UK Biobank

First Occurrence of Health Outcomes Defined by 3-character ICD10 code

https://www.ukbiobank.ac.uk

September 2019

This companion document provides supporting information for the release of broad health outcome indicators in UK Biobank.

Contents

1. Introduction

1.1. Health outcomes in the UK Biobank

The linkage of all UK Biobank (UKB) participants to their health-related records enables researchers to investigate a broad range of health outcomes. All conditions that lead to an interaction with part of the National Health Service may be represented in the linked data.[1] Linkages currently available (as of summer 2019) cover hospital inpatient data, coded primary care data, cancer and death registry data.

The linked data are real-world, administrative data captured during the delivery of care, and were not designed or structured to readily facilitate research. UK Biobank has created algorithms (combinations of clinical codes with rules for case inclusion/exclusion, where appropriate) for some specific health outcomes, which are available via the Data Showcase.[2] These algorithms have gone through an expert peer review and consensus processes, with information provided on the positive predictive value and other validation, where possible.

In the absence of robust definitions of the many thousands of other conditions of potential interest to researchers, we have generated data-fields to indicate the first occurrence of a set of diagnostic codes for a wide range of health outcomes across self-report, primary care, hospital inpatient data and death data, mapped to a 3-digit code of International Classification of Disease (ICD-10).[3]

1.2. Health outcome coding classifications in UK Biobank

There are two main classification systems of clinical coding used in the linked health data: ICD and Read.

- Hospital in-patient records (ICD10 and ICD9)
- Death records (ICD10)
- Cancer register (ICD10 and ICD9)
- Primary care records (Read v2 and Read CTV3)

Self-report of health conditions from the UK Biobank assessment clinics is also available. Participants were asked to report all diagnoses of conditions during the touchscreen questionnaire and the details were subsequently checked during the verbal interview with a nurse. These health outcomes are therefore not verified by a clinician nor provide a fully comprehensive health history, but reflect the best recollections of individual participants.

The 'first occurrence' fields define each health outcome by the 3-character codes within ICD10's diagnostic chapters, excluding cancer, and use mapped codes from the other classification systems mentioned above on which more details are provided below. We have excluded cancer codes from this mapping as these outcomes are comprehensively captured via the cancer register data.

Please note that these code lists have not been reviewed by clinicians or externally validated and should thus be thought of as a first pass to identify cases with any given condition.

[1] http://biobank.ndph.ox.ac.uk/showcase/exinfo.cgi?src=UnderstandingUKB
[2] http://biobank.ndph.ox.ac.uk/showcase/label.cgi?id=42
[3] https://www.who.int/classifications/icd/en/

2. Mapping between coding systems

2.1.The ICD10 'spine'

The whole ICD10 classification system consists of around 18,000 codes across 22 chapters. Chapters contain codes for related conditions and sub-types such as disorders of the circulatory (chapter IX) or respiratory systems (chapter X). The full code look-up for the latest release (version 5) can be downloaded from TRUD (Technology Reference Data Update Distribution, NHS Digital).[4]

ICD10 chapters XVIII to XXII (codes beginning R00 to Z99) were excluded as they encompass entities that are broader (e.g. symptoms, injury, external causes of morbidity, factors influencing health status) than the diagnostic health outcomes found in chapters I to XVII. As noted previously, codes for cancer (chapter II - codes beginning C00 to D48) are also excluded, as the most reliable source of this information is already available in UK Biobank via the cancer register data.[5] Thus, of the available 2,000 unique 3-character ICD10 codes, around 1,200 were used as the 'spine' for this set of health outcomes.

Table 1 shows several examples of the full 4-character ICD10 code with its description, and the associated 3-character ICD10 code used by UK Biobank to identify the "first occurrence of health outcome" fields. These are highlighted in the green cells. In the examples given, for each 3-character ICD10 code there are up to seven associated 4-digit ICD10 codes, which are included within that health outcome to indicate first occurrence for an individual.

ICD10 incorporates the asterisk (*) and dagger (†) system, whereby two codes are used as a pair to provide diagnostic information about an underlying condition and its manifestation in a particular body site that is considered a clinical problem in its own right. Examples can be found in the WHO documentation and online.[6] For further details about how these codes are used, please refer to the full WHO ICD10 guide.[7]

[4] https://isd.digital.nhs.uk/trud3/user/guest/group/0/pack/28

[5] http://biobank.ctsu.ox.ac.uk/crystal/label.cgi?id=100092

[6] https://www.ndc.scot.nhs.uk/Data-Dictionary/SMR-Datasets//General-Clinical-Information/Diagnostic-Section/Dagger-and-Asterisk-Pairs.asp

[7] https://www.who.int/classifications/icd/ICD-10_2nd_ed_volume2.pdf

Table 1. Selected examples of 4-character ICD10 codes and relation to 3-character level code, code range, chapter and description at each hierarchical level. The green cells indicate the level at which the health outcomes for the first occurrence are defined.

