**Rare Kidney Disease Biobank Data Management Plan**

# **A. Description of Rare Kidney Disease (RKD) Biobank**

Anti-Neutrophil Cytoplasmic Antibody (ANCA)-Associated Vasculitis (AAV) is a rare autoimmune disease that results in rapidly progressive kidney impairment, in addition to immune-mediated destruction of other organs. This affects about 20 per million population per year, as opposed to affecting a handful of people in the country, so is considered a “common-rare disease”. It causes severe multi-organ dysfunction, including irreversible kidney failure, lung haemorrhage, stroke and sino-nasal destruction. The prevalence is 300/million and hence can only be studied by a coordinated international network of centres. The Vasculitis Ireland Network was created in 2014 to harmonise care for vasculitis patients across Ireland. It is a member of ERN-RITA, the European Reference Network for rare immune disorders. Documenting and reporting on clinical outcomes and benchmarking against international norms using a dedicated registry are key components of ERN membership.

The primary aims of this initiative are to address fundamental questions about vasculitis epidemiology, facilitate conduct of phase II/III interventional studies by allowing easy identification of a suitable cohort, monitor use of novel biologic agents, compile sufficiently large cohorts to study immunogenetics, rapidly assess the clinical utility of new biomarkers for development in clinical trials as surrogate end-points, and to characterise ‘difficult to define’ disease subgroups including ANCA negative vasculitis and polyarteritis overlap syndromes.

## A.1 DeCOmPRESS Sub-Study

Coronavirus disease 2019 (COVID-19) is an infectious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The DeCOmPRESS study will rapidly inform management of immunosuppressed patients who contract COVID-19 by comprehensively profiling immunological manifestations and defining the natural history and of the disease in these patients. This project aims to determine if COVID-19 is more or less severe in these immunosuppressed patients. We aim to determine whether immunosuppressant therapy for chronic autoimmune disease protects against the cytokine storm associated with COVID-19 and reduces the severity of the clinical syndrome, thereby paradoxically improving rather than worsening clinical outcome.

We will obtain a granular clinical dataset from RKD recruits contracting COVID-19, which will be linked to an existing clinical phenotype and blood samples analysed by flow cytometry and ELISA to define the immunophenotype and cytokine profile. Importantly, use of a FAIR COVID-19 dataset (designed to be interoperable with international data collection initiatives) and deposition of data in an open science repository will allow ready integration with other studies to maximise the impact of this project.

We will incorporate a broad range of potential data streams, including: the Rare Kidney Disease registry; immunophenotyping and serum cytokine data from the Central Pathology Laboratory (CPL) at St. James’ Hospital, clinical data derived from the registry database, and patient-derived data streams using smart phone and wearable technologies (if the participant has also consented to join the aligned AVERT study <https://www.tcd.ie/medicine/thkc/avert/>).

# **B. RKD Biobank Information Governance Summary**



**Figure 1. Summary of key sites involved in the RKD registry and biobank.**

This data management policy will be followed by all collaborators on the RKD project and complemented by standard operating rules and codes of practice for the diverse research and technical actors in the project. It should be read in conjunction with the RKD data protection impact assessment. A uniform base standard for information security will apply across the project (designed to interface with the existing policies and practices at Trinity College Dublin [TCD]), which will provide a consistent assurance of privacy protection. TCD is the **data controller** of RKD registry and biobank data. The relevant hospitals are the **data controllers** of clinical data before it is pseudonymised and entered into the registry database. The basic RKD data governance principles are summarised in Table 1.

**Table 1:** RKD Data Governance Principles

|  |  |
| --- | --- |
| Integrity | We will practice integrity with their dealings with one another; we will be truthful and forthcoming when discussing constraints, options, and impacts for data-related decisions. |
| Transparency | Data Governance and Stewardship processes will exhibit transparency; it should be clear to all participants and auditors how and when data-related decisions and controls were introduced into the processes. |
| Auditability | Data-related decisions, processes, and controls subject to Data Governance will be auditable; they will be accompanied by documentation to support compliance-based and operational auditing requirements. |
| Accountability | Data Governance will define accountabilities for cross-functional data-related decisions, processes, and controls. |
| Stewardship | Data Governance will define accountabilities for stewardship activities that are the responsibilities of individual contributors, as well as accountabilities for groups of Data Stewards. |
| Checks-and-Balances | Data Governance will define accountabilities in a manner that introduces checks-and-balances between recruitment and analysis teams as well as between those who create/collect information, those who manage it, those who use it, and those who introduce standards and compliance requirements. |
| Standardization | Data Governance will introduce and support standardisation of data to maximise potential for sharing. |
| Change Management | Data Governance will support proactive and reactive Change Management activities for reference data values and the structure/use of master data and metadata. This will allow us to track reliably where changes were made. |

