# Chapeau UK Biobank! A revolution for integrated research on humans and large-scale data sharing

Thomas Bourgeron

INSTITUT DE FRANCE
Académie des sciences

# Comptes Rendus
# Biologies

Thomas Bourgeron

**Chapeau UK Biobank! A revolution for integrated research on humans and large-scale data sharing**

Volume 345, issue 1 (2022), p. 7-10

Published online: 11 May 2022

https://doi.org/10.5802/crbiol.76



*Les* Comptes Rendus. Biologies *sont membres du*
*Centre Mersenne pour l'édition scientifique ouverte*
www.centre-mersenne.org
e-ISSN : 1768-3238

*Comptes Rendus*
*Biologies*
**2022**, 345, nº 1, p. 7-10
https://doi.org/10.5802/crbiol.76

News and Views / *C'est apparu dans la presse*

# Chapeau UK Biobank! A revolution for integrated research on humans and large-scale data sharing

## *Chapeau UK Biobank ! Une révolution dans la recherche intégrée sur l'humain et le partage de données à grande échelle*

**Thomas Bourgeron** [a]

[a] Membre de l'Académie des sciences, Université de Paris et Institut Pasteur, Paris, France

*E-mail:* Thomas.bourgeron@pasteur.fr

**Keywords.** Biomedical research, Data sharing, Biobank, Genetics, Epidemiology, Human population.

**Mots-clés.** Recherche biomédicale, Partage de donnée, Biobanque, Génétique, Épidémiologie, Population humaine.

*Manuscript received 15th February 2022, accepted 23rd February 2022.*

It is obvious, but sometimes it needs to be reminded, researchers need data to generate and test novel hypotheses. The more the dataset is of high quality, the more robust the scientific results and predictions are produced. And the more widely the data is shared, the more verifiable and reproducible the results are. In human research, many initiatives have been made to produce large-scale data, ranging from national registries that include the entire population, to the establishment of cohorts of varying sizes. In 2006, the UK Biobank biomedical database and resource was created, and it has revolutionized research on several levels, particularly because the data is accessible to the scientific and medical communities [1].

The initial budget of UK Biobank was £62 million, funded by the UK government and the Wellcome Trust. The objective was to collect data on 500,000 participants aged 40–69 years old over four years and then follow them for at least 30 years. The database was opened to researchers in March 2012, hosting demographic, medical, psychological, socio-economic data (from interviews or online questionnaires), biological samples (e.g., urine and blood), genetic data and imaging data. Ten years later, more than 28,000 researchers had access to this data.

Why is UK Biobank revolutionizing research? Unlike national registries where data access is limited, researchers from all countries can access UK Biobank data after their research project was approved and their IT security to host the requested data was verified. Although genetic data is considered sensitive, UK Biobank has obtained informed consent from participants to share it. There is a periodical update sent to all registered researchers to ask them to remove participants who chose not to share their data anymore.

More than five petabytes of genetic data are available including >500,000 microarrays, >400,000 whole-exome and >200,000 whole-genome sequences [2]. These data allow researchers to explore the genetic contribution to many traits and diseases. Among the major results, this resource has been key to estimate the heritability of more than 4000 phenotypic traits and are available online at the Benjamin Neale laboratory (https://nealelab.github.io/UKBB_ldsc/index.html). Genetic variations have been identified as risk factors and predictive models have been proposed for diseases such as diabetes, cardiovascular disease and cancer [3]. You can interrogate if a gene is associated with a specific trait here https://azphewas.com/.

Epidemiological and genetic data are also linked to imaging data. The goal is to have magnetic resonance images (MRI) of the brain, heart, and abdomen for 100,000 participants (already 50,000 available). The combined analysis of genotyping and brain MRI data has allowed, among other things, to the estimation of the heritability of interindividual differences in brain anatomy [4, 5].

