Plan Overview

A Data Management Plan created using DMPonline

Title: Centre for Data Enriched Medicine (TeamPerMed)

Creator: Laura Breede

Principal Investigator: Mait Metspalu

Project Administrator: Riina Raudne, Laura Breede

Affiliation: University of Tartu

Funder: European Commission

Template: Horizon Europe Template

ORCID iD: 0000-0003-3099-9161

Project abstract:

Using advanced experience, medical research progress, and international know-how, the TeamPerMed project empowers the University of Tartu together with Tartu University Hospital to form the independent Centre for Personalised Medicine - TeamPerMed. This Center of Excellence (CoE) will become a multi-disciplinary unit integrating expertise in Genomics, IT, Clinical Medicine, and Socio-Economic Analysis. It will create a scalable framework for developing clinical guidelines and clinical decision support (CDS) tools that can be effectively integrated into EU healthcare systems. With the guidance from the leading expert partners, Erasmus Medical Centre and the University of Helsinki, a vibrant network of associated stakeholders, and the support of EC and the Republic of Estonia, TeamPerMed will become a recognized regional/EU leader in translating genomic and electronic health data to prevent the development of chronic or hereditary disease and offering high-quality personalized medical care. TeamPerMed will stimulate digital transformation in healthcare by using available databases and electronic health records to develop AI tools and protocols effective for prevention/early identification of individual health risks and their consecutive treatment. TeamPerMed CoE intellectual outputs including technology, tools, knowledge, and innovative healthcare practices, will boost Estonia’s and Europe’s competitiveness in the emerging personalised medicine sector, improving the quality of healthcare, prolonging longevity and maximizing health of European citizens, and bringing the advantages of future digital medicine to our citizens already today. The TeamPerMed Center will become an internationally recognized leader in the areas of personalized medicine and customized healthcare contributing directly to a better physical and mental wellbeing and indirectly to social and spiritual wellness, opening the door to a more harmonious society and paving the way for the ‘next generation

ID: 209353

Start date: 01-09-2023

End date: 31-08-2029

Last modified: 12-08-2026

Grant number / URL: https://cordis.europa.eu/project/id/101060011

Copyright information:

Centre for Data Enriched Medicine (TeamPerMed)

Data Summary

Will you re-use any existing data and what will you re-use it for?

TeamPerMed project will re-use previously existing data, including health, genomic, lifestyle and environment data from health registries, the Estonian Biobank (EstBB) and health records. In addition, we
will perform whole genome sequencing for 10 000 previously collected gene donor's DNA samples. All gene donors have signed consent forms.
The aim of the re-use of the data is to validate the ability to develop, improve and use AI-based risk prediction tools to identify high-risk individuals.

What types and formats of data will the project generate or re-use?

No new DNA will be collected during the project. For whole genome sequencing, long-read and short- read sequencing is used for already collected DNA samples of gene donors. Project will generate
sequencing data in .map and .ped format that later will be converted to .vcf, .bed and .bim format.
Electronic health records are imported and updated on participants based on the consent, increasing and benefitting from the secondary use of real worlds health data. New additional ethics approval is first
obtained for general population health data.
During the clinical studies will be collected different types of data that will be collected in REDCap database and clinical data will be stored in e-tervis portal:
1. Cholesterol level in blood samples (LDL, HDL), and blood sugar;
2. The patient's blood pressure is measured by the family doctor or nurse;
3. There will be two types of questionnaires: for patients to assess their health behavior, and for family doctors and nurses on training and study processes

What is the purpose of the data generation or re-use and its relation to the objectives of the project?

The generation of the sequencing data aims to develop tools that will allow earlier diagnosis of diseases and adoption of better treatment strategies, leading to better health outcomes, increased patient
survival rates and reduced treatment costs for healthcare systems.

What is the expected size of the data that you intend to generate or re-use?