ICD10 chapter	Grouped 3-character ICD10		3-character ICD10		4-character ICD10	
	Code range	**Description**	**Code**	**Description**	**Code**	**Description**
Chapter IX Diseases of the circulatory system	I20-I25	Ischaemic heart diseases	I21	Acute myocardial infarction	I21	Acute myocardial infarction
					I210	Acute transmural myocardial infarction of anterior wall
					I211	Acute transmural myocardial infarction of inferior wall
					I212	Acute transmural myocardial infarction of other sites
					I213	Acute transmural myocardial infarction of unspecified site
					I214	Acute subendocardial myocardial infarction
					I219	Acute myocardial infarction, unspecified
			I22	Subsequent myocardial infarction	I22	Subsequent myocardial infarction
					I220	Subsequent myocardial infarction of anterior wall
					I228	Subsequent myocardial infarction of other sites
					I229	Subsequent myocardial infarction of unspecified site
					I221	Subsequent myocardial infarction of inferior wall
	I26-I28	Pulmonary heart disease and diseases of pulmonary circulation	I26	Pulmonary embolism	I26	Pulmonary embolism
					I260	Pulmonary embolism with mention of acute cor pulmonale
					I269	Pulmonary embolism without mention of acute cor pulmonale

2.2.Read codes

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.[8] Read v2 and CTV3 were last updated in April 2016 and April 2018, respectively. Both versions of Read codes are now deprecated (as is the Read Browser) and no further updates will occur. From April 2018 SNOMED CT (https://digital.nhs.uk/snomed-ct) 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.

2.2.1. Read v2 mapping

TRUD also provides maps between Read2 and ICD10 codes. These mapped codes have been incorporated into the "first occurrence" fields if there was an unambiguous mapping to a 3-character ICD10 code. The types of included mappings are described in Table 2. Read v2 codes that mapped to more than one 3-character ICD10 code were not included. Note that one-to-one mappings for both 'asterisk' and 'dagger' codes as described in the previous section are retained, as are maps to codes paired with a plus sign in the mapping (see Table 2 for examples). This, and the loss of all Read v2 codes which mapped to more than one 3-character ICD10 code, may have implications for the interpretation of the data and researchers should investigate this thoroughly before analysing and interpreting these data.

Table 2. Retention of mapped Read v2 codes

Read v2 code	Mapped ICD10 code	Mapping type	Included?
Ayu00	A009	1:1 map - Read v2 term maps to a single ICD10 code	yes
A0…	A00-A09	Range map: Read v2 term maps to any in a range of ICD10 codes	yes - only if all matches were to the same 3-character ICD10 code
Ab1..	B358, B359	Combination – any of the ICD10 codes given are possible cross-maps for this Read v2 term	
D212.	D630a	Asterisk code (dagger code unspecified)	yes
D2120	A010d G01Xa, G01Xa A010d	Dagger and asterisk code pair – the listed codes can be accepted in any sequence	yes
Ayu15	A022d	Dagger code (asterisk code unspecified)	yes
J666.	A419+K839	Plus sign used when using two diagnostic codes where both are needed to describe the Read v2 term	yes – to both ICD10 codes

Overall, there are around 102,000 Read v2 codes, of which around 20,000 mapped to ICD10 3-character codes in the included chapters described above.

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

2.2.2. Read CTV3 mapping

There are around 277,000 unique CTV3 codes (many have primary and multiple secondary terms), of which around 27,000 mapped to ICD10 3-character codes. Note that CTV3 includes administrative, medication and procedure codes, which may also be used to indicate diagnoses. The "first occurrence" fields used only the disorders section for consistency with the Read v2 mappings.

2.3. ICD9 mapping

There are around 8,000 ICD9 codes, of which around 2,400 mapped to ICD10 3-character codes. As with other classification systems, cancer codes and ICD9 codes that mapped to more than one ICD10 3-character code were excluded.

2.4. Mapping self-report at UK Biobank assessment clinics

Participants were asked to report all health conditions diagnosed by a doctor when they attended the assessment clinic at recruitment (and at repeat assessments if they attended them) and their answers were verified with a nurse during the verbal interview (UKB Field 20002). We have created mappings from all reported diagnosed conditions to the 3-character ICD10 codes where appropriate.

