Plan Overview

A Data Management Plan created using DMPonline

Title: HDR UK Gut Reaction Hub

Creator: Neil Walker

Principal Investigator: Prof John Bradley

Data Manager: Neil Walker

Contributor: Jonathan Pilgrim

Affiliation: University of Cambridge

Funder: UKRI Future Leaders Fellowships

Template: UKRI Future Leaders Fellowships Template for a Data Management Plan

Project abstract:

Gut Reaction will create the world’s largest virtual repository of data from people with Inflammatory Bowel Disease (IBD). It will contain information on overlapping participants from 4 sources: the IBD BioResource (34,000 participants); IBD Registry (58,000) and (UK) IBD Genetics Consortium (20,000); with linked data on 7,000 patients from the IBD BioResource for whom detailed hospital records are also held. Researchers can search metadata about the available data sets and apply for access to the subsets of data to support their research, in a variety of secure settings.

ID: 82133

Start date: 01-10-2019

End date: 31-08-2022

Last modified: 20-09-2021

Copyright information:

The above plan creator(s) have agreed that others may use as much of the text of this plan as they would like in their own plans, and customise it as necessary. You do not need to credit the creator(s) as the source of the language used, but using any of the plan's text does not imply that the creator(s) endorse, or have any relationship to, your project or proposal

HDR UK Gut Reaction Hub

0. Proposal name

0. Enter the proposal name

HDR UK Gut Reaction Hub

1. Description of the data

1.1 Type of study

The Hub is collating and intersecting existing data from 3 sources of consented participants - IBD BioResource, IBD Registry and UK IBD Genetics Consortium - and seeking additional linkages to NHS Trust data for a subset of the former.

1.2 Types of data

There are 4 main sources of data in Gut Reaction:

Source institution	Programme	Target participant numbers	Data types	Additional funders
NIHR BioResource - https://bioresource.nihr.ac.uk/	IBD BioResource	34,000	• survey: Health & Lifestyle Questionnaire (H&LQ) • survey: Clinical report Form (CRF) • blood samples: DNA, plasma, sera • genotypic data: SNP chip, imputation data	NIHR, MRC
NHS Trusts	-	8,000	• electronic health records: Diagnostics, Prescriptions, Clinical Notes	NHS England
IBD Registry - https://ibdregistry.org.uk/	COVID-19, Patient Reported Outcome Measures (PROMs)	58,000	• survey: Self-Reported Risks & Outcomes	Crohn's & Colitis UK
Wellcome Sanger Institute - https://www.sanger.ac.uk/	UK IBD Genetics Consortium	30,000	• genotypic data: Whole Genome / Exome Sequencing	Wellcome

During the COVID-19 pandemic, the NIHR BioResource has also requested and received data (*"record linkage"*) on the IBD BioResource cohort:

- summaries of health records: NHS Digital - Hospital Episode Statistics (HES)
  - Admissions
  - Outpatient appointments
  - A&E attendances
  - Diagnostics & treatment coding
- SARS-CoV-2 testing results: Public Health England (PHE)
- intensive care records: Intensive Care National Audit and Research Centre (ICNARC)
- Mortality records: NHS Digital

Additional linkages will be sought.

The IBD Registry also routinely achieves the same NHS Digital linkages.

All the primary sources of data may be shared with other *bona fide* researchers worldwide, although the research setting may be prescribed e.g. to a Trustworthy Research Environment (TRE) under one of the partners' control; linkage data is only available for named use cases.

This Data Management Plan concerns the management of data at the NIHR BioResource, and how the data may be linked between sources: both the IBD Registry and the Wellcome Sanger Institute (and of course, NHS Trusts) have their own policies and procedures to handle primary data sources. See e.g. https://ibdregistry.org.uk/ibd-kpis/ and https://www.sanger.ac.uk/about/who-we-are/research-policies/ .