The expected size of the data within the project is:
• Sequencing data: we are going to re-use and generate new genomic data. The total expected size

of the data within the project would be 2100 TB. However, it might change if there is a need to vary some technological aspects during the project period.
• Health data - is obtained from health records collected through routine practice, ranging from 15-20 years and updated on annual basis for new updates. On average, this constitutes data from 3-5 encounters per year per donor. Health data is originally in various formats, central health data registry uses HL7 v3 CDA XML formats; hence ca 10 million documents related to discharge summaries and medical notes; procedures, prescriptions and medication purchases, as well as registry data from death, cancer, etc registries.

What is the origin/provenance of the data, either generated or re-used?

The origin of the data is:
1. DNA previously collected from gene donors.
2. Data generated by DNA sequencing.
3. Real world observational data from electronic health records, prescriptions, claims and registry entries.
4. Surveys collected from study participants.
5. Data generated by cost-effectiveness analysis.

To whom might your data be useful ('data utility'), outside your project?

Outside the TeamPerMed project the data would be useful for policymakers to show that our research can help to cost-effectively improve the quality of healthcare systems by influencing improvements to medical guidelines.
During the project we will organize training courses and workshops for healthcare providers to share more knowledge about personalized medicine and how these tools can greatly improve the work of healthcare professionals by helping to identify patients in need of healthcare interventions and supporting professionals in making decisions regarding interventions, treatment and medications, improving the healthand well-being of patients.
In addition to healthcare providers, we will educate the general public how project can improve the health and well-being of society. People will be educated about what it means to be in high-risk groups and the importance of taking action to prevent negative health outcomes.
Data generated will be useful for personalized medicine industry. They would have access to the Estonian Biobank's genomic, clinical and life-style data, highly developed e-health infrastructure. We can serve as both a gateway to the ecosystem for outsiders as well as a resource for those working within it, providing support for personalized medicine tool development, clinical and socio-economic validation and regulatory approval.

FAIR data

2.1. Making data findable, including provisions for metadata: Will data be identified by a persistent identifier?

Personal data will be identified by a persistent identifier; however, it is not accessible by third parties. The pseudonymization code is 16 digits and is generated by the computer using random numbers and letters. The code is stored securely on a server with restricted access and no internet

connection. Only authorized employees of the Estonian Biobank have access to personal data, which is necessary for updating the database. The personal data of the subjects are protected from parties not involved in the research, and the investigators have an obligation to ensure compliance with data protection requirements. Therefore, linking health data to an individual gene donor's person is not possible.
Researchers can analyze only pseudonymized data when conducting the study.

2.1. Making data findable, including provisions for metadata: Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

The project data is processed according to the Human Genes Research Act (HGRA), GDPR, Estonian Personal Data Protection Act and Estonian health Services Organisation Act and participants informed consent.
In the context of the TeamPerMed project the metadata will be the descriptions of the samples, methods used, discovered phenotypes and other relevant information. We are using OMOP (Observational Medical Outcome Partnership) Common Data Model (CDM), which is designed to standardize the structure and content of observational data and to enable efficient analyses that can produce reliable evidence; SNOMED, which determines global standards for health terms; LOINC-codes which are unique codes for laboratory analysis in Estonia; ICD-10 which stands for International Classification of Diseases 10th
revision; ATC which stands for Anatomical Therapeutic Chemical classification system for medications. OMOP CDM data allows for collaborations using common scripting, cohort definition and characterization, and patient level prediction tools, while only the aggregated data is shared with other partners.
Gene Ontology (GO)- the mission of the GO consortium is to develop an up-to-date, comprehensive, computational model of biological systems, from the molecular level to larger pathways, cellular and organism-level systems. The Gene Ontology knowledgebase provides a computational representation of our current scientific knowledge about the functions of genes (or, more properly, the protein and non-coding RNA molecules produced by genes) from many different organisms, from humans to bacteria. It is widely used to support scientific research and has been cited in tens of thousands of publications. GO ontology is the logical structure describing the full complexity of the biology, comprised of the ‘classes’ (often referred to as ‘terms’) for the many different kinds of biological functions, the pathways carrying out different biological programs, and the cellular locations where these occur; and equally important, the different types of specific relationships that indicate how each of these classes is related to other classes.
ISO 8602- standard that eliminates the confusion between European and American date formats.
Publication data: We use Dublin Core metadata standards for our files to provide basic information about them. README file will include information about the project, the author's information and contact
information and a general overview of the dataset and project results when necessary. No private date will be included in the publications, only aggregated summary tables and figures.

