International Journal of Population Data Science

Journal Website: www.ijpds.org

What makes administrative data "research-ready"? A systematic review and thematic analysis of published literature

Louise Mc Grath-Lone[1,*], Matthew A. Jay[2], Ruth Blackburn[1], Emma Gordon[3], Ania Zylbersztejn[2], Linda Wiljaars[2], and Ruth Gilbert[1,2]

Submission History

Submitted:	29/11/2021
Accepted:	8/02/2022
Published:	27/04/2022

[1]Institute of Health Informatics, University College London, UK
[2]Population, Policy and Practice Research and Teaching Department, Great Ormond Street Institute of Child Health, University College London, UK
[3]Administrative Data Research UK, Economic & Social Research Council, UK

Abstract

Introduction
Administrative data are a valuable research resource, but are under-utilised in the UK due to governance, technical and other barriers (e.g., the time and effort taken to gain secure data access). In recent years, there has been considerable government investment in making administrative data "research-ready", but there is no definition of what this term means. A common understanding of what constitutes research-ready administrative data is needed to establish clear principles and frameworks for their development and the realisation of their full research potential.

Objective
To define the characteristics of research-ready administrative data based on a systematic review and synthesis of existing literature.

Methods
On 29th June 2021, we systematically searched seven electronic databases for (1) peer-reviewed literature (2) related to research-ready administrative data (3) written in the English language. Following supplementary searches and snowball screening, we conducted a thematic analysis of the identified relevant literature.

Results
Overall, we screened 2,375 records and identified 38 relevant studies published between 2012 and 2021. Most related to administrative data from the UK and US and particularly to health data. The term research-ready was used inconsistently in the literature and there was some conflation with the concept of data being ready for statistical analysis. From the thematic analysis, we identified five defining characteristics of research-ready administrative data: (a) accessible, (b) broad, (c) curated, (d) documented and (e) enhanced for research purposes.

Conclusions
Our proposed characteristics of research-ready administrative data could act as a starting point to help data owners and researchers develop common principles and standards. In the more immediate term, the proposed characteristics are a useful framework for cataloguing existing research-ready administrative databases and relevant resources that can support their development.

Keywords
administrative data; research-ready; systematic review; thematic analysis

*Corresponding Author:
Email Address: l.mcgrath-lone@ucl.ac.uk (Louise Mc Grath-Lone)

https://doi.org/10.23889/ijpds.v7i1.1718

Introduction

Administrative data (i.e. information that is routinely collected by organisations for operational reasons) are a valuable resource for research in fields such as health, education, justice and social care [1]. One advantage of administrative data is that they are comprehensive, often involving a whole country, and therefore less affected by biases related to participation and reporting than primary data that are collected as part of a research study (e.g., through surveys and interviews) [2]. Administrative data also tend to have larger sample sizes than primary data which provides an opportunity to explore rare events and small sub-groups that may otherwise be difficult or unfeasible [3]. Longitudinal administrative data linked across different domains are particularly powerful as a research resource as they allow pathways throughout the life course to be examined and provide insights into how circumstances, experiences and opportunities are inter-related [1, 4]. In the UK, the re-use of administrative data for research purposes is broadly supported by the general public [5, 6]. Indeed, the Digital Economy Act 2017 includes specific provision for making de-identified administrative (non-health) data held by public authorities available for research purposes [7]. Yet, compared to other countries such as Sweden, administrative data are under-utilised as a research resource in the UK. This is due in part to technical and governance barriers which limit their use; for example, administrative data are often difficult and time-consuming to access [8, 9], costly [10], and can require extensive processing, cleaning and preparation before they are analysed.

There has been considerable financial investment by the UK government in administrative data research projects and initiatives in recent years [11]. One such example is the £90 million investment in the Administrative Data Research (ADR) UK programme (funded via the Economic and Social Research Council, part of UK Research and Innovation) which aims to create wider recognition of the potential benefits of research using administrative data across government, academia and society [12]. The ADR UK programme includes a dedicated Research-Ready Data Fund to enable the creation of novel, high-quality, long-term, linked administrative databases for research that can be re-used by multiple users [1]. This move towards long-term databases represents an important shift in the administrative data research landscape in the UK [13]. Previously, administrative data research tended to operate on a 'create and destroy' model, whereby data owners prepared bespoke extracts of anonymised administrative data for specific research projects, which were then destroyed at the end of the project cycle. Long-term, research-ready, administrative databases represent a unique opportunity to reduce duplication of effort for administrative data owners and researchers (e.g., in terms of negotiating access permissions and governance arrangements) leading to less waste of resources and greater use of data for the public benefit [13].

The increased use of Trusted Research Environments (TREs) in the UK to access administrative data for research purposes also represents a unique opportunity to develop ongoing research-ready databases. Previously, administrative data owners would securely transfer extracts of anonymised data to individual researchers who were responsible for their storage, analysis and destruction. This data release model of access meant that data owners needed to prepare multiple bespoke data extracts which is costly and inefficient [14] and researchers were duplicating efforts in terms of cleaning and preparation. It also meant that researchers would often encounter barriers to accessing the data as they were required to meet technological and governance standards specified by data owners. In recent years, there has been a move towards data owners facilitating access to administrative data via TREs, such as the Office for National Statistics Secure Research Service. This centralised model of data access presents an opportunity for data owners to make extracts of administrative data available securely to multiple researchers and for researchers to contribute to the development and enhancement of this data as an ongoing research resource.

Despite the increased interest and investment in making administrative data research-ready, there is no agreed definition of what this term means. A common understanding of what constitutes research-ready administrative data is a pre-requisite to establishing clear principles and frameworks which are needed to develop these long-term data resources and to realise their full research potential [6]. The aim of this study was to systematically review the available literature related to research-ready administrative data with a view to proposing an initial set of defining characteristics. The proposed characteristics could act as a starting point to initiate and frame discussions between data owners and researchers about what makes administrative data research-ready with the view to developing common principles and standards.

Methods

We carried out a systematic review of published, peer-reviewed literature related to research-ready administrative data with the aim of identifying a set of key characteristics that define an administrative dataset as being research-ready. The review protocol was pre-registered on Open Science Framework on 29$^{\text{th}}$ June 2021 and is publicly available [15]. There were no deviations from the pre-registered protocol.

Eligibility criteria

This review considered published literature that 1) related to research-ready administrative data, which was broadly defined as information collected for operational purposes by an organisation (commercial or non-commercial), including health data. Studies related to data collected specifically for research purposes (e.g., surveys) were not eligible for inclusion. The review was restricted to literature that was 2) written in English and 3) peer-reviewed (e.g., conference abstracts, letters, commentaries etc. were not eligible). There were no restrictions imposed in terms of study design, population, setting, timeframe or publication date. The minimum number of eligible studies required for data synthesis was pre-specified as two.

Information sources

On 29$^{\text{th}}$ June 2021, we searched seven electronic databases (Embase (via Ovid), Medline (via Ovid), Pubmed, Scopus, ProQuest Central, CINAHL, Web of Science Core Collections).

On 7th July 2021, we also carried out supplementary searches of Google Scholar (first 100 entries only) and the International Journal of Population Data Science (IJPDS) website, a subject-relevant journal that was only recently indexed in electronic databases [16] and therefore not fully accessible through the pre-specified electronic databases included in the main search.

Search strategy

The SPIDER framework (Sample, Phenomenon of Interest, Design, Evaluation and Research type) was used to develop the search strategy and select relevant search terms. Search strategies were applied to all fields, included notation to allow for differences in the spelling of search terms and used combination of search terms using Boolean operators, where possible. For example, in PubMed the search string was: ("research#ready") AND (administrative OR operational OR "routinely#collected" OR records OR data?). Supplementary Table 1 describes the exact search strings used for each database.

