ABSTRACT

Purpose The Scottish Diabetes Research Network (SDRN)-diabetes research platform was established to combine disparate electronic health record data into research-ready linked datasets for diabetes research in Scotland. The resultant cohort, ‘The SDRN-National Diabetes Dataset (SDRN-NDS)’, has many uses, for example, understanding healthcare burden and socioeconomic trends in disease incidence and prevalence, observational pharmacoepidemiology studies and building prediction tools to support clinical decision making.

Participants We estimate that >99% of those diagnosed with diabetes nationwide are captured into the research platform. Between 2006 and mid-2020, the cohort comprised 472,648 people alive with diabetes at any point in whom there were 4 million person-years of follow-up. Of the cohort, 88.1% had type 2 diabetes, 8.8% type 1 diabetes and 3.1% had other types (eg, secondary diabetes).

Findings to date There have been >50 publications using the SDRN-NDS. Examples of recent key findings include analysis of the incidence and relative risks for COVID-19 infection, drug safety of insulin glargine and SGLT2 inhibitors, life expectancy estimates, evaluation of the impact of flash monitors on glycaemic control and diabetic ketoacidosis and time trend analysis showing that diabetic ketoacidosis (DKA) remains a major cause of death under age 50 years. The findings have been used to guide national diabetes strategy and influence national and international guidelines.

Future plans The comprehensive SDRN-NDS will continue to be used in future studies of diabetes epidemiology in the Scottish population. It will continue to be updated at least annually, with new data sources linked as they become available.

STRENGTHS AND LIMITATIONS OF THIS STUDY

⇒ The cohort has nationwide coverage with >99% of all those with diabetes in Scotland. This includes 472,648 individuals from 2006 to 2020.
⇒ The cohort is updated annually with extensive linkage to existing healthcare data sources, negating requirements for de novo data collection.
⇒ Furthermore, it is extendable, with new datasets being easily linked as they are created using the national Community Health Index number.
⇒ The underpinning research data platform facilitates the use of a verifiable research pipeline, as it provides both the originating and cleaned data with a controlled and documented provenance pathway.
⇒ Limitations include the need for cleaning raw data values manually entered at the clinical interface; however, this cleaning is performed consistently during the database creation and not by each research analyst.

INTRODUCTION

In Scotland, a standardised electronic health record (Scottish Care Information-Diabetes (SCI-diabetes)) has been in use for patient care in diabetes since the late 1990s, gaining nationwide coverage by mid-2000s. The record uses a unique healthcare identifier, the Community Health Index (CHI) number, which is also used on all other administrative health datasets in Scotland.1 By linking these datasets together, we sought to generate a nationwide cohort of people with diabetes, updated annually with those newly diagnosed, and rich in a wide range of data.7 Such
a population-wide cohort provides invaluable information for a range of stakeholders. Uses of such data include but are not limited to (1) understanding current disease prevalence, healthcare burden and trends in disease incidence to inform resource allocation, (2) studying disease aetiology, for example, determinants of complications of diabetes including in relation to sex, ethnicity and social deprivation, (3) evaluation of new developments in care, for example, flash glucose monitors, (4) studying the real-world observational pharmacoepidemiology of diabetes drugs on outcomes, (5) understanding the natural history of disease, for example, the progression rates to type 2 diabetes in those with prior gestational diabetes, (6) building prediction tools for decision making, such as cardiovascular disease risk scores, and many more.

However, building a cohort and underpinning a research data platform from electronic healthcare records, as distinct from study-specific data collections as in a clinical trial, for example, brings several challenges. A key issue is how best to organise and control the vast amounts of data received from various sources, each with differing levels of consistency and historical meta-information. Another issue is that there will, in such data, be errors, and extensive data cleaning may be required. There is also a need to provide metadata to users on such extensive datasets, and the data must be held in a way that provides security and privacy. For a wide range of end-users of the database, data must be centrally provisioned in a common, consistent format that ensures the efficiency of the analytic code and provides a scalable, standardised structure for organising data in a way that can answer different research questions concurrently across teams of individuals. Such abstraction of data resources also enables common approaches to be adopted in downstream processes, including cohort generation, data analysis, automatic manuscript generation using R markdown\(^3\) and implementation of reproducible research frameworks.

In this paper, we (the Scottish Diabetes Research Network Epidemiology SDRN-EPI Group) provide a detailed description of the SDRN-diabetes research platform where SCI-Diabetes data (the spine of the database) have been linked to other data. We present details of the resulting cohort, the SDRN-National Diabetes Study (NDS) cohort summarising the data content in the cohort and its characteristics.

**COHORT GENERATION AND CHARACTERISTICS**

**Data sources/diabetes data**

As shown in figure 1, the main source of diabetes data comes from NHS Scotland’s national patient record for diabetes care called Scottish Care Information-Diabetes (SCI-Diabetes). SCI-Diabetes itself is used for delivering patient care in most specialist and some primary care settings, including hospital, adult and paediatric diabetes clinics, podiatry clinics, dietetic clinics, inpatient review, community diabetes and so on. All newly diagnosed persons coded with diabetes in primary care have a record created in SCI-Diabetes. For patients registered on the system, there is an automated nightly feed into SCI-Diabetes of key retrospective and prospective information relevant to diabetes care, including all prescribed drugs from all primary care practices. Key data items including laboratory tests relevant for diabetes management and retinopathy screening and grading outcomes are uploaded to the system via direct data transfer via web services from NHS laboratory data stores and the National Retinopathy screening programme. There are various dashboards and interfaces enabling clinicians to enter data and gain summaries of individuals and their overall clinic or regional population.

We estimate that the coverage of the diabetes non-temporary population with a diabetes diagnosis residing in Scotland by SCI-Diabetes is more than 99%. All general practices nationwide contribute data to the SCI-Diabetes database. Furthermore, in a validation study, we queried all national hospital admission records and prescribing databases in 2018/2019 for any evidence of diabetes and then established whether all such persons have a record in SCI-diabetes. There were just 3228 people (<1% of the total including people on SCI-diabetes who were alive at any point in 2018/2019) with evidence of diabetes but not on SCI-Diabetes. Confirming diabetes registration is an essential step for a person’s diabetes care since it forms the basis for invitation to the national retinal screening programme. Since 2% or less of retinopathy screening invitations are rejected on the basis of an incorrect assignment of diabetes where the person does not have diabetes, the positive predictive value of registration is 98%, and specificity is high.\(^4\)

**Other linked datasets**

Primary care is free at the point of delivery in Scotland. On registering with a primary care physician, all patients in Scotland are assigned a CHI number, which is used as the key identifier on all health record systems across the country. This allows linkage of the primary SCI-Diabetes patient datasets to other key sources of data for research purposes, for example, the Scottish Morbidity Records that cover inpatient and outpatient attendances, maternity and birth hospital data and cancer registry data. Also linked are dispensed drugs and devices, intensive care unit and microbiology data, births and deaths data from National Records of Scotland (NRS). See figure 1 for a full list of datasets. Online supplemental table 1 provides a listing of key variables available in the database.

**Provisioning of data for research and its governance**

Deidentified extracts of data from SCI-Diabetes containing a pseudonymised identifier are provided to the authorised group of research users, the SDRN-EPI group, via an approved, secured safe haven. For the same cohort of individuals, linked datasets are provided by the Public Health Scotland (PHS) Electronic Data Research and Innovation Service group. This is achieved by a transfer of
CHI numbers with their pseudonymised identifier to PHS. Deidentified data containing the pseudonymised identifier and not the CHI are then provided to SDRN-EPI for merging. Regular transfer of data is scheduled from each source, with each provider performing extraction and deidentification before transfer into the SDRN-EPI Safe Haven environment. Deidentification includes pseudonymisation of the CHI number, removal of any identifiable data and reduction in granularity of key dates (eg, date of birth) by resetting each to mid-month.

Access to the Scottish NHS diabetes data sources is granted to the SDRN epidemiology research purposes by approval from the Public Benefit and Privacy Panel for Health and Social Care (reference 1617-0147). All data are held in a secure safe haven environment. All users are trained in data governance and as all processing and computation take place centrally, no export of data from the safe haven environment is permitted. The SDRN epidemiology group is not authorised to secondarily provision data externally; however, researchers who have obtained local R&D sponsorship may contact the SDRN administrator (administrator-sdrn@dundee.ac.uk) regarding collaborations that fall within the remit of the SDRN epidemiology governance structure.

**Diabetes research data platform**

The data transfer process results in several very large flat text data files containing longitudinal point-in-time data for various measures, diagnoses and interventions. On receipt of these data, the SDRN data manager ensures all

---

**Figure 1** Scottish Diabetes Research Network data flow. SCI-DM, Scottish Care Information-Diabetes; SDRN, Scottish Diabetes Research Network.
meta information files are updated and correct. Subsequently, a new research database is generated with a three-stage build process converting the input data into a structured and strongly-typed relational database of longitudinal patient data. This data platform provides abstraction between research data and analysis. This is achieved by implementing two distinct software layers. The first layer performs extraction, transformation and loading (ETL) into a common data resource. The second, the analysis layer, enables research question-specific extraction and transformation from the common data resource. The data ETL is implemented in Python and R and takes disparate data sources, transforming them into a controlled, comprehensive, standardised relational research database with an accompanying electronic metadata dictionary. In the analysis layer, each database is designed to provide research question-specific longitudinal cohort datasets covering all areas of electronic health records through a standard interface with minimal latency. As the data layer is refreshed through time, the original analysis code is executed on the updated resource with minimal modification. The analysis layer is currently implemented in ‘R’, connecting to the data resource via Open Database Connectivity, with libraries and object-oriented code providing mechanisms for cohort definition and analysis dataset generation for the full gamut of epidemiological study designs.

