BMJ Health & Care Informatics

Findability of UK health datasets available for research: a mixed methods study

Emily Griffiths ,[1] Rebecca M Joseph ,[2] George Tilston,[3] Sarah Thew,[4] Zoher Kapacee,[1] William Dixon ,[1,5] Niels Peek [1,5]

To cite: Griffiths E, Joseph RM, Tilston G, *et al*. Findability of UK health datasets available for research: a mixed methods study. *BMJ Health Care Inform* 2022;**29**:e100325. doi:10.1136/bmjhci-2021-100325

Received 08 January 2021
Accepted 25 October 2021

[1]Centre for Health Informatics, School of Health Sciences, The University of Manchester, Manchester, UK
[2]School of Medicine, University of Nottingham, Nottingham, UK
[3]Informatics, Imaging, and Data Sciences, The University of Manchester, Manchester, UK
[4]Manchester Academic Health Science Centre, Manchester, UK
[5]NIHR Manchester Biomedical Research Centre, Manchester, UK

Correspondence to
Dr Niels Peek;
niels.peek@manchester.ac.uk

ABSTRACT

Objective How health researchers find secondary data to analyse is unclear. We sought to describe the approaches that UK organisations take to help researchers find data and to assess the findability of health data that are available for research.

Methods We surveyed established organisations about how they make data findable. We derived measures of findability based on the first element of the FAIR principles (Findable, Accessible, Interoperable, Reproducible). We applied these to 13 UK health datasets and measured their findability via two major internet search engines in 2018 and repeated in 2021.

Results Among 12 survey respondents, 11 indicated that they made metadata publicly available. Respondents said internet presence was important for findability, but that this needed improvement. In 2018, 8 out of 13 datasets were listed in the top 100 search results of 10 searches repeated on both search engines, while the remaining 5 were found one click away from those search results. In 2021, this had reduced to seven datasets directly listed and one dataset one click away. In 2021, Google Dataset Search had become available, which listed 3 of the 13 datasets within the top 100 search results.

Discussion Measuring findability via online search engines is one method for evaluating efforts to improve findability. Findability could perhaps be improved with catalogues that have greater inclusion of datasets, field-level metadata and persistent identifiers.

Conclusion UK organisations recognised the importance of the internet for finding data for research. However, health datasets available for research were no more findable in 2021 than in 2018.

Summary

What is already known?

- Science benefits hugely from the sharing and reuse of datasets.
- There are many barriers to reuse, one of which is researchers not knowing what datasets already exist that may be relevant to their analysis.

What does this paper add?

- Organisations say that they want to make datasets more findable online, but that the time and personnel to achieve this is often lacking.
- We assess findability of UK health datasets in online searches.
- We found that this aspect of findability is no better in 2021 than it was in 2018.
- Online catalogues of health data rarely include identifiers that would enable proper referencing or field level metadata to indicate suitability for reanalysis.

INTRODUCTION

With 65 million people, a single payer health system, a unique identifier for its citizens' health data, and long-standing population-wide electronic health records (EHRs), the UK is uniquely placed to harness insights from routinely collected health data. UK primary care has been an early adopter of information technology, with most practices computerising prescribing and clinical record keeping over the past 20 years.

EHRs are collected routinely as part of direct care in the National Health Service (NHS), with tens of millions of records in existing 'e-cohorts' based on geography or diagnosis.[1–4] An e-cohort can enable researchers to 'investigate the broadest possible range of social and environmental determinants of health and social outcomes by exploiting the potential of routinely collected datasets'.[5] Some e-cohorts thus include other detailed data, for example, the Wales E-Cohort for Children includes educational attainment.[6] There is an ambition to sequence 5 million NHS patients' genomes.[7] Reuse of such data is advancing research, from disease aetiology to drug discovery, translational research and public health. There is a drive across many fields towards the sharing and reuse of health data.[8 9]

Apart from several long-standing and widely used national e-cohorts, for example, the Clinical Practice Research Datalink (CPRD),[10 11] there exist regional e-cohorts[12–14] that are known anecdotally to researchers connected to data providers, but are less well known by the wider research community. Lack of familiarity with existing e-cohorts may reduce their utilisation for research, weaken transparency and replicability of research and lead to duplication of effort in generating new equivalent datasets.[8 15]

The FAIR principles[16] were developed to guide sharing of scientific data and maximise the discovery, evaluation and reuse of such data. These four principles state that published data should be findable, accessible, interoperable and reusable. This article focuses on the principle of findability. The FAIR principle of findability recommends that data (or metadata) should be:

- Assigned a unique and persistent identifier.
- Described by rich metadata which links explicitly to the data described.
- Indexed in a searchable resource.