2.1. Making data findable, including provisions for metadata: Will search keywords be provided in the metadata to optimize the possibility for discovery and then potential re-use?

Our metadata is findable in GIFT as a table where all the phenotypes and tables are described.
There are two options to review this data:

1. Table can be downloaded in necessary format and opened with respective software.
2. Table can be viewed in a browser.
Table doesn't have a search engine or keywords, however, it's possible to use filters to review only desired data.
In order to have an access to the table in GIFT, person has to apply for access rights to work with databases. Applicant has to send an e-mail to the coordinator of data releases with the following information:
1. The name and Estonian ID code of the applicant;
2. Existinf UT username;
3. UT e-mail address (preferably in form first name.surname@ut.ee)
4. What data is requested to access;
5. The name of the applicant's supervisor;
6. Add Estonian Committee on Bioethics and Human Research (EBIN) permit and project description as an attachment.
The coordinator of data releases checks whether the conditions for obtaining the access right(s) are met, formalizes the application, and organises the signing in the Document Management Information System (DHIS). After the signing, the coordinator of data releases directs the work of granting access(es) to the access providers in the DHIS environment.
Access providers create access(es) for applicant:
1. GEVA: access is provided by the OPPA administrator at the proposal of the release coordinator. The operational area (OPPA) system sends a notification to the applicant about the creation of access. The GEVA OPPA is for accessing the general phenotype database.
2. HPC: to grant an access to the HPC cluster, the head of the bioinformatics department sends a message to HPC support and adds the applicant as a co-recipient. Accesses are granted no longer than the validity of the EBIN permit. HPC support provides access data to the applicant via e-mail. HPC cluster is for working with genetic data.
In addition to GIFT table, we our data sets have been described in BBMRI cohort corresponding to OMOP standard.

2.1. Making data findable, including provisions for metadata: Will metadata be offered in such a way that it can be harvested and indexed?

EstBB data is stored in GEVA, that offers an option to use different filters to obtain needed datasets. Descriptions on our data can also be found in BBMRI cohort corresponding to the OMOP standard with assigned code. We use Dublin Core metadata standards for our files to provide basic information about them. README file will include information about the project, the author's information and contact information and a general overview of the dataset and project results when necessary. No private date will be included in the publications, only aggregated summary tables and figures.

2.2. Making data accessible - Repository: Will the data be deposited in a trusted repository?

Personal data is stored securely on a server with restricted access and no internet connection. Only authorized employees of the Estonian Biobank have access to personal data, which is necessary for updating the database. The personal data of the subjects are protected from parties not involved in the research, and the investigators have and obligation to ensure compliance with data protection requirements. Therefore, linking health data to an individual gene donor's person is not possible. Researchers can analyse only pseudonymised data when conducting the study.
Pseudonymised data is stored in GIFT and can be accessed only with the Estonian ID-card.

The health data will be stored according to the "Health Services Organisation Act" issued by the Ministry of Social Affairs of Estonia.
Health data processing happens in trusted, secure, private cloud instances of Estonian Scientific Computing infrastructure, the SAPU servers.

2.2. Making data accessible - Repository: Have you explored appropriate arrangements with the identified repository where your data will be deposited?

The study in conducted in accordance with the operating procedures of Estonian Biobank and relevant legislative acts. The personal data is stored securely on a server with restricted access and no internet connection. Only authorized employees of the Estonian Biobank have an access to personal data, which is necessary for updating the database.

2.2. Making data accessible - Repository: Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object?

GEVA- EstBB's data management system that creates pseudonymized or anonymized cohorts based on the genetic and phenotypic data provided by gene donors under consent, which are suitable for genetic research, conforming to international standards such as OMOP. These cohorts serve as the basis for genetic studies. GEVA ensures the privacy of gene donors and the lawful use of data. The database is based on Quretec's standard solution. All the data has assigned an identifier.
In collaboration with the EstBB data, the institute of computer science has developed an additional database that corresponds to the OMOP standards.

