UK Biobank

Primary Care Linked Data

Version 2.0

https://www.ukbiobank.ac.uk
April 2024

This companion document provides supporting information for the interim release of primary care linked data in UK Biobank. Version 2.0 includes reference to Resource 951, which contains reference data for numeric values accompanying TPP clinical codes.

Contents

1. Introduction

1.1. Primary care in the UK

Within the UK healthcare setting, individuals seeking advice or treatment for a health concern normally first meet with a family physician (known as a General Practitioner, or GP) or a nurse (for example, a Nurse Practitioner) at their local general practice. GPs can refer patients who require more specialised treatment (or further tests) to hospital or other community-based services. There is a wealth of information available within primary care records; most secondary care interactions are reported back to general practice and some illnesses tend to be managed entirely within a primary care setting.

The term 'primary care' is sometimes used more broadly to include other healthcare professionals such as pharmacists, dentists and opticians. The UK Biobank primary care data relates only to data recorded by health care professionals working at general practices.

1.2. Linkage to primary care data in UK Biobank

There is currently no national system for collecting or sharing primary care data. UK Biobank has been liaising with various data suppliers and other intermediaries (including the main primary care computer system suppliers in England) to obtain primary care data for UK Biobank participants, all of whom have provided written consent for linkage to their health-related records. To date, coded data have been obtained for approximately 45% of the UK Biobank cohort (~230,000 participants) and are now available as part of this interim release. Details of the data providers and the coding schema used are summarised in Table 1.

UK Biobank is currently in the process of securing access to data for the remaining cohort, mainly for participants registered with EMIS practices across England.

Table 1. System suppliers, participant numbers & coding classifications used

Country	GP Computer System Supplier	Approx no. of UK Biobank participants	Clinical coding classification	Prescription coding classification
Scotland [a]	EMIS [b] / Vision [c]	27,000	• Read v2	• Read v2 • British National Formulary (BNF)
Wales [d]	EMIS / Vision	21,000	• Read v2	• Read v2
England	TPP[e]	165,000	• Clinical Terms Version 3 (CTV3 or Read v3)	• BNF
	Vision	18,000	• Read v2	• Read v2 • Dictionary of Medicines and Devices (dm+d)

a. UK Biobank has engaged Albasoft (http://www.albasoft.co.uk/) (a third party data processor) to obtain data from GP practices in Scotland.

b. EMIS Health (https://www.emishealth.com/) is a computer system supplier to the NHS and provides the EMIS Web practice management system.

c. Vision Health (https://www.visionhealth.co.uk/) (previously InPS) is a computer system supplier and provides the Vision practice management system.

d. Data from Wales have been obtained via the SAIL Databank (https://saildatabank.com/) hosted by the University of Swansea.

e. TPP (https://www.tpp-uk.com/) is a computer system supplier and provides the SystmOne practice management system.

2. Format of the primary care data

The availability, completeness and level of detail in the data varies between systems and suppliers and we have purposefully limited the amount of data cleaning/curation for this interim release (See Appendix B for the limited validation checks performed on the data). The dataset contains variables that are considered the most important for epidemiological research: coded clinical events (including diagnoses, history, symptoms, lab results, procedures), prescriptions (i.e. medications that are prescribed but not necessarily dispensed) and a range of administrative codes (e.g. referrals to specialist hospital clinics). Non-coded, unstructured data (e.g. free-text entries, referral letters) are not included, with minor exceptions as described below.

As noted in UK Biobank is currently in the process of securing access to data for the remaining cohort, mainly for participants registered with EMIS practices across England.

Table *1*, the primary care computer system suppliers have adopted different coding classifications as part of their underlying data schema. In addition to these coding variations for clinical events, the different system suppliers use a range of coding classifications for prescriptions. For ease of use by UK Biobank researchers, data from the different sources are organised into three tables with harmonised variable names and data types with a field indicating the source.

- **Registration records**
    - ID, registration date and date of removal from practice lists. Multiple registration records are available per person from most (but not all) suppliers (see section 4.1 for further details).
- **Clinical events**
    - Date and clinical code (Read v2 or CTV3 [1]) for primary care events, such as consultations, diagnoses, history, symptoms, procedures, laboratory tests and administrative information. Where available, value fields are provided which may give further details. These particular fields have been modified to remove potentially identifiable information. See section 4.5 for more information.
- **Prescriptions**
    - Date, drug code (Read v2, BNF [2] and/or dm+d [3]) and, where available, drug name and quantity for medicines or devices prescribed in primary care. Drug name and quantity will assist with interpretation of drug code fields that have different levels of completeness. See sections 3.2 and 3.3 for more information.

3. Clinical coding classification systems

As part of the limited data curation that has been applied, multiple clinical coding data are provided per record where these are available. Given that some records include more than one clinical code these may be contradictory. Further, the nature of these data (including some local variation in code use) may mean that some codes do not match to official code lists. Researchers are strongly advised to test and interpret their findings appropriately. Each of the coding classification systems is described below with links to resources for more information.

3.1. Read v2 and CTV3

Read codes are a coded thesaurus of clinical terms used in primary care since 1985. There are two versions: version 2 (Read v2) and version 3 (CTV3 or Read v3). Both provide a standard vocabulary for clinicians to record patient findings and procedures. Read v2 and CTV3, together with a UK Read code browser, are available via the NHS Digital Technology Reference Data Update Distribution (TRUD) website[4]. Read v2 and CTV3 were last updated in April 2016 and April 2018, respectively. Both versions

[1] Read codes were updated biannually and distributed under Open Government License via the UKTC Terminology Reference data Update Distribution (TRUD) service - https://isd.digital.nhs.uk/

[2] British National Formulary (BNF) provides prescribing and pharmacology guidance on medicines used within the NHS https://www.bnf.org/

[3] dm+d provides a dictionary of descriptions and codes for medicines and devices used across the NHS.

[4] https://isd.digital.nhs.uk/trud3/user/guest/group/0/pack/9

are now deprecated (as is the Read Browser) and no further updates will occur. From April 2018, SNOMED CT [5] was introduced into primary care in a phased approach and it is intended by April 2020 that SNOMED CT will be fully incorporated across the wider NHS, including codes related to prescriptions.

Table *2* shows the number of records in the clinical and prescription tables in this interim release, including those that have missing or Read codes which do not match downloaded code lists (e.g. that may have been defined for local purposes). Although a large proportion of the Scottish prescribing records (74%) do not include a Read code, many of these records have a BNF code.

Table 2. Approximate number of records by supplier and coding schema, with amount of missing and unmatched codes

Country	GP Computer System Supplier	Table	Coding system	Total no of records	Unmatched codes	No code provided
Scotland	EMIS / Vision	clinical events	Read v2	11.4M	18k (0.2%)	0
		prescription	Read v2	4.3M	<1k	3.2M (74%)
Wales	EMIS / Vision	clinical events	Read v2	12.8M	0	0
		prescription	Read v2	7.5M	0	0
England	Vision	clinical events	Read v2	12M	128k (1%)	0
		prescription	Read v2	6.3M	<1k	124k (2%)
	TPP	clinical events	Read CTV3	87.5M	2.5M	0

Some Read v2 and CTV3 codes have the potential to be disclosive (e.g. occupational codes that may be attributable to an identifiable person). To investigate this further, codes were identified that appeared only once in each supplier's data extract (range: n = 214 (Scotland) to n=528 (TPP)) and code descriptions were manually reviewed for disclosive information, none of which were specific enough to identify an individual. We therefore assessed the risk of releasing identifiable information as low.