It is useful for other researchers trying to implement such datasets from electronic health records to have a working example of how some of the challenges of the use of electronic health records have been addressed that may be of general use for others in the field. We provide a detailed specification of the database and its construction in the supplemental appendix. This includes an overview of the SDRN-NDS data model in online supplemental figure 1 and details of an object oriented library used for converting database data into a longitudinal research form in online supplemental figure 2.

**Cohort characteristics**

Altogether, the Diabetes Research Platform contains data on 528721 individuals alive and with diabetes in Scotland at any time between 1 January 1984 and 8 April 2020, with data extracted between 3 August 2020 and 5 October 2020. The diabetes electronic health record in Scotland was used in some parts of the country since the mid-1990s but did not reach >95% coverage of the population of Scotland until 2006. For most analyses, therefore, we use the data from 2006 onwards which includes 472648 individuals. Here, we describe the data from the cohort alive with diabetes anytime between 2006 and mid-2020 (the last date on which extraction from raw clinical data occurred).

**Table 1** shows the prevalence of type 1 and type 2 diabetes from years 2006 to 2020. Mid-year population estimates were imported from NRS. Altogether, there were 472648 individuals with diabetes who were alive with diabetes at any time between 2006 and 2020 who were...
included in the cohort. There are 4 million person-years of follow-up.

As shown, 8.8% of those in this cohort have been assigned as having type 1 diabetes. The type of diabetes can be recorded in SCI-Diabetes by multiple sources (primary care physician, secondary care physician, community nurse and so on). Thus, there is a longitudinal record of type for any given person within which there can be consistent or inconsistent type assignment. In the research data platform, we therefore employ an algorithm to check type against other data on prescriptions and date of diabetes onset. For type 1 diabetes, for example, misclassification might be defined by (1) extensive use of oral antidiabetic medication and (2) no continuous insulin therapy within 1 year of diagnosis. Those who are initially assigned as type 2 are reassigned to type 1 only if they have no contrary prescription history, 183 days of insulin prescribed in the year since diagnosis and an age of onset below age 30 years. The application of the algorithm resulted in a reassignment of 10.5% of people in the cohort with an initial label of type 1 being reassigned to type 2 and 0.8% of type 2 being reassigned to type 1. Most of this reassignment refers to those with already prevalent diabetes when the SCI-Diabetes record system was being established. As shown in online supplemental table 2, there is a much lower reassignment of type for incident cases in recent years. Of the cohort, 3.1% have other types of diabetes, as shown in table 2. These comprise, for example, secondary diabetes, gestational diabetes and monogenic diabetes.

A summary of the cohort demographics by diabetes type is provided in table 2. There is a slight excess of males for both types of diabetes. For non-fixed characteristics, we show the median and IQR of values in the cohort, having computed the within-person results over the years that they are observed in the cohort. The average age during the follow-up period is 47 years for type 1 diabetes and 71 years for type 2 diabetes. The average duration of diabetes during the follow-up period is 18 years for type 1 diabetes and 9 years for type 2 diabetes. The geographic distribution follows that of the overall population of Scotland, with the majority of the population residing in the central belt of Scotland between Glasgow in the West and Lothian in the East. The social class categorisation used is the Scottish Index of Multiple Deprivation. This categorises the deprivation score of the area the person lives in. As can be seen particularly for type 2 diabetes, there is a social class gradient with 47% being in the most deprived two quintiles, where 40% would be expected if there were no social class disparity in prevalence. There is substantial missingness for ethnicity, which is optionally self-assigned by the person with diabetes during outpatient and hospital encounters.

Clinical characteristics are summarised in table 3, including the median (IQR) frequency of each measure each year from 2006 to 2020 and the percentage of missingness. As shown for most of these measures, on average, people have at least one reading per year for each year of follow-up. Thus, the database is a very rich source of longitudinal trajectories of these characteristics in diabetes. There is a high level of missingness for low-density lipoprotein (LDL) cholesterol as the default is to measure total cholesterol first, and LDL cholesterol is then measured contingent on the total-cholesterol value. Height is not typically measured annually as expected for adults. For retinopathy, we show the grading on the photographs taken in the national screening programme for which only those aged 12 years and upwards are eligible. Those denoted as attending the eye clinic have previously had gradings to indicate either maculopathy or at least preproliferative retinopathy. Many person-years of follow-up are missing the albuminuria status and foot screening data in part because some point of care tests are not captured into the system, but this can also be caused by clinicians failing to arrange the tests and patients not having an adequate urine sample at their clinic visit. However, there is a high capture of estimated glomerular filtration rate (eGFR) with, on average, at least one measure each year in those with type 1 and 2 measures per year in those with type 2 diabetes.

**PATIENT AND PUBLIC INVOLVEMENT**

The work of SDRN-Epi generating and using the national diabetes research platform is approved by the Public Benefit and Privacy Panel, which includes patient representatives. The Diabetes Informatics and Epidemiology team at the University of Edinburgh hosts a Patients Advisory Committee (PAC) that scrutinises and makes recommendations on the use of the data, priorities for research as well as advising on messaging key findings to the diabetes community. The SDRN also hosts PAC that comments and advises on specific research funding applications using the data.

**FINDINGS TO DATE**

The SDRN-Epi team have published more than 50 papers on the cohort from the National Diabetes Research Platform to date. These papers span a range of topic areas including evaluation of new technologies, modelling to underpin retinopathy screening intervals, complication risk prediction tools, observational pharmacoepidemiology, time trend analyses and much more. Several international collaborations have used the data. More recently, the database has been pivotal in generating data and an evidence base for COVID-19 prevention policies in people living with diabetes. We describe here a selection of the more recently published outputs from the platform.

With two recent analyses, we were able to reassure policymakers in the National Health Service in Scotland that investment in free provision of continuous subcutaneous insulin infusion (CSI) pumps and flash monitors is having an impact on important outcomes. We showed that flash monitor initiation was associated with clinically important reductions in HbA1c, especially in those with worst glycaemic control; an average fall of 15.5 mmol/mol

McGurnaghan SJ, et al. BMJ Open 2022;12:e063046. doi:10.1136/bmjopen-2022-063046
We also showed a striking 40% reduction in diabetic ketoacidosis incidence with flash monitor use. With CSII use, we also observed marked falls in HbA1c, especially in those with high baseline HbA1c, an average fall of 21.0 mmol/mol (1.9% units in those with a baseline >84 mmol/mol within a year of exposure) that was sustained.11 CSII was associated with a 39% reduction in DKA rates and a 33% reduction in severe hospitalised hypoglycaemia. Such data are key inputs to health economic analyses that justify increasing provision for the use of diabetes technologies to improve health outcomes.

We have demonstrated increases in the number of women with existing type 2 diabetes before pregnancy12 however, there are marked increases in birth weight in women with type 1 and type

