1.3 Healthcare Data resources and UK BioBank

Angela Wood and Fergus Imrie
amw79@medschl.cam.ac.uk

Aims

- Overview of major healthcare data resources
  - UK Biobank
  - Emerging Risk Factors Collaboration
  - NHS Digital Trusted Research Environment
  - INTERVAL
- Considerations and challenges in handling healthcare data resources
- Inspire new research ideas

Objectives

- Challenges of healthcare dataset (UK Biobank)
- Individual participant data meta-analysis (Emerging risk factors collaboration)
- Population-wide Electronic Health records (NHS Digital Trusted Research Environment)
- Large-scale multi-omics (INTERVAL)

Challenges of healthcare dataset

- Healthcare data from different sources;
  - Imaging
  - Text
  - Tabular (including omcis and patient health history)
  - Temporal….
- Source of such data: can impact the data and usability
  - EHR
  - Biobanks
  - Clinical trials
  - Medical studies
- UK BioBank
  - https://www.ukbiobank.ac.uk/
  - Over 0.5 mil volunteers in UK, 40 to 69 yo
  - Enrollment between 2006 to 2010
    - Invitation-based
  - Follow-up of up to 30 years after enrollment
  - Information from Biobanks: Heterogeneity of information
    - Questionnaires
      - Lifestyle, physical activity, habits, diet and social/professional status
      - **Test of memory** (source of reporting bias) –Text
    - Interview
      - Clinical history questions (diagnosis, symptoms, and tests)
      - Conducted by nurses (source of reporting bias)
    - Physical Measurement
      - Body composition, visual and auditory acuity, bp, heigh, weight etc.

            - imaging
        - Basic screenings/ tests
            - FEV, FVC, ultrasound bone densitometry
        - Fitness test
            - ECG — waveforms
        - Samples
            - Blood cell counts, blood and urine composition, DNA information
            - Omics test undertaken – Genetics
    - Features can be broadly categorized into:
        - Demographics – questionnaire
        - Physical measurements and body compositions – physical measurements
        - Clinical history – questionnaire, nurse interview
        - Symptoms
        - Diagnostic tests and biomarkers – Basic screening/tests, fitness test, samples
        - Physical activity - fitness test
        - Psychology - questionnaire
        - Diet and nutrition - questionnaire
        - Social and environment - questionnaire
    - Data collection
        - 22 Assessment centers
        - Allowing regional variation
            - Problem with harmonizing across regions
            - Problems with equipment and methods
        - Enrolled over four years between 2006 to 2010
            - Time: source of differences between samples
- Medical dataset vs ML datasets

Medical datasets	**ML datasets**
Example: BreCaHAD https://bmcresnotes.biomedcentral.com/articles/10.1186/s13104-019-4121-7	Example: ImageNet https://www.image-net.org/
Often relatively small	Can be very large
Dirty: different conditions, missing data, missing outcomes etc.	Clean
Multimodal (heterogenous)	Unimodal
Broad range of purposes: Make discoveries, test hypothesis, insurance “If we could do something” Render the results to be flawed/ impossible/	To test algorithms = often have performed preliminary evaluations

meaningless	

- Unique challenges in healthcare data
    - Multiple streams of measurement
    - Sparse, irregularly, and informatively sampled measurements
    - Multiple outcomes of interest
        - Various events of interests
        - Various morbidities (i.e. not just cats vs dogs)
    - True clinical states are sometime unobserved (e.g. onset of disease)
    - Many possible patterns (heterogenous phenotypes, comorbidities)
- Accessing healthcare data
    - Strict regulations due to valid concerns regarding privacy
    - Strong regulators (e.g. HIPAA and GDPR) not allowing direct share private data to ML community from data holders (e.g. hospitals)

Assessing cardiovascular risk using multiple studies: The Emerging Risk Factors Collaboration