This project aimed to describe the current findability of routinely collected e-cohorts from the UK to a person (a researcher or interested citizen) using internet search engines. Specific objectives were: (1) to identify current approaches and potential barriers to increasing findability by surveying established organisations that facilitate access to health data (including e-cohorts) for research, and (2) to assess the findability of a target list of e-cohorts directly through internet searches and indirectly via online health data catalogues and see how findability changed between 2018 and 2021.

METHODS

Assessing approaches to findability at UK organisations supplying data to researchers

One route of access to routinely collected data for research is via organisations acting as data curators, providers, safe havens or research services. We wanted to understand what these organisations do to make their datasets findable and what obstacles they face in doing so. The datasets available may extend beyond health, but all are confidential datasets based in UK public sector organisations so findability practices should be transferable.

We conducted telephone surveys with staff from such organisations. We contacted the organisations with a participant information sheet via email, using publicly available contact information. These organisations were those of which the authors were aware, through their prior research or through participation in national initiatives such as the Farr Institute[17] or Safe Data Access Professionals.[18] As well as organisations specialising in health research, we included five that host other types of confidential data to understand their practices (eg, Her Majesty's Revenue and Customs (HMRC) Data Lab; see

Table 1 List of public sector organisations that took part in the surveys

Repository	Description	URL
Health Data Finder for Research	Health data finder is a metadata catalogue aiming to inform potential users about health datasets that are available for use in research	www.hdf.nihr.ac.uk
UK Data Service*	The UK Data Service enables access to a range of datasets, primarily in the field of social and economic research; funded by the Economic and Social Research Council (ESRC)	https://www.ukdataservice.ac.uk/
Consumer Data Research Centre (CDRC)*	The CDRC enables access to routinely collected consumer data; funded by the ESRC	https://www.cdrc.ac.uk
Urban Big Data Centre (UBDC)*	The UBDC enables access to urban-related data; funded by the ESRC	https://www.ubdc.ac.uk
Administrative Data Research Network (ADRN)*	The ADRN was a service funded by the ESRC to enable secure access to datasets	https://adrn.ac.uk
Electronic Data Research and Innovation Service (eDRIS)	eDRIS is a service coordinating access to the national Scottish health datasets	https://www.isdscotland.org/Products-and-Services/eDRIS
Health Informatics Centre—Trusted Research Environment (University of Dundee)	A data safe haven run as part of the University of Dundee, affiliated with National Health Service (NHS) Tayside and NHS Fife; the service coordinates access to local health datasets	https://www.dundee.ac.uk/hic/hicsafehaven
NHS Greater Glasgow and Clyde Safe Haven	A data safe haven and data service coordinating access to local health datasets	https://www.nhsggc.org.uk/about-us/professional-support-sites/nhsggc-safe-haven
CALIBER (University College London)	A platform for sharing data and methodologies; linked primary care, secondary care (hospital admissions), mortality and cancer registry data	https://www.ucl.ac.uk/health-informatics/caliber
Her Majesty's Revenue and Customs (HMRC) Data Lab*	A service providing secure access to deidentified HMRC data	https://www.gov.uk/government/organisations/hm-revenue-customs/about/research#the-hmrc-datalab
Connected Health Cities (CHC) North East and North Cumbria	CHC is a programme in the North of England which aims to use local health data and technology to improve health services; North East and North Cumbria are developing infrastructure to connect local hospitals with their trustworthy research environment—this will include development of a metadata catalogue	https://www.connectedhealthcities.org/connected-health-cities/cumbria-and-north-east-england
CHC Connected Yorkshire	Connected Yorkshire is based across Leeds, Sheffield and Bradford and works with the established Born in Bradford cohort; the dataset information described in this paper relates to the Born in Bradford study	https://www.connectedhealthcities.org/connected-health-cities/yorkshire-humber

*Not primarily health organisations.

asterisks in table 1). Up to two follow-up emails were sent to centres that did not initially respond.

Semistructured telephone surveys were conducted by RMJ and EG in April and May 2018 and focused on how organisations currently make their data findable, future plans to increase findability and any barriers to making data more findable. The HMRC Data Lab responded via email. An interview data collection sheet was developed from discussion among coauthors based on a preliminary interview with Electronic Data Research and Innovation Service conducted jointly by RMJ and EG. Notes were taken by RMJ or EG during each survey. Results were compiled by summarising and counting responses.

Assessing findability of e-cohorts for health research

We used several approaches to explore findability of e-cohorts from the perspective of health researchers. First, we quantified how frequently e-cohorts appeared in a series of internet searches. Second, we searched the health data catalogues for the prespecified e-cohorts, and for those e-cohorts that were present in the health data catalogues, we assessed whether the e-cohorts met the FAIR criteria of having rich metadata and a persistent identifier.