2.2. Making data accessible - Data:

Will all data be made openly available? If certain datasets cannot be shared (or need to be shared under restricted access conditions), explain why, clearly separating legal and contractual reasons from intentional restrictions. Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if opening their data goes against their legitimate interests or other constraints as per the Grant Agreement.

Only summary-level research data is planned to be published in international peer-reviewed scientific journals. Genetic, health and personal data of the participants are used in this research and therefore can't be shared with external parties. Only pseudonymised data will be shared with third parties.
The project data is processed according to the Human Genes Research Act (HGRA), GDPR, Estonian Personal Data Protection Act and Estonian health Services Organisation Act and participants informed consent.

2.2. Making data accessible - Data:

If an embargo is applied to give time to publish or seek protection of the intellectual property (e.g. patents), specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

The data will be shared internally between the consortium members. Consortium Agreement

conditions apply.

2.2. Making data accessible - Data:

Will the data be accessible through a free and standardized access protocol?

Due to the nature of the data, it will not be accessible through a free and standardized access protocol.

2.2. Making data accessible - Data:

If there are restrictions on use, how will access be provided to the data, both during and after the end of the project?

nstitution accessing the data has to have an approval of the Estonian Committee on Bioethics and Human Research to conduct their study with the pseudonymised data of the Estonian Biobank. They have to submit an application to have an access to the SAPU. SAPU is an isolated environment which connection to the outside world is secured through VPN, firewall and IDS. Within SAPU, controlled data will flow according to set rules.
We will upload only pseudonymised data to SAPU, all the data processing will take place in SAPU. Persons who were gave an access to SAPU, can't move data out of the SAPU. In order to move the results of the data processing they need to contact us to move the results outside the SAPU.
Access to health data will be available according to relevant legislation acts.

2.2. Making data accessible - Data:

How will the identity of the person accessing the data be ascertained?

In Estonia, the accessing person is identified by the electronic ID-card and the access is only for authorised individuals. Within Europe, it's also possible to use Life Science Research Infrastructures (LS-RI) or ELIXIR-Europe to prove the identity of the person.
Data located in the HPC server will be accessed using SSH Keys.

2.2. Making data accessible - Data:

Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)?

In order to get an access to the data, the applicant must have to have an ethics approval from the Estonian Committee on Bioethics and Human Research. In addition, the application has to have an approval of the scientific council of the Institute of Genomics of the University of Tartu.

2.2. Making data accessible - Metadata:

Will metadata be made openly available and licenced under a public domain dedication CC0, as per the Grant Agreement? If not, please clarify why. Will metadata contain information to enable the user to access the data?

According to the Personal Data Protection Act it's mandatory to share public information on what data is used.

2.2. Making data accessible - Metadata:

How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available?

Data will be available as long the obtained ethics approval allows. If the data will be part of the Estonian Biobank it will be stored indefinitely, unless they are destroyed based on the gene donor's request.

2.2. Making data accessible - Metadata:

Will documentation or reference about any software be needed to access or read the data be included? Will it be possible to include the relevant software (e.g. in open source code)?

All data will be accessible through open source software.

2.3. Making data interoperable:

What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data interoperable to allow data exchange and re-use within and across disciplines? Will you follow community-endorsed interoperability best practices? Which ones?

We are using OMOP (Observational Medical Outcome Partnership) Common Data Model (CDM), which is designed to standardize the structure and content of observational data and to enable efficient analyses that can produce reliable evidence. BBMRI is using OMOP CDM and as we are part of it's consortium, we are following the same standards.
Estonia has launched OHDSI Eestonia node; while EMC represents OHDSI European central node; the OHDSI OMOP data is promoted both in Estonia and Europe-wide.

2.3. Making data interoperable:

In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining or extending them?

Estonian Biobank collects medical data from different sources and uses common ICD-10, ATC and NCSP ontologies. However, we use local ontologies like Estonian Health Insurance Fund's pricelist and codes of hospital-specific medical services. These will also added into OMOP as external vocabularies. There is no direct mapping to SNOMEd but we will explore those possibilities as well.

