# ON LiMiT Data Management Plan

Kristiane Beicher Daniel Ibsen Luke Johnston

# Table of contents

# Welcome!

> **Note**
>
> This is **version 0.3.0** of the Data Management Plan. Every change to this document will update the version number. To see a list of the changes made, check out Appendix B.

> **Tip**
>
> Some of the content in this document, in particular the diagrams, may be best viewed online at https://dmp.onlimit.org/.

ON LiMiT (Optimal Non-pharmacological Lifestyle Modifications in people with Type 2 diabetes) is a large Danish intervention study investigating the maintenance of remission of type 2 diabetes through weight loss, diet, and exercise. The project runs from 2025 to 2030 and will involve 1,500 participants with type 2 diabetes.

Although the ON LiMiT study has been classified as not being a clinical drug trial by the Danish Medical Agency (DMA), we have chosen to align with best practices for clinical trials wherever possible. While a Data Management Plan (DMP) is not mandatory for a study such as ON LiMiT, it is strongly recommended for all research studies. This DMP ensures transparency, compliance, and integrity in handling participant data throughout the project lifecycle.

We follow a modified version of a DMP template from DeiC, mainly around how we describe access to data.

## Collaborations

The project is being conducted in Aarhus, Odense, and Copenhagen and is a collaboration between Steno Diabetes Center Aarhus, Steno Diabetes Center Odense, Steno Diabetes Center Copenhagen, Bispebjerg Hospital, and the Department of Nutrition, Exercise and Sports and the Center for Basic Metabolic Research at the University of Copenhagen.

## Governance and maintenance

This document will be maintained by members of the study team and has been approved for publication on the project website by the Project Manager on behalf of the Chief Investigator.

## Funding

The project is funded by approximately DKK 102 million from the Novo Nordisk Foundation (Grant Number NNF24SA0097844).

# Part I

# Sections

# 1 Data sources and flow

ON LiMiT collects data from multiple sources, including biological samples, questionnaires, and electronic tools. Data originates from participants, healthcare professionals, and bioanalysts, and is stored securely on GenomeDK servers for subsequent analysis. The diagram below illustrates the context of ON LiMiT, key users, and data providers:

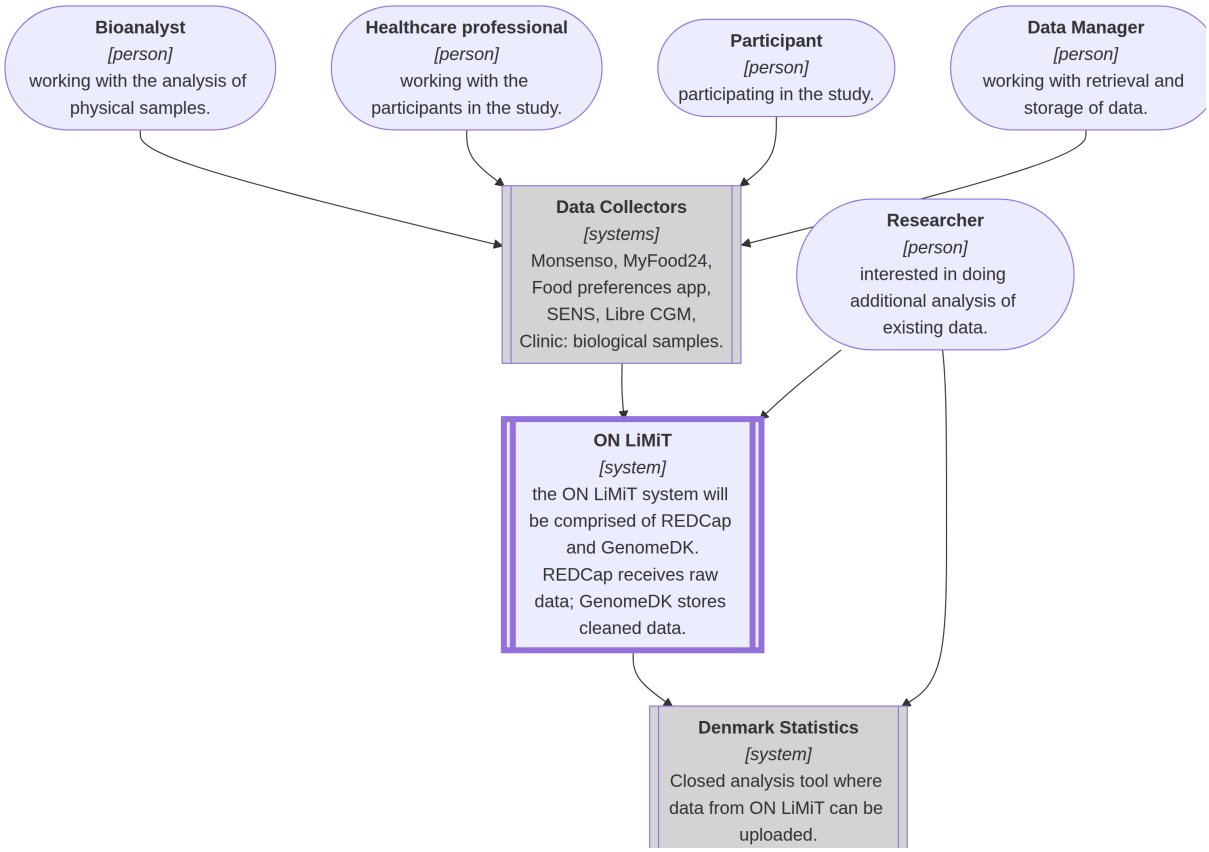


Figure 1.1: C4 Context diagram showing a basic overview of ON LiMiT, its anticipated users, and data providers.