3.2. BNF (British National Formulary)

The BNF is the standard list of medicines, dressings and appliances prescribed in the UK. It is published as a reference guide in both online [6] and paper versions and contains information on, for example, dose, side effects and price for over 70,000 items. Code lists are updated annually and can be downloaded from the NHS Business Services Authority (NHSBSA). [7]

The BNF, in its standard form, does not cover all items prescribed by the NHS, and many items are listed in appendices as opposed to formal chapters. To address this, the NHSBSA has developed pseudo-BNF

[5] The NHS flavour of SNOMED CT used across the UK is managed by NHS Digital https://digital.nhs.uk/snomed-ct

[6] https://www.bnf.org/

[7] https://apps.nhsbsa.nhs.uk/infosystems/welcome

codes and chapters (18-23) covering dressings, other drugs, preparations and appliances. Other online resources [8] [9] [10] give further information including how the NHSBSA assigned codes to BNF items.

3.2.1. Interpreting BNF codes

The format and detail of BNF codes varied by supplier and researchers are advised to that care must be taken to correctly interpret them, particularly when data from multiple sources are combined.

The full BNF presentation code, as provided by the NHSBSA for chapters 1-15 and 18-19, is fifteen characters in length (note that there are no chapters numbered 16 or 17). Table 3 shows an example for Yaltormin SR 500mg tablets, an antidiabetic drug with the chemical substance metformin hydrochloride.

Table 3. Example of a fifteen-digit BNF code and its interpretation in BNF chapters 1-15 & 18-19

Detail level in BNF	Relevant character(s) in BNF code	Example code: Yaltormin SR 500mg Tablets	Description
Chapter	1 & 2	**06**01022B0BPAAAS	Chapter 6, endocrine system
Section	3 & 4	06**01**022B0BPAAAS	Section 1, drugs used in diabetes
Paragraph	5 & 6	0601**02**2B0BPAAAS	Paragraph 2, antidiabetic drugs
Subparagraph	7	060102**2**B0BPAAAS	Subparagraph 2, biguanides
Chemical substance	8 & 9	0601022**B0**BPAAAS	Metformin hydrochloride, all other biguanides with this chemical substance are coded as B0
Product name	10 & 11	0601022B0**BP**AAAS	Yaltormin SR, all other Yaltormin SR products are coded as BP
Further product information (e.g. capsule, tablet, liquid, strength)	12 & 13	0601022B0BP**AA**AS	Yaltormin SR 500mg = AA, note: 750mg = AB & 1000mg = AC. Letters denoting strength do not always refer to the same dosage in other drugs, hence AA does not always refer to 500mg tablets for other medicines
Equivalent products	14 & 15	0601022B0BPAA**AS**	All biguanides (tablets) with the chemical substance metformin hydrochloride, with a dosage of 500mg, e.g. Meijumet 500mg tablets, are coded as AS

Chapters 20-23 follow a similar coding format and are eleven characters in length. They relate to dressings and appliances, hence no information on chemical substance and dose is necessary.

BNF codes are provided in this data interim release for TPP and Scotland; however, the formatting of these codes differs by source. Information on the structure and detail in these two sources is described below.

3.2.2. BNF coding in Scottish prescription data

[8] https://www.nhsbsa.nhs.uk/sites/default/files/2017-04/BNF_Classification_Booklet-2017_0.pdf

[9] https://ebmdatalab.net/prescribing-data-bnf-codes/

[10] https://digital.nhs.uk/data-and-information/areas-of-interest/prescribing/practice-level-prescribing-in-england-a-summary/practice-level-prescribing-glossary-of-terms

The method of BNF coding used in the Scottish GP data does not always follow the standard formatting provided by the NHSBSA. Codes range from one to fifteen characters in length, with an associated drug name or description to help with interpretation.

Table ***4*** shows examples of typical codes found in the Scottish GP data and their level of detail.

Table 4. Examples of BNF codes in the Scottish GP data

Character length	Example code	Approx. occurrences	Level of detail covered by code
Null	-	49k	-
1	3	<10	Chapter
2	23	<100	Some codes relate to chapter, others follow nonstandard format
4	0411	134k	Chapter, section
6	020201	638k	Chapter, section, paragraph
8	04080100	1,299k	Chapter, section, paragraph is accurate in many cases. Note that codes of this length do not always map to those provided by the NHSBSA
11	22600506000	9k	11 character codes relate to chapters 20 – 23 only, covering dressings and appliances
15	0704020N0AAABAB	2,168k	Full presentation code, all levels of detail
Total	-	**4,297k**	-

3.2.3. BNF codes in TPP data

BNF codes in the TPP extract follow the format 00.00.00.00.00. However, the coding structure does not always map to codes provided by the NHSBSA. The first six digits of the code typically relate to BNF chapter, section and paragraph in the NHSBSA code lists, although this is not consistent. Digits 7 and 8 do not appear to correspond to subparagraphs in the NHSBSA codes, and digits 9 and 10 are always coded as 00. To support analysis, the associated drug name or description for each BNF code is included.

3.3. dm+d (Dictionary of Medicines and Devices)

The prescription data from Vision (England) contains dm+d codes (as well as Read v2 codes) to record medicines prescribed to patients. The dm+d dictionary[11] has been developed for use throughout the NHS (primary and secondary care) to identify specific medicines and devices used in the treatment of patients and consists of a dictionary containing unique identifiers and associated text descriptions.

The dm+d model consists of five components:

- a Virtual Therapeutic Moiety (VTM) - the substances intended for use in the treatment of a patient
- Virtual Medicinal Product (VMP) - the properties of one more AMPs

[11] https://digital.nhs.uk/data-and-information/information-standards/information-standards-and-data-collections-including-extractions/publications-and-notifications/standards-and-collections/scci0052-dictionary-of-medicines-and-devices-dm-d

- Actual Medicinal Product (AMP) - a single dose unit of an actual product known to have been available from a specific supplier
- Virtual Medicinal Product Pack (VMPP) - the properties of one or more equivalent AMPPs
- Actual Medicinal Product Pack (AMPP) - the packaged product supplied for direct patient use.

An example of the dm+d component structure for a packet containing 56 tablets of Yaltormin 500mg is shown in Table 5. Note the generic name appears in the VTM, VMP and VMPP dm+d components while the brand name is used in the AMP and AMPP components.

Table 5. Example dm+d codes, components and descriptions

dm+d code	dm+d component	Description
109081006	VTM	Metformin
386047000	VMP	Metformin 500mg modified-release tablets
35547511000001101	AMP	Yaltormin SR 500mg tablets (Wockhardt UK Ltd)
8990611000001109	VMPP	Metformin 500mg modified-release tablets 56 tablets
35547911000001108	AMPP	Yaltormin SR 500mg tablets (Wockhardt UK Ltd) 56 tablet

3.4. Clinical code look-ups and mapping files

In order to facilitate research on these data, clinical code lists have been compiled from TRUD and NHSBSA (Appendix C) [see Resource 592]. TRUD has historically provided information on how to map from Read v2 and CTV3 to other clinical coding systems. However, this information is now being archived as Read versions are deprecated and SNOMED CT is adopted. The accuracy of code lists, definitions and maps should be verified by specialists as part of any analysis undertaken on these data.

Selected code lists developed by UK Biobank which identify specific health outcomes are also available [see Resource 594]. These include a limited set of validated algorithmically-defined health outcomes some of which are already published [12] and a broader set of outcomes mapped to 3-character ICD-10 [see Resource 593].

Other sources of health outcome code lists which may be useful include CALIBER [13] and the Clinical Codes repository. [14] As noted above, any code list used to analyse these data should be considered and verified by appropriate specialists.

