Research Projects Using Data from the Blood Donors Studies BioResource			
Last updated: 13.04.23			
*** = project has been approved, but data not yet accessed.			
Project Title	**Principal Applicant**	**Principal Applicant Institution**	**Lay Summary**
NMR metabolites: Correlates and associations with incident disease outcomes in the INTERVAL study	Dr Stephen Kaptoge	University of Cambridge	Metabolomics profiling and analysis of blood samples has been successfully applied to investigate a variety of diseases. Many metabolites have been associated with various disease phenotypes and outcomes, including obesity, diabetes, hypertension, cardiovascular diseases (CVD), cancer, mortality, etc. However, some key features of the metabolite associations have not been explored in detail (e.g. dose-response relationships) by the available observational studies or independently replicated in studies of healthy participants. Causality also remains uncertain for many metabolites. Each molecular phenotype can, in principle be mapped onto the genome using advanced technologies. On the horizontal axis from the DNA to phenotype, the metabolome is at closest proximity to the phenome, and thus a detailed study of metabolite correlates and disease associations may potentially yield novel insights on disease aetiology.
Association of genomic scores with multi-omics protein, lipid, metabolite and cell traits profiles	Professor Mike Inouye	University of Cambridge	Genetic scores constructed to predict lifetime risk of cardiovascular and other diseases may also reveal new insights into the underlying biological causes of disease development and progression, which may lead to improved treatment strategies. Here, we propose to use the INTERVAL cohort of healthy blood donors to identify proteins and other molecules that differ in people with increased genetic predisposition to heart diseases and others, and evaluate whether these differences translate into increased disease risk over the intervening period since the INTERVAL study collection. This may allow us to identify new proteins and molecules that play a causal role in disease development, which may in the future inform development of new prevention and treatment therapies.

Addressing COVID-19 using INTERVAL/COMPARE	Professor Adam Butterworth	University of Cambridge	COVID-19, caused by the SARS-CoV-2 virus, emerged as a serious public health threat in January 2020 and reached pandemic status in March 2020. For health departments and national governments worldwide COVID-19 has rapidly become the top priority, with thousands of researchers repurposing their resources and capabilities. In this pandemic environment, speed and coordination are paramount. Academic and hospital departments are responding accordingly, with many creating more flexible research frameworks and teams which can quickly focus and refocus on time-critical tasks. In leveraging the data within our Blood Donors Studies BioResource, the aims of this proposal are to enable our researchers to rapidly and efficiently: (1) understand the susceptibility, pathogenesis, outcomes and complications associated with COVID-19, (2) identify and characterise causal and non-causal (predictive) biomarkers for COVID-19 and (3) create tools and pipelines for rapid COVID-19 analyses which can be updated in real-time.
EpiCov	Professor Willem Ouwehand	University of Cambridge	COVID-19 is an infectious disease caused by the coronavirus (SARS-CoV-2), related to those which caused the previous outbreaks of Severe Acute Respiratory Syndrome (SARS) and Middle East Respiratory Syndrome (MERS). Within this research project we aim to answer two questions: (1) given a set of clinical variables, measured on a patient prior to admission, what is the probability that the patient is infected with COVID-19? and (2) given a set of clinical variables measured on an inpatient, and in the context of given treatments, what is the expected course of the patient's disease? We have received permission to create a patient-level database from our electronic health record (EHR) systems in order to answer these questions and to inform the acute management of patients during the pandemic. Data from the Blood Donors Studies BioResource will increase the power of our analyses (by increasing the number of participants) and diversity of data available for analyses.
Blood traits in predisposition to colorectal cancer	Professor Ian Tomlinson	University of Edinburgh	The aim of our project is to examine the possibility that blood cell traits predispose to colorectal cancer (CRC). We are interested in determining whether specific blood cell traits may heighten the risk of developing CRC and are therefore interested in both blood cell and DNA genotyping data for patients. However, covariation between blood traits means that accurately determining causality is difficult. We hope to use individual-level single nucleotide polymorphism (SNP; “snips”) from patients in the INTERVAL study to build associations with blood traits and develop models that may be used to predict levels of blood traits in our CRC Genome Wide Association Study (GWAS) dataset. This will allow for a greater understanding as to the role of blood trait variation in CRC predisposition and, if there is a role, which traits specifically are most relevant.