## 1.1 Technical infrastructure

The Container diagram shows the main components of the ON LiMiT data infrastructure, including data sources, storage, and analysis tools. It provides an overview of how the data is collected, stored, and made available for analysis.

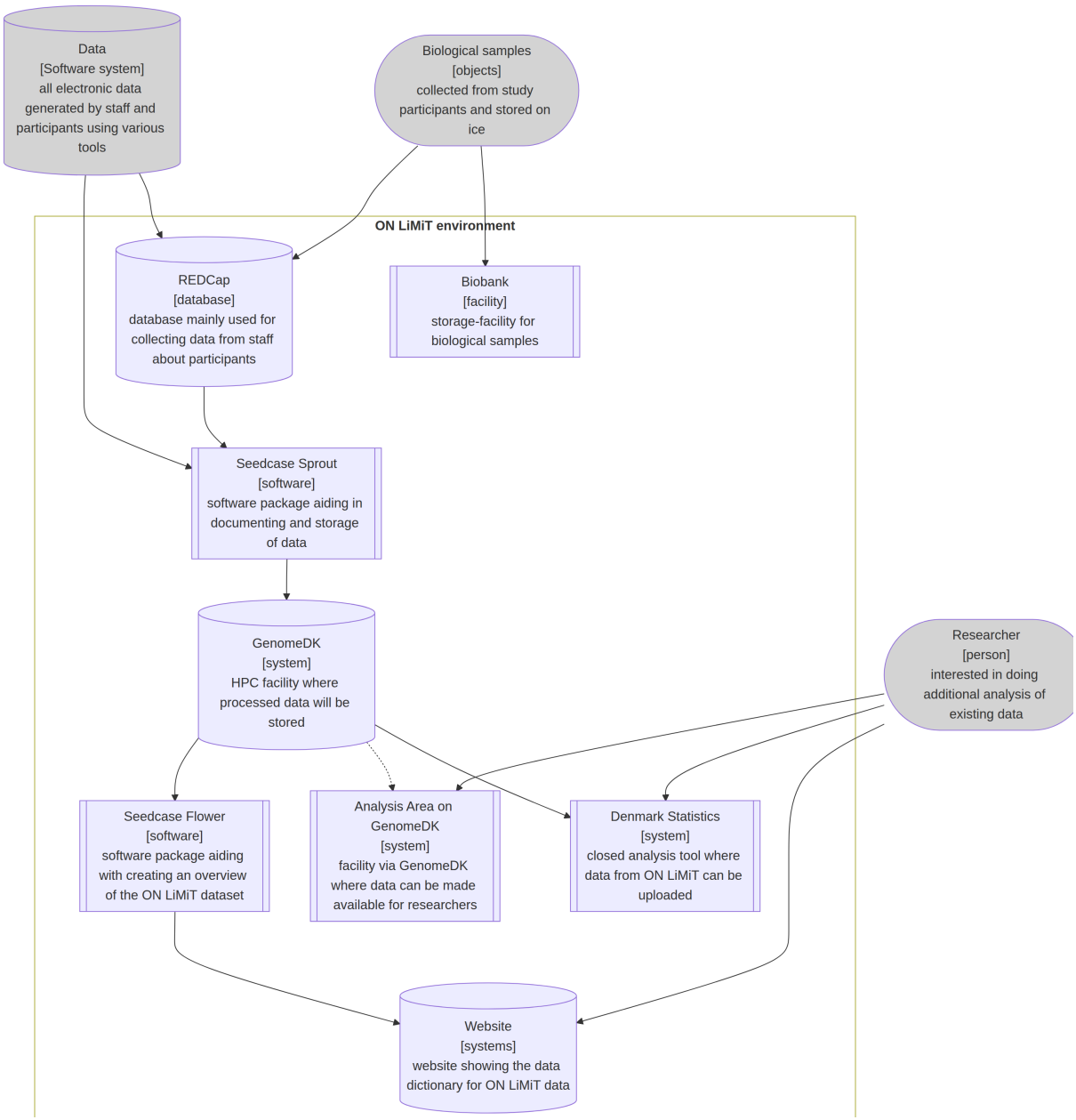


Figure 1.2: C4 Container diagram showing the main components of ON LiMiT data infrastructure, including data sources, storage, and analysis tools.

# 2 Data collection

Data for the ON LiMiT study will be collected from multiple sources, including participants, healthcare professionals, lab technicians, and electronic devices. This section describes what, where, and how these sources of data will be collected.

## 2.1 Sources and methods

Participants will provide data through a variety of sources, the primary being the Liva Healthcare app which will record lifestyle information in REDCap. Connected devices such as smart watch and smart scales will transmit data directly to the Liva app.