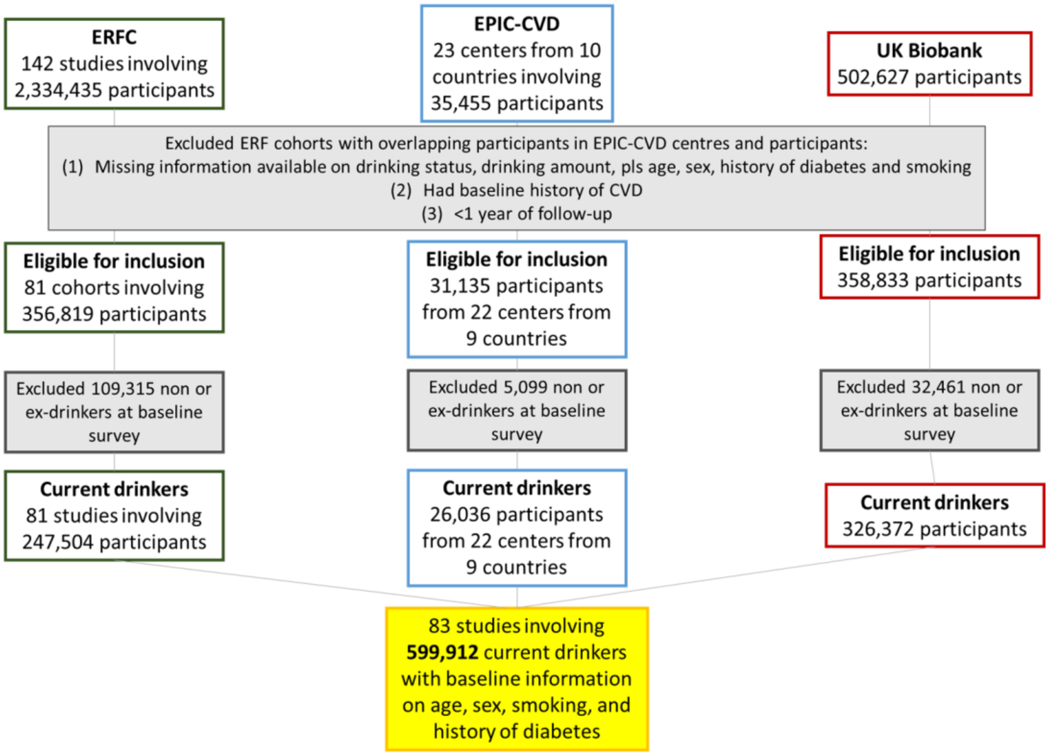
- Motivation
    - Enhance precision/ reduce overfitting/ increase generalizability
- Challenges
    - Harmonization of information
    - Combining analysis of different study designs
    - Accounting for measurement errors
    - Adjusting for known or potential confounders observed in a subset of studies
    - Assessing effect modification (within or between studies)
    - Dealing with missing data
- Emerging risk factors collaboration (ERFC)
    - https://www.phpc.cam.ac.uk/ceu/erfc/
    - Consortium of > 130 prospective studies from 30 countries
        - The Emerging Risk Factors Collaboration: analysis of individual data on lipid, inflammatory and other markers in over 1.1 million participants in 104 prospective studies of cardiovascular diseases
        - https://link.springer.com/article/10.1007/s10654-007-9165-7
    - Collated and harmonized individual-participant data (IPD) from ~2.5M participants
    - Aim: to study risk factors for cardiovascular disease and cause-specific mortality in greater detail:
        - Circulating lipid markers
            - Major lipids, apolipoproteins, and risk of vascular disease. JAMA, 2009
                - https://jamanetwork.com/journals/jama/fullarticle/184863
            - Lipoprotein(a) concentration and the risk of coronary heart disease, stroke, and nonvascular mortality. JAMA, 2009

    - https://jamanetwork.com/journals/jama/article-abstract/184315
  - Lipoprotein-associated phospholipase A(2) and risk of coronary disease, stroke, and mortality: collaborative analysis of 32 prospective studies. Lancet, 2010
    - https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(10)60319-4/fulltext
  - Lipid-related markers and cardiovascualar disease prediction. JAMA 2012
    - https://jamanetwork.com/journals/jama/fullarticle/1187927
- Inflammatory markers
  - C-reactive protein concentration and risk of coronary heart disease, stroke, and mortality: an individual participant meta-analysis. Lancet, 2010
    - https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(09)61717-7/fulltext
  - Interleukin-6 receptor pathways in coronary disease: a collaborative meta-analysis of 82 studies. Lancet 2012
    - https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(11)61931-4/fulltext
  - C-reactive protein, fibrinogen, and cardiovascular disease prediction. N Engl J Med 2012
    - https://www.nejm.org/doi/full/10.1056/nejmoa1107477
- Glycaemia markers
  - Surveillance intervals for small abdominal aortic aneurysms: a meta-anaylsis. JAMA. 2013
    - https://jamanetwork.com/journals/jama/fullarticle/1656254
- Adiposity markers
  - Adult height and the risk of cause-specific death and vascular morbidity in 1 million people: individual participant meta-analysis. Int J Epidemiol 2012
    - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3465767/
  - Global, regional, and national prevalence of overweight and obesity in children and adults during 1980–2013: a systematic analysis for the Global Burden of Disease Study 2013. The Lancet. 2014
    - https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(14)60460-8/fulltext
- Diabetes
  - Leucocyte Telomere Length and Risk of Type 2 Diabetes Mellitus: New Prospective Cohort Study and Literature-Based Meta-Analysis. Lustig AJ, editor. PLoS ONE
    - https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0112483

