AUTOIMMUNE RELAPSE PREDICTION USING MULTIPLE PARALLEL DATA SOURCES

Data Management Plan

Mark little
www.thkc.ie

Brief description of AVERT

ANCA vasculitis is a relapsing and remitting rare autoimmune disease that results in rapidly progressive kidney impairment, in addition to immune-mediated destruction of other organs. Epidemiological data exist supporting a strong environmental impact on autoimmunity in ANCA vasculitis, although the nature of these triggers is rarely defined in detail. The rapidly emerging discipline of data science - alongside massive increases in computing capability, machine learning and artificial intelligence - is poised to allow the incorporation of such highly complex health data environments, and the generation of outputs with potential applicability in personalised medicine. By adopting an unbiased approach, we aim to integrate a wide array of unstructured data streams to define the signature of relapse of the disease. We believe this approach will represent a new paradigm in managing chronic conditions governed by interaction between patient-level factors and their environment.

This is a five-year project. It ultimately aims to allow for the tailoring of therapy to the risk in the individual at a given point in time, and will use a big data approach to incorporate a broad range of potential data streams, including: the Rare Kidney Disease registry; the HSE's Computerised Infectious Disease Reporting (CIDR) system; online pollution/weather data streams; and patient-derived data streams using smart phone and wearable technologies.

The primary objectives of this ambitious programme, titled "AVERT", are therefore to:

- Establish and refine an ethically robust and scalable cloud-based data architecture, supporting rapid curation of a wide array of unstructured environmental data streams for integration with electronic health records (EHR) and data derived directly from the patient. This will be modular, thereby facilitating the addition of future data streams and public data resources. It will continue to exist, grow and learn beyond the lifetime of the proposed project, and will be applicable to other chronic diseases.
- Generate exploratory models of the association between environmental, electronic health record and direct patient-derived data using vasculitis relapse as the primary outcome variable and incorporating prior knowledge of influences of relapse, validating these findings in an independent cohort.
- Use these statistical models to generate machine learning algorithms that support prediction of vasculitis relapse risk, and
- Ultimately incorporate these into a prototype physician dashboard.

As part of this, we have been engaging with several bodies, such as Met Éireann and the Environmental Protection Agency to gain access to data measuring relevant environmental and external factors that may influence this risk. Data relating to influenza outbreaks and the CIDR database will support better understanding of the pathogens present in the patient's environment, and the potential interaction between these and other factors.

This data management plan refers to the development of the statistical model(s) prior to their use in the dashboard. Issues such as obtaining patient consent for use of data to develop a clinical tool can be

considered beyond its scope, and will need to be revisited in future if necessary. It is a living document and will be updated periodically.

