




## Article

# Challenges Encountered and Lessons Learned When Using a Novel Anonymised Linked Dataset of Health and Social Care Records for Public Health Intelligence: The Sussex Integrated Dataset

Elizabeth Ford <sup>1,\*</sup> , Richard Tyler <sup>2</sup>, Natalie Johnston <sup>3</sup>, Vicki Spencer-Hughes <sup>4</sup>, Graham Evans <sup>4</sup>, Jon Elsom <sup>5</sup>, Anotida Madzvamuse <sup>6,7,8,9</sup>, Jacqueline Clay <sup>2</sup>, Kate Gilchrist <sup>3</sup>  and Melanie Rees-Roberts <sup>10</sup> 

- <sup>1</sup> Department of Primary Care and Public Health, Brighton and Sussex Medical School, Room 104 Watson Building, Village Way, Falmer, Brighton BN1 9PH, UK  
<sup>2</sup> West Sussex County Council, Chichester PO19 1RQ, UK  
<sup>3</sup> Brighton and Hove City Council, Brighton BN3 3BQ, UK  
<sup>4</sup> East Sussex County Council, Lewes BN7 1UE, UK  
<sup>5</sup> East Sussex Health Trust, St Leonards-on-Sea, East Sussex TN37 7PT, UK  
<sup>6</sup> Department of Mathematics, University of Sussex, Brighton BN1 9PH, UK  
<sup>7</sup> Department of Mathematics University of British Columbia, Vancouver, BC V6T 1Z2, Canada  
<sup>8</sup> Department of Mathematics and Applied Mathematics, University of Pretoria, Pretoria 0132, South Africa  
<sup>9</sup> Department of Mathematics and Applied Mathematics, University of Johannesburg, Auckland Park 2006, South Africa  
<sup>10</sup> Centre for Health Services Studies, University of Kent, Canterbury CT2 7NZ, UK  
\* Correspondence: e.m.ford@bsms.ac.uk



**Citation:** Ford, E.; Tyler, R.; Johnston, N.; Spencer-Hughes, V.; Evans, G.; Elsom, J.; Madzvamuse, A.; Clay, J.; Gilchrist, K.; Rees-Roberts, M. Challenges Encountered and Lessons Learned When Using a Novel Anonymised Linked Dataset of Health and Social Care Records for Public Health Intelligence: The Sussex Integrated Dataset. *Information* **2023**, *14*, 106. <https://doi.org/10.3390/info14020106>

Academic Editors: Mario Ciampi and Mario Sicuranza

Received: 19 December 2022

Revised: 18 January 2023

Accepted: 3 February 2023

Published: 8 February 2023



**Copyright:** © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

**Abstract:** Background: In the United Kingdom National Health Service (NHS), digital transformation programmes have resulted in the creation of pseudonymised linked datasets of patient-level medical records across all NHS and social care services. In the Southeast England counties of East and West Sussex, public health intelligence analysts based in local authorities (LAs) aimed to use the newly created “Sussex Integrated Dataset” (SID) for identifying cohorts of patients who are at risk of early onset multiple long-term conditions (MLTCs). Analysts from the LAs were among the first to have access to this new dataset. Methods: Data access was assured as the analysts were employed within joint data controller organisations and logged into the data via virtual machines following approval of a data access request. Analysts examined the demographics and medical history of patients against multiple external sources, identifying data quality issues and developing methods to establish true values for cases with multiple conflicting entries. Service use was plotted over timelines for individual patients. Results: Early evaluation of the data revealed multiple conflicting within-patient values for age, sex, ethnicity and date of death. This was partially resolved by creating a “demographic milestones” table, capturing demographic details for each patient for each year of the data available in the SID. Older data ( $\geq 5$  y) was found to be sparse in events and diagnoses. Open-source code lists for defining long-term conditions were poor at identifying the expected number of patients, and bespoke code lists were developed by hand and validated against other sources of data. At the start, the age and sex distributions of patients submitted by GP practices were substantially different from those published by NHS Digital, and errors in data processing were identified and rectified. Conclusions: While new NHS linked datasets appear a promising resource for tracking multi-service use, MLTCs and health inequalities, substantial investment in data analysis and data architect time is necessary to ensure high enough quality data for meaningful analysis. Our team made conceptual progress in identifying the skills needed for programming analyses and understanding the types of questions which can be asked and answered reliably in these datasets.

**Keywords:** health data; electronic health records; data linkage; data quality; public health

## 1. Introduction

In England, one of four countries of the United Kingdom (UK), digital transformation plans have been underway in the National Health Service (NHS) for several years, first within Sustainability and Transformation Plans (STPs) announced in NHS planning guidance in 2015 [1]. STPs were succeeded by integrated care systems, made up of NHS organisations and upper-tier local authorities (LAs), with 42 systems covering the whole of England in 2018/2019 planning guidance [2]; in July 2022, these then became known as Integrated Care Partnerships (ICPs; made up of NHS integrated care boards and upper-tier LAs), putting them on a statutory footing [3,4].

Digital transformation programmes are designed to bring data and technology together to provide better health and care services for their ICP populations. In the counties of East and West Sussex, on the Southeast coast of England, the digital strategy has resulted in the development of the Sussex Integrated Dataset (SID). SID includes de-identified and linked health and social care data from the whole population to help commissioners and service providers better understand the Sussex population, identify those at risk of poor health and design and run better services to improve population health [5].

ICPs are responsible for the health and care of their population and therefore include all NHS providers in the region as well as local authorities who are responsible for delivering social care and public health interventions and improvements [6]. Part of this responsibility includes establishing effective public health intelligence (PHI), including using relevant health and social care data [7]. The local authorities in Sussex (East Sussex County Council (ESCC), West Sussex County Council (WSCC), and Brighton and Hove City Council (BHCC)) were particularly interested in using the newly developed SID for executing their public health intelligence responsibilities. Whilst some public health intelligence tasks involve record-level datasets, the majority utilise aggregated data outputs that have been extensively cleaned, and often data are accompanied by metadata and caveats. SID, contrastingly, at project start, was an uncurated “data lake” undergoing minimal cleaning, and formed of multiple patient-level entries for most of the Sussex population across all health and social care services, recorded over several years.