All coded experimental data pertaining to the DeCOmPRESS sub-study will be stored on TCD’s Microsoft OneDrive service. OneDrive is Microsoft's cloud-based file storage service, which allows syncing and sharing of files between computers and mobile devices, and is the cloud computing service of choice of TCD. These files can be accessed from anywhere you have an internet connection. Data hosted on OneDrive will be securely hosted by Microsoft in Europe in compliance with relevant legislation (see TCD Data Protection Procedural Guidelines, accessible at: <https://www.tcd.ie/info_compliance/data-protection/policy/>). TCD is the data controller of DeCOmPRESS data.

At times, cross-centre research collaborations will require data sharing, potentially between European countries. RKD data management and sharing policies will be applied to such data transfers, which are detailed in the data sharing policy.

# **C. Documentation and Metadata**

Each data document generated as part of the RKD study (see [Section D.1](#_D.1_Types_of)) will include a metadata descriptor. Metadata and data field descriptors for REDCap data exports are available and have already been incorporated into the RDF uplift process, where this approach to data integration is employed. Benefits of this approach include data storage in an interoperable format, quicker and more intuitive querying of the data, consistent combination of diverse data sources, and ease of reporting.

# **D. Data Collection**

## D.1 Types of Data Collected and Created

All data will be collected according to the FAIR data principles; findability, accessibility, interoperability, and reusability. Data are pseudonymised by use of the **RKD Main Study ID** primary identifier; this is assigned as the next available ID on the REDCap database. Access to relevant data sources will be controlled by the PI. The key data files generated are:

* A site **recruitment log,** which incorporates identifiable data that allows local re-identification of recruits.
* Pseudonymised **clinical characteristics** of patients and controls. These are stored in the central RKD REDCap database.
* Pseudonymised **experimental data**, e.g. immunophenotyping and serum cytokine results from CPL analysis, biomarker data, flow cytometry data.
* Pseudonymised **biobank data**, describing the provenance, status and location of biobank aliquots. These are stored in a Freezerworks database.
* Uplifted Resource Description Framework (RDF) data, comprising **fused clinical and experimental data** and held in a dedicated triplestore database.

The RKD registry database is hosted by TCD IT service providers, as its security and access permissions are managed by them. This database is periodically backed up on an external third-party server located in Dublin.

The DeCOmPRESS sub-study also incorporates a data stream from the patientMpower vasculitis support app in those participants who have consented to inclusion in the aligned AVERT project (DMP for this study located here: <https://www.tcd.ie/medicine/thkc/assets/pdf/Data%20Management%20Plan_v1.1.pdf>). As a sample sub-study, the data streams for DeCOmPRESS are summarised in Figure 2.



**Figure 2:** Overview of Types of Data Collected in the DeCOmPRESS sub-study

### D.1.1 Patient Recruitment Log

These logs are maintained on each hospital site in a firewall-protected folder on the hospital server. They are held on excel spreadsheets and include the following information: RKD ID, consent version, date of recruitment and identifiable parameters (name, DOB, MRN, contact details). It is accessible only by local study PI and research nurse.

DeCOmPRESS sub-study subjects are all current recruits of the RKD Biobank. Upon contracting COVID-19, RKD recruits will be recognised by their local medical team and their RKD ID registered with the DeCOmPRESS research group, who will maintain a dedicated log of (pseudonymised) study recruits.