Selection process

Records from the main and supplementary searches were exported to Mendeley (reference management software) and automatically de-duplicated. All titles and abstracts were then screened for inclusion independently by two reviewers (LMcGL and MAJ) using the pre-specified eligibility criteria. Discrepancies between reviewers were resolved by discussion. Full texts of all eligible studies were retrieved and then screened for inclusion independently by two reviewers (LMcGL and MAJ) with any discrepancies resolved by discussion. Snowball screening of all included full texts was conducted by one reviewer (LMcGL) on 31st July 2021 and included all cited (backwards) and citing articles (forwards).

Data collection process and data items

The following descriptive data from the included studies were extracted into Excel by one reviewer (LMcGL): country of data collection, domain of administrative data, name and details of research-ready administrative data, first author, year of publication and purpose of the study. Data extraction was checked by a second reviewer (AZ or LW).

Synthesis methods

One reviewer (LMcGL) carried out a thematic analysis of aggregate findings from across all the identified studies [17]. All publications were read repeatedly and a set of initial codes were generated using NVivo. These codes were then refined and grouped into themes which were tested for validity through one-to-one discussions between LMcGL and the co-authors who are experienced in the field of administrative data research. LMcGL then carried out a narrative synthesis of the identified themes in relation to the pre-specified review question: What makes administrative data research-ready? All co-authors contributed to the final narrative summary (as presented in the Results).

Reporting bias and certainty assessment

When preparing the narrative synthesis, potential bias that may be introduced by the range of included studies was considered, particularly the domain of administrative data and their country of origin. Because this was an exploratory study that aimed to propose a set of key characteristics of research-ready administrative data as a starting point for discussion and consensus building in the wider community, there was no assessment of the strength or quality of the body of evidence conducted as part of this review.

Results

Study selection

The study selection process is summarised in Figure 1. In the main search, 622 records were retrieved from the seven included databases. After de-duplication, 464 titles and abstracts were screened and 384 were excluded. Full texts of the remaining 80 sources were retrieved for further screening. The supplementary search of Google Scholar and the IJPDS website yielded 50 records that had not previously been identified in the main search. Based on title and abstract screening, 40 were excluded and the full texts of the remaining 10 sources were retrieved. Overall, 58 of the 90 full-texts retrieved were excluded. Snowball screening was carried out on the 32 included full-texts. Following backward and forward screening of 953 cited and 908 citing articles, a further 6 eligible publications were identified. In total, 2,375 records were screened and 38 relevant publications related to research-ready administrative data were identified and included in the thematic analysis.

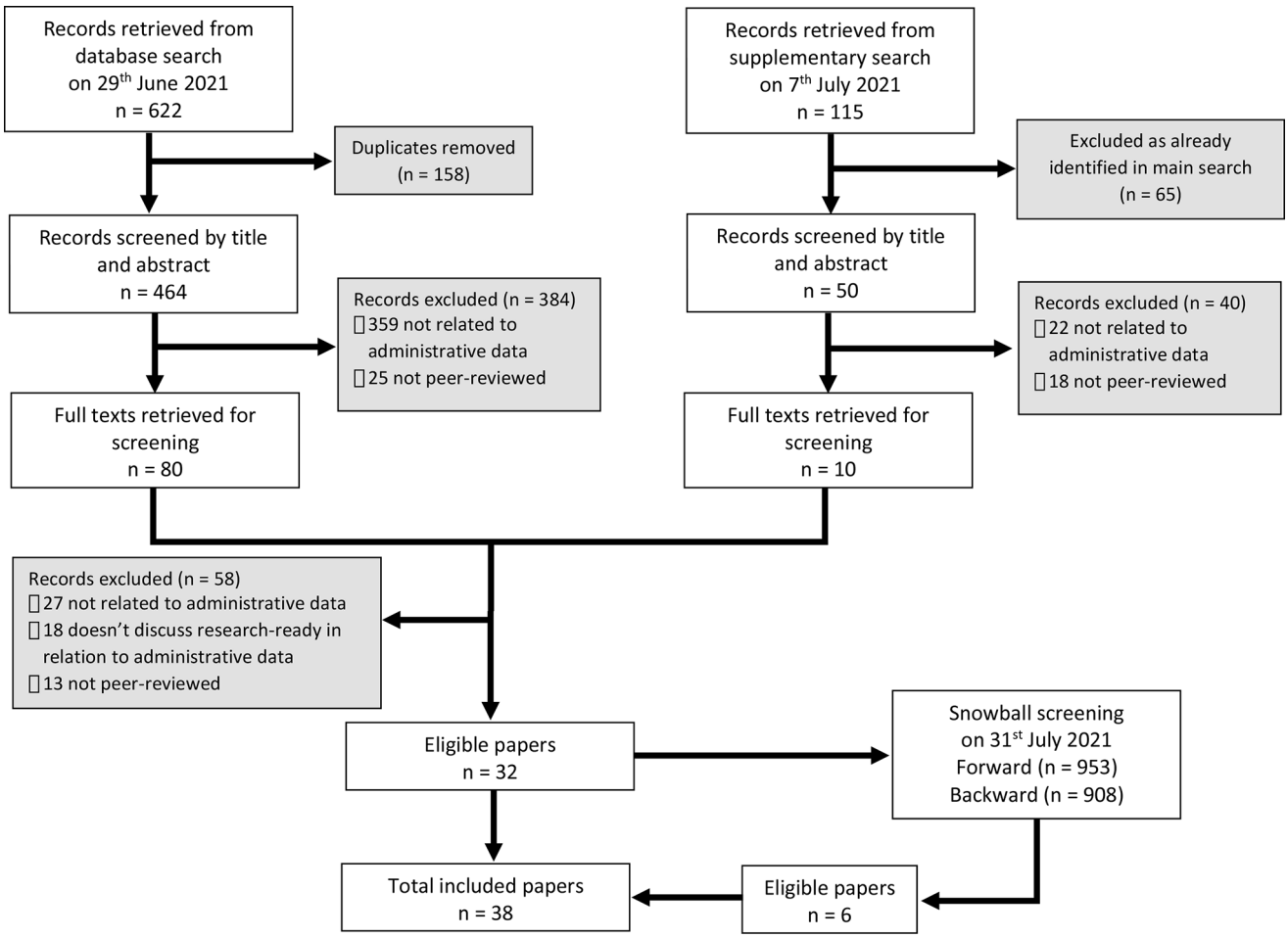
Study characteristics

Key characteristics of the included studies [18–54] are summarised in Table 1. The included studies were published between 2012 and 2020 and related to administrative data in the UK ($n = 12$), US ($n = 11$), Australia ($n = 7$), Canada ($n = 4$), Brazil ($n = 1$), China ($n = 1$), New Zealand ($n = 1$) and Taiwan ($n = 1$). The majority of studies related to administrative health data and to government or public sector data, but there were also some examples from private organisations, including online supermarkets, professional sports organisations and private health insurers. The majority of publications had the primary aim of profiling or describing the development of research-ready administrative data sources, infrastructure and research tools ($n = 23$).

Narrative summary of thematic analysis

Looking across the body of included literature, it was evident that there are differences in researchers' understanding of the term "research-ready" based on the varied range of data that were described as such. In some studies, only administrative data that was ready for analysis or query was considered research-ready. For example, the UK National Joint Registry described their research-ready data source as a "pre-cleaned dataset ready to 'plug and play'...[which] reduce[s] the burden on researchers by providing a single, 'clean' source of data" [47]. Similarly, Hilder

Figure 1: Flow diagram of study selection



et al. described linked administrative health and justice data as not research-ready because "additional effort is needed on the part of researchers to validate and prepare the data for epidemiological analysis" [45]. In contrast, other studies considered administrative data that required further processing and preparation before they could be queried or analysed as research-ready. For example, the Critical Care Health Informatics Collaborative (CCHIC) dataset was described as "a 'warts and all' version" of electronic health records [44]. Other research-ready datasets were described as containing known duplicate records, errors and unstandardised variables that required further decision-making and processing by researchers before they could be analysed [20, 46]. The heterogeneity of data described as research-ready identified in this review highlights that there is an important and crucial difference between making administrative data research-ready for broad purposes and making it analysis-ready for a specific research question.