INTERVAL dataset: QC of Olink protein biomarkers	Professor Adam Butterworth	University of Cambridge	We require an up-to-date dataset in order to quality control (QC) new protein data.

Regulation of gene expression in health and disease	Dr Emma Davenport	Wellcome Sanger Institute	Our group uses functional genomics datasets to understand the role of genetics in patient heterogeneity with a particular focus on infectious and autoimmune diseases, including sepsis and systemic lupus erythematosus (SLE). We routinely apply functional genomics strategies such as expression quantitative trait locus (eQTL) mapping across the genome to understand the role of regulatory genetic variants in disease susceptibility and outcome. We have recently generated genotyping and sequencing of RNA datasets for hundreds of patients with suspected infection presenting to the emergency room and those that develop sepsis and are admitted to intensive care. We are currently mapping eQTL genome-wide in these datasets. We propose to firstly use the INTERVAL cohort to improve our eQTL mapping approaches. We will then use the INTERVAL cohort to represent the regulation of gene expression under healthy conditions and compare the signals we observe in our disease cohorts to determine if the regulation of particular genes or pathways is modulated by disease.
Antibody testing for SARS-CoV-2 with Public Health England	Professor Emanuele Di Angelantonio	University of Cambridge / NHS Blood and Transplant	We possess blood samples from the COMPARE study (2016-17) which are negative for SARS-CoV-2. We can help Public Health England (PHE) in the development of serodiagnostic tests indicative of past SARS-CoV-2 diagnosis by designing a study using the COMPARE samples to evaluate the false positive rate and potential determinants. We have therefore sent COMPARE samples to PHE and will need a dataset from the Blood Donors Studies BioResource to help analyse results from the sample assays.
TRACK-COVID feasibility pilot	Professor Emanuele Di Angelantonio	University of Cambridge / NHS Blood and Transplant	We require a dataset to assess results obtained from a few samples from this live study (TRACK-COVID).
Contribution of polygenic scores to haematological disease	Professor Nicole Soranzo	Wellcome Sanger Institute	Extreme blood count values characterise many haematological cancers such as polycythemia vera (PV) and essential thrombocythemia (ET). It is becoming clear that both germline variants (variations in DNA sequence transmitted from parent to offspring) and somatic variants (genetic material that can't be passed to future generations) act in modulating blood cell traits and in contributing to disease. We propose to comprehensively assess the germline contribution to haematological disease by quantifying the predictive power of blood counts genetic scores (GS) to disease status and severity in a cohort of 2,035 samples with myeloproliferative neoplasms (MPN) from a UK collection. As part of the project, we had initiated an international collaboration to run a case-control analysis for MPN and, as such, we would need to identify a set of matched controls from the INTERVAL genotyped samples and perform a genome-wide association study (GWAS) based on HRC-imputed genotype data.

Phenotypic effects of somatic clonal expansions in healthy individuals' blood cells	Dr Dragana Vuckovic	Wellcome Sanger Institute / Imperial College London	Somatic variation occurs randomly in the DNA, beginning in prenatal development. Most somatic mutations appear and disappear with the normal cell cycle; however, some confer selective advantage leading to cellular expansion and to the formation of a somatic clone (Ju et al. 2017). Detection of somatic clones traditionally required very deep DNA sequencing (e.g. 500x coverage), limiting the analysis to small numbers of samples. More recently it has been shown that coding somatic mutation can be detected reliably from RNA sequencing, which comes at a very deep coverage as a standard. This pipeline was applied to the GTEx resource and provided a comprehensive catalogue of somatic variation across healthy tissues proving that somatic mutations are acquired as part of the normal ageing process (Yizhak et al. 2019). Genome sequencing has greatly improved our understanding of somatic mutation in cancer (Martincorena and Campbell 2015). Carriers of cancer-related mutations develop cancer only in a small fraction of cases and in the presence of multiple drivers and other exposure factors. The functional role of somatic mutations in healthy tissue is largely unknown. Some evidence exists that acquired mutations in hematopoietic stem cells (HSCs) and subsequent clonal haematopoiesis might increase cardiovascular risk and drive atherosclerosis (Fuster and Walsh 2018). The haematopoietic system is of particular interest as HSCs are highly proliferative and it has been estimated that by age 50 each individual accumulates on average 5 coding mutations in each HSC (Welch et al. 2012). Here we propose to exploit the somatic pipeline for RNA sequencing to characterise somatic mutations in ~5,000 INTERVAL participants. We then propose to test for associations between the somatic variation and the rich collection of blood phenotypes available for the same individuals, with the aim to characterize any observable phenotypic effects.