We aimed to replicate searches that might be carried out by a researcher trying to find data for their research or a member of the public curious about routine health information that is used in research. The study team, which has significant experience of research with health data and was involved in national initiatives such as the Farr Institute[17] and Health Data Research UK,[19] compiled a list of UK health-related e-cohorts known to them, without consulting the internet. This list served as targets for our searches (table 2), including well-known national datasets (eg, CPRD) and smaller, regional datasets of which the team had prior knowledge. The list also contained a number of data organisations, which provide access to e-cohorts.[20] Two kinds of search were performed to try to find these datasets.

Search using general internet searches

Search engines Google and Bing were searched separately in March 2018 (by EG and RMJ) and May 2021 (by EG and GT) using each of the following terms: health data research; acute care research datasets; community care research datasets; electronic health records; health datasets; health records research; hospital research datasets; primary care research datasets; secondary care research datasets and tertiary care research datasets (figure 1).

We used plain text search terms (no wildcards) to replicate simple searches the way someone might initially explore the public internet for relevant websites. We avoided terms such as 'case control study' or 'clinical cohort' as these relate to particular study designs, whereas we wanted to find routinely collected datasets. We wanted to replicate a well-motivated search and give a good chance of finding relevant results so we reviewed multiple pages of search results up to the hundredth listing. Search results were screened for reference to the target datasets (figure 1, step 1b). These references were either direct (the search result was itself the target's website) or indirect (a link in the search result led to the target).

Search using research data catalogues

To identify existing catalogues of UK health data, Google was searched using the terms 'health data catalogue' or 'research data catalogue' (omitting the quotation marks). The first 100 search results were screened for our targets (figure 1, step 2b).

Search using Google dataset search engine

After our 2018 searches were conducted, a new search engine was available from Google dedicated to finding datasets. In 2021, two authors (GT and EG) each searched for our 10 search terms in Google Dataset Search and reviewed the top 100 search results for our 13 target datasets.

Findability was assessed according to the following criteria:

1. Was a direct link found from Google or Bing searches?
2. Was there any indirect link to the e-cohort from the Google/Bing search results which might prompt a researcher to investigate further?
3. Was the e-cohort listed in one of the catalogues that were found by searching the internet for health data catalogues? If so, as defined by the FAIR principles, what depth of metadata were available and was there a persistent identifier?[20]

Data sharing

We have made data freely available online on Mendeley and Figshare including survey participant information sheet and summary notes (https://data.mendeley.com/datasets/j49bgj7nmn/1), 2018 internet search results (https://data.mendeley.com/datasets/fp9mpj3t9r/1) and 2021 search results and protocol (https://doi.org/10.48420/14791590). Original survey notes have not been shared to protect respondent confidentiality.

RESULTS

Survey findings: current practice as reported by established organisations

Of the 18 centres contacted, 12 agreed to be surveyed (table 1) and 6 did not respond. Of the 12 organisations that responded to the survey, 11 reported to share public-facing information about the available datasets (for Connected Health Cities North East and North Cumbria, a catalogue was under development at the time of the interview, now available at https://github.com/connectedhealthcities/nenc-chc). Some had different levels of access where more sensitive information was restricted to an approved audience. Metadata were provided in various ways, including through interactive catalogues (based on a number of software packages), static websites, PDFs and Excel files. The UK Data Service, Consumer Data Research Centre and Administrative Data Research Network used the DDI (Data Documentation Initiative) metadata standard[21] to describe the datasets. The other

Table 2 Description of target UK e-cohorts assessed for findability in direct and indirect searches