Despite the lack of a common understanding of the term 'research-ready' in the literature, five key characteristics of research-ready administrative data emerged from our thematic analysis (Figure 2).

Accessible

Administrative data are collected by organisations for internal purposes. Across the body of included literature, a common theme was that in order for administrative data to be considered research-ready, they must be accessible to researchers external to that organisation. A key aspect of accessibility is that administrative data must be findable. This could be through data owners depositing information and metadata on online repositories (e.g., UK Data Archive [55] or Health Data Research Innovation Gateway [56]) or by publishing a data profile that describes the data source.

To make administrative data accessible, it is also important that permissions and governance arrangements for their re-use for research purposes are in place. This is particularly important for linked administrative data as it can take considerable time for researchers to negotiate permissions with multiple organisations [8]. Clear procedures and prerequisites for obtaining access to data for research purposes are also needed. This requires setting standards related to who can use the data and for what purposes, as well as for how data are securely shared with researchers (e.g., only available in designated TREs).

Ensuring that research-ready administrative data are de-identified also emerged as an important factor related to accessibility. De-identification was described as contributing to the accessibility of administrative data by helping to maintain the privacy and confidentiality of individuals included in a dataset and lowering the risks associated with making it available for research purposes. All research-ready administrative data identified in our review that related to

Table 1: Summary of included publications, by country and domain of administrative data

Country	Domain	Name of research-ready administrative data source	Details of research-ready administrative data source	First author (Year)	Purpose of study
Australia	Health	Rural Acute Hospital Database Register (RAHDaR)	Longitudinal health data from 10 hospitals in South West Victoria Australia.	Kloot (2019)	To profile a data resource.
				Peck (2020)	To examine patterns of childhood injuries presenting to rural Urgent Care Centres.
				Terry (2020a)	To examine whether asthma presentations to rural hospitals are represented in national datasets.
				Terry (2020b)	To examine whether asthma presentations to rural hospitals are represented in national datasets.
				Peck (2021)	To compare socioeconomic characteristics of children with injury-related emergency presentations at a rural Urgent Care Center versus Emergency Department.
	Utilities	Visualising Victoria's Groundwater web portal	Groundwater monitoring data for the State of Victoria, Australia.	Dahlhaus (2016)	To evaluate the social impact of a data visualisation portal, including the support of decision making.
	Multiple	Mothers and Gestation in Custody (MAGIC) cohort	Linked government data related to health and justice for women in New South Wales.	Hilder (2016)	To determine pregnancy prison exposure for incarcerated women.
Brazil	Multiple	Centre for Data and Knowledge Integration for Health (CIDACS)	Several linked government databases, including health, social benefits and housing.	Barreto (2019)	To profile a data resource (including the establishment and operations of CIDACs and efforts to obtain high-quality administrative data for research).
Canada	Health	Canadian Forces Health Information System	Demographic and health data for members of the Canadian Forces, linked to census data and government water data.	Batsos (2021)	To investigate the association between municipal water fluoridation and dental health.
	Welfare	Ontario Child Abuse and Neglect Data System (OCANDS)	Provincial linked data related to child welfare.	Fallon (2017)	To promote and demonstrate the policy and practice value of analysing longitudinal administrative data related to child welfare.
	Multiple	Institute for Clinical Evaluative Sciences (ICES)	Provincial linked health and other administrative data.	Schull (2020)	To profile an administrative data research network.
				Walker (2018)	To describe the linkage of the Indian Register database to the Ontario Registered Persons Database within the context of Indigenous data sovereignty principles.
China	Health	Chinese Electronic Health Records Research in Yinzhou (CHERRY)	Regional, longitudinal health data.	Lin (2018)	To describe the methods for establishing an electronic health cohort in one region of China.
New Zealand	Health	Vascular Risk in Adult New Zealanders (VARIANZ) dataset	National, linked administrative health datasets.	Mehta (2019)	To profile a data resource.
Taiwan	Health	National Health Insurance Re-imbursement Database (NHIRD)	National, longitudinal health and prescription data.	Wang (2014)	To determine the risk of urothelial cancer associated with aristolochic acid-related Chinese herbal medicines among end-stage renal disease patients.
		Critical Care Health Informatics Collaborative (CCHIC)	Multi-centre database of longitudinal health data from adult Intensive Care Units.	Harris (2018)	To profile a data resource (including creation and curation using a set of open-source tools).
				Tissot (2020)	To evaluate the simulated performance of an algorithm for identifying patients for recruitment into a clinical trial.
		Cardiovascular disease research using linked bespoke studies and electronic records (CALIBER)	National linked health and other administrative data.	Denaxas (2012)	To profile a data resource.