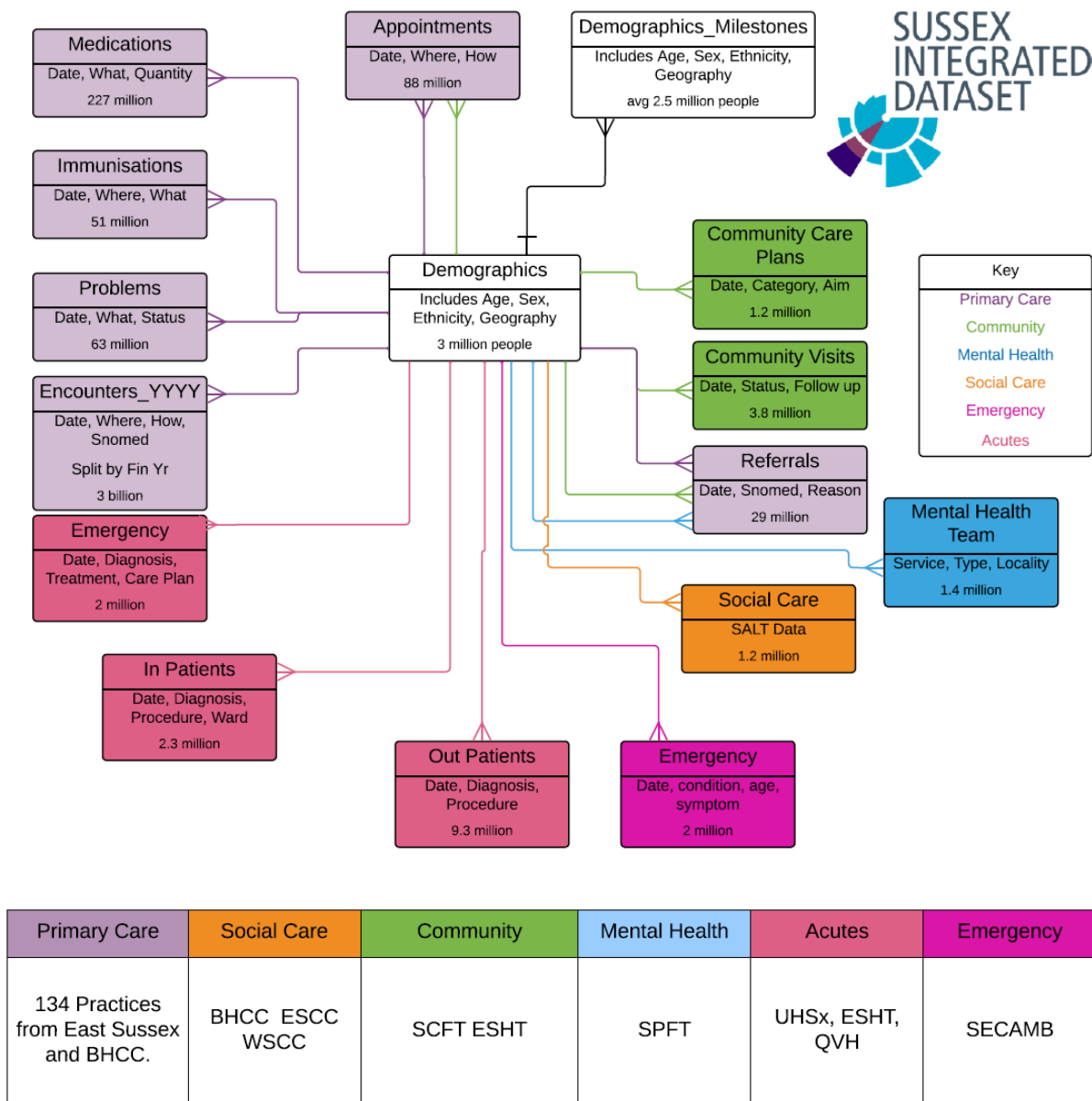
Public health intelligence teams across Sussex planned to make use of the SID to inform strategy and to better understand the health and care needs of the local population. They wanted to use the SID to track individuals’ patterns of service use, understand the distribution of long-term conditions within the population, build a complex description of co-morbidities’ clustering and accumulation over time, and learn methods to present multidimensional longitudinal data in a way that would be easily understood by decision-makers. The overall aim of securing this intelligence would be to reduce the gap in healthy life expectancy between the least and most deprived populations by designing interventions aimed at delaying the age of onset of multiple long-term conditions.

In this project, we aimed to work across the interface of academic data scientists and local authority analysts to investigate the structure of the SID data for the first time and to prepare the team for using the SID to answer the public health intelligence aims described above. In this article, we report on what we found in the dataset, what problems needed to be addressed, and how these were solved. Our aim is to share this learning with other teams using novel health and care-linked datasets for the first time.

## 2. Materials and Methods

### 2.1. The Data Source: Sussex Integrated Dataset

Sussex ICP has secured data flows from nearly all NHS and local authority (LA) providers across Sussex, including GP practices (97% sign up at the end of 2022), Adult Social Care, Community Care, Mental Health, Acute Trusts (Hospital Care) and Emergency Care (Ambulance Service). The data flows into SID are depicted in Figure 1.



**Figure 1.** The data flowing into SID as of November 2022. Abbreviations: BHCC—Brighton and Hove City Council, ESCC—East Sussex County Council, WSCC—West Sussex County Council, SCFT—Sussex Community Foundation NHS Trust, ESHT—East Sussex Health Trust, SPFT—Sussex Partnership Foundation NHS Trust, UHSx—University Hospitals Sussex, QVH—Queen Victoria Hospital, East Grinstead, SECAMB—South East Coast Ambulance Service.

From GP practices, data flows from the primary care electronic health records (EHR) source system (either TPP or EMIS) to the NHS Commissioning Support Unit and from there via a pseudonymisation process into the SID. Data is supplied monthly. Use of the pseudonymiser in this way ensures that no identifiable data leaves the provider’s premises. Once the data lands in the SID, automatic processing validates and cleans data items as necessary and provides a reporting or semantic layer against which analysts can issue queries. The creation of the semantic layer has taken a significant amount of time to develop, and this process is still ongoing. The initial data set that the team worked on lacked much of what is now considered standard, such as clustered code sets, enhanced data quality and reporting objects designed specifically for grouping and aggregation; these are now in place but were developed alongside or after this project.

Only structured health and care data, making use of clinical coding systems (ICD-10 [8], SNOMED-CT [9], Read [10–12] and OPCS codes [13]), are flowed into the SID. Both Read and SNOMED alphanumeric or numeric codes are accompanied by plain English descriptive terms giving the clinical concept represented by the code. No unstructured text data (clinic notes, letters, or reports) are contributed to SID. Data are de-identified and pseudonymised by removal of all structured fields containing names, addresses and dates of birth (only age at event date is supplied). A unique patient can be linked across multiple data sources by their SID ID, created from an encrypted version of their unique NHS identifier plus their date of birth, although neither piece of information is imported into the SID. Every clinical event is accompanied by two date stamps: the event date (when the clinical event occurred) and the process date (when it was recorded in the EHR).

## 2.2. How Data Was Accessed

Access to the SID is achieved via submitting a Data Access Request (DAR) to be assessed against a set of “guard rails” by the SID Analytics Working Group. DARs can be submitted by analysts employed by any data-contributing Sussex-based NHS or LA organisation, who are all joint data controllers for the SID.

The DAR for this project specified that LA analysts would interrogate the SID to:

- (1) Identify efficient methods for evaluating the number of data sources for each individual;
- (2) Evaluate data quality and completeness of diagnostic codes for long term conditions and whether the number of people with a condition in SID was equivalent to numbers reported through traditional routes;
- (3) Develop early methods for describing multiple service use by an individual;
- (4) Explore the application of longitudinal modelling to identify individuals developing 2 or more long-term conditions at different ages, their socio-demographic risk factors and their service use;
- (5) Develop advanced presentation methods to communicate intelligence from these models to decision-makers in an easy-to-understand way.