1.3 Format and scale of the data

Name	Programme	Purpose	Dataset	Scale	Format
RedCap	IBD BioResource	Online survey tool	CRF H&LQ	GB	Relational Database Available as csv
OpenClinica	IBD BioResource	Online clinical trial management tool - *see note below*	CRF H&LQ	GB	Relational Database Available as csv
CiviCRM	IBD BioResource	Recruitment database	Demographics Consent H&LQ subset	GB	Relational Database Available as csv
Microsoft 365	IBD BioResource	Document store	Consent forms	TB	Scanned images, PDFs Not available
i2b2	IBD BioResource	Cohort discovery tool - *snapshot collation of above*	Demographics CRF H&LQ	GB	Relational Database Available as csv
University of Cambridge High Performance Computing Service	IBD BioResource	Big data computing environment	Genetic whole genome, whole exome sequence data	TB	BAMs, CRAMs & VCFs Accessed *in situ*, or via European Genome-Phenome Archive (EGA) managed access repository
proprietary LIMS	IBD BioResource	Samples database	Sample details	GB	Relational Database Available as csv

All data and samples from the IBD BioResource are captured at the time of recruitment, excepting genetic data, which is generated as sufficiently large batches are assembled.

All but the HPC and i2b2 (which is re-built each week as a snapshot) have audit capabilities to allow long-term curation of data. All can be output in non-proprietary formats. In creating a snapshot, i2b2 codes items to clinical ontologies: SNOMED-CT and Human Phenotype Ontology (HPO).

Note on use of OpenClinica:

Some of the H&LQ data for the IBD BioResource is also captured on paper in OCR-ready forms. These are scanned - using software from SRCapture - and loaded into OpenClinica and CiviCRM.

2. Data collection / generation

2.1 Methodologies for data collection / generation

Data collection / generation is ongoing during the period of this grant funding.

The following table uses ontologies from the HDR UK Innovation Gateway - https://www.healthdatagateway.org/ - where metadata concerning these datasets is lodged. A wider metadata dataset conforms to standards used by the UK Data Archive - https://www.data-archive.ac.uk/managing-data/standards-and-procedures/metadata-standards/

Survey data from IBD BioResource and IBD Registry are taken at particular timepoints, especially recruitment, and are managed and curated by their respective data management teams. Genetic data from both IBD BioResource and the Wellcome Sanger Institute are taken through standard QC - the former based on UK Biobank's pipeline as described in https://biobank.ctsu.ox.ac.uk/crystal/crystal/docs/genotyping_qc.pdf , the latter as described in https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7035382/

Source institution	Programme	Data types	Source	Collection Situation
NIHR BioResource	IBD BioResource	• Health & Lifestyle Questionnaire (H&LQ) • Clinical Report (CRF) • Demographics	• PAPER BASED • ELECTRONIC SURVEY	• CLINIC • COMMUNITY • HOME
NIHR BioResource	IBD BioResource	• Genetics (SNP, SNP Imputation) • Samples (DNA, PLASMA, SERUM)	• MACHINE GENERATED • Laboratory Information Management System (LIMS)	• OTHER • CLINIC
NHS Trusts	*Collated by* NIHR BioResource	• Diagnostics • Prescriptions • Clinical Notes	• EHR	• ACCIDENT AND EMERGENCY • OUTPATIENTS • IN-PATIENTS
IBD Registry	COVID-19, Patient Reported Outcome Measures (PROMs	• Self-Reported Risks & Outcomes	• ELECTRONIC SURVEY	• HOME
Wellcome Sanger Institute	UK IBD Genetics Consortium	• whole genome/exome sequencing	• MACHINE GENERATED	• OTHER

2.2 Data quality and standards

We follow the following principles of data quality:

- Accuracy – data should be sufficiently accurate for their intended purposes.
- Validity – data should be recorded and used in compliance with relevant requirements, including the correct application of any rules or definitions.
- Reliability – data should reflect stable and consistent data collection processes across collection points and over time.
- Timeliness - data should be captured as quickly as possible after the event or activity and be available for the intended use quickly and frequently enough to support information needs and to influence service or management decisions.
- Relevance – data should be relevant to the purposes for which they are used. This entails periodic review of requirements to reflect changing needs.
- Completeness – Data requirements should be clearly specified based on the information needs of the organization and data collection processes matched to these requirements.