Additional data will be gathered with a SENS accelerometer and a blinded Libre CGM which will be uploaded either into REDCap or directly to GenomeDK. Some scanner results will be uploaded via REDCap (DXA and Fibro) others will be uploaded directly to GenomeDK (VO2Max and Finapres).

We will also be collecting data from MyFood24, which is a website where participants will be asked to keep a very detailed food diary for three days at a couple of points during the study. Some participants will also be asked to do a Food Preference test which we expect to upload into REDCap.

Healthcare professionals and lab technicians will enter clinical and biological sample data into REDCap forms, and laboratory data may be uploaded from internal systems where applicable.

Further details can be seen in the diagram below.

## 2.2 Frequency of collection

The aim is to collect participant app and web based data (Liva, RECap, MyFood24) on a monthly basis.

Sensor and scanner data (SENS, Libre, Fibro scans, and DXA scans) will be uploaded at clinic visits, as we expect to sign off on a visit as soon as it is finished.

## 2.3 Data formats

- REDCap: Structured tabular data in CSV imported to GenomeDK via API.
- Liva app: CSV exports via FTP to GenomeDK. The watch and scales data are directly imported into Liva app.
- MyFood24: CSV import from secure website into GenomeDK.
- Scanners: CSV files uploaded into REDCap or directly to GenomeDK.

## 2.4 Estimated raw data volumes for the feasibility study

- REDCap clinical data: Unknown size at present.
- Liva app data: Unknown size at present.
- MyFood24 dietary logs: Unknown size at present.
- Sensor data (SENS, Libre): Unknown size at present.
- Scanner data:
  - VO2Max: 20KB times 1 reading per visit, 4 visits = 20x4x24 = 2MB.
  - Finapres: 60KB times 8 readings per visit, 6 visits = 60x8x6x24 = 69MB.

## 2.5 Data aggregation and storage

The REDCap database serves as the central hub for most clinical and scanner data, including manual uploads from devices such as SENS and Libre, and potentially iMotions. The Liva app aggregates participant lifestyle data and connected device readings. There will also be test results (like VO2Max and Finapres) that will be sent to GenomeDK directly, and this is the same as with data from MyFood24.

GenomeDK's servers will be the final repository for processed and collated data, ensuring a central location for secure storage of data and computational capacity for analysis.

See the Storage, backup, recovery, and access for further details.

## 2.6 Quality assurance

We will set up and run automated checks in REDCap to ensure completeness and consistency of clinical data. The person responsible will be the Data Architect.

Uploaded sensor device data will be checked against expected ranges and timestamps. The people responsible will be the on-site staff who are using REDCap during data collection.

We will run checks to identify any missingness in the data or any anomalies and report on the results (if any). We will do any follow-up procedures to correct for these issues, which we will run during transfer to GenomeDK. The person responsible for this will be the Data Architect.

## 2.7 Visualisation of data flow

Underneath the diagram is a more detailed description of the components and their interactions.

### 2.7.1 REDCap

REDCap is a secure web application for building and managing online surveys and databases. It is used to collect data from participants and healthcare professionals in the ON LiMiT project. REDCap will have two main functions, the first is to collect data from clinicians and participants via forms and questionnaires, and the second is to collate data from other systems, likely through APIs or csv uploads.