Once the DAR was approved, the team gained access to SID via virtual machines. Database access security constraints ensured that only the data specified as needed in the DAR could be viewed. Access to the cloud-based virtual machines was controlled through an existing East Sussex Health Trust (ESHT) Azure tenancy.

Data were initially queried using Jupyter notebooks (a web-based interactive computing platform) and analyses written in SQL and R languages (R version R.4.1.0) before workflows were set up in R Studio to establish a connection to the database, execute queries, and complete further analyses. Packages used in R included tidyverse, dbplyr, zoo, epitools, DBI, odbc, keyring, httr and rvest. Data manipulation was run using R scripts so that the team could record their steps, annotate workings and, importantly, share the whole analysis process rather than just the output. Credentials were encrypted in R environment variables, and as data objects were stored in the memory of the R session, it meant that local copies of records were not stored on the machines.

## 2.3. Data Quality Assessment Methods, including Comparison of other Datasets to SID Outputs

The database was so large and heterogeneous that it was not possible to view or browse the whole dataset or even summarise it with simple descriptive figures. For example, there were over 3 billion consultation or encounter records. This was challenging as we could not be sure of what we saw on the screen; queries always represented the raw data or the breadth of possible values. There were often parsing errors that stripped out or added incoherent values during querying. Therefore, we compared values delivered from queries (e.g., how many patients have diabetes?) with expected values from other openly available sources of data held on the Sussex population. We used the following datasets for comparisons:

1. Disease prevalence from the UK Quality and Outcomes Framework (QOF) [14];
2. National Diabetes Audit public reports on prevalence [15];

3. National Cancer Registration and Analysis Service (NCRAS) incidence figures [16]
4. Cardiovascular Disease Prevention Audit (CVDPREVENT), produced by the Office for Health Improvement and Disparities and the NHS Benchmarking Network [17];
5. Number of maternity admissions in secondary care from local analysis of Hospital Episode Statistics and comparisons of overall numbers of emergency and elective admissions [18];
6. Numbers of patients registered to primary care (GP) organisations [19];
7. Office for National Statistics Death Data [20].

We also worked with local stakeholders, such as commissioners, to corroborate the number of referrals to mental health and social care services.

#### *2.4. Methods for Identifying Multiple Long-Term Conditions—Development and Validation of Code Lists*

We developed methods to identify patients having certain conditions in the absence of disease registers. We extracted this information from primary care encounter data, as well as linked hospital records using a set of machine-readable codes (a code list) to represent each clinical entity (e.g., SNOMED, Read Codes, ICD Codes, OPCS codes). The primary care data were split into multiple tables: appointments, encounters, referrals, medications, and problems. In the first instance, we searched the problems table for evidence (using the diagnosis codes in code lists) of each condition.

We used two open-access British code list repositories (HDR Phenotype Library [21] and Open Codelists [22]) to source lists of clinical codes representing 11 long-term conditions of particular interest to the local authorities, identified as important in their Joint Strategic Needs Assessment [23] and also because there were some established prevalence figures collected in the aforementioned data sources. The 11 conditions were: asthma, atrial fibrillation, cancer, chronic obstructive pulmonary disease (COPD), dementia, depression, diabetes, epilepsy, heart failure, serious mental illness (SMI) and stroke.

To validate the code lists, we used the numerators from disease registers in QOF as a benchmark prevalence estimate. The primary care data in SID contained Read codes and SNOMED codes, and we utilised code lists from both coding systems to be as exhaustive as possible in searching. In most cases, the SNOMED code represented the same condition or definition as the Read code, but in some cases, only one field was complete.

The second step in validating code lists was to use text pattern matching on the descriptive terms (rather than just the codes) to identify if some custom codes were being used. This highlighted several instances of relevant codes which were added to the open-source code lists. This worked better for some conditions that were easier to define or had relatively few codes (e.g., asthma and atrial fibrillation compared to cancer, coronary heart disease, or heart failure, which had thousands of codes). We applied this approach with caution as there were a number of negated clinical entities in codes (e.g., “no evidence of asthma,” “not asthma”), so we made sure only relevant, positive codes were included. In the case where an individual could have two conflicting codes (asthma/not asthma), codes were sequenced by activity date so that if asthma was coded more recently, then the individual was designated currently asthmatic in the absence of a record showing resolved asthma. If the individual had a positive asthma code first, followed by a “no asthma” code second, then we deduced resolved asthma.

Some Read codes have special characters such as full stops (e.g., “Atrial Fibrillation and Flutter” is “G573.” and “Exercise induced Asthma” is “173A.”). We found that parsing factors for data as it was processed into SID had led to special characters being removed from codes (for some but not all records). Therefore, we needed to create multiple versions of codes that included and excluded any special characters.

Having applied code lists to extract cases from the problem table, we additionally ran the lists on the encounter tables, as we discovered these were much more likely to be populated with diagnosis codes. The challenge here was that the encounter table was much larger in terms of number of records than the problem table, and the processing power in

the virtual machines for accessing the data was not suitable for running queries over the encounter tables. Therefore, we ran queries over a random sample of 1000 patients and manually screened results to find and exclude additional relevant or non-relevant codes.

Once the improved SID-specific code lists were finalised, the team used these to query much larger cohorts of patients in the database to retrieve all relevant populations with the conditions of interest. The number of long-term conditions per patient by age group was calculated. We plotted the proportion of patients in a GP practice at each age group with 1, 2, 3, 4 or  $\geq 5$  conditions. These are well-established plots that generally have a predictable shape (see, for example, [24]).

### 2.5. Visualisation of Multidimensional Longitudinal Data

Patient journey plots (sometimes known as theographs) of service use were made to identify what events and observations took place over time. These plots were created as a way of summarising a lot of information about a patient's interaction with Sussex services over time in a simple way; to demonstrate the value and potential of visualising data held across the database for a given patient identifier. These plots are increasingly used in case reviews from care providers [25].

### 2.6. Ethics Statement

In the UK, NHS patient data which is rendered functionally anonymous and curated for public health purposes does not need Research Ethics Committee (REC) approval [26]. Data in SID is stripped of all identifiers such as name, date of birth (only age at consultation date is given), address, etc., so that patients become anonymous. Data users sign a formal agreement that no attempts at re-identification will be made. Certain highly sensitive data, such as HIV status or termination of pregnancy, is not routinely extracted from clinics and held within SID. As data is processed (lawfully) without patient consent, the SID team is committed to engaging with Sussex citizens about how they would like their data to be used for public health purposes and research and what safeguards they would like to see [27].

## 3. Results

### 3.1. Data Quality Activities to Identify an Analysable Cohort

In SID, hundreds of providers across multiple data systems were adding record-level data to an expanding "data lake." As the team started to work with the data, it became apparent that an understanding of how data moves from the provider to the database is necessary in order to resolve the challenges of partial and sometimes conflicting metadata and demographic data.

