British Heart Foundation Data Science Centre
Led by Health Data Research UK
Kidney Research UK

FAQs: BHF Data Science Centre Kidney Catalyst Funding Call

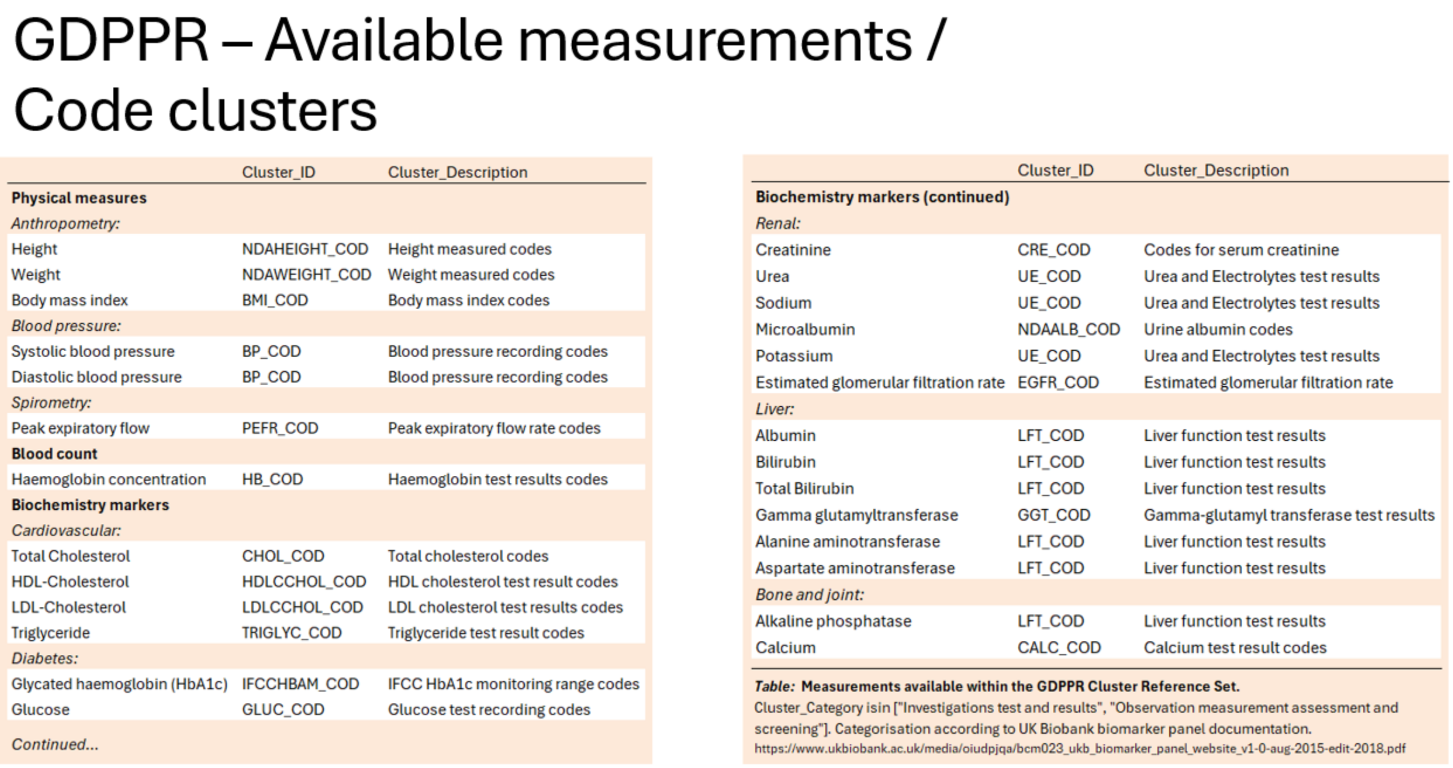
Data availability and feasibility

Q1: Are sociodemographic variables available in the NHS England SDE?