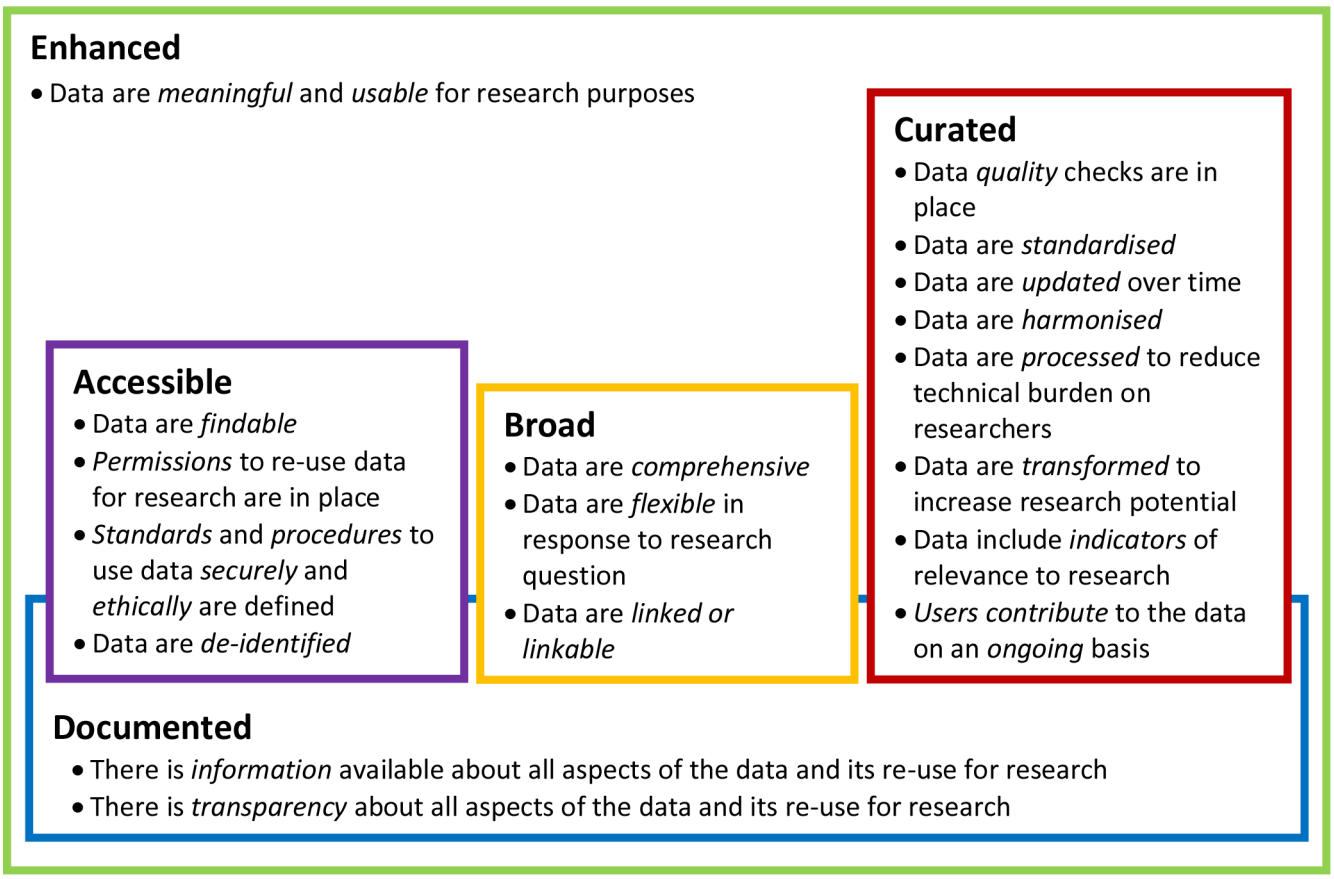
Continued

Table 1: Continued

Country	Domain	Name of research-ready administrative data source	Details of research-ready administrative data source	First author (Year)	Purpose of study
UK	Health	Farr Institute	National health data.	Hemingway (2020)	To profile a data science research institute.
		National Joint Registry	National health data related to joint replacement surgery.	Porter (2019)	To profile a data resource.
		No name specified	Research database developed by one adult Intensive Care Unit.	McWilliams (2019)	To profile a data resource (including creation and curation).
		Viral Hepatitis Central Data Repository	Multicentre database of health data.	Smith (2020)	To describe the development of a pipeline to collate electronic clinical data for viral hepatitis research.
		Multiple electronic primary care databases	National health data related to primary care.	Springate (2017)	To describe an algorithm for extracting information from electronic health records.
	Multiple	The Secure Anonymised Information Linkage (SAIL) Databank	National linked health and other administrative data.	Atkinson (2017)	To develop and validate an algorithm for determining smoking status and behaviour over the life course.
				Lyons (2021)	To describe the creation of an electronic cohort (the Wales Multimorbidity e-Cohort (WMC)).
		Administrative Data Research Northern Ireland	National linked health and other administrative data.	O'Reilly (2020)	To profile an administrative data research network.
	Food pricing	Food DB	Longitudinal data related to food and drink products available in online supermarkets.	Bhatnagar (2021)	To compare product availability, nutritional information, front-of-pack labelling, price and price promotions for food and drink products between physical and online supermarkets.
	Health	The Flatiron Health dataset	Longitudinal electronic health records data related to cancer patients from practices from different regions in the US.	Abernethy (2017)	To describe overall survival with current first-line treatment for patients presenting with stage IV non-small cell lung cancer.
				Li (2019)	To examine the management patterns and outcomes of patients with epidermal growth factor receptor tyrosine kinase inhibitors in non-small-cell lung cancer.
		Epic electronic health record platform	Electronic health records in a large tertiary paediatric healthcare system in Southern California.	Bush (2017)	To gather the perspectives of physicians and surgeons on structured data entry in electronic medical records.
US		Epic electronic health record platform	Electronic health records in a large healthcare system in Ohio.	Milinovich (2018)	To profile a data resource.
		Optum Labs database	Multicentre database of health and insurance claims data.	Wallace (2014)	To profile a data resource.
		Commercial Claims Database	Longitudinal health insurance claims from 3 of the largest insurers in the US.	Newman (2014)	To profile a data resource.
	Higher education	Database for Research on Academic Medicine (DREAM)	Undergraduate and graduate medical education outcomes data from one university.	Wilhite (2020)	To profile a data resource (including creation and curation).
	Injury surveillance	National Football League (NFL) Injury Surveillance Program	Linked health and sports-related data for professional NFL players.	Dreyer (2019)	To profile a data resource (including creation and curation).
	Patents	US Patent Assignment Dataset (UPAD) and Patent Examination Research Dataset (PatEx)	National, longitudinal, relational databases of patents applications and patents in the US.	Graham (2018a) Graham (2018b))	To profile a data resource. To profile a data resource.
		US Patent Assignment Dataset (UPAD)	National, longitudinal, relational database of patents applications and patents in the US.	Li (2020)	To analyse the effects of pharmaceutical company characteristics on the timing of drug patent purchases.

UK = United Kingdom; US = United States.

Figure 2: Key characteristics of research-ready administrative data



Italicised words/phrases are the initial codes that emerged from the thematic analysis of the included literature, which were then grouped into themes.

people, were de-identified. However, it is important to consider that the relative importance of de-identification in making data research-ready identified in this review may be influenced by the inclusion of health data in our definition of administrative data.

Based on the information provided in the included literature, it was not evident what level of "de-identification" was required for administrative data to be considered research-ready. Indeed, as data availability, linkage and analytical methods advance, it is increasingly difficult to have truly anonymised data without artificially modifying or degrading the data, which affects its utility for research. Establishing standards for de-identification of administrative data that balance the need to preserve confidentiality and maintain research utility will require further engagement with data owners and research users, as well as the public. For example, if de-identified administrative data are made available through a TRE that is accredited under a suitable process (for example, the Digital Economy Act 2017), they can be considered functionally anonymised through a combination of the actions taken to de-identify them and the secure environment in which they are accessed.

Broad

Primary data are collected with specific consent for research purposes; however, when re-using administrative data for research, the types of questions that can be answered are constrained by the legal basis underlying their collection, as well as the variables that have been collected. It is important that the scope of administrative data are kept as broad as possible to maximise their utility for research purposes. This would include making data available for all individuals in a population (not just a sample), for all years for which data are collected, and for all variables, while acknowledging the need to subsequently minimise the data that are accessed to support a specific research purpose. Data should also include as detailed information as possible as this will ensure that researchers have the flexibility to create derived variables with differing sensitivity and specificity to meet the needs of their individual research question [43]. Facilitating linkage to other data sources will also increase the breadth of administrative data and their utility for research purposes.

When establishing governance arrangements for research-ready administrative data, it is also important to keep these as broad as possible. Creating and maintaining datasets that are available for re-use for a broad range of research purposes is more sustainable and cost-effective than the traditional, project-specific 'create and destroy' model that has previously been the standard for administrative data research in the UK [1, 13]. It is also potentially less risky to allow re-use of pseudonymised linked data than to use identifiable information to re-link the same data for multiple times for different purposes.

Curated

To be research-ready, administrative data must be curated for research purposes (i.e. managed and looked after to make it more useful [57]). For example, updating a research-ready administrative dataset as more recent data become available will ensure that it remains a relevant research resource. A crucial step in curating administrative data for research purposes is ensuring that they are of good quality [1]. This may be achieved through validation checks at collection and/or post-collection processing to correct inaccuracies. These types of measures would rectify common issues researchers encounter when working with administrative datasets (such as incorrect variable formats and erroneous dates). This would reduce duplication of effort by researchers and avoid inconsistencies in cleaning these types of errors.

Other data curation activities could include standardising variables to a common data model to enable linkage with other data sources or harmonising longitudinal variables that have changed over time [13]. It could also include transforming multiple years of cross-sectional data into a longitudinal research resource or deriving additional variables relevant to research, such as indicators for groups, states or phenotypes of interest. However, this would require substantial time and resource investment and it is not feasible for data providers to anticipate *a priori* the range of transformations, derived variables and other processing that researchers may require to address their specific research questions. A collaborative approach to data curation that allows data users to contribute to the development of research-ready administrative data will be more sustainable. This will require standardised mechanisms for knowledge to be fed back into a dataset by research users, thereby improving and expanding its utility as a long-term research resource. For example, data providers could encourage users to publish protocols, research outputs and data insights or to submit code for deriving new variables for peer review with a view to incorporating these into the research-ready dataset (as per the CALIBER dataset [43] for example).