HDR-UK Multiomics Cohorts Consortium: linking multiomics data to health outcomes	Professor Adam Butterworth	University of Cambridge	Health Data Research UK have awarded funding for a National Implementation Project on Multiomics, led by Dr Butterworth. The project aims to capitalise on the UK's strengths in population cohorts, multiomics and genomics to identify novel causal pathways for disease. INTERVAL is the key cohort led by the University of Cambridge that will contribute to this national initiative, which includes several other institutions and cohorts from across the UK. Initially we will look at associations of molecular phenotypes with health outcomes, which we will then integrate with genetic data to understand whether the associations reflect cause-and-effect relationships. We will also look at comparing findings across different measurement methods using the several different platforms used to measure proteins and metabolites/lipids in INTERVAL.

***Anti-interferon autoantibodies in severe COVID-19	Professor Ken Smith	University of Cambridge	Some patients admitted to hospital with severe COVID-19 symptoms have been shown to have anti-cytokine antibodies (ACAs) present in their serum prior to COVID-19 infection. We propose to test for the presence of ACAs in pre-pandemic serum samples from INTERVAL participants admitted to hospital with COVID-19 infection, with the aim of better understanding whether or not patients who have anti-interferon autoantibodies and severe COVID-19 had pre-existing autoantibodies, or whether these developed over the course of the disease.
The genetics of inflammatory disease susceptibility, progression and drug response	Dr Carl Anderson	Wellcome Sanger Institute	We are using the INTERVAL genome and exome sequence data as populations controls for our genome and exome sequences from tens of thousands of inflammatory bowel disease patients. We anticipate that this will enable us to identify new regions of the genome significantly influencing both susceptibility to and progression of inflammatory bowel disease. We are also using the data to test new statistical and computation methods, including scaling of variant call across hundreds of thousands of individuals, genetic risk score estimation, and rare variant association testing. Moving forward, we would like to use GWAS, WES and WGS from INTERVAL, STRIDES and COMPARE as controls for our IBD studies (including, but not limited to those involving the IBD BioResource).
Exploring the genetic architecture of blood cell granularity	Professor Andrew Mumford	University of Bristol	Platelets are small blood cells whose main function is to assist blood clotting but also assist in regulating inflammation, immunity and cell growth. Platelets perform these functions by releasing tiny granules that contain a complex mixture of biological substances. It is already known that genetic disorders in which platelet granules are reduced result in abnormal bleeding. Platelet granules also contribute to a range of common disorders, including cardiovascular disease, the leading worldwide cause of death. The research objective is to identify new genes that are important for the formation of platelet granules. My approach builds on recent technological advances in genomics, particularly research by my collaborators in which small differences in the size and number of blood cells in a group of 170,000 people were linked to several hundred naturally occurring changes in DNA (genetic variants). This research has also now shown that differences between people in the number of granules in blood cells are also linked to genetic variants, and we are exploring these variants first computationally then experimentally in the laboratory.

Host genetics of COVID-19 phenotypes – collaboration with the COVID-19 Host Genetics Initiative	Professor Adam Butterworth	University of Cambridge	Researchers around the world are interested to know why people seem to respond differently when they get infected with SARS-CoV2, the coronavirus that causes the COVID-19 disease. For example, why do some people get only mild symptoms, whereas others have to go to hospital, and can even end up in intensive care or die? An international collaboration ("the COVID-19 Host Genetics Initiative") has formed to try to combine results from studies in many countries that all aim to answer the question: "How does someone's DNA affect their response to being infected with SARS-CoV2?". We propose to use data from the Blood Donor Studies to help answer this question by conducting analyses of participants' DNA taken from their blood samples and linking this with data on (a) who has tested positive for SARS-CoV2, (b) who was admitted to hospital due to COVID-19, (c) who was admitted to intensive care due to COVID-19, and (d) who died due to COVID-19.
