**CPRD Aurum linkage data set 21**

1. Datasets and tables

|  |  |  |  |
| --- | --- | --- | --- |
| **Dataset** | **Description** | **Fields** | **Observations** |
| CHESS | COVID-19 Hospitalisations in England Surveillance System  Laboratory-confirmed cases | 90 + 1\* | 13,343 |
| HES APC | Hospital Episode Statistics Admitted Patient Care  Tables:  1. Patient (hes\_patient)  2. Hospitalisations (hes\_hospital)  3. Episodes (hes\_episodes)  4. Diagnoses (hes\_diagnosis\_epi, hes\_diagnosis\_hosp, hes\_primary\_diagnosis\_hosp)  5. Procedures [operations] (hes\_procedures\_epi)  6. Augmented care (hes\_acp)  7. Critical care (hes\_ccare)  (We haven’t requested 8. Health Resource Group or 9. Maternity) | 6  11  21  8, 6 +1\*, 6 +1\*  10  21\*\* + 1\*  36 + 1\* | 1,333,911  15,003,568  18,426,574  97,225,204,  71,105,386,  15,592,213  27,030,397  60,025  276,801 |
| IMD | Tables:  1. Patient (IMD 2015)  2. Practice | 3  3 | 1,465,167  1,441 |
| ONS death | Office for National Statistics death data  (We haven’t requested causes of neonatal death) | 26 | 348,895 |
| SGSS | Second Generation Surveillance System [COVID testing]  Just positive tests from PCR swab testing (not antibodies) | 10 + 1\* | 27,813 |

See ‘Linkage datatypes’ spreadsheet for fields and their datatypes

\*Extra field added by me (primary\_key)

\*\*NB: our data download does not have ‘acpdqind’ field in Data Specification (field no. 22)

2. Lookup tables

|  |  |  |
| --- | --- | --- |
| **Lookup table** | **Used in which tables above** | **No. of codes** |
| acpdisp | HES APC (hes\_acp table, acpdisp field) | 10 |
| acploc | HES APC (hes\_acp table, acploc field) | 18 |
| acpout | HES APC (hes\_acp table, acpout field) | 5 |
| acpplan | HES APC (hes\_acp table, acpplan field) | 3 |
| acpsour | HES APC (hes\_acp table, acpsour field) | 11 |
| acpspef | HES APC (hes\_acp table, acpspef field) | 209 |
| admimeth | HES APC (admimeth fields in hes\_hospital and hes\_episodes)  Note from data spec: From April 2013, National Codes 2A, 2B, 2C and 2D have been introduced to replace National Code 28 'Other means'.  **NB: has 3 columns – middle column = admimeth\_type** | 22 |
| admisorc | HES APC (admisorc fields in hes\_hospital and hes\_episodes) | 24 |
| ccadmisorc | HES APC (hes\_ccare table, ccadmisorc field) | 5 |
| ccadmitype | HES APC (hes\_ccare table, ccadmitype field) | 6 |
| ccapcrel | HES APC (hes\_ccare table, ccapcrel field) | 14 |
| ccdisdest | HES APC (hes\_ccare table, ccdisdest field) | 6 |
| ccdisloc | HES APC (hes\_ccare table, ccdisloc field) | 10 |
| ccdisstat | HES APC (hes\_ccare table, ccdisstat field) | 11 |
| ccsorcloc | HES APC (hes\_ccare table, ccsorcloc field) | 12 |
| ccunitfun | HES APC (hes\_ccare table, ccunitfun field) | 22 |
| classpat | HES APC (hes\_episodes table, classpat field) | 7 |
| disdest | HES APC (disdest fields in hes\_hospital and hes\_episodes) | 25 |
| dismeth | HES APC (dismeth fields in hes\_hospital and hes\_episodes) | 7 |
| epitype | HES APC (hes\_episodes table, epitype field) | 6 |
| firstreg | HES APC (hes\_episodes table, firstreg field) | 4 |
| gen\_ethnicity | HES APC (hes\_patient table, gen\_ethnicity field; hes\_episodes table, ethnos field) | 10 |
| icd10\* | HES APC (ICD field in hes\_diagnosis\_epi and hes\_diagnosis\_hosp; ICD\_PRIMARY fields in hes\_primary\_diag\_hosp)  ONS death (all cause fields – use with decimal aka Code)  COVID emergency codes used in UK (https://hscic.kahootz.com/gf2.ti/f/762498/96552069.1/PDF/-/COVID19\_NCCS\_01\_April\_2021.pdf):  March 2020:  U07.1: COVID-19 virus identified  U07.2: COVID-19 virus not identified (includes clinically-epidemiologically diagnosed, probable and suspected)  November 2020:  U07.3: Personal history of COVID-19  U07.4: Post COVID-19 condition  U07.5: Multisystem inflammatory syndrome associated with COVID-19  February 2021:  U07.6: Need for immunization against COVID-19  U07.7: COVID-19 vaccines causing adverse effects in therapeutic use | 17,872 |
| intmanig | HES APC (hes\_episodes table, intmanig field) | 7 |
| mainspef | HES APC (hes\_episodes table, mainspef field) | 93 |
| nhs\_indicator | ONS (nhs\_indicator field) | 3 |
| opcs4\* | HES APC (hes\_procedures\_epi table, OPCS field) | 11,251 |
| orgsup | HES APC (hes\_acp table, orgsup field) | 7 |
| tretspef | HES APC (hes\_ episodes table, tretspef field)  **NB: has 3 columns – 2 description columns for pre- and post- March 2004** | 177 |
| trustid | CHESS (trustid field) | 109 |
| unitbedconfig | HES APC (hes\_ccare table, unitbedconfig field) | 4 |