Information governance summary

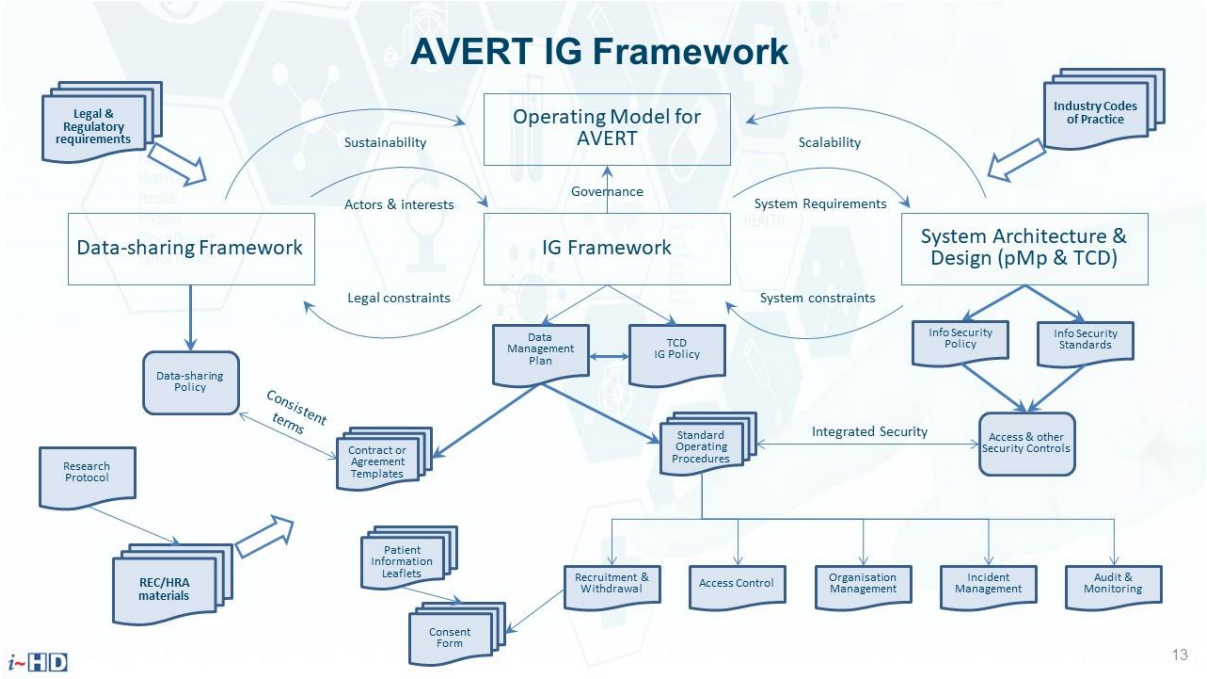

Figure 1 - Organogram of AVERT information governance

AVERT aims to conduct novel bioinformatics research, making use of an existing clinical research registry. AVERT will draw on existing such instruments that are being collated and further developed by IGB partner i~HD at a European level, which will set a default standard across the project. At times cross-centre research collaborations will require data sharing, potentially between European countries. Overarching AVERT policies and measures will be applied to such data transfers (or to distributed querying). To ensure that the necessary research can be undertaken across AVERT, and yet to minimise, control and audit the disclosure of subject-level data, AVERT will conduct a Privacy Impact Assessment to inform this overarching information governance policy. This policy will be complemented by standard operating rules and codes of practice for different, well-defined, roles for the research and technical actors in the project. An information security policy will be implemented across all of the cross-centre information flows and repositories. A uniform base standard for information security will apply across the project (designed to interface with the existing policies and practices at TCD), to provide a consistent assurance of privacy protection across all AVERT research. The IGB members will have oversight of the security practices, and the remit to investigate any issues or concerns that arise during the project. We will support and have oversight of consent and ethical approvals in place, and support good practices in capturing any new subject consent.

Good practices in data sharing, including a standard data sharing agreement, will be used across the project which defines in advance the data to be shared, the data access mechanisms, the security

measures to be applied, and also includes predefined agreements about intellectual property, publication and authorship etc.

IGB partner i~HD already has a portfolio of policies and governance instruments, including a data sharing template, that have been developed and refined through previous EC and IMI projects (including EHR4CR and EMIF) which respond to the obligations in the new European GDPR and conform to the IMI Code of practice on secondary use of medical data in European scientific research projects.

The basic AVERT data governance principles can be summarised as follows:

1. Integrity

We will practice integrity with their dealings with each other; we will be truthful and forthcoming when discussing constraints, options, and impacts for data-related decisions.

2. 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.

3. 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.

4. Accountability

Data Governance will define accountabilities for cross-functional data-related decisions, processes, and controls.

5. 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.

6. 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.

7. Standardization

Data Governance will introduce and support standardisation of data to maximise potential for sharing.

8. 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.