#### 3.1.1. Conflicting Demographic Values

We identified multiple conflicting values within individuals for age (year of birth was redacted, so only an age value was given at each event or consultation), ethnicity, registered GP practice, geographical area of residence (known as lower-layer super output area (LSOA)), and sex. There were concerns surrounding which demographic values to choose as a patient's true value when multiple values were presented (particularly in the same year by different providers).

Around 5% of the SID IDs with primary care demographics had different "current" age values. We challenged the assumption that the most recently processed record was presented as the truest reflection, as some providers (particularly new providers) may submit historical or backdated activity records. Instead, we made the assumption that GP practice data would be the best source of a current age as it was the most frequently and recently updated record in the database. With the conflicting age values, we decided to use the process date stamp and a notional age from the primary care record to derive an age for longitudinal analyses (such as the age at diagnosis or age at hospital admission). This

value was up to 1 year out (i.e., a patient could be 40 years + 11 months), so we weighted everyone to be in the middle of their given age (e.g., 40 = 40.5).

For each patient, the data processing team attempted to fill a new table called “demographic milestones.” This was primarily based on primary care data and had a single record per patient per calendar year representing the best assumption of age, the GP practice they were registered to, and where they lived at the beginning of that year (information was taken as close to the 1st of January as possible). This was a compromise in terms of building a longitudinal picture resulting in changes (such as a change in GP practice, address move or a death) within years not identified until the following milestone record.

### 3.1.2. Conflicting Date Stamps and Sparse Historical Events

The event date was often poorly complete, with much of the data either missing or given the date at which it was processed into the EHR. In many cases, the two values would be the same if data were processed into the EHR during or soon after the activity. However, with backdated/historical records added by new providers, as well as a focus on the age at which events occurred, we quickly identified that many more clinical events (such as strokes) were occurring (and an increasing trend over time) in the most recent five years of data for SID than in earlier years; this represented either that strokes occurring further back in time were missing, or that they had been misattributed to the time in which they entered the database. Published incidence figures indicate a modestly increasing but relatively stable incidence of strokes each year [14]. While primary care practices might be expected to hold full lifetime data on their patients, we found data was particularly sparse for patients who had moved into the area with a history of service use in other parts of the country. We, therefore, concluded that lower incidence in previous years was likely to be due to the absence of record rather than the absence of an event. This reduced our ability to use SID for historical prevalence or incidence estimates.

### 3.1.3. Death

It was difficult to establish which patients had died and when they died. Checking against publicly available death data [20], we identified an under-recording of death in SID patient-level records. If not indicated as deceased, these patients might inadvertently be categorised as “alive and healthy” due to their lack of healthcare usage, or they might have moved out of the area but been retained on a Sussex GP’s register.

We also found some cases where patients had multiple dates of death attributed to them. This could be attributed to some providers adding a date of notification of death and others the occurrence or even the registration of death. Similarly to an age value, we weighted primary care as likely the most accurate value for the date of death, only using other sources of information if primary care data was missing.

### 3.1.4. Visitors to Sussex

Sussex, being a rural and coastal county, attracts a large number of visitors (around 62 million tourism visits are taken in Sussex annually). Rother, Eastbourne, Hastings, Chichester and Brighton and Hove all have a ratio of tourist trips to the local resident population of more than 40:1 [28]. These visitors may need on-the-day and emergency healthcare. This tourism use of healthcare in Sussex was evidenced by there being more than 3 million SID IDs (based on individual NHS numbers and DOB information) in comparison to an estimated 1.8 million currently GP-registered people across Sussex. Around 2.2 million of the SID IDs were identified as being current or former Sussex residents. Around 40% patients in SID had a single entry in the database; this proportion roughly corresponds to the excess in patient IDs that were found.

Of course, new patients will enter the database (either moving into the area or as babies registering with services), and some patients will die; these patients (as well as those visitor service users) may be of genuine interest, depending on the questions being asked of the dataset. As such, some SID IDs will legitimately have only one record, but it was important to distinguish whether a patient should have just one or a handful of records (i.e., they were a visitor, and as such, can be excluded if analysing case history/health/life events) or whether their history is absent due to errors in data processing or in porting their record between GP practices when they moved.

This was particularly important in our attempts to build a picture of patient journeys, their long-term conditions and multimorbidity based on the evidence of diagnostic codes in encounters/activity in the absence of disease registers; we could have mistakenly attributed a data-omitted/visiting patients as healthy and low service using, inflating the picture of healthy patients within Sussex.

How to distinguish tourists or visitors from “healthy” residents with no service use was identified as an important task. In the processing of the demographic milestones table, the SID ID was checked to see if there was any evidence of a primary care registration (and if that registration was to a Sussex practice) as well as whether the patient’s residence (aggregated to lower-layer super output area (LSOA)) was in the county. Therefore, the demographic milestones table was instrumental in reducing the decision-making burden on analysts to search every raw data record linked to an ID because it sequenced records and automated processes to decide which record was relevant for that time period. The table meant we could quickly make a decision about excluding patients who had missing data, where they were identified as non-residents. This process also helped us explain why our multimorbidity analyses substantially over-represented patients with few or no long-term conditions because we originally included health tourists.

### 3.1.5. Duplication

SID data storage requirements were estimated to grow by 100 GB every 3 months across the different tables of data as new providers came on board and fed historical data as well as new regular feeds. Exploratory work identified that about 10% of the data was duplicated to some degree. New records were sometimes updates of older, perhaps incomplete records. However, new data did not replace old data, it was simply added on top, and we had to adjust our analyses to recognise this and to ensure we were not overcounting patients or events. Scripts to achieve these adjustments (identification and removal of duplicates) were automated so that processing and transformation decisions to choose which records should be counted would subsequently be run automatically on each data update.