Yes, sociodemographic variables such as birth date, sex, ethnicity, and Lower Layer Super Output Area (LSOA) are available. LSOA can be mapped to regions and the Index of Multiple Deprivation (IMD). These variables can be sourced from various datasets, including primary care, secondary care, death records, and others. The curated assets provided by the BHF Data Science Centre includes regularly updated tables including Key Patient Characteristics table.

Q2: What is the completeness of primary care data in the NHS England SDE?

GPES Data for Pandemic Planning and Research (GDPPR) is based on a subset of SNOMED codes (~40,000 out of 900,000), primarily covering events related to diagnoses, biomarkers, and prescribed medications. The codes are categorised into code clusters, a CSV file containing code clusters can be requested by emailing bhfdsc@hdruk.ac.uk, and the image below summarises some of the most commonly used biomarker code clusters.



GDPPR – Available measurements / Code clusters

	Cluster_ID	Cluster_Description
Physical measures		
Anthropometry:		
Height	NDAHEIGHT_COD	Height measured codes
Weight	NDAWEIGHT_COD	Weight measured codes
Body mass index	BMI_COD	Body mass index codes
Blood pressure:		
Systolic blood pressure	BP_COD	Blood pressure recording codes
Diastolic blood pressure	BP_COD	Blood pressure recording codes
Spirometry:		
Peak expiratory flow	PEFR_COD	Peak expiratory flow rate codes
Blood count		
Haemoglobin concentration	HB_COD	Haemoglobin test results codes
Biochemistry markers		
Cardiovascular:		
Total Cholesterol	CHOL_COD	Total cholesterol codes
HDL-Cholesterol	HDLCCHOL_COD	HDL cholesterol test result codes
LDL-Cholesterol	LDLCCHOL_COD	LDL cholesterol test results codes
Triglyceride	TRIGLYC_COD	Triglyceride test result codes
Diabetes:		
Glycated haemoglobin (HbA1c)	IFCCHBAM_COD	IFCC HbA1c monitoring range codes
Glucose	GLUC_COD	Glucose test recording codes

Continued...

	Cluster_ID	Cluster_Description
Biochemistry markers (continued)		
Renal:		
Creatinine	CRE_COD	Codes for serum creatinine
Urea	UE_COD	Urea and Electrolytes test results
Sodium	UE_COD	Urea and Electrolytes test results
Microalbumin	NDAALB_COD	Urine albumin codes
Potassium	UE_COD	Urea and Electrolytes test results
Estimated glomerular filtration rate	EGFR_COD	Estimated glomerular filtration rate
Liver:		
Albumin	LFT_COD	Liver function test results
Bilirubin	LFT_COD	Liver function test results
Total Bilirubin	LFT_COD	Liver function test results
Gamma glutamyltransferase	GGT_COD	Gamma-glutamyl transferase test results
Alanine aminotransferase	LFT_COD	Liver function test results
Aspartate aminotransferase	LFT_COD	Liver function test results
Bone and joint:		
Alkaline phosphatase	LFT_COD	Liver function test results
Calcium	CALC_COD	Calcium test result codes

Table: **Measurements available within the GDPPR Cluster Reference Set.**
Cluster_Category isin ["Investigations test and results", "Observation measurement assessment and screening"]. Categorisation according to UK Biobank biomarker panel documentation.
https://www.ukbiobank.ac.uk/media/oiudpjqa/bcm023_ukb_biomarker_panel_website_v1-0-aug-2015-edit-2018.pdf

The completeness of these measurements depends also on the cohort and study period. GDPPR includes only patients who were actively registered at participating practices and were alive on or born after 1 November 2019. Records of patients who died before this date are available in other datasets, such as the ONS Civil Registration of Deaths or Hospital Episode Statistics (HES) but are not included in GDPPR. Additionally, certain code clusters, such as those for biomarkers, have time-based cut-offs, typically with a 2-year lookback period from the first reporting period end date. For example, records containing codes related to creatinine (CRE_COD cluster) have shown an upward trend starting in mid-2018 to 2019, with low numbers before these dates. The period after 1 November 2019 may provide a more comprehensive view of primary care measurements compared to earlier periods. Despite the mentioned limitations, GDPPR has been successfully utilised by the majority of projects within the consortium to date and serves as the main source for primary care diagnostics, prescriptions, and biomarker phenotyping.