Ongoing curation of research-ready datasets, and the publication of the curation processes, will promote bodies of knowledge being built up around the datasets and allow researchers to build on what has been done before more effectively, resulting in more research being done, faster. These bodies of knowledge could also help to break down barriers for and support new researchers (both early career researchers and experienced researchers from different fields) to start using administrative data to address their research questions. This will be particularly important for linked administrative datasets that bring together information from different domains for the first time and will require interdisciplinary research teams. It is important that curation activities are embedded into the ongoing development of research-ready administrative data. Without curation, "data curation debt" - the amount of work required to bring a dataset up to a point that is acceptable to research users [58] - begins to accumulate. If data curation debt goes unchecked, the utility of the data for research purposes is reduced and, eventually, the amount of work required becomes so great that the investment to bring it up to usable standards can't be justified.

Documented

To ensure an administrative dataset is research-ready, researchers require information about key aspects of the data in the form of documentation. For example, researchers will need background information about the context and purposes of the data collection. They will also require information on the creation and processing of an administrative dataset, as well as the content, coverage, quality and completeness of included variables. This type of information is important for understanding the research possibilities (and limitations) of an administrative dataset [43, 59]. Documentation is particularly important for data that involves linkage (e.g., longitudinal records within a single dataset or linkage between multiple datasets) [45]. Documentation could include data catalogues, user guides, descriptive notes, technical reports and data resource profiles.

Researchers will also need information about data governance and access, including details of the application and assessment process. It is also important that there is transparency about how the administrative data have been used in research. For example, this may include publishing details about the applications that have been made to use the data and maintaining a register of project protocols and research outputs (e.g., reports and academic articles). This type of documentation will build evidence of the utility of the administrative dataset for research purposes and avoid duplication of effort by researchers. It could also serve to improve the visibility and transparency of administrative data research, thereby building public confidence and trust.

Documentation should be revised on an ongoing basis to ensure it remains up-to-date and relevant. An archive of documentation should also be maintained. This is particularly important for data owners to ensure that information about datasets is not lost over time (e.g., through staff turnover or institutional changes, such as mergers or rebrands). Users of research-ready administrative databases could be invited to contribute to the documentation to reduce the burden of documentation on the data owner.

Enhanced

An overarching theme that emerged from our analysis of the included literature was that, in comparison to the raw administrative data collected by organisations, research-ready administrative data are enhanced to make them usable for research purposes. For example, Denaxas *et al.* described CALIBER as a research-ready data source because it "curates data from multiple electronic health record sources, generating research-ready data from raw data" [43]. Similarly, Harris *et al.* highlighted that the administrative data on which their research-ready critical care dataset is created "is frequently unusable [for research purposes] in its raw form" [44]. However, there were no specific criteria that qualified a dataset as enhanced. Instead, this enhancement was achieved via the aforementioned characteristics of research-ready administrative data. For example, establishing permissions to re-use administrative data (accessible), making a comprehensive range of data available (broad), validating data quality (curation) and

producing metadata (documentation) all serve to enhance administrative data and make it usable for research purposes.

Discussion

The aim of our study was to propose a set of key characteristics for research-ready administrative data based on evidence from the available literature. Our review identified a small, but growing body of relevant publications. In this literature, there was no common understanding of what the term research-ready means vis-à-vis administrative data. Based on thematic analysis of the available body of literature, we identified five key characteristics of a research-ready administrative data. We propose that research-ready administrative data are (a) accessible, (b) broad, (c) curated and (d) documented which contributes to them being (e) enhanced for research purposes. When these characteristics are achieved the result is well-defined, research-ready administrative databases.

The characteristics of research-ready administrative data that we identified are closely related. In particular, documentation emerged as both a defining characteristic in its own right and one that underpins and contributes to the other characteristics. For example, for administrative data to be findable (accessible) they need to be well-described in online repositories (documented). There can also be tension between the defining characteristics; for example, excessive processing of an administrative dataset with the aim of improving its quality (curated) could inadvertently curtail the range of research that is possible (broad). The challenge in making administrative data research-ready is in striking the balance between these interlinked and sometimes opposing characteristics.

In several included publications, the term research-ready was conflated with being ready for statistical analysis. Under this conceptualisation, administrative data were considered research-ready when it absolved researchers of the need to carry out computationally intensive and non-trivial, technical tasks related to data cleaning and preparation [36, 43, 45]. For example, Harris *et al.* described the development of their research-ready critical care dataset as an attempt to overcome the issue of "the pace of research [being] mired by the need to repeatedly prepare and clean the data" [44]. However, there are benefits to researchers needing to carry out some data cleaning and preparation themselves, such as greater familiarity with the data, as well as more control and flexibility in their research. The research process often requires choices to be made and researchers will want (and need) to make their own decisions as to how to clean and prepare the data for their own purposes [60] and in relation to the needs and priorities of their specific research questions. For example, such choices might include how to derive variables which do not have agreed definitions, such as an inpatient admission [61], a chronic health condition [62] or a child re-entering care [63]. As such, there is an important and crucial difference between making administrative data *research-ready* for broad research purposes and making it *analysis-ready* to address a specific research question. Researchers should expect that some cleaning and preparation is required on their part when working with research-ready administrative data. Equating the term research-ready to readiness for statistical analysis also transfers the burden of cleaning and preparation from the individual researcher to the data owner. This approach is likely to hinder the development of research-ready administrative data resources, given that data owners cite the time and costs required to prepare data for research purposes as a major barrier to making data available for research purposes [1].

The lack of common understanding of what constitutes research-ready administrative data highlights that there are likely to be disparities in expectations among data owning organisations wanting to open up their data to researchers, those running TREs facilitating access to this data, and researchers themselves. Given the nebulous nature of the term, perhaps it is more appropriate to consider research-readiness as a continuum, along which data sources' readiness may vary based on their complexity and maturity in relation to research? For example, the National Joint Registry has been used for research purposes for more than 15 years and has a comparatively narrow scope (patients receiving treatment for joint diseases); therefore, it may be reasonable for researchers to expect to access a pre-cleaned dataset ready to "plug and play" [47]. However, for novel data sources or linkages between multiple data sources, such an expectation would be unreasonable and would ultimately lead to delays in data being used for research purposes. Conceptualising research-readiness as a continuum or scale may also be helpful for evaluating the current utility of existing data resources [64] and managing researcher expectations. Enhancing data for research purposes should be seen as an ongoing and collaborative process between research users and data owners that results in sustainable, long-term research resources that benefit both parties.

A strength of this study is that it had a narrowly focussed review question and rigorous, pre-specified and pre-published methods [15], including a comprehensive search strategy. However, given that the field of administrative data research is fast growing, our decision to not include grey literature in our search may have excluded additional relevant publications. A further limitation is that well-established administrative data resources that have been used extensively in research were not identified in our systematic search, because there are no peer-reviewed articles that explicitly describe them as research-ready (e.g. the Manitoba Population Research Data Repository in Canada [65] or the Population Health Research Network in Australia [66]). However, as the motivation for this study was to gain clarity on the characteristics of data that researchers perceive to be "research-ready", it was necessary to include this phrase as a search term. In the future, this work could be extended by identifying a set of well-established administrative data resources and assessing whether their defining characteristics are captured by the proposed five characteristics, as a form of validation. An important next step for this work will also be to explore how research-ready administrative data have been achieved, paying particular attention to diversity of approach between the four nations of the UK. For example, this could involve stakeholder engagement and documentary analysis to produce case studies of the development of existing data repositories, such as the Secure Anonymised Information Linkage (SAIL) Databank [67] or OpenSAFELY [68].