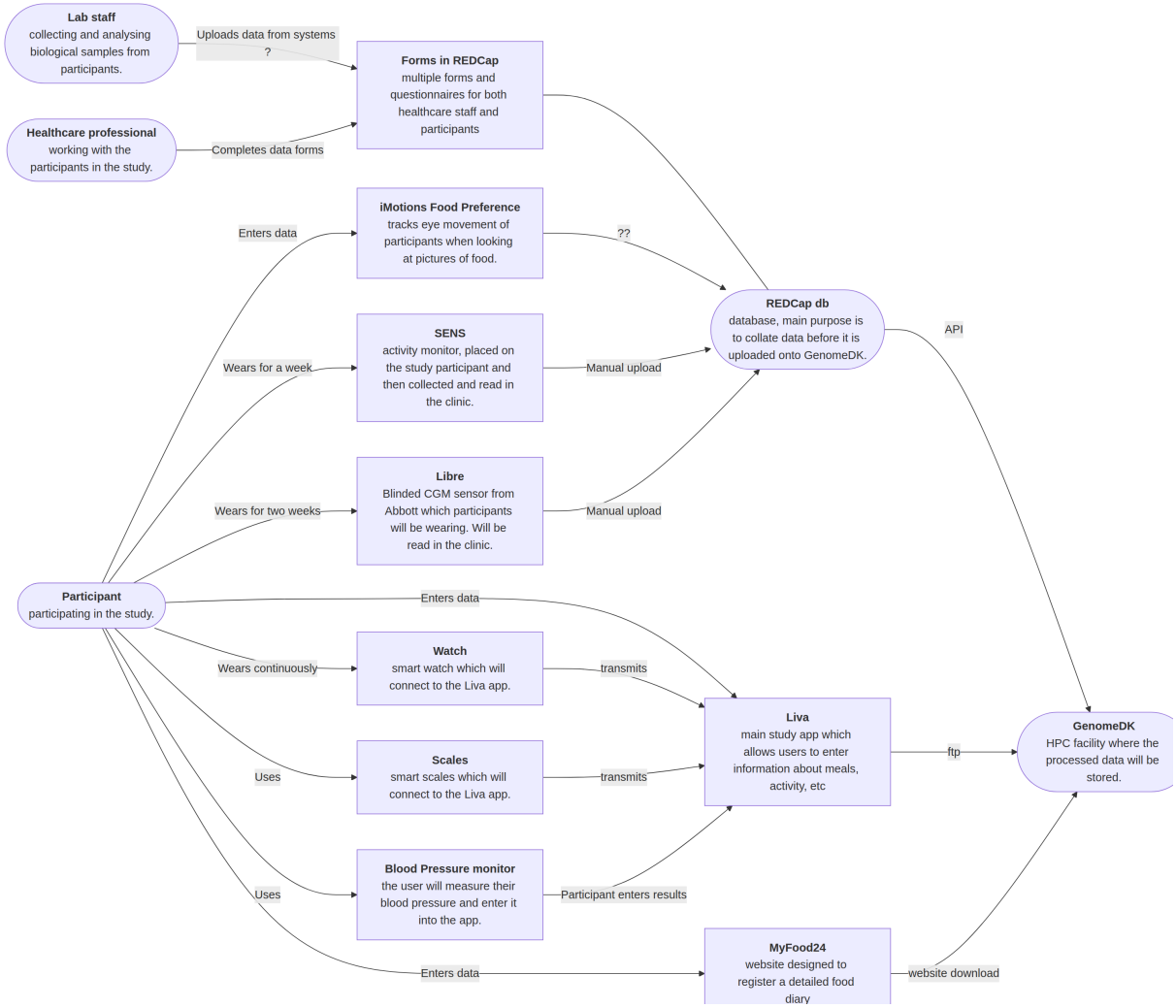


Figure 2.1: Data flow in the ON LiMiT study, who generates data, where is does it enter, and how is it transmitted to storage.

### 2.7.2 iMotions Food Preference

iMotion Food Preference is a system that tracks the eye movements of participants when they are looking at pictures of food. The data collected by this system can be pulled out as csv files and uploaded to the REDCap database or directly to the GenomeDK server.

### 2.7.3 SENS

SENS is an activity monitor that will be placed on the study participants for a week. The data collected by a SENS device will be extracted at clinic visits and the data will be uploaded to the REDCap database. The SENS device will track the participants' physical activity levels in selected weeks of the study.

### 2.7.4 Libre

Libre is a blinded continuous glucose monitoring (CGM) sensor from Abbott that will be worn by the participants for a week. The data collected by the Libre device will be extracted at clinic visits and uploaded to the REDCap database. The Libre device will track the participants' glucose levels in selected weeks of the study.

### 2.7.5 Liva Healthcare app

Liva is a mobile app that will be used by the participants to track their physical activity (via Garmin activity watch), adherence to the assigned diet intervention (via short diet screener), body weight (via Gamin scale), answer other small questionnaires, follow the study schedule, and interact with study personnel via a chat function. The app will likely connect to the REDCap database via an API, allowing the data to be automatically uploaded to the database.

### 2.7.6 Garmin watch and scales

The Garmin watch and scales will be used by the participants to track their physical activity and weight. The data collected by these devices will be transmitted to the Liva app where it will be collated and then transmitted to GenomeDK. The watch will track the participants' physical activity levels, while the scales will track their weight, both will submit their measurements to the Liva app.

# 3 Documentation and metadata

Clear and comprehensive documentation is essential to ensure that the data collected during the study is FAIR (Findable, Accessible, Interoperable, and Reusable) not only today but also in the future. This section outlines the approach to documenting all datasets and associated metadata, including the tools, standards, and processes that will be applied. By maintaining detailed descriptions of variables, data sources, and collection methods, and by implementing robust metadata management practices, we want to support data integrity, making it easier to use the data, and comply with FAIR principles.

## 3.1 Documentation

All data stored on the GenomeDK servers will be generated during the study, either by study staff or directly by participants. No external datasets (such as national registers) will be stored within the ON LiMiT environment. The primary sources of data include:

- Participant input via the Liva Healthcare app and responses to questionnaires.
- Medical devices and scanners, which will produce structured data files. A detailed list of scanners and devices will be maintained as part of the technical documentation.
- Additional sensors and site-level data collection, uploaded as CSV files by study staff.

Data generation will occur throughout the two-year participation period for each individual. While there are plans to link study data with register data, this work will take place on Statistics Denmark's servers. Access to the data will be strictly controlled through GenomeDK. Researchers must submit an application to the Publications Panel to gain access, ensuring transparency and governance.

## 3.2 Metadata strategy

To ensure data usability and interoperability, comprehensive metadata will be created and maintained throughout the study:

- Field-level documentation: We will write detailed variable descriptions using the *Field Annotation* functionality in REDCap. This includes adherence to naming conventions defined by internal rules.
- Supplier data dictionaries: All external data providers (e.g., Liva Healthcare, MyFood24) will supply data dictionaries in CSV format. These will be integrated into the metadata repository by the data architect.
- Device and sensor data: The site staff and the data architect will collaboratively document the CSV filesoriginating from scanners and sensors to ensure completenessand clarity.