[12] http://biobank.ndph.ox.ac.uk/showcase/label.cgi?id=42
[13] https://www.caliberresearch.org/portal/documentation/introduction
[14] https://clinicalcodes.rss.mhs.man.ac.uk/

4. Data quality

‘Real world’ administrative, routinely collected data (such as this interim GP data release) have enormous potential to support research with far reaching benefits to human health. By their nature, analysing and interpreting these data within the context of health research requires careful consideration of their content, structure and crucially, an in-depth understanding that they were collected for an entirely different purpose: recording the delivery of patient care in thousands of different centres across the UK countries operating within their own NHS systems.

Although this interim GP data release has involved some alterations to the raw data to facilitate its research use (e.g. selecting variables, appending tables from different sources requiring aligned variable names), minimal data cleaning has been undertaken (see Appendix B) in order to retain as much useful information that is as close to its original form as possible. While this approach avoids inserting unintentional bias into the data, it leaves significant risk of data quality issues that must be taken into account in all analyses. Several key data quality issues are described in detail below.

4.1.Registration records

Information on participant registrations varies by data supplier, in that Vision (England) provided a single registration record per person while the other suppliers provided multiple records per participant, and a small number of participants with data in the TPP extract do not have a registration record. Therefore variable numbers of registration records are included in this release, reflecting the providers’ extracts. The start date of coverage is not known for all participants, nor is the completeness of coverage of their primary care health records until the extract date (see section 4.2 for more information). Appropriate analytical techniques must therefore be adopted to deal with the impact of unknown timelines and potential absences of coverage which these data likely encapsulate.

In hospital episode statistics, and its equivalents in Scotland and Wales, the format and timing of submission of data on patient care are standardized. The practice of electronic coding in primary care has increased over time and may be influenced by local procedures, requirements around reporting (e.g. Quality Outcomes Framework [15] (QOF)) and other factors. Both primary and secondary care systems are subject to a range of potential biases and fluctuations over time due to national and local policy initiatives and local processes and procedures. Their completeness and accuracy (relative to the actual health experiences of the individuals represented in the coded data) cannot be assumed and is expected to differ between systems and over time.

Figure 1 (clinical) and Figure 2 (prescriptions) show the trend in availability of data from each source by year. To facilitate comparison, each bar indicates the number of records per year as a proportion of the total number of records from that source.

The completeness of transfer of data when a patient moves between practices is unknown.

[15] https://digital.nhs.uk/data-and-information/data-collections-and-data-sets/data-collections/quality-and-outcomes-framework-qof

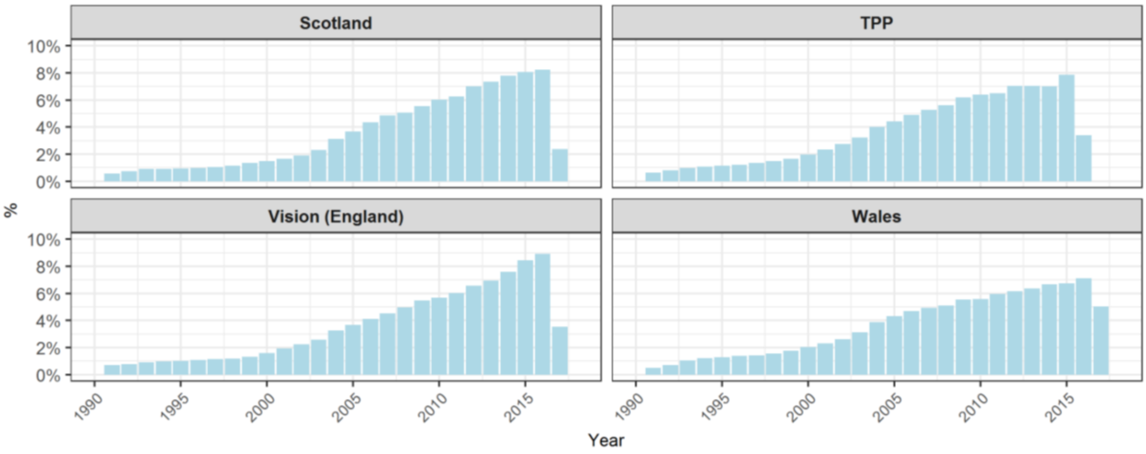


Figure 1. Proportion of participants with clinical data by year and source.[16]

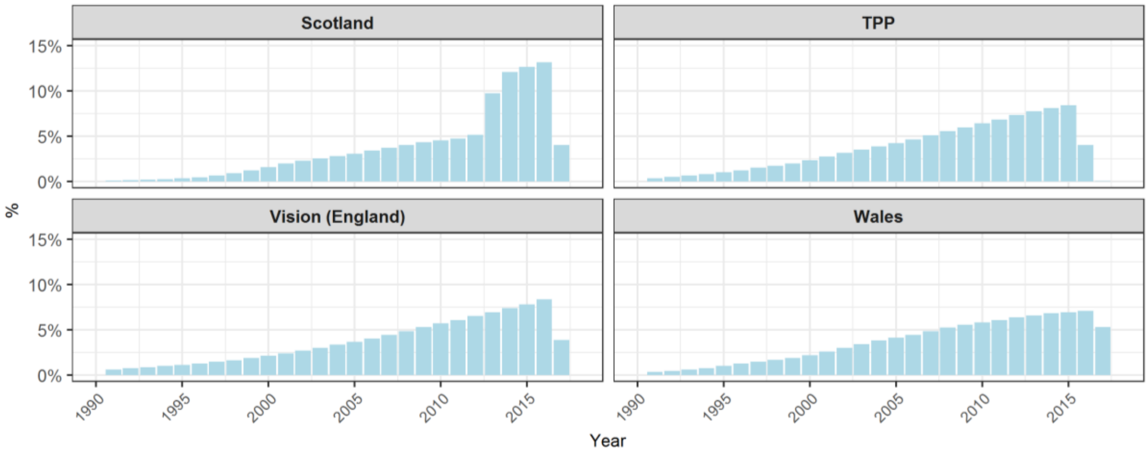


Figure 2. Proportion of participants with prescription data by year and source.

4.2.Extract date

Data were extracted from supplier's computer systems using different approaches, in each case making a single extract date or cut-off point impossible to determine. Table 6 provides information on the month of extract and the dates during which an observable reduction in the number of records was noted compared to the same dates in the previous year, for each supplier.

[16] In Figures 1 & 2, the graph x-axis is cut-off at 1990 for clearer presentation however a small number of earlier records are available.

Table 6. End of coverage (extract date) and period of reduced data observation by supplier

Country	GP Computer System Supplier	Extract date	Records below expected (relative to previous year)
Scotland	EMIS / Vision	May 2017	19th Apr – 4th May 2017
Wales	EMIS / Vision	Sep 2017	18th – 27th Sep 2017
England	Vision	Jun/Jul 2017	25th May – 14th Jun 2017
	TPP	Aug 2016	14th – 16th Jun 2016

The reduction in number of records by supplier is shown in Figure 3. Care should be taken to incorporate varying durations of coverage and follow-up in analyses, both within and between sources as some participants' data may have been extracted days or weeks apart.

