Plan Overview

A Data Management Plan created using DMPonline

Title: REGRISK

Creator: Ana-Teresa Maia

Principal Investigator: Ana-Teresa Maia

Data Manager: Ana-Teresa Maia

Affiliation: Other

Funder: European Commission

Template: DCC Template

ORCID iD: 0000-0002-0454-9207

Project abstract:

An effective population breast cancer (BC) risk screening will decrease the over 350,000 new annual cases in Europe but requires full knowledge of the genetic and environmental risk factors. Despite the progress already achieved, half of the familial cases are still unexplained by known genetic risk factors, and new study approaches are needed urgently. Previous BC GWAS and follow-up studies suggest that most of the known genetic risk is due to variants regulating the level of expression of target genes in cis. Therefore, we hypothesise that the most efficient approach to tackle BC's missing heritability is to focus risk studies on variants that show greater cis-regulatory potential.

The central goal of REGRISK is to comprehensively identify the cis-regulatory variants (rVars) associated with BC risk, both in high-risk families and sporadic cases. We will achieve this by integrating allelic expression (AE) analysis, the most robust method to detect and map rVars, into case-control studies. We will analyse individuals from high-risk families, carrying or not a known pathogenic variant in breast cancer genes, and sporadic cases. Due to the increased statistical power of using a continuous variable like AE ratios in association studies, the number of samples required is an order smaller than for the current genotype-based GWAS, facilitating the collection of an informative set of samples. Furthermore, the consortium team members have pioneered the discovery and characterisation of genetic risk factors for BC, particularly those with cis-regulatory effects for which they have developed mathematical models to map causal rVars.

REGRISK results will contribute to an improved understanding of the biological mechanisms underlying breast cancer risk and will provide risk biomarkers with great potential to shift the paradigm of genetic risk testing for breast cancer, while also providing direct answers to the families enrolled.

ID: 41507

Start date: 01-01-2023

End date: 31-12-2027

Last modified: 21-04-2022

Copyright information:

REGRISK

Data Collection

What data will you collect or create?

This is the data we will work with, some will be generated in the scope of this project, other is already available from partner SFC.

Type of Data	Format	Size	Origin	Existing
RNAseq	bam	4TB	UALG	No
DNAseq	bam	4TB	UALG	No
microarray	IDAT	1TB	IC	No
clinical	csv	1GB	IC	Yes

Data Volume:
Note what volume of data you will create in MB/GB/TB. Indicate the proportions of raw data, processed data, and other secondary outputs (e.g., reports).
Consider the implications of data volumes in terms of storage, access and preservation. Do you need to include additional costs?
Consider whether the scale of the data will pose challenges when sharing or transferring data between sites; if so, how will you address these challenges?
Data Format:
Clearly note what format(s) your data will be in, e.g., plain text (.txt), comma-separated values (.csv), geo-referenced TIFF (.tif, .tfw).
Explain why you have chosen certain formats. Decisions may be based on staff expertise, a preference for open formats, the standards accepted by data centres or widespread usage within a given community.
Using standardised, interchangeable or open formats ensures the long-term usability of data; these are recommended for sharing and archiving.
See UK Data Service guidance on recommended formats or DataONE Best Practices for file formats.

Data Description:
Give a summary of the data you will collect or create, noting the content, coverage and data type, e.g., tabular data, survey data, experimental measurements, models, software, audiovisual data, physical samples, etc.
Consider how your data could complement and integrate with existing data, or whether there are any existing data or methods that you could reuse.
Indicate which data are of long-term value and should be shared and/or preserved.
If purchasing or reusing existing data, explain how issues such as copyright and IPR have been addressed. You should aim to minimise any restrictions on the reuse (and subsequent sharing) of third-party data.

How will the data be collected or created?

Patient samples will be processed and RNAseq, DNAseq and Oncoarrays will be carried out at UALG and IC. Following X&Y protocol.

Data Quality:
Outline how the data will be collected and processed. This should cover relevant standards or methods, quality assurance and data organisation.
Indicate how the data will be organised during the project, mentioning, e.g., naming conventions, version control and folder structures. Consistent, well-ordered research data will be easier to find, understand and reuse.
Explain how the consistency and quality of data collection will be controlled and documented. This may include processes such as calibration, repeat samples or measurements, standardised data capture, data entry validation, peer review of data or representation with controlled vocabularies.
See the DataOne Best Practices for data quality.

