique values in the latest year, suggesting use of a nationally standardised list on collection.

Descriptive analysis

Using OpenSAFELY-TPP, descriptive analysis of the characteristics of the patients who receive HCD can be carried out for the first time on a large scale, to inform related analysis on COVID-19. In this Data Note, we have provided some summary analysis of the demographic characteristics of patients in the HCD dataset (including age, sex, ethnicity and geographic location) and compare this patient group to other patients registered at TPP practices.

Using OpenSAFELY-TPP, we produced a descriptive analysis to better understand the demographic characteristics of patients that appear in the HCD dataset and how these patients compare to others registered at TPP practices. This analysis was

Table 2. Completeness of each variable in the national high-cost drugs dataset.

Variable Name	Records for FY 2018/19				Records for FY 2019/20			
	Total records	% missing	% Numeric (and not missing)	Number of unique values	Total Records	% missing	% Numeric (and not missing)	Number of unique values
PersonAge	2,799,394	1.8%	100.0%	114	3,984,198	2.9%	100.0%	125
PersonGender	2,799,394	0.0%	100.0%	3	3,984,198	0.0%	100.0%	4
Activity Treatment FunctionCode	2,799,394	9.2%	100.0%	144	3,984,198	6.0%	100.0%	143
Therapeutic IndicationCode	2,799,394	49.6%	12.6%	7,230	3,984,198	68.3%	36.0%	5,130
HighCostTariff Excluded DrugCode	2,799,394	29.0%	78.9%	12,687	3,984,198	43.5%	87.9%	11,722
DrugName	2,799,394	1.0%	0.0%	20,698	3,984,198	1.4%	0.1%	19,609
RouteOf Administration	2,799,394	7.3%	2.3%	1,213	3,984,198	41.8%	69.9%	625
DrugStrength	2,799,394	24.1%	11.8%	22,887	3,984,198	31.4%	13.1%	9,091
DrugVolume	2,799,394	42.2%	30.3%	12,168	3,984,198	60.8%	22.7%	7,472
DrugPackSize	2,799,394	21.8%	35.6%	7,952	3,984,198	24.7%	48.2%	7,156
DrugQuantity OrWeight Proportion	2,799,394	6.8%	65.5%	10,349	3,984,198	7.4%	65.5%	13,871
UnitOf Measurement	2,799,394	100.0% ¹	23.4%	57	3,984,198	77.7%	84.5%	853
Dispensing Route	2,799,394	12.4%	97.7%	235	3,984,198	24.6%	99.1%	36
HomeDelivery Charge	2,799,394	1.8%	100.0%	6,190	3,984,198	3.8%	100.0%	11,639
TotalCost	2,799,394	0.2%	100.0%	160,060	3,984,198	0.0%	100.0%	169,294
Derived SNOMED	2,799,394	92.3%	100.0%	205	3,984,198	91.7%	100.0%	218
DerivedVTM	2,799,394	30.6%	100.0%	684	3,984,198	36.5%	100.0%	737
DerivedVTM Name	2,799,394	30.6%	0.0%	682	3,984,198	36.5%	0.0%	736

¹ Rounded to nearest decimal point - there are 657 non-missing records in total out of 2,799,394

restricted to patients who were registered at a TPP practice between 1st January 2020 and 31st March 2020 inclusive. Any patient who appeared in the HCD dataset between 1st October 2019 and 31st March 2020 were counted as in the HCD population, all other patients were included in the comparator population (not in HCD).

Information governance and ethical approval

This study was approved by the Health Research Authority (REC Reference 20/LO/0651) and by the LSHTM Ethics Board (Reference 21863).

NHS England is the data controller; TPP is the data processor; and the key researchers on OpenSAFELY are acting on behalf of NHS England. This implementation of OpenSAFELY is hosted within the TPP environment which is accredited, the ISO 27001 information security standard and is NHS IG Toolkit compliant^{14,15}; 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, the platform is via a virtual private network (VPN) connection, restricted, a small group of researchers; the researchers hold contracts with NHS

England and only access the platform, 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 counts¹⁶. The OpenSAFELY research platform adheres, the obligations of the UK General Data Protection Regulation (GDPR) and the Data Protection Act 2018. Since March 2020, the Secretary of State for Health and Social Care used powers under the UK Health Service (Control of Patient Information) Regulations 2002 (COPI), require organisations to process confidential patient information for the purposes of protecting public health, providing healthcare services, the public and monitoring and managing the COVID-19 outbreak and incidents of exposure; this sets aside the requirement for patient consent¹⁷. Taken together, these provide the legal bases, link patient datasets on the OpenSAFELY platform. GP practices, from which the primary care data are obtained, are required, share relevant health information, support the public health response, the pandemic, and have been informed of the OpenSAFELY analytics platform.

Data access and verification

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-TPP is drawn from General Practice data across England where TPP is the Data Processor. TPP developers (CB, JC, JP, FH, and SH) initiate an automated process to create pseudonymised records in the core OpenSAFELY-TPP database, which are copies of key structured data tables in the identifiable records. These 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. DataLab developers and primary investigators (BG, LS, CEM, SB, AJW, KW, WJH, HJC, DE, PI, SD, GH, BBC, RMS, ID, KB, EJW and CTR) 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-TPP platform as a whole is available for review on Github.

