Supplementary File 1. International databases and their key features

A. Healthcare Cost and Utilization Project (HCUP) :
https://www.ahrq.gov/data/hcup/index.html

Overview: The HCUP contains the National Inpatient Sample (NIS), Nationwide Emergency Department Sample (NEDS), and Nationwide Readmissions Database (NRD). The NIS is the largest publicly available hospital inpatient care database in the United States.
ICD Codes: ICD-9 and ICD-10, depending on year
Cost: Access to a single year of NIS data ranges from $350 to $625 per year.

B. Academic, national, and hospital electronic health records (EHRs)
Overview: Research hospitals continuously acquire patient encounters and grow their EHR databases. Many of the top systems include Vanderbilt University Medical Center, Geisinger Health System, Kaiser Permanente, and the Cleveland Clinic, to name a few. Government-level databases include the Center for Medicare and Medicaid Services (https://www.cms.gov/Research-Statistics-Data-and-Systems/CMS-Information-Technology/AccesstoDataApplication/index.html)

ICD Codes: ICD-9 and ICD-10, depending on system and year. Many systems, such as at Vanderbilt University Medical Center, contain mixed codes and include a "research" version of codes (called "phecode", see https://phewascatalog.org/phecodes).

Cost: Access commonly requires a data use agreement and collaboration with a clinician or research professor at the institution. The National Patient-Centered Clinical Research Network (https://pcornet.org/clinical-research-network/) is a portal to the network of health systems and research opportunities. The costs vary: data use may be free if the researcher is affiliated with an academic institution; however, some institutions may require a fee for use or offer a fee-for-service model to work with researchers employed at their institutions.

C. UK Biobank - https://www.ukbiobank.ac.uk/
Overview: The UK Biobank is a research resource on more than 500,000 men and women from the UK population, aged 40–69 at the time of assessment, from 2007–2010.

ICD Codes: ICD-9 and ICD-10, depending on year

Access: The UK Biobank is available to all researchers across universities and commercial companies, regardless of geographic location, but after an approval process.

Researchers who are granted access to this resource will be required to publish their findings and return their results to the UK Biobank. The average time from application submission to data release is roughly 24 weeks.

Cost: A fixed charge of 250 Pounds for every research application. Upon approval, data access is 1500 Pounds. Additional costs will be required for access to bulk genetic data access and biological samples. Reduced fees are possible for researchers in low to low-middle income countries.

D. China Kadoorie Biobank - https://www.ckbiobank.org/site/

Overview: The China Kadoorie Biobank (CKB) is a research resource on 510,000 adults across China. The recruitment of this cohort occurred from 2004–2008. Since that time, about 98% of the still-alive population is being more closely monitored through linkage to established registries and health insurance databases.

ICD Codes: ICD-10

Access: The CKB is available to researchers upon approval. Researchers are defined as being at an academic institution, health service, or charitable research organization. The proposal must have clearly-defined research objectives and amenable to peer-review journal publication. The average time from application submission to data release is 16 weeks.

Cost: A fixed charge of 750 Pounds for the application process and preparation of the dataset is required for each successful data request.

E. Estonian Biobank (https://www.geenivaramu.ee/en/access-biobank)

Overview: The Estonian Biobank comprises more than 51,000 gene donors in Estonia, with 50,000–100,000 more individuals expected to be genotyped during 2018–2019. This biobank is also linked to national registries and hospital databases for phenotypic information, including procedures and medications.

ICD Codes: ICD-10

Access: According to the data application procedures, it is not clear if researchers need to be members of academic institutions. Even so, it is unclear which phenotype data include ICD-10 codes. Phenotype data that are available for public use may only be restricted to lifestyle and demographic data.

Cost: No information exists regarding the price; however, the applicant is expected to pay for data access, regardless of scientific or commercial purpose. Additionally, the results of the research project must be shared with the Estonian Genome Center.

F. Genes and Health - http://www.genesandhealth.org/

Overview: The Genes and Health dataset is an open access data resource for genetic and health information on 100,000 people of Pakistani and Bangladshi heritage in the UK. Sequencing data, including exome sequencing and SNP array genotyping, are linked to the UK National Health Service electronic health records.

ICD Codes: Phecodes (https://phewascatalog.org/phecodes) mapped from ICD-10 codes.

Access: Summary genetic variant and genotype counts are freely available. Requests for phenotype data, including procedure codes, require a formal application and approval process. Given the limited resources of this resource, research studies are prioritized if they include objectives around diabetes, cardiovascular disease, and mental health. Access is possible to any academic and life science industry research partners.

Cost: No costs are explicitly mentioned. This resource may be either free or cheaper than other resources. However, information on their website implies they have limited resources to handle large numbers of research and access requests.

G. Resources which require additional collaboration:

Finngen - https://www.finngen.fi/en/Forresearchers

- This resource is very new and is not yet open access.
- Only researchers who are part of the current set of consortium partners may be granted access to this resource. Thus, collaborations with consortium partners may be necessary (https://www.finngen.fi/en/node/16)

H. Resources without evidence of ICD Codes

Lifelines cohort study and biobank –
https://www.lifelines.nl/researcher/biobank-lifelines/data-and-samples

- Prior research using this resource inferred cardiovascular disease ICD-10 codes based on questionnaires; however, there is no evidence that codes themselves are stored in this biobank.

European Prospective Investigation into Cancer and Nutrition – (https://epic.iarc.fr/index.php)

- Any codes related to cancer and/or mortality
- Most of the data collection occurred before structured EHRs, and thus codes would be limited to ICD-9 if any were even collected