Documentation and Metadata

What documentation and metadata will accompany the data?

Clinical metadata

- patient identifier
- genome assembly version
- chr
- position
- consequence
- disease status
- family ID

(describe below)

Indicate standards for originating metadata

Metadata & Documentation:
What metadata will be provided to help others identify and discover the data?
Researchers are strongly encouraged to use community metadata standards where these are in place. The Research Data Alliance offers a Directory of Metadata Standards. Data repositories may also provide guidance about appropriate metadata standards.
Consider what other documentation is needed to enable reuse. This may include information on the methodology used to collect the data, analytical and procedural information, definitions of variables, units of measurement, any assumptions made, the format and file type of the data and software used to collect and/or process the data.
Consider how you will capture this information and where it will be recorded, e.g., in a database with links to each item, in a 'readme' text file, in file headers, etc.

Ethics and Legal Compliance

How will you manage any ethical issues?

Use of all samples have been approved by Ethical Board of Hospitals X&Y
Questions to consider:
Have you gained consent for data preservation and sharing?
How will you protect the identity of participants if required? e.g. via anonymisation
How will sensitive data be handled to ensure it is stored and transferred securely?
Guidance:
Ethical issues affect how you store data, who can see/use it and how long it is kept. Managing ethical concerns may include: anonymisation of data; referral to departmental or institutional ethics committees; and formal consent agreements. You should show that you are aware of any issues and have planned accordingly. If you are carrying out research involving human participants, you must also ensure that consent is requested to allow data to be shared and reused.

Ethics:
nvestigators carrying out research involving human participants should request consent to preserve and share the data. Do not just ask for permission to use the data in your study or make unnecessary promises to delete it at the end.
Consider how you will protect the identity of participants, e.g., via anonymisation or using managed access procedures.
Ethical issues may affect how you store and transfer data, who can see/use it and how long it is kept. You should demonstrate that you are aware of this and have planned accordingly.
See UK Data Service guidance on consent for data sharing.
See ICPSR approach to confidentiality and Health Insurance Portability and Accountability Act (HIPAA) regulations for health research.

How will you manage copyright and Intellectual Property Rights (IPR) issues?

IPR will be dealt with by the participating institutions, which have dedicated departments for IP management.

Questions to consider:
Who owns the data?
How will the data be licensed for reuse?
Are there any restrictions on the reuse of third-party data?
Will data sharing be postponed / restricted e.g. to publish or seek patents?
Guidance:
State who will own the copyright and IPR of any data that you will collect or create, along with the licence(s) for its use and reuse. For multi-partner projects, IPR ownership may be worth covering in a consortium agreement. Consider any relevant funder, institutional, departmental or group policies on copyright or IPR. Also consider permissions to reuse third-party data and any restrictions needed on data sharing.

State who will own the copyright and IPR of any existing data as well as new data that you will generate. For multi-partner projects, IPR ownership should be covered in the consortium agreement.
Outline any restrictions needed on data sharing, e.g., to protect proprietary or patentable data.
Explain how the data will be licensed for reuse. See the DCC guide on How to license research data and EUDAT's data and software licensing wizard.

Storage and Backup

How will the data be stored and backed up during the research?

Data will be stored at local and offsite servers, for at least 10 years
servers are organised at redundancy level 10

Questions to consider:
Do you have sufficient storage or will you need to include charges for additional services?
How will the data be backed up?
Who will be responsible for backup and recovery?
How will the data be recovered in the event of an incident?
Guidance:
State how often the data will be backed up and to which locations. How many copies are being made? Storing data on laptops, computer hard drives or external storage devices alone is very risky. The use of robust, managed storage provided by university IT teams is preferable. Similarly, it is normally better to use automatic backup services provided by IT Services than rely on manual processes. If you choose to use a third-party service, you should ensure that this does not conflict with any funder, institutional, departmental or group policies, for example in terms of the legal jurisdiction in which data are held or the protection of sensitive data.