| Table 2 | Cohort demographics for people included in SCI-diabetes any time between 2006 and 2020 by diabetes type |
|---------|--------------------------------------------------|
|          | Type 1        | Type 2        | Other         | Total diabetes population |
| Total included | 41 814 (8.8) | 416 291 (88.1) | 14 543 (3.1) | 472 648 |
| Sex (female)    | 18 608 (44.5) | 185 265 (44.5) | 6346 (43.6) | 210 219 (44.5) |
| Age (years)     | 47.1 (30.3, 61.5) | 71.3 (61.2, 79.9) | 64.0 (52.2, 74.7) | 69.8 (58.7, 79.0) |
| Age at diabetes diagnosis (years) | 22.0 (11.5, 36.7) | 60.0 (50.6, 69.0) | 56.9 (44.1, 67.9) | 58.4 (47.6, 68.1) |
| Diabetes duration (years) | 18.5 (8.4, 30.0) | 9.2 (4.6, 15.0) | 5.3 (1.9, 10.8) | 9.6 (4.6, 15.8) |
| Follow-up (years since 2006) | 13.5 (6.5, 14.8) | 8.1 (4.2, 12.8) | 5.4 (2.3, 10.3) | 8.3 (4.2, 13.3) |
| Ethnicity       |                  |                |              |                      |
| White           | 33 704 (80.6)   | 301 587 (72.4) | 10 172 (69.9) | 345 463 (73.1)       |
| South Asian     | 426 (1.0)       | 10 047 (2.4)   | 262 (1.8)     | 10 735 (2.3)         |
| Black           | 203 (0.5)       | 1572 (0.4)     | 60 (0.4)      | 1835 (0.4)           |
| Chinese         | 70 (0.2)        | 1313 (0.3)     | 40 (0.3)      | 1423 (0.3)           |
| Other           | 1267 (3.0)      | 14 222 (3.4)   | 425 (2.9)     | 15 914 (3.4)         |
| Unknown         | 6144 (14.7)     | 87 550 (21.0)  | 3584 (24.6)   | 97 278 (20.6)        |
| Health Board    |                  |                |              |                      |
| Greater Glasgow & Clyde | 8394 (20.1) | 89 664 (21.5) | 3413 (23.5) | 101 471 (21.5) |
| Lothian         | 6627 (15.9)     | 56 658 (13.6)  | 2456 (16.9)   | 65 741 (13.9)        |
| Lanarkshire     | 5412 (12.9)     | 53 801 (12.9)  | 1873 (12.9)   | 61 086 (12.9)        |
| Grampian        | 4415 (10.6)     | 40 928 (9.8)   | 1120 (7.7)    | 46 463 (9.8)         |
| Tayside         | 2944 (7.0)      | 33 511 (8.1)   | 1108 (7.6)    | 37 563 (7.9)         |
| Ayrshire & Arran| 3051 (7.3)      | 33 662 (8.1)   | 801 (5.5)     | 37 514 (7.9)         |
| Fife            | 3029 (7.2)      | 30 562 (7.3)   | 843 (5.8)     | 34 434 (7.3)         |
| Highland        | 2644 (6.3)      | 24 877 (6.0)   | 1229 (8.5)    | 28 750 (6.1)         |
| Forth Valley    | 2392 (5.7)      | 23 824 (5.7)   | 625 (4.3)     | 26 841 (5.7)         |
| Dumfries & Galloway | 1320 (3.2) | 13 734 (3.3)  | 446 (3.1)     | 15 500 (3.3)         |
| Borders         | 943 (2.3)       | 9722 (2.3)     | 443 (3.0)     | 11 108 (2.4)         |
| Western Isles   | 285 (0.7)       | 2054 (0.5)     | 57 (0.4)      | 2396 (0.5)           |
| Orkney          | 168 (0.4)       | 1723 (0.4)     | 55 (0.4)      | 1946 (0.4)           |
| Shetland        | 186 (0.4)       | 1540 (0.4)     | 58 (0.4)      | 1784 (0.4)           |
| Deprivation index |                  |                |              |                      |
| Quintile 1 (most deprived) | 8524 (20.4) | 99 606 (23.9) | 3598 (24.7) | 111 728 (23.6) |
| Quintile 2      | 8392 (20.1)     | 95 063 (22.8)  | 3215 (22.1)   | 106 670 (22.6)       |
| Quintile 3      | 7798 (18.6)     | 83 471 (20.1)  | 2959 (20.3)   | 94 228 (19.9)        |
| Quintile 4      | 7301 (17.5)     | 71 981 (17.3)  | 2486 (17.1)   | 81 768 (17.3)        |
| Quintile 5 (least deprived) | 6693 (16.0) | 56 744 (13.6) | 1970 (13.5)  | 65 407 (13.8)  |
| Unknown         | 3106 (7.4)      | 9426 (2.3)     | 315 (2.2)     | 12 847 (2.7)         |

Categorical values are shown in N (%) and continuous values are median IQR across the cohort in the full period. Number of measures are median IQR across the cohort by year. The follow-up period from 2006 to 2020 includes 14% incident cases of diabetes and 13% who died. SCI-diabetes, Scottish Care Information-diabetes.
Table 3  Cohort summary clinical measurements from 2006 to 2020 by diabetes type

|                          | Type 1       | Type 2       | Other        | Total diabetes population | Missing |
|--------------------------|--------------|--------------|--------------|---------------------------|---------|
| HbA1c measures (yearly)  | 2.0 (1.0, 3.0)| 2.0 (1.0, 2.0)| 1.0 (<1, 2.0) | 2.0 (1.0, 2.0)            | 1.2     |
| HbA1c (mmol/mol)         | 68 (58, 80)  | 55 (47, 68)  | 52 (43, 69)  | 56 (48, 70)               |
| HbA1c (%)                | 8.37 (7.46, 9.52) | 7.18 (6.45, 8.37) | 6.95 (6.08, 8.46) | 7.27 (6.52, 8.51) | 11 |
| Height measures (yearly) | 1.0 (<1, 2.0) | <1 (<1, 1.0) | <1 (<1, 1.0) | <1 (<1, 1.0)             | 2.2   |
| Height (m)               | 1.70 (1.62, 1.77) | 1.67 (1.60, 1.75) | 1.68 (1.60, 1.75) | 1.68 (1.60, 1.75) | 2.5   |
| Weight measures (yearly) | 2.0 (1.0, 3.0) | 1.0 (1.0, 2.0) | 1.0 (<1, 2.0) | 1.0 (1.0, 2.0)            | 1.3 |
| Weight (kg)              | 76 (64, 89)  | 84 (71, 98)  | 77 (64, 91)  | 83 (70, 97)               | 1.3   |
| BMI measures (yearly)    | 1.0 (1.0, 2.0) | 1.0 (<1, 2.0) | 1.0 (<1, 2.0) | 1.0 (<1, 2.0)            | 6.3   |
| BMI (kg/m²)              | 26 (23, 30)  | 30 (26, 34)  | 27 (23, 32)  | 29 (26, 34)               | 30.2  |
| Systolic BP measures (yearly) | 2.0 (1.0, 3.0) | 2.0 (1.0, 3.0) | 1.0 (<1, 3.0) | 2.0 (1.0, 3.0)            | 2.4   |
| Systolic BP (mm Hg)      | 130 (120, 141)| 133 (123, 142)| 131 (120, 141)| 133 (123, 142)            | 2.5   |
| Diastolic BP measures (yearly) | 2.0 (1.0, 3.0) | 2.0 (1.0, 3.0) | 1.0 (<1, 3.0) | 2.0 (1.0, 3.0)            | 2.0   |
| Diastolic BP (mm Hg)     | 76 (69, 82)  | 76 (70, 81)  | 77 (70, 82)  | 76 (70, 81)               |
| HDL cholesterol measures (yearly) | 1.0 (<1, 1.0) | 1.0 (<1, 2.0) | 1.0 (<1, 1.0) | 1.0 (<1, 2.0)            | 1.0   |
| HDL cholesterol (mmol/L) | 1.4 (1.2, 1.8) | 1.1 (1.0, 1.4) | 1.2 (1.0, 1.5) | 1.2 (1.0, 1.4)            | 1.4   |
| LDL cholesterol measures (yearly) | <1 (<1, 1.0) | <1 (<1, 1.0) | <1 (<1, 1.0) | <1 (<1, 1.0)            | 1.0   |
| LDL cholesterol (mmol/L) | 2.3 (1.8, 3.0) | 2.0 (1.5, 2.7) | 2.1 (1.6, 2.8) | 2.0 (1.5, 2.7)            | 1.5   |
| Total cholesterol measures (yearly) | 1.0 (<1, 2.0) | 1.0 (1.0, 2.0) | 1.0 (<1, 2.0) | 1.0 (<1, 2.0)            | 1.0   |
| Total cholesterol (mmol/L) | 4.5 (3.8, 5.2) | 4.1 (3.4, 4.9) | 4.3 (3.6, 5.1) | 4.1 (3.5, 4.9)            | 4.9   |
| eGFR measures (yearly)   | 1.0 (<1, 2.0) | 2.0 (1.0, 3.0) | 1.0 (<1, 3.0) | 2.0 (1.0, 3.0)            | 3.0   |
| eGFR (mL/min/1.73 m²)    | 97 (77, 114) | 75 (54, 91)  | 85 (66, 100) | 77 (56, 93)               |
| Albuminuric status       |              |              |              |                          |        |
| Grading frequency (yearly) | <1 (<1, 1.0) | 1.0 (<1, 1.0) | <1 (<1, 1.0) | 1.0 (<1, 1.0)            | 1.0   |
| Normal                   | 19272 (46.1) | 185021 (44.4)| 5692 (39.1)  | 209985 (44.4)             |       |
| Micro                    | 7332 (17.5)  | 98402 (23.6) | 2381 (16.4)  | 108115 (22.9)             |       |
| Macro                    | 2342 (5.6)   | 24635 (5.9)  | 567 (3.9)    | 27544 (5.8)               |       |
| Unknown                  | 12868 (30.8) | 108233 (26.0)| 5903 (40.6)  | 127004 (26.9)             |       |
| Retinopathy              |              |              |              |                          |        |
| Grading frequency (yearly) | 1.0 (<1, 1.0) | 1.0 (<1, 1.0) | 1.0 (<1, 1.0) | 1.0 (<1, 1.0)            | 1.0   |
| None                     | 14659 (35.1) | 25744 (61.8) | 8962 (61.6)  | 281069 (59.5)             |       |
| NPDR—mild/background     | 10828 (25.9) | 59757 (14.4) | 1540 (10.6)  | 72125 (15.3)              |       |
| NPDR—severe maculopathy | 1141 (2.7)   | 3512 (0.8)   | 81 (0.6)     | 4734 (1.0)                |       |
| NPDR—severe              | 484 (1.2)    | 2334 (0.6)   | 52 (0.4)     | 2870 (0.6)                |       |
| PDR—proliferative        | 73 (0.2)     | 398 (0.1)    | <10 (<1)*    | <482 (0.1)*               |       |
| Not eligible             | 1335 (3.2)   | 25 (<1)      | 35 (0.2)     | 1395 (0.3)                |       |
| Unknown                  | 3404 (8.1)   | 53982 (13.0) | 3042 (20.9)  | 60428 (12.8)              |       |
| Tobacco smoking status   |              |              |              |                          |        |
| Current smoker           | 8233 (19.7)  | 66863 (16.1) | 3300 (22.7)  | 78396 (16.6)              |       |
| Ex-smoker                | 16058 (38.4) | 218246 (52.4)| 5866 (40.3)  | 24017 (50.8)              |       |
| Never smoked             | 15642 (37.4) | 129463 (31.1)| 4538 (31.2)  | 149643 (31.7)             |       |
| Unknown                  | 1881 (4.5)   | 1719 (0.4)   | 839 (5.8)    | 4439 (0.9)                |       |

Continued
2 diabetes. Rates of stillbirth were 4 and 5 times those of the background population in women with type 1 and type 2 diabetes, respectively. We have further explored the importance of glycaemic control and adiposity in stillbirth.