The literature we identified in this systematic review was mainly from the UK and US (23/38 included publications). This may be an artefact of the restriction to English-only publications. However, it may also reflect to a certain extent the natural distribution of administrative datasets that are described as research-ready. For example, in Nordic countries, where there is already a long and well-established tradition of using administrative data for research [69], as well as operational and statistical purposes, the term research-ready may not be used. Research-ready administrative data may, therefore, be only an interim measure until the use of linked administrative data becomes so normalised across government and research that data providers have the capacity to document, curate, link and enhance data for widespread use. In the meantime, as part of this journey in the UK, one solution may be for data owners to make data available after minimal, well-documented processing and to establish a culture of reproducibility and transparency in relation to how their data are subsequently cleaned and prepared by researchers prior to analysis. For example, researchers could be encouraged to share best practice guidelines or code related to data cleaning and preparation. This would require collaboration between data owners, research institutions and funding bodies as the current climate of 'publish or perish' does little to incentivise the publication of research resources that fall outside the traditional peer-reviewed journal article. Much more needs to be done to foster a culture which recognises and rewards the value of contributions that promote reproducible research and reduce duplication of effort across the research community [70], such as depositing code for re-use, contributing to metadata and developing dataset documentation. One example, suggested by Denaxas *et al.* [4], would be to embed the practice of citing algorithms, code and other tools used to create and prepare research-ready datasets in recognition of their contribution to a study in the same way that researchers cite published articles.

Conclusion

Administrative data are an extremely valuable, but under-utilised, research resource in the UK. There has been considerable interest and financial investment in making administrative data research-ready in recent years, but, there is no clear definition of what this entails. Our findings should help to frame discussions between data owners and researchers about how we conceptualise research-ready administrative data and provide opportunities to develop common principles and standards. The characteristics we have identified in this analysis could act as a starting point to develop a set of principles and common standards for research-ready administrative data which could then be used to evaluate the utility of administrative data resources [64]. In the more immediate term, our proposed characteristics could act as a useful framework for drawing together and cataloguing existing resources relevant to research-ready administrative data owners and users (e.g., the GUILD reporting guidelines for linked data [71] and HDR UK metadata standards [72] which are useful resources related to documenting different aspects of a research-ready dataset). This would also allow areas where there are gaps in terms of best practice guidelines to be identified and addressed through collaborative efforts of data owners and researchers.

Acknowledgements

This work is supported by ADR UK (Administrative Data Research UK), an Economic and Social Research Council (part of UK Research and Innovation) programme [grant number ES/V000977/1]. This research was also supported in part by the National Institute for Health Research (NIHR) Great Ormond Street Hospital Biomedical Research Centre and the Health Data Research UK [grant number LOND1], which is funded by the UK Medical Research Council and eight other funders. This research benefits from and contributes to the NIHR Children and Families Policy Research Unit, but was not commissioned by the NIHR Policy Research Programme. MJ, AZ, LW and RG are in part supported by the NIHR Children and Families Policy Research Unit. The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care. RB is supported by a UKRI Innovation Fellowship funded by the Medical Research Council [grant number MR/S003797/1].

The authors would like to thank Emily Oliver and Paul Jackson (ADR UK) for their helpful comments on earlier versions of this work.

Statements on conflicts of interest

None to be declared.

Ethics statement

No ethical approval was required for this research as it was a systematic review of published literature.

References

1. ADR UK. Annual Report 2018–19 Administrative data is an invaluable resource for public good. Let's use it. [Internet]. Swindon, UK.: 2019. Available from: tinyurl.com/bmfwwekr

2. Connelly R, Playford CJ, Gayle V, Dibben C. The role of administrative data in the big data revolution in social science research. Soc. Sci. Res. 2016;59:1–12. https://doi.org/10.1016/j.ssresearch.2016.04.015

3. Drake B, Jonson-Reid M. Some Thoughts on the Increasing Use of Administrative Data in Child Maltreatment Research. Child Maltreat. 1999;4:308–15. https://doi.org/10.1177/1077559599004004004

4. Denaxas SC, Morley KI. Big biomedical data and cardiovascular disease research: Opportunities and challenges. Eur. Hear. J. - Qual. Care Clin. Outcomes 2015;1:9–16. https://doi.org/10.1093/ehjqcco/qcv005

5. Waind E. International Journal of linking and use of administrative data for research. Int. J. Po 2020;5. https://doi.org/10.23889/ijpds.v5i3.1368

6. Health Data Research UK. Realising patient and NHS benefits from health and care data – from policy to practice [Internet]. 2020. Available from: https://acmedsci.ac.uk/file-download/73707502

7. Great Britain. Digital Economy Act 2017. London: Stationery Office; 2017. Available from: https://www.legislation.gov.uk/ukpga/2017/30/contents/enacted

8. Morris H, Lanati S, Gilbert R. Challenges of administrative data linkages: experiences of Adminstrative Data Research Centre for England (ADRC-E) researchers. Int. J. Popul. Data Sci. 2018;3:97. https://doi.org/10.23889/ijpds.v3i2.566

9. Taylor JA, Crowe S, Espuny Pujol F, Franklin RC, Feltbower RG, Norman LJ, et al. The road to hell is paved with good intentions: The experience of applying for national data for linkage and suggestions for improvement. BMJ Open 2021;11:1–10. https://doi.org/10.1136/bmjopen-2020-047575

10. Gilbert R, Goldstein H, Hemingway H. The market in healthcare data. BMJ [Internet] 2015;351:1–2. Available from: https://doi.org/doi:10.1136/bmj.h5897

11. Department for Business Innovation and Skills. £73 million to improve access to data and drive innovation [Internet]. 2014 [cited 2021 9]; Available from: https://www.gov.uk/government/news/73-million-to-improve-access-to-data-and-drive-innovation

12. ADR UK. ADR UK secures £90 million funding extension as Government plans better use of data [Internet]. 2021; Available from: https://www.adruk.org/news-publications/news-blogs/adr-uk-secures-gbp90-million-funding-extension-as-government-plans-better-use-of-data-450/

13. Jones KH, Heys S, Tingay KS, Jackson P, Dibben C. The Good, the Bad, the Clunky. Int. J. Popul. Data Sci. 2019;4. https://doi.org/10.23889/ijpds.v4i1.587

14. UK Health Data Research Alliance. Trusted Research Environments (TRE) [Internet]. 2020. Available from: https://ukhealthdata.org/wp-content/uploads/2020/07/200723-Alliance-Board_Paper-E_TRE-Green-Paper.pdf

15. Mc Grath-Lone L. Systematic review protocol: What makes administrative data research ready? [Internet]. 2021. Available from: osf.io/vrqfe

16. International Journal of Population Data Science. IJPDS is indexed in PubMed Central [Internet]. 2020; Available from: https://ijpds.org/announcement/view/21

17. Braun V, Clarke V. Thematic analysis. In: Cooper H, Camic PM, Long DL, Panter AT, Rindskopf D, Sher KJ, editors. APA handbook of research methods in psychology, Vol 2: Research designs: Quantitative, qualitative, neuropsychological, and biological. American Psychological Association; 2012. p. 57–71. https://doi.org/10.1037/13620-004

18. Abernethy AP, Arunachalam A, Burke T, McKay C, Cao X, Sorg R, et al. Real-world first-line treatment and overall survival in non-small cell lung cancer without known EGFR mutations or ALK rearrangements in US community oncology setting. PLoS One 2017 Jun.;12. https://doi.org/http://dx.doi.org/10.1371/journal.pone.0178420

19. Atkinson MD, Kennedy JI, John A, Lewis KE, Lyons RA, Brophy ST. Development of an algorithm for determining smoking status and behaviour over the life course from UK electronic primary care records. BMC Med. Inform. Decis. Mak. 2017;17. https://doi.org/http://dx.doi.org/10.1186/s12911-016-0400-6

20. Graham SJH, Marco AC, Myers AF. Patent transactions in the marketplace: Lessons from the USPTO Patent Assignment Dataset. J. Econ. Manag. Strateg. 2018;27:343–71. https://doi.org/10.1111/jems.12262