The following therefore has been considered across all service areas:

- Staff are made aware by their line manager of their responsibilities in relation to data quality.
- Commitment to data quality is clearly stated in job descriptions for all relevant roles.
- Staff have the relevant skills and competencies to fulfil their role in ensuring good quality data.
- Staff receive appropriate training and guidance.
- Training needs are identified through the appraisal process and built into personal development.
- Data quality is a key part of the induction process.
- Commitment to data quality is clearly communicated through the organization

The following are the systems and processes in the bioresource.

All clinical and administrative records must be input into approved systems. The use of any IT system to record service user data, other than those listed in section 1.3 above, is to be avoided.

The data entry systems will be configured, where possible, to ensure that the business processes are followed.

In particular, that the system is configured to follow the participant pathway. The collection and input 'trigger points' will be identified and referenced in training materials. All changes to the clinical and administrative systems will be quality controlled to assure standards concerning the accuracy of recording data.

Fields will be made mandatory where a data item must be collected in all circumstances. The need to make further fields mandatory is kept under review subject to the necessary criteria.

Data Quality is also achieved on the bases on data types:

1. For Demographic and sensitive personal information, All administration and clinical staff are responsible for checking demographic details with the participants and volunteers at all appropriate attendances. Where changes are identified they should follow the NIHR BioResource procedures for ensuring that the change is recorded appropriately. It is vital that all demographic data is recorded accurately, completely and kept as up-to-date as possible.

2. Clinical coding is practiced in all datasets to make sure the information is of the highest standard.

3. The responsibility and ownership of data rests with the system user who must ensure that any errors are corrected promptly at source. Where validation reports are available from systems for use by clinical, managerial and data quality staff, these should be used to check for inaccurate, incomplete or untimely data.

Furthermore, Data Quality incidents are also part of the NIHR BioResource data quality management process. When serious data quality incidents occur or are identified, they should be reported immediately using the organizations incident reporting system and corrective action commenced.

No level of inaccuracy should be viewed as acceptable. Data quality reports are available to help staff identify data quality issues.

Careful monitoring and error correction supports good data quality. However it is more effective and efficient for data to be entered correctly in the first instance. In order to help achieve this, procedures must exist within the BioResource so that staff can be trained and supported in their work.

Situations that could arise due to insufficient information being recorded or inaccuracies in the patient details, would require an incident to be entered in the Incident Log such as:

- Attempts to contact participants / volunteers who are now deceased (this is due to not being notified of the status of the participant but is still an IG incident)
- Duplicate participant records
- System inaccessibility
- Database rollbacks and restores

3. Data management, documentation and curation

3.1 Managing, storing and curating data

Currently there are 5 filestores where IBD BioResource data may reside: at AIMES data centre in Liverpool - https://aimes.uk/; at the University of Cambridge High Performance Computing Service (HPC) - https://www.hpc.cam.ac.uk/; in designated SharePoint sites within Microsoft365; on designated areas of the University of Cambridge Clinical School Computing Service network (CSCS) - https://cscs.medschl.cam.ac.uk/; and on paper in a locked cupboard in a locked office on the Cambridge Biomedical Campus. Of these neither CSCS nor the HPC may be used for identifiable data.

These are the main data ingest sources:

1. Consent/Contact details, filled on paper and sent to the BioResource for data entry. It is stored in the recruitment database (CiviCRM).
2. Consent/Contact details/Health & Lifestyle Questionnaire (H&LQ)/Case Report Form (CRF), on paper and sent to the BioResource for data entry. It is scanned using OCR and stored in CiviCRM. All phenotype information is extracted, cleansed and stored in a separate database known as OpenClinica.
3. H&LQ/CRF is also entered by participants using REDCap, an online survey tool and a holding application/data stored via a file storage at AIMES.
4. Consent/Contact details/CRF, participants recruited and registered at the NHS Trusts are registered on local Electronic Health Records. This data is stored at AIMES. All phenotype is then cleansed and stored in OpenClinica. The participant list is reconciled via email with the relevant NHS Trust.
5. Data about samples collected arrives from the laboratory that receives and processes them - the National Biosample Centre at Milton Keynes - and is stored at AIMES.
6. A project is underway to collect genetic data on all participants in the IBD BioResource. Here samples are sent to Thermofisher in the US, and data returned to the HPC.