Data Collection

Types of data collected and created

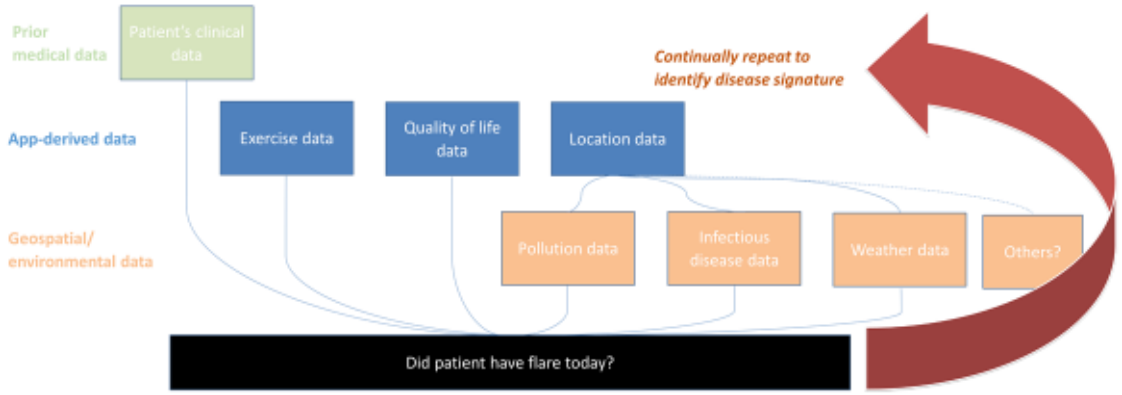

Figure 2 - Overview of types of data collected

All data will be collected according to the FAIR data principles, namely *Findability, Accessibility, Interoperability, and Reusability*. The primary means by which this will be achieved will be by uplifting study data into the RDF data model for storage in a "triplestore" database.

Data can be broadly grouped into two categories: data that we will collect and generate ourselves, and data collected elsewhere that we have access to, and which we will link to.

1. *Data we will collect*
    a. We have commissioned PatientMPower (pMp) to develop an app which will monitor:
        i. Participants' daily activity levels (step count, as measured by smartphone accelerometer)
        ii. health related quality of life (HRQoL), using EQ-5D-5L (entered directly by recruit)
        iii. GPS location (as determined by smartphone GPS). This will be recorded once per day.
    b. AVERT will not collect genetic information in the first instance, but may in the future incorporate de-identified genomic, proteomic and other biological "-omic" data sources, pending revision of this DMP. AVERT will not collect or process data concerning sexual lifestyle, ethnicity, political opinion, religious or philosophical conviction. However, it will collect potentially sensitive healthcare data from the RKD registry. It does not involve direct observation of participants.

2. *Data collected elsewhere*
    a. For recruits with location data, we can link this with (largely publicly available) data on relevant environmental factors, such as local pollution levels, weather conditions and infection rates.
    b. We will also have access to deidentified clinical data, recorded as part of the Rare Kidney Disease (RKD) registry; we shall link these data with other AVERT data streams.

Data will be stored using a Resource Description Framework (RDF) model to facilitate analysis and data sharing.

Mechanisms for collection and creation of data

1. Data generated **directly** by AVERT will be collected from:
    a. an app developed for AVERT, measuring patient-reported outcome measurements, exercise and their location data.
    b. For a limited number of recruits without smartphones who nonetheless wish to take part in the study, periodic HRQoL data can be inputted via desktop computer or tablet, or by direct interview with the study research nurse. Exercise and precise location information will not be available to researchers for these patients.
2. Data generated by **RKD registry**. Clinical data will be created independently of AVERT and stored in a separate database. We will not be responsible for collection of these data. Clinical data will be eligible for updating (such as where clinical diagnoses have been revised by clinicians). Such data will replace earlier data sources in our records.
3. Data obtained from **publicly available** electronic archives. Met Eireann and EPA data are freely available on both organisations' websites. The Health Service Executive have agreed to provide regular data updates on the Irish sentinel GP influenza-like illness (ILI) consultation rates and Computerised Infectious Disease Reporting (CIDR). Other similar sources, some not yet foreseen, may be added modularly as they become available. Historical, publicly available data (such as weather readings) will be recorded here.
4. Biological data generated from **genomic, proteomic, metabolomic** and other similar sources. These data will be generated and annotated by research groups within or outwith TCD using biological samples provided by AVERT recruits. These will be de-identified prior to transfer into the AVERT data architecture.

Documentation and Metadata

The big data nature of the AVERT project poses specific challenges. We propose to use an RDF approach as the overall model for the project's knowledge management. This approach will enable the storage of data in a common format, which will allow the data to be more easily understood by both machines and the human observers who are responsible for overseeing the analyses. Ontology languages can be built upon this, allowing for quicker and more intuitive querying of the data than might otherwise be possible for data of this size, facilitating consistent combination of diverse data sources

and supporting the application of queries to the model, thereby generating new knowledge. The RDF approach will allow for a simple modular framework, allowing us to add further data sources in the future as they become available.