- Cardiometabolic multimorbidity
  - Association of cardiometabolic multimorbidity with mortality. JAMA. 2015
    - https://jamanetwork.com/journals/jama/fullarticle/2382980
- Alcohol
  - Risk thresholds for alcohol consumption: combined analysis of individual-participant data for 599 912 current drinkers in 83 prospective studies
    - https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(18)30134-X/fulltext
- Depression
  - Association between depressive symptoms and incident cardiovascular diseases. JAMA. 2020
    - https://jamanetwork.com/journals/jama/fullarticle/2774050#:~:text=In%20a%20pooled%20analysis%20of%20563%20255%20participants%20in%2022,magnitude%20of%20associations%20was%20modest.

- Aetiological hypothesis and risk prediction assessment
- Methodological developments occurring in parallel as necessary

- Exemplar: Risk thresholds for alcohol consumption
  - Rationale:
    - Low-risk limits recommended for alcohol consumption vary substantially across different national guidelines
  - Aim:
    - To define thresholds associated with lowest risk for all-cause mortality and



  - No overlapping participants across the three groups
  - No missing data
  - People with no existing CVD
- Harmonization of information across studies
  - Months taken to harmonizing
  - Checking with study coordinators for agreement

Methods to record alcohol consumption	**Different types of alcohol**	**Various recoding formats**
Self-administered Interview-led questionnaires Food frequency questionnaires Dietary recall surveys	Beer, wine, cider, spirits/liquor, alcopops, long drink, fortified wine, liqueur, sake, shochu, tharra, aperitif/digestif	Amount in a given period Frequency of drinks in a given period Categories for amount of frequency
↓		
Harmonised and cross-referenced into the following variables: Amount, status, duration, stop age, start age, years stopped, usage frequency and Categorised as “never”, “never/ex”, “ex”, “ex/current”, and “current” drinkers		
↓		
UK standard scale of grams/ week (1 unit = 8 grams of ethanol)		

- ERFC covering the highest recruitment amongst ERFC, EPIC-CVD, UK Biobank:
- Consumption divided into 4 categories:
    - Between 7 to 10% include those drinking more than recommended
    - Majority drinking within guidelines
- Smokers
- UK biobank more healthy than others

- Individual participant data meta-analysis strategy
    - Prospective studies: cohort (78)/ nested case-control (4)/ case-cohort(1)

2-Stage analysis strategy
Stage 1: Estimate study-specific risk ratios Cox model/ (un)conditional logistic model/ Prentice-Weighted Cox model Stratified by sex, centre Adjusted for age, smoking and diabetes
↓
Stage 2: Pool estimates by random-effects meta-analysis

- Accounting for measurement error in reported drinking
    - Publications of:
        - **Measurement error as an explanation for the alcohol harm paradox: analysis of eight cohort studies**
            - https://academic.oup.com/ije/article/49/6/1836/5913111?login=false
        - AJE 2013 (?)
    - Measurement error/ within-person variability in exposure/confounders
        - Biased associations in analysis using only single measurements
        - Often quantified by regression dilution ratio (RDR)
    - To correct bias, we estimated long-term “usual” alcohol consumption
        - Multi-level regression calibration
        - 152,640 serial assessments in 71,011 individuals from 37 studies
        - Regress re-survey measures (or lifetime alcohol consumption available in EPIC-CVD) on baseline alcohol consumption, adjusted for relevant covariates with random effects for study and re-survey
        - Estimate conditional expectations of usual levels and include in regression models
- Study and re-survey regression dilution ratios
    - Average regression dilution ratio around 0.54
    - Enhanced precision to assess less common outcomes
- Key points
    - Described:
        - Challenges arising from disparate study designs