Restless Legs Syndrome	Professor David Roberts	University of Oxford / NHS Blood and Transplant	Restless legs syndrome (RLS) is defined by unpleasant sensations in limbs while at rest and relieved by movement. RLS is linked to sleep disturbance and abnormal movement of legs at night. These symptoms are associated not only with inherited traits (your genes), but also with low iron stores. However, the severity, duration and persistence of these potentially serious symptoms in blood donors have not been described. Here, we will select donors with different severities of RLS symptoms. We will examine these donors with RLS to define objective measurements of periodic movement of legs during sleep. We will determine: (1) the association of RLS with abnormal movement of legs at night (PLMS), (2) the association of RLS severity and PLMS with sleep disturbance and fatigue and (3) the contribution of low iron stores and genetic factors to these symptoms. Access to the variety of data in the BioResource will help us understand the need for interventions and provide a foundation for trials to reduce RLS symptoms in people that donate blood.
The role of clonal haematopoiesis in immune-mediated inflammatory diseases	Dr Carl Anderson	Wellcome Sanger Institute	The overall objective of this study is to test the hypothesis that mutations acquired with age are associated with immune mediated inflammatory diseases (IMID), such as Inflammatory Bowel Disease (IBD). This is set to be achieved by detecting mutations via deep targeted sequencing and comparing their prevalence in blood samples from IBD patients with late disease onset with age-matched healthy controls from the INTERVAL study.
***Phasing sequenced samples from PGP-UK data set	Professor Richard Durbin	University of Cambridge	We sequenced the genomes of ~90 people from the Personal Genome Project (PGP)-UK cohort, around 5 years ago at the Sanger Institute, and now wish to use these to test and demonstrate new methods to study genetic history and population structure. To do this we need to first separate their maternal and paternal sequences, which requires a large reference panel of comparator genomes. Because the INTERVAL genome sequences were processed the same way, they would be an ideal reference panel.

Donor characteristics associated with slow or rapid recovery of pre-donation haemoglobin amongst repeat blood donors	Dr Lois Kim	University of Cambridge	In the UK, donors current have their haemoglobin level tested on arrival at blood donation centres; those not meeting the threshold (12.5g/dl for women and 13/5g/dl for men) are not permitted to donate and are deferred. Since many donors are repeat donors who return regularly, this is an important check to ensure that their haemoglobin levels have recovered to a safe level between donations. However, it may also be of interest to identify characteristics of returning donors in whom haemoglobin has not recovered to the level observed at the previous visit. This would indicate a slower recovery process, which may warrant longer inter-donation intervals to avert future deferral. Conversely, this would also help identify those with a rapid recovery process in whom shorter inter-donation intervals may be appropriate.
***GWAS relating to high titre anti-A/B status of blood donors	Dr Rebecca Cardigan	University of Cambridge / NHS Blood and Transplant	Clinical guidelines recommend that components that are rich in plasma are ABO blood group identical between donor and patient to reduce the risk of a haemolytic transfusion reaction occurring. When ABO identical units are not available, components are selected that have low levels (titres) of anti-A and B to reduce the risk of reactions. To enable this NHSBT screens all donations for anti-A/B antibodies and classifies donations into “high” or “low” titre. Preliminary analysis suggests that most donors are consistently high or low titre over time, with only a small fraction of donors (<5%) switching groups on a regular basis. This suggests that there could be a genetic component involved in determination of an individual’s anti-A/-B titre levels. Using the genetic data available on INTERVAL, STRIDES and COMPARE donors in combination with the routine anti-A/-B titre data available on NHSBT electronic donor record, we aim to conduct a case control GWAS study to investigate the genetic basis of anti-A/-B titre in donors.
INTERVAL dataset: QC of Cambridge Protein Arrays protein biomarkers	Professor Adam Butterworth	University of Cambridge	We require an up-to-date dataset in order to quality control (QC) new protein data.
Prioritisation of cardiovascular disease genetic risk-loci for potential recall-by-genotype studies	Dr David Stacey	University of Cambridge	Cardiovascular disease (CVD) risk is impacted by a combination of genetic and environmental factors. It is hoped that a better understanding of the genetic factors involved will inform drug target selection and the development of new drug treatments. With this project, we aim to identify specific genetic variants affecting CVD-risk genes and that may be amenable to recall-by-genotype studies (i.e., where volunteers are selected based on the presence or absence of genetic variants of interest) in INTERVAL in order to investigate their biological impact.