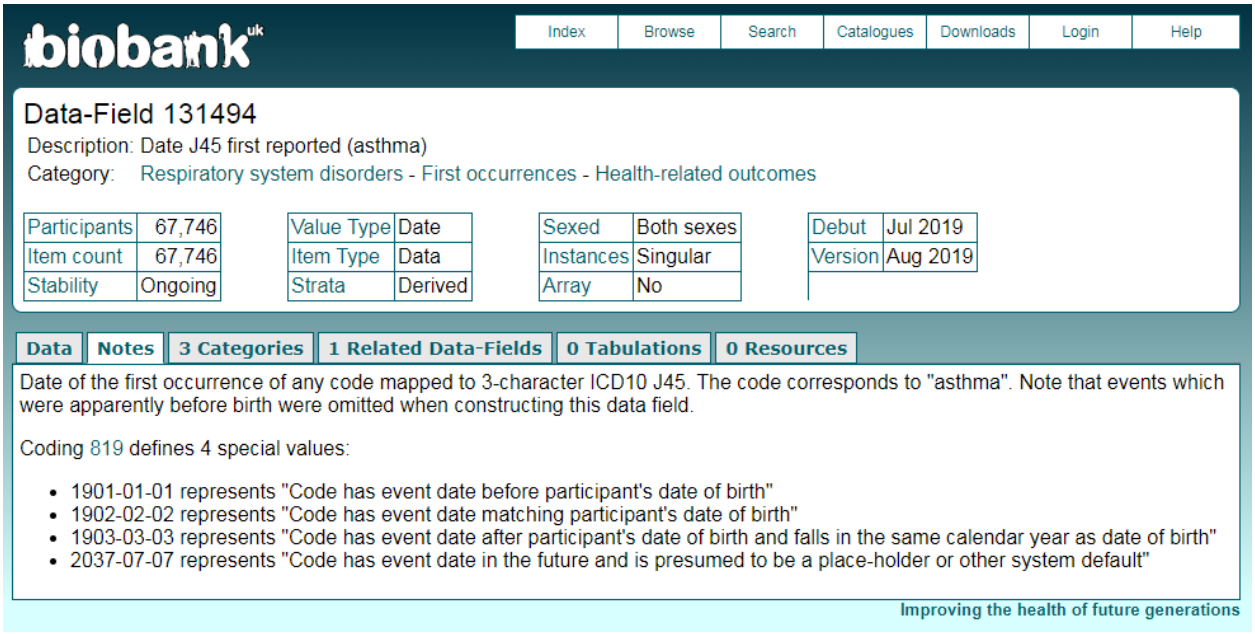
3. Populating the first occurrence of health outcome fields on Showcase

For each of the approximately 1,200 health outcomes defined by 3-character ICD10 codes, two data fields have been created in Data Showcase Category 1712 – first occurrences: the earliest date and the source of the code.

3.1. Earliest date of first occurrence

This data-field shows the earliest date that the 3-character ICD10 (or one of its mapped codes as described in the previous section) was recorded through either self-report at any assessment centre, inpatient hospital data, primary care or death record data.

For example, Data-field 131494 - Date J45 first reported (asthma)



The selection of date information from each data source is described in Table 3.

Table 3. Details of date field by source

Source	Date field	Notes
Death register	Data-field 40000	This data-field provides the date of death.
Hospital inpatient	epistart, admidate, epiend, disdate (in the hesin table)	Hospital inpatient data does not record the diagnosis date directly, but rather information about the dates the hospital episode started and ended and the dates the hospital admission started and ended. Given that a hospital stay can include multiple episodes, the episode start date (epistart) has been used as the best proxy for the diagnosis date. If this was missing, we used the admission date (admidate), episode end date (epiend) or discharge date (disdate) instead. Please refer to the Inpatient Data Dictionary (Resource 141140) for details of how availability of these date fields varies systematically by data provider.
Primary care	Event date (event_dt in the gp_clinical table)	Date event was recorded in primary care. We have altered some dates in relation to participant date of birth as follows: • where clinical event or prescription 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 • Where the date is in the future (and is presumed to be a place-holder or other system default) it has been changed to 07/07/2037. For the purpose of identifying first occurrence dates in the primary care

Source	Date field	Notes
		data, records with the 'special' event date 01/01/1901, or those with the value 01/01/1900 (suggesting the date was unknown or missing) were excluded. The other 'special' event dates are treated as regular dates, but researchers are advised to replace these with appropriate dates based on the month and year of birth (data-fields 52 and 34 respectively).
Self-report	Derived from data-field 20008	This data-field gives the interpolated year when non-cancer illness first diagnosed. During the verbal interview at the UK Biobank assessment centre participants were asked when they had first been diagnosed (by a doctor) with each condition they self-reported. Participants could provide a year, or their age at diagnosis. These were then converted to interpolated year. • If the participant gave a calendar year, then the best-fit time is half-way through that year. For example if the year was given as 1970, then the value presented is 1970.5 • If the participant gave their age then the value presented is the fractional year corresponding to the mid-point of that age. For example, if the participant said they were 30 years old then the value is the date at which they were 30years+6months. • Interpolated values before the date of birth were truncated forwards to that time. • Interpolated values after the time of data acquisition (i.e. the date of the assessment centre visit) were truncated back to that time. For the purposes of identifying first occurrences, the interpolated years were converted to dates, and if the date corresponding to a self-reported medical condition was 'unknown' (-1) or 'prefer not to answer' (-3) then these were ignored. Data from all assessment centre visits are included as some participants have attended more than one assessment centre (e.g. the Baseline visit and Imaging assessment centre). Please note that the transformation of the self-reported diagnosis age/year, coupled with participant recall means the diagnosis dates from self-report should be viewed as approximate.