In a recent time trends analysis, we focused on trends in mortality under the age of 50 years, as overall mortality trends are overwhelmingly determined by cardiovascular disease trends in older persons. Yet, young deaths contribute enormously to overall years-of-life lost. We showed that absolute mortality has fallen, but the relative impact of type 1 diabetes on mortality below 50 years has not improved; the standardised mortality ratio relative to the background population was approximately stable at 3.1 and 3.6 in men and 4.09 and 4.16 in women for 2004 and 2017, respectively. Diabetic ketoacidosis or coma deaths accounted for 22% of deaths under age 50 years and the rate did not decline significantly in that period. The vast majority of such deaths (79.3%) occurred out of hospital, emphasising the need for community recognition and prevention of DKA. This work influenced the recent Scottish Government Diabetes Improvement Plan for the next 5 years with the launch of a new DKA national education campaign.

During the first wave of the COVID-19 pandemic, we quickly produced a report for Government and Diabetes Charity stakeholders, later published as a manuscript, showing elevated relative risks of severe COVID-19 in those with type 1 (2.4-fold) and type 2 diabetes (1.4-fold). Before that, most estimations of the risks were simple descriptions of the proportions of hospitalised patients with diabetes. We showed that there was wide variation in risk in those with diabetes and that risk was highly predictable (C-statistic 0.89), and we produced a tool (https://diabepi.shinyapps.io/covidrisk/) to facilitate conversations on COVID-19 risk between clinician and their patients. The data we produced were pivotal in undertaking retinal screening every 2 years rather than every year for those with two baseline reports of no retinopathy. This has led to a change in the National DRS policy in Scotland. SDRN data have also been the first comprehensive national data to demonstrate a reduction in amputation rates with a 29.8% reduction in all amputations for people with diabetes between the years of 2004 and 2008. In addition, SDRN data have allowed Scotland to be the first country to report on comprehensive national data on the incidence of foot ulceration at 1.1%, with first time ulceration at 0.7%. People with foot ulcers are 2–5-fold more likely to die than to undergo amputation, and those with high risk feet are 9-fold more likely to die than undergo amputation which has major implications for health planning.

Other examples of recent work include descriptions of:
1. marked and widening socio-economic inequalities in type 2 diabetes prevalence in Scotland
2. prevalence of remission of type 2 diabetes
3. variation in glycaemic control of type 1 diabetes by age and national/regional data sources

**STRENGTHS AND LIMITATIONS**

The strengths of this cohort are its large size (over 2 billion health data records from over 472,648 individuals to date), the nationwide coverage, the long period of follow-up, the frequency and, by definition, completeness of capture of data items given comprehensive coverage of electronic records. Other key strengths include the extensive data linkages to other datasets and that the data are regularly updated. A major strength is that this is built on existing healthcare data and does not require any de novo data collection.

Furthermore, it is extendible, with new datasets being easily linked as they get created by using the national CHI number. An example of this was the rapid recent linkage to national virology to capture all SARS-CoV-2 tests done nationally.

Key strengths of the underpinning research data platform and attendant tools are that it encapsulates much of the required cleaning and complexity away from the end user. It presents metadata simply; it has in-built source code control, it allows rapid creation of the necessary longitudinal subsets of records for a given analysis and

### Table 3 Continued

|                  | Type 1 | Type 2 | Other | Total diabetes population | Missing |
|------------------|--------|--------|-------|---------------------------|---------|
| Categorical values are shown in N (%) and continuous values are median IQR across the cohort in the full period. Number of measures are median IQR across the cohort in the full period. Missingness is the percentage of the cohort missing a measure in the full period. Categorical values are shown as unknown for missing non-routine measures. Normal albuminuria is an albumin/creatinine ratio <30, micro is 30–300 and macro is >300 mg/L. Please see the supplemental material for an explanation of retinopathy grading.
| Disclosure control applied for small number of individuals
| BMI, body mass index; BP, blood pressure; DKA, Diabetic Ketoacidosis; eGFR, Estimated Glomerular Filtration Rate; GP, General Practitioner; HDL, High-density lipoprotein; LDL, Low-density lipoprotein; NPDR, Nonproliferative Diabetic Retinopathy; PDR, Proliferative Diabetic Retinopathy. |
it facilitates the use of a verifiable research pipeline as it offers full traceability to originating precleaned data.

Limitations include the inherent limitations of basing a cohort on electronic health records. There will inevitably be incorrect raw data values entered at the clinical interface that require cleaning, along with changes to lab reference ranges within various health boards, incomplete metadata and inconsistent data due to new systems being introduced in the earlier years. Another challenge is that for key data concepts, the underpinning raw data source for example, assay method and normal range may change over time. For example, albuminuria status might be measured by albumin concentrations or albumin creatinine ratios at differing points in time, and how this is handled must be captured in the metadata. Another limitation is that we are dependent on the timescales of upstream data providers; ideally, we would like to refresh the data every few months, but currently, it typically happens annually. Since the cohort is limited to people with a current or previous diabetes diagnosis, any analysis requiring a non-diabetic comparative group will require further linkage to the general population without a history of diabetes. Finally, many cohort studies with dedicated data collection systems will use the health record as the gold standard or ‘ground truth’ against which to check the accuracy of their data. Here, we are using this gold standard health record itself as the data source and, therefore, must use internal consistency and validity checks, as exemplified by our diabetes type algorithm, to establish ground truth.

Author affiliations
1MRC Institute of Genetics and Cancer, The University of Edinburgh, Edinburgh, UK
2Nine uwells Hospital, The Scottish Diabetes Research Network, Dundee, UK
3School of Health and Life Sciences, Glasgow Caledonian University, Glasgow, UK
4Institute of Cardiovascular and Medical Sciences, University of Glasgow, Glasgow, UK
5Edinburgh Centre for Endocrinology, Western General Hospital, Edinburgh, UK
6Institute of Applied Health Sciences, University of Aberdeen, Aberdeen, UK
7BHF Glasgow Cardiovascular Research Centre, University of Glasgow, Glasgow, UK
8Institute of Health and Wellbeing, University of Glasgow, Glasgow, UK
9Department of Medicine, Ninewells Hospital and Medical School, Dundee, UK
10Division of Molecular & Clinical Medicine, School of Medicine, University of Dundee, Dundee, UK
11Usher Institute, College of Medicine and Veterinary Medicine, The University of Edinburgh, Edinburgh, UK
12Department of Public Health, NHS Fife, Kirkcaldy, UK

Twitter Naveed Sattar @MetaMedTeam

Acknowledgements We acknowledge with gratitude the contributions of people with diabetes, NHS staff and organisations (the Scottish Care Information-Diabetes development team and Steering Group, the Scottish Diabetes Group, the Scottish Diabetes Survey Group, the diabetes managed clinical networks) involved in providing data, setting up, maintaining and overseeing collation of data for people with diabetes in Scotland. Data linkage is performed by colleagues at Public Health Scotland. We acknowledge the financial support of NHS Research Scotland (NRS), through Diabetes Network. The NHS Research Scotland Diabetes Network (formerly Scottish Diabetes Research Network) receives funding from the Chief Scientist Office of the Scottish Government.

Collaborators The SDRN-EPI team welcomes external collaborative research proposals that use the research data platform. Such proposals are welcomed by academic researchers, commercial entities and other stakeholders, for example, policymakers. There must be a valid scientific question in all collaborations, and no right of veto over the publication of results will be granted.

Contributors SP is Clinical Lead for SCI-Diabetes. SMcG, LAKB, PMcK and HC designed the platform. SMcG, LAKB and HC performed data transformation. SW, LAKB and HC obtained ethical and governance approvals. TC, JM, AB, AC, NS, JMck, JP, SP, RL, KH, DMcA, GL and EP commented on database design and contributed to source data generation. SMcG and LAKB conducted the analyses in this manuscript. SMcG and HC wrote the initial draft of this manuscript. TC, JM, AB, AC, NS, JMck, JP, SP, RL, KH, DMcA, GL, SW and EP edited the manuscript and revised it critically for important intellectual content. HC is responsible for the overall content as the guarantor. All authors approved the final version and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Funding Work described here has been supported by Chief Scientist Office Scotland (Ref. EIM/47) and by Diabetes UK (Ref. 17/0005627).

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement The SDRN-EPI team welcomes external collaborative research proposals that use the research data platform. SDRN-EPI are not data custodians and are not permitted to directly provision data externally. However, the component datasets can be obtained by data governance trained bone fide researchers through the Public Benefit and Privacy Panel for Health and Social Care. See https://www.informationgovernance.scot.nhs.uk/ppbphsc/ for how to apply.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, any changes made indicated, and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/4.0/.