- Metadata packaging: We will use software developed from the Seedcase Project to generate and store metadata in a standardized data package format. These software will also structure the data in a standardized format and will run checks on the data to confirm that it corresponds to the metadata. The data package will be stored on GenomeDK, where the data will be version controlled and a backup kept in case of errors or issues. The metadata will be connected to GitHub to increase accessibility and discoverability as well as improve future collaboration on analyses of the data.

## 3.3 Versioning and updates

Metadata will be treated as a living resource. Updates will occur whenever new variables are introduced, existing fields are modified, or additional data sources are integrated. Version control will be implemented through the Seedcase software, ensuring that:

- Each metadata package is assigned a unique version identifier.
- Changes are logged with timestamps and descriptions.
- Previous versions remain accessible for audit and reproducibility purposes.

This approach guarantees traceability and supports long-term data integrity.

## 3.4 Data discovery and accessibility

To facilitate data discovery, a dedicated website will be developed. This site will provide:

- A searchable catalog of all variables.
- Clear descriptions and metadata for each variable.
- A foundation for the Publications Panel application process, ensuring researchers can identify relevant data before requesting access.

This approach guarantees that data is well-documented, discoverable, and compliant with FAIR principles.

# 4 Ethics, security, and legal compliance

Ethical and legal considerations, including related to security, are central to responsible data management, particularly in health research involving sensitive personal information. This section outlines how we are planning to handle ethical standards, legal obligations, and data security measures.

## 4.1 Ethics and privacy

All participants will provide informed consent, which includes a clear statement that their data will be retained for a minimum of 15 years. The consent form also specifies that data may be shared with other research institutions, but strictly for research purposes.

To protect the privacy of our participant, personally identifiable data (PID) will never be made available to researchers during the analysis phase. PID will only be accessible to staff responsible for direct participant contact, those managing instruments within REDCap, or for managing the final data structure. For all analytical purposes, pseudonymised or anonymised data will be used. Requests for data extracts must be submitted via a formal application form, which requires applicants to justify their need for specific data points, especially if they involve PID.

In general, the principles for data extraction follows the current guidelines from Statistics Denmark. Individual-level data cannot be exported from the project database at GenomeDK. Only aggregated results and analysis scripts not including any personal identifiable data can be extracted.

## 4.2 Legal compliance

The project complies with all relevant legal obligations concerning data handling and sharing. A joint data controller agreement is in place between all participating sites, enabling secure and lawful data exchange within the main study and any sub-studies. Data from REDCap will be transferred to GenomeDK (GDK) via a secure API. Data from LIVA will be made available through an AWS FTP server and transferred directly to GDK. This should ensure that we comply with all legal requirements.

## 4.3 Data security

All data analysis will be conducted within the secure environment of GenomeDK (with the exception of the work that will take place in the secure environment at Statistics Denmark) . For each approved data access request, a dedicated project folder will be created. Researchers will be informed (via the data access application process) that raw data must not be extracted from GDK and that only analysis results and scripts may be exported. PID stored on GDK will be further protected through encryption using standard scripts available within the platform. If a study requires linkage with additional datasets from Statistics Denmark, PID will only be

used for secure transfer purposes and not for direct analysis. Following the conclusion of the feasibility study, we will evaluate whether to delete PID from REDCap or remove the entire dataset. A similar process will be applied to the main study once all data has been successfully transferred and verified on GDK.

Any suspected breaches of data security must and will be escalated immediately to the relevant point of escalation at Aarhus University or relevant authority.

# 5 Storage, backup, recovery, and access

As part of a DMP, consideration should be given to the physical storage and security of the data that a research project generates. This should be done to minimize the risk of accidental data loss, be it from overwriting raw data with an edited copy, or the loss of either a laptop or the breakdown of a disk. Some data will be easier to replace, data which contains measurements done in a clinic and entered directly on a laptop will be harder to recover than a data set downloaded from an online repository. It is therefore important when looking at a data management plan that we are careful with how we store data we have generated within the study, and that we do our due diligence in making sure that data we store is recoverable, should a disk fail, or a server stop working.

## 5.1 Storage of data and metadata

There are three main places we will store data from ON-LiMiT. The first two, where the data is entered or generated are REDCap and the LIVA application. These are not final storage areas, but they will hold data for long enough that we will need to take them into consideration. The final storage place for data will be GenomeDK, which also is the place where we are most responsible for setting up not only the storage, but also the subsequent safety and security of the data.

### 5.1.1 GenomeDK (GDK)

GDK provides several storage areas. We will use the project storage in their Open zone for the ON LiMiT project (see a description of the zones here). These locations are designed for collaborative data and are eligible for backup. We do not need a Data Processing Agreement with GenomeDK as our host organisations Aarhus University (AU) and Aarhus University Hospital (AUH) already have an agreement in place which will cover us.

### 5.1.2 REDCap

REDCap is a secure web application for research databases and e-forms. It stores data in a PostreSQL database and related files in a controlled file area ("File Repository"). AU provides REDCap facilities for both the University and AUH employees. The File Repository is intended for storing data and study files, but it should not be the primary long-term file storage for bulk data; we will export and archive data into our primary storage (GDK) using the API facility provided by REDCap. As with GDK we don't need a Data Processing Agreement for the same reasons as above.

