UK Biobank / Report for the ICO / Incident involving data from feedback forms on UK Biobank's website / reference IC-43177-Z7PO

Summary

There has been an inadvertent disclosure of personal identifiers (such as names and email addresses) left by a total of 6,152 participants, researchers and members of the public on feedback forms on the externally managed UK Biobank website**.** As a result of this incident, the Chief Information Officer has resigned, and UK Biobank has engaged two external IT security-expert organisations to undertake a thorough review of its systems. In addition, the management of the website has been moved in-house and a new Chief Information Officer will be appointed in due course.

It should be emphasised that this incident does not involve any data from the entirely separate UK Biobank research database or any other UK Biobank IT systems.

The UK Biobank website (*the Website*) was, until very recently, developed and managed by a third-party company (*the Website Management Company)* on behalf of UK Biobank (http://www.ukbiobank.ac.uk). The actual hosting of *the Website* is managed separately by another third-party organisation. *The Website* is the principal public communications interface used by UK Biobank. As many such websites do, it incorporated feedback forms for visitors to *the Website* (be they researchers, participants or members of the public) to provide their feedback to UK Biobank about items that they may have viewed on *the Website*.

It should be further emphasised that this feedback is predominantly administrative and is not linked with UK Biobank's collection of health-related information from participants (which is contained within the main research database) used to support research intended to improve diagnosis, prevention and treatment of a wide range of diseases.

The incident in question relates to disclosure of personal identifiers collected on these forms, along with free text comments. The scope and type of data which were disclosed is summarised below and described in more detail in Annexes B & C.

The incident has arisen as follows (along with other relevant background):

- *The Website* was launched in 2005. It had a substantive rebuild in 2012, based on WordPress (a common website platform at that time). From 2012, generic feedback forms were included on the Website and the responses to these feedback forms were processed by the UK Biobank Participant Resource Centre[1] and the UK Biobank Communications Team: for example, to answer a question relating to material on the Website;
- The data from these feedback forms was aggregated and retained in a password-protected file. The password was available to the UK Biobank Communications Team and to the Website Management Company and thus the data was only accessible, in a secure manner, to a small number of authorised users;
- *The Website* was updated in January 2019 by the Website Management Company in order to use a newer version of WordPress. In the course of this update, the Website Management Company copied the aggregated feedback form data into an unprotected text-based CSV file;
- The CSV file was placed by the Website Management Company into a file location on *the Website*, which was accessible within the WordPress Media Library (and thus capable of being

[1] The UK Biobank Participant Resource Centre (PRC) is based in Cardiff and is the primary interface between UK Biobank and its participants for all participant queries.

identified and indexed by Internet search engines) and was not deleted after the update completed;

- However, the CSV file was not visible in the published architecture for *the Website* within the World Press Media Dashboard. UK Biobank was thus unaware of the existence of the CSV file;
- The CSV file was indexed by search engines and subsequently accessed by third parties. The web logs are incomplete (since they only cover a rolling 14-day period), so it is not possible to be sure how many times the CSV file had been accessed (in addition to the search engine robots) and by whom;
- UK Biobank was alerted to the existence of the CSV file on 16th June 2020, by a bioinformatics provider who had found a link to the CSV file via a Google search. As described below, the file was taken down immediately;
- However, a UK Biobank participant (an individual who is a volunteer in the UK Biobank research project) did contact UK Biobank in January 2020 to alert us to the possibility that their email address had been exposed (as they had used the email address solely for providing feedback to UK Biobank and subsequently received an unsolicited email to this email address);
- At the time, UK Biobank's then Chief Information Officer investigated but failed to identify this problem on the website. After UK Biobank was alerted in June 2020, the Chief Information Officer resigned;
- UK Biobank contacted the Website Management Company to better understand exactly how they allowed this disclosure of personal identifiers to happen. The Website Management Company has responded and acknowledged that the cause was an act of human error. The Website Management Company has no further or ongoing involvement in the Website.