ORCID ids
Stuart J. McGurnaghan http://orcid.org/0000-0002-3292-4633
Luke A. K. Blackbourn http://orcid.org/0000-0002-4234-8040
Thomas M. Carapoppa http://orcid.org/0000-0001-9009-9179
Joseph Mellor http://orcid.org/0000-0003-1452-887X
Anna Barnett http://orcid.org/0000-0002-6345-0194
Andy Collier http://orcid.org/0000-0001-5220-6244
Naveed Sattar http://orcid.org/0000-0002-1604-2593
John McKnight http://orcid.org/0000-0002-8214-7625
John Petrie http://orcid.org/0000-0002-4894-9819
Sam Philip http://orcid.org/0000-0001-6164-211X
Robert Lindsay http://orcid.org/0000-0002-9868-5217
Katherine Hughes http://orcid.org/0000-0001-5278-5282
David McAllister http://orcid.org/0000-0003-3550-1764
Graham P Leese http://orcid.org/0000-0003-0570-5678
Ewan R Pearson http://orcid.org/0000-0001-9237-8585
Sarah Wild http://orcid.org/0000-0001-7824-2569
Paul M McGuire http://orcid.org/0000-0002-5217-1034
Helen M Colhoun http://orcid.org/0000-0002-8345-3288

REFERENCES
1 Womersley J. The public health uses of the Scottish community health index (ch). J Public Health 1996;18:485–72.
2 Scottish Diabetes Research Network Epidemiology Group. SDRN-nds national diabetes dataset. Available: https://orcid.org/0000-0002-3365-3441 [Accessed 15 Mar 2022];
3. Allaire JJ, Xie J, et al. Rmarkdown: dynamic documents for R, 2022.

4. Livingstone SJ, Levin D, Looker HC, et al. Estimated life expectancy in a Scottish cohort with type 1 diabetes, 2008-2010. *JAMA* 2015;313:37-44.

5. Van Rossum G, Drake Jr FL. *Python tutorial*. The Netherlands: Centrum voor Wiskunde en Informatica Amsterdam, 1995.

6. R Core Team. R: a language and environment for statistical computing. Online. Available: https://www.R-project.org/ [Accessed 11 Nov 2020].

7. National Records of Scotland. *Population estimates time series data*, 2021. https://www.nrscotland.gov.uk/files//statistics/population-estimates/mid-20/mid-year-pop-est-20-time-series-5.xlsx.

8. Scottish Government. Scottish index of multiple deprivation. Available: https://www.gov.scot/collections/scottish-index-of-multiple-deprivation-2020/ [Accessed 11 Nov 2020].

9. Scottish Diabetes Research Network. Scottish diabetes research network publications. Available: https://www.ed.ac.uk/mrc-human-genetics-unit/research/colhoun-group/sdm-type1-bioresource/scottish-diabetes-research-network-sdrn/ [Accessed 24 Feb 2022].

10. Jeyam A, Gibb FW, McKnight JA, et al. Flash monitor initiation is associated with improvements in HbA1c levels and DKA rates among people with type 1 diabetes in Scotland: a retrospective nationwide observational study. *Diabetologia* 2021;64:1320–31.

11. Jeyam A, Gibb FW, McKnight JA, et al. Marked improvements in glycaemic outcomes following insulin pump therapy initiation in people with type 1 diabetes: a nationwide observational study in Scotland. *Diabetologia* 2022;65:159–72.

12. Mackin ST, Nelson SM, Kerssens JJ, et al. Diabetes and pregnancy: national trends over a 15 year period. *Diabetologia* 2018;61:1081–8.

13. Mackin ST, Nelson SM, Wild SH, et al. Factors associated with stillbirth in women with diabetes. *Diabetologia* 2019;62:1938–47.

14. O’Reilly JE, Blackbourn LAK, Caparrotta TM, et al. Time trends in deaths before age 50 years in people with type 1 diabetes: a nationwide analysis from Scotland 2004-2017. *Diabetologia* 2020;63:1626–36.

15. Scotland Government. Diabetes care - diabetes improvement plan 2021 to 2026. Available: https://www.gov.scot/publications/diabetes-improvement-plan-diabetes-care-scotland-commitments-2021-2026/ [Accessed 21 Feb 2022].

16. McGurnaghan SJ, Weir A, Bishop J, et al. Risks of and risk factors for COVID-19 disease in people with diabetes: a cohort study of the total population of Scotland. *Lancet Diabetes Endocrinol* 2021;9:82–93.

17. Leese GP, Stratton IM, Land M, et al. Progression of diabetes retinal status within community screening programs and potential implications for screening intervals. *Diabetes Care* 2015;38:889–94.

18. Kennon B, Leese GP, Cochrane L, et al. Reduced incidence of lower-extremity amputations in people with diabetes in Scotland: a nationwide study. *Diabetes Care* 2012;35:2588–90.

19. Chamberlain RC, Fleetwood K, Wild SH, et al. Foot ulcer and risk of lower limb amputation or death in people with diabetes: a national population-based retrospective cohort study. *Diabetes Care* 2022;45:83–91.

20. Vadiveloo T, Jeffcoate W, Donnan PT, et al. Amputation-free survival in 17,353 people at high risk for foot ulceration in diabetes: a national observational study. *Diabetologia* 2022;65:159–72.

21. Prigge R, McKnight JA, Wild SH, et al. International comparison of glycaemic control in people with type 1 diabetes: an update and extension. *Diabet Med* 2022;39:e14766.

22. Captieux M, Fleetwood K, Kennon B, et al. Epidemiology of type 2 diabetes remission in Scotland in 2019: a cross-sectional population-based study. *PLoS Med* 2021;18:e1003828.

23. Wang J, Wild SH. Marked and widening socioeconomic inequalities in type 2 diabetes prevalence in Scotland. *J Epidemiol Community Health* 2021, doi:10.1136/jech-2021-217747. [Epub ahead of print: 11 Oct 2021].
Supplementary appendix

Diabetes Research Data Platform

Summary

The data transfer process results in several very large flat text data files containing longitudinal point in time data for various measures, diagnoses, and interventions. Upon receipt of this data, the SDRN data manager ensures all meta information files are updated and correct. Subsequently, a new research database with a three-stage build process converting the input data into a structured and strictly typed relational database of longitudinal patient data is generated. This data platform provides abstraction between research data and analysis. This is achieved by implementing two distinct software layers. The first layer performs extraction, transformation and loading (ETL) into a common data resource. The second analysis layer performs research question-specific extraction and transformation from the common data resource. The data ETL is implemented in Python[1] and R[2], and takes disparate data sources, transforming them into a controlled, comprehensive, standardised relational research database with an accompanying electronic meta data dictionary. In the analysis layer, each database is designed to provide research question specific longitudinal cohort datasets covering all areas of electronic health records through a standard interface with minimal latency. As the data layer is refreshed through time, the original analysis code is executed on the updated resource with minimal modification. The analysis layer is currently implemented in ‘R’, connecting to the data resource via Open Database Connectivity (ODBC), with libraries and object oriented code providing mechanisms for cohort definition and analysis dataset generation for the full gamut of epidemiological study designs.

As this may be useful for other researchers trying to implement such databases from electronic health records, we provide a detailed specification of the databasing process below.

Clinical Coding Systems

In the United Kingdom, several clinical coding systems have been introduced and subsequently superseded from as early as the 1950s to the present day. These coding systems ensure that a common dictionary or vocabulary is used across clinical systems and archives, thus minimising freeform text entry in digital records, and reducing misclassification. It is because of these coding systems that it is possible to categorise and organise specific areas of the electronic health records for research purposes.

For drugs and medical devices, a plethora of coding systems have been employed and our database is able to handle these all. First, British National Formulary (BNF) codes were introduced in 1949 as a reference for prescribing healthcare professionals in the UK. This was followed by the Anatomical Therapeutic Classification (ATC) codes in 1976 for international drug research purposes. Read Codes were introduced in the
1980s as clinical terminology in primary and secondary care, the latter of which became the standard in the NHS until recently. All NHS systems in England have now adopted a comprehensive system called SNOMED-CT dm+d in 2020. Although Scotland’s timetable for full adoption of this standard remains to be finalised, the diabetes research platform has already implemented this system as the research standard.

For diagnoses and procedure codes, the International Classification of Disease (ICD-10, soon to be ICD-11, diagnoses) and Office of Population Censuses and Surveys’ Classification of Surgical Operations (OPCS procedures) are used. Again, different releases of these coding systems have been introduced through time, with the diabetes research platform providing current ICD-10 and OPCS-4 mappings along with historical ICD-9 and OPCS-3 for older records.

For each release of the diabetes research platform, the most recent terminology dictionaries and mappings are sourced from the NHS’s Terminology Reference Update Distribution (TRUD) service and from the World Health Organisation (WHO). These reference libraries and mappings are combined into a framework that provides a hierarchical path-based notation, enabling pre-defined high-level, non-vendor specific drugs, chemical compound based drug class and phenotypic selection. Selection of any drug class, phenotype or sub-phenotype within the hierarchy automatically selects associated drugs, devices, and diagnostic codes across the various clinical coding systems, resulting in a comprehensive code list for each phenotype or drug class. The code lists are used to build longitudinal views of people being in receipt of the drug or reaching a particular outcome. The classes and phenotypes are developed either from existing definitions, or by agreed definitions provided by an expert group of clinicians, epidemiologists, and data scientists. Having these fixed definitions ensures a consistent, reliable, and reusable approach to drug and phenotype definitions within the database and thus across research groups. This enables the rapid generation of cohorts and association testing based on the generated code lists.