The data management and analysis code for this paper was led by AR with contributions from WH, BMK, SD, PI and DE.

Software and reproducibility

Data management was performed using Python, with analyses carried out using R. All of the code used for data management and analyses is openly available for inspection and re-use from the OpenSAFELY-TPP high cost drugs - research GitHub repository (*Extended data*)⁹. More information on data access and verification is available in the supplementary material.

Results

Variable overview

Patient ID, Financial Year and Financial Month. The HCD dataset contains submissions from April 2018 to March 2021. In FY 2018/19, there were 2.8 million submissions for 1.1 million unique patient IDs. The average number of submissions per patient over the year was 2.6. In FY 2019/20 there were 4.0 million submissions for 1.3 million unique patient IDs. The average number of submissions per patient over the year was 3.1. However, there are only a small number of submissions for FY 2020/21, and these are prospective submissions submitted before the patient had received the medicine. We recommend that these records are ignored and not used in any analysis. The patient ID in the HCD dataset is used to match the information from this dataset to other patient-level data included in the OpenSAFELY-TPP environment. This ID allows OpenSAFELY-TPP users to include information from other data sources on the platform (e.g. hospital episodes or COVID-19 testing) in any analysis of HCD use.

The financial year and financial month variables in the HCD dataset are stored separately, which makes analysis over a specific time interval a little more complex than if it were combined as a single variable. The OpenSAFELY-TPP cohort extractor has been developed so that users can query dates easily and the translation from conventional date format to separate FY and financial month filters is done in the background of the OpenSAFELY-TPP cohort extractor.

Drug name. The drug name variable is a mandatory part of the submission (where relevant) and can be used in OpenSAFELY-TPP study definitions to provide information on the HCD a patient has been prescribed in a given time period. This variable can be queried to produce: a flag to indicate a patient was ever prescribed a medicine between two dates; the first date a patient was prescribed a medicine between two dates and the last date a patient was prescribed a medicine between two dates. This variable is populated for 99% of records.

There are almost 21,000 unique values for the drug name variable in FY 2018/19 and almost 20,000 in FY 2019/20. The majority of these are not in the NHS-mandated dm+d format.

This variation in the recording of drug names in submissions means that codelists cannot be created using existing data definitions (e.g., dm+d or BNF) and bespoke codelists need to be created to try and pick up all possible variants of a drug name. These bespoke codelists are created by carrying out keyword searches on the list of unique values in the DrugName variable. The range of values in the DrugName variable is dealt with by building bespoke codelists rather than via the OpenSAFELY-TPP study definition.

Not only is there variation in how the same medicine is referenced by different providers, but we also found occurrences of misspelled drug names. For example, when constructing a bespoke code list for the medicine dupilumab, we included the

misspelling dipilumab as that appeared in the DrugName variable values.

Taking the medicine adalimumab as an example, based on a keyword search, there were around 460 different ways that adalimumab was described in the DrugName variable, including various brand names⁸. The keywords used were adalimumab, amgevita, hyrimoz, humira, idacio and imraldi. The search ignored whether letters were uppercase or lowercase. The twenty most common names that appear through this search are shown in [Table 3](#).

Descriptive analysis

Age. A higher proportion of patients in the HCD dataset belonged to older age bands (50+) than the other patients registered at TPP practices ([Figure 1](#), [Table 4](#)). For females, 67.8% of the patients in the HCD dataset were in older age bands compared to 39.2% of patients not in the HCD dataset. For males, 70.0% of the patients in the HCD dataset were in older age bands compared to 36.7% of patients not in the HCD dataset.

Ethnicity. A higher proportion of patients in the HCD dataset were in the White ethnicity group compared to the other patients registered at TPP practices ([Figure 2](#), [Table 5](#)). For females, 67.5% of the patients in the HCD dataset were in the White ethnicity group compared to 63.7% of patients not in the HCD dataset. For males, 65.9% of the patients in the HCD dataset were in the white ethnicity group compared to 59.7% of patients not in the HCD dataset.

Geographical variation - STP. We looked at geographical variation by grouping patients by sustainability and transformation partnerships (STPs). The proportion of an STP population that appears in the HCD dataset ranged from around 0.5% to 1.5% ([Table 6](#)).

Data validation

Summary

The national HCD dataset provides information on prescriptions and spend on HCD at patient level for the FY 2018/19 and 2019/20. There were 4.0 million submissions for 1.3 million

Table 3. Top twenty names that appear in the DrugNames variable using the adalimumab keyword search as described above.