Regarding COVID-19, all registered researchers received messages informing them of the collection of new data collected in relation to COVID-19 and encouraging them to submit research projects on the pandemic. More than 740 researchers responded, and more than 148 papers were published. A specific UK Biobank study also collected new data on 20,000 volunteers who were either original UK Biobank participants or their children or grandchildren over the age of 18. A total of 6.6% of participants had already been infected by May/June 2020 and this rate increased to 8.8% by the end of November 2020. This study was one of the first to show that antibodies produced following natural infection can protect most people from further infection for at least 6 months (The UK Biobank SARS-CoV-2 Serology Study report is available here: https://www.ukbiobank.ac.uk/media/x0nd5sul/ukb_serologystudy_report_revised_6months_jan21.pdf).

I have only scratched the surface of the findings and the possibilities offered by UK Biobank. I have not mentioned all the richness and reliability of the data as well as the next objectives such as the linkage with other registries such as death, cancer, … I also omitted the ability to share results of the research projects directly via the UK Biobank portal, and the description of the limitations and biases of representativeness of UK Biobank which are well documented [6]. All this information is available on the UK Biobank website https://www.ukbiobank.ac.uk/. Again, for scientists to work, there is a need for large amounts of high quality and accessible data. There is an urgent need to support such initiatives in other countries to replicate the UK Biobank results, to increase the diversity of people studied and to detect associations that are specific to countries with different health systems.

## Conflicts of interest

The author has no conflict of interest to declare.

## *French version*

C'est une évidence, mais il faut parfois le rappeler : pour trouver, les chercheurs ont besoin de données. Plus les données sont de qualité et en grande quantité, plus les résultats scientifiques et les prédictions sont solides. Plus les données sont partagées, plus les résultats obtenus sont vérifiables et réplicables. Concernant la recherche sur la population humaine, de nombreuses initiatives ont été mises en place afin d'obtenir des données à grande échelle allant de registres nationaux incluant l'ensemble de la population, à l'établissement de cohortes de plus ou moins grandes tailles. En 2006, la base de données biomédicales et de ressources UK Biobank est créée et a révolutionné la recherche à plusieurs niveaux, en particulier en rendant les données accessibles aux scientifiques [1].

Le budget initial de UK Biobank était de 62 millions de livres, financées par le Royaume-Uni et le Welcome Trust. L'objectif était de collecter en quatre ans des données sur 500 000 personnes de 40 à 69 ans, puis de suivre ces dernières sur au moins 30 ans. La base a été ouverte aux chercheurs en mars 2012 et elle héberge, entre autres, des données démographiques, médicales, psychologiques, socio-économiques (issues d'interviews ou de questionnaires en ligne), des prélèvements biologiques

(ex. urine et sang), des données génétiques et des données d'imagerie. Dix ans plus tard, plus de 28 000 chercheurs ont eu accès à ces données.

UK Biobank révolutionne la recherche car contrairement aux registres nationaux dont les données sont peu accessibles, la base est accessible aux chercheurs de tous les pays après validation de leur projet de recherche et vérification de la sécurité informatique dont ils disposent pour héberger les données qu'ils souhaitent récupérer. Bien que les données génétiques soient considérées comme sensibles, UK Biobank a obtenu le consentement éclairé des participants pour les partager. Il y a d'ailleurs continuellement des mises à jour pour indiquer si des participants ne veulent plus participer à la recherche.

Plus de cinq petabytes de données génétiques sont disponibles. Elles incluent plus de 500 000 puces à ADN, plus de 400 000 exomes complets et plus de 200 000 génomes complets [2]. Ces données permettent d'explorer la contribution des variations génétiques dans la vulnérabilité aux maladies. Parmi les avancées majeures, cette ressource a permis de préciser l'héritabilité de plus de 4000 caractères phénotypiques. Ces résultats sont accessibles en ligne sur le site du laboratoire de Benjamin Neale (https://nealelab.github.io/UKBB_ldsc/index.html). Des variations génétiques ont été identifiées comme facteurs de risques et des modèles prédictifs ont été proposés pour des maladies comme le diabète, les maladies cardiovasculaires et le cancer [3]. Vous pouvez interroger si un gène est associé à un trait spécifique ici : https://azphewas.com/.

