Q&A

The role of UK Biobank

An interview with Rory Collins, Head of Oxford University's Nuffield Department of Population Health and British Heart Foundation Professor of Medicine and Epidemiology

Precision Medicine has been fueled by convergent technology developments – first, the ability to generate massive amounts of data at relatively low cost and second, the capacity to move, store, and share data among many organizations. Several groups had the foresight to initiate biobanks as an archive of tissue samples to generate data from current and future analytical assays. One can see in hindsight the value of setting up the UK Biobank, though setting one up in 2005 required remarkable insight. This was at a time when many technologies to process samples and the subsequent deluge of data were not yet available, let alone what questions could be addressed.

We asked Rory Collins,[1] who was instrumental in establishing the UK Biobank, to address a few questions on the role of the biobank in public health. The UK Biobank was founded with the mission of improving the prevention, diagnosis and treatment of a wide range of serious and life-threatening illnesses – including cancer, heart diseases, stroke, diabetes, arthritis, osteoporosis, eye disorders, depression and forms of dementia.

1. Design and Quality

Could you comment on the initial design schema and quality criteria, especially for the fields of population health and epidemiology? Beyond, say, size and diversity, were there public health goals stated from the outset?

What questions did you and your colleagues have in mind as you planned the biobank?

A. Many people in the UK and beyond had been arguing that there was a need for some very large deeply characterized population cohorts with secure long-term follow-up of health outcomes. As you indicate, it was extraordinary prescience at the UK Medical Research Council (MRC) and the Wellcome Trust that led to the decision to provide funding for UK Biobank at a level that would allow a large, deep, long-term cohort to be built, with the proviso that the data would be made available to the global research community. Rather than being designed to study some particular risk factors or some particular health conditions, the aim was to make UK Biobank as widely applicable as possible.

How have the design and quality criteria (e.g., sample types, data format) changed over the years?

A. The initial investment by MRC and WT to create a large and deep resource, combined with ready access by all types of researchers, has leveraged substantial additional funding from government, charity, and industry. That has, in particular, allowed the samples from all 500,000 participants to be turned into data, both genetic (genotype then exome sequence and now whole genome sequence) and other assays (initially haematology and biochemical measures known to be relevant, but more and more into more agnostic proteomic and metabolomic measures).

In addition, following the rather limited physical assessment of all participants at the baseline visit, we're now conducting detailed imaging of 100,000 of them (and, subsequently, will do repeat imaging) which will provide not only further measures that may be relevant to the development of health outcomes, but may also serve as intermediate health outcomes in their own right.

As with the ways in which the UK Biobank data have been used in lots of interesting and unexpected ways by the broad range of researchers who are accessing it, enhancement of UK Biobank has been driven by the imaginations and expertise of the external research community. Our role within in the UK Biobank team is to turn the aspirations of researchers into realities … as well as understanding that the UK Biobank resource is there for them to use, we need the research community to tell us how to make it even more valuable for their research.

UK's population is highly diverse. Could you comment on how the UK's diversity is reflected in the Biobank's samples?

A. In order that a prospective cohort can provide generalisable information about the associations of various exposures (whether genetic, lifestyle or environmental) with subsequent health outcomes, what is needed is heterogeneity of exposures. What's not needed is that participants in a cohort are representative of any particular population. For example, if 80% of a particular population smoked cigarettes, a cohort that included 80% of participants who smoked (i.e., was representative) may well provide less reliable and generalisable information about the health effects of smoking than one that included 50% of participants who smoked and 50% who did not (i.e., was deliberately non-representative) which would tend to have greater statistical power.

What is needed to understand the relevance of risk factors in different circumstances is to have large enough numbers of individuals who have different levels of many different exposures to be able to assess whether associations differ in different types of people (e.g., men versus women, younger versus older, more versus less deprived, etc). And where that is not possible – for example, by recruiting sufficiently large numbers of people from various minority groups – then the solution is not to make undue efforts to recruit a "representative" proportion from such groups (which may well still be too small to be scientifically valuable), but instead to establish studies in places where such individuals are not a minority and, by doing, potentially also increase heterogeneity in other ways (e.g. extending the range of other exposures or diseases that can be studied reliably)

Another aspect of diversity is that different tissues are well-represented in the biobank. Can you please comment on the diversity with respect to tissue types – e.g., blood vs tissue biopsies?