The timeline of how UK Biobank has investigated this matter, the steps it will take both to mitigate the impact of the data incident and to contact the individuals affected by the data incident, and how it will ensure that such an incident does not occur again, is set out below (with more detail provided in Appendix A):

- 16th June: the unprotected CSV file was taken down within 90 minutes of being alerted to their existence, and have subsequently confirmed there are no other such data on *the Website*;
- 17th June: UK Biobank commenced investigation of the incident to determine the scope of data disclosed, the extent of any accesses, and its root cause. An initial notification was made to the ICO;
- 18th June: in parallel to its own internal security checks of *the Website* and all other UK Biobank IT systems, UK Biobank engaged external IT/security experts (NCC Group) to investigate exactly what had happened and to undertake a full review and penetration testing;
- 28th June: UK Biobank also engaged additional external IT/security experts (IBM) to extend and accelerate the external third-party review;
- 10th July: UK Biobank has submitted this report to the ICO, and UK Biobank will notify the Charity Commission in due course (UK Biobank is a registered charity);
- By 17th July: NCC will provide its full report of the security review and penetration testing of *the Website*, including forensic analysis and audit log recovery;
- By 17th July: UK Biobank's Principal Investigator and CEO, Professor Sir Rory Collins, intends to write to the individuals in whom personal identifiers beyond name were disclosed (of the 6,152 individuals included in the exposed file there are approximately 60 instances where there are insufficient details for a response to be produced) to explain exactly what information about each of them has been disclosed. In case any of them wish to discuss the matter with UK Biobank, UK Biobank's direct contact details will be provided, including the

telephone number of UK Biobank's PRC. The PRC Staff will be fully briefed to field relevant questions and queries and will be able to forward any calls to more senior UK Biobank staff (including the CEO) as required.

- By 20th July: a summary description of the incident will be published on UK Biobank's website and UK Biobank intends to make a copy of this report publicly available;
- By 20th July (or shortly thereafter): IBM will have completed its review and penetration testing of all UK Biobank IT systems and provided a report detailing corrective action (if any) to be taken.

Full Report

This report is structured in the following sections:

- Background to UK Biobank
- How the incident arose;
- The incident itself (as revealed by the investigation);
- The extent of the data involved;
- What steps did UK Biobank take to contain the incident;
- To what extent have individuals being affected;
- Further investigations and measures being taken (to prevent a recurrence).

1. Background information on UK Biobank and processes in place to protect personal data held by UK Biobank

1.1 UK Biobank is a charitable company limited by guarantee established as a medical research resource by the Wellcome Trust and the Medical Research Council.

1.2 UK Biobank involves 503,000 UK volunteer participants who joined the project between 2006-2010 when they answered questions about their lifestyle, had various measurements made and provided blood samples, as well as giving UK Biobank permission to follow their health via their medical records, so that these data could be made available in a properly de-identified form to medical researchers around the world.

1.3 UK Biobank is now the largest intensively characterized health database in the world. It is being used by more than 15,000 scientists globally on more than 1,500 research projects, many of which have already made major contributions to our understanding of the causes of many different diseases (including, most recently, COVID-19). Further information is available at https://www.ukbiobank.ac.uk/.

1.4 UK Biobank takes information security and data protection extremely seriously. UK Biobank is ISO 27001 certified such that UK Biobank's Information Security Management System (ISMS) is audited regularly by external auditors and is subject to external penetration testing – the most recent external penetrance testing took place in June 2019 – to ensure that UK Biobank has the policies and procedures needed to maintain that certification along with the risk controls (legal, physical and technical) necessary for robust IT security management.

1.5 All staff involved in the UK Biobank project receive updated Information Security and GDPR training on a regular basis.