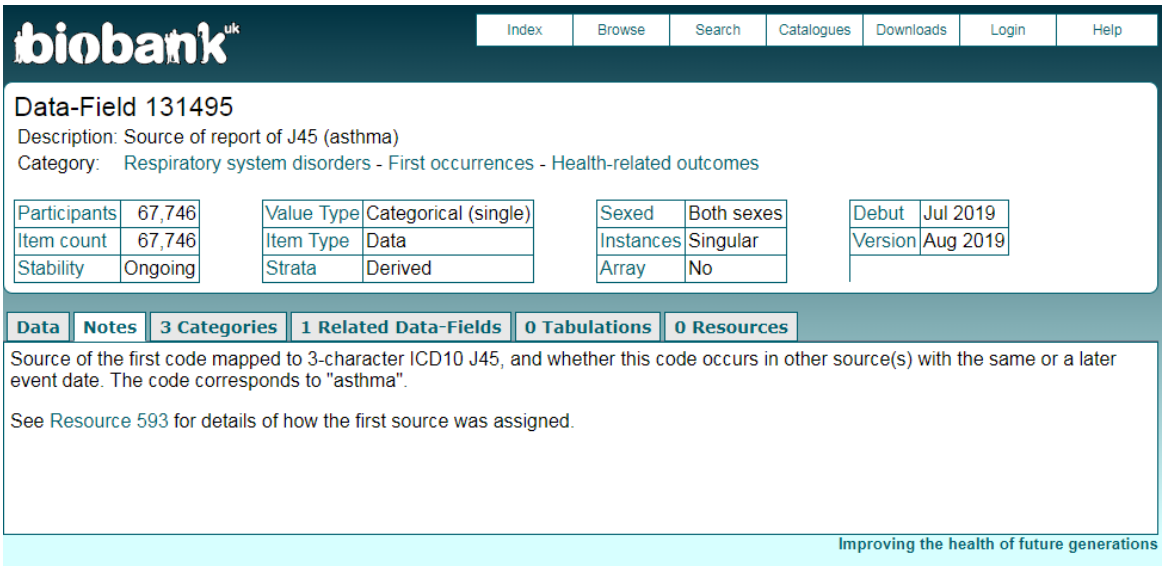
Further details on the mapping of the diagnosis information in each source to 3-character ICD10 codes and information about how the earliest occurrence dates have been extracted from each source can be found in the Appendix:

- Extraction of first occurrence data from Death Register data
- Extraction of first occurrence data from Primary care data
- Extraction of first occurrence data from Hospital inpatient data
- Extraction of first occurrence data from self-report data

3.2. Source data of first occurrence

This data-field is an integer indicating the source in which the earliest instance of each 3-character ICD10 (or mapped) code was recorded (e.g. hospital inpatient, primary care, death record, or self-report) and whether the code was recorded in at least one other source (with matching event date or a subsequent event date).

For example, Data-field 131495 – source of report of J45 (asthma)



The source definitions are provided in Table 4 and are available on Showcase as Data-Coding 2171.

Table 4. Definition of integer field which indicates source of code mapped to 3-character ICD10 code

Code	Name	Definition
20	Death register only	Code only recorded on the death record
21	Death register and other source(s)	Code recorded on the death record, and at least one other source (i.e. primary care, hospital admissions or self-report). [9]
30	Primary care only	Code only recorded within primary care records
31	Primary care and other source(s)	Code recorded within primary care and at least one other source (i.e. death, hospital admissions or self-report)
40	Hospital admissions data only	Code only recorded within hospital admissions records
41	Hospital admissions data and other source(s)	Code recorded within hospital admissions and at least one other source (i.e. death, primary care or self-report)
50	Self-report only	Code only recorded within self-report
51	Self-report and other source(s)	Code recorded within self-report and at least one other source (i.e. death, primary care or hospital admissions)

If the same 'first occurrence date' was recorded in more than one source, the source mentioned in the source field explicitly, e.g. Primary care in code 31, is selected based on which 'first source' features highest in this ordered list:

[9] The diagnosis code would only appear for the first time on a death record and in self-report if the participant died on the day they attended the clinic, or there was an error in the dates recorded in one or more of the sources.