DrugName	Frequency count
ADALIMUMAB	34,498
HC ADALIMUMAB IMRALDI 40 mg Injection Pre Filled Pen	17,949
ADALIMUMAB (D2E7) - HOMECARE 40 mg Preloaded Pen	11,495
ADALIMUMAB (IMRALDI) (HOMECARE)	8,205
ADALIMUMAB (IMRALDI)	7,984
HC ADALIMUMAB HUMIRA 40 mg Injection Pre Filled Pen	7,167
Adalimumab 40mg/0.8ml solution for injection pre-filled disposable devices	6,809
ADALIMUMAB REFERENCE PRICE	4,810
HOMECARE ADALIMUMAB (IMRALDI)	4,373
ADALIMUMAB (HUMIRA)_(HOMECARE) 40 mg in 0.4mL Pre-filled Injection Pen	3,649
ADALIMUMAB(AMGEVITA)	3,005
HOMECARE IMRALDI (ADALIMUMAB)	2,926
Adalimumab	2,461
HOMECARE - ADALIMUMAB (AMGEVITA) 40 mg in 0.8ml Auto Injector Pen	2,440
ADALIMUMAB (HUMIRA) (HOMECARE)	2,251
ADALIMUMAB(HUMIRA)	1,968
Adalimumab (Homecare)	1,820
HOMECARE AMGEVITA (ADALIMUMAB)	1,703
ADALIMUMAB - IMRALDI (HOMECARE)	1,477
HOMECARE ADALIMUMAB!40mg/0.8mL! PEN (HYRIMOZ)	1,466

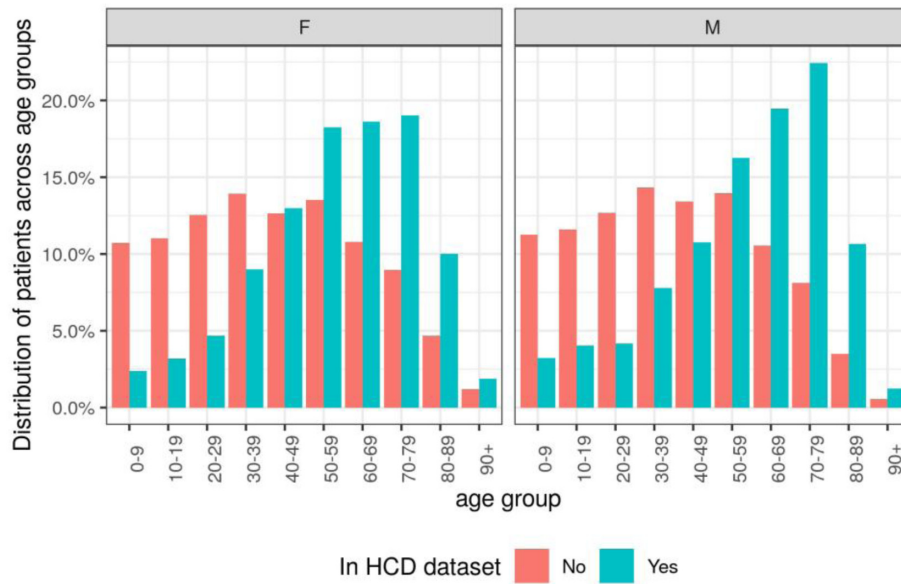


Figure 1. Distribution of patients across age groups, categorised by sex and whether patient appeared in the high cost drugs (HCD) dataset.

Table 4. Distribution of patients across age groups, categorised by sex and patient appearance in the high-cost drugs (HCD) dataset.

Age group	In HCD dataset		Not in HCD dataset	
	Number in each age group	% of total in each age group	Number in each age group	% of total in each age group
Females				
Total	101,596	NA	11,623,335	NA
0-9	2,437	2.4%	1,244,818	10.7%
10-19	3,235	3.2%	1,282,360	11.0%
20-29	4,736	4.7%	1,459,202	12.6%
30-39	9,125	9.0%	1,617,461	13.9%
40-49	13,200	13.0%	1,468,794	12.6%
50-59	18,547	18.3%	1,570,566	13.5%
60-69	18,897	18.6%	1,251,879	10.8%
70-79	19,338	19.0%	1,042,945	9.0%
80-89	10,163	10.0%	545,253	4.7%
90+	1,918	1.9%	140,057	1.2%
Males				
Total	96,061	NA	11,640,791	NA
0-9	3,109	3.2%	1,311,321	11.3%
10-19	3,883	4.0%	1,348,521	11.6%
20-29	4,012	4.2%	1,474,829	12.7%
30-39	7,461	7.8%	1,669,359	14.3%

Age group	In HCD dataset		Not in HCD dataset	
	Number in each age group	% of total in each age group	Number in each age group	% of total in each age group
40-49	10,323	10.7%	1,563,269	13.4%
50-59	15,606	16.2%	1,624,034	14.0%
60-69	18,688	19.5%	1,229,951	10.6%
70-79	21,559	22.4%	944,582	8.1%
80-89	10,221	10.6%	408,320	3.5%
90+	1,199	1.2%	66,605	0.6%