An explicit goal of AVERT is to explore the use of RDF to represent clinical research metadata, which will help to articulate and clarify the structure (and, crucially, the *provenance*) of the datasets. It will also allow for ontological concepts to be classified using well-understood generic representations. For example, "gender" can be understood with reference to previously defined definitions of gender from prior ontologies (such as FOAF). Where possible, previously defined ontologies can be used. For others, we will have to generate our own, and give clear descriptions of each, which will be made available to relevant users of the data in future.

Responsibilities and Resources

Who will be responsible for data management?

The Information Governance Board (IGB) will be responsible for the ethical handling of data and for data security. However, all members of the team (and anyone with access to the data) must act responsibly, and will be trained in Data Governance principles.

The IGB will be comprised of four members, in the first instance Mark Little (PI), Lucy Hederman (Data steward), Dipak Kalra (i~HD) and a patient representative (VIA). A quorum of 3 is necessary for any decision to take place. Members can be replaced with unanimous assent of other members. The activities and opinions of the Board will be made appropriately transparent through the project web site.

Who will be responsible for data erasure following a request to do so?

In accordance with GDPR legislation, if a participant requests for their data to be erased, this will be managed by the PI. This request may be submitted through the app, or following direct contact with investigators. A log of such requests will be maintained by the study research nurse. Upon receipt of such a request, the PI will delete the data linked to the relevant AVERT study ID in the ADAPT centre study servers, as well as instructing the research nurse to delete the record from the recruitment log, and to destroy any linked paper documentation. This will be undertaken within the time frame stipulated by GDPR.

As indicated in the information leaflet, it will not be possible to erase data that have already been used in a scientific manuscript or collaboration.

Who will be responsible for transfer of data back to the participant following a request for same?

GDPR calls for mechanisms to be put in place whereby the participant can request a copy of their data. This can be triggered by contacting the study nurse. The PI will be responsible for exporting the coded data from the ADAPT server. This will be in MS excel format and will be transferred to the research nurse, 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.

Who will be responsible for rectifying inaccurate data?

If the participant identifies inaccuracies in the data help by the AVERT study team, they can notify the research nurse about this.

- If the inaccuracy exists within the participant's demographics (eg, gender, date of birth, address), this will be amended directly within the recruitment log; a record will be recorded of this event.
- If the inaccuracy exists within the pMp app data, the participant can send a message directly from the app to the pMp team, who will be responsible for rectifying it.
- If the inaccuracy exists within the clinical data derived from the RKD registry, this will be alerted to the RKD study coordinator, who will be responsible for rectifying it.

Ethics and Legal Compliance

Management of ethical issues

Ethics approval will be sought at the Tallaght/St James' Hospital ethics board. The PI will ensure that consent document wording is mirrored across paper and app in the event of ethics amendments.

Consent will be sought from patients for both perpetual data preservation and for the sharing with relevant research groups (as approved by the information governance board, described below). The building of appropriate infrastructure for eventual sharing of this data is an important aim of the study. The key ethical issue for this project is protection of patient data, including data that is identifiable and potentially sensitive (such as registry-derived clinical information or location data). RKD clinical data will be linked with detailed personal data coming from an app and will be de-identified.