For more details visit:

- GDPPR dataset insights
- NHS England's guide for GDPPR

Q3: What is the time span covered by the datasets in the NHS England SDE?

For details on the earliest record dates in each dataset and trends in data coverage, visit the Dataset Summary Dashboard and Dataset overviews and coverage. The most recent data extracts were delivered in December 2024, containing complete records up to approximately October 2024. By the time projects commence in May 2025, it is expected that data covering the entirety of 2024 will be available for both primary and secondary care datasets.

Q4: Which data sources can be used to extract COVID-19 infections and vaccinations in the NHS England SDE?

COVID-19 infection and hospitalisation events are extracted from several datasets, including GDPPR, HES, Pillar 2 (antigen) and Pillar 3 (antibody) test results, Second Generation Surveillance System (SGSS), COVID-19 Severe Acute Respiratory Infection (SARI)-Watch, and Intensive Care National Audit and Research Centre (ICNARC) data. COVID-19 vaccination events are also available. Intensive Therapy Unit (ITU) admissions can be sourced from HES Critical Care (HES CC) data, with COVID-related admissions available in SARI-Watch and ICNARC datasets.

Q5: Which datasets can be used to extract prescribed and dispensed Medications in the NHS England SDE?

There are three key sources for medication data:

- Prescribed medications in GDPPR which might be affected by the limitations in SNOMED codes or time-based cut-offs.
- The primary care dispensed medicines dataset is based on BNF codes. We recommend prioritising the utilisation of this dataset for medication phenotyping/trends.
- Electronic Prescribing and Medicines Administration (EPMA) dataset, which includes prescribed and administered medicine in hospitals (i.e., in HES datasets). However, EPMA only covers 15% of the medications in HES with variable data quality.

Q6: How can I identify comorbidities and other phenotypes within the NHS England SDE?

Clinical outcomes, phenotypes, and comorbidities (e.g., CKD, CVD) along with their associated event dates are primarily derived from primary and secondary care datasets. These events are identified using clinical code lists based on coding systems such as SNOMED, ICD-10, OPCS, and others. Code lists can be defined within your study or selected from sources such as the HDR UK Phenotype Library or OpenSAFELY's OpenCodelists. Further detailed phenotyping can be achieved using biomarker variables from primary care data, trajectory of events, or prescribed/dispensed medications.

Q7: Which data analysis environments and software packages are available in the NHS England SDE?

The analysis environment is hosted on a Spark cluster accessible via Databricks, supporting PySpark, Python, Spark SQL, and R Studio Server. R Studio Desktop, and STATA (upon request) are also available for non-distributed data analysis. Most common statistical and machine learning packages in Python and R are pre-installed. However, GPU access requires additional approval, may involve costs outside the Kidney Data Science Catalyst funding scope, and the provision of pre-trained models into the environment is restricted. For more information, please refer to NHS England Secure Data Environment (SDE) user guidance.

Other Useful links

- BHF Data Science Centre Dataset Summary Dashboard: Contains data dictionary, data coverage/trends, and overall data summary.

- BHF Data Science Centre Health Data Science documentation page: Contains information on datasets and curated assets

- Dataset overviews and coverage

- List of all datasets

- CVD-COVID-UK/COVID-IMPACT consortium:

- Consortium webpage

- Consortium GitHub
- Consortium protocol

- CVD-COVID-UK/COVID-IMPACT collection on Health Data Research Gateway

- NHS England Secure Data Environment (SDE) user guidance

- GPES Data for Pandemic Planning and Research (GDPPR):

- GDPPR dataset insights
- NHS England’s guide for GDPPR
- GDPPR SNOMED code clusters

Funding

Q8: Do PI costs count as eligible costs as part of this funding call?

No, PI costs are not eligible costs, unless the PI is the person conducting the analysis. This call will fund salary costs for staff working directly on the research project (e.g., research analysts, project management staff).

Q9: Do we need to budget for data access fees to the SDE?

No, these costs will be covered separately by the BHF Data Science Centre.