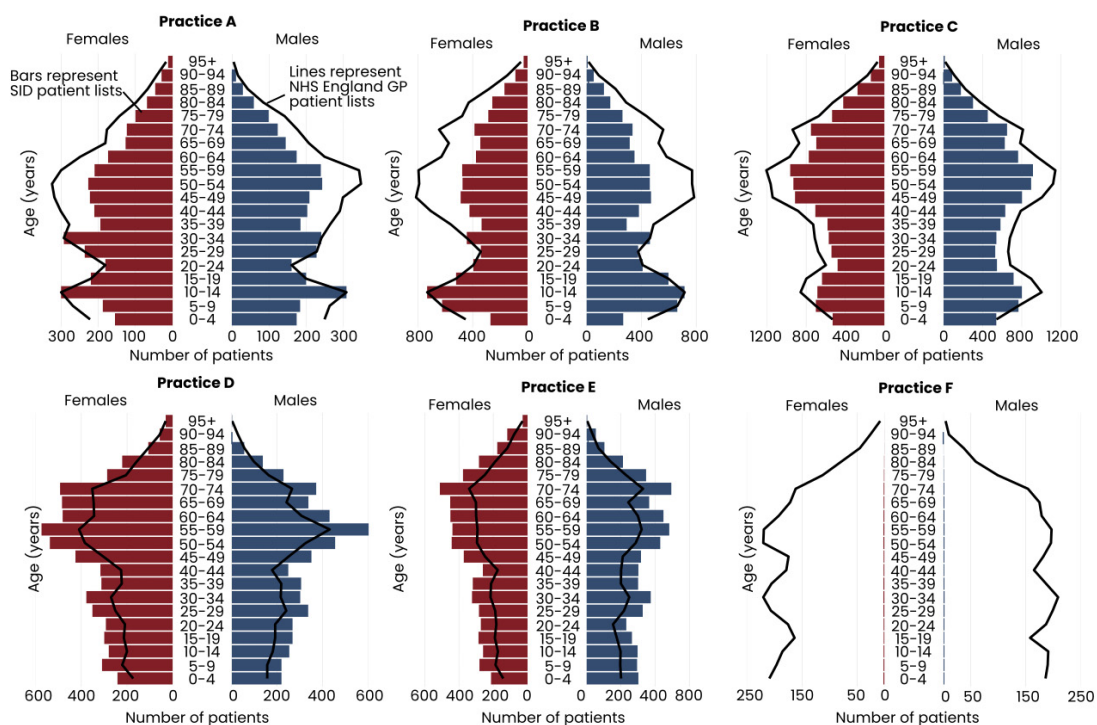
### 3.1.6. GP Practice List Sizes

To understand the completeness of primary care data imports into SID and reconcile who our “resident and registered” populations were, we compared published patient list numbers from NHS Digital [19] with the number of records linked to the same organisation in SID.

This exercise showed that some GP practices were submitting patient lists that were much larger or smaller than expected. For one practice, their original uploads contained records four times the number of patients on lists from NHS Digital. These discrepancies were often the result of incomplete closure or de-registering from one practice to another.

We ran data visualisations to compare the age/sex structure of data imported into SID to published outputs from NHS Digital to further support identifying if there were extra patients in particular age groups (such as older cohorts not being de-registered). The visualisation took the form of population pyramids (Figure 2), with bars representing SID data and lines representing published list size data.





**Figure 2.** Population structure of six primary care organisations in SID compared to published list sizes. SID data (Jan 2021 extract) represented in bars and NHS digital data (Jan 2021 extract) represented in lines. Note: the scales on each pyramid are different.

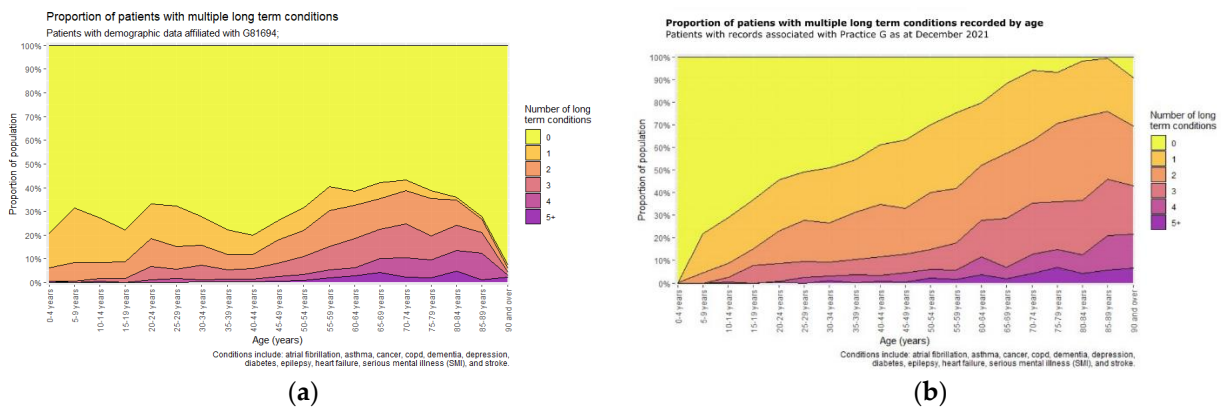
Sharing this report and the analysis code with the data team at SID prompted the team to investigate some of the discrepancies further. These discrepancies led to the SID data team identifying a data processing error linked to practices re-registering existing patients with a start date on a subsequent record before the end date of the initial registration. Once this was rectified, some of the GP practice list sizes came much closer to those expected from NHS Digital data.

### 3.2. Identifying Multiple Long-Term Conditions

As the data quality issues were addressed, we began to create usable datasets to analyse patient journeys and life events with the goal of assessing multimorbidity and patient encounters with health services over their life course.

We found the open-source code lists we initially chose were poor at identifying patients. Searching for the codes in the generic code lists, with the aim of identifying full disease cohorts, led to an initial underrepresentation of patients when matched to expected numbers from outside sources. We found that this was due to conditions being coded with Read or SNOMED codes but usually not both. We also found that symptom and process codes were used rather than established or audited diagnostic codes. After multiple iterations of code lists, we were satisfied that our SID-specific code lists were returning roughly accurate numbers of patients with each condition of interest.

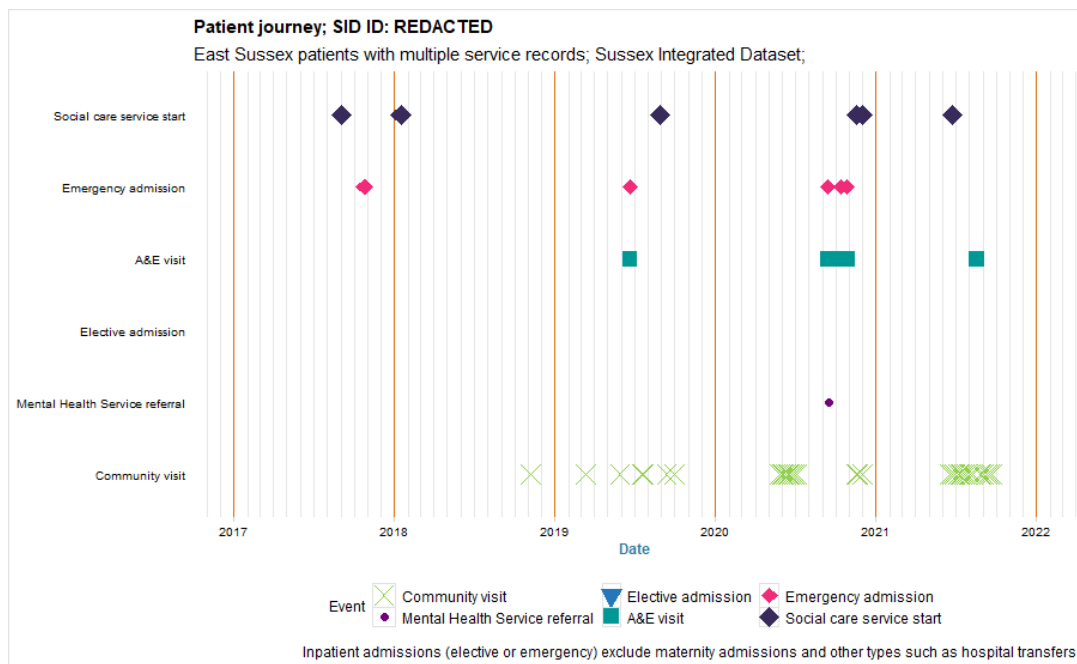
We plotted the proportion of patients at each age with 0, 1, 2, 3, 4, or  $\geq 5$  conditions. Initially, the number of conditions by age was flat and did not increase with age (Figure 3a). Following exploration of the patients with little or no health history (and as such no evidence of diagnoses), likely recent arrivals or visitors to Sussex, as described above, these patients were removed, and a more expected plot was achieved with older age groups having higher numbers of long-term conditions (Figure 3b).