1. Death register,
2. Primary Care,
3. Hospital admissions,
4. Self-report.

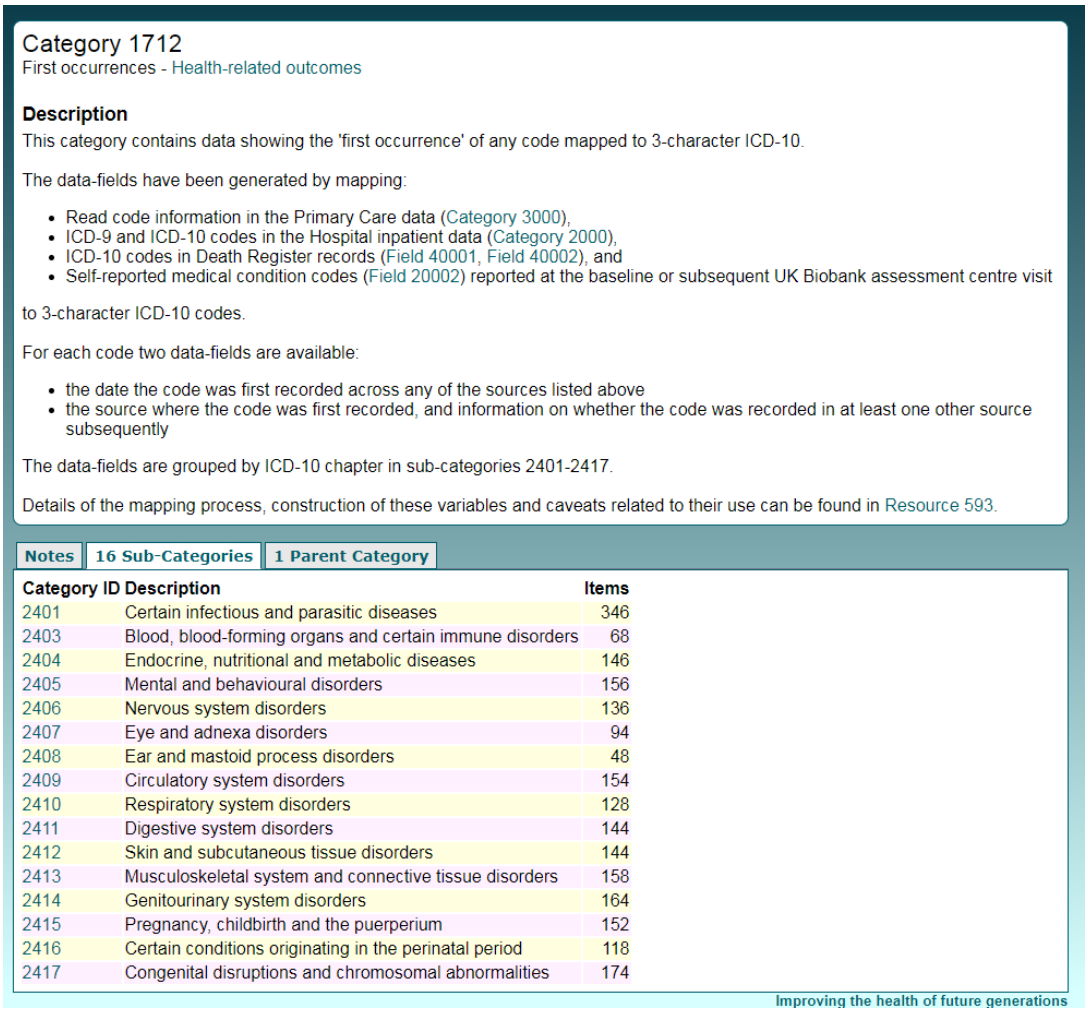
The ordering of this list was selected arbitrarily.

For example, if a participant has the first occurrence dates for asthma as shown in the table below, then the source of the first occurrence would be recorded as 31 = 'Primary care and other', because although the earliest first occurrence date of 23/04/2006 is present in both the Primary care and Hospital admissions data, the Primary Care data source features earlier in the ordered list (above).

Data source	First occurrence date within data source
Death register	17/02/2010
Primary care	23/04/2006
Hospital admissions data	23/04/2006
Self-report	-

The coding of the source field gives the flexibility to identify all participants with any code that maps to a particular 3-character ICD10 code, excluding participants who only have a record of the code in one source. For example, given the lack of explicit clinical confirmation, for some analyses it might be desirable to exclude participants for whom the only record of the participant having the condition is through self-report data from the assessment centre.

The data-fields are grouped into sub-categories related to the ICD10 chapters which incorporate diagnostic information as described above (i.e. Chapters I to XVII, excluding Chapter II Neoplasms).



Category 1712
First occurrences - Health-related outcomes

Description

This category contains data showing the 'first occurrence' of any code mapped to 3-character ICD-10.

The data-fields have been generated by mapping:

- Read code information in the Primary Care data (Category 3000),
- ICD-9 and ICD-10 codes in the Hospital inpatient data (Category 2000),
- ICD-10 codes in Death Register records (Field 40001, Field 40002), and
- Self-reported medical condition codes (Field 20002) reported at the baseline or subsequent UK Biobank assessment centre visit

to 3-character ICD-10 codes.

For each code two data-fields are available:

- the date the code was first recorded across any of the sources listed above
- the source where the code was first recorded, and information on whether the code was recorded in at least one other source subsequently

The data-fields are grouped by ICD-10 chapter in sub-categories 2401-2417.

Details of the mapping process, construction of these variables and caveats related to their use can be found in Resource 593.

Notes | 16 Sub-Categories | 1 Parent Category

Category ID	Description	Items
2401	Certain infectious and parasitic diseases	346
2403	Blood, blood-forming organs and certain immune disorders	68
2404	Endocrine, nutritional and metabolic diseases	146
2405	Mental and behavioural disorders	156
2406	Nervous system disorders	136
2407	Eye and adnexa disorders	94
2408	Ear and mastoid process disorders	48
2409	Circulatory system disorders	154
2410	Respiratory system disorders	128
2411	Digestive system disorders	144
2412	Skin and subcutaneous tissue disorders	144
2413	Musculoskeletal system and connective tissue disorders	158
2414	Genitourinary system disorders	164
2415	Pregnancy, childbirth and the puerperium	152
2416	Certain conditions originating in the perinatal period	118
2417	Congenital disruptions and chromosomal abnormalities	174

Improving the health of future generations

Figure 1. Screenshot of category 1712 on Data Showcase listing the ICD10 chapter names- clicking on each chapter links to summary fields for each of the health outcomes defined by 3-character ICD10 code