2. UK Biobank's investigation - into the circumstances of the data incident

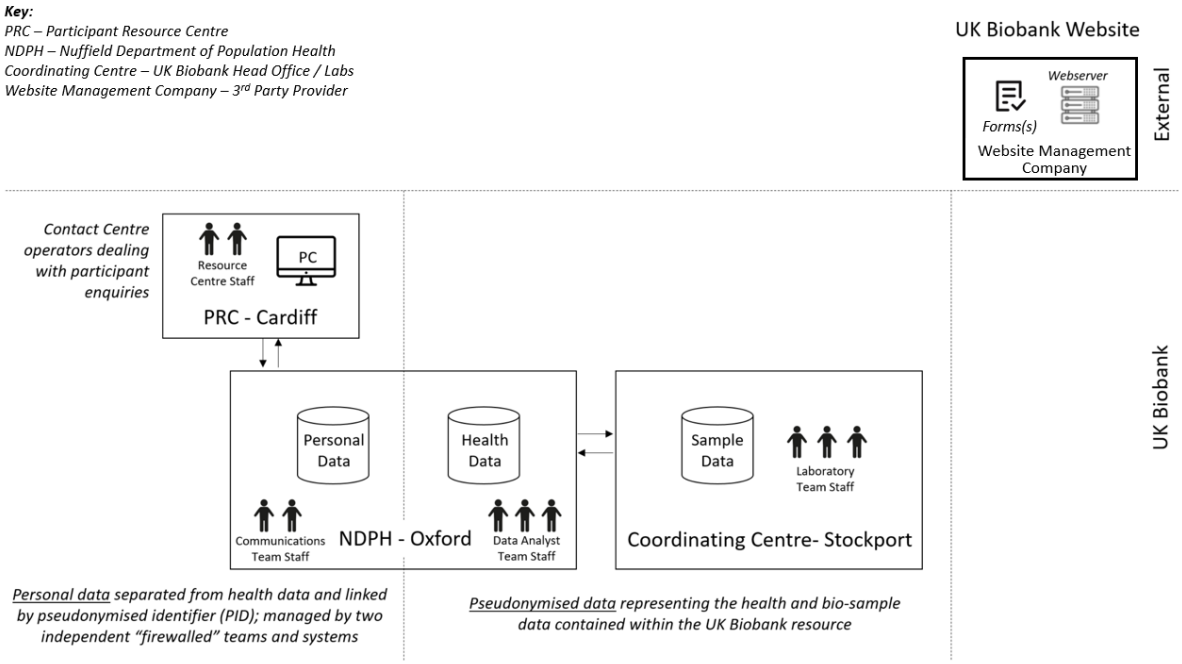
2.1 *The Website* was launched in 2005. Since this time, *the Website* has been developed and managed on UK Biobank's behalf by the Website Management Company. These arrangements were renewed on an annual basis. *The Website* had a substantive rebuild in 2012, based on WordPress (a common website platform at that time), and has been subsequently extended.

2.2 Generic feedback forms are available on *the Website* for visitors (participants, researchers and members of the public) to provide feedback to UK Biobank. The type of feedback provided is predominantly administrative, such as requests to attend events, comments on events and/or

opinions about UK Biobank and its activities. The feedback forms and their contents are stored and processed in a manner which is entirely separate from UK Biobank's internal IT systems (such as the research database).

2.3 Once a feedback form had been submitted, the feedback form data (comprising identifiable data, such as name and email address, and/or free-text comments) were forwarded to the PRC (which is based in Cardiff and is the primary interface between UK Biobank and its participants for all participant queries).

2.4 The aggregated feedback data were stored as a password-protected file and accessible only to the Website Management Company and the UK Biobank Communications Team (which is based in the Nuffield Department of Population Health at the University of Oxford https://www.ndph.ox.ac.uk). A schematic overview of UK Biobank's internal organisational architecture (relevant to this matter) is shown below.



2.5 Access to any personal data held by UK Biobank is limited to those employees where it is strictly necessary for their role. As such, only the Website Management Company, the PRC and the UK Biobank Communications team would have had access to the details contained in the feedback forms.

2.6 The incident was caused by a technical upgrade to *the Website* performed by the Website Management Company in January 2019 when the version of the website platform software was changed from WordPress v4 to WordPress v5.

2.7 During this technical upgrade, the Website Management Company migrated the file containing the feedback form data (comprising identifiable data, such as name and email address, and/or free-text comments) into the new website using an automated plug-in tool.

2.8 During the migration, the password was removed from the file and then exported into an unprotected CSV format file. The unprotected CSV file was then placed in an externally referenceable web folder before using an automated plug-in tool to import the file into the access-controlled database within the updated WordPress v5 platform.

2.9 Although the CSV file was in an externally referenceable web folder (a folder within the WordPress Media Library), it did not appear within the WordPress Admin Dashboard and so was not described within any of the website architecture. Consequently, UK Biobank had no visibility (or knowledge of) of its existence (something which was confirmed by the Website Management Company).

3. UK Biobank's investigation - access to the CSV file

3.1 As soon as the existence of this unprotected CSV file was drawn to UK Biobank's attention (by a bioinformatics company that had located it inadvertently during a Google search on 16th June 2020), the file was removed, and UK Biobank initiated a thorough investigation.

3.2 In terms of assessing the amount of access there has been to this CSV file, *the Website* server only stores the two most recent weeks (the first two weeks of June 2020) of audit logs (on the basis that it is a non-transactional system).