**Figure 3.** (a) Multiple long-term conditions by age group in one GP practice (all contributed patients); (b) Multiple long-term conditions by age group in one GP practice (only SID patients with >1 SID entry).

3.3. Visualisation of Multiple Service Use over Time

We produced multiple plots of service interactions for individual patients (see an example in Figure 4). From feedback from commissioners and potential future users of SID outputs, this was chosen as a novel way of summarising patient histories (and highlighting data omitted patients) and helped to express the breadth of data within SID and how the use of various services interacted. It showcased encounters with all services, including primary care, community services, hospital admissions, and social service referrals.



**Figure 4.** Plot of Service Interaction: showing an example patient’s interaction with selected Sussex services based on data supplied into SID between 2017 and 2022.

This patient journey plot was particularly useful for identifying patterns of events for a small cohort of patients (e.g., focussing on a single organisation, with patients aged 85+ years who had had at least one fall in the previous year and multiple long-term conditions).

#### 4. Discussion and Reflections

This article reports the first attempt by any team to interrogate and look comprehensively at the data imported into SID. This activity resulted in significant contributions to improvements in the way SID imports and structures data and laid the foundation for understanding how to build a semantic layer and data models for the analysis and interpretation of data. Key lessons learned were that we quickly needed a practical way of prioritising the most likely value out of competing demographic variables from various sources. In SID, hundreds of health care providers across multiple data systems were adding patient level data to an expanding “data lake.” In the early days of understanding the data, we spent much energy and time trying to understand which of multiple conflicting values should be used, not just for a single patient but to create rules for automated processing across a whole database. Creating an automated demographics table, which was populated with weighted values for each calendar year, helped both with identifying reliable and stable demographics and also helped with identifying patients who died or who did not live in the county. Creating scripts to achieve this data quality improvement held the team back from achieving their full aims in terms of producing a robust analysis of the population, because SID was at an earlier stage of development than initially realised. One benefit, however, was that the early stage of the SID development meant that changes, new features and processing scripts could be implemented in near real-time. The strong relationships formed with the data team within the ICP, accompanied by their willingness to work collaboratively together to solve problems, meant that whilst we did not accomplish the full set of aims of the project around complex analyses and modelling, we laid the foundations for a longer-term ambition of work with a community of development analysts. Further lessons around data quality were that we could not rely on historical data to contain a complete picture of patient health, and we could not rely on open-source code lists to accurately identify patients with particular conditions in the Sussex systems. These issues were identified, and work was possible to rectify the difficulties by using other published data sources to which we could compare the results of SID data queries. Lastly, we achieved significant progress in understanding the skills and analysis steps required to make use of integrated datasets for public health purposes in Sussex and highlighted technical skills gaps in public health intelligence teams which are now being addressed.

Key analysis techniques developed through this work were: methods for adapting open-source code lists of conditions to ensure completeness, identifying deceased patients or visitors to the county, and using visualisations of the timeline from a single patient identifier or a small cohort of patients to look at similarities and differences in patterns of service use over time. This latter method was limited to small numbers of patients. At scale, capturing the general pattern of several hundred patient journeys visually remains difficult, but conceptual progress was made on how to record the sequence of interactions with service and diagnosis rather than simply noting the co-occurrence of events. For example, some consultations include key decision points by healthcare professionals, and pulling these out from among various events may be key to understanding patients’ overall care pathways within the systems. This technique will be highly valuable for using integrated data to improve joined-up service and system planning. We are continuing to develop methods for quantifying patterns, such as fields denoting the sequence of health events (e.g., two emergency admissions preceding a service referral or number of diagnosed conditions before a health event). This inevitably reduces the nuances of the health histories of individuals but offers a method of analysing patterns at scale and will form the basis of future work.

The project resulted in greater than anticipated learning about the impact of data architecture and information governance decisions on the functionality of an integrated dataset and the ability to carry out analysis for public health purposes, as well as the quality assurance required to develop and maintain integrated datasets, and therefore resulted in a significant contribution to the development of SID as a resource for public health and the NHS in Sussex.

Teams curating other UK data resources have presented limited information on assessing and overcoming their data quality issues in published data resource profiles [29–33], but to our knowledge, few papers have discussed these early quality issues centred around the curation of the database, and their solutions, in-depth. International research has reported on assessments of domains of data quality such as completeness, consistency and accuracy in health records [34–36], but mainly these reports focus on the correspondence between the information in the record and the state of health or illness of the patient. Our paper addresses the data curation and processing issues encountered when a newly linked database of multiple sources of health data is created for the first time. Some studies have reported, like us, that “off-the-shelf” code lists do not find all expected cases of data and that it is hard to interpret the reasons for apparently missing cases of conditions under study [37]; teams have proposed methods for developing more comprehensive code lists which rely on additional or contextual codes [38,39]. Other studies have identified that incorrect or missing time stamps on data entry mean that mapping the process of patients through the healthcare system can be unreliable and have proposed methods for taking this into account [40]. Furthermore, data quality issues with EHR-based research were revealed during the COVID-19 pandemic when two high-profile COVID-19 papers based on EHR data were swiftly retracted [41]. This led a consortium of EHR researchers to write a guide to appraising EHR data quality and aptitude for answering clinical questions [41]. They urge researchers to take notice of issues such as documentation of data handling and data completeness related to data type (e.g., signs and symptoms missing from coded data because they are captured in unstructured clinic notes). Future research, which brings together recognised quality issues and provides frameworks for reporting on data processing and curation steps in health databases (collectively known as data provenance) would be of value to the health data community.

#### *4.1. Strengths and Limitations*

Data quality in SID was at a much lower level than we initially anticipated, substantially limiting our initial aims to use the data for complex longitudinal analyses. One example was the lack of date of birth; the decision not to include DOB in the SID was made on the basis of patient privacy and information governance rules. Age defined as a year was not sensitive enough for many analyses, such as potential years of life lost or healthy life expectancy at any granular level. Instead, we sought to create nominal years of birth and crude ages (in years) to enable certain analyses, such as age at the onset of conditions.

We spent much time trying to identify if we had the right number and age/sex spread of patients in the SID as submitted from general practices. Our team identified substantial unexpected discrepancies between published GP patient lists and patient IDs in the SID. Although identifying these errors took up research time and effort, we were able to share code and findings with the SID data team. The use of annotated R scripts for our exploratory work meant that the data team could follow our annotated analysis step by step and reproduce the results we had seen. They identified an error in data onboarding and processing which was subsequently rectified. Creating annotated and sharable code was, therefore, a strength, and integral to the success of the project and the development of the database.