2.3. Making data interoperable:

Will your data include qualified references[1] to other data (e.g. other data from your

project, or datasets from previous research)?

[1] A qualified reference is a cross-reference that explains its intent. For example, X is regulator of Y is a much more qualified reference than X is associated with Y, or X see also Y. The goal therefore is to create as many meaningful links as possible between (meta)data resources to enrich the contextual knowledge about the data. (Source: https://www.go-fair.org/fair-principles/i3-metadata-include-qualified-references-metadata/)

We are working on including qualified references to other data. Possible ways to do so is to have different filters, for example: relation (is original form of, is complied by etc.), identifier, scheme (DOI, ISBN, GRID etc.), resource type (dataset, event, image, diagram etc.).

2.4. Increase data re-use:

How will you provide documentation needed to validate data analysis and facilitate data re-use (e.g. readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, units of measurement, etc.)?

To validate data analysis and facilitate data re-use, we provide readme files with information on methodology, codebooks, data cleaning, analysis, variable definitions etc.

2.4. Increase data re-use:

Will your data be made freely available in the public domain to permit the widest re-use possible? Will your data be licensed using standard reuse licenses, in line with the obligations set out in the Grant Agreement?

Our data will not be made freely available in the public domain. We are not allowed to make all the data publicly available, this is forbidden with different legislative acts, such as Personal Data Protection Act, Human Gene Research Act, GDPR. Genetic data is considered to be special category of data and processing it is regulated.

2.4. Increase data re-use:

Will the data produced in the project be useable by third parties, in particular after the end of the project?

Only pseudonymised data will be usable by third parties. In addition, they need an approval of Estonian Committee of Bioethics and Human Research.

2.4. Increase data re-use:

Will the provenance of the data be thoroughly documented using the appropriate standards?

Appropriate standards will be followed for the provenance of data.

2.4. Increase data re-use:

Describe all relevant data quality assurance processes.

The data quality will be assured according to ISO 17025:2017. Information security (InfoSec) wise we are planning to obtain the ISO 21001 standard.

2.4. Increase data re-use:

Further to the FAIR principles, DMPs should also address research outputs other than data, and should carefully consider aspects related to the allocation of resources, data security and ethical aspects.

The data sharing will follow the legislative acts (GDPR, Human Genes Research Act, Personal Data Protection Act), as well as Consortium Agreement and Estonian Committee on Bioethics and Human Research approval.

Other research outputs

In addition to the management of data, beneficiaries should also consider and plan for the management of other research outputs that may be generated or re-used throughout their projects. Such outputs can be either digital (e.g. software, workflows, protocols, models, etc.) or physical (e.g. new materials, antibodies, reagents, samples, etc.).

In given project we will collaborate with the High Computing Center (UTHPC) in order to store and analyse pseudonymised data. The UTHPC is a consortium inside of the University of Tartu with an aim to develop and maintain the infrastructure for scientific computing. One of the goals of the TeamPerMed project is to improve data processing and storage capacity. UTHPC is also a coordinator of the Estonian Scientific Computing Infrastructure roadmap activity.
During the project we will develop educational materials for healthcare professionals, students and public in general. Physicians, the gatekeepers for PerMed, will be introduced to novel approaches in genomics/pharmacogenetics, disease risk assessment and the use of decision support tools. They also will receive training on methods of patient counseling, and motivational interviewing to ensure patient acceptance and satisfaction with PerMed. Training will be offered in cooperation with the Ministry of Social Affairs and National Institute of Health and will be adjusted for use in the Nordic/Baltic countries thereafter. We will develop a PerMed course, which will introduce PerMed to medical and other students, focusing on both sides of the topic: genomics/data and clinical/health. We will organize TeamPerMed workshops, lectures and seminars engaging general public.
For the successful implementation of the project a clinical protocols for the pilot studies will be established in collaboration with clinical partners of the project. The protocol will be designed to be as accurate as possible in order to use it in the future.

Beneficiaries should consider which of the questions pertaining to FAIR data above, can apply to the management of other research outputs, and should strive to provide sufficient detail on how their research outputs will be managed and shared, or made available for re-use, in line with the FAIR principles.