See ‘Linkage lookup datatypes’ spreadsheet for fields and their datatypes

\*Not uploaded to MySQL on Slade

3. Known quality issues and other notes

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Dataset** | **Linkage to CPRD Aurum data** | **Area coverage** | **Time coverage/Changes in recording over time** | **Other notes** |
| CHESS | Via NHS number only | 109 of 152 NHS Acute Trusts in England  Many London trusts didn’t report in early phases | Started ~06/04/2020 but retrospective reporting required from 01/01/2020; up to 29/09/2020 inclusive.  From looking at data: ‘labtestdate’ (which apparently is a ‘required field’ but is missing/invalid (01/01/2001) for 1,572 out of 13,343 and invalid ) range from 01/01/2020 to 28/09/2020. ‘Weekno’ (week of admission) has no missingness and ranges from 11-40.  Eligibility for SARS-CoV-2 testing has changed during pandemic (and CHESS is only lab-confirmed cases):  • 02-15/03/2020: only ICU/HDU patients with influenza like illness (ILI) / lower respiratory tract infections (LRTI) / pneumonia  • 15/03/2020-27/04/2020: all hospitalisations with ILI/LRTI/pneumonia  • 27/04/2020: all non-elective hospital admissions including with no COVID-19 symptoms | Three components:  • Aggregate COVID-19 related hospital and ICU/HDU admission - estimated 70-75% of known cases captured  • COVID-19-related ICD - estimated 20-30% of eligible admissions  • Subset of 24 NHS Trusts (‘Sentinel’) individually reported COVID-19 related hospital admission – not complete for all, and in total represents <20% of all England COVID-19 related hospital admissions  80% patid-casid pairs have duplicates where record updated  Some ‘-11’ ageyear values – removed when imported onto MySQL.  Some 01/01/2001 labtestdate values – removed when imported onto MySQL.  NB: added extra column for primary key. |
| HES | 8-step algorithm\*; only patients scoring >6 provided. Score in hes\_patient.match\_rank.  hes\_patient.gen\_hesid = ID of patient in HES data. Can be multiple rows per gen\_hesid with different patid (ID of patient in CPRD Aurum primary care data) if multiple patients match. n\_patid\_hes = number of patid matched with gen\_hesid. CPRD recommend removing if >20.  Patients linked to multiple gen\_hesid not provided. | All English NHS health care providers | April 1997 – Oct 2020 inclusive. Data up to March 2020 are final. Data for 2020/21 (currently April 2020 – Oct 2020) are provisional (likely to be incomplete).  Critical care data is in ‘Augmented Care’ up to 2007/2008. This has ‘known data quality issues’. Note that lookups for acploc, acpdisp, acpout, acpplan, acpspef, and orgsup do not contain all the values in the augmented care (ACP) table with epistart dates 1998-2000 (see ‘Linkage datatypes’ spreadsheet). We don’t have acpqind variable in data specification.  Firstreg field in hes\_episodes also contains entries from 1999-2006 with values not in the lookup table. | Ethnicity recorded at each episode in hes\_episodes.ethnos. hes\_patient.gen\_ethnicity = CPRD-derived using majority episode ethnicity from all set 19 HES data (APC, OP, A&E, PROMs and DID).  If episode spans end of March-start of April, appears as two separate episodes. If patient passed between consultants or organisations, new episode. Up to 20 diagnoses and 24 procedures per episode.  Other issues:  Some 15/10/1582 and 01/01/1600 dates.  • Episodes where admission < epistart  • Unfinished episodes  • Duplicate records which vary only by unique episode identifier (epikey).  NB: hes\_diagnosis\_hosp, hes\_primary\_diag\_hosp, hes\_acp, hes\_ccare: primary keys as described as CPRD data specifications do not define unique rows in the data, so have added extra column with auto-incrementing primary key. |