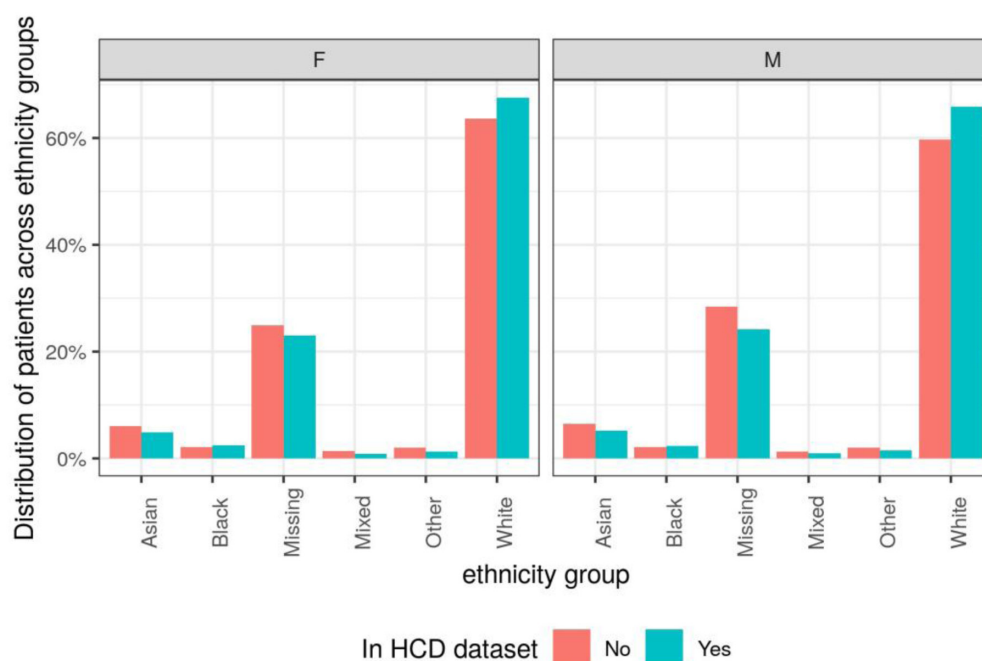


Figure 2. Distribution of patients across ethnicity groups, split by sex and whether patient appears in high-cost drugs (HCD) dataset.

unique patient IDs, with an average number of submissions per patient over the year of 3.1. This data is now available with the OpenSAFELY-TPP framework, linked to other NHS records at patient level, alongside reusable code to undertake analyses related to COVID-19. We observed substantial variation in missing data between specific fields in the data (0% gender - 77.7% unit of measurement) and consequently have made three fields available: FinancialMonth, FinancialYear and Drug-Name. The first output using this data in OpenSAFELY-TPP has already been published: a research paper on the association between the use of immune modifying medicines to treat immune-mediated inflammatory diseases and severe COVID-19 outcomes¹⁸. This report can support those undertaking further analysis on COVID-19 using the HCD in OpenSAFELY-TPP.

Strengths and weaknesses

The national HCD dataset includes information on all HCD, rather than being limited to a specific class of medicines or disease as some other data collections are, such as national disease registries. At the outset of the pandemic, there was no data available on medicines supplied by hospitals; the provision of HCD allows researchers and the NHS to capitalise on new information to inform analysis. We utilised an existing data collection, efficiently re-using information already collected from NHS providers, but did not add any further burden to hospitals. Access to the national HCD dataset via OpenSAFELY-TPP means that this dataset can be analysed alongside a range of other patient level information, meaning that analysis of drugs prescribed and clinical outcomes is straightforward to run, and does not involve any additional time to source data.

Table 5. Distribution of patients across ethnicity groups, categorised by sex and whether patient appears in high-cost drugs (HCD) dataset.

Ethnicity group	In HCD dataset		Not in HCD dataset	
	Number in each age group	% of total in each age group	Number in each age group	% of total in each age group
Females				
Total	101,596	NA	11,623,335	NA
Asian	4,978	4.9%	700,909	6.0%
Black	2,467	2.4%	247,046	2.1%
Missing data	23,373	23.0%	2,891,214	24.9%
Mixed	897	0.9%	155,264	1.3%
Other	1,270	1.3%	228,596	2.0%
White	68,611	67.5%	7,400,306	63.7%
Males				
Total	96,061	NA	11,640,791	NA
Asian	4,971	5.2%	749,327	6.4%
Black	2,238	2.3%	247,208	2.1%
Missing data	23,263	24.2%	3,310,054	28.4%
Mixed	900	0.9%	150,515	1.3%
Other	1,383	1.4%	230,568	2.0%
White	63,306	65.9%	6,953,120	59.7%

Table 6. Proportion of sustainability and transformation partnerships (STP) TPP patient population that appear in the high-cost drugs dataset.