4. Using health outcomes defined by 3-character ICD10

4.1. Creating sub-cohorts

The first occurrence data-fields are provided to support UK Biobank researchers who wish to quickly identify sub-cohorts of individuals with respect to health outcomes not already covered by detailed algorithms[10] or code lists available elsewhere.[11] [12] The diagnostic ICD and Read codes mapped to each 3-character ICD10 code have not been individually curated to sensitively or specifically pick out individual conditions and any results should be analysed carefully. As with all administrative data (which are subject to a range of potential biases and risk of being incomplete), the presence of a code should not be assumed to infer a confirmed diagnosis. Similarly, its absence does not guarantee that the participant was not affected by the health outcome, especially where the data source is missing (i.e. primary care

[10] http://biobank.ctsu.ox.ac.uk/crystal/label.cgi?id=42
[11] http://biobank.ndph.ox.ac.uk/showcase/refer.cgi?id=594
[12] https://www.caliberresearch.org/portal

for some participants) or the code was present but there was an invalid date, e.g. an event date before date of birth.

All clinical code classification lookups and maps between systems are available with detailed source information.[13] Further information from participants' episode-level hospital and primary care data may provide details that allow researchers to assess whether or not individuals should be assigned as 'cases' with a health outcome or not depending on their planned analyses.

4.2. Prevalent and incident 'cases'

This 'first occurrence' date may be compared with the participant's baseline assessment centre attendance date as a proxy indicator for 'prevalent' and 'incident' cases of a given health outcome in the linked data. Given the nature of these data as described above, and the unknown veracity of self-report dates, such analyses should be undertaken with care. In particular, 'first occurrence' dates that are special date values, e.g. 02/02/1902 and 0/3/03/1903 should be replaced with appropriate dates based on the participant's month and year of birth.

4.3. Data availability

The (approximately) 1,200 broad health outcomes identified in the first occurrences fields combine clinical codes found in data from a number of different sources in UK Biobank, the availability of which varies by participant. The amount of linked inpatient and death record data for each participant is dependent on its availability from their country's provider[14] and the individual's interactions with health services. Primary care data are currently available for about 45% of the UK Biobank cohort and hence coded data from this source cannot contribute to the 'first occurrence' data-fields for some individuals. As a result, many conditions will either be unascertained (if they are only found in primary care) or will be ascertained at a later date (e.g., if it is coded through hospital admissions data, as either a primary or underlying condition after they were diagnosed through primary care). Self-reported conditions are available for all participants from the initial recruitment visit to the assessment centres, while others have returned for repeat assessment and/or an imaging visit. Consequently the first occurrence fields provide a summary of what has been recorded about participants' health status, based on data that were available in the resource when the code maps were incorporated.

4.4. Data cleaning

The health outcome fields use record level data as inputs without applying any comprehensive cleaning to remove or correct inconsistent dates or other data quality issues. However, the fields ignore event dates in the primary care data that have been amended to protect the date of birth (see Table 3), and those that were in the future when incorporated into the dataset. Consequently some first dates reported in the algorithm fields might be subject to recording errors. The self-report date provided by participants for each health outcome is prone to recall error. Researchers should consider the relative reliability of dates and note that the first occurrence date provided in the Showcase field may be the earliest 'valid' (i.e. non-missing) date, but may not be the most reliable to indicate when a diagnosis first occurred.

---

[13] http://biobank.ndph.ox.ac.uk/showcase/refer.cgi?id=592

[14] http://biobank.ndph.ox.ac.uk/showcase/exinfo.cgi?src=Data_providers_and_dates

4.5. Code mapping completeness

The code mappings behind the first occurrence fields were used exactly as downloaded from source. Information on the evolution of each classification system and the methodologies for mapping between them is available in the source material.[15] The veracity of the mapping between systems is likely to be subject to debate by experts and unlikely to be fully comprehensive. Researchers should verify the underlying codes with respect to any particular health outcome before drawing conclusions on the data.