Consent procedures

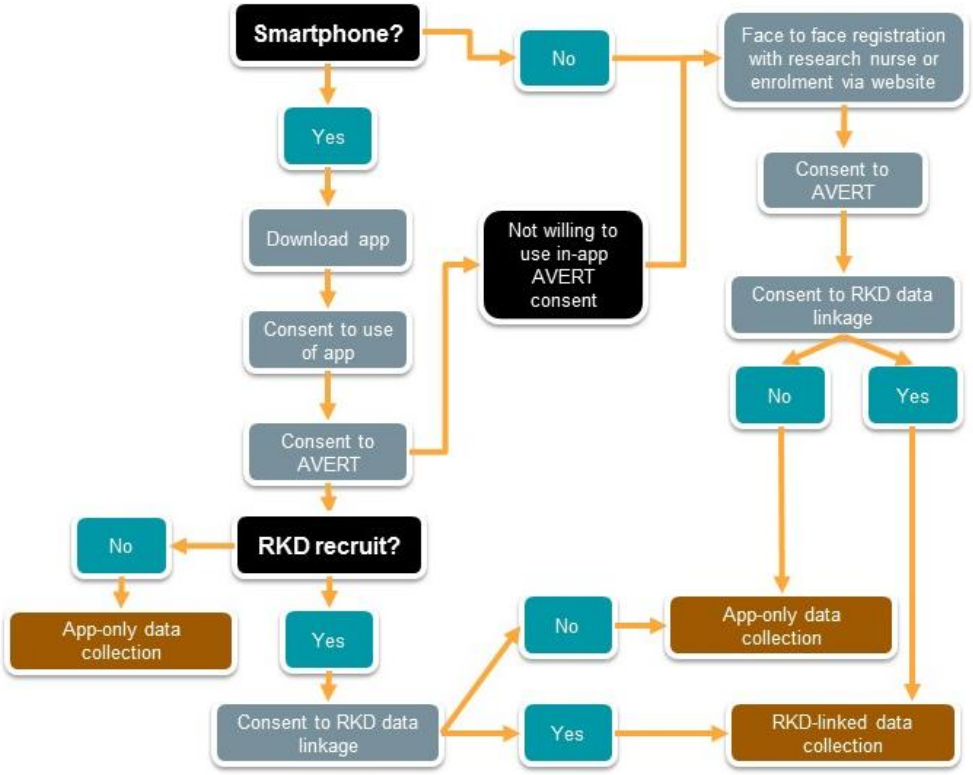

Fig 3: Summary of pathways to consent. These are also described in SOP 2.

In-app consent

The workflow and governance of AVERT consent is optimal when in-app consent is used.

The patient downloads the app from IOS or Android Play (search for patientMpower), completes the consent to use the app and provides registration details. This allows the individual to use the app as a clinical support tool, but does not permit research.

The patient is then invited to participate in the AVERT study. This has 2 levels:

a. Basic enrolment in AVERT, which allows activity and location data collected through the app to be used in the AVERT study
b. Additional linkage of these data to the RKD registry clinical dataset. These are separated because recruits should be able to opt out of sharing their clinical data, and because many AVERT recruits will be outside the RKD study or, indeed, outside Ireland (in which cases there will be no clinical dataset to link to).

If the individual has previously been recruited to the RKD registry, they enter their RKD study ID when prompted. This also serves as the AVERT study ID. If the patient does not know their study ID they can:

- Refer to the communication from their local hospital, which will include the ID number
- Contact the study coordinator on rkdnurse@tcd.ie

The patient can take as long as they wish to consider the study information, which will be provided both through the app and downloadable via the website. Parts a) and b) can be completed at separate times. For example, part b) can be completed at a later time point during a face to face interaction with a research nurse or other study personnel.

Face to face consent of prior RKD recruits

Where possible, the app should be used to manage consent process as this allows for a dynamic consent process over the life of the study.

Informed consent will be obtained by one of the lead investigators or a fully qualified member of the research team such as a research nurse or research registrar (as defined by delegation log).

The nature and objectives of the study will be explained to the patient. Risks and benefits of participation will be discussed and patients will be made aware that participation or otherwise will not alter their healthcare in any way.

The nurse will guide the patient in downloading the app (if not already done so) and completing the appropriate consent items as per section 1. If the patient has not got a smart phone or does not wish to complete electronic consent, a paper consent process can be followed:

- The RKD study number will be written on the consent form.
- The consenting patient, research nurse and/or doctor will sign the consent form.
- The patient should date their own consent form.
- The Research Nurse will photocopy the form twice, maintaining one copy in the CRF, one copy in the patient's medical chart, and giving one copy to the patient.
- Irrespective of the mode of consent, enrolment into the AVERT study is recorded in the patient's notes.

If the patient decides to withdraw from the study at any stage, the research nurse/research team member will document this decision clearly in the patient's medical notes and CRF and ECRF, detailing the reason if known.