Data goes through a life-cycle: its acquisition is recorded; it is (save the big data in HPC) loaded into audited databases; and curated to make data releases. Those releases are also recorded in detail, and through a data access register at https://bioresource.nihr.ac.uk/studies/?speciality=&studytype=Data%2Bonly&tag= . All data sources are backed up. The AIMES data is snapshotted and stored as encrypted files at AWS, in their London, UK data centre; HPC data is uploaded to the Hinxton, UK instance of the European Genome-Phenome Archive - https://ega-archive.org/ - from where it may be accessed under managed access. While outside the scope of this DMP, data for the IBD Registry is also held at AIMES, in an independent tenancy. The IBD Registry routinely uses a Trustworthy Research Environment (TRE) from which data may not be downloaded. The main sources of data ingest are:

1. Patient Reported Outcome Measures (PROMs) received from participants, via REDCap
2. Linked health record data from NHS Digital, based on the record of consenting participants in clinic

Data for the Wellcome Sanger Centre is held in their own data centre and is processed on their own high performance computing cluster - https://www.sanger.ac.uk/group/information-communications-technology/ For Gut Reaction, a copy of standard file formats generated by the Sanger, and post-QC, are held by the NIHR BioResource at the University of Cambridge High Performance Computing Service - https://www.hpc.cam.ac.uk/

Linkage:

Linkage between NIHR BioResource participants and NHS Trusts, is achieved by (securely) reminding recruiting Trusts of the participant identifiers and personal details of their recruits.

Linkage between NIHR BioResource and Wellcome Sanger Centre data is achieved through sharing of identifiers and data, under contract: genetic data is not personal data if it cannot be linked to the person, which makes sharing data easier where a participant has consented to one party (and is known) and not to the other (and is not known).

Linkage between the IBD Registry and NIHR BioResource, is harder, as the data shared would still be personal data when de-identified. Linkage is achieved through a method of comparing hashed personal data before data is released: if the hash does not match, it is not the same person. The hash cannot be reversed to re-discover personal details. This privacy-preserving method has been described widely in e.g. https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-017-0437-1

Linkage between IBD Registry and Wellcome Sanger Centre, where the participant is also not in the NIHR BioResource, is not possible: the latter has insufficient personal details to create a hash, and the former has no genetic data.

3.2 Metadata standards and data documentation

Metadata for the primary sources of data is captured in the HDR UK Innovation Gateway - https://www.healthdatagateway.org/ , in the Gut

Reaction Hub's own collection - https://web.www.healthdatagateway.org/collection/8070361309216243

Additional documentation is held on the Gut Reaction website - https://bioresource.nihr.ac.uk/centres-programmes/ibd-bioresource/gut-reaction-data-sources (temporarily hosted on the NIHR BioResource website, August 2021). This includes PDFs of the data capture forms, data catalogues with data profiling, Venn diagrams to show the overlap between datasets, and usage metrics.

One use of metadata standards of note: the Gut Reaction Cohort Discovery Tool uses data dictionaries to map data into an i2b2 data warehouse - https://www.i2b2.org/ . Data from self-report Health & Lifestyle Questionnaires and clinical Case Report Forms is mapped to SNOMED-CT codes for the purpose of recording diagnoses, procedures and medications. Upper level medication classes are recorded using the Anatomical Therapeutic Chemical (ATC) classification system. Rare Disease abnormalities are coded according to the Human Phenotype Ontology (HPO).

3.3 Data preservation strategy and standards

All 3 of the partners described have long-term aspirations for the data they hold:

1. the NIHR BioResource uses data to invite participants from the IBD BioResource to experimental medicine studies
2. the IBD Registry's core business is around clinical evaluation and audit, and changes in patient treatment and outcomes over time
3. the UK IBD Genetics Consortium is building ever larger cohorts of participants to investigate more fine-grained aspects of disease using more subtle genomic techniques.