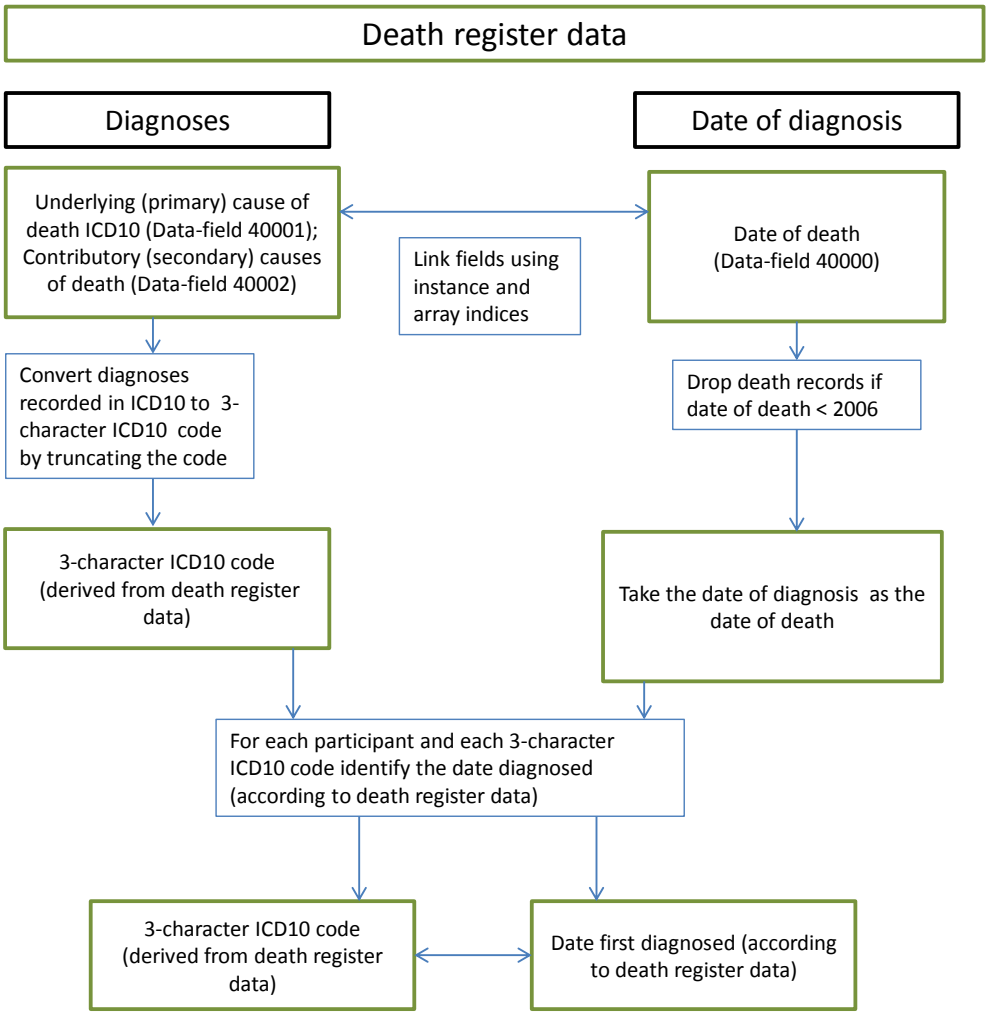
[15] https://isd.digital.nhs.uk/trud3/user/authenticated/group/0/pack/9