STP code	STP name	Number of patients in HCD dataset	Total patients registered at TPP practice	% of patients in HCD dataset
E54000005	West Yorkshire and Harrogate (Health and Care Partnership)	12,447	2,304,250	0.5%
E54000006	Humber, Coast and Vale	6,175	1,045,049	0.6%
E54000007	Greater Manchester Health and Social Care Partnership	2,230	219,453	1.0%
E54000008	Cheshire and Merseyside	1,787	160,579	1.1%
E54000009	South Yorkshire and Bassetlaw	8,222	1,080,379	0.8%
E54000010	Staffordshire and Stoke on Trent	766	83,909	0.9%
E54000012	Joined Up Care Derbyshire	6,139	878,428	0.7%
E54000013	Lincolnshire	9,770	659,544	1.5%
E54000014	Nottingham and Nottinghamshire Health and Care	6,434	923,839	0.7%
E54000015	Leicester, Leicestershire and Rutland	13,503	1,030,600	1.3%
E54000016	The Black Country and West Birmingham	4,510	351,433	1.3%

STP code	STP name	Number of patients in HCD dataset	Total patients registered at TPP practice	% of patients in HCD dataset
E54000017	Birmingham and Solihull	6,199	536,086	1.2%
E54000020	Northamptonshire	3,349	613,173	0.5%
E54000021	Cambridgeshire and Peterborough	5,257	866,038	0.6%
E54000022	Norfolk and Waveney Health and Care Partnership	12,359	870,625	1.4%
E54000023	Suffolk and North East Essex	4,580	757,888	0.6%
E54000024	Bedfordshire, Luton and Milton Keynes	6,923	1,004,396	0.7%
E54000025	Hertfordshire and West Essex	5,496	808,624	0.7%
E54000026	Mid and South Essex	13,625	1,188,216	1.1%
E54000027	North West London Health and Care Partnership	11,558	1,533,808	0.8%
E54000029	East London Health and Care Partnership	647	104,099	0.6%
E54000033	Sussex and East Surrey Health and Care Partnership	9,138	886,111	1.0%
E54000035	Surrey Heartlands Health and Care Partnership	286	42,681	0.7%
E54000036	Cornwall and the Isles of Scilly Health and Social Care Partnership	2,131	252,544	0.8%
E54000037	Devon	7,873	780,333	1.0%
E54000040	Bath and North East Somerset, Swindon and Wiltshire	9,380	909,687	1.0%
E54000041	Dorset	5,558	783,453	0.7%
E54000042	Hampshire and the Isle of Wight	4,189	596,569	0.7%
E54000043	Gloucestershire	5,455	506,331	1.1%
E54000044	Buckinghamshire, Oxfordshire and Berkshire West	236	26,080	0.9%
E54000049	Cumbria and North East	11,369	1,649,249	0.7%
NA	NA	70	8,754	0.8%

However, there are several caveats that need to be considered when using this resource. The national HCD dataset in OpenSAFELY-TPP is comprehensive, however due to the scale and speed at which it was assembled, it is possible that unknown inconsistencies or omissions may have occurred. The inputs to the DrugName and other variables are not standardised at a national level, which means there is a wide range of values (many thousands over a financial year). This is a feature of the data collection process. This means that the creation of bespoke codelists is required each time the HCD dataset is used for new analysis, and there can be misspellings of drug names. Alongside this non-standardised input, some variables have a high proportion of missing records. Finally, the dataset in OpenSAFELY-TPP is currently limited to a one-off collection covering submissions from FY 2018/19 and FY 2019/20.

Findings in context

To date there has been limited research conducted using patient-level HCD data in the UK. The national HCD dataset covers all HCDs, which means that, for the first time, researchers can produce analyses covering large numbers of patients, over one million unique patients in each year. As an example, the first analysis using the HCD dataset within OpenSAFELY-TPP was conducted to ascertain the risk of severe COVID-19 outcomes associated with immune-mediated inflammatory diseases and immune modifying therapies: a nationwide cohort study in 17 million individuals¹⁸. We are unaware of any other use of comprehensive and routinely collected data on medicines supplied by hospitals to individual patients in England. There are several large clinical registry studies in England focused on specific diseases or medicines; although detailed and

comprehensive, they are limited by underreporting, loss to follow-up and absence of information from elsewhere in the NHS¹⁹⁻²¹. Combining detailed clinical registry data with the data available in OpenSAFELY may enhance the quality and robustness of analysis that can be achieved.

Policy implications

In March 2020, at the outset of the COVID-19 pandemic, there was no routine comprehensive hospital medicines data from the UK available to researchers and organisations. Since the onset of the pandemic, the NHS has improved access to information on the usage of medicines in hospitals, the NHS Business Services Authority now publish a monthly summary of the volume of medicines issued in hospitals²² which is publicly available and NHS Digital are developing the electronic prescribing and medicines administration (EPMA) data collection with a subset of hospital data now available²³. The availability of the national HCD dataset through OpenSAFELY-TPP adds to this collection of knowledge, and researchers can access this resource, along with all other OpenSAFELY-TPP data sources, by following the OpenSAFELY access process²⁴.

However, the current dataset in OpenSAFELY-TPP is a one-off collection covering submissions from FY 2018/19 and FY 2019/20, and there is no process in place as at time of writing to routinely update the information available in the HCD dataset. Whilst this is very useful for assessing events and outcomes early in the COVID-19 pandemic, a routine update of the data is needed to assess current high-priority questions and future important questions. For example, a routine update to this data will allow assessment of COVID-19 vaccine effectiveness in people using high-cost medicines or indeed people with a recorded diagnosis likely to be treated with a HCD. Our work demonstrates that it is possible for the NHS to collate the data at a national level and we strongly recommend that a routinely updated version of the HCD dataset is produced and made available to all interested users, including via the OpenSAFELY-TPP platform.