Therefore, we assume that data we collect will have long-term value.

For the NIHR BioResource, we protect our day-to-day data holdings in three main ways:

- We follow best technical practice in how we handle information:
  - we encrypt data when we have to move it
  - we keep data in secure data centers – both physically secure against intruders, and electronically secure against hackers
  - we keep personal details separate to other forms of information
  - we monitor who can access what.
- We train our staff carefully, so they know what they need to do to keep information safe. We do this to NHS standards, using NHS training materials
- We check these standards are met.

For long-term preservation, data will be placed in standard formats in managed access repositories. A substantial amount of genetic data is already available (as VCFs and CRAMs) at the EGA - see https://ega-archive.org/dacs/EGAC00001000259

The NIHR BioResource has ethical approval to keep (and therefore allow access to) data for 10 years after the study has finished (to November 2032 in the first instance). Practically, this would involve placing data under the guardianship of Cambridge University Hospitals NHS Foundation Trust (CUH) who are the Data Controller.

4. Data security and confidentiality of potentially disclosive information

4.1 Formal information/data security standards

The NIHR BioResource self-assesses annually as meeting the requirements of the NHS Digital Data Security and Protection Toolkit (Category 3), registration EE133801-NIHR-NBR. The bulk of our personal data is held at AIMES Ltd, which is ISO 27001 and Cyber Essentials Plus certified - see https://aimes.uk/p-accreditation/ . Scanned images of consent forms and similar are held in private areas on Microsoft365, which is ISO 27001 certified. The University of Cambridge HPC - where genetic data sits - is developing a Secure Research Computing Platform with ISO 27001 certification: https://docs.hpc.cam.ac.uk/srcp/index.html . We have a pilot tenancy on this service, sit on the project board that is seeing it through to completion, and will move our genetic data under this service when that becomes possible.

4.2 Main risks to data security

Risks are managed through Data Protection Impact Assessments.

Our current template considers these risks as a starting point:

Risk description	**Risk likelihood** 1: Rare 2: Unlikely 3: Possible 4. Likely 5. Almost certain	**Risk severity** 1. Very low 2. Low 3. Medium 4. High 5. Very high	**Residual risk (before mitigation)**	**How will the likelihood of this risk be mitigated?**	**Risk likelihood (after mitigation)** 1: Rare 2: Unlikely 3: Possible 4. Likely 5. Almost certain	**How will the severity of this risk be mitigated?**	**Risk severity (after mitigation)** 1. Very low 2. Low 3. Medium 4. High 5. Very high	**Residual risk (after mitigation)**
Information: inappropriate protection (lack of encryption etc)	**Possible**	**Very high**	**12**	**Mandating encrypted transfers, as part of data exchange document**	**Rare**	**Encrypt in transit**	**Very low**	**1**