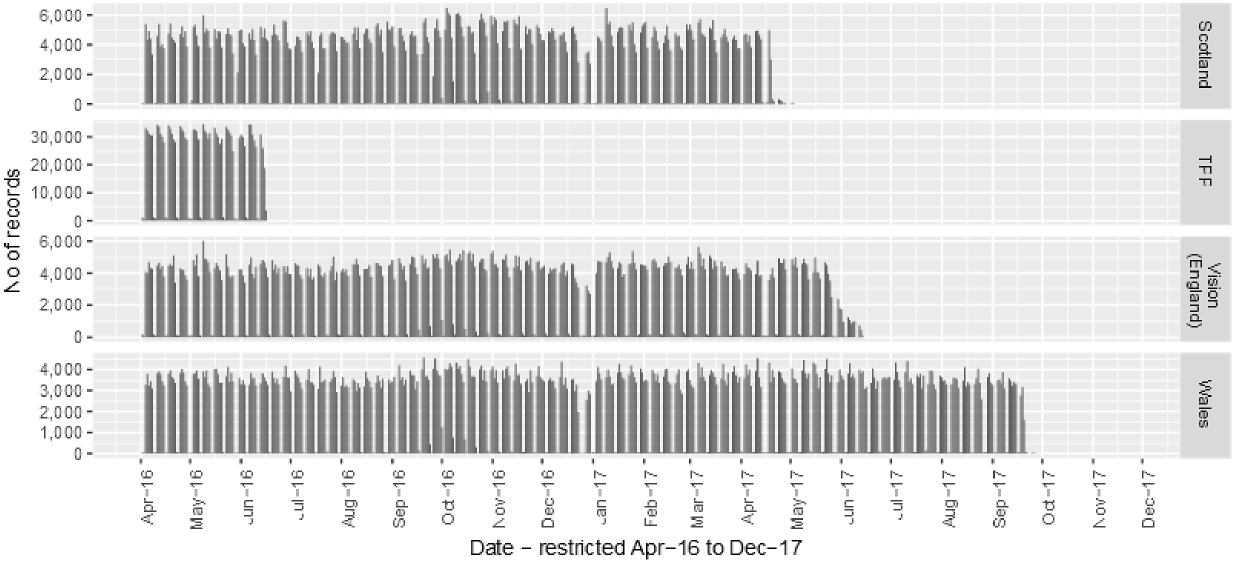


Figure 3. Total number of daily records from each data supplier

4.3. Deceased participants

Approximately 3% of participants in data extracts from TPP and Wales are known (from UK Biobank's linkages to death registry data) to have died prior to the data extraction, as shown in Table 7. A similar percentage of deceased participants is present in Scottish registration and clinical records, but only around half of that number are present in the Scottish prescription data, which may be due to a system-wide block of missing records prior to 2012 (Figure 2 top left panel). There are a very small number of known deceased participants in the English (Vision) data, suggesting that the clinical and prescription data from around 500 people (who would have died before the extraction occurred) are missing from the extract. This is a significant data quality issue which will lead to biased analyses if not properly taken into account. In particular, researchers looking into conditions where mortality is a significant outcome may wish to exclude the sub-group of individuals with linked data from this supplier, but all researchers should consider the impact of these missing individuals.

Table 7. Approximate number and proportion of deceased UK Biobank participants in each table.

Source		Table	Approx. number of deceased	Approx. number in extract	Percentage
Wales		registration	750	21k	3.6%
		clinical events	750	21k	3.6%
		prescription	750	20k	3.6%
Scotland		registration	700	27k	2.7%
		clinical events	700	27k	2.7%
		prescription	350	25k	1.4%
England	TPP	registration	5,300	167k	3.2%
		clinical events	5,300	165k	3.2%
		prescription	5,200	160k	3.2%
	Vision	registration	<10	19k	<0.03%
		clinical events	<10	18k	<0.03%
		prescription	<10	18k	<0.03%

4.4. Immunisation / vaccination records

Data on immunisations are available from England (Vision), Scotland and Wales and are included in the clinical events table. However immunisation records from TPP did not include a clinical code, so they are not included.

4.5. Free-text fields

In the clinical events table, there are three 'value' fields which may provide further detail on certain types of event e.g. a blood pressure measurement or lab test result. Suppliers' extracts contain only numeric data in one (TPP)[17] or two (England (Vision), Wales) 'value' fields. In Scotland, however, all three 'value' fields incorporate free text. These are mainly local system codes; the indicators "Y" or "N"; or units relating to a numerical value. In order to retain as much potentially useful information as possible, manual checks on the contents were carried out on the 1,600 values which contained some text and 80 potentially disclosive values were removed, mostly containing the name or initials of a doctor, patient name, phone number or location.

A similar exercise was undertaken on a small number of (n=151) free-text values in the quantity field for prescriptions from TPP.

[17] To accompany a later primary care data release (for COVID-19 research) TPP provided to UK Biobank unit of measurement information as a separate "numeric reference" table, which is available as Resource 951. Within this table, TPP also provided a "numeric precision" value for relevant clinical codes; this represents the appropriate number of digits in the value to be used after the decimal. Researchers might consult normative data on the measure of interest, to verify the accuracy of information in this resource, rather than rely entirely on the assumption that all details within have remained unchanged between the TPP releases of 2016 and 2020 respectively.

4.6. Dates

These data are provided in a form which is as close as possible to how they were issued from their source supplier, in order to avoid potential systematic error or bias by attempting to 'clean' them by removing or altering invalid or erroneous information. However, to protect individuals, alterations have been made to dates in relation to participant date of birth as follows:

- where clinical event or prescription date precedes participant date of birth it has been altered to 01/01/1901.
- Where the date matches participant date of birth it has been altered to 02/02/1902.
- Where the date follows participant date of birth but is in the year of their birth it has been altered to 03/03/1903.
- Where the date was in the future this has been changed to 07/07/2037 as these are likely to have been entered as a place-holder or other system default.

Researchers are advised to take steps to ensure that information related to clinical event or prescription date is analysed appropriately.

These quality issues, and other artefacts in the data, have implications for the conclusions that can be drawn; care should be taken not to make generalisations based on the assumption that this is a complete and error-free dataset. For example, the absence of a diagnostic code for any individual cannot be interpreted with 100% certainty to mean that they did not have that condition. As previously noted, this interim release includes data on approximately 231,000 UKB participants, or just under half of the UKB cohort. It should not be assumed that analyses conducted on these participants' data can be generalised to the whole cohort, or the UK population as a whole. Appropriate analytical techniques must be employed to account for missing or unreliable data. This includes identifying:

- duplicate information (e.g. when a participant moves between general practices whose data is captured by different data suppliers)
- erroneous information (e.g. dates or codes entered incorrectly)
- inconsistent information (e.g. variation in timing or content of records between sources)
- data gaps (e.g. absent or suppressed information, or variation in completeness of available data).

Researchers who use this interim release are invited to feed back to UK Biobank on their experiences with these data. This, and other exploration done within UK Biobank, will be used to develop additional guidance for future releases. A list of frequently asked questions will be compiled and updated regularly based on feedback – see the Researcher section of our website.

5. Data available in UK Biobank

The interim release data schema is shown below indicating the availability of variables by source, and the unified UK Biobank variable names.