Les données épidémiologiques et génétiques sont aussi reliées à des données d'imagerie. L'objectif de UK Biobank est d'avoir des images par résonance magnétique (IRM) du cerveau, du cœur et de l'abdomen pour 100 000 participants (50 000 données sont déjà accessibles). L'analyse combinée des données de génotypage et d'IRM cérébrales a permis, entre autres, d'estimer l'héritabilité des différences interindividuelles pour l'anatomie du cerveau [4, 5].

Concernant la recherche sur la COVID-19, tous les chercheurs dont les projets de recherche avaient été acceptés par UK Biobank recevaient des messages les alertant du recueil de nouvelles données en rapport avec la COVID-19 et les incitant à déposer de nouveaux projets de recherche. Plus de 740 chercheurs ont répondu et plus de 148 articles ont été publiés. Une étude ancillaire de UK Biobank a aussi collecté des données sur 20 000 volontaires qui étaient soit des participants originaux de UK Biobank, soit leurs enfants ou leurs grands enfants de plus de 18 ans. Au total 6,6 % de ces participants avaient déjà été infectés en mai/juin 2020, et ce taux est passé à 8,8 % à la fin du mois de novembre 2020. Cette étude a été l'une des premières à montrer que les anticorps produits à la suite d'une infection naturelle peuvent protéger la plupart des personnes contre une infection ultérieure pendant au moins 6 mois. (Le rapport de la UK Biobank SARS-CoV-2 Serology Study est disponible ici : https://www.ukbiobank.ac.uk/media/x0nd5sul/ukb_serologystudy_report_revised_6months_jan21.pdf).

Je n'ai effleuré que la surface des découvertes issues des données de la base UK Biobank. Je n'ai pas mentionné toute la richesse et la fiabilité des données accessibles, ainsi que les prochains objectifs. Par exemple, le chaînage avec d'autres registres, tels que celui des décès, des cancers, … Je n'ai pas non plus mis en avant le partage des métadonnées issues des projets (déposées par les chercheurs et accessibles via le portail de UK Biobank), ni fait la description des biais de représentativité de UK Biobank, qui existent bien évidement mais qui sont bien documentés [6]. Toutes ces informations sont disponibles sur le site web officiel de la base : https://www.ukbiobank.ac.uk/. Je le répète, pour que les scientifiques puissent travailler, il faut des données en grande quantité, de très bonnes qualités et accessibles. Il est urgent de soutenir de telles initiatives dans d'autres pays afin de répliquer les résultats issus de UK Biobank, d'augmenter la diversité des personnes étudiées et de pouvoir détecter des effets spécifiques de certains pays ayant des systèmes de santé différents.

## Conflit d'intérêt

L'auteur n'a aucun conflit d'intérêt à déclarer.

## References

[1] C. Sudlow *et al.*, "UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age", *PLoS Med.* **12** (2015), article no. e1001779.

[2] C. Bycroft *et al.*, "The UK Biobank resource with deep phenotyping and genomic data", *Nature* **562** (2018), p. 203-209.

[3] P. Glynn, P. Greenland, "Contributions of the UK biobank high impact papers in the era of precision medicine", *Eur. J. Epidemiol.* **35** (2020), p. 5-10.
[4] A. Biton *et al.*, "Polygenic architecture of human neuroanatomical diversity", *Cereb. Cortex* **30** (2020), p. 2307-2320.
[5] L. T. Elliott *et al.*, "Genome-wide association studies of brain imaging phenotypes in UK Biobank", *Nature* **562** (2018), p. 210-216.
[6] A. Fry *et al.*, "Comparison of sociodemographic and health-related characteristics of UK biobank participants with those of the general population", *Am. J. Epidemiol.* **186** (2017), p. 1026-1034.