We noticed further data limitations in the number of historical health events and diagnoses; these were recorded much less than expected in data 5 years or older. Further exploration on whether the rise was due to more accurate recording in recent years and whether lower incidence in previous years was due to an absence of records rather than the absence of an event would be helpful, especially if SID will be used to track the prevalence of long-term conditions and multi-morbidity over time. In the future, as more data is received, there will be greater confidence that the database contains full health and social care activity for a given patient, and the limitations of using historical data for each patient will be more fully understood.

#### 4.2. Key Learning for Public Health and Future Plans

The exploration of data in the SID has allowed the focussing of the types of public health questions which can realistically be asked of these newly linked datasets. For example, dates of condition onset were particularly hard to identify with any confidence, although existence of conditions was fairly easy to establish. Further challenges to overcome to gain meaningful and quality research outcomes for public health planning include understanding the completeness of a patient's history of service use and gaining confidence that the right cohort of patients has been identified without browsing the whole dataset. We hope that with further work, PH teams can use these data sets to understand multimorbidity and complex service use across the Sussex population in a way that has not been done before. To do this, we have identified specific learning needs within the teams for programming and data retrieval skills that will enable complex analyses of linked data. Developing further the partnership between public health and academia can further this learning and bring in new skills and analysis techniques; this may become formalised in a training programme for SID analysts. While the early stages of using these new datasets may be slow, we hope that in time they will become trustworthy and high-quality data sources for identifying communities or groups who would benefit from public health interventions to reduce health inequalities and can be used for evaluating the impact of public health interventions as they are rolled out. Collaborative working with clinicians, commissioners, analysts and PH strategic leads will enable linked data to lead to changes in the health and care system.

#### 4.3. Conclusions

Building a secure and usable health data infrastructure and community, essentially from scratch, is a huge undertaking. In this initial data exploration, we made progress toward understanding linked NHS and social care data in our geography and towards creating processes that will improve its quality for all future uses. We will continue to build and work with our emerging community of university researchers, local authority public health teams, NHS dataset teams, and public representatives to foster methodological and analytical skills and capacity and identify additional data sources to augment health data for public health planning and policy.

**Author Contributions:** Conceptualization, E.F., A.M., V.S.-H. and M.R.-R.; methodology, E.F., A.M., R.T., N.J., J.E., V.S.-H., G.E., J.C. and K.G.; validation, J.E., R.T. and N.J.; formal analysis, R.T. and N.J.; investigation, J.E.; resources, E.F. and M.R.-R.; data curation, J.E.; writing—original draft preparation, E.F. and R.T.; writing—review & editing, N.J., V.S.-H., G.E., J.C., K.G. and A.M.; supervision, E.F., A.M., V.S.-H., G.E., K.G. and J.C.; funding acquisition, E.F., M.R.-R. and A.M. All authors have read and agreed to the published version of the manuscript.

**Funding:** This study was funded by a grant from the National Institute of Health Research Public Health Research Programme (NIHR 133761). The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care.

**Data Availability Statement:** The data in the Sussex Integrated Dataset are currently only available for analysis to employees of joint data controller organisations. These are limited to the member organisations of the Sussex Integrated Care Partnership called “NHS Sussex.”

**Acknowledgments:** We acknowledge the kind contributions of the Data Intensive Science Centre at the University of Sussex for resourcing a “data sprint” event and staff and students at the University of Sussex who gave us early advice on programming techniques for accessing the data.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. The King's Fund. Sustainability and Transformation Plans (STPs) Explained: The King's Fund. 2017. Available online: <https://www.kingsfund.org.uk/topics/integrated-care/sustainability-transformation-plans-explained> (accessed on 15 December 2022).
2. NHS Providers. NO TRUST IS AN ISLAND: A Briefing For Governors on Working Collaboratively in Health and Care Systems: NHS Providers. 2018. Available online: <https://nhsproviders.org/stp-governor-briefing> (accessed on 15 December 2022).
3. NHS Digital. ICS Implementation NHS Digital. 2022. Available online: <https://digital.nhs.uk/services/ics-implementation> (accessed on 15 December 2022).
4. NHS England. Integrated Care Boards. 2022. Available online: <https://digital.nhs.uk/services/organisation-data-service/integrated-care-boards> (accessed on 15 December 2022).
5. Sussex Health and Care. Our Care Connected: Sussex Health and Care. 2022. Available online: <https://www.sussex.ics.nhs.uk/our-vision/priorities-and-programmes/digital/our-care-connected/> (accessed on 15 December 2022).
6. UK Parliament. Local Authorities' Public Health Responsibilities (England) London: House of Commons Library. 2014. Available online: <https://researchbriefings.files.parliament.uk/documents/SN06844/SN06844.pdf> (accessed on 15 December 2022).
7. Department of Health. Local Public Health Intelligence: Department of Health. 2012. Available online: [https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\\_data/file/212959/Public-health-intelligence-all-factsheets.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/212959/Public-health-intelligence-all-factsheets.pdf) (accessed on 15 December 2022).
8. Centers for Disease Control and Prevention National Center for Health Statistics. International Classification of Diseases, Tenth Revision (ICD-10). 2021. Available online: <https://www.cdc.gov/nchs/icd/icd-10-cm.htm> (accessed on 15 December 2022).
9. SNOMED International. Use SNOMED CT. 2022. Available online: <https://www.snomed.org/snomed-ct/Use-SNOMED-CT> (accessed on 15 December 2022).
10. Booth, N. What are the Read Codes? *Health Libr. Rev.* **1994**, *11*, 177–182. [PubMed]
11. Chisholm, J. The Read clinical classification. *Br. Med. J.* **1990**, *300*, 1092. [CrossRef] [PubMed]
12. Stuart-Buttle, C.D.; Read, J.D.; Sanderson, H.F.; Sutton, Y.M. A language of health in action: Read Codes, classifications and groupings. In Proceedings of the A Conference of the American Medical Informatics Association AMIA Fall Symposium, New Orleans, LA, USA, 11–15 November 1996; pp. 75–79.
13. NHS Digital. DAPB0084: OPCS Classification of Interventions and Procedures. Available online: <https://digital.nhs.uk/data-and-information/information-standards/information-standards-and-data-collections-including-extractions/publications-and-notifications/standards-and-collections/dapb0084-opcs-classification-of-interventions-and-procedures> (accessed on 15 December 2022).
14. NHS Digital. Quality and Outcomes Framework, 2020-21: NHS Digital 2021. Available online: <https://digital.nhs.uk/data-and-information/publications/statistical/quality-and-outcomes-framework-achievement-prevalence-and-exceptions-data/2020-21> (accessed on 15 December 2022).
15. NHS Digital. National Diabetes Audit, 2019-20, Type 1 Diabetes: NHS Digital 2021. Available online: <https://digital.nhs.uk/data-and-information/publications/statistical/national-diabetes-audit/national-diabetes-audit-2019-20-type-1-diabetes> (accessed on 15 December 2022).
16. National Cancer Registration and Analysis Service. Welcome to CancerData: CancerData. 2022. Available online: <https://www.cancerdata.nhs.uk/> (accessed on 15 December 2022).
17. Office for Health Improvement and Disparities, NHS Benchmarking Network. Quality Improvement Tool. 2021. Available online: <https://www.cvdprevent.nhs.uk/quality-improvement?period=4> (accessed on 15 December 2022).
18. NHS Digital. Hospital Episode Statistics (HES). 2019. Available online: <https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics> (accessed on 15 December 2022).
19. NHS Digital. Patients Registered at a GP Practice NHS Digital. 2022. Available online: <https://digital.nhs.uk/data-and-information/publications/statistical/patients-registered-at-a-gp-practice> (accessed on 15 December 2022).
20. Office for National Statistics. Deaths Broken down by Age, Sex, Area and Cause of Death. 2022. Available online: <https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths> (accessed on 15 December 2022).
21. Health Data Research in UK. The HDR UK Phenotype Library London UK: Health Data Research UK. 2022. Available online: <https://phenotypes.healthdatagateway.org/> (accessed on 15 December 2022).
22. OpenCodelists. OpenCodelists: Bennett Institute for Applied Data Science, University of Oxford. 2022. Available online: <https://www.opencodelists.org/> (accessed on 15 December 2022).
23. Brighton and Hove City Council. Joint Strategic Needs Assessment Brighton, UK: Brighton and Hove City Council. 2022. Available online: <https://www.brighton-hove.gov.uk/joint-strategic-needs-assessment> (accessed on 15 December 2022).
24. Office for National Statistics. Living Longer: How our Population is Changing and Why it Matters. 2021. Available online: <https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/ageing/articles/livinglongerhowourpopulationischangingandwhyitmatters/2018-08-13#what-are-the-implications-of-living-longer-for-society-and-the-individual> (accessed on 15 December 2022).
25. Imperial College Health Partners. How to Use Theographs to Better Understand Individual Stories and Improve Patient Care. 2019. Available online: <https://imperialcollegehealthpartners.com/gps-and-commissioners-are-increasingly-interested-in-using-theographs/> (accessed on 15 December 2022).