E-cohort	URL	Responsible organisation	Description	Number of 2018 search results: direct (indirect)	Number of 2021 search results: direct (indirect)
Clinical Practice Research Datalink	https://www.cprd.com/	MHRA (Medicines and Healthcare Regulatory Agency)/National Institute for Health Research	Primary care research dataset with linkage to additional datasets	Bing 9 (1) Google 1 (3)	Bing 14 (8) Google 13 (33)
The Health Improvement Network	https://www.cegedim-health-data.com/cegedim-health-data/thin-the-health-improvement-network	Cegedim*	Primary care research dataset	Bing 3 (0) Google 0 (4)	Bing 1 (2) Google 1 (13)
QResearch	https://www.qresearch.org/	The University of Oxford; EMIS (Egton Medical Information Systems)*	Primary care research dataset	Bing 5 (0) Google 1 (4)	Bing 0 (3) Google 2 (15)
ResearchOne	http://www.researchone.org/	TPP SystmOne*	Primary care research dataset	Bing 4 (0) Google 1 (1)	Bing 0 (1) Google 4 (6)
Consultations in Primary Care Archive	https://www.keele.ac.uk/mrr/cipcadatabase/	Keele University	Primary care research dataset	Bing 0 (1) Google 0 (1)	Bing 0 (0) Google 0 (0)
Hospital Episode Statistics	https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics	National Health Service (NHS) Digital	Secondary care dataset	Bing 7 (3) Google 0 (4)	Bing 6 (3) Google 1 (8)
Salford Integrated Record	http://www.salfordccg.nhs.uk/download.cfm?doc=docm93jijm4n524.pdf&ver=680	Salford Royal NHS Foundation Trust	Integrated primary and secondary care dataset	Bing 0 (0) Google 0 (1)	Bing 0 (0) Google 0 (0)
Prescribing Information System	https://www.ndc.scot.nhs.uk/National-Datasets/data.asp?SubID=102	NHS Scotland	National prescribing dataset	Bing 0 (1) Google 0 (1)	Bing 0 (1) Google 0 (0)
SAIL databank	https://saildatabank.com	Swansea University; NHS Wales; Health and Care Research Wales	Linked health and other routinely collected datasets	Bing 4 (0) Google 1 (6)	Bing 1 (6) Google 3 (11)
NHS Lothian Research Safe Haven/The University of Edinburgh	https://www.accord.scot/researcher-access-research-data-nrs-safe-haven/safe-haven-network	NHS Lothian, University of Edinburgh, Edinburgh Napier University, Queen Margaret University	Service coordinating access to linked health datasets across the Lothian region, Scotland	Bing 0 (2) Google 1 (1)	Bing 0 (0) Google 0 (0)
Grampian Data Safe Haven	https://www.abdn.ac.uk/iahs/facilities/grampian-data-safe-haven.php	NHS Grampian and the University of Aberdeen	Service coordinating access to linked health datasets across the Grampian region, Scotland	Bing 0 (1) Google 1 (2)	Bing 0 (0) Google 0 (0)
Health Informatics Centre—Trusted Research Environment (University of Dundee)	https://www.dundee.ac.uk/hic/hicsafehaven	University of Dundee	Service coordinating access to linked health datasets across the Tayside region, Scotland	Bing 2 (1) Google 2 (1)	Bing 1 (0) Google 0 (0)
NHS Greater Glasgow and Clyde Safe Haven	https://www.nhsggc.org.uk/about-us/professional-support-sites/nhsggc-safe-haven	NHS Greater Glasgow and Clyde and the Robertson Centre for Biostatistics, University of Glasgow	Service coordinating access to linked health datasets across the Greater Glasgow and Clyde region, Scotland	Bing 2 (2) Google 1 (1)	Bing 0 (0) Google 0 (0)

Most of the organisations are public sector.
*Commercial organisations.

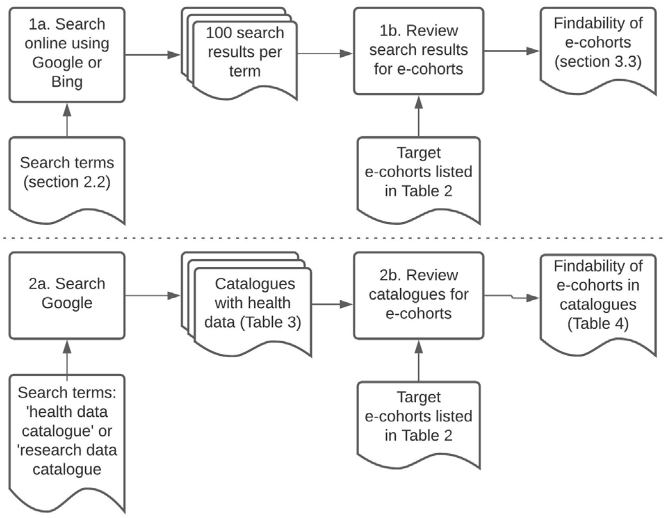
1a. Search online using Google or Bing
100 search results per term
1b. Review search results for e-cohorts
Findability of e-cohorts (section 3.3)
Search terms (section 2.2)
Target e-cohorts listed in Table 2
2a. Search Google
Catalogues with health data (Table 3)
2b. Review catalogues for e-cohorts
Findability of e-cohorts in catalogues (Table 4)
Search terms: 'health data catalogue' or 'research data catalogue
Target e-cohorts listed in Table 2

Figure 1 Internet search process—looking for health datasets via two popular, general search engines (1) and via catalogues (2).

nine organisations did not use a standard metadata schema.

Respondents talked about many other means of increasing findability, including using social media, newsletters, scientific articles and conference presentations to publicise their datasets. They were also interested in finding out what researchers wanted; three used Google Analytics to understand what people were looking for and others described discussions with researchers to better understand their needs. One organisation described a more proactive approach, using calls for expressions of interest to find and support researchers interested in using their data. For further details on approaches to findability, see the supplementary files available on Mendeley Data (https://doi.org/10.17632/j49bgj7nmn.1).

Perceived challenges to findability according to established organisations

Respondents were also asked for perceived barriers to data findability. This prompted a broad range of responses, which are summarised below and detailed in the supplementary files available on Mendeley Data (https://doi.org/10.17632/j49bgj7nmn.1). Issues include: datasets submitted with poor quality metadata, no widely adopted metadata standards or cataloguing technologies. Shortages in expertise and time were also cited, as was the view that data providers and funders did not prioritise curation of metadata and that the role of data curators is underappreciated. Many respondents recognised that more support was needed to curate good quality metadata. The challenges of dealing with the inherent variability of routinely collected health data for both curators and researchers and lack of appropriate metadata standards for health data were also raised.