During the project we are going to develop different protocols and models.
We are developing four TeamPerMed optimized PerMed research protocols based on 4 different working group: data, genomics, clinical and socio-economic. CoE will work to harmonize genomics research of all partners by sharing best practices and engaging in mutual learning. Four written protocols for methods will be produced to guide TeamPerMed research over time, enabling greater collaboration between partners.
TeamPerMed partners will share experiences in running localized clinical trials on personalized medicine tools such as PRS, PGx, decision support tools, medical devices, etc. Clinical trial methodologies for each type of study will be modified as needed and reflection papers based on these studies will be developed to facilitate regulatory approval.
2 PerMed genomic risk prediction tools/clinical decision support tool prototypes will be validated/developed. The curricula in partner institutions will be reviewed to identify where gaps exist in terms of the PerMed topic. A strategy will be developed that identifies these gaps and offers a strategy for filling them. This will include updating existing curricula by creating new interdisciplinary courses (including MOOCs) that combine the different elements of personalised medicine (genomics, data clinical, health, etc.).
We will develop or upgrade training materials upon the protocols, guidelines and materials developed by 4 research working groups and the research results of the clinical pilot studies. Training materials will be made available in different formats and via various media channels. All training materials will be created in Estonian and English, so they could be adapted to other national contexts.
TeamPerMed videos and marketing materials will be developed to support the marketing activities of the Centre of Excellence. The website of the CoE is designed for an easy transition into the permanent Centre for Personalised Medicine site.
Draft national guidelines on how to implement genetics-informed CVD risk evaluation in the medical system. White papers to the policy-makers on improving healthcare, research and innovation policy.

Allocation of resources

What will the costs be for making data or other research outputs FAIR in your project (e.g. direct and indirect costs related to storage, archiving, re-use, security, etc.) ?

Storage costs per year are 110€/TB, personnel costs related to optimizing pipelines and updating the server’s software will be 2000€/year. We are continuously updating our security measures to keep the data safe and secure on our servers. During the project we will use resources allocated for the implementation of the project. After the end of the project the University of Tartu will cover the corresponding costs.

How will these be covered? Note that costs related to research data/output management are eligible as part of the Horizon Europe grant (if compliant with the Grant Agreement conditions)

Costs related to the FAIR (findable, accessible, inoperable, reusable) will be covered by the European Union funding during the project period (01.09.2023-31.08.2029).
After the end of the TeamPerMed project the corresponding costs will be covered by the University of Tartu.

Who will be responsible for data management in your project?

The Institute of Genomics of the University of Tartu will be responsible for data management, and this includes data capture, metadata production, data quality and data sharing.
University of Tartu High Performance Computing Center is responsible for data storage and backup.
The clinical trial will require the involvement of the University of Tartu as the EGC's data controller as well as partner institutions and family physicians that will be conducting the clinical trial. The family physicians will be engaged under a contract for services (researcher's contract).
All TeamPerMed consortium members are responsible for correct and cooperative data management according to the principles listed in the consortium agreement.

How will long term preservation be ensured? Discuss the necessary resources to accomplish this (costs and potential value, who decides and how, what data will be kept and for how long)?

After the end of project, the resources for long term preservation will be covered by the University of Tartu. Approximate cost for storage is 110€/TB.
Under the Human Genes Research Act §18 Storage of tissue samples, descriptions of DNA and descriptions of state of health:
(1) Tissue samples, descriptions of DNA and descriptions of state of health shall be preserved in pseudonymised form or unpseudonymised form without a term if these must not be destroyed in the form provided for in subsection 21 (1) or (2) of this Act.
For data collected during research, the data will be stored in secure servers of the University of Tartu for up to 5 years after the end of the clinical trial. Once 5 years has passed, all data collected and stored about the clinical trial will be destroyed.

Data security

What provisions are or will be in place for data security (including data recovery as well as secure storage/archiving and transfer of sensitive data)?