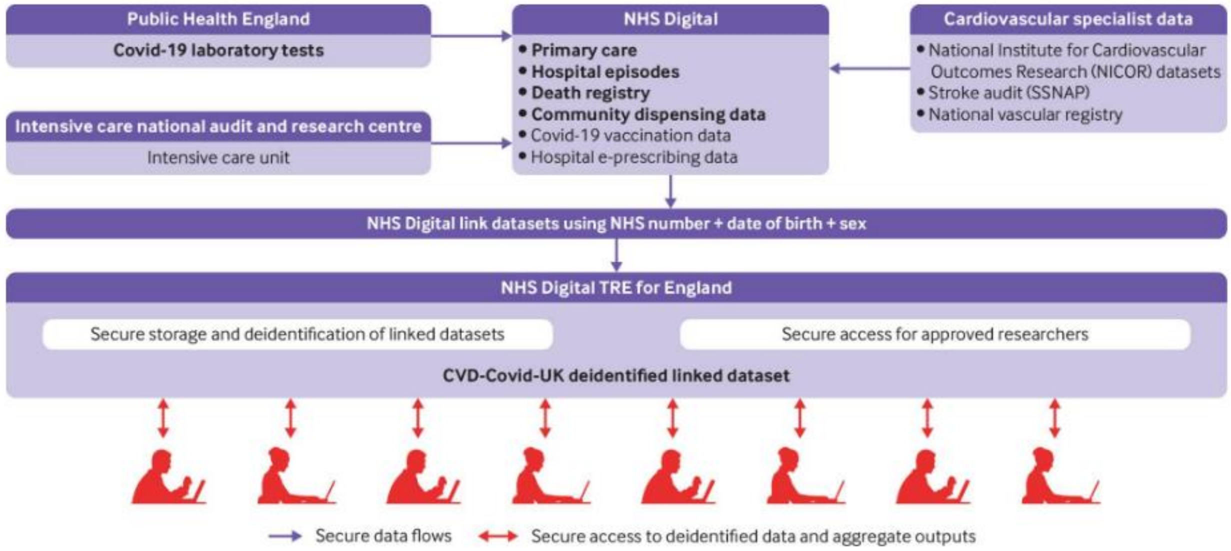
        - Data available for enhanced precision
        - Handling measurement error using repeated measures in sub-samples
        - Computational limitations individual participant data meta-analysis with large datasets
        - Translation of findings in to clinically useful interpretations
    - Stat program:
        - http://www.phpc.cam.ac.uk/ceu/erfc/programs//

Part 3: Population-wide Electronic Health Records Exemplar: NHS Digital Trusted Research Environment

- BHF data science centre:
    - Interrogating linked health data from >60 million people to better understand CVD
    - https://www.hdruk.ac.uk/helping-with-health-data/bhf-data-science-centre/
- UK-wide network of national Trusted Research Environment (TREs)

[Diagram]

- NHS Digital’s new Trusted Research Environment
    - **Linked electronic health records for research on a nationwide cohort of more than 54 million people in England: data resource**
    - https://www.bmj.com/content/373/bmj.n826

