UK Biobank: a proposal to significantly advance research that improves public health and patient care via the coded primary care data of consented participants

During 2006 to 2010, 503,000 men and women aged 40-69 years agreed to join the UK Biobank study. As part of study recruitment, all of the participants gave their consent for *"…access to my medical and other health-related records, and for long-term storage and use of this and other information about me, for health-related research purposes (even after my incapacity or death)."*

Participants are regularly informed about progress with the use of the resource for health-related research in accordance with their consent, as well as being asked to contribute further information (e.g. completing on-line questionnaires, wearing various devices, attending an imaging assessment). Participants have also been informed that UK Biobank is following their health through linkage to NHS records, including obtaining data from their GPs during the COVID pandemic, which are made available to researchers. Participant focus groups have found that they expect their primary care data to be made available routinely, and many are surprised and disappointed when told it is not happening.

> *"My wife and I are both participants in UK Biobank. It is over a decade since we willingly and enthusiastically undertook the battery of tests and questionnaires ….*
>
> *We understood that the project would also have access to our medical records to be able to include our data for whatever research purposes that might be needed.*
>
> *I had a routine appointment with my GP today and I mentioned our participation in the UK Biobank project. He was not aware of it and was quite sure that our surgery was not providing any data to the project. This was as surprising as it was disappointing."*
>
> Participant Focus Group

Value of primary care data for research to improve public health and patient care

Access to coded primary care data (i.e. codes related to diagnoses, prescriptions, referrals, etc.) under a COPI notice during the pandemic has demonstrated the value of being able to combine primary care data with other sources of health outcome data (e.g. hospitalisation admissions, cancer and death records) that have already been made available to UK Biobank by the NHS centrally. Over 200 papers about COVID have been published using these linked healthcare records, many of which required the primary care data to investigate the role of co-morbidities and medications as determinants of severe COVID-19 (e.g., Pavey, et al. 2022; Yu, et al. 2021; Xiang, et al. 2021).

Extending this access to coded primary care data for broad research purposes (i.e. not solely for COVID-19 research) in accordance with the participants' consent would enable new understanding of causes and development of disease, and would support new approaches to prevention and treatment of a wide range of conditions, especially those managed largely outside of hospital (e.g., arthritis and other causes of pain, dementia and other neurodegenerative conditions, impaired vision or hearing, many respiratory conditions, heart failure and mental health problems).

Such conditions have been systematically under-represented in large-scale epidemiological studies. Securing access to primary care data for consented research resources (such as UK Biobank) offers an unparalleled opportunity to redress this imbalance, with the potential for major impact on public health. For example, the inclusion of the coded primary care data in UK Biobank would result in an approximate doubling of cases of depression and dementia that can be identified (see Figure), as well as allowing detection of less severe cases at an earlier stage, enabling studies across the full spectrum of disease severity to further the understanding of disease progression.