### 5.1.3 LIVA

Liva Healthcare states compliance with GDPR and ISO/IEC 27001/27002 and NHS Data Security & Protection standards on their website. We understand that data is being stored on AWS services that are physically based in Frankfurt, Germany. The data will be made available to us via AWS on a ftp connection on request, we hope to transfer the data directly to folders on GDK, from where we will transform and load them into storage.

## 5.2 Backup and recovery

When looking at backup and recovery we again have two situations. With the LIVA system we will need to trust that what is stated in the contract is also what LIVA is doing, and that backup and recovery is possible without any significant loss of data. We also need to trust REDCap's system, as the setup is done entirely without any user engagement.

Where we will play a role is with GenomeDK, as described below it is the responsibility of the user to set up the files for backup, although it is not possible for us to check that a recovery is possible, as backup is still done by the GDK team.

### 5.2.1 GenomeDK

GDK runs its own disk-based backup at a remote site. By default nothing is backed up. There is guidance given on the GDK website which details how the backup should be set up. It is the responsibility of the individual study teams to set this up correctly, but help can be requested from the GDK team (in particular to check that the folders have been set up correctly).

Backups are conducted once per week and retained for 14 days and it is possible to review recent backup runs. Though, it is recommended that larger files are kept out of the backup structure and only added in compressed format. At present we don't believe that we will have any files that require this, but it will be something that we will keep under observation (see section on data volume).

We will back up the original raw data, the cleaned data in Parquet format, and the metadata. As we do not expect to do any analysis in the main study folder (see section on Sharing Data) we don't expect to be backing up any analysis scripts or results here. This will be done in separate project folders which will be set up to only back up scripts and results.

### 5.2.2 REDCap

There isn't any information on the website for REDCap at AU about backup and recovery, but we have received the following reply from the team maintaining the servers in answer to our questions about the procedures in place.

> "AU-IT performs daily backups of the REDCap application and the data contained within it. All backups are managed under IBM's Tivoli Storage Manager and stored in encrypted form. Ubuntu servers are configured to automatically install security updates. A monthly vulnerability scan is carried out using Tenable Security Centre. In accordance with Aarhus University's annual information security cycle, a system and data classification is conducted at least once a year, followed by a risk assessment. In addition to the regular annual classification, a new risk assessment is carried out whenever significant changes are made to the system."

### 5.2.3 LIVA

Backups, retention, recovery, and export options are governed by our contract with them. The Liva contract draft should specify backup frequency, encryption, location (EU/EEA), recovery time, and access/audit reporting. We will confirm these parameters during contract finalization and update this section once the contract is signed.

## 5.3 Access to the main data structures

It should be a limited number of people who have access to the full set of research data, which is currently the project manager and the Seedcase Project team, which includes the data architect, will have access to data on GDK as they will be helping with the transformation of raw data to Parquet files, and the creation of metadata. It should be made clear that no data analysis is to take place using the data without prior approval of project synopsis from the Publication Panel and from Project Management Team. There will also be staff at LIVA which will have access to their back-end data, this is covered by the contract.

### 5.3.1 GDK

Access is managed on a project folder basis. GDK maintains logs and states it records login and file access events to help ensure GDPR compliance. We have been informed that we will be receiving regular reports from the system once we start using the site. These will be reviewed and signed off by the project team when we receive them.

### 5.3.2 REDCap

REDCap provides a logging module that automatically records project-level activity (e.g., data exports, data changes, user creation/deletion). We will review logs on a monthly basis and after any major change. This will be done by the project team on the back of an automated email sent out by REDCap encouraging project owners to check through their log files and access controls once a month.

### 5.3.3 LIVA

We will rely on the Liva contract agreement for access control, audit reporting cadence, and data-sharing/export mechanisms. We expect that Liva will be able to provide us with either access to a list of users with data viewing access, or will be able to provide such a list on demand.

# 6 Selection and preservation

The long-term preservation of research data and biological samples is important to make sure we have reproducibility and compliance with ethical and legal requirements. This section outlines the strategy for selecting which data and samples will be retained, how it will be stored and for how long, as well as the mechanisms for ensuring secure storage and accessibility. All processes adhere to the FAIR principles and comply with GDPR to protect participant privacy and data integrity.

## 6.1 Retention and storage

As a general rule, all data generated by or from participants will be retained and stored on GenomeDK's secure servers. Exceptions may include more technical/system data such as log files, which have limited long-term value.

- Study data: All core datasets will be preserved for a minimum of 15 years after study completion. The long-term plan is to integrate these datasets into a joint Steno Database, currently under development, to ensure continuity beyond the funding period.
- Biobank samples: We will initially retain biological samples until the end of the current funding period. Plans are in place to establish a Steno Biobank before this point, where we will transfer ON LiMiT samples for extended preservation.

## 6.2 Study timeline

Data collection for the feasibility study will span approximately 13 months (12 months per participant), while the main study will involve a two-year participation window per participant, with the last participant's final visit expected in 2032. Retention commitments extend well beyond this timeline to support future research.