If during the study, it is discovered that the patient was enrolled but did not meet the inclusion criteria, this will be documented clearly in the CRF and ECRF and the principal investigator will be informed.

Data linkage

Linkage allows the patient-reported data to be linked to clinical data derived from the RKD registry. The principal purpose of this is to characterise flare occurrences, and to incorporate additional clinical data such as baseline vasculitis phenotype, medication use and laboratory results.

Linkage of activity, location and patient reported outcomes to environmental and clinical (RKD registry) data feeds is managed by the research nurse using an encrypted database on a password protected server.

The AVERT study ID is auto-generated by the pMp software, or from a central list of IDs (managed by the research nurse) for non-app/website users. The RKD study ID is entered directly into the app or web page upon registering for the AVERT study; this is held separately from the other app data and is sent separately to the research nurse. The research nurse manually links the AVERT and RKD codes. Only data coded with the AVERT study ID will be used for subsequent analysis.

Researchers will therefore have access to relevant and suitably transformed RKD data for patients, alongside the app data.

Anonymisation procedures and data security throughout the data life cycle

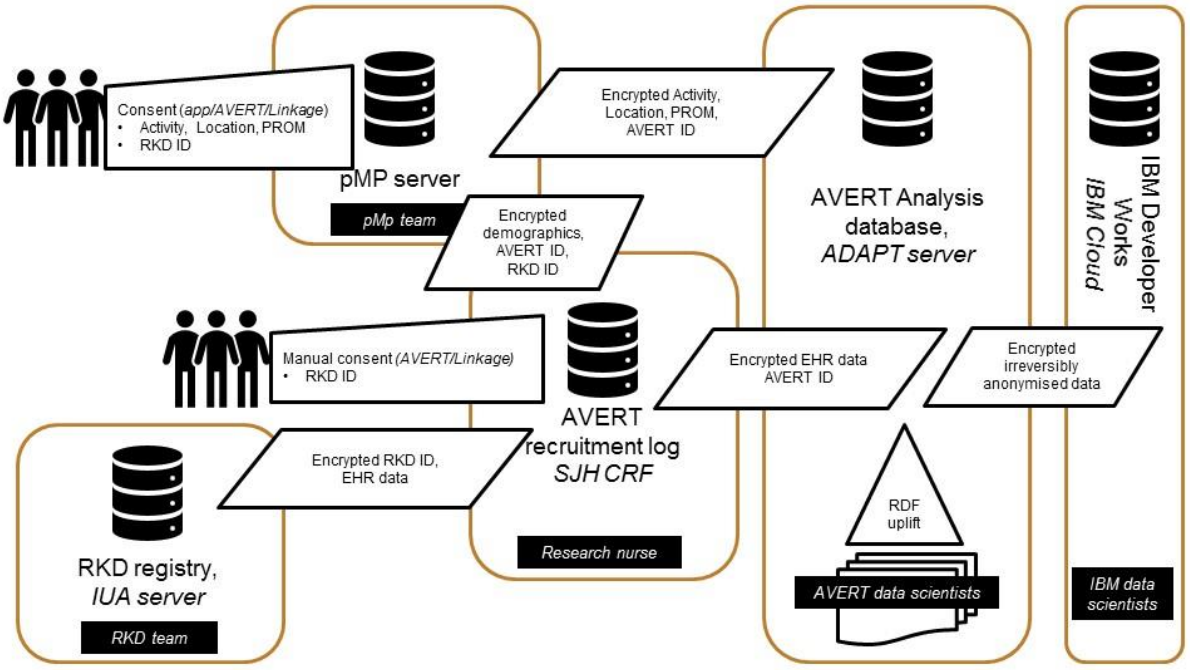

Fig 4: Summary of AVERT data sources and their storage sites, along with a description of the people with access to each site. IUA=Irish Universities Association, CRF = Clinical Research Facility, RDF = Resource Data Framework. The brown lines indicate firewalls.

