SLEEP REVOLUTION

Revolution of sleep diagnostics and personalized health care based on digital diagnostics and therapeutics with health data integration

D1.7:

Data Management Plan generated for “Pilot on Open Research Data”

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement no. 965417

SLEEP REVOLUTION

Deliverable D1.7: Data Management Plan generated for "Pilot on Open Research Data". Version 1

Grant Agreement nr.	965417
Start date	01.03.2021
Project title	Revolution of sleep diagnostics and personalized health care based on digital diagnostics and therapeutics with health data integration
Project acronym	SLEEP REVOLUTION
Type of Action	Horizon 2020, RIA – Research and Innovation Action
Duration	48 months
Prepared by	María Óskarsdóttir & Jacky Mallett
Date of delivery	25.11.2021
Nature	ORDP: Open Research Data Pilot
Work Package	WP 1 – Project Coordination
Lead Beneficiary	Reykjavik University
Reviewed/Editor	Erna Sif Arnardottir, Timo Leppänen & Lara Jonasdottir

Contents

Executive summary

This deliverable is the first version of the SLEEP REVOLUTION Data Management Plan and covers the way that research data is handled during and after the project, what data will be collected, processed or generated, the methodology & standards to be applied and the kind of data sharing. Moreover, the methodology of how data will be shared and/or how to make them available to users is also presented.

Introduction

SLEEP REVOLUTION is a Horizon 2020 funded project that aims at revolutionising current obstructive sleep apnea (OSA) diagnosis and treatment.

Data managed by the project is predominantly medical and research data (including sleep studies, medical diagnoses and treatment data, questionnaires, cognitive tests etc). This data is a mixture of datasets provided by consortium beneficiaries for a retrospective data analysis (medical data), and prospective studies with the informed consent of research participants performed as part of the project (research data). A key deliverable of the project is the creation of a secure repository, established at Reykjavik University, which will be used to store this data on behalf of the SLEEP REVOLUTION Consortium, and to provide computer resources for analysing and developing algorithms using this data. These deliverables are part of Work Package 2 of the SLEEP REVOLUTION.

Guiding principles

The guiding principle for data collection and analysis as part of the SLEEP REVOLUTION project is to provide a secure researcher access of relevant beneficiaries to the medical and research data, while preserving the data access, privacy, control, and other General Data Protection Regulation (GDPR) rights of all subjects with respect to their data[1]. To achieve this a dedicated High-Performance Computing (HPC) cluster (from here onwards *cluster*) has been established at Reykjavik University to provide a central secure data store, and a shared compute resource. This facility will provide both storage and computer resources which allow researchers to access and work with the data in an efficient and collaborative manner, whilst preserving the security of the data. Only authorised researchers will be able to access the data (an authorisation system has been constructed in Deliverable 2.1). Only data that that has been pseudonymised, and for which written informed consent for open data sharing has been received from the research participants will be shared publicly as a part of the European Open Science Cloud (EOSC) as lined out in Deliverable 2.2. This data sharing is in line with the FAIR Guiding Principles – Findable, Accessible, Interoperable and Reusable.

To protect the privacy of individual research participants, only pseudonymised data will be held on the cluster, without any personally identifiable information. The cluster will be held in a secure part of the Reykjavik University network and monitored continuously for abnormal behaviour such as large data transfers, inappropriate access, and usage.

Before any publications (e.g. scientific papers and public deliverables) with the potential of containing any personally identifiable data are released, it will be ensured that all data included

[1] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 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. https://eur-lex.europa.eu/eli/reg/2016/679

is anonymised, aggregated, and/or analysed in such a way that none of the content can be traced back to an individual research participant.

Permissions for collecting and handling personal data

All data collected from partners in the project will be in accordance with applicable ethical permissions, data protection authority permissions and medical board requirements in the respective countries of the data collection and in agreement with the Grant Agreement, Consortium Agreement and Joint Controller Agreement of the SLEEP REVOLUTION Consortium.

Data summary

The purpose of the data collection

The SLEEP REVOLUTION project will collect a large dataset of medical and research data based on multi-sensor sleep recordings and other clinically relevant information to perform detailed analysis and algorithm development.

The purpose of this data collection is to provide medical and research data to improve the diagnosis and treatment of OSA and related sleep disorders. This data will be used for research into OSA and the development of automatic algorithms to analyse the data to improve diagnosis and treatment options. The project will both generate prospective datasets from sleep studies and other data from research participants and provide a centralised repository for existing data sets from our collaborators for a retrospective data analysis.