Information: inaccessibility due to encryption/key issues	**Possible**	**Low**	**4**	**Technological management of keys**	**Rare**	**N/A**	**Low**	**2**
Information: loss or theft by employees or contractors at data centre	Possible	Very high	12	ISO27001 controls re staffing at AIMES; staff contracts and IG training at NBR	Rare	N/A	Very High	8
Information: loss or theft by employees or contractors at backup site	Possible	Very high	12	ISO27001 controls re staffing at AWS	Rare	Encryption of offsite backups	High	5
Information: unauthorised disclosure (social engineering, eavesdropping etc)	Unlikely	Very high	10	IG Training, ISO27001 controls re access to premises	Rare	N/A	Very high	8
Information: unauthorised or incomplete changes	Possible	Medium	7	Automation of processes, authentication/authorization	Rare	Automated change reporting	Low	2
Information: risks whilst on mobile equipment	Almost certain	Very high	25	2-factor authentication required for mobile devices	Unlikely	Encryption of staff devices. Use of tracking and “bricking” software for devices used in clinics	Low	3
Information: improper disposal	Rare	Very high	8	ISO27001 controls re disposal	Rare	Encryption of offsite backups	High	5
Information: theft by an application or system	Possible	Very high	12	DPIAs considered on commissioning new applications and systems	Unlikely	ISO27001 controls of outgoing data	Low	3
Information: theft or mis-use by a third party	Unlikely	Very high	10	ISO27001 controls re access	Rare	ISO27001 controls of outgoing data	Low	2
Information: leakage from data centre	Unlikely	Very high	10	Security training of privileged individuals, automation of processes, authentication/ authorization	Rare	N/A	Very high	8
Information: leakage of data by email	Likely	Very high	18	IG training, automation of processes, authentication/ authorization	Rare	N/A	Very high	8
Information: leakage by post or fax	Unlikely	Very high	10	IG training, automation of processes, authentication/ authorization	Rare	N/A	Very high	8
Information: leakage by social media networks	Almost certain	Very high	25	IG training, automation of processes, authentication/ authorization. Website black-listing possible with Sophos firewall at AIMES	Rare	N/A	Very high	8
Information: accidental/malicious deletion	Possible	High	8	IG training, ISO27001 controls re staffing, use of audit in applications	Rare	Application-specific and DR/BC backups	Very low	1
Information: risks from changes to international data protection legislation	Unlikely	Low	3	N/A	Unlikely	N/A	Low	3
Software: use of non-current versions	Unlikely	High	10	ISO27001 controls: patching regime	Rare	N/A	High	5

Information: Unauthorised access	Almost certain	Very high	25	ISO27001 controls, authentication/authorization	Rare	AD and role-based access to applications	Low	2
Media: loss or theft	Unlikely	Very high	10	IG training	Rare	Encryption of removal media	Low	2
Hardware: temporary loss of service	Likely	Very high	18	ISO27001 controls	Unlikely	SLA with AIMES	Low	3

5. Data sharing and access

5.1 Suitability for sharing

All the data partners in Gut Reaction - NIHR BioResource, IBD Registry Ltd, Wellcome Sanger Institute, NHS Trusts - hold confidential personal data. This is not shared, without explicit consent and with regulatory approval.

However, participants *have* consented that their de-identified data may be shared with certain safeguards and aggregate data may be published.

5.2 Discovery by potential users of the research/innovation data

Potential users of the data in the Gut Reaction Hub will be offered different avenues to discover the research/innovation content.

Current routes to discover are:

- Metadata is available through the HDR UK Innovation Gateway - https://www.healthdatagateway.org/ , under the Gut Reaction Hub Collection - https://web.www.healthdatagateway.org/collection/8070361309216243
- Additional documentation is found on the Gut Reaction website at https://bioresource.nihr.ac.uk/centres-programmes/ibd-bioresource/gut-reaction-data-sources (temporary page)
- The existence of the Hub may also be discovered through the NIHR BioResource website - https://bioresource.nihr.ac.uk/centres-programmes/ibd-bioresource/gut-reaction/
- A data descriptor paper, to be published in a peer-reviewed paper, is also underway.

That this data can be requested is described on these sites, and through a dedicated data access page - https://bioresource.nihr.ac.uk/using-our-bioresource/academic-and-clinical-researchers/apply-for-bioresource-data/ (temporary page)

5.3 Governance of access

Since the start of the Gut Reaction Hub, in October 2019, data has been requestable through the pre-existing NIHR BioResource data access method - https://bioresource.nihr.ac.uk/using-our-bioresource/academic-and-clinical-researchers/apply-for-bioresource-data/ . This has a Data Access Committee (DAC) deciding on academic applications, with criteria set out below, with escalation to the NIHR BioResource Steering Committee for contentious applications, and/or applications from industry.

In Autumn 2021 this is being replaced with a Gut-Reaction-specific DAC, which will allow more frequent committee meetings, and much greater patient and public involvement in specific decisions - to date most of the involvement is through Patient Advisory Committee discussion of principles.