AVERT data will be located within the following domains (figure 4):

1. **pMp server**: identifiable data entered by the recruit at the time of app consent (required for management of the app), location data, activity data, HrQOL data, RKD ID (entered by recruit at time of consent for RKD linkage).
    a. App and identifiable data accessible to pMp staff
    b. RKD ID not accessible to pMp staff
2. **RKD registry** on IUA server: coded de-identified data describing clinical parameters over time; this falls under a separate data governance framework.

    a. Accessible to RKD registry study personnel (and not AVERT study personnel)
3. **AVERT study recruitment log**: identifiable data obtained from app and local RKD recruitment log, RKD clinical data (for the purpose of transforming to AVERT study ID prior to transmission to AVERT database).
    a. Accessible to AVERT research nurse and one delegate. Tightly controlled access procedures and data security.
4. **AVERT analysis database**: de-identified app and RKD data, coded with AVERT study ID; environmental data
    a. Accessible to AVERT data scientists, access controlled by AVERT information governance team
5. **IBM Developer Works** cloud-based server: irreversibly anonymised app and RKD data
    a. Accessible to IBM data scientists

Given that location data are collected by the app, the project data remains sensitive and cannot therefore be considered as truly anonymous. Therefore, the storage and security procedures described in this section are key to AVERT data governance, as is the need for researchers with access to such data to behave responsibly and with integrity. AVERT researchers commit to the highest standards of data security and protection in order to preserve the personal rights and interests of study participants. They will adhere to the provisions set out in the

- 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, due to be 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 reviewed by the relevant authorities as well as the IGB to ensure only adequate and relevant information will be recorded.

- Data are processed only for the purposes outlined in the patient information leaflets. Use for other purposes will require explicit patient approval. Also, data are not transferred to any laboratories or places out-side the consortium without patient consent.
- The IGB will maintain accurate logs that define access clearance for various AVERT data stores.

Copyright and Intellectual Property Rights (IPR) issues

Personal data and data generated by the AVERT participants will be owned by the participants. Data deriving from analysis of these primary data will be owned by the Chronic Disease Informatics Group in the Trinity Health Kidney Centre. No personal data will be used for commercial purposes, although knowledge derived from the research using the personal data may be brought forward to such use as appropriate.

Storage and Backup

Data will be stored in the locations described in Figure 4. Each of these locations is protected by a firewall.

Data security statements:

1. **pMp server** (August 2017)
    a. **Introduction**. patientMpower is a platform to empower patients to take control of health conditions. This is enabled by collecting and processing data relevant to their condition in the form of home diagnostics, surveys, and activity data. Due to the nature of this data, we take issues of security and consent very seriously and have implemented various technology and process safeguards to ensure confidentiality of patient's data.
    b. **Security Details**. patientMpower is designed with stringent security protocols. Our solution is hosted in Microsoft Azure, and we have built our security layers on top of the standard security which Microsoft offers:
        i. The platform uses a PostgreSQL database which is backed up nightly
        ii. The Virtual Machine running the platform is snapshotted weekly
        iii. Operating System is Ubuntu, running unattended security upgrades to remain up to date with current vulnerabilities
        iv. Running fail2ban which protects against brute force attacks on server
        v. SSH access to server with IP whitelists and public/private key access only
        vi. Built-in firewalls
        vii. Encrypted data storage
    c. **Staff access** to the PostgreSQL database and content system is restricted and monitored:
        i. A unique username and password for each user.

ii. Audit and accounting of all access to the system is recorded. In the event of any staff looking at data without proper authorisation, there is an audit trail of what data was viewed

iii. Data transfer between the patient mobile device and cloud server is sent securely via TLS, and our cloud infrastructure uses an Extended Validation SSL Certificate issued by Digicert

iv. Data on the server is encrypted, only authenticated users can access the server

d. **Anonymisation.** For patientMpower research studies all data are anonymised using the study ID format agreed with the research centre. Separate login access is provided to any study staff in order to access the anonymised records.

e. **Hosting.** patientMpower is cloud hosted and utilises Microsoft Azure for all our hosting requirements. Our solution and data are hosted within the European Economic Area. We have signed EU Model Contract Clauses and a Data Processing Amendment which means all data must remain in countries which meet the EU's "adequacy" standard for privacy protection.