3.3 As a result, UK Biobank commissioned an external forensic investigation by the NCC Group http://www.nccgroup.co./uk/ in order to attempt recovery of deleted audit logs and to determine the likely number of data downloads (see Section 7 below).

3.4 The available weblogs for the 2-week period between 3 June and 16 June 2020 show that there were a number of web access requests to the file during this period. However, it is difficult to determine whether this activity is the result of automated search robot activity or by third-party download. NCC's preliminary findings are that a significant proportion of activity is related to automated search robot activity and their investigation is seeking to further quantify this.

3.5 UK Biobank has been able to recover historic access logs for *the Website* for the period from 7 January 2019 to 22 April 2019 (as these had been archived in a separate area of *the Website* server). They indicate that within this period there were zero access requests for the CSV file (even from search engine robots).

3.6 However, it is clear that sometime between the end of April 2019 and January 2020, the file was accessed not just by search engine indexing robots but also by third parties. This is because a UK Biobank participant contacted UK Biobank in January 2020 about unsolicited email that they had received at an email address which they stated (and this was backed up by a level of detail) had been used exclusively for feedback communication with UK Biobank.

3.7 As the email contact details of this UK Biobank participant were contained in the CSV file, it is reasonable to conclude that this file has been accessed by a third party. Although, UK Biobank's then CIO investigated this report he did not find the cause because, although he did check the security of the UK Biobank research database and IT systems (which were not the source), he did not consider possible disclosure from *the Website*.

4. The extent of the identifiable data that has been disclosed

4.1 The CSV file, which is an aggregation of the individual feedback forms (there are slightly different forms for different events), has been reviewed and categorised by type of individual (i.e. participant, researcher or member of the public) and by type of information (i.e. contact details, date of birth, UK Biobank study participant identifier (*PID*), free text comment).

4.2 The types of data are set out in more detail in Appendix C. Of note, is that are a number of instances (398 to be precise) of the inclusion of Special Category Data (such as participant reference to a self-reported health condition).

5. How UK Biobank discovered the data incident and what immediate steps it took to contain the incident

5.1 The detailed timeline of actions is set out further in Appendix A. As soon as UK Biobank was alerted to the breach on 16 June 2020 – UK Biobank's Data Protection Officer was alerted within 30 minutes of the initial notification – the unprotected CSV file containing the feedback form data was immediately removed within 90 minutes from the initial notification.

5.2 Access to *the Website* feedback forms was disabled, the forms themselves were removed from *the Website*, and copies of any feedback form data deleted from the WordPress platform.

6. Has there been any detriment to the affected individuals as a result of the incident and what steps will we take as regards the individuals affected to mitigate any detriment

6.1 UK Biobank's working assumption is that these data have been accessed by third parties (who have neither the consent nor the authorisation to do so) since UK Biobank is aware of at least one individual who has been contacted via unsolicited email (which email originated from the CSV file) by third parties without their consent.

6.2 However, it is difficult to determine how many individuals have been impacted and it is not clear how much to extrapolate from the fact that only one individual has contacted us to date.

6.3 For 6 UK Biobank participants, the disclosed data does contain sufficient information for an unauthorised external party to sign into the UK Biobank Participant Portal and amend certain of their contact details. UK Biobank has reviewed the activity relating to all 6 such accounts and there does not appear to be any suspicious activity. As a precautionary step, UK Biobank has disabled their logins with a redirection message (should they try to log-in) to contact the Participant Resource Centre, so that they can be issued with new login details.

6.4 After the submission of this report, UK Biobank's CEO will write to all 6,152 affected individuals (save those limited number, 60 or so, without any useable contact details) with a description of the type of information about them, which has been disclosed. In case they wish to discuss the matter with UK Biobank, contact details of the PRC will be provided. The PRC staff will be

fully briefed to address questions and queries from affected individuals and as required, will be able to forward any calls to more senior UK Biobank staff (including the CEO) as required.

6.5 This communication will also set out what UK Biobank has done to prevent any further harm to them taking place and also to identify any possible further detriment (such as unsolicited third-party contacts) so that they can take steps to mitigate such detriment (for example, by blocking unsolicited emails or considering changing their email address if necessary).