The application form follows the "5 Safes" model - https://blog.ukdataservice.ac.uk/access-to-sensitive-data-for-research-the-5-safes/ - establishing that the applicant is a bona fide researcher ("safe person"); with a project in the public interest ("safe project"); with proportionate data ("safe data"); with analysis occurring in a secure environment ("safe setting"); and with non-disclosive and non-harmful outcomes ("safe outputs").

Criteria used are modelled on those of METADAC - https://www.metadac.ac.uk/data-access-committee/application-assessment-criteria/ : g*iven the meeting is quorate and has sufficient representation of different domains (data, IG, ethics):*

- The application has been submitted by *bona fide* researchers with sufficient experience and seniority to carry out the work proposed, or with supervisors with the same
- The form is filled in properly
- We have the data requested
- There is negligible risk that the application will produce information that may allow individual study participants to be identified
- There is no substantive risk that the application might harm individuals in the study, or the study as a whole
- The application does not violate (or potentially violate) any of the consents given by the participants or their guardians
- The application does not violate (or potentially violate) any of the ethical permissions granted to the study from which data or samples are requested
- The application addresses topics that fall within the acknowledged remit of the study, as understood by participants
- We believe the recipients will handle the data appropriately and securely
- There is no substantive risk that the application might upset or alienate study members or of reducing their willingness to continue as participants
- The application includes a good quality plain language summary

- The request is for a reasonable level of data access (Not more than 3 paper's worth/ the level of data requested is justified in the application).

Applications may be approved, rejected, escalated or conditionally approved. The usual reasons for conditional approval are:

- plain language guides written as scientific abstracts
- lack of detail on data security in the proposed analysis setting.

We are developing our own advice on plain language writing - and see https://www.metadac.ac.uk/files/2017/06/v1.0-Plain-language-guidance-for-METADAC-applications.pdf

Plain language summaries are key for 3 reasons:

1. They permit public and patient representatives to understand research proposals
2. They may be posted online - e.g. at https://bioresource.nihr.ac.uk/studies/ - for transparency
3. For the NIHR BioResource, they are submitted annually to the Research Ethics Committee to show use of data under its Research Tissue Bank designation.

If an application is rejected, other than for lack of detail, we also consider whether it addresses topics that fall within the strategic goals of the BioResource. If it does, we will work harder with applicants to refine applications.

Data is held by the NIHR BioResource (IBD BioResource, NHS Trust and UK IBD Genetics Consortium data) and the IBD Registry (their own data).

Genetic data is being exported to the European Genome-Phenome Archive (EGA) - https://ega-archive.org/ - which both acts as a backup site, and a source of managed-access data. The Wellcome Sanger Centre Data Sharing Policy describes this - https://www.sanger.ac.uk/wp-content/uploads/Data_Sharing_Policy_and_Guidelines_July_2018.pdf

The majority of the self-report or clinical data is suitable and is documented to the standard required for the UK Data Archive - https://www.data-archive.ac.uk/ . This archive is suitable for deposition for e.g. a Scientific Data data descriptor article. We have enquired whether the "safeguarding" and "controlled" managed-access options in deposition - see https://ukdataservice.ac.uk/help/deposit-data/faqs-on-depositing-data-with-the-uk-data-service/ - would permit us to continue to run our Gut Reaction DAC, as this will be required by a patient-centric committee.

5.4 The study team's exclusive use of the data

Every dataset has a release cycle, with a plan of release and updates. Most of our datasets are updated quarterly. An example is found below:

Source institution	Data Source	Dataset time lag	Publishing frequency	Follow Up
NIHR BioResource	Case Report Forms	1-2 MONTHS	QUARTERLY	1 - 10 YEARS
NIHR BioResource	Health and Lifestyle Questionnaire	1-2 MONTHS	QUARTERLY	1 - 10 YEARS
Wellcome Sanger Institute	UK IBD Genetics Consortium	6 MONTHS PLUS	IRREGULAR	1 - 10 YEARS

For more information see the Gut Reaction collection at the HDR UK Innovation Gateway https://web.www.healthdatagateway.org/collection/8070361309216243

Both the IBD BioResource (as part of NIHR BioResource) and the UK IBD Genetics Consortium (as hosted at Wellcome Sanger Centre) have their own research programmes, and data may be embargoed for some months. This usually coincides with the QC cycle of the data - i.e. the QC is only complete when enough research has been done to identify the subtle errors and batch effects common in medical research. However, in line with UKRI policy this QC/embargo period cannot extend beyond 12 months.