When asked about plans to improve findability, respondents covered topics as diverse as making better use of existing web tools (cited most often), improving metadata quality, offering more support to research users and overlapping with other developments in the repository operations such as data linkage or migration (cited least often). Some organisations reported actively exploring new tools to replace their existing catalogues. Respondents highlighted that a good catalogue needs to contain entries for a wide range of datasets and have a usable search tool, developed with an understanding of researchers' needs.

Findability of target e-cohorts and data organisations using general internet search engines

Internet searches in 2018 found direct links to the websites of 8 of the 13 target e-cohorts listed in table 2. When clicking on links within each search result, all 13 targets were indirectly findable. For further details see the supplementary files available on Mendeley Data (https://doi.org/10.17632/fp9mpj3t9r.1).

In 2021, there were direct links to 7 of the 13 target e-cohorts listed in table 2, but when clicking on links within each search result 8 were indirectly findable. See supplementary files available on Figshare (https://doi.org/10.48420/14791590).

Findability of target e-cohorts and data organisations using health data catalogues

In 2018 we identified nine catalogues of UK-based e-cohorts through internet searches (table 3). Six catalogues referred to 1 or more of the 13 target e-cohorts listed in table 2, while 3 catalogues did not reference any of the targets. In 2021 two of those nine catalogues were inaccessible, and, among the remaining seven catalogues, one listed more target e-cohorts (from one in 2018 to four in 2021).

In 2018 all the catalogues included dataset-level metadata (descriptive, structural or administrative metadata about the dataset). The Health Data Finder, particular entries in the NHS England Data Catalogue, the Perinatal Mental Health (published by Public Health England) and Social Services Improvement Agency Data Catalogue had field-level metadata (descriptive, structural or administrative metadata held at the level of individual fields). None of the catalogues attached DOIs to their entries. The results are summarised in table 4. In 2021, among the seven catalogues still accessible, their findability in terms of metadata detail and identifiers was unchanged. Nine additional catalogues were found in the searches in 2021, seven of which included persistent identifiers but not always field level metadata and only two included target e-cohorts.

Findability of target e-cohorts and data organisations in 2021 using Google dataset search

Using the Google dataset search, all but 1 of our 10 searches produced over 100 results (searching for 'tertiary care research datasets' only produced 30 results). Among all available search results up to 100, 3 of the 13 target datasets were found once (HES, CPRD and SAIL).

Table 3 Catalogues of UK-based e-cohorts found through general internet search engines in 2018 and the number of target e-cohorts within them in 2018 and 2021

Catalogue	Web link (correct in March 2018 at the time of searching)	Number of targets found (2018)	Number of targets found (2021)
Health Data Finder for Research	http://www.hdf.nihr.ac.uk/	2	NA
Children and young people's health data catalogue 2009	http://www.childhealthresearch.eu/research/add-knowledge/Health/Data/Catalogue__2.pdf/at_download/file	0	NA
NHS Digital: Data and information	https://digital.nhs.uk/data-and-information/	1	1
Perinatal mental health: national datasets	https://www.gov.uk/government/publications/perinatal-mental-health-national-datasets (also linked to https://fingertips.phe.org.uk/profile-group/mental-health/profile/perinatal-mental-health)	1	1
NHS England Data Catalogue	https://data.england.nhs.uk/dataset	1	1
National Data Catalogue Scotland	http://www.ndc.scot.nhs.uk/	1	1
Asthma UK Data Catalogue	https://www.aukcar.ac.uk/asthma-observatory/data-catalogue	1	5
Urban Big Data Centre Health and social care data	http://ubdc.ac.uk/data-services/data-catalogue/health-and-social-care-data/	0	0
Social Services Improvement Agency Data Catalogue	http://www.dataunitwales.gov.uk/SharedFiles/Download.aspx?pageid=30&mid=64&fileid=22	0	0

Catalogues no longer accessible in 2021 are marked as NA.

DISCUSSION

We sought to understand how easily a person could discover e-cohorts from the UK via internet search engines. We used a telephone survey to understand how organisations try to make data findable and measured how findable e-cohorts were across two internet search engines. In our survey, findability was recognised as valuable, however those managing e-cohorts were still exploring how to harness the power of the internet to improve findability. Using internet search engines, we found a wide range of e-cohorts and catalogues, but between 2018 and 2021 neither the findability of target e-cohorts in the top 100 results nor in catalogues had improved. If anything, findability had decreased slightly. Target e-cohorts were less findable using a new, dedicated dataset search than a general internet search engine. While established national e-cohorts were found directly through search engines, several catalogues and smaller, local or specialist e-cohorts were only found indirectly through other webpages. A crucial factor appears to be the coverage of e-cohorts listed in catalogues or specialist search tools.