| Patient and practice IMD | Directly use patient or practice postcode so no linkage issues | England only (patient), or all UK (practice; we actually only have England and NI) | England: IMD 2015  NI: MDM (Multiple Deprivation Measure) 2017  so if area has rapidly changed, might not be correct for a different time. | England: based on 2011 LSOA boundaries; postcode-LSOA map used is from Nov 2019 NHS Postcode Directory (NHSPD)  NI: based on ?2011? SOA boundaries; not sure about postcode-SOA lookup used  Missing data if new/invalid postcode, or (patient only) postcode outside England.  Practice: Since January 5th 2015, GP practices in England are free to register new patients who live outside their practice boundary area  NB: 1 = LEAST DEPRIVED |
| ONS death | As per HES: 8-step algorithm\*; only patients scoring >6 provided. Score in match\_rank.  gen\_death\_id = ID of patient in death data. Can be multiple rows per gen\_death\_id with different patid (ID of patient in CPRD Aurum primary care data) if multiple patients match. n\_patid\_death = number of patid matched with gen\_death\_id. CPRD recommend removing if >20.  Patients linked to multiple gen\_hesid not provided. |  | January 1998 – November 2020.  Late registration means proportion of deaths captured is lower for the last year of the coverage period and likely to differ by age and cause of death. Especially last 2 weeks of data. | COVID-19 deaths: more conditions; could be due to better reporting or genuinely more co-morbidities. Non-COVID-19 deaths have more ‘ill-defined’ causes during pandemic.  Fewer suicides April-June 2020; unlikely to be real; due to delays in coroner service.  Recorded as ICD-9 prior to 2001.  Some codes not in general ICD-10 classification:  • No longer in use (e.g. I50.2, J84.2)  • Refer to place of accident (e.g. T58.2 [School, other institution and public administrative area], T71.1 [Residential Institution])  • U50.9: deaths involving adjourned inquests. Final ICD added once inquest complete.  • R97: ‘Cause Unknown’ and is processed differently to other R99 codes (Other ill-defined and unspecified cause of mortality).  Use date of registration of death where date of death missing (dod\_partial is NOT helpful – only present for 2 entries and just gives year which is the same as year of registration date). |
| SGSS | Via NHS number only.  n\_patid\_spec is how many patients linked to same specimen record. May be correct if patient moved practice (get new patid in CPRD). However, they recommend removing if >20. | Majority of tests in London at start of pandemic. Different locations started testing/reporting at different times. May be some geographical bias, particularly in early data. | 01/03/2020 – 29/09/2020 inclusive | Just positive tests from PCR swab testing (not antibodies). Only pillar 1 testing (diagnostic testing for those with a clinical need – usually in hospital, and health and care workers) done in NHS/PHE labs or forwarded to NHS. Not private testing or pillar 2 (community testing e.g. regional test sites, mobile testing units, satellite test centres and via home tests).  Some duplicates (mostly differing only by care home variable).  Some ‘-1’ in age field – removed on import to MySQL.  NB: added extra column for primary key (patid is meant to be primary key but are duplicates as above). |

\*

1. Exact NHS no, sex, DOB, postcode
2. Exact NHS no, sex, DOB
3. Exact NHS no, sex, postcode, partial DOB
4. Exact NHS no, sex, partial DOB
5. Exact NHS no, postcode
6. Exact sex, DOB, postcode (where NHS number does not contradict the match, the DOB is not 1st of January & the postcode not on the communal establishment list)
7. Exact sex, DOB, postcode (where NHS number does not contradict the match & the DOB is not 1st of January)
8. Exact NHS number