5.5 Restrictions or delays to sharing, with planned actions to limit such restrictions

Participants in the NIHR BioResource, IBD Registry and UK IBD Genetics Consortium have all consented to the sharing of de-identified data with *bona fide* researchers worldwide, for research in the public interest. For the NIHR BioResource, there is a Participant Privacy Notice available at https://bioresource.nihr.ac.uk/about-us/governance-and-ethics/privacy-notice/ that reminds participants both of what they signed up to, and their rights to withdraw without reason. IBD Registry similarly outline their commitment to privacy - https://ibdregistry.org.uk/privacy-policy/

There are limits to these consents both by expectation - participants would not expect any sharing to be disclosive; and legal - some datasets (such as those received through NHS Trusts) may not be shared beyond a safe setting in the UK under the control of the data recipient, in this case NIHR BioResource.

Data sharing is usually with data at the participant/sample record level. However, NIHR BioResource is currently commissioning a Cohort Discovery Tool in a Trustworthy Research Environment (TRE) that will generate record counts.

The Data Access Application form currently in use by Gut Reaction - linked from https://bioresource.nihr.ac.uk/using-our-bioresource/academic-and-clinical-researchers/apply-for-bioresource-data/ - is clear on IP and copyright, as well as Data Controllership.

5.6 Regulation of responsibilities of users

The Data Access Agreement (DAA) current in use - linked from https://bioresource.nihr.ac.uk/using-our-bioresource/academic-and-clinical-researchers/apply-for-bioresource-data/ - has a full schedule of responsibilities, and is signed by an authorised legal representative of the applying institution.

Data access is not permitted without a signed, approved DAA, unless the need for such has been obviated by a wider contractual agreement.

6. Responsibilities

6. Responsibilities

Gut Reaction has a Chief Data Officer - Neil Walker. He is a University of Cambridge employee, working for both Gut Reaction and NIHR BioResource, where he is IT Lead and Information Governance Responsible Officer.
With the assistance of 2 senior managers, Neil oversees a team of 12 data scientists of various specialisms (including 3 dedicated solely to Gut Reaction) and 5 IT staff. These are responsible - for data held at the NIHR BioResource - for:

- study-wide data management
- metadata creation
- data security
- quality assurance of data.

The data security aspect is substantially supported by infrastructure suppliers, particularly AIMES Ltd, a secure data centre based in Liverpool, UK.

7. Relevant policies

7. Relevant institutional, departmental or study policies on data sharing and data security

Policy	URL or Reference
Data Management Policy & Procedures	This document represents the policy. Procedures are not public.
Data Security Policy	NIHR BioResource policy on IG - linked from https://bioresource.nihr.ac.uk/about-us/governance-and-ethics/
Data Sharing Policy	The NIHR BioResource is bound by its funder's policy - https://www.nihr.ac.uk/documents/nihr-position-on-the-sharing-of-research-data/12253 . Our implementation of this policy is laid out in this DMP
Institutional Information Policy	NIHR BioResource policy on IG - linked from https://bioresource.nihr.ac.uk/about-us/governance-and-ethics/
Other	
GDPR	NIHR BioResource policy on GDPR: https://bioresource.nihr.ac.uk/about-us/governance-and-ethics/gdpr/

8. Author and contact details

8. Author of this Data Management Plan (Name) and, if different to that of the Principal Investigator, their telephone & email contact details

The person leading the team responsible for this Data Management Plan, is Neil Walker, Chief Data Officer of the HDR UK Gut Reaction Hub.
Email: neil.walker@bioresource.nihr.ac.uk
Tel: +44 1223 254906 (although, during the pandemic, this is not a reliable contact method)