All project data is stored, and all analyses are performed in the sensitive data private research environment (SAPU) provided by the University of Tartu High Performance Computing Center.
During the project, a dedicated SAPU environment is used, which is not connected to other SAPU environments. Detailed information about the High Performance Computing Center of the University of Tartu can be found at https://hpc.ut.ee and additional information about the SAPU can be found at https://docs.hpc.ut.ee/public/services/SAPU
Special and separate user accounts are created in the SAPU environment only at the request and approval of the project manager as needed (e.g. new researchers are added to the research group).
The SAPU environment is available to authorized persons only during the active analysis phase, and at other times the environment is turned off and it is not possible to enter SAPU even if there is an active user account. Third parties (including other employees of the University of Tartu) do not have access to the SAPU environment.
Encrypted and automatic backup copies of SAPU environment are made regularly (once a week) to the tape robot of the High Performance Computing Center of the University of Tartu, which is physically located in another location (another data center). The High Performance Computing Center of the

University of Tartu is responsible for making encrypted automatic backup copies of SAPU. Backups are encrypted and only the last three backups are kept. Incremental backup is used for backups, and it is guaranteed that previous backups cannot be changed.
At the end of the project, the SAPU along with the data will be deleted. All SAPU backup copies will also be deleted. Deletion is carried out by the High Performance Computing Center of the University of Tartu in cooperation with the responsible researcher to ensure permanent and irreversible data deletion.
Data deletion is documented, and a data deletion act is drawn up.
Information technology security measures:
- The requirements established by the Estonian information security standard (E-ITS), which are equivalent to ISO/IEC 27001 requirements, are followed.
- Servers are regularly tested, updated and monitored.
- Monitoring, machine learning and various penetration tests are used to detect vulnerabilities. Among other things, already existing vulnerability detection software (for example Nessus) is used to detect vulnerabilities, and various vulnerability lists are continuously monitored.
- Assigning rights is based on the principle of minimality, and administrator access is not enabled by default.
- Servers are regularly scanned and network traffic is monitored on an ongoing basis.
- By default, all activities that are not directly necessary to perform the work are prohibited.
- Encrypted backups are used and they are physically located in another location.
- All user activities in the SAPU environment are logged.
- All potential security incidents and attempts to find security vulnerabilities are logged (for example, login attempts, requests to different ports, changes in user rights, etc.).
- Moving information/data out of the SAPU environment is only possible if the responsible researcher has reviewed the relevant data and given consent.
- It is not possible to copy information/data from the SAPU environment ("copy" command).
- The SAPU environment is located behind a separate firewall.
- Internet access from the SAPU machine is completely closed and it is not possible to make requests to the Internet.
- The SAPU environment has pre-installed software, and the user cannot install the software in the environment himself.
- Moving data (including analysis results) out of the SAPU environment requires the consent of a third party (data owner).
- During periods when the SAPU environment is not in use, the environment is switched off and it is not possible to enter the environment.
- In place are technical and organizational measures to ensure information security and data protection at the High Performance Computing Center of the University of Tartu. A selection of technical and organizational measures (for security reasons, not all technical and organizational measures have been disclosed) can be found on the website of the High Performance Computing Center of the University of Tartu https://hpc.ut.ee/terms/information-security

Ethics

Are there, or could there be, any ethics or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).

In order to conduct the clinical pilot studies, we have to get ethics approval from the Estonian Committee on Bioethics and Human Research. Data collection and sharing will be in accordance

with relevant legislative acts.
1. Data that will be re-used during the project:Genomic data collected previously from the gene donors will be re-used within the project. All gene donor have signed a consent form that allows to use their data in research. All project re-using the data have to obtain an ethics approval from the Estonian Committee on Bioethics and Human Research. Data collection, storage and re-use is in accordance with relevant national and European legislative acts, such as Human Genes Research Act (HGRA), Personal Data Protection Act and GDPR.
2. Data that will be collected and created during the project:
The basis for the processing of genetic data is provided by the Human Genes Research Act (‘Inimgeeniuuringute seadus’ – IGUS, adopted on 13.12.2000, RT I 2000, 104, 685 (current version: RT I, 14.03.2014, 30)).
For each subject, a personal data file will be prepared, both on paper and in an electronic format, containing the subject’s informed consent form and the data collection form each visit.
Data for the clinical trial will be collected in personally identifiable form. An electronic data collection and analysis platform will be developed for the trial for collecting and analyzing data. The EGC will assign a personal identification code (ID) to each subject, which will enable each subject’s health data to be monitored over time but will not enable the individuals to be identified outside the framework of the trial. Any release and analysis of data will be based only on the ID codes and not on the names of the subjects. To ensure confidentiality, access to personal data (name, date of birth, other personal and medical data) will be restricted to the researchers directly responsible for conducting the trial. The linking of the personal data in the EGC’s database with the EHIF’s database and the register of causes of death will be performed according to a specific procedure agreed with the EGC and the administrators of the databases to ensure the inaccessibility of personal data to third parties. The EGC has received approval from the Research Ethics Committee of the University of Tartu for linking the data with these databases.