Table 8. Registrations

Name in UKB	Description	Name – England (Vision)	Name – England (TPP)	Name - Scotland	Name - Wales
eid	Participant identifier	pid	pid	pid	pid
reg_date	Registration date	REGDATE	RegistrationDate	RegDate	FROM_DT
deduct_date	Deduction date	DEDUCTIONDATE	DeductionDate	DeductionDate	TO_DT

Table 9. Clinical events

Name in UKB	Description	Name - England (Vision)	Name - England (TPP)	Name - Scotland	Name - Wales
eid	Participant identifier	pid	pid	pid	pid
event_dt	Date clinical code was entered	ODATE	EventDate	StartDate	EVENT_DT
read_2	Read v2	READCODE	n/a	ReadCode	EVENT_CD
read_3	CTV3 (Read v3)	n/a	CTV3ConceptId	n/a	n/a
value1	Value recorded 1	NUMRESULT	NumericValueRecorded	Data1	VALUE1
value2	Value recorded 2	NUMRESULT2	n/a	Data2	VALUE2
value3	Value recorded 3	n/a	n/a	Data3	n/a

Table 10. Prescriptions

Name in UKB	Description	Name - England (Vision)	Name - England (TPP)	Name - Scotland	Name - Wales
eid	Participant identifier	pid	pid	pid	pid
issue_date	Date prescription was issued	ISSUEDATE	MedicationStartDate	IssueDate	EVENT_DT
read_2	Read v2	READCODE	n/a	ReadCode	EVENT_CD
bnf_code	BNF code	n/a	BNFChapterId	BNF	n/a
dmd_code	DM+D code	DMDCODE	n/a	n/a	n/a
drug_name	Drug name	n/a	DrugName	Drugname	n/a
quantity	Quantity issued	SUPPLY	DrugQuantity	qty	n/a

6. Data organisation

Data received from the four GP data providers have been combined into three tables as below.

Table 11. Participant registration records

Registrations table: gp_registrations	
Column name	**Description**
eid	Participant identifier
data_provider	1= England(Vision), 2= Scotland, 3 = England (TPP), 4 = Wales
reg_date	Registration date
deduct_date	Deduction date – date of individual's removal from GP list

Table 12. Clinical event records

Clinical (events) table: gp_clinical	
Column name	**Description**
eid	Participant identifier
data_provider	1= England(Vision), 2= Scotland, 3 = England (TPP), 4 = Wales
event_dt	Date clinical code was entered
read_2	Read v2
read_3	CTV3 (Read v3)
value1	Value recorded 1
value2	Value recorded 2
value3	Value recorded 3

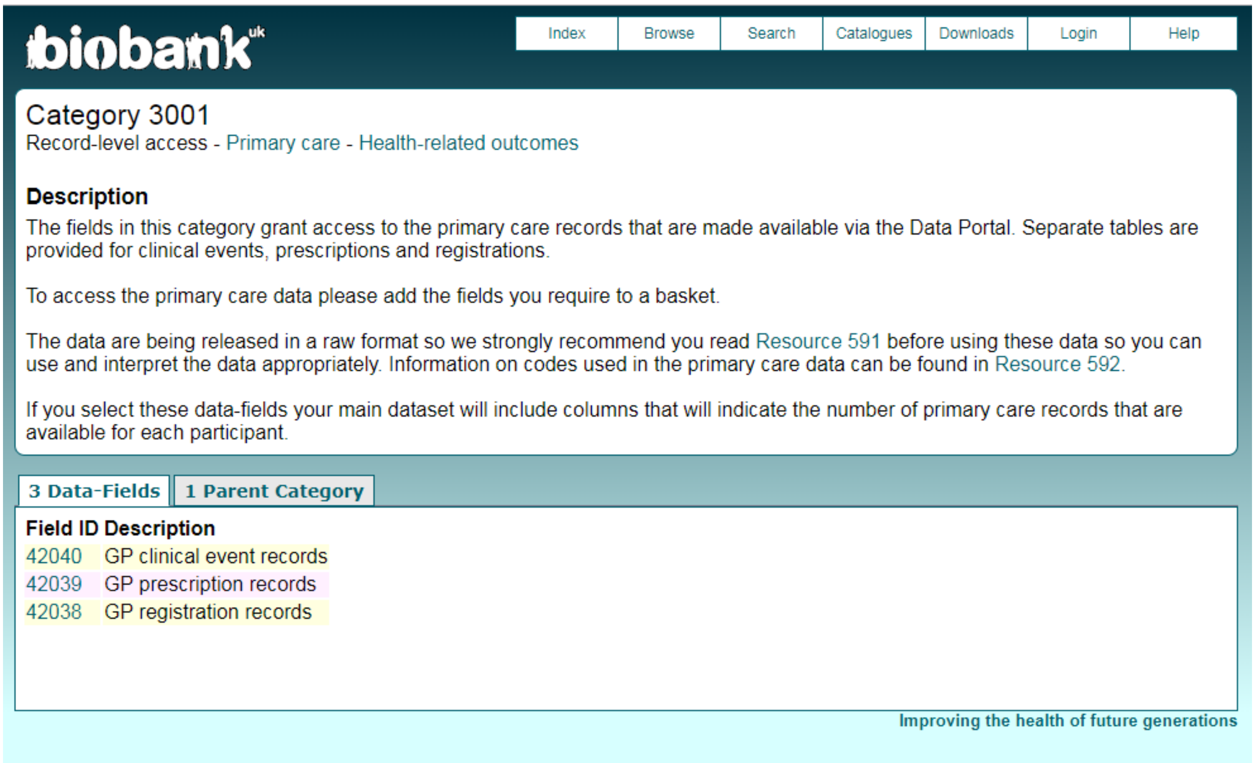
Table 13. Prescription records

Prescriptions table: gp_scripts	
Column name	**Description**
eid	Participant identifier
data_provider	1= England(Vision), 2= Scotland, 3 = England (TPP), 4 = Wales
issue_date	Date prescription was issued
read_2	Read v2
bnf_code	BNF code
dmd_code	DM+D code
drug_name	Drug name
quantity	Quantity issued

7. Access to approved datasets

7.1 Selecting the record-level access fields on Showcase

To access record level primary care data you must include the relevant data-field(s) from Category 3001 – Record-level access on Data Showcase in your basket of variables.



Category 3001

Record-level access - Primary care - Health-related outcomes

Description

The fields in this category grant access to the primary care records that are made available via the Data Portal. Separate tables are provided for clinical events, prescriptions and registrations.

To access the primary care data please add the fields you require to a basket.

The data are being released in a raw format so we strongly recommend you read Resource 591 before using these data so you can use and interpret the data appropriately. Information on codes used in the primary care data can be found in Resource 592.

If you select these data-fields your main dataset will include columns that will indicate the number of primary care records that are available for each participant.

3 Data-Fields | 1 Parent Category

Field ID	Description
42040	GP clinical event records
42039	GP prescription records
42038	GP registration records

Improving the health of future generations

There is a separate data field for each of the three Primary care tables (Table 14).

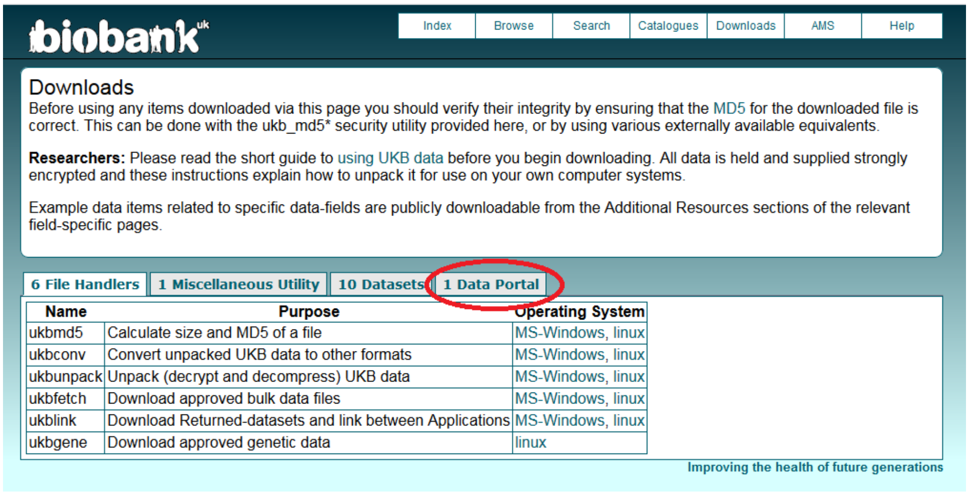
Table 14. Fields required to access primary care data tables from the Data Portal

Primary care table on the Data Portal	Data Showcase	
	Data field ID	Description
gp_registrations	42038	GP registrations records
gp_scripts	42039	GP prescription records
gp_clinical	42040	GP clinical event records

7.1 Navigating to the Data Portal to view the Primary care tables

Once your basket of variables has been approved, and you have received your notification email containing the key file you should:

1. log-in to the Application Management System: https://bbams.ndph.ox.ac.uk/ams/
2. Go to the Projects section for your Application and click on the Data tab. Then click on the "Go to Showcase download page" button which is at the bottom of the screen in the Data Download section. The Data Portal tab should display, as below.



