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 datases from-e 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In the United Kingdom, several clinical coding systems have been introduced and subsequently superseded from as early as the 1950s to the present day.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

McGurnaghan SJ, *et al. BMJ Open* 2022; 12:e063046. doi: 10.1136/bmjopen-2022-063046

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

infarction 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 hemi-sphere 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 bot 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. Scottishi 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.

Supplementary Table 1: Key database variables

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	hba	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 consumpation
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-tsh	Thyroid stimulating hormone
observation	biochem-tt3	Total triiodothyronine (TT3)
observation	biochem-free_tt3	Free total triiodothyronine (TT3)
observation	biochem-bglu	Random blood glucose value
observation	biochem-bglu_fasting	Fasting blood glucose level
observation	biochem-fasting_bglu_diag	Fasting blood glucose level at diagnosis of diabetes
observation	biochem-bglu_2hr_gt_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-haemoglobin	Haemoglobin 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)*

Table	Item Name	Description
observation	biochem-other-gada_interpr	Interpretation of the subject's anti-glutamic acid acid decarboxylase (GAD) antibody test result
observation	biochem-other-ica_interpr	Interpretation of the subject's anti-islet cell 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 absense 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 abnormal.
observation	lower_limb-ftdef	Record of whether there is any foot deformity present.
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-diabrel	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_ulcer-active	Record of any active ulcers on the foot on the given date
observation	lower_limb-foot_ulcer-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 painful neuropathy.
observation	lower_limb-neurothes_assess	Neurothesiometer Assessment Result
observation	lower_limb-loss_protective_sens	Records whether or not the patient has loss of protective sensation in the foot
observation	neuropathy-bowel_dysmobility	Record of bowel dysmobility
observation	eye-vision-va	Cleaned eye visual acuity of the patient recorded in the corrected state
observation	eye-vision-va_corrected	Whether the corresponding eye-vision-va record for this patient was taken with corrected eyesite i.e. wearing glasses

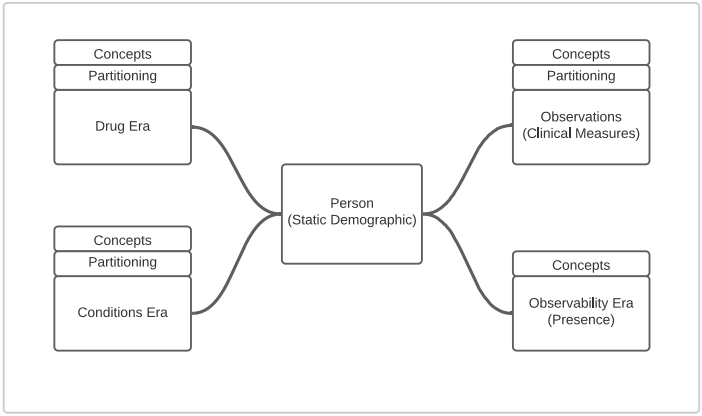
Supplementary Table 1: Key database variables *(continued)*

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__scar	Cleaned records 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 patients 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	hps-ecoss-covid__test__status	COVID-19 test status
device__era	devices	Eras for common devices such as pumps
observation	device-insulin__pump	Record of the insulin pump used by the patient
device	pump	Insulin pump era data
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	gro	GRO/NRS (death registry)

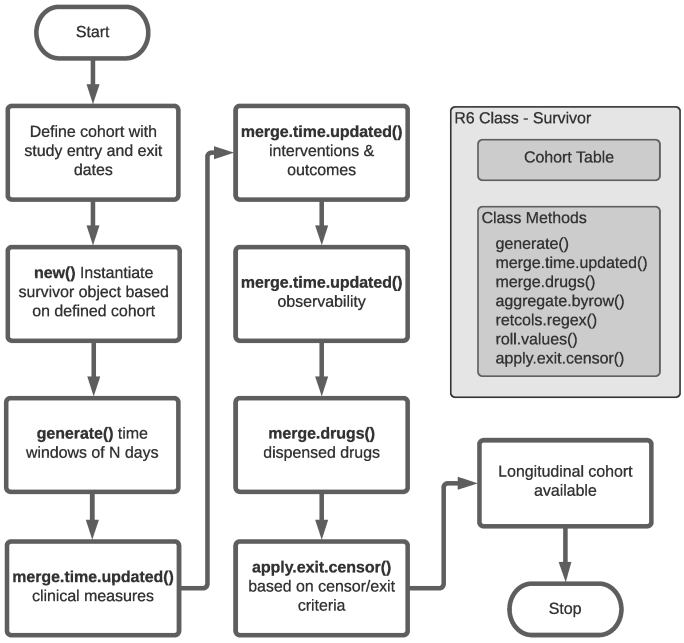
Supplementary Table 2: Diabetes type algorithm re-assignment by year

Year of diagnosis	Age category at diagnosis	Type 1 (N)	Type 1 Re-categorised (%)	Type 2 (N)	Type 2 Re-categorised (%)	All Types (N)	All types Re-categorised (%)
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	11.81	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	26.00	17413	1.78	18600	2.54
2017	0-15	362	0.55	9	33.33	378	2.38
	16-35	356	6.46	703	4.69	1124	5.96
	36+	297	22.56	16732	1.94	17926	2.59
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	14264	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	2.46

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