### D.1.2 Patient Clinical Data – REDCap

REDCap is a secure electronic data capture (EDC) web platform for building and managing online databases and surveys. Pseudonymised clinical data for all encounters with RKD recruits are recorded in REDCap project “Rare Kidney Disease – RIT Production”. The data dictionary is included in appendix 1. This database operates on multiple-access levels and only the PI and RKD Research Nurse have complete access to the database and permission to modify. Access to patient clinical records is available to the medical team caring for the patient and local study site managers. The RKD registry database is hosted by TCD Research-IT service providers (TCHPC), and its security and access permissions are managed by them. The database is protected behind host and institutional firewalls. This database is backed up daily on an external third-party server located in Dublin.

### D.1.3 Experimental Data

TCD researchers who generate experimental data deriving from analysis of RKD samples or data store this dedicated TCD SharePoint folders, which is the preferred storage solution of the university. Each researcher has their own access-controlled folder. These data include spreadsheets (excel, CSV, .xlsx), Graphpad Prism files (.pzfx), flow cytometry files (.fcs), word documents, image files (.jpeg, .png, .ndpi), PowerPoint files (.pptx), transcriptomic data files (.cel). Sample analysis results from patient samples analysed by CPL will be transferred to the designated DeCOmPRESS OneDrive folder described in [Section G.3.3](#_G.3.3_Immunophenotyping,_Serum). Access to this folder is reserved for the PI, PM, CPL Chief Medical Scientist, and designated CPL Laboratory Technicians. Results will be added as they become available and version control managed by the DeCOmPRESS PM. Coded biological samples are processed and stored temporarily at the CPL Bioresource using industry standard TelePath software. Following batched multiplex analysis, coded serum sample aliquots will be stored in the RKD biobank using standard RKD protocols. Archived RKD Biobank samples and freshly obtained serum will be used for serology analysis.

### D.1.4 Biobank Sample Data

When sample aliquots are archived in the biobank freezer, the provenance data are recorded in the RKD Freezerworks database, which is accessible to the PI, biobank manager and selected delegates. These data include arm to freezer time, sample transport conditions, sample processor identity, aliquot volume, presence/absence of protease inhibitor, freeze-thaw cycles and an audit of sample access. These data are pseudonymised.

### D.1.5 Fused Experimental/Clinical Data

Data integrated into ADAPT’s analysis networks will be stored using a Resource Description Framework (RDF) model to facilitate analysis and data sharing. The RDF model stores data in a “triplestore” database. This is a purpose-built database for the storage and retrieval of triples through semantic queries. Semantic triples codify statements about the data in the form of subject–predicate–object expressions. The RDF model facilitates enhanced data analysis and sharing. Examples where this approach is taken include the AVERT study (where RKD clinical data are fused with environmental and app data) and the DeCOmPRESS sub-study (where experimental, clinical and app data are integrated). All such data are coded using the RKD ID.

# **E. Responsibilities and Resources in the RKD Biobank**



**Figure 3:** Overview of RKD policies and responsibilities

## E.1 Data Management

The HRB’s “Guidance for Health Researchers” (accessible at: https://www.hrb.ie/funding/gdpr-guidance-for-researchers/) advises research organisations and researchers on data protection obligations under GDPR, the Data Protection Act 2018 and the new Health Research Regulations 2018. As per these guidelines any project processing personal data for the purposes of health research must assign a data controller and a data processor. There remains both individual and organisational responsibility to comply with these regulations. All members of the study team (and anyone with access to the data) must act responsibly and will be trained in Data Governance principles. Inclusion of patient representatives in the steering committee will ensure that the best interests of patients are adhered to.

A data controller is the individual (or entity) who controls and is responsible for the collection, storage and/or use of personal information, and determines the purposes for which, and the way in which, personal data is processed. GDPR Article 5(2) requires that the data controller be responsible for, and be able to demonstrate, compliance with the principles of data protection. The RKD data controller is TCD, acting through the RKD PI (Prof. Mark Little).

TCD researchers will act as the initial Data Processors and will be responsible for managing access to the relevant data files for the appropriate stakeholders. Additional project members will be assigned Data Processor status as novel analyses are required. A data processor may hold or process personal data on behalf of the data controller but does not exercise responsibility for or control over the personal data. The Data Processors will immediately notify the controller of any data breaches. If RKD data are shared with third parties, these will then become data controllers of the received data. These interactions will be governed by formal material and data sharing agreements.