***A study of the correlation between red cell count parameters in donors and their corresponding red cell components	Dr Rebecca Cardigan	University of Cambridge / NHS Blood and Transplant	NHS Blood and Transplant (NHSBT) produce >1Million red cell concentrates (RCC) per year, from whole blood donations. Around 4-5% of RCC are tested for volume, haematocrit and haemoglobin (Hb) content as part of quality monitoring (QM). Within that distribution of Hb values, some fall below the acceptable level of 50g/unit. Such reduced Hb yield may be caused by a donor of below average Hb having their donation processed by a method producing lower yield RCCs. To potentially correct this, we could direct donations of predicted lower Hb into the higher yield processing stream. NHSBT does not currently know how closely donor Hb values correspond to RCC Hb yields, as accurate blood counting is not among the tests routinely performed on its donors. However, blood counts were performed on NHSBT donors during INTERVAL. In order to estimate how much NHSBT RCC QM values are affected by donor blood values , we wish to obtain information on INTERVAL donors (including blood counts and factors likely to play a role, such as sex, age, blood group and details of the donation process).
INTERVAL dataset: QC of Caprion Proteomics data	Professor Adam Butterworth	University of Cambridge	We require an up-to-date dataset in order to quality control (QC) new protein data.
INTERVAL dataset: QC of Genos glycans data	Professor Adam Butterworth	University of Cambridge	We require an up-to-date dataset in order to quality control (QC) new glycan data.
Evaluation of post-donation testing strategies for blood donors	Professor Emanuele Di Angelantonio	University of Cambridge / NHS Blood and Transplant	We will use accurate blood measurements taken from analysis after repeated blood donations, together with key characteristics of blood donors, to model recovery after donation. We will then use this information to develop and evaluate alternative strategies for blood donation in England.
STRIDES post-donation testing study	Dr Lois Kim	University of Cambridge	The STRIDES post-donation testing study is collecting repeat haemoglobin measurements from blood donors in order to inform an evaluation of post-donation testing policies, with the aim to identify strategies to reduce donor deferrals. However, staffing issues following COVID have meant that for the majority of donors, only on-site measurements have been collected for those deferred, whereas more accurate off-site measurements were collected for donations. In order to conduct an appropriate analysis of this study it is necessary to convert between the two types of measurement. This is not possible using the STRIDES data since the vast majority of donors do not have both measurements. We are therefore seeking data from COMPARE to inform this aspect of analysis.

Identifying associations between signatures of platelet reactivity in CBC scattergrams and disease risks	Dr Will Astle	University of Cambridge / NHS Blood and Transplant	We will combine data from the Cambridge Platelet Function Cohort and INTERVAL/COMPARE to develop a signature of platelet reactivity from Sysmex haematology analyser scattergrams. We will derive a genetic predictor of this signature using INTERVAL and subsequently perform association analyses of this genetic predictor with clinical outcomes in UK Biobank.
Genetic associations of age acquired skewed X chromosome inactivation and their link to disease	Dr Emma Davenport	Wellcome Sanger Institute	Due to the difference in number of genes between the X and Y chromosomes in mammals, in every female cell, one of the two X chromosomes is chosen at random to be silenced. This is termed Chromosome X inactivation (XCI). The selection of which X to silence is a random process during development and therefore across a tissue, such as the blood, it is expected that each X chromosomes will be silenced in 50% of cells. However, as females age, some individuals develop a "skewed" pattern of XCI which is very different from the expected 50:50 ratio, and this has been linked with some chronic diseases such as cancer. Many studies have shown that development of skewed XCI is influenced by genetics, but little research has been done to identify which genetic variants influence skewed XCI. Understanding which genetic differences between people increases the chances of skewed XCI developing as they age will will allow us to better understand who is at risk.