Data Layer

To achieve extraction and transformation, the Diabetes Research Platform uses a three stage process coded in Python for importing, cleaning and generating derivative data. The stages are managed using a controller program with its own backend database which contains controlled recipes for each research database build release. Each recipe contains the sequence of import, cleaning, and algorithmic modules required to convert the disparate source data to the final research database. The modules themselves are coded in a combination of Python and R with the stages outlined as follows:
Data Import

The first stage deals with the import of the data into an electronic database. The process handles any localisation issues such as date and time formatting, inconsistent fields, field counts and file encoding issues. Dictionary files provide mappings of dataset fields and data types by way of a standardised meta-information format. The goal of this initial stage is to maximise data entry, while minimising evaporation of data when changing from disparate semi-structured data to relational database table. After the processing and import is completed, the result is a series of source data tables that are classed as ‘dirty’ or ‘red’ tables. They are not yet in a state suitable for research but are available for further processing and traceability/validation purposes.

Data Cleaning & Validation

The first phase of data processing takes place in the clinical dataset and further information, explaining that manual chart review is no longer undertaken, is provided in previous publications [3,4].

The key component of quality control in the clinical SCI-diabetes dataset arises from the fact that it is used as the basis for retinopathy screening. This means that there is a major incentive for clinicians to ensure that diagnostic codes for diabetes are applied so that the eligible population (people over 11 years of age) are invited for retinopathy screening. Receipt of an inappropriate invitation to retinopathy screening by a person who does not have diabetes as a consequence of erroneous coding provides an opportunity to correct their record. At present validation of diabetes status against prescribing data has only been feasible in the research extract of SCI-diabetes, but there are plans to develop this within primary care data.

The SCI-diabetes system supports quality improvement at several levels, including Health Board (14 across Scotland), hospital clinic level, general practice (GP) level (approximately 1000 across Scotland) or GP cluster level. GP clusters generally include five to eight GP practices in a close geographical location and there is a Cluster Quality Lead. SCI-diabetes users have access to a dashboard that allows comparison of performance in terms of completion of processes of care and treatment outcomes with other regions/domains of care. Users can also run queries, for example, to identify patients under their care that have not received a particular process of care. The dashboard has been designed to align with the nine key processes of care and treatment outcomes identified by the National Institute of Health and Care Excellence[5] and in Scotland’s Diabetes Improvement Plan [6].

Further cleaning of the clinical data is performed at the research extract level. At this stage, a controller program applies a series of simple, followed by more complex modular cleaning programs which not only clean the data based on a series of rules but also provide summary of the results for QC purposes, which are then automatically added to the data dictionary. The process follows the quality control steps outlined by the ODHSI[7], but integrates these during the processing of the data, rather than
using post-processing steps. The conversion and quality control programs perform tasks as varied as plausibility checks, to complex within-person longitudinal cleaning by identifying potentially spurious data by, for example, ensuring that impossible or out-of-range-given-the context values are removed, so as not to disrupt any subsequent analysis downstream. The result of all cleaning and QC steps is a series of clean, ‘green tables’, which are used by the analysis code layer.

**Derivative Data Generation**

After import and cleaning, derivative data is generated. Derivative fields are based on more complex algorithms which require several cleaned source fields to provide a pre-defined derivative or phenotype which will be used by more than one researcher. These derivative fields generally originate from the research layer and promoting such phenotypes to the data platform ensures that each subsequent implementation in other analyses follows the original specification.

**Diabetes Research Database Data Model**

Since the early 2000’s, data models from standards such as the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM)[8] and Clinical Data Interchange Standards Consortium Operational Data Model (CDISC ODM)[9] have been available for adoption by clinical systems to store and provision structured clinical data for research purposes. However, these standards are complex, and implementation often demands a holistic approach across clinical systems, requiring large scale projects which can take many years and considerable cost to realise. For those countries and organisations with healthcare systems that have not yet fully adopted such standards, a simpler, scalable, intermediate approach for integrating existing electronic health record data for research purposes is required. Accordingly the database follows a simplified five table Observational Medical Outcomes Partnership (OMOP)-like format as shown in Supplementary Figure 1, with four of the five tables structured in very long form with many millions of partitioned rows and minimal column counts. This design ensures efficient data transfer by eliminating redundant payload data, while indexing partitions that are relevant to the associated research questions.

**Person table**

The person table is a short table, with one person per row. It contains static information about the person at the time the data was extracted, including date of birth, date of death, gender, ethnicity, diabetes type and date of diabetes diagnosis.
Drug era table

The drug era table is a long format table and holds information about all outpatient prescribed and dispensed drugs. The raw prescription information is converted into a person time series with one row per drug, with the initial prescription date being the start date and the end date calculated from drug duration which is determined using the number of repeated prescriptions, the quantity, and the dosing instructions. When the prescribed daily dose is not available, the WHO-defined daily dose (DDD) is used.

Observability (contributing electronic health record data)

The observability table provides a longitudinal view of when a patient was observable in the database based on their presence in the aforementioned drug era table or having a routine measurement taken (i.e., BMI, HbA1c, etc.). Each row has a start date and end date for each observability signal type and periods of unobservability. This allows researchers to exclude or censor people who have become lost to follow-up for some reason, e.g., they have emigrated.

Observations

The observation table holds cleaned and derived values from clinical measurements. This includes clinical measurements such as blood pressure, lower limb examination data (e.g., monofilament test or pedal pulse data), laboratory data (such as HbA1c, renal function), eye screening data (such as retinopathy grading), and lifestyle habits (alcohol, smoking, and exercise). This is held in a person time-series, with each row representing a different measurement, the date it was taken and its value in standard units.

Conditions

The condition table contains diagnosis and procedure data from SMR, NRS, and PHS in a person time series. The diagnosis data is transformed from raw tables containing one row per admission and up to nine diagnoses into a person time series with one row per diagnosis. For hospital inpatient records, start and end dates of conditions are based on the duration of hospital stay, whereas for death diagnoses and outpatient records, the start and end date is the date of death or clinic appointment.
**Metadata dictionary**

Meta-information on every variable available on the research platform is available to researchers through the data dictionary, which consists of a backend that builds the data dictionary database table during data extraction and processing, and an interactive searchable javascript front end. Each table in the research database has an associated metadata document, holding information about the provenance of each variable in the table, minimum/maximum expected values, lab specifications, any controlled vocabularies used, and any other information that may be of use to a researcher. This metadata also provides links to the group’s wiki, which goes into more depth on some of the more complicated variables and cleaning routines. When an Extract, Transform and Load (ETL) module is run to import a table into the research database, or to carry out data cleaning, this metadata is automatically extracted, along with information about the module that is being run. The modules also produce graphical summaries of the data that is being extracted or processed, such as availability over time, the general distribution of the data, and reliability. All this information is automatically collated and stored in a data dictionary database table. Researchers requiring access to the data dictionary do so through a bespoke searchable javascript front-end hosted on an internal website, which queries the data dictionary table to produce a graphical tree-based view describing all the information available in the research database, where the data is derived from, what processing it has undergone, and anything else that may be relevant to the researcher to ensure proper use of the data.

**Research Layer**

The research layer includes an internally developed R package which is designed to simplify user interaction with the back-end database system, providing a standard interface to define cohorts specific to any research question, and provide standard mappings of defined phenotypes to the relevant components of the longitudinal electronic health record.

**Defined Phenotypes**

Both published and internally defined phenotypes are integrated into the research platform. Examples of external algorithms are the Charlson Comorbidity Index, the Michigan Diabetic Neuropathy Score, and the Framingham CVD Risk Score. Internal phenotypic algorithms are developed with the combined knowledge of clinicians, epidemiologists, and data scientists. This internal phenotyping leverages all available information, while accounting for specific features and nuances in the underlying input data. Two examples of such algorithms are the diabetes type algorithm and the date of diabetes onset, with both datapoints being subject to variability in the longitudinal electronic records. These algorithms can be summarised as follows:

The date of diabetes onset is determined through an algorithm that finds the earliest date out of the following: 1) first date mentioned by consistent clinical assignment;
2) first date where there are two hospital admission records in a 3 year period with a diabetes-relevant ICD code, 3) first date of two consistently prescribed diabetes medications within one year, and 4) first HbA1c observation greater than 48 mmol/mol followed by another within 15 months.

The algorithm for diabetes type is validated using the longitudinal health record to evaluate longitudinal consistency of the clinician’s assignment of type along with any persistent prescribing for diabetes drugs, while accounting for data source completeness over time. We use the clinician-defined diabetes type unless there is evidence of misclassification. For type 1 diabetes, misclassification might be defined by 1) extensive use of oral anti-diabetic medication and 2) no continuous insulin therapy within one year of diagnosis. The application of the algorithm results in a reassignment of 10.5% of people assigned type 1 to type 2 and 0.8% of type 2 to a type 1.