Many authors have argued for improved findability, but empirical studies to assess findability have been rare and have not previously been done for UK health data. In the FAIR principles,[16] findability requires that datasets have a globally unique and persistent identifier, are described with rich metadata which explicitly include that identifier and are registered or indexed in a searchable web catalogue. In the UK, there have been government-commissioned reports into how FAIR research information is, which recognised the importance of a sector-specific approach but said little about health and did not measure findability.[22] Wilkinson *et al* proposed a set of metrics and a design framework for a FAIRness assessment[23] and this framework has been applied to omics data.[24] That assessment takes a machine-led approach, that is, whether a dataset is findable, accessible, interoperable and reusable without human intervention. We took an alternative starting point, assessing findability using the searches that might be carried out by a person trying to find e-cohorts. The importance of the public internet in providing search engines that index metadata to make data findable has been recognised,[25] although others have highlighted challenges to implementing the FAIR principles for online searches.[26] Such publications describe and debate what findability is or should be, but they do not offer an empirical assessment of findability and their claims that improving findability for machines will improve findability for humans are untested. A toolkit was published in 2019[27] that includes at least three metrics of whether or how easily datasets and other resources can be found using internet searches[28]; our methods fall in this vein. Looking back to just before our first online searches, a

Table 4 Assessment of findability within catalogues, including whether the catalogue listed target e-cohorts from table 2 (see figure 1)

	Catalogue name	Target e-cohorts listed	Searchability	Metadata	Unique and persistent identifier
Found in 2018 but not in 2021	Health Data Finder for Research	Clinical Practice Research Datalink (CPRD) Hospital Episode Statistics (HES)	Can filter	Dataset and field level	No
	Children and young people's health data catalogue 2009	–	Downloadable file	Dataset level	No
Found in 2018 and 2021	NHS Digital: Data and information	HES	Search bar; Can filter	Dataset level	No
	NHS England Data Catalogue	HES	Search bar; Can filter	Dataset and field level	No
	Perinatal mental health: national datasets	HES	Downloadable file	Dataset and field level	No
	Asthma UK Data Catalogue	HES In addition in 2021: SAIL QResearch Clinical Practice Research Datalink (CPRD) PIS	Search bar; Dropdown list	Dataset level	No
	Urban Big Data Centre Health and social care data	–	Dropdown list	Dataset level	No
	Social Services Improvement Agency Data Catalogue	–	Downloadable file	Dataset and field level	No
	National Data Catalogue Scotland	PIS	a-z listing	Dataset level	No
Not found in 2018, found in 2021	DataCat (University of Liverpool)	–	Search bar; Can filter	Dataset level	Yes
	ORDA (University of Sheffield)	–	Search bar; Can filter	Dataset level	Yes
	UK Data Archive	–	Search bar; Can filter	Dataset level	Yes
	University of Lancaster	–	Search bar; Can filter	Dataset level	Yes
	Mauro Data Mapper/ Oxford Metadata Catalogue	–	Dropdown list	Dataset and field level	Yes
	Zenodo	–	Search bar; Can filter	Dataset level	Yes
	Health Innovation Gateway	CPRD PIS SAIL HES Grampian	Search bar; Can filter; Dropdowns; Highlight new datasets	Dataset level	Yes
	Social Care Wales	–	Search bar; filter; show all	Dataset level	No
	ONS Secure Research Service	HES	Spreadsheet	Dataset level	No

For catalogues found in 2018, these were revisited in 2021; two were inaccessible, the other eight were unchanged in terms of metadata detail and presence of identifiers.

paper from 2016 envisaged a community to advance the FAIR principles (including searchability) in the life sciences,[29] and in 2017 researchers highlighted the need for better web-based identifiers for life sciences datasets[30] and for improved online discoverability and standardisation for UK health data.[31] Our 2021 results show many of those lessons still need to be heeded.

Our finding that some regional e-cohorts had by 2021 become less findable than national counterparts and that some catalogues had become inaccessible has implications for those working to increase data findability. Community efforts and standardisation have been advocated by researchers as the best way to implement the FAIR principles.[32] One approach has been to collate metadata centrally, as was done recently for opthalmology.[33] Centralised repositories and dedicated data search tools may be increasingly important for fostering findability as more and more datasets are described online, however we found that not all available datasets are currently listed. Search engines, which are increasingly embedded into catalogues as well as being available for the general internet searches we conducted, enhance the findability of some datasets more than others. For example, CPRD was the most findable of our target e-cohorts in 2018 and 2021 and even increased its presence in search results, while some other target e-cohorts became less findable. As well as creating hubs, we suggest that the health data community also discusses variability in the findability of datasets and use benchmarks for online findability to assess progress.