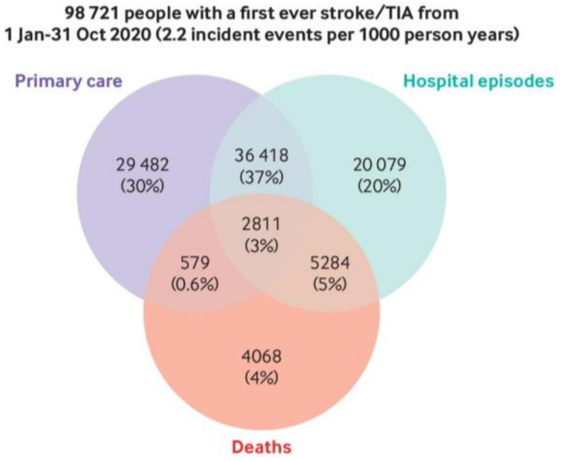

Public Health England
Covid-19 laboratory tests
Intensive care national audit and research centre
Intensive care unit
NHS Digital
• Primary care
• Hospital episodes
• Death registry
• Community dispensing data
• Covid-19 vaccination data
• Hospital e-prescribing data
Cardiovascular specialist data
• National Institute for Cardiovascular Outcomes Research (NICOR) datasets
• Stroke audit (SSNAP)
• National vascular registry
NHS Digital link datasets using NHS number + date of birth + sex
NHS Digital TRE for England
Secure storage and deidentification of linked datasets
Secure access for approved researchers
CVD-Covid-UK deidentified linked dataset
Secure data flows
Secure access to deidentified data and aggregate outputs

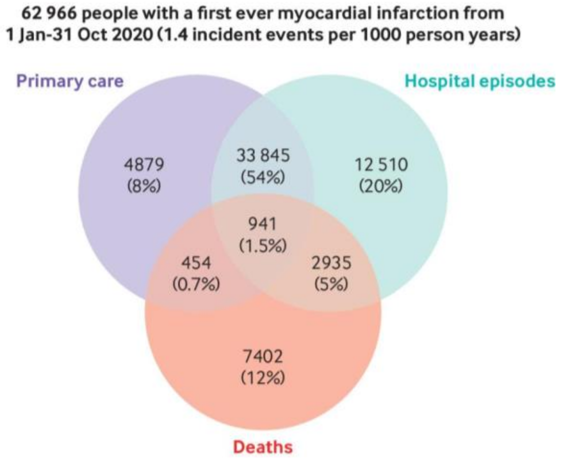
    - Enables whole population research:
        - >55 million people alive on 1st Jan 2020 -> >95% of population
        - Statistically powerful
        - Comprehensive information on characteristics and health outcomes
        - Includes all age groups, ethnicities, geographic locations, socioeconomic, health and personal characteristics
        - All datasets updated monthly

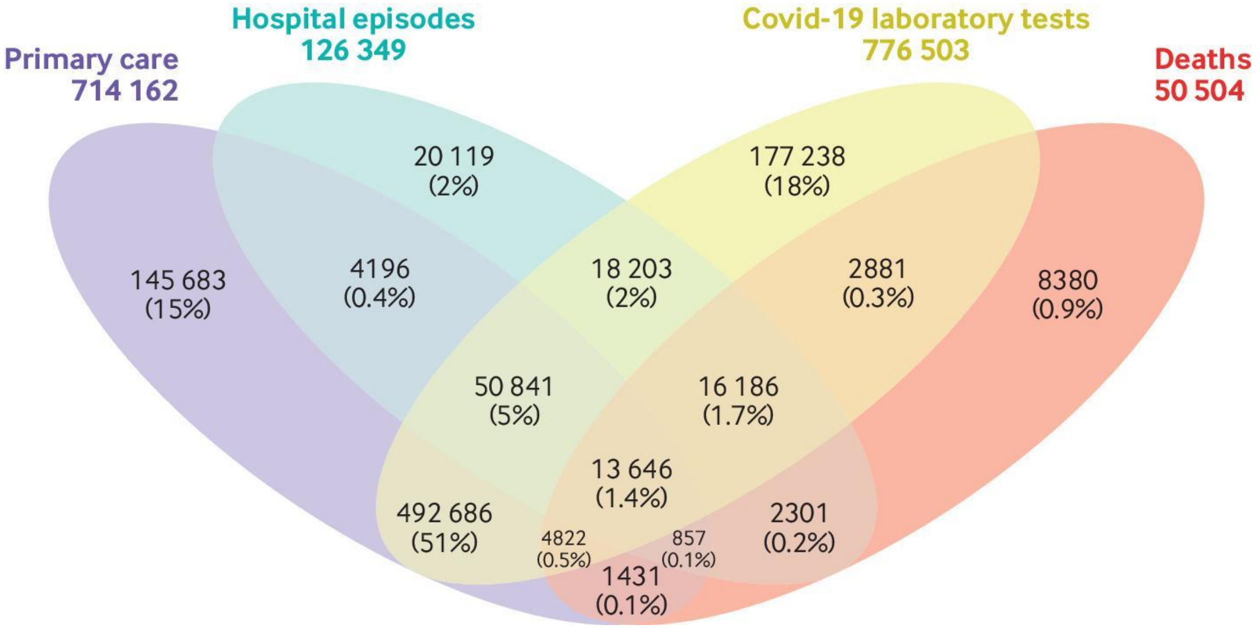
- CVD-COVID-UK/COVID-IMPACT consortium: enabling across to UK population linked health data
    - https://www.hdruk.ac.uk/projects/cvd-covid-uk-project/