## E.2 Data Erasure Requests

Patients are free to deny the use of their samples for this project and to withdraw from the study at any stage. In accordance with GDPR legislation, if a participant requests for their data to be erased, this will be managed by the Data Controller (PI). This request may be following direct contact with medical or research teams or the RKD Research Nurse. A log of such requests will be maintained by the Biobank manager. Upon receipt of such a request, all RKD study ID will be deleted from the appropriate files and records. The patient will be removed from the study recruitment log and any linked paper documentation destroyed. This will be undertaken within the time frame stipulated by GDPR. As indicated in the RKD PIL, it will not be possible to erase data that have already been used in a scientific manuscript or collaboration. If desired, the participant will be given a copy of their data on a USB memory stick.

## E.3 Data Requests

Chapter 3 of the GDPR details Data Subject Right provisions and calls for mechanisms to be put in place whereby the participant can request a copy of their data. This request may be facilitated following direct contact with medical or research teams or the RKD Research Nurse. Such requests will be noted in the RKD recruitment log. The Biobank manager is responsible for compiling patient data exports and formatting an individualised report. This will be in CSV or PDF format with accompanying metadata and will be transferred to the local research nurse/medical team, who will link to identifiable data and transfer the file to the requester on a USB memory stick. This will be undertaken within the time frame stipulated by GDPR.

## E.4 Data Review

The PI will be responsible for managing version control of RKD documents and files and for completing regular QC reviews of data accuracy. Regular audits of access will ensure efficient reporting of results as detailed in the Data Sharing Plan. Where inaccuracies in the data are identified by other RKD study team members this will be amended in the relevant file and recorded in the recruitment log.

# **F. Ethics and Legal Compliance**

All participants have provided consent for research projects deriving from the stored samples and data. The RKD project has oversight by the RKD Steering Committee and sub-studies are approved by the SJH/Tallaght Ethics Committee. For example, the DeCOmPRESS sub-study has received approval: 2020-04 List 13 – Amendment (22). A log of such approvals is maintained by the RKD PI.

## F.1 Informed Consent

All patients enrolled to the RKD Biobank provide informed consent for their data and specimens to be used for research purposes. Full details are provided in the RKD Biobank protocol. Patient information leaflets and RKD protocols are available to download at: <https://www.tcd.ie/medicine/thkc/research/rare.php>*.*

## F.2 Data Linkage

Linkage allows patient-reported data to be linked to clinical data derived from the RKD registry and immunological data from patient sample analysis. The purpose of this is to characterise the immune response to COVID-19 and how AAV treatment regimens affect the disease course. The RKD ID is the primary identifier and unifies data collected from multiple sources. RKD IDs can only be re-identified via the local RKD Biobank recruitment log, to which only the RKD Research Nurse and site PI have access. Study team researchers will have access to the relevant and suitably transformed data with prior approval from the PI. The PI is responsible for ensuring correct data linkage across sources, and the validity and consistency of results and data. In some cases (with complex data streams, such as the AVERT and DeCOmPRESS studies), data linkage is undertaken in the ADAPT RDF triplestore database.

## F.3 RKD Personnel

Researchers and study team members with access to this data have a responsibility to work with integrity. RKD researchers commit to the highest standards of data security and protection in order to preserve the personal rights and interests of study participants. All RKD personnel who have access to identifiable patient data are required to undergo GCP training and ensure that this training is maintained up to date. They will adhere to the provisions set out in the:

* General Data Protection Regulation (GDPR) (Regulation (EU) 2016/679) as enacted in May 2018, which strengthens and unifies data protection for all individuals within the European Union (EU). It also addresses the export of personal data outside the EU.
* Directive 2006/24/EC of 15 March 2006 on the retention of data generated or processed in connection with the provision of publicly available electronic communication services or of public communications networks.
* Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications) and
* Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data.

In addition, data governance will comply with the stipulations of the General Data Protection Regulation legislation enacted in 2018. To ensure the confidentiality, accuracy and security of data, the following measures will be taken eCRFs and other forms needed for the collection of patient data will be unified and exported in appropriate formats for review by the relevant authorities as required.