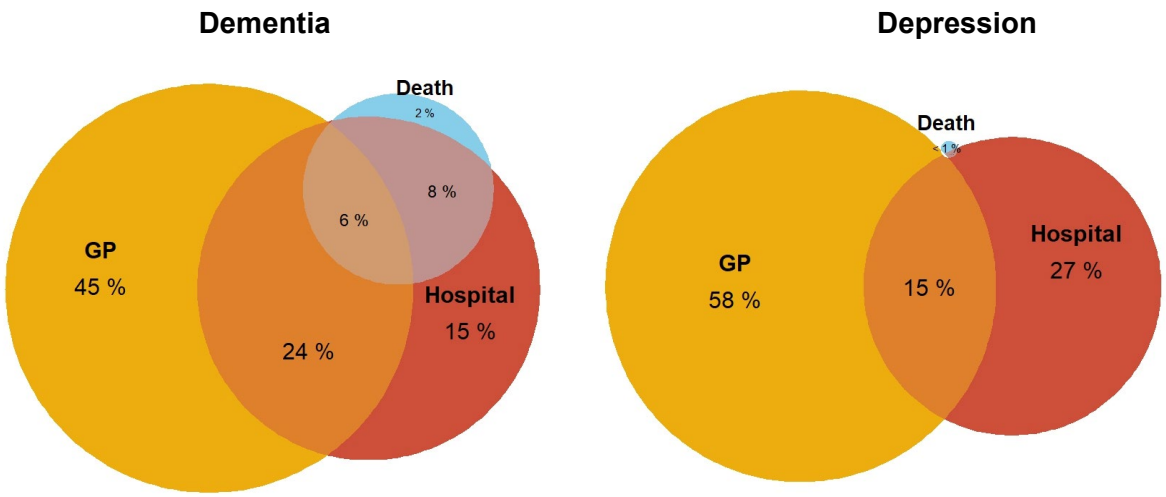

Figure: Exemplar health outcomes for which primary care data dramatically increase the numbers of cases that can be detected compared with other sources of healthcare data

How the breadth of UK Biobank data leads to identification of disease subtypes

The breadth and depth of data available in UK Biobank are already enabling novel approaches that go well beyond the simple combination of linked healthcare data to sub-classify health outcomes more specifically (see exemplars in the box below).

- Genetic data can accurately discriminate between type I and type II diabetes, indicating that type I diabetes is misdiagnosed in middle and old age (Thomas, et al. 2018).
- Distinct co-morbidity patterns that have discrete genetic profiles can discriminate diabetes subtypes even further (such as early vs. late-onset type II diabetes) (Jiang, et al. 2023).
- Genetic factors and metabolic traits can differentiate early vs. late-onset and atopic vs. non-atopic asthma subtypes (Zhu, et al. 2020).
- Machine learning approaches can classify over 1500 broad health conditions based on genomic, clinical and other phenotypic data contained in UK Biobank (Yang, et al. 2023).

Such multi-modal approaches hold huge promise for better classifying disease subtypes based on their genetic and molecular profiles. In particular, the emergence of novel artificial intelligence approaches will accelerate the use of orthogonal datasets (such as genomic, proteomic and other -omics datasets) for such purposes. More distinct disease sub-classification could revolutionise prediction, diagnosis, prognosis and treatment of disease in healthcare settings. But without being able to integrate primary care data integrated within UK Biobank, the opportunity will be missed to drive this revolution with the full range of disease presentations from across both primary and secondary care settings.

Issues with obtaining access to primary care data for consented UK Biobank participants

General Practitioners are currently the data controllers for primary care data. Previous attempts by UK Biobank to obtain their explicit agreement to approve the release of coded data for participants have been unsuccessful (<20% agreeing to do so, with the vast majority non-responsive) despite the explicit consent from all participants.

The main reason behind their lack of agreement is not known for certain but, as well as being busy and having other priorities, it appears to be largely due to concerns about data sharing responsibilities. Experience with the COPI notice reinforces this conclusion, since it removed the need for GPs to be

responsible for the data sharing decision, instead relying on a Secretary of State instruction to the EMIS and TPP system suppliers to make the coded data directly available to UK Biobank. Apart from a few Practices that sought further information, no issues were raised with the primary care data being made available to UK Biobank in this way.

Most recently, with the endorsement of the RCGP, a letter has been sent by NHS England to General Practitioners in England asking them to approve the release of coded data to UK Biobank for consented participants registered at their Practices. Given the low response rate in previous pilot studies, it would seem unlikely that comprehensive access to primary care data across England will be achieved. Indeed, one week after the mailing, only 55 (of several thousand) TPP practices had agreed to release of the data. Incomplete data collection may not be random, leading to biases in analyses if participants for whom primary care data are not available differ materially from those for whom they are available.

Alternative approaches involving summary "composite" health outcomes limit research value

NHSE has enabled access for OpenSAFELY to primary care records of 59 million people across England without consent, or explicit agreement from their General Practices, by issuing a Data Provision Notice (DPN). Currently, use of these data is limited to COVID-related research (as it derived from a previous COPI notice), but OpenSAFELY is seeking an extension for other research purposes.