Describe where the data will be stored and backed up during the course of research activities. This may vary if you are doing fieldwork or working across multiple sites so explain each procedure.
Identify who will be responsible for backup and how often this will be performed. The use of robust, managed storage with automatic backup, for example, that provided by university IT teams, is preferable. Storing data on laptops, computer hard drives or external storage devices alone is very risky.
See UK Data Service Guidance on data storage or DataONE Best Practices for storage.
Also consider data security, particularly if your data is sensitive e.g., detailed personal data, politically sensitive information or trade secrets. Note the main risks and how these will be managed. Also note whether any institutional data security policies are in place.
Identify any formal standards that you will comply with, e.g., ISO 27001. See the DCC Briefing Paper on Information Security Management - ISO 27000 and UK Data Service guidance on data security.

How will you manage access and security?

Participating institutions have dedicated cyber security plans which we will follow

Questions to consider:
What are the risks to data security and how will these be managed?
How will you control access to keep the data secure?
How will you ensure that collaborators can access your data securely?
If creating or collecting data in the field how will you ensure its safe transfer into your main secured systems?
Guidance:
If your data is confidential (e.g. personal data not already in the public domain, confidential information or trade secrets), you should outline any appropriate security measures and note any formal standards that you will comply with e.g. ISO 27001."

Describe where the data will be stored and backed up during the course of research activities. This may vary if you are doing fieldwork or working across multiple sites so explain each procedure.
Identify who will be responsible for backup and how often this will be performed. The use of robust, managed storage with automatic backup, for example, that provided by university IT teams, is preferable. Storing data on laptops, computer hard drives or external storage devices alone is very risky.
See UK Data Service Guidance on data storage or DataONE Best Practices for storage.
Also consider data security, particularly if your data is sensitive e.g., detailed personal data, politically sensitive information or trade secrets. Note the main risks and how these will be managed. Also note whether any institutional data security policies are in place.
Identify any formal standards that you will comply with, e.g., ISO 27001. See the DCC Briefing Paper on Information Security Management - ISO 27000 and UK Data Service guidance on data security.

Selection and Preservation

Which data are of long-term value and should be retained, shared, and/or preserved?

Questions to consider:
What data must be retained/destroyed for contractual, legal, or regulatory purposes?
How will you decide what other data to keep?
What are the foreseeable research uses for the data?
How long will the data be retained and preserved?
Guidance:
Consider how the data may be reused e.g. to validate your research findings, conduct new studies, or for teaching. Decide which data to keep and for how long. This could be based on any obligations to retain certain data, the potential reuse value, what is economically viable to keep, and any additional effort required to prepare the data for data sharing and preservation. Remember to consider any additional effort required to prepare the data for sharing and preservation, such as changing file formats.

What is the long-term preservation plan for the dataset?

Questions to consider:
Where e.g. in which repository or archive will the data be held?
What costs if any will your selected data repository or archive charge?
Have you costed in time and effort to prepare the data for sharing / preservation?
Guidance:
Consider how datasets that have long-term value will be preserved and curated beyond the lifetime of the grant. Also outline the plans for preparing and documenting data for sharing and archiving. If you do not propose to use an established repository, the data management plan should demonstrate that resources and systems will be in place to enable the data to be curated effectively beyond the lifetime of the grant.
Where will the data be deposited? If you do not propose to use an established repository, the data management plan should demonstrate that the data can be curated effectively beyond the lifetime of the grant.
It helps to show that you have consulted with the repository to understand their policies and procedures, including any metadata standards, and costs involved.
An international list of data repositories is available via re3data and some universities or publishers provide lists of recommendations e.g., PLOS ONE recommended repositories.

Outline the plans for data sharing and preservation - how long will the data be retained and where will it be archived? Will additional resources be needed to prepare data for deposit or meet any charges from data repositories?
See the DCC guide: How to appraise and select research data for curation or DataONE Best Practices: Identifying data with long-term value.

Data Sharing

How will you share the data?