A large effort as a result of the COVID-19 pandemic has given momentum to new findability tools, such as Health Data Research UK with their new catalogue: the Innovation Gateway.[34] COVID-19 data were listed in the catalogue and already found in our 2021 searches. The pace and scale of these developments, which are already producing research insights, are impressive. This may be helped by a more coordinated effort in the NHS under the UK government's data strategy.[35] Such efforts need continued support to enhance coverage, for example, to include more of our target e-cohorts or newer e-cohorts such as OpenSAFELY[4] and to boost metadata quality and accessibility.

Our work has some limitations. First, although we tried to contact as many organisations as possible across the UK, not all the ones we contacted were able to participate, and we may have missed some others. We can only speculate on how this has affected our results; it is possible that organisations that did not respond are stretched and chose to prioritise other work over our survey into findability. Second, our prior knowledge of the target e-cohorts probably made it easier for us to find them. Third, when screening search results, we reviewed 100 results per search (approximately 10 pages), two or three pages might be more realistic. We may therefore have overestimated the findability of UK e-cohorts. Fourth, the proprietary nature of search engines makes their operations unclear, for example, the consistency of the search rankings among different users[36] or how algorithms may have altered findability between 2018 and 2021. Google and Bing limit automated processing of their search tool[26] and manually checking 100 results per search was time intensive.

There are opportunities to extend our approach in further research. It would be useful to study how researchers find and access e-cohorts in practice. The use of wildcards to make searches more flexible, analysis of rankings and use of other search engines could be adopted in future. Comparison across organisations of the investment (time, money) and competencies of personnel working to make e-cohorts findable and accessible could reveal the most efficient methods to inform successful strategies for improving findability.

Based on our findings, we recommend that UK e-cohorts implement the following features to improve their findability: create a unique and persistent identifier, have richer metadata descriptions and ensure they are indexed in a searchable resource either through search engine optimisation of their own website or through catalogues that are highly ranked by search engines.

Twitter Emily Griffiths @emble64, George Tilston @Tilstongeorge and Niels Peek @NielsPeek

Acknowledgements The authors wish to thank Ben Green, Sarah Al-Adely and Will Hulme for their advice in the planning and data collection phases of this project.

Contributors EG, RMJ, NP, ZK and WD designed the study. ST made substantial contribution to summary and analysis of the survey information. GT made substantial contribution to the collection of 2021 search results. EG acted as guarantor. All authors contributed to drafting the paper and providing critical comments, approved the final submission and are accountable for this work.

Funding NP's work was partially funded by the National Institute for Health Research (NIHR) Greater Manchester Patient Safety Translational Research Centre and the NIHR Manchester Biomedical Research Centre. All the authors except GT were supported by Connected Health Cities, which was a Northern Health Science Alliance (NHSA) led programme funded by the Department of Health and delivered by a consortium of academic and NHS organisations across the north of England. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR, the Department of Health or NHSA.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval The University of Manchester Research Ethics Committee does not require approval for studies in which participants are interviewed on subjects within their professional competence, provided that participants are contacted using publicly available details and are provided with an information sheet and data are not attributable to individuals. Those who participated had the opportunity wherever possible to review the manuscript before submission.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available in a public, open access repository. Data are available freely online, including internet search results and summary survey notes. Original survey notes have not been shared to protect respondent confidentiality. Mendeley Data—internet search results: https://data.mendeley.com/datasets/fp9mpj3t9r/1. Mendeley Data—semistructured interviews: https://data.mendeley.com/datasets/j49bgj7nmn/1. Additional data following reviewer suggestions are also available via Figshare: https://doi.org/10.48420/14791590.

ORCID iDs
Emily Griffiths http://orcid.org/0000-0001-7603-3552
Rebecca M Joseph http://orcid.org/0000-0002-0147-0712
William Dixon http://orcid.org/0000-0001-5881-4857
Niels Peek http://orcid.org/0000-0002-6393-9969

REFERENCES

1 Chen Y-C, Wu J-C, Haschler I, *et al*. Academic impact of a public electronic health database: bibliometric analysis of studies using the general practice research database. *PLoS One* 2011;6:e21404.
2 Chaudhry Z, Mannan F, Gibson-White A, *et al*. Outputs and growth of primary care databases in the United Kingdom: bibliometric analysis. *J Innov Health Inform* 2017;24:942.
3 Vezyridis P, Timmons S. Evolution of primary care databases in UK: a scientometric analysis of research output. *BMJ Open* 2016;6:e012785.
4 OpenSAFELY. Available: https://opensafely.org/ [Accessed 01 Dec 2020].
5 Hyatt M, Rodgers SE, Paranjothy S, *et al*. The Wales electronic cohort for children (WECC) study. *Arch Dis Child Fetal Neonatal Ed* 2011;96:Fa18.