Analysis of RNA-sequencing data from the INTERVAL cohort	Professor Mike Inouye	University of Cambridge	eQTL and sQTL mapping from RNA-sequencing data provide a way to better understand the genetics of transcription regulation. In this project, we wish to analyse the RNA-sequencing data from the INTERVAL cohort, and integrate other datasets from multiple sources, including proteomics, metabolomics, and phenotypic information. Such study would be great resource for the broader scientific community to better understand the role of genetic variants in disease development through the regulation of genetic expression and may help identifying new therapeutic targets.
NMR metabolites: Correlates and associations with incident disease outcomes	Dr Stephen Kaptoge	University of Cambridge	Each molecular phenotype can, in principle be mapped onto the genome using advanced technologies. On the horizontal axis from the DNA to phenotype, the metabolome is at closest proximity to the phenome, and thus a detailed study of metabolite correlates and disease associations may potentially yield novel insights on disease aetiology. Therefore, we aim to assess or replicate correlates and associations of NMR-based metabolites and major disease outcomes using data available in the INTERVAL cohort.

***Accessing CALIBER endpoints to identify disease risk outcomes (genetic loci associated with COVID-19 traits)	Dr Elodie Persyn	University of Cambridge	In the original project, we analysed RNA-sequencing data and 'omics data (proteomics and metabolomics) to identify genetic variations regulating gene expression in blood. We then extended analyses, using Electronic Health Record data, to identify disease risk variants. That analysis revealed genetic loci with shared associations with COVID-19 disease traits. We are now accessing COVID-19 Test Results Data to further investigate the biological mechanisms of these genetic loci.
INTERVAL dataset: QC of SomaLogic protein data	Professor Adam Butterworth	University of Cambridge	It is now possible to measure the levels of thousands of proteins in a blood sample from an INTERVAL study participant. We have worked with a company in the US ('SomaLogic') to measure ~7,000 proteins in blood samples from ~10,000 INTERVAL participants. We will use these data to: - Understand the relevance of proteins and biological pathways to health and disease - Identify the genetic variations in our DNA that influence levels of proteins in the blood - Work out which proteins cause diseases, such as heart disease, stroke and cancer - See if protein measurements can improve existing methods for predicting who will (and who won't) develop various diseases - See how this way of measuring proteins compares to other methods already used in samples from INTERVAL participants
***Factors influencing red blood cell side-scatter, a flow-cytometric proxy of oxygen-unloading rate	Professor Pawel Swietach	University of Oxford	To adequately oxygenate tissues, red blood cells (RBCs) must carry enough oxygen on haemoglobin and then release this within the timeframe of a typically capillary transit (i.e. a few seconds). Whereas the first process is well described in terms of haemoglobin concentrations, the latter is not routinely measured because it requires complex devices. We developed a technique, called single-cell oxygen saturation imaging, to show that oxygen unloading from RBCs is slower than previously estimated and can even become rate-limiting for tissue oxygenation. Our preliminary observations indicate that the shape of RBCs, as measured by flow-cytometric side-scatter, is a good proxy of the oxygen-unloading rate, and we are seeking BioResource datasets (INTERVAL, COMPARE, STRIDES) to determine: (i) the robustness and reproducibility of side-scatter calibrations, (ii) how side-scatter varies between individuals, (iii) the factors that determine its magnitude and (iv) how it influences tissue oxygenation alongside the blood's carrying capacity.
Drug target validation for cardiovascular and neurological disease	Dr Amand Schmidt	University College London	We are running a number of drug target validation projects on cardiovascular disease and neurological disease, using protein and mRNA linked genetics data. Analyses typically include Genome-Wide Association Studies (GWAS), Mendelian randomisation and colocalization. We are accessing the INTERVAL Whole Genome Sequencing (WGS) and Whole Exome Sequencing (WES) data to enlarge our dataset and improve our analyses.

Genome-wide association study and Mendelian randomization investigation of vasovagal syncope	Dr Elias Allara	University of Cambridge	Vasovagal syncope' is the most common type of fainting and a known adverse effect of blood donation. It is currently unclear how genetic factors contribute to an individual risk of developing vasovagal syncope. Understanding a blood donor's genetic risk of vasovagal syncope could help prevent this adverse effect from occurring. In this study, we aim to identify the genetic determinants of vasovagal syncope by analysing data from approximately 25,000 participants enrolled in the COMPARE trial and by combining these results with those from other large studies conducted in Denmark, the Netherlands and the UK.