Will informed consent for data sharing and long term preservation be included in questionnaires dealing with personal data?

Informed consent for data sharing and long term preservation was signed by the gene donors before the collection of the DNA sample.
Additional informed consent will be collected from the study participants before the clinical study. It will be signed on paper by all participants of the clinical trial. The informed consent form for subjects includes consent for encoding and decoding of data. Under § 24 (2) 4) of the Human Genes Research Act (IGUS, adopted on 13.12.2000, RT I 2000, 104, 685 (current version: RT I, 14.03.2014, 30)), decoding is permitted on the proposal of the data controller and with the approval of the Ethics Committee to identify a gene donor, contacting them or updating, supplement or verifying a description of their state of health with their written consent. Only gene donors who meet the inclusion criteria will be included in the trial. Gene donors who have prohibited their data from being updated, supplemented, or verified will not be included in the trial. Based on the consent form signed by individuals participating in the clinical study, only the persons conducting the study (family physicians and nurses) and members of the research team authorized under the ethics approval, have access to the study data. Third parties do not have access to that data. Only analyzed data, which are anonymized and not individually identifiable, are shared with them at the end of the study.

Other issues

Do you, or will you, make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones (please list and briefly describe them)?

We will follow different procedures for data management.
Human Genes Research Act (HGRA)- The objectives of this Act are to regulate the establishment and maintenance of a Gene Bank, to organise the genetic research necessary therefor, to ensure the voluntary nature of gene donation and the confidentiality of the identity of gene donors, and to protect persons from misuse of genetic data and from discrimination based on interpretation of the structure of their DNA and the genetic risks arising therefrom.

European GDPR- Subject-matter and objectives
1. This Regulation lays down rules relating to the protection of natural persons with regard to the processing of personal data and rules relating to the free movement of personal data.
2. This Regulation protects fundamental rights and freedoms of natural persons and in particular their right to the protection of personal data.
3. The free movement of personal data within the Union shall be neither restricted nor prohibited for reasons connected with the protection of natural persons with regard to the processing of personal data.

Estonian Personal Data Protection Act- This Act regulates:
1) protection of natural persons upon processing of personal data to the extent in which it elaborates and supplements the provisions contained in Regulation (EU) 2016/679 of the European Parliament and of the Council on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (OJ L 119, 04.05.2016, pp. 1-88);
2) protection of natural persons upon processing of personal data by law enforcement authorities in the prevention, detection and proceedings of offences and execution of punishments.

Estonian Health Services Organisation Act- Scope of application of Act:
1) This Act provides the organisation of and the requirements for the provision of health services, and the procedure for the management, financing and supervision of health care.
2)This Act applies to the organisation of the provision of health services in the area of government of the Ministry of Defence with the specifications arising from the Defence Forces Service Act and Estonian Defence Forces Organisation Act.
(3) The provisions of the Administrative Procedure Act apply to administrative proceedings prescribed in this Act, taking account of the specifications provided for in this Act.
Estonian information security standard (E-ITS)- The purpose of E-ITS is to develop and promote the level of information security of the Estonian public authorities as well as private businesses. Previously, the information security system ISKE was used for the same purpose. The intention is also to make dealing with information security more manageable for smaller organisations.The requirements established by the Estonian information security standard (E-ITS), which are equivalent to ISO/IEC 27001 requirements, are
followed.