A. We collected about 55mls of blood at the baseline assessment visit which was processed in a range of different ways in order to support a wide range of assays that it was thought might be done on them at some time in the future. Spot urine samples were also collected in all the participants, along with saliva samples – at the request of dental researchers – from the last 200,000 or so participants who were recruited. The assay strategy that has developed is one of seeking to analyse all 500,000 participants, either in one go (as we did with the haematologic, biochemical and genotype assays) or in a series of tranches of randomly selected samples (as with the sequencing projects, as well as the ongoing lipidomic and proteomic assays).

We're collecting many of the same samples at the imaging assessment visits of 100,000 participants so that the relevance of different rates of changes in various non-genetic factors to the subsequent development of disease can be assessed. However, we are open to suggestions from the research community about what else (e.g., cells) should be collected from participants attending for imaging.

2. Multi-omic data: genomics, proteomics, metabolomics releases

Multi-omics data was not well annotated (or functions understood) at the time of setting up the UK Biobank. What role did the UK Biobank play in defining, filling in, and expanding such information?

Can you comment on how multi-omic information has evolved over the years to create a molecular profile of individuals in the UK Biobank – e.g., classifying and treating tumors by genotype?

A. An extensive set of pilot studies was conducted before UK Biobank started to help develop sample collection, processing and storage procedures that would support a wide range of different kinds of assay at some point in the future. Of course, knowing what we know now – 15 years later – there are samples that were collected that are less valuable for the intended purpose ... so we're looking at ways that they might be repurposed. And, there are sample types that we wished that we had collected which we didn't, so we're very interested in considering what might be collected at the imaging visits to enhance the ability to characterize those UK Biobank participants.

In addition to an increasing range of genomic and other –omic assays that are now being conducted on the stored samples, we are also investigating ways in which to enhance the characterization of health outcomes that occur. That includes obtaining more detailed information from health record systems for particular outcomes (such as imaging data for someone who has had a stroke in order to be able to determine the stroke subtype), but we are also looking at getting more information about cancers not just from record systems (with histology data now increasingly enhanced in the UK by sequence data) but also by aiming to obtain tumour samples for subsequent assessments.

3. Impact on therapeutic areas covered by the UK Biobank

Are certain diseases targeted by the biobank – e.g., by prevalence in the UK? or other public health metric?

Has enough time passed to see the impact of the population health programs on biobank data – e.g., heart disease or smoking-related conditions? Or vice-versa, the use of biobank data to influence public health programs (e.g., NICE guidelines or other)?

A. No, UK Biobank has been deliberately set up to allow researchers to study as wide as possible range of exposures for as wide as possible range of health outcomes. And we are always looking at ways to increase the range of exposures assessed (for example, web-questionnaires about diet, occupation, etc; remote monitoring of activity, etc; and linkage to socio-economic and environmental data).

As the duration of follow-up increases then so too does UK Biobank's ability to support research into the determinants of a wider and wider range of health conditions that have occurred in sufficiently large numbers. Perhaps the most obvious way in which UK Biobank may influence public health programmes soon is in the application of the polygenic risk score (PRS) concept – which derived initially from analyses in UK Biobank – to preventive strategies: for example, fine-tuning screening strategies which are currently driven to a large extent by age (such as offering breast cancer screening at younger ages to women with a high PRS-predicted risk).

4. Precision Medicine: Developing and prescribing drugs, vaccines, and biologics by population molecular profiles

Can you cite examples of how the UK Biobank has contributed to the UK Precision Medicine program – e.g., companion diagnostics or novel therapies?

A. I don't think that UK Biobank will necessarily contribute to so-called "precision medicine" strategies per se, but instead will help researchers to understand disease processes better and, by so doing, identify innovative ways to prevent and treat diseases.

The application of PRSs – which derived from research using UK Biobank – may well lead to what might be called "Precision Population Health" (for example, determining who to offer screening for breast cancer earlier or, for those at low predicted risk, later; who to offer more than fecal occult blood tests in screening programmes for colorectal cancer; who to recommend take cholesterol-lowering therapy from a much younger age), which may have a much bigger impact on the health of the population than determining who to treat with a specific treatment (albeit that too is valuable).

Having said that, from the large-scale sequencing data in UK Biobank, we are now starting to see novel targets for health conditions that could also be of value for large numbers of people.

You advised the All of US[3] governance bodies in the US. Would you be able to share a high-level summary of the advice you provided?

A. Aim to ensure that the cohort is able to provide generalisable evidence about the association of many different exposures for many different diseases rather than aiming to be "representative" of the population at a particular moment in time – especially since, by the time sufficient duration of follow-up has occurred to study the determinants of disease in All of Us, the population will have changed, so the cohort may no longer be representative, but it could still be generalisable.