    - bhfdsc@hdruk.ac.uk
    - Population coverage:
        - England (NHSD): >55M
        - Scotland (Safe Haven): >5M
        - Wales (SAIL): >3M
    - Consortium
        - >250 members
        - >40 NHS and academic organisations
    - Analysts:
        - >70 analysts in the TREs
        - TRE(s) accessible by approved researchers
    - Project:
        - 30+ approved projects, more coming
        - Protocols/ algorithms in GitHub
        - Published outputs…
- CVD-COVID-UWCOVID-IMPACT projects:
    - Methods
        - Data management and analysis methods
        - High-throughput phenotyping approaches
        - Improving methods to minimise bias in ethnicity data
    - Medicines
        - Effects of ACE inhibitors & ARBs on COVID-19
        - Impact of COVID-19 on managing BP and lipids
        - Assessing COVID-19 impact through medicines
        - Antipsychotic prescribing during the pandemic and
        - cardiovascular risk in patients with dementia
        - Evaluation of antithrombotic use on COVID-19 outcomes
        - Repurposing medicines to prevent COVID-19
    - Others
        - COVID-19 infection, vaccination and vascular risk
        - Direct and indirect effects of COVID-19 in people with
        - cardiovascular disease
        - COVID and cardiovascular disease risk prediction
        - Impact of COVID-19 on Congenital Heart Disease (CHD)
        - patients undergoing cardiac surgery
        - Influence of multi-morbidity on outcomes of COVID-19
        - Predicting severe COVID-19 in people with rare diseases Genomics of multi-morbidity and susceptibility to COVID-19
        - Longer-term effects of COVID-19 in non-hospitalised people
        - Evaluating how palliative and end of life care teams have responded to COVID-19
        - Coronary revascularisation and outcomes before and after the COVID-19 pandemic

        - Children admitted to hospital with COVID-19 — risk factors, risk groups and NHS care utilization
        - Understanding the increased risk of severe COVID-19 in people with intellectual & developmental disabilities
        - Risks of cardiovascular disease in people with COVID-19 and pre-existing respiratory disease
        - Impact of COVID-19 on eye disease
        - Impact of COVID-19 on heart failure
        - Impact of COVID-19 on people with diabetes
- Novel and key benefits of population-wide data for research:
    - Scale and depth
    - Generalisability
    - Public health policies
- Challenges of using population-wide data for research
    - ~65 million people alive on 1st Jan 2020, registered with an NHS general practitioners in England, Scotland and Wales
    - Consistent data curation pipelines and quality checks
    - Defining population of interest
    - Phenotyping diseases and conditions
    - Study designs
    - Analytical approaches and interpretation
    - Distributing analytical pipelines between systems
    - Computationally efficient analyses
    - Open access Protocols/algorithms in GitHub
- Linking data from different healthcare settings to ascertain incident cardiovascular events
    - https://www.bmj.com/content/373/bmj.n826