Object Oriented Libraries

The research layer includes object oriented code libraries to aid in the conversion of relational data into a longitudinal research compatible format of adjustable person time intervals. One such library detailed in Supplementary Figure 2, implements a ‘survivor’ software class, which results in the creation of a survivor data object. The initial object of the class is instantiated with the definition of the cohort. This includes individuals that meet the study inclusion criteria along with their baseline or study entry variables and their study entry and expected study exit date (administrative censoring). These data are held in the format of an R data table (data.table) with one row per person containing multiple static data points. After the initial object definition, a class-based method is used to expand the single row into a series of N rows where each row represents a contiguous interval of time from that person’s entry to the study until their exit (e.g. 28 days). The form of the dataset is now multiple rows per person, with each row including baseline static values. Each row from 1:N-1 represents a full time interval with the exception of the last row N, where the person may have exited the study before the end of the interval and in which case it is truncated. Once each person has this time series, other methods are applied to merge time updated electronic health record data into each interval. Depending on the requirements of the study, parameters are passed to the method to specify any aggregation requirements. For example, if the time interval accounts for several months, where there may be several records of HbA1c results, the row for that interval may aggregate to the minimum, maximum and mean values within the interval. Where results are binary, for example in the case of being prescribed a drug or particular outcome, the initial starting date may be included along with current and ever/never binary terms. Once all variables have been merged into the intervals, the class includes a method to censor each person based on the criteria set out in the study. If a person’s time interval meets the censoring criteria, their time series is truncated, with the final interval being adjusted based on the new exit dates. The class method then allows recalculation of various variables to account for the censoring, for example by reclassifying someone as never having a myocardial
infection within the study period if a myocardial infarction occurred after that person was censored but before the administrative end of the study.

The format of the final dataset lends itself to almost any type of epidemiological analysis required of the diabetes dataset. Several objects may be created per study, enabling the calculation of results in different windows of time. This may include observation windows prior to the study, providing entry values based on lookback windows or defining prior disease states which may have occurred before the start of the study e.g., defining primary vs. secondary cardiovascular disease prevention.

With the data in this longitudinal format, not only are cross-sectional views available at any point in a person's trajectory, but also cumulative and time updated views. This flexibility provides the data in a format that will satisfy the demands of the simplest to the most complex analyses. For count based data used in Poisson survival analysis models, the flexibility in modifying the interval enables accurate predictive performance calculations to be achieved.
Scottish Diabetic Retinopathy Grading

The categories of retinopathy and maculopathy used in the cohort study have been harmonised with the nomenclature of Wilkinson et al[10]. The corresponding definitions of retinopathy and maculopathy used by the Scottish DRS[11] and the comparable ETDRS scale levels for each category are:

a) None - referring to no diabetic retinopathy anywhere having grades R0 and M0
b) Mild/Background NPDR - mild non-proliferative diabetic retinopathy corresponding to presence of - anywhere in a fundus image - at least one of: dot haemorrhages, microaneurysms, hard exudates, cotton wool spots, blot haemorrhages, superficial/flamed shaped haemorrhages. This is comparable to ETDRS scale level 20.

c) Moderate NPDR or Observable Maculopathy - moderate non-proliferative diabetic retinopathy corresponds to four or more blot haemorrhages in one hemisphere of the fundus images only comparable to ETDRS scale level 35, 43, and 47 and observable maculopathy is any hard exudates further than 1 disc diameter but within 2 disc diameter of the fovea.

d) Referable Maculopathy - this corresponds to any hard exudates or blot haemorrhages within a disc diameter of the fovea.

e) Severe NPDR - severe non-proliferative diabetic retinopathy corresponds to the presence of either: four or more blot haemorrhages in both hemi-spheres; venous beading; or intraretinal microvascular abnormalities and is comparable to ETDRS scale level 53.

f) Proliferative PDR - proliferative diabetic retinopathy corresponds to the presence of active new vessels or vitreous haemorrhage and is comparable to ETDRS scale levels 61, 65, 71, 75, 80, and 85.
References

[1] Van Rossum G, Drake Jr FL. Python tutorial. Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands; 1995.

[2] R Core Team. R: A language and environment for statistical computing. Online; Accessed 11th Nov 2020; https://www.R-project.org/.

[3] Cunningham S, McAlpine R, Leese G, Brennan G, Sullivan F, Connacher A, et al. Using web technology to support population-based diabetes care. Journal of Diabetes Science and Technology 2011;5:523–34. doi:10.1177/193229681100500307.

[4] Boyle DI, Cunningham SG. Resolving fundamental quality issues in linked datasets for clinical care. Health Informatics Journal 2002;8:73–7. doi:10.1177/146045820200800205.

[5] Overview Type 2 diabetes in adults: Management Guidance NICE n.d.

[6] Scotland Government. Diabetes care - diabetes improvement plan 2021 to 2026. Online; Accessed 21st Feb 2022; https://www.gov.scot/publications/diabetes-improvement-plan-diabetes-care-scotland-commitments-2021-2026/.

[7] Informatics OHDS and. Chapter 15 Data Quality The Book of OHDSI. n.d.

[8] Observational health data sciences and informatics. Observational medical outcomes partnership common data model. Online; Accessed 11th Nov 2021; https://www.ohdsi.org/data-standardization/the-common-data-model/.

[9] Clinical Data Interchange Standards Consortium. CDISC data exchange standards. Online; Accessed 11th Nov 2021; https://www.cdisc.org/standards.

[10] Wilkinson CP, Ferris FL, Klein RE, Lee PP, Agardh CD, Davis M, et al. Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. Ophthalmology 2003;110:1677–82. doi:10.1016/S0161-6420(03)00475-5.

[11] Scottish Diabetic Retinopathy Screening Service. Scottish diabetic retinopathy grading scheme. Online; Accessed 21st Feb 2022; https://www.ndrs.scot.nhs.uk/wp-content/uploads/2013/04/Grading-Scheme-2007-v1.1.pdf.
| Table | Item Name | Description |
|-------|-----------|-------------|
| person | serialno | Unique patient identifier used as a key for the patients data across tables |
| person | date_of_birth | Date of birth |
| person | date_of_death | Date of death |
| person | diag_support | How many supporting em_type earliest mention fields supported this data being set |
| person | dm_type | Cleaned diabetes type |
| person | earliest_mention | An algorithm based earliest mention of diabetes symptoms. |
| person | em_type | The earliest mention type is the information type that supported the adjustment of the earliest mention date |
| person | ethnic | Patient’s ethnicity based on various sources detailed in algorithm |
| person | gender | The patient’s gender based on the CHI database |
| person | hha | Health Board Authority of Residence |
| observation | body_mass-height | Height |
| observation | longitudinal-body_mass-height | Longitudinally-cleaned height |
| observation | body_mass-weight | Weight |
| observation | longitudinal-body_mass-weight | Longitudinally-cleaned weight |
| observation | body_mass-bmi | BMI |
| observation | longitudinal-body_mass-bmi | Longitudinally-cleaned BMI |
| observation | body_mass-activity | Activity status |
| observation | lifestyle-smoker | Tobacco consumption at date of contact including smoking |
| observation | lifestyle-cigs_day | Average number of cigarettes smoked per day |
| observation | lifestyle-alcohol | Alcohol intake per average week |
| observation | lifestyle-stopsmok | Date the patient stopped smoking |
| observation | lifestyle-alcohol_days | Number of days per average week alcohol consumed |
| observation | lifestyle-alcohol_status | Record of individual’s current alcohol consumption |
| observation | education-structured_program | Education structured_program - specific program data subject is enrolled in |
| observation | education-level | Structured education level |
| observation | education-status | Whether education has been offered |
| observation | blood_pressure-sbp | Systolic blood pressure |
| observation | blood_pressure-dbp | Diastolic blood pressure |
| observation | biochem-lipid-hdl | HDL cholesterol |
| observation | biochem-lipid-ldl | LDL cholesterol |
| observation | biochem-lipid-trig | Triglycerides |
| observation | biochem-lipid-tchol | Total cholesterol |
Supplementary Table 1: Key database variables (continued)

| Table | Item Name | Description |
|-------|-----------|-------------|
| observation | biochem-hba1c | Glycated Haemoglobin |
| observation | longitudinal-biochem-hba1c | Longitudinally-cleaned HbA1c |
| observation | biochem-creatinine | Serum creatinine |
| observation | biochem-creatinine_hospital | Serum creatinine - records taken when subject was in hospital |
| observation | biochem-creatinine_dialysis | Serum creatinine - any records taken at or after the start of RRT |
| observation | derived-rrt_longitudinal_startdate | Date of each data subject’s first record of the start of RRT, if any |
| observation | biochem-gfr | Estimated Glomerular Filtration Rate |
| observation | derived-eGFR | eGFR derived from creatinine level |
| observation | biochem-albumin-ratio | Urinary albumin/creatinine ratio |
| observation | biochem-albumin-concentration | Urinary albumin concentration |
| observation | biochem-albumin-night_rate | Timed overnight albumin excretion rate |
| observation | biochem-albumin-stage | Microalbuminuria stage |
| observation | biochem-albumin-level | Albumin level |
| observation | derived-albuminuric__longitudinal_grading | Longitudinally-cleaned albuminuric grading |
| observation | derived-albuminuric__longitudinal_clinical_grading | Longitudinal smoothed albuminuric grading |
| observation | biochem-crcl | Creatinine clearance rate |
| observation | biochem-ft4 | Free thyroxine level |
| observation | biochem-tt3 | Total triiodothyronine (TT3) |
| observation | biochem-free_tt3 | Free total triiodothyronine (TT3) |
| observation | biochem-hblu | Random blood glucose value |
| observation | biochem-bglu_fasting | Fasting blood glucose level |
| observation | biochem-bglu_fasting_bglu_diag | Fasting blood glucose level at diagnosis of diabetes |
| observation | biochem-bglu_2hr_bglu_diag | 2hr oral glucose tolerance test at diagnosis |
| observation | biochem-blood_cpep-fasting | Fasting C-Peptide |
| observation | biochem-blood_cpep-one_hr | C-Peptide one hour after oral glucose |
| observation | biochem-blood_cpep-two_hr | C-Peptide two hours after oral glucose |
| observation | biochem-blood_cpep-random | Random C-Peptide |
| observation | biochem-blood_cpep-post_meal | C-Peptide post meal |
| observation | biochem-urine_cpep-cpep_creat_ratio | Urinary C-Peptide/Creatinine ratio |
| observation | biochem-urine_prot-24g24 | Urinary protein 24hr |
| observation | biochem-urine_prot-pcr | Urinary protein/Creatinine ratio |
| observation | biochem-urine_prot-total_prot | Total urinary protein |
| observation | biochem-urine_prot-dipstick | Urinary protein dipstick |
| observation | biochem-other-creatinine | Creatinine level |
| observation | biochem-other-ast | Aspartate aminotransferase level |
| observation | biochem-other-alt | Alanine aminotransferase level |
| observation | biochem-other-ggt | Gamma glutamyltransferase level |
### Supplementary Table 1: Key database variables (continued)