# **G. Data Storage and Backup**

Data pertaining to the RKD study will be held in hospital and TCD locations as summarised in Figure 4. Data held within TCD will be pseudonymised.



**Figure 4:** Summary of RKD data flows

## G.1 TCD Network Security Statement

1. Information is a critical asset of Trinity College Dublin hereafter referred to as the ‘University’. Accurate, timely, relevant, and properly protected information is essential to the success of the University’s academic and administrative activities. The University is committed to ensuring all accesses to, uses of, and processing of University information is performed in a secure manner.
2. Trinity College Dublin is committed to adopting a security model in line with the ISO27001/ISO27002 international best practice standards.
3. Technological Information Systems hereafter referred to as Information Systems’ play a major role in supporting the day-to-day activities of the University. These Information Systems include but are not limited to all Infrastructure, networks, hardware, and software, which are used to manipulate, process, transport or store Information owned by the University.
4. The object of this Information Systems Security Policy and its supporting technical requirements policy is to define the security controls necessary to safeguard University Information Systems and ensure the security confidentiality and integrity of the information held therein.
5. The Policy provides a framework in which security threats to University Information Systems can be identified and managed on a risk basis and establishes terms of reference, which are to ensure uniform implementation of Information security controls throughout the University.
6. The University recognises that failure to implement adequate Information security controls could potentially lead to:
	1. Financial loss
	2. Irretrievable loss of Important University Data
	3. Damage to the reputation of the University
	4. Legal consequences
7. Therefore, measures must be in place which will minimise the risk to the College from unauthorised modification, destruction or disclosure of data, whether accidental or deliberate. This can only be achieved if all staff and students observe the highest standards of ethical, personal and professional conduct. Effective security is achieved by working with a proper discipline, in compliance with legislation and University policies, and by adherence to approved University Codes of Practice.
8. The Information Systems Security Policy and supporting policies apply to all staff and students of the University and all other users authorised by the University.
9. The Information Systems Security Policy and supporting policies do not form part of a formal contract of employment with the University, but it is a condition of employment that employees will abide by the regulations and policies made by the University from time to time. Likewise, the policies are an integral part of the Regulations for Students
10. The Information Systems Security Policy and supporting policies relate to use of:
11. All University networks connected to the University Backbone
12. All University-owned/leased/rented and on-loan facilities.
13. To all private systems, owned/leased/rented/on-loan, when connected to the University network directly, or indirectly.
14. To all University-owned/licensed data/programs, on University and on private systems.
15. To all data/programs provided to the University by sponsors or external agencies.
16. The objectives of the Information Systems Security Policy and supporting policies are to:
	1. Ensure that information is created used and maintained in a secure environment.
	2. Ensure that all of the University’s computing facilities, programs, data, network and equipment are adequately protected against loss, misuse or abuse.
	3. Ensure that all users are aware of and fully comply with the Policy Statement and the relevant supporting policies and procedures.
	4. Ensure that all users are aware of and fully comply with the relevant Irish and European Community legislation.
	5. Create awareness that appropriate security measures must be implemented as part of the effective operation and support of Information Security.
	6. Ensure that all users understand their own responsibilities for protecting the confidentiality and integrity of the data they handle.
	7. Ensure all University owned assets have an identified owner /administrator
17. The University Board has approved the Information Systems Security Policy and supporting technical policy. The Board has delegated the implementation of the Information Systems Security Policy, to the heads of academic and administrative areas. The Director of IT Services and his/her delegated agents will enforce the Information Systems Security Policy and associated supporting policy.

## G.2 Microsoft Office OneDrive/SharePoint

OneDrive/SharePoint is a file hosting and synchronization service operated by Microsoft as part of its web version of Office. OneDrive allows users to sync and share files between computers and mobile devices from anywhere with an internet connection. Document edits are instantly saved online and can be viewed and edited in real time by anyone with access to the document enabling live collaboration with colleagues. Version and access control of files stored on OneDrive will be the responsibility of the PI. OneDrive data is securely hosted by Microsoft in Europe in compliance with relevant legislation.