Data availability

OpenSAFELY: <https://opensafely.org/>

The project contains the following underlying data:

- All data were linked, stored and analysed securely within the OpenSAFELY platform. Data include pseudonymized 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. Detailed pseudonymised patient data is potentially re-identifiable and therefore not shared.
- For security and privacy reasons, OpenSAFELY is very different to other approaches for EHR data analysis. The platform does not give researchers unconstrained access to view large volumes of pseudonymised

and disclosive patient data, either via download or via a remote desktop. Instead we have produced a series of open source tools that enable researchers to use flexible, pragmatic, but standardised approaches to process raw electronic health records data into “research ready” datasets, and to check that this has been done correctly, without needing to access the patient data directly. Using this data management framework we also generate bespoke dummy datasets. These dummy datasets are used by researchers to develop analysis code in the open, using Github. When their data management and data analysis scripts are capable of running to completion, and passing all tests in the OpenSAFELY framework, they are finally sent through to be executed against the real data inside the secure environment, using the OpenSAFELY jobs runner, inside a container using Docker, without the researcher needing access to that raw potential disclosive pseudonymised data themselves. The non-disclosive summary results output tables, logs, and graphs are then manually reviewed, as in other systems, before release.

- As part of building that resource for the community, over the next six months we are working with NHS England to cautiously on-board a small number of external pilot users to develop their analyses on OpenSAFELY. This process is described in further detail on our webpage, here: <https://opensafely.org/onboarding-new-users/>.

Extended data

Analysis code available from: <https://github.com/opensafely/high-costdrugs-research>

Archived analysis code as at time of publication: <https://doi.org/10.5281/zenodo.5747620>

License: MIT

Acknowledgements

We would like to thank all Commissioning Support Units for their collaborative work in assembling the high-cost drugs dataset and we are very grateful for NECS, the North East CSU, for processing it into a single standardised dataset. We would like to thank the TPP Technical Operations team for all the support received throughout this work, and for generous assistance from the information governance and database teams at NHS England/NHSX. We also want to thank Alison Kay and Graeme Hood for their expertise whilst collating the dataset.

The findings and conclusions in this report are those of the authors and do not necessarily represent the views of the funders.

References

1. **Prescribing Costs in hospitals and the Community 2019-2020.** NHS digital. (accessed 6 Apr 2021).
[Reference Source](#)
2. Goldacre B, MacKenna B: **The NHS deserves better use of hospital medicines data.** *BMJ.* 2020; **370**: m2607.
[PubMed Abstract](#) | [Publisher Full Text](#)
3. Matthews A, Donaldson LJ, Evans SJ, *et al.*: **Safety of medicines delivered by homecare companies.** *BMJ.* 2018; **361**: k2201.
[PubMed Abstract](#) | [Publisher Full Text](#)
4. Department of Health and Social Care: **High cost drugs.** GOV.UK, 2012; (accessed 6 Apr 2021).
[Reference Source](#)
5. **OpenSAFELY: Home,** (accessed 20 Apr 2021).
[Reference Source](#)
6. Williamson EJ, Walker AJ, Bhaskaran K, *et al.*: **Factors associated with COVID-19-related death using OpenSAFELY.** *Nature.* 2020; **584**(7821): 430–6.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
7. England NHS: **NHS England » NHS England drugs list.** (accessed 6 Apr 2021).
[Reference Source](#)
8. **Drugs Patient Level Contract Monitoring Data Set.** (accessed 6 Apr 2021).
[Reference Source](#)
9. OpenSAFELY GitHub: **Exploring the high cost drugs dataset to understand data coverage and quality.** OpenSAFELY.
10. **Activity treatment function code.** (accessed 6 Apr 2021).
[Reference Source](#)
11. **OpenSAFELY documentation.** (accessed 6 Apr 2021).
[Reference Source](#)
12. **OpenSAFELY Codelists.** (accessed 6 Apr 2021).
[Reference Source](#)
13. OpenSAFELY Codelists: **High cost drugs - Adalimumab.** (accessed 6 Apr 2021).
[Reference Source](#)
14. **BETA – Data Security Standards.** NHS Digital. (accessed 30 Apr 2020).
[Reference Source](#)
15. **Data Security and Protection Toolkit.** NHS Digital. (accessed 30 Apr 2020).
[Reference Source](#)
16. **ISB1523: Anonymisation Standard for Publishing Health and Social Care Data.** NHS Digital. (accessed 30 Apr 2020).
[Reference Source](#)
17. Secretary of State for Health and Social Care - UK Government: **Coronavirus (COVID-19): notification to organisations to share information.** 2020.
[Reference Source](#)
18. MacKenna B, Kennedy NA, Mehkar A, *et al.*: **Risk of severe COVID-19 outcomes associated with immune-mediated inflammatory diseases and immune modifying therapies: a nationwide cohort study in the OpenSAFELY platform.** *medRxiv.* 2021.
[Publisher Full Text](#)
19. Burden AD, Warren RB, Kleyn CE, *et al.*: **The British Association of Dermatologists' Biologic Interventions Register (BADBIR): design, methodology and objectives.** *Br J Dermatol.* 2012; **166**(3): 545–54.
[PubMed Abstract](#) | [Publisher Full Text](#)
20. **BSRBR-RA Study.** homepage, (accessed 13 Oct 2021).
[Reference Source](#)
21. **UK IBD Registry - IBD Registry.** 2020; (accessed 13 Oct 2021).
[Reference Source](#)
22. **Secondary Care Medicines Data (SCMD).** Open Data Portal BETA, (accessed 6 Apr 2021).
[Reference Source](#)
23. **Electronic prescribing and medicines administration (EPMA) data.** NHS Digital. (accessed 6 Apr 2021).
[Reference Source](#)
24. OpenSAFELY: **Onboarding new users to OpenSAFELY.** (accessed 22 Apr 2021).
[Reference Source](#)