21. Hemingway H, Lyons R, Li Q, Bucha I, Ainsworth J, Pell J, et al. A national initiative in data science for health: an evaluation of the UK Farr Institute. Int. J. Popul. Data Sci. 2020;5:20. https://doi.org/10.23889/ijpds.v5i1.1128

22. Terry DR, Peck B, Kloot K. The data deficit for asthma emergency presentations might surprise you: How RAHDaR addresses the data chasm. Rural Remote Health 2020;20:8–13. https://doi.org/10.22605/RRH5776

23. Li Y, Appius A, Pattipaka T, Feyereislova A, Cassidy A, Ganti AK. Real-world management of patients with epidermal growth factor receptor (EGFR) mutation-positive non–small-cell lung cancer in the USA. PLoS One 2019 Jan.;14. https://doi.org/http://dx.doi.org/10.1371/journal.pone.0209709

24. Li Y, Rizzo JA. Timing and payoff of patent purchases: the role of firm size and composition. Appl. Econ. 2020;52:5894–908. https://doi.org/10.1080/00036846.2020.1778159

25. Lin H, Tang X, Shen P, Zhang D, Wu J, Zhang J, et al. Using big data to improve cardiovascular care and outcomes in China: a protocol for the CHinese Electronic health Records Research in Yinzhou (CHERRY) Study. BMJ Open 2018;8. https://doi.org/http://dx.doi.org/10.1136/bmjopen-2017-019698

26. Lyons J, Akbari A, Agrawal U, Harper G, Azcoaga-Lorenzo A, Bailey R, et al. Protocol for the development of the Wales Multimorbidity e-Cohort (WMC): Data sources and methods to construct a population-based research platform to investigate multimorbidity. BMJ Open 2021;11. https://doi.org/10.1136/bmjopen-2020-047101

27. Mehta S, Jackson R, Exeter DJ, Wu BP, Wells S, Kerr AJ. Data resource: Vascular Risk in Adult New Zealanders (VARIANZ) datasets. Int. J. Popul. Data Sci. 2019;4. https://doi.org/10.23889/ijpds.v4i1.1107

28. Milinovich A, Kattan MW. Extracting and utilizing electronic health data from Epic for research. Ann. Transl. Med. 2018;6. https://doi.org/10.21037/atm.2018.01.13

29. Newman D, Herrera CN, Parente ST. Overcoming barriers to a research-ready national commercial claims database. Am. J. Manag. Care [Internet] 2014 Nov.;20:eSP25–30. Available from: https://www.ajmc.com/view/overcoming-barriers-to-a-research-ready-national-commercial-claims-database

30. Barreto ML, Ichihara MY, Almeida BA, Barreto ME, Cabral L, Fiaccone RL, et al. The centre for data and knowledge integration for health (CIDACS): Linking health and social data in Brazil. Int. J. Popul. Data Sci. 2019;4. https://doi.org/10.23889/ijpds.v4i2.1140

31. O'Reilly D, Bateson O, McGreevy G, Snoddy C, Power T. Administrative Data Research Northern Ireland (ADR NI). Int. J. Popul. Data Sci. 2019;4. https://doi.org/10.23889/ijpds.v4i2.1148

32. Peck B, Terry DR, Kloot K. Understanding childhood injuries in rural areas: Using Rural Acute Hospital Data Register to address previous data deficiencies. EMA - Emerg. Med. Australas. 2020;32:646–9. https://doi.org/10.1111/1742-6723.13565

33. Peck B, Terry D, Kloot K. The socioeconomic characteristics of childhood injuries in regional victoria, australia: What the missing data tells us. Int. J. Environ. Res. Public Health 2021;18. https://doi.org/10.3390/ijerph18137005

34. Schull M, Azimaee M, Marra M, Cartagena R, Vermeulen M, Guttmann A. ICES: Data, Discovery, Better Health. Int. J. Popul. Data Sci. 2020;4. https://doi.org/10.23889/ijpds.v4i2.1135

35. Smith DA, Wang T, Freeman O, Crichton C, Salih H, Matthews PC, et al. National Institute for Health Research Health Informatics Collaborative: Development of a pipeline to collate electronic clinical data for viral hepatitis research. BMJ Heal. Care Informatics 2020;27. https://doi.org/10.1136/bmjhci-2020-100145

36. Springate DA, Parisi R, Olier I, Reeves D, Kontopantelis E. rEHR: An R package for manipulating and analysing electronic health record data. PLoS One 2017;12:1–25. https://doi.org/10.1371/journal.pone.0171784

37. Tissot HC, Shah AD, Brealey D, Harris S, Agbakoba R, Folarin A, et al. Natural Language Processing for Mimicking Clinical Trial Recruitment in Critical Care: A Semi-Automated Simulation Based on the LeoPARDS Trial. IEEE J. Biomed. Heal. Informatics 2020;24:2950–9. https://doi.org/10.1109/JBHI.2020.2977925

38. Walker JD, Pyper E, Jones CR, Khan S, Chong N, Legge D, et al. Unlocking First Nations health information through data linkage. Int. J. Popul. Data Sci. 2018;3. https://doi.org/10.23889/ijpds.v3i1.450

39. Wallace PJ, Shah ND, Dennen T, Bleicher PA, Crown WH. Optum Labs: Building A Novel Node In The Learning Health Care System. Health Aff. 2014 Jul.;33:1187–94. https://doi.org/10.1377/hlthaff.2014.0038

40. Wang S-M, Lai M-N, Wei A, Ya-Yin C, Pu Y-S, Pau-Chung C, et al. Increased Risk of Urinary Tract Cancer in ESRD Patients Associated with Usage of Chinese Herbal Products Suspected of Containing Aristolochic Acid. PLoS One 2014 Aug.;9. https://doi.org/http://dx.doi.org/10.1371/journal.pone.0105218

41. Batsos C, Boyes R, Mahar A. Community water fluoridation exposure and dental caries experience in newly enrolled members of the Canadian Armed Forces 2006-2017. Can. J. Public Health 2021;112:513–20. https://doi.org/10.17269/s41997-020-00463-7

42. Wilhite JA, Altshuler L, Zabar S, Gillespie C, Kalet A. Development and maintenance of a medical education research registry. BMC Med. Educ. 2020;20:1–12. https://doi.org/http://dx.doi.org/10.1186/s12909-020-02113-5

43. Denaxas SC, George J, Herrett E, Shah AD, Kalra D, Hingorani AD, et al. Data resource profile: Cardiovascular disease research using linked bespoke studies and electronic health records (CALIBER). Int. J. Epidemiol. 2012;41:1625–38. https://doi.org/10.1093/ije/dys188

44. Harris S, Shi S, Brealey D, MacCallum NS, Denaxas S, Perez-Suarez D, et al. Critical Care Health Informatics Collaborative (CCHIC): Data, tools and methods for reproducible research: A multi-centre UK intensive care database. Int. J. Med. Inform. 2018;112:82–9. https://doi.org/10.1016/j.ijmedinf.2018.01.006

45. Hilder L, Walker JR, Levy MH, Sullivan EA. Preparing linked population data for research: Cohort study of prisoner perinatal health outcomes. BMC Med. Res. Methodol. 2016;16:1–11. https://doi.org/10.1186/s12874-016-0174-7

46. McWilliams C, Inoue J, Wadey P, Palmer G, Santos-Rodriguez R, Bourdeaux C. Curation of an intensive care research dataset from routinely collected patient data in an NHS trust. F1000Research 2019;8:1–11. https://doi.org/10.12688/f1000research.20193.1

47. Porter M, Armstrong R, Howard P, Porteous M, Wilkinson JM. Orthopaedic registries - the UK view (National Joint Registry): Impact on practice. EFORT Open Rev. 2019;4:377–90. https://doi.org/10.1302/2058-5241.4.180084