It has been proposed that the OpenSAFELY platform could address UK Biobank's needs (and those of other consented cohorts) by applying algorithms within OpenSAFELY to combine coded primary care data in order to provide derived health outcome codes to UK Biobank. However, although OpenSAFELY should be commended for its approach to generating reproducible code and algorithms for analysing primary care data, there are a number of reasons why it cannot fulfil the needs of UK Biobank, particularly that it will not offer the ability to interact with the raw data, pose novel questions and find new methods of analysis. Our large and diverse community of researchers has already shown the value of this using existing data. Leaving aside the lack of scalability of the OpenSAFELY approach (e.g. contrast the approximately 60 publications that have emerged so far, largely authored by the internal team, according to the OpenSAFELY website versus the 2,000 papers published based on UK Biobank by external researchers globally in 2022 alone), these limitations include:

- *Unsophisticated analyses:* As so many types of researchers use UK Biobank there is great variation in need, and using derived variables severely limits the scope of what can be achieved. Creating and extracting composite health outcomes (or other derived variables) using algorithms that combine sources of information solely within primary care data would not take advantage of the power of innovative approaches (including artificial intelligence) that use the richness of the wider data available within UK Biobank (e.g. genomic sequencing, proteomics, metabolomics, imaging, etc.) to further characterise health outcomes in an agnostic manner. As an analogy, it would be akin to only allowing known algorithms to be applied to the MRI images of UK Biobank participants in order to derive brain volume rather than the approach that has been taken of enabling the development and application of analytic methods to generate thousands of image-derived variables (which, for example, have been found to be informative about dementia and other neurodegenerative conditions: Elliott, et al. 2018). If, on the other hand, de-identified record-level data were made available directly within the UK Biobank resource (as is the case with the other NHS health record data), then researcher innovation, expertise and added value back to participants and the wider public would be maximised (and would leverage the considerable public and charitable investment made so far within UK Biobank).
- *Data governance:* The issues that OpenSAFELY has been designed to address (i.e. unconsented research use of primary care data by researchers in a trusted and auditable manner) are already solved within UK Biobank as a trusted research environment (i.e. approved *bona fide* researchers working on approved research applications using de-identified data, and with regular reporting and publishing of research findings fully aligned with the explicit participant consent). A model whereby OpenSAFELY would provide composite health outcomes to UK Biobank for consented

participants based on aggregating primary care records is not fundamentally different in research governance terms from coded primary care data being provided directly to UK Biobank from the system suppliers. In both cases, specific information about the health conditions of particular individuals would be made available in the UK Biobank resource for approved researchers to use. However, as well as providing no added benefits or assurance, the OpenSAFELY approach would lead to an effective monopoly within OpenSAFELY rather than taking advantage of the expertise in the wider global researcher community to interrogate the primary care data in innovative ways (in particular, in combination with other data available in UK Biobank: as in the examples above).

Consequently, from both the perspective of optimising the potential for improving public health and patient care through use of UK Biobank for research in accordance with the explicit wishes of participants, and from a research governance perspective, direct access to coded primary care data is warranted (as is already the case with other health data being provided to UK Biobank by the NHS).

Conclusion

In conclusion, the addition of coded primary care data has been shown unequivocally to increase the numbers of many different cases of particular conditions (especially at an earlier stage of development of a disease) that can be identified, and the participants have given their explicit consent for access to all of their medical and other health-related records (and UK Biobank's recent communications with them confirm the persistence of this consent). However, UK Biobank has demonstrated in a series of pilot studies that securing Practice-by-Practice agreement cannot ever be successful.

Consequently, UK Biobank proposes that NHSE issue a DPN (or some such central instruction) for provision of the coded primary care data for its consented participants registered with Practices in England. [Primary care data are available to UK Biobank for participants in Scotland and Wales through central systems.] Such a provision would be consistent with the participants' consent (as confirmed by the Information Commissioner), would reduce data governance concerns for General Practices, and would substantially enhance the ability of UK Biobank to support a wide range of innovative health-related research in accordance with the participants' wishes.