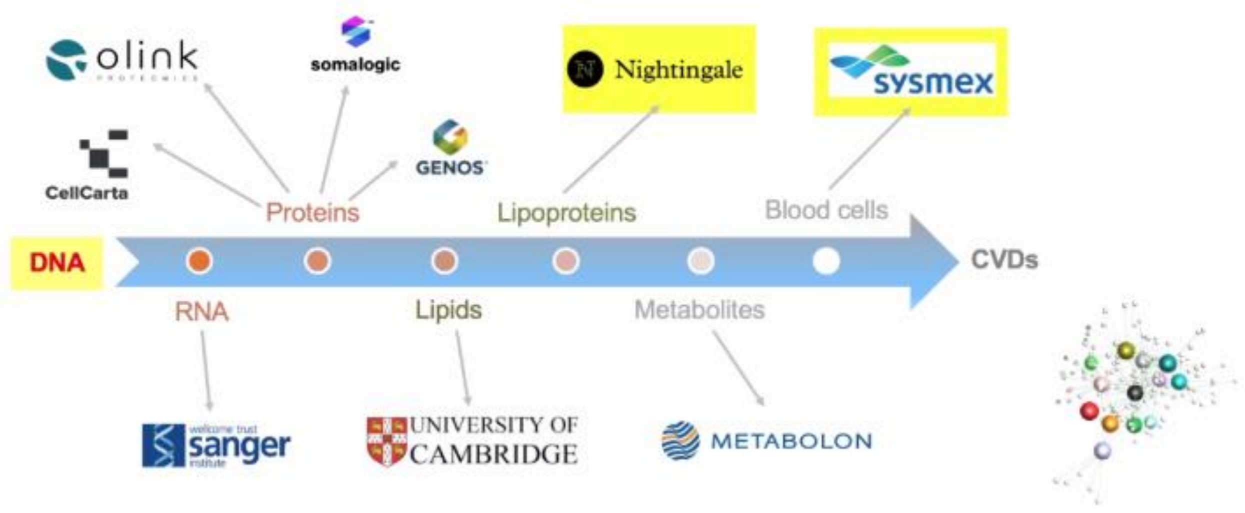


        - 3.5 million people with COVID-19 by mid Feb 2021:
            - 3.1 million with a positive test
            - 2.4 million diagnosed in primary care
            - 364,000 hospitalized
- Higher risks of major vascular and arterial disease following hospitalised COVID-19
    - Non-hospitalised also remain higher risk (double) of vascular events of up to 2 years
    - Knight et al., 2022 Circulation (?)
- Key points:
    - During the COVID-19 pandemic, it has become possible to conduct research of clinical and policy relevance at UK population-wide scale using rich, diverse linked national health data
    - A critical enabler has been the establishment in 2020-21 of NHS Digital's new trusted research environment for England
    - This has enabled many, well-powered studies and insights
    - All analyses require essential data curation/wrangling tasks (at least 80% of research project time)
    - Transparency in all stages of analyses — reproducible research
    - National coordination, a team science approach and public support essential
    - Hot press: CODE-EHR best practice framework for the use of structured electronic healthcare records in clinical research, BMJ 2022

Part 4: Large-scale multi-omics: Exemplar “INTERVAL”

- INTERVAL trial:
    - https://www.intervalstudy.org.uk/
    - Randomized trial assessing how often blood donors can safely give whole blood

    - In addition to questions that can be answered by the randomised trial, created a bioresource of 50,000 trial participants to address other epidemiological questions, particularly those relating to genetics
- Domains of the expressed genome: study at scale with Interval bioresource:
    - https://www.phpc.cam.ac.uk/ceu/interval-bioresource/
    - All 50,000 participants:
        - Basic lifestyle and self-reported health information using web-based questionnaire



Example of studies from INTERVAL:

- The allelic landscape of human blood cell trait variation and links to complex disease. Cell 2016
    - https://www.cell.com/cell/fulltext/S0092-8674(16)31463-5?_returnURL=https%3A%2F%2Flinkinghub.elsevier.com%2Fretrieve%2Fpii%2FS0092867416314635%3Fshowall%3Dtrue
    - Astle et al.
- Genomic atlas of the human plasma proteome. Nature 2018
    - https://www.nature.com/articles/s41586-018-0175-2
    - Sun et al.
- Genomic and drug target evaluation of 90 cardiovascular proteins in 30,931 individuals. Nature Metabolism 2020
    - https://www.nature.com/articles/s42255-020-00287-2
    - Folkersen et al.
- Whole-exome sequencing identifies rare genetic variants associated with human plasma metabolites. AJHG 2022
    - https://www.cell.com/ajhg/fulltext/S0002-9297(22)00157-4
    - Bomba et al.
- Machine learning optimized polygenic scores for blood cell traits identify sex-specific trajectories and genetic correlations with disease. Cell Genomics 2022
    - https://www.cell.com/cell-genomics/pdf/S2666-979X(21)00107-5.pdf
    - Xu et al.

Final remarks:

- Large-scale data resources more widely available and accessible
  - With restrictions
- Unique challenges:
  - Esp. from routine health data (those not collected for research)
- Data generally need to be pre-processing steps and data quality checks
  - Project specific
- Get in touch:
  - amw79@medschl.cam.ac.uk