## 6.3 Access and future use

To maximise reuse, we will develop a searchable website to enable discovery of available variables and datasets, as described in the Documentation and metadata section. Initially, access is expected to be of particular interest to the study team. In accordance with the Publications Guidance, data will be made available to a wider research community once primary findings have been published, subject to an application and approval process.

## 6.4 Long-term strategy

Wherever possible, data that cannot be re-measured will be preserved beyond the funding period, either on GenomeDK servers or within the planned Steno Diabetes Centre database. This ensures that valuable datasets remain accessible for secondary analyses and future research, supporting sustainability and compliance with FAIR and GDPR principles.

# 7 Data sharing

Sharing data from research projects is essential for promoting transparency, reproducibility, and collaboration within the scientific community. By making data findable, accessible, interoperable, and reusable (FAIR), we enable other researchers to build upon existing work, validate and verify findings, and generate new insights. This section outlines how data and metadata from the project will be shared, under what conditions, and with what safeguards in place.

## 7.1 How will the data and metadata be shared?

We will make the metadata describing the structure, variables, and scope of the dataset publicly available through one of our project websites, using Seedcase software to build and structure it. This will allow researchers (internal and external) to explore the metadata before submitting an access request. The actual research data will not be publicly accessible and will only be shared upon formal application and approval, and only through a secure server.

## 7.2 When will the metadata be available?

We aim to start metadata publication as soon as data collection starts. A preliminary release will be based on the feasibility study, which we plan to use as a test case for the metadata infrastructure. While we primarily intend to use this dataset to identify any issues with the two most demanding arms of the study, it will also be used as a testing ground for the metadata sharing process. We do not expect high external demand for the feasibility data, but it will help ensure readiness for the main study.

## 7.3 Who is the audience for the data?

We anticipate that most data access requests will come from researchers affiliated with the study or collaborating institutions. However, external researchers may also apply, provided they meet the criteria outlined in the data access guideline. The audience may include clinical researchers, epidemiologists, or data scientists researching diabetes-related topics, as well as anthropologist or other social scientists looking at behaviour in relation to food and exercise for people with type 2 diabetes.

## 7.4 How will researchers get access to the data?

Researchers must submit a formal application detailing their planned use of the data, including which specific data points they require and why. Applications involving sensitive data categories must include a clear justification for their necessity (for instance in cases where the data is to be uploaded to Denmark Statistics). We will develop a guide on the application process, including eligibility criteria and review procedures and that we will make available as a website.

## 7.5 How will the data be made available?

Once the steering committee has approved an application, we'll make the requested data available in a dedicated project folder on GenomeDK (GDK). Raw data must remain on the GDK server, and only analysis scripts and results may be exported. Researchers will need to create an account with GDK to access the folder and work on their analysis environment. In some cases, it may be possible to upload a dataset to Statistics Denmark, under the SCDA Statistics Denmark project database. We only consider this option for projects requiring linkage with other national datasets.

## 7.6 How will access be controlled?

There are three main ways to access data, and only one is used for data analysis. Data will be available to a few members of the study team on REDCap where we do data collection. All data collected by the LIVA app will be available to clinicians involved with the running of the study in the LIVA system. Finally, the research data we have collected, cleaned, and stored on GDK servers is also where the analyses will take place. Access to the finalised research data will be controlled by the central study team on GDK, with each successful data application getting their own project folder with only the data requested. All the programming scripts used to create these data sets will be stored in the main study folder on GDK to ensure transparency and reproducibility. We expect that those who have access to the data collected within REDCap will not use the data for analysis and/or publication. The data collected by LIVA will be available through the use of the Clinician Portal and LIVA Reporting Module, which is encouraged for monitoring participant progress but should not be used for publication purposes.

## 7.7 What kind of IP or license will be used?

A formal decision regarding intellectual property rights and licensing for shared data is still pending. This will be finalized before the main study data is made available to external researchers.

## 7.8 Additional considerations

We may need to share some data from the feasibility study and analyse it before the launch of the main study, but the formal end points in the study registration on ClinicalTrials.com should help clarify what data points will be needed for this. This process should be planned and initiated as soon as possible. A data sharing agreement is already in place between the five participating centres, outlining responsibilities and procedures for internal data access and collaboration.

# 8 Responsibilities and resources

Effective data management relies to a great extent on clear guidelines as well as clear definitions of roles and responsibilities. By defining responsibilities and ensuring the right expertise and tools are in place, then data quality, compliance, and the long-term value of the project's outputs should be ensured. This section outlines the key roles, and resources dedicated to managing data within the study.

## 8.1 Responsibilities

This section outlines who is responsible for key aspects of data handling (from collection and documentation to long-term storage and sharing) and how collaboration between people and institutions supports the integrity and usability of the data throughout the study. For now we'll be looking at the first 6 years of the study (the active phase), but data will be stored for at least 9 years more (minimum of 15 years in total), and this will need to be addressed further down the line.

### 8.1.1 Data management roles

The project has appointed a dedicated Data Architect at Steno Diabetes Center Aarhus (SDCA) who holds primary responsibility for the overall data management strategy. This includes overseeing data capture, metadata production, data quality assurance, secure storage, backup and recovery procedures, long-term archiving, and data sharing. The data architect will make sure that data flows are secure and well-documented.