RNAseq data will be deposited onto EGA, code will be made available thorugh github, reserach outuputs will be available open source via biorxiv, figshare

***

Questions to consider:
How will potential users find out about your data?
With whom will you share the data, and under what conditions?
Will you share data via a repository, handle requests directly or use another mechanism?
When will you make the data available?
Will you pursue getting a persistent identifier for your data?
Guidance:
Consider where, how, and to whom data with acknowledged long-term value should be made available. The methods used to share

data will be dependent on a number of factors such as the type, size, complexity and sensitivity of data. If possible, mention earlier examples to show a track record of effective data sharing. Consider how people might acknowledge the reuse of your data.

How will you share the data e.g. deposit in a data repository, use a secure data service, handle data requests directly or use another mechanism? The methods used will depend on a number of factors such as the type, size, complexity and sensitivity of the data.
When will you make the data available? Research funders expect timely release. They typically allow embargoes but not prolonged exclusive use.
Who will be able to use your data? If you need to restrict access to certain communities or apply data sharing agreements, explain why.
Consider strategies to minimise restrictions on sharing. These may include anonymising or aggregating data, gaining participant consent for data sharing, gaining copyright permissions, and agreeing a limited embargo period.
How might your data be reused in other contexts? Where there is potential for reuse, you should use standards and formats that facilitate this, and ensure that appropriate metadata is available online so your data can be discovered. Persistent identifiers should be applied so people can reliably and efficiently find your data. They also help you to track citations and reuse.

Are any restrictions on data sharing required?

Questions to consider:

What action will you take to overcome or minimise restrictions?
For how long do you need exclusive use of the data and why?
Will a data sharing agreement (or equivalent) be required?
Guidance:

Outline any expected difficulties in sharing data with acknowledged long-term value, along with causes and possible measures to overcome these. Restrictions may be due to confidentiality, lack of consent agreements or IPR, for example. Consider whether a non-disclosure agreement would give sufficient protection for confidential data.

How will you share the data e.g. deposit in a data repository, use a secure data service, handle data requests directly or use another mechanism? The methods used will depend on a number of factors such as the type, size, complexity and sensitivity of the data.
When will you make the data available? Research funders expect timely release. They typically allow embargoes but not prolonged exclusive use.
Who will be able to use your data? If you need to restrict access to certain communities or apply data sharing agreements, explain why.
Consider strategies to minimise restrictions on sharing. These may include anonymising or aggregating data, gaining participant consent for data sharing, gaining copyright permissions, and agreeing a limited embargo period.
How might your data be reused in other contexts? Where there is potential for reuse, you should use standards and formats that facilitate this, and ensure that appropriate metadata is available online so your data can be discovered. Persistent identifiers should be applied so people can reliably and efficiently find your data. They also help you to track citations and reuse.

Responsibilities and Resources

Who will be responsible for data management?

Questions to consider:

Who is responsible for implementing the DMP, and ensuring it is reviewed and revised?
Who will be responsible for each data management activity?
How will responsibilities be split across partner sites in collaborative research projects?
Will data ownership and responsibilities for RDM be part of any consortium agreement or contract agreed between partners?
Guidance:

Outline the roles and responsibilities for all activities e.g. data capture, metadata production, data quality, storage and backup, data archiving & data sharing. Consider who will be responsible for ensuring relevant policies will be respected. Individuals should be named where possible.

Outline the roles and responsibilities for all activities, e.g., data capture, metadata production, data quality, storage and backup, data archiving & data sharing. Individuals should be named where possible.
For collaborative projects you should explain the coordination of data management responsibilities across partners.
See UK Data Service guidance on data management roles and responsibilities or DataONE Best Practices: Define roles and assign responsibilities for data management.

What resources will you require to deliver your plan?

Questions to consider:

Is additional specialist expertise (or training for existing staff) required?
Do you require hardware or software which is additional or exceptional to existing institutional provision?
Will charges be applied by data repositories?

Guidance:

Carefully consider any resources needed to deliver the plan, e.g. software, hardware, technical expertise, etc. Where dedicated resources are needed, these should be outlined and justified.

Carefully consider and justify any resources needed to deliver the plan. These may include storage costs, hardware, staff time, costs of preparing data for deposit and repository charges.
Outline any relevant technical expertise, support and training that is likely to be required and how it will be acquired.
If you are not depositing in a data repository, ensure you have appropriate resources and systems in place to share and preserve the data. See UK Data Service guidance on costing data management.