f. **Staff Guidelines.** patientMpower operates a separate test infrastructure which is used by developers, testers and for demonstration purposes. Only authorised staff have access to the 'production environment'. Any access in the production environment is logged, and comments are logged where a staff member accesses patient data (e.g. due to a support request)

2. Recruitment log

a. This will be maintained on a password-protected SQL database on a computer in the St James's Hospital Clinical Research Facility. The database will sit within the TCD network. Only the AVERT research nurse and one delegate will have access to this.

b. This will also be backed up on a regular (approximately monthly) basis on a password-protected external hard drive. This will not be used for other purposes and will not be connected to the internet except as part of the downloading of relevant data.

c. TCD network security statement

i. 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.

ii. Trinity College Dublin is committed to adopting a security model in line with the ISO27001/ISO27002 international best practice standards.

iii. 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.

iv. 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.

v. 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.

vi. 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

vii. 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.

viii. 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.

ix. 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

x. The Information Systems Security Policy and supporting policies relate to use of:

    1. All University networks connected to the University Backbone
    2. All University-owned/leased/rented and on-loan facilities.
    3. To all private systems, owned/leased/rented/on-loan, when connected to the University network directly, or indirectly.
    4. To all University-owned/licensed data/programs, on University and on private systems.

5. To all data/programs provided to the University by sponsors or external agencies.

xi. 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

xii. 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.

3. ADAPT server

a. This is located on the TCD Virtual Machine and Docker cluster
b. *Hardware*:
    i. 4 Virtual Machine (VM) nodes and 4 storage nodes
c. *Software*:
    i. OS (hosts and storage nodes): Debian 9
    ii. VM cluster: OpenNebula 5
    iii. Container hosts: Docker
d. *Storage*: Ceph
    i. This will also be backed up daily and stored on a dedicated back-up server.
e. *Security details*:
    i. There are 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;

ii. For Apache web servers, we use a tool called Nikto (https://cirt.net/nikto2) to scan every month all the websites hosted in our cluster for known vulnerabilities;

iii. 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.

f. *Access control*:

i. Only the requesting user can login and obtain a shell on the VMs.

ii. Connections to services hosted on the VM can be restricted by source IP as requested by the user.

Access management

Only those with explicit need to see specific data will have access to it (SOP4). This access will be managed by the Information Governance Board (IGB) using password protection, and aligning to the United States National Institute of Standards and Technology (NIST) digital authentication guidelines, NIST SP 800-63B-3. A log will be maintained of access rights.

Access privilege revocation: Upon discontinuation of a contract with TCD, access to email and the TCD network is automatically disabled. This process will be linked to additional disablement of AVERT data storage access profiles and passwords.

Auditing and Reporting

The IGB will be responsible for conducting at least six-monthly audits of data access (or for appointing an independent party to conduct such audits). These will include an assessment of all individuals who have accessed each data location, and the nature of all data (eg identifiable v de-identified) in each location.

Data selection and preservation

Data destined to be retained, shared, and/or preserved

One of the aims of AVERT is to establish a scalable cloud-based data architecture so that such studies can be carried out and refined in future. Because it is unclear at this point which data sources will be the most relevant, or which further data sources may become available in future (potentially making seemingly insignificant variables more important), we intend to retain these data. The nature of machine learning means that such data should be kept for the foreseeable future so that it can continue to be added beyond the lifetime of the project; each flare event provides the opportunity to refine the algorithms. Because of the nature of the data, however, it can only be shared with trusted organisations, as approved by the IGB and described in the data sharing plan.

Long-term preservation plan for the dataset

This of course poses ethical issues about who should be responsible for its management beyond the timescale of the study. The data will be held on the ADAPT server for as far as is currently foreseeable. The IGB will be responsible for its management, and will also have a responsibility to continue to ensure that it continues to replace its own members when they leave so that suitable decision making over the data's management can take place. At some appropriate time in future, the structure of the board itself may be altered, for example if the data (and its management) is to be merged with that of another dataset. Such a move would require the unanimous agreement of members of the Board. Ultimately, if the AVERT group decide to disband we will deposit the datasets and algorithm / algorithm provenance in an accredited archive, respecting data protection requirements.