6 Gabbe BJ, Brooks C, Demmler JC, *et al*. The association between hospitalisation for childhood head injury and academic performance: evidence from a population e-cohort study. *J Epidemiol Community Health* 2014;68:466–70.
7 United Kingdom Department of Health and Social Care. Matt Hancock announces ambition to map 5 million genomes - GOV. UK, 2018. Available: https://www.gov.uk/government/news/matt-hancock-announces-ambition-to-map-5-million-genomes
8 Figueiredo AS. Data sharing: convert challenges into opportunities. *Front Public Health* 2017;5:327.
9 Chan A-W, Song F, Vickers A, *et al*. Increasing value and reducing waste: addressing inaccessible research. *Lancet* 2014;383:257–66.
10 CPRD. Clinical practice research Datalink | CPRD, 2021. Available: https://www.cprd.com/
11 NHS Digital. Hospital episode statistics (hES). Available: https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics
12 SAIL Databank. SAIL Databank - The Secure Anonymised Information Linkage Databank, 2021. Available: https://saildatabank.com/
13 Salford Clinical Commissioning Group. Sharing patient information locally Salford integrated record. Available: http://www.salfordccg.nhs.uk/download.cfm?doc=docm93jijm4n524.pdf&ver=680 [Accessed 18 Jan 2019].
14 New JP, Leather D, Bakerly ND, *et al*. Putting patients in control of data from electronic health records. *BMJ* 2018;360:j5554.
15 Tenopir C, Allard S, Douglass K, *et al*. Data sharing by scientists: practices and perceptions. *PLoS One* 2011;6:e21101.
16 Wilkinson MD, Dumontier M, Aalbersberg IJJ, *et al*. The fair guiding principles for scientific data management and stewardship. *Sci Data* 2016;3:160018.
17 The Farr Institute. Farr Institute | the farR Institute of health informatics research. Available: https://twitter.com/farrinstitute
18 Secure Data Group. Safe data access professionals, 2021. Available: https://securedatagroup.org/
19 HDRUK. Health data research UK | HDR UK, 2021. Available: https://www.hdruk.ac.uk/
20 Paskin N. Toward unique identifiers. *Proc IEEE Inst Electr Electron Eng* 1999;87:1208–27.
21 DDI Alliance. Welcome to the data documentation initiative | data documentation initiative. Available: https://www.ddialliance.org/
22 Open Research Data Taskforce. Realising the potential: final report of the open research data Task force, 2018. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/775006/Realising-the-potential-ORDTF-July-2018.pdf
23 Wilkinson MD, Sansone S-A, Schultes E, *et al*. A design framework and exemplar metrics for fairness. *Sci Data* 2018;5:180118.
24 Berrios DC, Beheshti A, Costes SV. Fairness and usability for open-access omics data systems. *AMIA Annu Symp Proc* 2018;2018:232–41.
25 Mons B. FAIR Science for Social Machines: Let's Share Metadata Knowlets in the Internet of FAIR Data and Services. *Data Intell* 2019;1:22–42.
26 Jacobsen A, de Miranda Azevedo R, Juty N, *et al*. Fair principles: interpretations and implementation considerations. *Data Intell* 2020;2:10–29.
27 Clarke DJB, Wang L, Jones A, *et al*. FAIRshake: toolkit to evaluate the fairness of research digital resources. *Cell Syst* 2019;9:417–21.
28 Team Nitrogen. FAIRshake. Available: https://fairshake.cloud/?q=search&metrics=1
29 McQuilton P, Gonzalez-Beltran A, Rocca-Serra P, *et al*. BioSharing: curated and crowd-sourced metadata Standards, databases and data policies in the life sciences. *Database* 2016;2016:baw075.
30 McMurry JA, Juty N, Blomberg N, *et al*. Identifiers for the 21st century: how to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data. *PLoS Biol* 2017;15:e2001414.
31 Salazzo D, Miller D. Open data in the health sector: users, stories, products and recommendations. open healthcare, 2017. Available: https://openhealthcare.org.uk/open-data-in-the-health-sector/
32 Sansone S-A, McQuilton P, Rocca-Serra P, *et al*. FAIRsharing as a community approach to Standards, repositories and policies. *Nat Biotechnol* 2019;37:358–67.
33 Khan SM, Liu X, Nath S, *et al*. A global review of publicly available datasets for Ophthalmological imaging: barriers to access, usability, and generalisability. *Lancet Digit Health* 2021;3:e51–66.
34 HDRUK. HDRUK innovation gateway | Homepage, 2020. Available: https://www.healthdatagateway.org/
35 UK Department for Digital, Culture, Media, and Sport. National data strategy, 2020. Available: https://www.gov.uk/government/publications/uk-national-data-strategy/national-data-strategy
36 Levene M. *An Introduction to Search Engines and Web Navigation. Wiley & Sons*. 2nd edition, 2014. https://ebookcentral.proquest.com/lib/manchester/reader.action?docID=573905