6.6 To assist with this communication, UK Biobank has engaged its retained external legal advisors, Eversheds Sutherland LLP, to review and categorise the data and to advise on mitigation and management of the incident. UK Biobank is working closely with Eversheds Sutherland LLP to ensure that our remediation processes are fully in line with the prevailing legal requirements.

6.7 UK Biobank also intends to publish a summary statement on its website – and intends to publish this report as well – explaining what has happened, so that any UK Biobank participant or researcher who has not been affected by the data breach, but is nonetheless concerned, can readily understand what has taken place (and what UK Biobank has done to remedy the situation).

7. What further action is being taken and what measures are being put in place to prevent a similar event occurring in the future

7.1 UK Biobank would like to emphasise that it takes data protection extremely seriously and implements extensive measures throughout its operations to ensure compliance with data protection obligations and requirements. The security of personal data that UK Biobank stores and processes is pivotal to the success of the organisation. This situation, which has revealed deficient processes on *the Website,* means that UK Biobank must be as vigilant in all areas of its activities as it has been to date in the area of its main database and IT systems.

7.2 UK Biobank's internal investigation will continue under the control of the Deputy CEO, and it will make a full report to UK Biobank's Board. This will include a comprehensive review of:

- Management and oversight of external service providers, particularly those who are involved in the processing of any personal data and to ensure that all appropriate GDPR Controller / Processor requirements are in place;
- Interactions between relevant UK Biobank departments to ensure that risk and security are being properly communicated, assessed and managed; and
- Further enhanced training for all staff on information governance and data protection.

7.3 Enquiries with the Website Management Company are ongoing, although any further arrangements with the Website Management Company have been terminated and *the Website* is in the process of being migrated internally.

7.4 In terms of the external NCC investigation, this will be completed with the aim of:

- Confirming the nature and cause of the breach;
- Conducting a full security audit and penetration testing of *the Website*; and

- Undertaking a forensic analysis of the webserver to attempt recovery of additional log files.

7.5 In terms of the external IBM investigation, this will be completed after having:

- Conducted a full security audit (including external and internal penetration testing) of all of UK Biobank IT systems at the Coordinating Centre (Stockport), NDPH (Oxford) and Participant Resource Centre (Cardiff); and
- Conducted a comprehensive search of the dark and deep web (DDW) using search techniques to ascertain whether there is evidence that these feedback form data have been shared.

Appendix A. Detailed timeline and Steps Taken

Date	Description
16/06/2020	15:35 – Email received by the Access team from a 3rd party informing us that they had been able to access a CSV file found through a Google search
	16:01 – Access team leader forwarded the email to the UK Biobank Data Protection Officer (DPO) task id in line with potential data breach procedure
	16:05 – Initial response from DPO task id seeking further information
	16:09 – Example rows from CSV file confirmed identifiable data
	16:10 – Escalated to DPO and CIO for investigation and immediate removal of file
	16:18 – Escalated by the DPO to the Senior Management Team and the Communications team with specific request to remove file immediately from *the Website* and then investigate further
	16:38 – *The Website* hosting provider contacted to request removal and investigation
	16:51 – *The Website* hosting provider confirmed removal of the file
	17:03 – Confirmation of deletion circulated within UK Biobank
	17:16 – Acknowledgement and request to *the Website* hosting provider to provide further information on what had happened and how. Initial internet search to check whether CSV file was being openly hosted
	18:07 – *The Website* hosting provider requested to review web server audit logs to confirm if the file had been remotely accessed
	18:15 – Initial report from *the Website* hosting provider detailing that the file related to a number of "Gravity Forms" (a plugin used to create online web feedback forms) that had been used on the site at various times. The file timestamp was 29th January
	20:19 – *The Website* provider confirmed this date coincided with an upgrade of *the Website* from WordPress 4 to WordPress 5 and must have been the result of temporary data being created by the Website Management Company and not cleaned up afterwards.
17/06/2020	*The Website* hosting provider makes available server logs for review
	Initial review of logs from June 2020 showed that the file had been remotely accessed
	Communications team confirmed all feedback forms had been disabled on *the Website*
	Investigation started to understand how and why the feedback form data had been exposed
	ICO notified of data breach
18/06/2020	Confirmation that submitted feedback form data had been deleted from *the Website* database
	Detailed review of audit logs commenced