Downloads

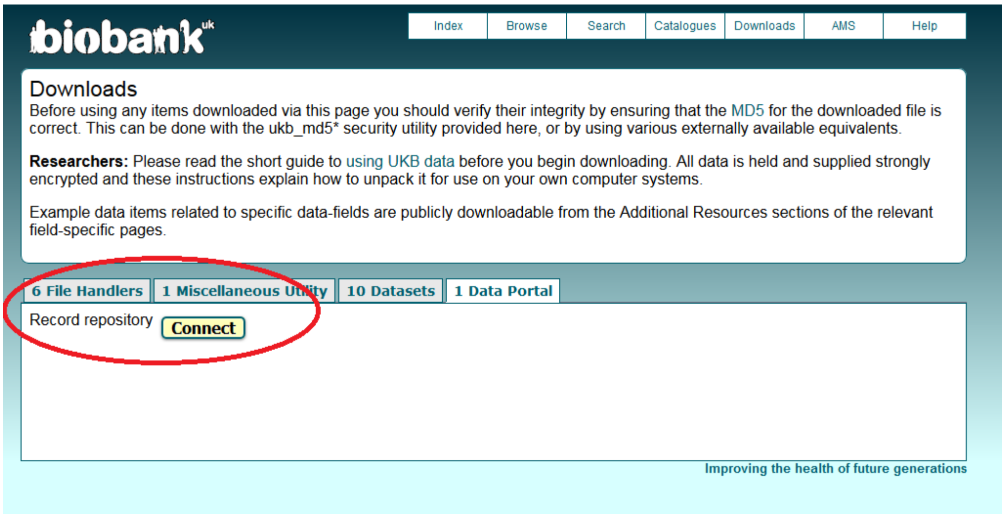
Before using any items downloaded via this page you should verify their integrity by ensuring that the MD5 for the downloaded file is correct. This can be done with the ukb_md5* security utility provided here, or by using various externally available equivalents.

Researchers: Please read the short guide to using UKB data before you begin downloading. All data is held and supplied strongly encrypted and these instructions explain how to unpack it for use on your own computer systems.

Example data items related to specific data-fields are publicly downloadable from the Additional Resources sections of the relevant field-specific pages.

Name	Purpose	Operating System
ukbmd5	Calculate size and MD5 of a file	MS-Windows, linux
ukbconv	Convert unpacked UKB data to other formats	MS-Windows, linux
ukbunpack	Unpack (decrypt and decompress) UKB data	MS-Windows, linux
ukbfetch	Download approved bulk data files	MS-Windows, linux
ukblink	Download Returned-datasets and link between Applications	MS-Windows, linux
ukbgene	Download approved genetic data	linux

3. Click ‘Connect’ to access the Record repository.



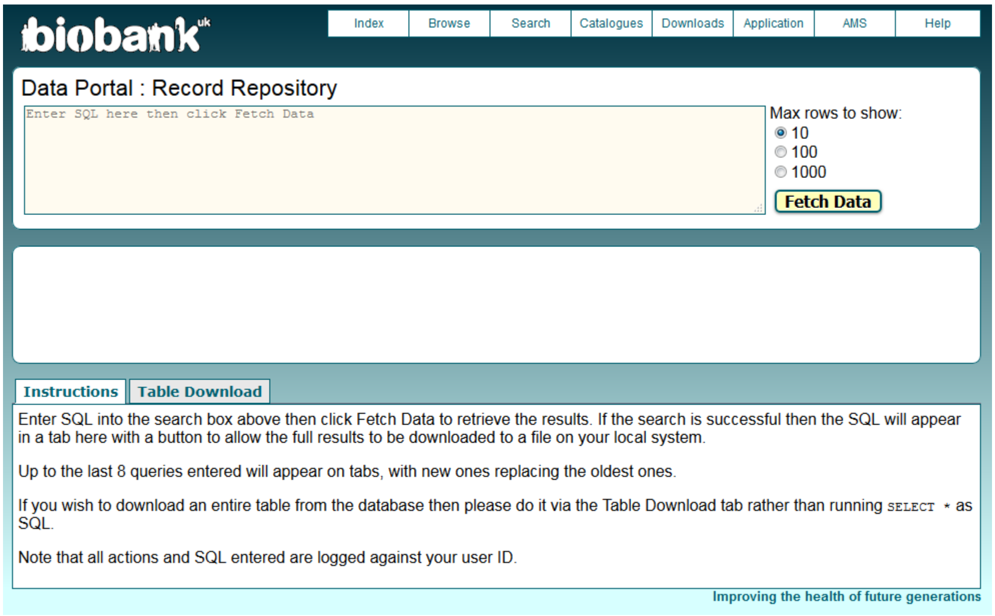
Downloads

Before using any items downloaded via this page you should verify their integrity by ensuring that the MD5 for the downloaded file is correct. This can be done with the ukb_md5* security utility provided here, or by using various externally available equivalents.

Researchers: Please read the short guide to using UKB data before you begin downloading. All data is held and supplied strongly encrypted and these instructions explain how to unpack it for use on your own computer systems.

Example data items related to specific data-fields are publicly downloadable from the Additional Resources sections of the relevant field-specific pages.

4. The Data Portal: Record Repository screen should now be visible:



Data Portal : Record Repository

Enter SQL here then click Fetch Data

Max rows to show:
● 10
○ 100
○ 1000

Enter SQL into the search box above then click Fetch Data to retrieve the results. If the search is successful then the SQL will appear in a tab here with a button to allow the full results to be downloaded to a file on your local system.

Up to the last 8 queries entered will appear on tabs, with new ones replacing the oldest ones.

If you wish to download an entire table from the database then please do it via the Table Download tab rather than running `SELECT *` as SQL.

Note that all actions and SQL entered are logged against your user ID.