### 8.1.2 REDCap setup and study team contributions

While the data architect is ultimately responsible for the setup and configuration of REDCap, much of the practical implementation will be carried out by some members of the study team at Steno Diabetes Center. These team members will also contribute to identifying and describing specific data points registered in REDCap.

### 8.1.3 Metadata and Documentation

Metadata will be organised using Seedcase Sprout, based on the data recorded in REDCap (for data registered there) and a data dictionary which will be developed in collaboration with the LIVA team as the study progresses.

This will result in metadata that is consistent, comprehensive, and aligned with FAIR principles.

### 8.1.4 Long-term data handling and storage

The data architect is also responsible for overseeing the work carried out by the Seedcase team in converting raw data into Parquet files for long-term storage. It is essential that this process preserves the integrity of the original data, and the data architect will ensure that no information is lost or altered during the conversion.

### 8.1.5 Sub-study data integration

Any sub-study that generates data is expected to submit its dataset in full to the main project, accompanied by appropriate metadata. Although the SDCA team is not responsible for collecting this data, the data architect and the Publications Panel share responsibility for ensuring that sub-study teams understand their obligations. This includes verifying that all relevant data is submitted to the main study and properly integrated into the central repository on GenomeDK.

### 8.1.6 Compliance

The data architect and project manager will jointly ensure that the data management practices outlined in this plan are adhered to throughout the project lifecycle. Regular reviews will be conducted, and updates to the DMP will be made as needed.

### 8.1.7 Contingency planning

The study as a whole will run over 6 years, so it is unlikely that all original staff will still be present at the end of the study period. At present there is no formal guidance on how we expect to cover either long-term absence or people leaving the study, but it is expected that the core study team, data architect and project manager alongside the Project Secretary and the Chief Investigator will be able to cover most of the urgent tasks if needed.

A core principle of the project is documentation and this is ensured by using GitHub and sharepoint across study sites. A minimum of documentation from each team member, depending on role, is required in case of people leaving.

Funding has been obtained from the Novo Nordisk Foundation to conduct the study. Funding for training and upskilling any replacement staff is part of this funding. However, as the project is ongoing, further funding will also be sought to accompany future needs. This is the responsibility of the Project Management Panel to ensure.

# A Contributing

## A.1 Issues and bugs

The easiest way to contribute is to report issues or bugs that you might find while reading through the website. You can do this by creating a new issue on our GitHub repository.

## A.2 Adding or modifying content

To contribute to `dmp`, you first need to install uv and justfile. We use uv and justfile to manage our project, such as to run checks and test the template. Both the uv and justfile websites have a more detailed guide on using uv, but below are some simple instructions to get you started.

It's easiest to install uv and justfile using pipx, so install that first. Then, install uv and justfile by running:

```
pipx install uv rust-just
```

We keep all our development workflows in the `justfile`, so you can explore it to see what commands are available. To see a list of commands available, run:

```
just
```

As you contribute, make sure your changes will pass our checks by opening a terminal so that the working directory is the root of this project's repository and running:

```
just run-all
```

When committing changes, please follow the Conventional Commits specification in your commit messages. Using this convention allows us to be able to automatically create releases based on the commit message by using Commitizen. If you don't use Conventional Commits when making a commit, we will revise the pull request title to follow that format, as we use squash merges when merging pull requests, so all other commits in the pull request will be squashed into one commit.

# B Changelog

Since we follow Conventional Commits, we're able to automatically create formal "releases" of the website based on our commit messages. Releases in the context of websites are simply snapshots in time of the website content. We use Commitizen to automatically create these releases using SemVer as the version numbering scheme.

Because releases are created based on commit messages, a new release is created quite often—sometimes several times in a day. This also means that any individual release will not have many changes within it. Below is a list of the releases we've made so far, along with what was changed within each release.

## B.1 0.3.1 (2025-12-15)

### B.1.1 Refactor

- rearrange content on the landing page (#148)
- match the headers with the others to be consistent (#154)

## B.2 0.3.0 (2025-12-15)

### B.2.1 Feat

- naming convention for files and variables (#146)

### B.2.2 Fix

- correct list formatting (#160)

## B.3 0.2.1 (2025-12-12)

### B.3.1 Fix

- section should be hidden for now (#161)

## B.4 0.2.0 (2025-12-08)

### B.4.1 Feat

- data selection and preservation (#134)
- data sources section (#136)
- expand on landing page (#129)
- documentation and metadata section (#128)
- add data collection section (#122)
- section on data sharing (#120)
- responsibilities section (#91)
- ethics section (#90)
- storage and backup section (#86)
- add 404 page to website (#69)
- connect `onlimit-theme` (#45)
- add some basic text to landing page (#40)
- add some basic text to landing page

### B.4.2 Fix

- correct link to metadata section (#153)
- intro with official translation (#138)
- convert to Mermaid so it can be bigger (#89)
- link to image was incomplete (#61)
- typos found from `typos` check (#34)
- switch to `onlimit` (#35)
- add gitignore file

### B.4.3 Refactor

- split into sections on the sidebar (#151)
- use full description in title, not abbreviation (#152)
- switch to more appropriate CC-BY license (#24)
- switch to more appropriate CC-BY license