Open Peer Review

Current Peer Review Status:   

Version 1

Reviewer Report 06 June 2022

<https://doi.org/10.21956/wellcomeopenres.19195.r50570>

© 2022 Page K et al. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Katie Page 

Centre for Health Economics Research and Evaluation, University of Technology Sydney, Sydney, NSW, Australia

Lutfun Hossain

Centre for Health Economics Research and Evaluation, University of Technology Sydney, Sydney, NSW, Australia

This data report outlines and describes the collation of a national open-source database for high costs drugs in the NHS England. Due to complex funding arrangements, much of this information was stored locally. The aim was to produce a comprehensive large-scale database on the use of high-cost medicines to better understand the impacts of these medicines on COVID-19 outcomes. It is an ambitious but worthwhile project achieved in an impressive timeframe, and it is well described. It clearly identifies a gap in researchers' access to detailed and reliable medicines information on high-cost drugs on a widespread scale. The authors also provide access to all code used for data management and analysis. Further efforts in this space are needed to enable researchers to conduct worthwhile studies on medicine use at a population level.

The dataset is comprehensive and covers the time from 2018 to 2020 (2 financial years) with over 1 million unique patients per year from 5.4 million submissions for payment for high-cost drugs. There are a total of 21 variables but only 3 are currently available for analysis in the open access data analysis platform (OpenSAFELY). Descriptive analysis shows that older (70+ years), white individuals were overrepresented in the database. A strength of the database is that the high-cost drug records can be linked to other patient records in primary and secondary care in OpenSAFELY, allowing a range of linkage studies not previously possible.

Overall, the paper is well written and the processes for data collation are clear. There are a few things we noted:

We would have liked a more detailed description of the range and mean costs of such drugs. What is the minimum cost of drugs to be subsidized outside of hospital contracts or tariffs? There is a brief description in the abstract but some further elaboration in the introduction would be

beneficial. Further information about the type and class of drugs in the HCD list would be useful because such details would allow international researchers a better understanding and comparison with their own countries' databases.

The open access element is a real strength of the study, but we are curious as to why only three variables can be accessed in such a way and there is no cost variable so far. Are the authors proposing to add to these variables over time? The amount of missing data for total cost is low so we wondered why that was not included with the other three variables? Is this a limitation of OpenSAFELY? As health economists, including cost variables alongside clinical information, is of paramount importance.

As noted by the authors the drug name variable, which is mandatory for reporting and has only 1% of missing data, is not in a standardized format. The researchers have provided links to access codelists that have been created for medicines in the research to date. For medicines outside of this list, researchers will need to create and use their own bespoke code or codelists to access all the data they might need. This could be burdensome or act as a barrier to use for some researchers. Nevertheless, it is a comprehensive list and with the right search tools relevant information can be extracted. Suggestions by the research team, or open source code on the most useful way to use and extract this information would be valuable.

The dataset is just a one-off compilation and covers only 2 years' worth of data which can be useful in the early stages of the pandemic. Updates will be essential for tracking the longer-term outcomes linked to COVID-19 including long COVID and vaccine impacts. Codelists will need to be updated on an ongoing basis to ensure all high-cost drug data can continue to be accessed.

Overall, it is very promising that these substantial logistical and data compilation efforts have resulted in such a useful and easy to access resource. It will be interesting to see how these data are used as we look forward to seeing the outcomes of these future studies.

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?

Yes

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

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Dr Katie Page is a senior health economist and academic psychologist with expertise in the use and evaluation of administrative health data and economic evaluations. Dr Lutfun Hossain is a pharmacist and researcher with expertise in quality use of medicines.

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 19 May 2022

<https://doi.org/10.21956/wellcomeopenres.19195.r50328>

© 2022 Jack R. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Ruth H. Jack 

Centre for Academic Primary Care, Lifespan and Population Health, School of Medicine, University of Nottingham, Nottingham, UK

This data note describes the first instance of high-cost drug (HCD) prescribing data from NHS hospitals in England being collected together and made available for research. The combined dataset covers the financial years 2018/19 to 2019/20, and it is possible to link records to other datasets within OpenSAFELY in order to determine other variables, including hospital episodes and patient characteristics. Descriptions of the variables in the HCD dataset, including the number of unique values and completeness, provide a helpful guide for where improvements in data collection are needed to enable more of this dataset to be used in research.