7.2 Using the Data Portal

The Data Portal offers three different ways of working with the Primary Care data:

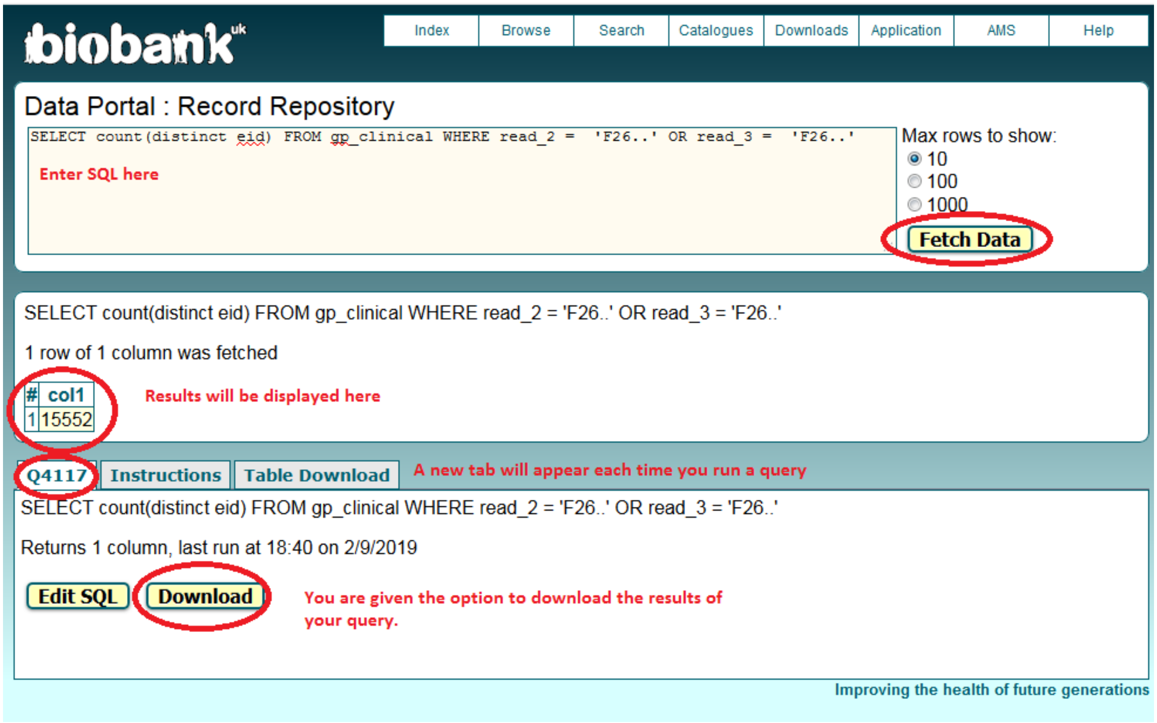
1. view the data in situ (suitable for simple, exploratory queries only),
2. download the results of simple queries,
3. download complete tables.

Some SQL examples are provided in **7.3 SQL examples**.

1. To **view data** within the Data Portal you can enter SQL code and then click the **'Fetch Data'** button. You can choose to restrict the number of rows of data that are displayed to 10, 100 or 1000. Your results will appear in a panel below the SQL box and each time you enter a new query a new tab will be added to the bottom panel.

   Please note that due to the size of the tables, some queries will not return any results before your web browser times-out. In these cases no results are returned and no error message is generated. If this happens you will need to either modify your SQL query so that it is computationally faster, or opt to download the complete table(s).

2. To **download the results of your SQL query**, enter the SQL code and click 'Fetch Data' to view the data. Your results will appear in a panel below the SQL box, and a new tab with a **download** button will be visible in the bottom panel. The results will be provided as a tab separated text file (.txt).



Data Portal : Record Repository

```
SELECT count(distinct eid) FROM gp_clinical WHERE read_2 =  'F26..' OR read_3 =  'F26..'
```

Enter SQL here

Max rows to show:
● 10
○ 100
○ 1000

SELECT count(distinct eid) FROM gp_clinical WHERE read_2 = 'F26..' OR read_3 = 'F26..'

1 row of 1 column was fetched

#	col1
1	15552

Results will be displayed here

A new tab will appear each time you run a query

SELECT count(distinct eid) FROM gp_clinical WHERE read_2 = 'F26..' OR read_3 = 'F26..'

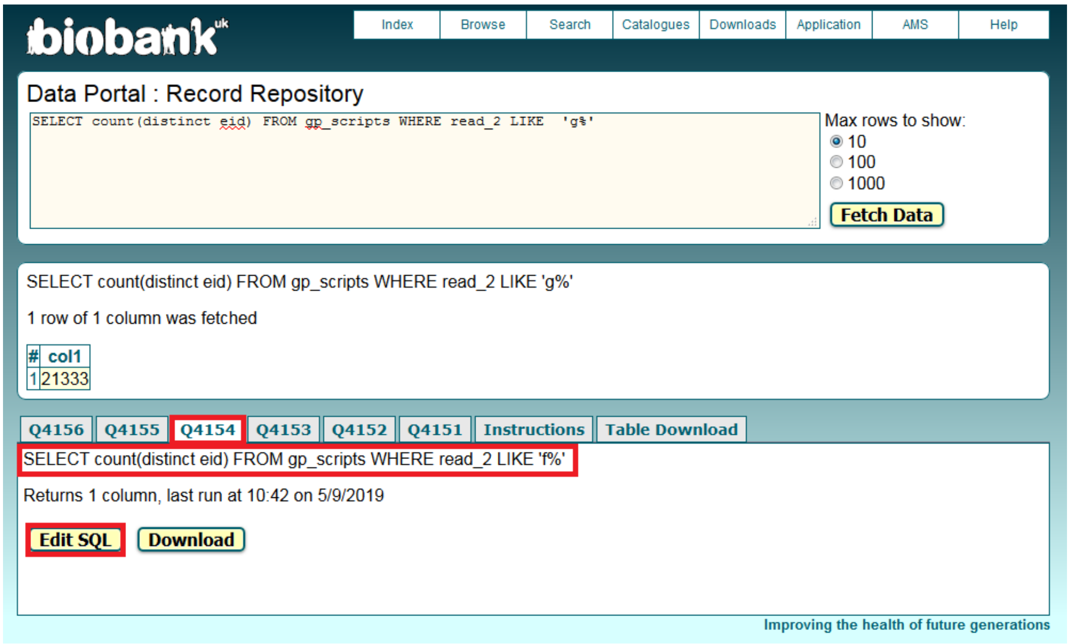
Returns 1 column, last run at 18:40 on 2/9/2019

You are given the option to download the results of your query.

Improving the health of future generations

The ‘Edit SQL’ button copies the SQL displaying in the tab visible in the bottom panel to the box in the top panel so you can edit the query. This allows you to adjust SQL queries without having to re-enter the full query again.

For example, in the screenshot below the last query run is displayed in the top and middle panels, but the bottom panel is displaying a tab relating to an earlier query (Q4154). To modify the query to count participants who have a read_2 code beginning ‘f9’, rather than all read_2 codes beginning ‘f’, click the ‘Edit SQL’ button and the query highlighted in red in the bottom panel will be moved to the top panel, allowing the SQL code to be amended. The ‘Fetch Data’ button must be clicked each time you run a query.