48. Terry D, Peck B, Kloot K, Hutchins T. Pediatric emergency asthma presentations in

Southwest Victoria: a retrospective cross-sectional study 2017 to 2020. J. Asthma 2020; https://doi.org/10.1080/02770903.2020.1845725

49. Bhatnagar P, Scarborough P, Kaur A, Dikmen D, Adhikari V, Harrington R. Are food and drink available in online and physical supermarkets the same? A comparison of product availability, price, price promotions and nutritional information. Public Health Nutr. 2020;1–19. https://doi.org/10.1017/s1368980020004346

50. Bush RA, Kuelbs C, Ryu J, Jiang W, Chiang G. Structured Data Entry in the Electronic Medical Record: Perspectives of Pediatric Specialty Physicians and Surgeons. J. Med. Syst. 2017 May;41:1–8. https://doi.org/http://dx.doi.org/10.1007/s10916-017-0716-5

51. Dahlhaus P, Murphy A, MacLeod A, Thompson H, McKenna K, Ollerenshaw A. Making the invisible visible: the impact of federating groundwater data in Victoria, Australia. J. Hydroinformatics 2016 Mar.;18:238–55. https://doi.org/http://dx.doi.org/10.2166/hydro.2015.169

52. Dreyer NA, Mack CD, Anderson RB, Wojtys EM, Hershman EB, Sills A. Lessons on Data Collection and Curation From the NFL Injury Surveillance Program. Sports Health 2019;11:440–5. https://doi.org/10.1177/1941738119854759

53. Fallon B, Filippelli J, Black T, Trocme N, Esposito T. How Can Data Drive Policy and Practice in Child Welfare? Making the Link in Canada. Int. J. Environ. Res. Public Health 2017 Oct.;14:1223. https://doi.org/http://dx.doi.org/10.3390/ijerph14101223

54. Graham SJH, Marco AC, Miller R. The USPTO Patent Examination Research Dataset: A window on patent processing. J. Econ. Manag. Strateg. 2018;27:554–78. https://doi.org/10.1111/jems.12263

55. UK Data Archive. Find data [Internet]. 2020 [cited 2020 9]; Available from: https://www.data-archive.ac.uk/find/

56. Health Data Research Alliance. Health Data Research Innovation Gateway [Internet]. [cited 2021 9]; Available from: https://www.healthdatagateway.org/

57. Horowitz BM. Policy Issues Regarding Implementations of Cyber Attack: Resilience Solutions for Cyber Physical Systems. In: Lawless W, Mittu R, Sofge D, Moskowitz IS, Russell S, editors. Artificial Intelligence for the Internet of Everything. 2019. p. 87–100. https://doi.org/10.1016/b978-0-12-817636-8.00005-3

58. Butters OW, Wilson RC, Burton PR. Recognizing, reporting and reducing the data curation debt of cohort studies. Int. J. Epidemiol. 2020;49:1067–74. https://doi.org/10.1093/ije/dyaa087

59. Johnson KE, Kamineni A, Fuller S, Olmstead D, Wernli KJ. How the Provenance of Electronic Health Record Data Matters for Research: A Case Example Using System Mapping. eGEMs (Generating Evid. Methods to Improv. patient outcomes) 2014;2:4. https://doi.org/10.13063/2327-9214.1058

60. Hotz VJ, Goerge R, Balzekas J, Margolin F, Balzekas JD, Bradburn N, et al. Adminstrative data for policy-relevant research: Assessment of current utility and recommendations for development. [Internet]. 1991. Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.421.1385&rep=rep1&type=pdf

61. Herbert A, Wijlaars L, Zylbersztejn A, Cromwell D, Hardelid P. Data Resource Profile: Hospital Episode Statistics Admitted Patient Care (HES APC). Int. J. Epidemiol. 2017;46:1093-1093i. https://doi.org/10.1093/ije/dyx015

62. Wijlaars LPMM, Gilbert R, Hardelid P. Chronic conditions in children and young people: Learning from administrative data. Arch. Dis. Child. 2016;101:881–5. https://doi.org/10.1136/archdischild-2016-310716

63. Mc Grath-Lone L, Dearden L, Harron K, Nasim B, Gilbert R. Factors associated with re-entry to out-of-home care among children in England. Child Abuse Negl. 2017;63:73–83. https://doi.org/10.1016/j.chiabu.2016.11.012

64. Health Data Research UK. Data Utility Framework [Internet]. 2020. Available from: https://www.hdruk.ac.uk/wp-content/uploads/2020/11/201105-Updates-to-the-Data-Utility-Framework-v2.pdf

65. Katz A, Brownell M, Enns JE, Nickel NC. Closing the loop: From system-based data to evidence-influenced policy and practice. Int. J. Popul. 2022;7:4. https://ajph.aphapublications.org/doi/10.2105/AJPH.2008.156224

66. Flack F, Smith M. The Population Health Research Network - Population data centre profile. Int. J. Popul. Data Sci. 2019;4. https://doi.org/10.23889/ijpds.v4i2.1130

67. Jones KH, Ford D V., Thompson S, Lyons RA. A profile of the SAIL databank on the UK secure research platform. Int. J. Popul. Data Sci. 2019;4. https://doi.org/10.23889/ijpds.v4i2.1134

68. The DataLab. About OpenSAFELY [Internet]. 2022; Available from: https://www.opensafely.org/about/

69. Laugesen K, Ludvigsson JF, Schmidt M, Gissler M, Valdimarsdottir UA, Lunde A, et al. Nordic health registry-based research: A review of health care systems and key registries. Clin. Epidemiol. 2021;13:533–54. https://doi.org/10.2147/CLEP.S314959

70. Open Research Data Task Force. Realising the potential Final report of the Open Research Data Task Force [Internet]. 2018. Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/775006/Realising-the-potential-ORDTF-July-2018.pdf

71. Gilbert R, Lafferty R, Hagger-Johnson G, Harron K, Zhang LC, Smith P, et al. GUILD: GUidance for Information about Linking Data sets. J. Public Heal. (United Kingdom) 2018;40:191–8. https://doi.org/10.1093/pubmed/fdx037

72. Milward A, Tripathi A, Jones M. Metadata Specification to support Innovation Gateway Minimum Viable Product (MVP). 2019; Available from: https://hdronboard.metadata.works/#/122894/dataClass/123142/main?path=122894-all-123142

Abbreviations

ADR UK:	Administrative Data Research UK
CINAHL:	Cumulative Index to Nursing and Allied Health Literature
GUILD:	Guidance for information about linking data sets
IJPDS:	International Journal of Population Data Science
TRE:	Trusted Research Environment
UK:	United Kingdom
US:	United States

Supplementary Table 1: Search strings used in systematic review

Main search (28thJune 2021)	
Electronic database	**String**
Embase (via Ovid) MEDLINE (via Ovid)	("research-ready" or "research ready") AND (administrative OR operational OR routinely-collected OR "routinely collected" OR records OR data?)
PubMed	("research-ready" or "research ready") AND (administrative OR operational OR routinely-collected OR "routinely collected" OR records OR data?)
Scopus	ALL (({research-ready} OR {research ready}) AND (administrative OR operational OR {routinely-collected} OR {routinely collected} OR records OR data*))
ProQuest Central	ft(research N/0 ready) AND ft(administrative OR operational OR (routinely N/0 collected) OR records OR data*) NOT ab(pilot OR experiment* OR participants OR randomised) RESTRICTED TO scholarly journals
CINAHL Plus	TX ("research ready") AND TX (administrative OR operational OR "routinely collected" OR records OR data*))
Web of Science Core Collection	ALL = (research-ready) AND ALL = (administrative OR operational OR "routinely collected" OR records OR data*)
Supplementary search (7th July 2021)	
Website	**String**
Google Scholar	("research-ready" or "research ready") AND (administrative OR operational OR routinely-collected OR "routinely collected" OR records OR data?)
International Journal of Population Data Science	(" research-ready" OR "research ready")