This is an incredibly useful piece of work, demonstrating the feasibility of collating such information but highlighting the problems when non-standard values are used. The example of drug names given, where around 460 ways of describing the drug adalimumab, shows how improvements in standardised data collection are needed, as well as open access collaboration of code lists already created. Due to the incompleteness of many of the variables, only a small proportion are of sufficient quality to be made available for research. This dataset is currently a one-off, made possible by the pressing need to understand the COVID-19 pandemic. A routinely updated dataset with more information in a useable format would be a valuable resource for understanding the fuller picture of healthcare within England.

The processes used are clearly described, with references to publicly available codes, analysis and other resources.

I have only a few minor revisions to suggest:

1. Adding a key to Table 1 to describe what the variable types are might be helpful.
2. The variable type is missing for UnitOfMeasurement in Table 1.
3. It would be more useful to have the number of non-missing (possibly excluding not known) records for each variable in Table 2 rather than repeating the total number of records in each financial year.
4. In 2019/20 there are four unique values for PersonGender, but only three are listed in Table

- 1 – what was the extra value?
5. On page 8 there are some sentences that aren't clear: "The OpenSAFELY research platform adheres, the obligations[...]" and "GP practices, from which the primary care data are obtained, are required, share relevant health information[...]".
6. It's not clear where the "supplementary material" referred to in the 'Software and reproducibility' section is.
7. The column headings in Table 5 refer to age, when the table shows ethnic group data.

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?

Yes

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

Yes

Competing Interests: No competing interests were disclosed.**Reviewer Expertise:** Epidemiology using large databases of electronic health records**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.**

Reviewer Report 12 May 2022

<https://doi.org/10.21956/wellcomeopenres.19195.r50060>

© 2022 Day R. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**Richard Day** ¹ Department of Clinical Pharmacology & Toxicology, St Vincent's Hospital, Sydney, NSW, Australia² St Vincent's Clinical Campus, University of New South Wales, Sydney, NSW, Australia

A comprehensive knowledge of medications being prescribed and dispensed across health care systems, to which people, with which characteristics, in which settings, with which outcomes and at what cost, is an important goal from many perspectives. Global and geographic trends in outcomes, costs, and access can provide critical information that can lead to better system

responses. In the UK, a national, patient-level, data set of all medications used that includes those prescribed in NHS hospitals was not available for interrogation pre the COVID pandemic.

Rowan *et al.*'s report describes the construction of a national 'high cost drugs' (HCD) data set from UK NHS, complete, patient level, hospital data. The 'pseudonymised' patient's data comprising the HCD data set were managed with the OpenSAFELY-TPP, electronic health record platform (OSP) established across a large percentage (>40%) of GP practices in England. OSP analyses in primary care access each patient's, pseudonymised, complete, electronic primary care health record data allowing patient level data-linkages to be made to secondary databases and services.

The current exercise is designed to demonstrate to researchers the HCD data set and the availability of the 'extended' OSP resource and how this resource might be used. The case to build the extended OSP prevailed because of the COVID pandemic and the acute need for health system-wide data on impacts of multiple, new, potentially critical, medication-related interventions and consequences, often with very limited background information to guide decisions. To progress from no availability of a national data set for medicines used in NHS hospitals to a HCD data set since the start of the pandemic is a remarkable achievement especially given the overwhelming complexity of supply and funding arrangements for medications in NHS hospitals. All steps and methods in building this resource are made available for researchers to evaluate via 'open-source' access. Considerable skills and resources are likely required.

The example discusses the characteristics of the HCD built and now ready for limited use. There were 2.8 million 'submissions' (think purchases of a high cost drug) in NHS hospitals for 1.1 unique patients for FY 2018/9 rising to 4million for 1.3 million individuals in FY 2019/20, a substantial experience to interrogate. **Why did the 'purchases' increase?** The premise behind the HCD is that it likely reflects the significant, long-term, inflammatory disease group whose outcomes, experiences and connections to secondary services and primary care could be significantly impacted by COVID. High cost drugs for these conditions are prescribed largely by specialist services in hospitals, not the community, in the UK. Three of the 21 variables in the HCD made available for the study at this preliminary stage of development show that in comparison to the remainder of the OSP cohort the HCD patients were older and more likely White. The 21 variables to describe this HCD cover information needed to understand 'high cost drug' use and outcomes associated in individual patients across the NHS hospital and primary care system in UK. For example what conditions are being treated, what 'high cost drugs' have been chosen, what dose selected, etc. Not a lot new emerged of special interest in this example, however, that was not the point. The resource and its features and how to access the resource and HCD was the goal and this has been achieved.

In summary, an important resource produced with impressive speed, promises deeper understanding and insights around impacts of medicine use and external impacts, such as the COVID epidemic, on individual patients and the NHS is now available for interrogations. Results from these studies are eagerly anticipated.

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?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Personalising medications according to the principles of 'quality use of medicines'

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.