Data Portal : Record Repository

```
SELECT count(distinct eid) FROM gp_scripts WHERE read_2 LIKE  'g%'
```

Max rows to show:
● 10
○ 100
○ 1000

SELECT count(distinct eid) FROM gp_scripts WHERE read_2 LIKE 'g%'

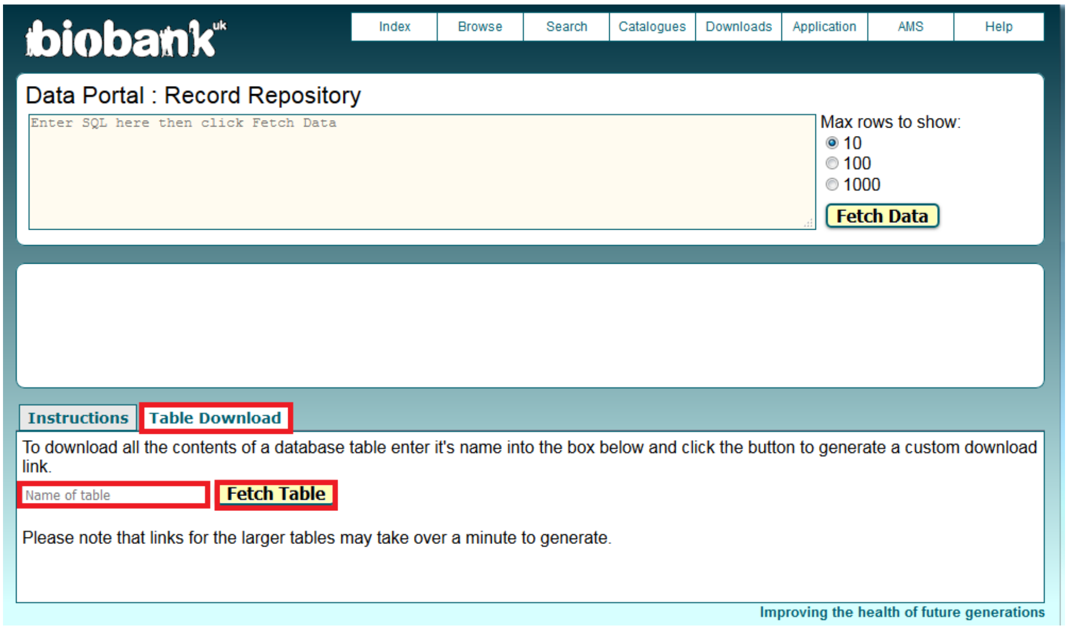
1 row of 1 column was fetched

#	col1
1	21333

SELECT count(distinct eid) FROM gp_scripts WHERE read_2 LIKE 'f%'

Returns 1 column, last run at 10:42 on 5/9/2019

3. To **download complete tables** click on the **‘Table Download’ tab** in the bottom panel, enter the name of the table you wish to download (e.g. gp_clinical) and click on the **‘Fetch Table’** button. This will generate a custom download link that you can paste into a web browser and a wget command for those using a linux system. The resulting dataset will be provided as a tab separated text file (.txt). Please note it can take some time to download the complete tables.



Data Portal : Record Repository

Enter SQL here then click Fetch Data

Max rows to show:
● 10
○ 100
○ 1000

To download all the contents of a database table enter it's name into the box below and click the button to generate a custom download link.

Name of table

Please note that links for the larger tables may take over a minute to generate.

4. Please note that if you include data-fields **42038**, **42039** and **42040** in your basket the fields will also appear as columns in the main dataset corresponding to that basket, with the values indicating for each participant the total number of records that participant has in the registrations, prescriptions and clinical tables, respectively.

7.3 SQL examples

The examples below are designed to illustrate the structure of the tables and basic SQL commands, rather than selection of Read codes, so the Read codes selected might not capture all cases.

Please note that due to the size of the tables, some queries can take some time to run and for some more computationally demanding queries browsers may time-out before any results are returned!

1. **Viewing all the primary care clinical records for participants with Read (version 2) code 'H060.'**

```
select * from gp_clinical where read_2 = 'H060.'
```

2. **Counting how many participants have a diagnosis of migraine recorded in either Read version 2 or Read version 3/CTV3**

```
select count(distinct eid)
from gp_clinical
where read_2 = 'F26..' or read_3  = 'F26..'
```

Note: for some conditions there will be different codes for Read version 2 and Read version 3.

3. **Viewing all the prescription records where the participant has been prescribed an opioid analgesic and this has been recorded in Read version 2**

```
select top 100 * from gp_scripts where left(read_2, 2) = 'dj'
```

Notes:

- *The top 100 * option in SQL can be useful for limiting the results of a query*
- *For some queries the semi-hierarchical structure of Read codes can be used to avoid having to list a long list of codes. In the query below we search for all Read 2 codes that begin dj.*
- *Be careful as nested within the hierarchy there are sometimes specific Read codes that indicate a person did not have the condition!*
- *The query below will not include participants whose medical records have been obtained from TPP as the prescriptions in those records are coded with BNF instead.*

4. **In the Welsh primary care records only, finding which medications participants with a diagnosis of asthma were prescribed; limiting the results to medications issued on the same day as the clinical event was recorded.**

```
select a.eid, a.event_dt, b.read_2 as read_2_script
from gp_clinical a join gp_scripts b
on (a.eid = b.eid) and (a.event_dt = b.issue_date)
where left(a.read_2, 3) = 'H33' and a.data_provider = 4
```

Notes:

- *In the interim Primary Care data release, consultation IDs have not been provided, however you might be able to link prescriptions with consultations by using the clinical event date (event_dt) and the issue date (issue_date).*
- *The results from this query are not necessarily prescribed for the diagnosis of asthma, as the participant might have attended with more than one complaint.*

8. Further information

An overview of all the linked health data which is available via the UK Biobank Showcase can be found in Resource 596.

Appendices

A. Glossary of terms

Term	Definition
BNF	British National Formulary
CTV3	Clinical Terms Version 3
dm+d	Dictionary of Medicines and Devices
GP	General Practice / Practitioner
ICD-9 / ICD-10	International Classification of Disease version 9 and 10 [18]
NHSBSA	NHS Business Services Authority
OPCS-4	Originally: Office of Population Censuses and Surveys Classification of Surgical Operations and Procedures version 4 [19], the coding system retains the name of the now defunct department (OPCS)
QOF	Quality and Outcomes Framework [20]
SNOMED CT	Originally: Systematized Nomenclature Of Medicine, but lost that meaning when merged with CTV3: SNOMED Clinical Terms, shortened to SNOMED CT
TPP	The Phoenix Partnership
TRUD	NHS Digital Technology Reference Data Update Distribution
UKB	UK Biobank

B. Validation checks

The data being made available in the initial primary care data release has been subject to minimal data cleaning.

Table 15. Data validation steps carried out on raw GP data

Check	Details
Event date	Where event date related to participant's date of birth it was replaced with a dummy date: • where the date precedes participant date of birth it has been changed to 01/01/1901. • Where the date matches participant date of birth it has been changed to 02/02/1902. • Where the date follows participant date of birth but is in the year of their birth it has been changed to 03/03/1903. Event dates in the future have been changed to 07/07/2037.
Disclosive text	Where free-text data was detected and its content was assessed as containing potentially disclosive information about the participant's identity or location, this was removed. The rest of the record was retained.

[18] https://www.who.int/classifications/icd/icdonlineversions/en/

[19] https://isd.digital.nhs.uk/trud3/user/guest/group/0/pack/10

[20] https://digital.nhs.uk/data-and-information/data-collections-and-data-sets/data-collections/quality-and-outcomes-framework-qof

C. Supporting material

Clinical coding classification lists have been compiled [see Resource 592] along with detailed information on their source, version number and content of fields. As Read v2 and CTV3 are now deprecated some of these are no longer available directly from source.

Table 16. Source of clinical coding classification systems included in supporting material

Type of information	Source	Detail
Code description look-ups	TRUD	• Read v2, Read v2 drugs, CTV3 [21] • dm+d [22]
	NHSBSA	• BNF [23]
Read code mappings	TRUD	Read v2 to: • CTV3 • BNF • ICD-9 • ICD-10 • OPCS-4)
		CTV3 to: • ICD-9 • ICD-10 • OPCS-4

[21] https://isd.digital.nhs.uk/trud3/user/guest/group/0/pack/9

[22] https://isd.digital.nhs.uk/trud3/user/guest/group/0/pack/26

[23] https://apps.nhsbsa.nhs.uk/infosystems/welcome