Appendix

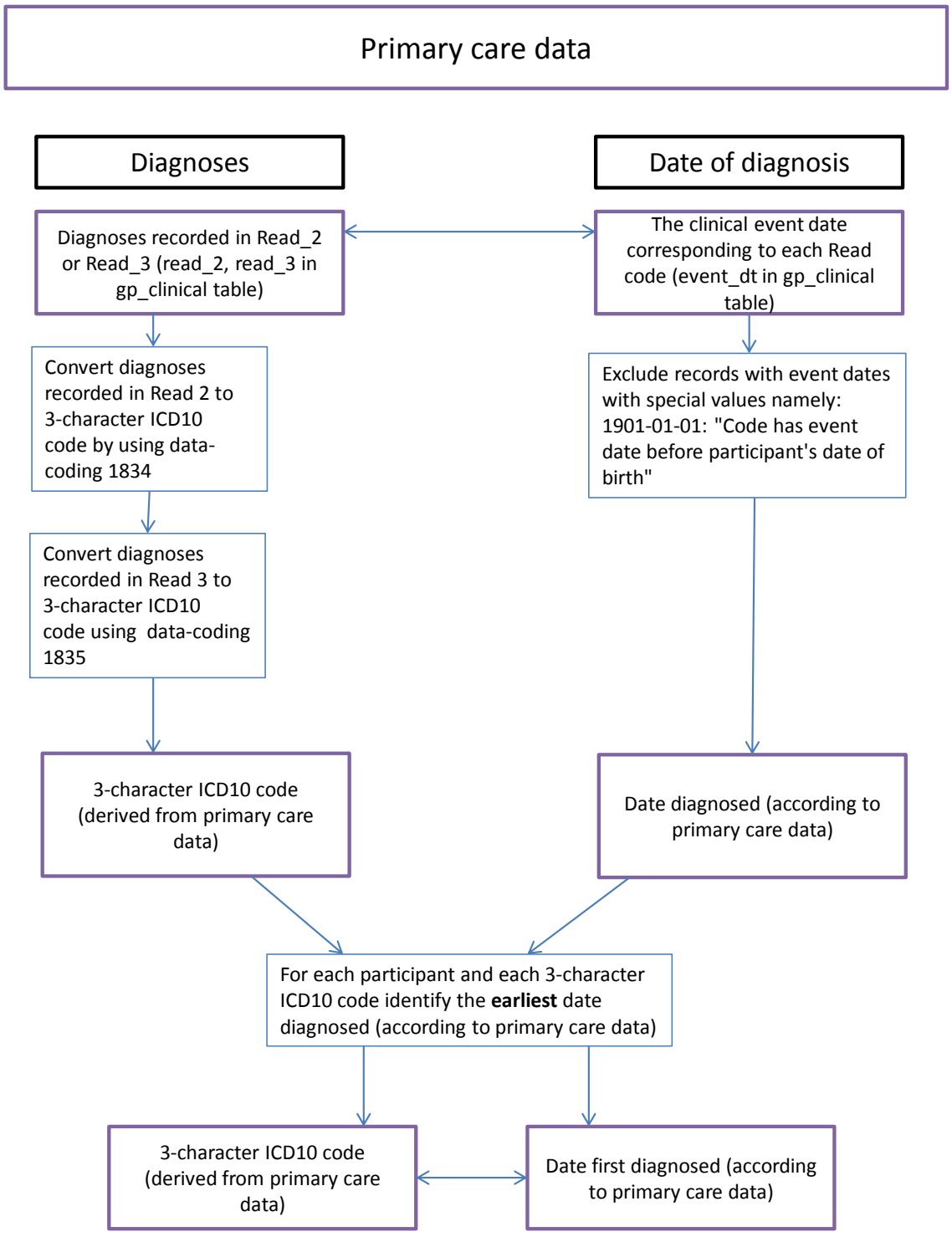
Glossary

Term	Definition
CTV3	Clinical Terms Version 3
ICD9 / ICD10	International Classification of Disease version 9 and 10
NHSBSA	NHS Business Services Authority
TRUD	NHS Digital Technology Reference Data Update Distribution
UKB	UK Biobank

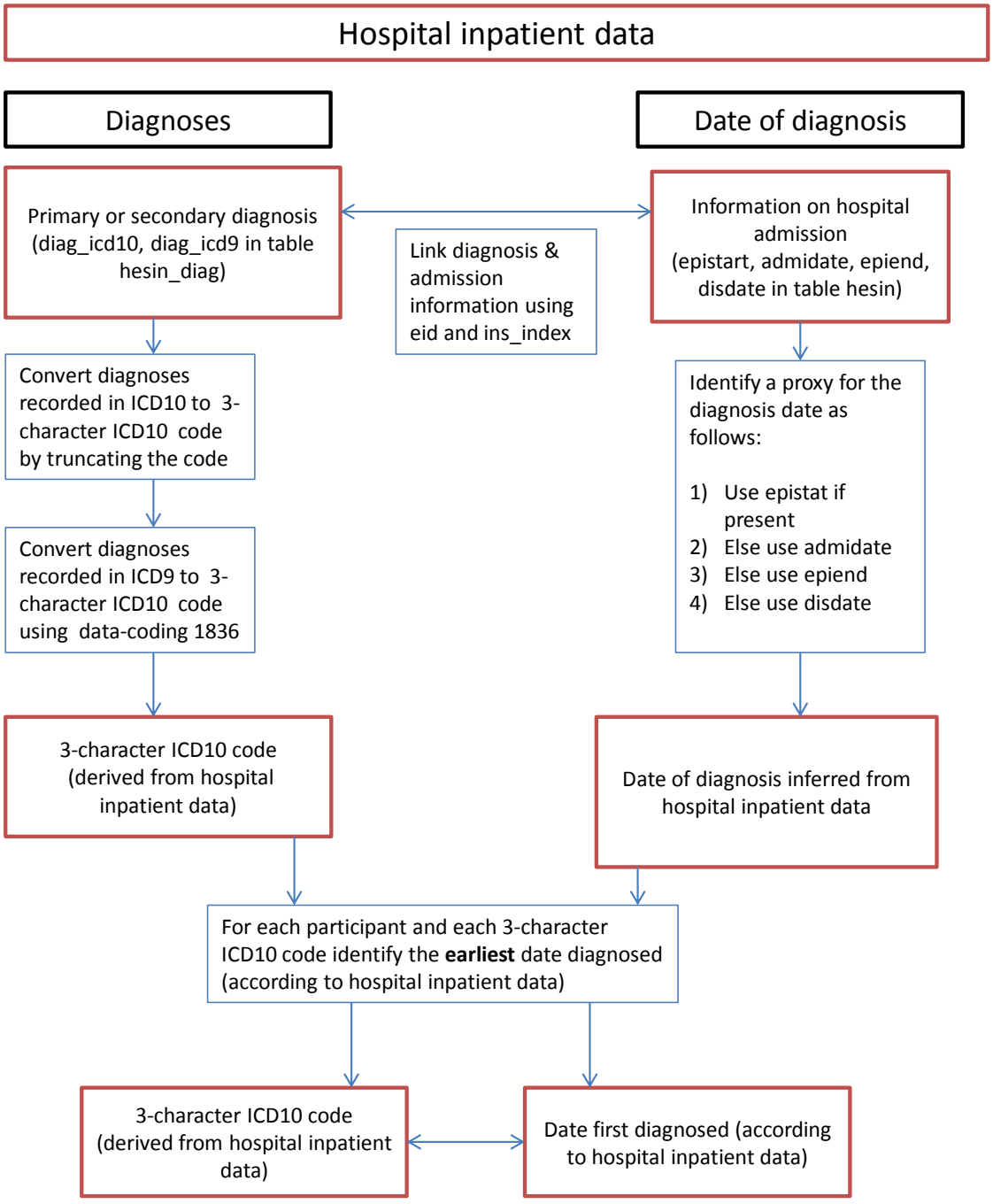
Extraction of first occurrence data from Death Register data



Extraction of first occurrence data from Primary care data



Extraction of first occurrence data from Hospital inpatient data



Extraction of first occurrence data from self-report data