	Copy of the deleted file obtained from *the Website* hosting provider
	NCC Group contacted to conduct full penetration test and code-level site review of *the Website*
	Emails relating to WordPress 4 to WordPress 5 upgrade and issues experienced with Gravity Forms migration provided and reviewed
19/06/2020	Full audit log analysis completed; however, audit logs were only available during the period January 2019 to April 2019, and June 3rd to June 17th 2020
	Proposal for penetration test of *the Website* received from NCC
20/06/2020 to 21/06/2020	Initial review of exposed file completed
	Analysis shows records exposed relate up to 6,152 data subjects
22/06/2020	6 of the 6,152 records were identified as containing sufficient information (Name, PID, Date of Birth) that would allow access to the UK Biobank Participant Portal (which allows review and update of contact details, and acceptance to take part in enhancement projects, but no access to other data or functionality).
	Emergency change implemented to revoke any further access to the Participant Portal for these 6 participants (with any further access requests directed to contact the Participant Resource Centre)
23/06/2020	Scope of NCC security review extended to include forensic analysis of webserver image to attempt recovery of overwritten audit logs, plus request for site-wide full penetration testing and security audit
	Details of backups and snapshot available for forensic investigation provided by *the Website* hosting provider
	Content review of imaging sub-site completed: no file containing personal data found
24/06/2020	UK Biobank's CEO instructed the Deputy CEO to take over the investigation of the incident in light of the CIO's failure to investigate the earlier report adequately
	Forensic analysis proposal received and committed
	Request for provision of access to server for imaging
	Attempt to speak with the Director of the Website Management Company that undertook *the Website* upgrade in 2019 (response from the Website Management Company received on 6th July 2020)
25/06/2020	Review of root cause and assessment of Gravity Importer Plugin

	External legal firm (Eversheds) engaged to undertake full review and characterisation of exposed file
26/06/2020	Imaging of webserver complete to allow forensic analysis to commence and penetration testing of *the Website* to start
	NCC unable to commit resources to start on the Friday, and unable to commit resources over the weekend
	Decision to decommission imaging microsite
27/06/2020 to 28/06/2020	Further analysis of access logs for 406 data subjects whose PID had been included in the exposed file; determination of remediation approach to re-issue new PIDs for these participants
	Further proposal sought from IBM
29/06/2020	Scoping call with IBM to progress full site-wide penetration testing and security audit
	Additional assistance provided by IBM Incident Response Team to review steps being taken and any gaps
	IBM X-Force IRIS team commences review of open and dark web for any references to hosting of the UK Biobank exposed file
30/06/2020	Outline proposal for full security audit and penetration testing received; IBM confirms ability to work at risk and work commenced that day
	Initial report from open and dark web review provides reassurance with no evidence of the exposed feedback data being openly shared.

Appendix B. Details of exposed feedback form data within file

Note: CSV stands for Comma Separated Value and is used to store in a text file, with one entry per row e.g.: "John","Smith","Birmingham","B11 1AA","My feedback comment is"

Filename	Data Subjects
The Website general feedback form [1 feedback form]	3731
Request to attend an event (e.g. scientific conference or public meeting) [25 feedback forms]	2304
Request for feedback on frequency and content of UK Biobank updates [1 feedback form]	25
Feedback from researchers for specification of a particular sample assay [1 feedback form]	67
Request for researchers to be included on update communications [1 feedback form]	25
Total Data Subjects	**6152**

Appendix C. Summary characterisation of the data contained within the exposed file

The tabulation below is based on a preliminary analysis of the records. The numbers presented below may change as a result of the validation work, but such changes should not be substantial.

Number of data subject records exposed 6152 (approximately 60 instances of no actual contact details)

of which, there were
4589 participants
1186 researchers
369 member of the public
8 UK Biobank staff (for test purposes)

Inclusion of contact details:
6146 emails
2589 physical addresses
256 telephone numbers

Inclusion of PIDs and/or Date of Birth:
406 participant ID (PID) numbers[2]
46 dates of birth
6 both PID and DOB[3]

Inclusion of free text with health information: 398 contain such information, for example reference to a self-reported health condition (such as diabetes or rheumatoid arthritis)

[2] The PID is an identifying number that each participant is uniquely assigned by UK Biobank; it is used internally as a pseudonymised identifier to link the information UK Biobank holds to each study participant

[3] The combination of these two data items means that a participant is able to sign into a UK Biobank portal used for the purpose of maintaining their contact details (such as email address)