26. NHS Health Research Authority. Guidance for Using Patient Data. 2022. Available online: <https://www.hra.nhs.uk/covid-19-research/guidance-using-patient-data/> (accessed on 15 December 2022).
27. Ford, E.; Rees-Roberts, M.; Stanley, K.; Goddard, K.; Giles, S.; Armes, J.; Ikhile, D.; Madzvamuse, A.; Spencer-Hughes, V.; George, A.; et al. Understanding how to build a social licence for using novel linked datasets for planning and research in Kent, Surrey and Sussex: Results of deliberative focus groups. *Int. J. Popul. Data Sci.* **2023**, *5*, 13. [[CrossRef](#)]
28. Blue Sail. Sussex Visitor Economy Baseline Report. 2021. Available online: <https://www.experiencewestsussex.com/wp-content/uploads/2022/03/Sussex-Visitor-Economy-Baseline-Review.pdf> (accessed on 15 December 2022).
29. Henson, K.E.; Elliss-Brookes, L.; Coupland, V.H.; Payne, E.; Vernon, S.; Rous, B.; Rashbass, J. Data resource profile: National cancer registration dataset in England. *Int. J. Epidemiol.* **2020**, *49*, 16–16h. [[CrossRef](#)] [[PubMed](#)]
30. 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. [[CrossRef](#)] [[PubMed](#)]
31. Herrett, E.; Gallagher, A.M.; Bhaskaran, K.; Forbes, H.; Mathur, R.; van Staa, T.; Smeeth, L. Data resource profile: Clinical practice research datalink (CPRD). *Int. J. Epidemiol.* **2015**, *44*, 827–836. [[CrossRef](#)] [[PubMed](#)]
32. Wolf, A.; Dedman, D.; Campbell, J.; Booth, H.; Lunn, D.; Chapman, J.; Myles, P. Data resource profile: Clinical Practice Research Datalink (CPRD) Aurum. *Int. J. Epidemiol.* **2019**, *48*, 1740–1740g. [[CrossRef](#)] [[PubMed](#)]
33. Lewer, D.; Bourne, T.; George, A.; Abi-Aad, G.; Taylor, C.; George, J. Data Resource: The Kent Integrated Dataset (KID). *Int. J. Popul. Data Sci.* **2018**, *3*, 427. [[CrossRef](#)] [[PubMed](#)]
34. Botsis, T.; Hartvigsen, G.; Chen, F.; Weng, C. Secondary use of EHR: Data quality issues and informatics opportunities. *Summit Transl. Bioinform.* **2010**, *2010*, 1. [[PubMed](#)]
35. Orfanidis, L.; Bamidis, P.D.; Eaglestone, B. Data quality issues in electronic health records: An adaptation framework for the Greek health system. *Health Inform. J.* **2004**, *10*, 23–36. [[CrossRef](#)]
36. Weiskopf, N.G.; Weng, C. Methods and dimensions of electronic health record data quality assessment: Enabling reuse for clinical research. *J. Am. Med. Inform. Assoc.* **2013**, *20*, 144–151. [[CrossRef](#)] [[PubMed](#)]
37. De Lusignan, S.; Hague, N.; van Vlymen, J.; Kumarapeli, P. Routinely-collected general practice data are complex, but with systematic processing can be used for quality improvement and research. *Inform. Prim. Care* **2006**, *14*, 59–66. [[CrossRef](#)] [[PubMed](#)]
38. De Lusignan, S.; Jones, S.; Liaw, S.; Michalakidis, G. Defining datasets and creating data dictionaries for quality improvement and research in chronic disease using routinely collected data: An ontology-driven approach. *Inform. Prim. Care* **2012**, *19*, 127–134. [[CrossRef](#)] [[PubMed](#)]
39. Nicholson, A.; Ford, E.; Davies, K.; Smith, H.; Rait, G.; Tate, R.; Petersen, I.; Cassell, J. Optimising Use of Electronic Health Records to Describe the Presentation of Rheumatoid Arthritis in Primary Care: A Strategy for Developing Code Lists. *PLoS ONE* **2013**, *8*, e54878. [[CrossRef](#)] [[PubMed](#)]
40. Perimal-Lewis, L.; Teubner, D.; Hakendorf, P.; Horwood, C. Application of process mining to assess the data quality of routinely collected time-based performance data sourced from electronic health records by validating process conformance. *Health Inform. J.* **2016**, *22*, 1017–1029. [[CrossRef](#)] [[PubMed](#)]
41. Kohane, I.S.; Aronow, B.J.; Avillach, P.; Beaulieu-Jones, B.K.; Bellazzi, R.; Bradford, R.L.; Brat, G.A.; Cannataro, M.; Cimino, J.J.; García-Barrio, N.; et al. What Every Reader Should Know About Studies Using Electronic Health Record Data but May Be Afraid to Ask. *J. Med. Internet Res.* **2021**, *23*, e22219. [[CrossRef](#)] [[PubMed](#)]

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.