| Item Name                                      | Description                                                                 |
|------------------------------------------------|-----------------------------------------------------------------------------|
| observation biochem-other-gada_interpr         | Interpretation of the subject’s anti-glutamic acid decarboxylase (GAD)       |
| observation biochem-other-ica_interpr          | antibody test result                                                        |
| observation biochem-other-wbc                  | White cell count                                                            |
| observation biochem-other-platelets            | Platelets                                                                   |
| observation biochem-other-alk_phos             | Alkaline phosphatase level                                                  |
| observation biochem-other-bilirubin            | Bilirubin level                                                             |
| observation biochem-other-sodium               | Sodium level                                                                |
| observation biochem-other-potassium            | Potassium level                                                             |
| observation lower_limb-ftrisk                  | Risk grading of the foot of a patient with diabetes mellitus                |
| observation lower_limb-pulses                  | Record of presence or absence of foot pulses.                              |
| observation lower_limb-ftmfil                  | Record of whether foot sensation to monofilaments is present or absent.    |
| observation lower_limb-ftvibr                  | Record of whether foot vibration sensation is normal or absent.             |
| observation lower_limb-ftsens                  | Record of whether foot sensation (monofilament and vibration) is normal or |
| observation lower_limb-ftdef                   | abnormal                                                                   |
| observation lower_limb-callus                  | Record of whether a foot callus is present.                                |
| observation lower_limb-amputation-amput        | Record of any lower limb amputation procedure(s) performed on the patient   |
| observation lower_limb-amputation-diabred       | Record of whether or not the corresponding amputation is diabetes related.  |
| observation lower_limb-amputation-amput_earliest| Record of the earliest lower limb procedure performed on the patient        |
| observation lower_limb-foot_ulecer-active      | Record of any active ulcers on the foot on the given date                   |
| observation lower_limb-foot_ulecer_previous    | Record of any previous ulcers on the foot before the given date             |
| observation lower_limb-painful_neuropathy      | Record of whether or not the patient has symptoms present that are due to   |
| observation lower_limb-neurothes_assess        | painful neuropathy                                                          |
| observation lower_limb-loss_protective_sens    | Neurothesiometer Assessment Result                                          |
| observation neuropathy-bowel_dysmobility       | Record of bowel dysmobility                                                |
| observation eye-vision-vu                      | Cleaned eye visual acuity of the patient recorded in the corrected state   |
| observation eye-vision-vu_corrected            | Whether the corresponding eye-vision-vu record for this patient was taken   |
|                                                 | with corrected eyesite i.e. wearing glasses                                 |

13
| Table Item Name | Description |
|-----------------|-------------|
| observation eye-globe-cataract | Presence of cataracts |
| observation eye-globe-retinal_exam-retina-retina | Cleaned record of whether or not the patient has diabetic retinopathy |
| observation eye-globe-retinal_exam-retina-nd_ret | Cleaned record of any retinal lesions identified and their nature |
| observation eye-globe-retinal_exam-retina-ret_scrn | Cleaned record of completion of an episode of retinal screening by an accredited method |
| observation eye-globe-retinal_exam-macula-macula | Cleaned record of whether or not the patient has diabetic maculopathy |
| observation eye-globe-retinal_exam-laser-laser_scrar | Cleaned record of whether or not laser photocoagulation scar(s) are visible at a grading examination |
| observation derived-retinopathy_longitudinal_grading | Longitudinally-cleaned retinopathy grading |
| observation eye-other-eye_meth | The method used to examine the patient's eye(s) |
| observation eye-drs-suspended | Reason for suspension from DRS |
| observation eye-drs-suspended_request_status | DRS Suspension Request Status |
| observation eye-drs-suspended_start | Start date of suspension from DRS |
| observation eye-drs-suspended_end | End date of suspension from DRS |
| observation eye-other-qa_indicator | Indicator provided by DRS to show if this particular result has been QA'd or not |
| observation eye-drs-screening_status | DRS screening status on given date |
| observation eye-other-laser_eye_therapy | Record of laser eye therapy |
| observation hpe-econo-covid_test_status | COVID-19 test status |
| device_era devices | Eras for common devices such as pumps |
| device_pump | Record of the insulin pump used by the patient |
| drug_era drugs | Eras for drug prescribing and dispensing |
| condition smr00 | SMR00 (hospital outpatient) |
| condition smr01 | SMR01 (hospital inpatient) |
| condition smr02 | SMR02 (maternity inpatient) |
| condition smr06 | SMR06 (Scottish Cancer Registry) |
| condition gho | GHO/NRS (death registry) |
## Supplementary Table 2: Diabetes type algorithm re-assignment by year

| Year of diagnosis | Age category | Type 1 (N) | Type 1 Recategorised (%) | Type 2 (N) | Type 2 Recategorised (%) | All Types (N) | All types Recategorised (%) |
|-------------------|--------------|-----------|--------------------------|-----------|--------------------------|---------------|----------------------------|
| 2010              | 0-15         | 389       | 5.14                     | 11        | 18.18                    | 407           | 6.39                       |
|                   | 16-35        | 366       | 11.20                    | 602       | 6.31                     | 1031          | 9.02                       |
|                   | 36+          | 308       | 37.66                    | 18563     | 1.71                     | 19438         | 2.42                       |
| 2011              | 0-15         | 378       | 2.38                     | 6         | 0.00                     | 389           | 3.08                       |
|                   | 16-35        | 381       | 11.84                    | 593       | 5.56                     | 1032          | 8.43                       |
|                   | 36+          | 313       | 28.75                    | 17733     | 1.50                     | 18669         | 2.14                       |
| 2012              | 0-15         | 420       | 2.86                     | 6         | 0.00                     | 437           | 2.97                       |
|                   | 16-35        | 371       | 14.02                    | 689       | 5.52                     | 1123          | 9.80                       |
|                   | 36+          | 312       | 32.05                    | 18733     | 1.62                     | 19668         | 2.27                       |
| 2013              | 0-15         | 328       | 2.13                     | 6         | 33.33                    | 347           | 3.17                       |
|                   | 16-35        | 329       | 12.46                    | 612       | 4.74                     | 996           | 8.13                       |
|                   | 36+          | 328       | 32.01                    | 18884     | 1.79                     | 19960         | 2.51                       |
| 2014              | 0-15         | 376       | 0.53                     | 11        | 18.18                    | 398           | 1.51                       |
|                   | 16-35        | 319       | 8.15                     | 672       | 4.46                     | 1062          | 6.87                       |
|                   | 36+          | 294       | 31.97                    | 17051     | 1.77                     | 18030         | 2.46                       |
| 2015              | 0-15         | 348       | 1.15                     | 8         | 25.00                    | 366           | 1.91                       |
|                   | 16-35        | 322       | 11.49                    | 662       | 4.23                     | 1055          | 7.77                       |
|                   | 36+          | 296       | 25.34                    | 17730     | 1.79                     | 18810         | 2.39                       |
| 2016              | 0-15         | 401       | 2.24                     | 8         | 0.00                     | 427           | 3.04                       |
|                   | 16-35        | 334       | 7.19                     | 704       | 6.39                     | 1112          | 6.83                       |
|                   | 36+          | 300       | 26.90                    | 17413     | 1.78                     | 18669         | 2.54                       |
| 2017              | 0-15         | 356       | 6.46                     | 703       | 4.69                     | 1124          | 5.96                       |
|                   | 16-35        | 297       | 32.56                    | 16732     | 1.94                     | 17926         | 2.59                       |
|                   | 36+          | 362       | 0.55                     | 9         | 33.33                    | 378           | 2.38                       |
| 2018              | 0-15         | 375       | 0.80                     | 7         | 14.29                    | 394           | 2.03                       |
|                   | 16-35        | 331       | 10.27                    | 571       | 3.68                     | 980           | 6.33                       |
|                   | 36+          | 248       | 17.34                    | 14284     | 1.91                     | 15304         | 2.37                       |
| 2019              | 0-15         | 357       | 1.96                     | 16        | 0.00                     | 385           | 1.82                       |
|                   | 16-35        | 336       | 9.82                     | 604       | 2.48                     | 1026          | 5.36                       |
|                   | 36+          | 301       | 22.59                    | 15222     | 1.83                     | 16308         | 2.54                       |

*Number and percentage of people reassigned by the diabetes type algorithm provided by year and broad age band. Note that ‘all types’ includes other types of diabetes. The percentage recategorised is from the original diabetes type to another diabetes type.*
Supplementary Figure 1: Scottish Diabetes Research Network data model
Supplementary Figure 2: Longitudinal cohort generation using survivor class