## G.3 Data Security Statements

### G.3.1 RKD Registry

1. Identifiable patient/control data is pseudonymised after recruitment by assigning a study ID; their consent forms with medical record numbers will be stored separately in a secure facility at the local hospital site.
2. Pseudonymised data will be uploaded on the REDCap database which will be mapped to a dedicated password protected computer using IP address. The database will be protected behind host and institutional firewall with access to dedicated personnel only.
3. For further information please refer to Data Protection Impact Assessment.

### G.3.2 RKD Biobank Samples

1. Coded biological samples are processed and stored centrally at the RKD biobank and archived by the biobank technician using industry standard Freezerworks software; only the biobank technician and Lead study PI have access to the software.
2. This software is maintained on a stand-alone TCD-networked and password-controlled laptop, which is stored in a locked cupboard in the Trinity Translational Medicine Institute. It is not allowed to leave this building.

### G.3.3 Experimental Data

1. This data will be stored in password-protected SharePoint/OneDrive folders with strict access control.
	1. OneDrive security is maintained by TCD in accordance with the TCD Network Security Statement
	2. OneDrive data is securely hosted by Microsoft in Europe in compliance with relevant legislation.

### G.3.4 RKD ADAPT Triplestore Database

The ADAPT server is located on the TCD Virtual Machine and Docker cluster.

1. Hardware
	1. 4 Virtual Machine (VM) nodes and 4 storage nodes
2. Software
	1. OS (hosts and storage nodes): Debian 9
	2. VM cluster: OpenNebula 5
	3. Container hosts: Docker
3. Storage: Ceph
	1. This will be backed up daily and stored on a dedicated back-up server.
4. Security Detail
	1. Two firewalls
		1. Between our subnet and the host School of Computer Science and Statistics network that filters connections on some ports
		2. TCD firewall that blocks all incoming connections and filters some outgoing connections;
	2. For Apache web servers, we use the [Nikto](https://cirt.net/nikto2) tool to scan all the websites hosted in our cluster once per month for known vulnerabilities;
	3. For all webservers, we expose them through our reverse proxy, and the reverse proxy logs every connection and can restrict incoming connections depending on source IP.
5. Access Control
	1. Only the requesting user can login and obtain a shell on the VMs.
	2. Connections to services hosted on the VM can be restricted by source IP as requested by the user.

## G.4 Access Management

The PI will be the only study member with full read & write access to all data files in the RKD study. Patient identifying information will only be accessible to local medical teams and not stored in any TCD RKD data files. Historical clinical data for patient group assignment and analysis are stored in the RKD database. This database operates on multiple-access levels and only the lead study PI and Research Nurse will have complete access to the database and permission to modify. Access to relevant documents (and new access requests) will be managed by the Project Manager, approved by the Lead Study PI, and maintained in an access log. An assessment of all individuals who have accessed each data location, and the nature of all data (e.g. identifiable vs de-identified) will be included in regular study reports report.

# **H. Data Selection and Preservation**

All data collected in the course of the RKD study will be retained in the secure structures as defined above. Data will be stored in accordance with FAIR principles. This will allow results from this work to be interoperable with international data collection initiatives and maximise the impact of the project. Organisation or study groups requesting access to RKD data or samples will be reviewed and approved/rejected by the PI and RKD steering committee as described in the RKD data sharing agreement.

Study data will be held on TCD servers for the duration of the project and beyond for as far as is currently foreseeable. The PI will ultimately be responsible for its management and storage beyond the study duration. Upon completion of specific projects, study datasets and will be deposited in a predetermined, accredited archive, respecting data protection requirements. Linked clinical samples will be stored indefinitely in appropriate storage conditions.

# **I. Dissemination**

A fundamental mission of the RKD study is the timely and open-access dissemination of study results. This is routinely achieved through production of an annual newsletter and publication of study results in open access journals. Audits of data access will be a fundamental part of this reporting procedure. All research outputs will be reported in real time in open access formats and shared with the relevant stakeholders in line with the [HRB Joint statement](https://wellcome.ac.uk/coronavirus-covid-19/open-data) on sharing research data and findings relevant to the novel coronavirus (COVID-19) outbreak. The study protocol will be published on the HRB open research. Detailed processes for data sharing, dissemination and exploitation are described in the data sharing agreement