Aim to ensure that all of the participants provide all of the exposure data and biological samples that are being sought, rather than having them pick and choose such that few provide everything, since it is important to bear in mind that only a small proportion will develop any particular disease and you, therefore, want to have complete exposure information in all (or almost all) of them.

Aim to ensure that it is straightforward to obtain information about health outcomes that occur during the subsequent 10-20 years in all of the participants since they do not provide useful information (but do use resources) if their health cannot be followed securely long-term.

5. COVID-19 and heart conditions UK Biobank COVID-19 hub[4]

Of more recent note, the UK Biobank has established a COVID-19 hub. Can you describe briefly the goal of the hub in gathering and disseminating information about COVID-19 in the UK?

COVID-19 impacts many organs in the body, especially the heart and lungs. What have you learned from the biobank on the cardiovascular health of those who were infected by the virus? Have you seen different impacts by virus variants? Or in different population cohorts – e.g., age, sex?

A. With such rich data already on 500,000 incredibly altruistic participants in UK Biobank, it seemed sensible to enhance the data for research into COVID-19 by adding virus testing data plus primary care data, which was made available under emergency legislation, and making these data rapidly available to approved researchers around the world in the hope that they might identify ways to help manage the pandemic.

We have also obtained antibody results from 200,000 of the participants who agreed to do the test in order to identify those who had been infected, with the intention that this information could support research into so-called "long COVID".

Among the 50,000 participants imaged prior to the pandemic, we are doing repeat imaging in about 1500 participants who have been infected with SARS-CoV-2 and 1500 matched participants without evidence of infection. By doing so, we will be able to make available unique pre-and post-infection imaging data, which will allow robust and unconfounded investigation of the effects of infection on body systems (by contrast with almost all other studies which only have post-infection imaging data in highly selected individuals without well-matched controls).

6. Since the UK Biobank is based in Manchester, England, has Brexit had an impact on the biobank's ability to share or exchange information with any EU or US partners?

A. There is little about Brexit that is positive, but it has not affected UK Biobank's ability or desire to be a resource that is of value globally. Indeed, when we launch UK Biobank's Research Analysis Platform in September, we hope that that will make the data and the compute required to analyse it even more accessible, further democratizing its availability for health-related research that is in the public interest.

7. Has the UK Biobank been used for forensic research – e.g., identifying criminals, disaster victims, or family relationships? What legal, scientific, and ethical considerations (e.g., consent) concerns are weighed in allowing access to biobank data for forensic purposes?

A. No. In our original consent, we made it clear to participants that their data and samples would only be made available for health-related research. In particular, we stated explicitly that we would resist any attempt to access the UK Biobank resource for other purposes (such as forensic investigation).

8. Would you like to make any final comments?

A. Yes, the success of UK Biobank depends on three things. First, the extraordinary altruism of the very special 500,000 people who agreed to join the project and who continue to contribute to it (with, remarkably, only about 1000 having withdrawn). Second, the initial foresight of the MRC and the WT to set up UK Biobank and, subsequently, their continued support (along with additional support from government, charities and industry) to help make it better. And, third, the research community which is using the UK Biobank data to make discoveries which will benefit patients and the wider public. UK Biobank is only of value if the data are used, and the more they are used and the more imaginatively that they are used by scientists from different disciplines, (that is, not just traditional medical and biological scientists, but also data scientists and engineers), the better.

The success of UK Biobank depends entirely on the altruism of the 500,000 participants, the unwavering support of the funders, and the imaginations of the researchers who use the data to improve health. JoPM

Rory Collins

Rory studied Medicine at St Thomas's Hospital Medical School, London University (1974-1980), and Statistics at George Washington University (1976-7) and at Oxford University (1982-3).

In 1985 he became co-director, with Professor Sir Richard Peto, of the University of Oxford's Clinical Trial Service Unit & Epidemiological Studies Unit (CTSU). In 1996, he was appointed Professor of Medicine and Epidemiology at Oxford, supported by the British Heart Foundation.

He became Principal Investigator and Chief Executive of the UK Biobank[2] prospective study of 500,000 people in September 2005. From July 2013, he became the Head of the Nuffield Department of Population Health at Oxford University.

His work has been in the establishment of large-scale epidemiological studies of the causes, prevention and treatment of heart attacks, other vascular disease, and cancer. He was knighted in 2011 for his services to science.

References

1. Rory Collins' Nuffield department of population health website, https://www.ndph.ox.ac.uk/team/rory-collins
2. UK Biobank website: https://www.ukbiobank.ac.uk/; https://www.ctsu.ox.ac.uk/research/uk-biobank
3. All of US website: https://allofus.nih.gov/
4. UK Biobank COVID-19 hub: https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank/covid-19-hub