The types and formats of data in the project

The project will collect different data related to sleep research. This consists of questionnaires, sleep diary data, cognitive tests, data from wearables (e.g., smartwatches) and diagnostic sleep monitoring equipment (both polysomnography and home sleep apnoea testing). Data will be received in several formats, including proprietary, which will then be converted to the appropriate open standard formats used for clinical data collection such as EDF, JSON, and CSV for long term storage.

Use of existing data

A number of existing sleep studies will be provided by different beneficiaries for analysis as part of the retrospective data analysis work package. This data will be screened, pseudonymised and securely transferred for storage on the secure cluster provided for this project. The precise contents of this data are being determined, and details will be provided in the next review of this document.

The origin of the data

The data will be collected from prospective sleep studies conducted by over 25 beneficiaries across Europe as well as from existing sleep studies from mostly the same beneficiaries in the project for retrospective analysis. The precise origin, type and format of each dataset will be documented here when this is available.

The expected size of the data

Approximately 30,000 retrospective sleep studies and related data will be collected onto the secure storage for this project, with about half of the sleep studies being 0.3 GB per study and the other half being 1GB per study, a total of 19.5 TB. For the prospective studies, approximately 1300 research participants will be included, each with a total of 6 nights of sleep studies for 1 GB each as well as additional data, approximately 1 GB per person, or a total of 9.1 TB. The total data size is therefore estimated at 28.6 TB or approximately 30 TB.

Data utility

The data will be useful during the project for the beneficiaries of the SLEEP REVOLUTION as per the SLEEP REVOLUTION Grant Agreement. The results from the research using the data are expected to contribute to the clinical knowledge, diagnosis, and treatment for OSA and other sleep disorders benefitting both researchers, medical professionals, hospitals, patients, society, and industry. Novel methods for early and improved diagnosis as well as targeted treatment of OSA contain a great potentiality for being greatly beneficial for patients and for the whole society. Patient level benefits of the new technology, diagnostic and treatment methods arising from this project will be assessed both in terms of improved outcomes and quality of life, as well as an assessment of costs, cost-benefits and cost-effectiveness at a societal level. The known direct beneficial mechanisms of new technology for the patients, include the increased treatment adherence and symptoms improvement. The aim is to maximize societal benefits while the correct patient group will the diagnosed early and the treatment will be administrated to those who are in need.

FAIR data

2. 1. Making data findable, including provisions for metadata

Data collected by the project can only be accessed by researchers who belong to the specific beneficiaries of SLEEP REVOLUTION and who have been appropriately authorised for access to specific datasets (as detailed in Deliverable 2.3). Within the project, appropriate descriptive disciplinary metadata will be used in a standardised way with appropriate digital objective identifier (DOI) and other descriptors being assigned to all data collected by the project.

The dataset and the repository will be listed as part of the European Open Science Cloud (EOSC), and relevant access will be provided to authorised researchers in line with Deliverable 2.2 where research participants have given explicit consent for the public reuse of their data. Currently, SLEEP REVOLUTION has a pending application status to be a provider and early adopter of the EOSC and we will strive to uphold the FAIR (Findability, Accessibility, Interoperability, and Reusability) data principles.

Naming conventions

Naming conventions for the data will identify each dataset, study and data with a unique label determined by the project. For the retrospective datasets provided by different beneficiaries, which are being reused, existing naming convention will be preserved and extended as appropriate. Software is currently being developed to perform this automatically.

Search keywords

Discipline-appropriate search keywords will be provided to optimize possibilities for re-use of data.

Version numbers

The project will use version numbers, change control, and associated tracking for all data and software within the project.

Metadata

Metadata will be created in the project. We are currently determining the most appropriate clinical metadata standard to be used, after analysis of the retrospective data, and this will be followed in all relevant work packages.

2.2. Making data openly accessible

Access to the data used by the project will be through a secure cluster at Reykjavik University. This cluster provides a shared data repository for all beneficiaries in the project in conjunction

with compute resources that provide support for the development of machine learning algorithms. Researchers will not be allowed to copy data from the cluster to their personal computers, algorithms will be developed and run against the data on the cluster, which will only be available for the use of this project. Data access is controlled by the SLEEP REVOLUTION Joint Controller Agreement which restricts access to the relevant work packages of different beneficiaries.

The retrospective data will not be made openly available for legal reasons, as this is medical data for which an informed consent of the research participant for open data sharing cannot be obtained. This is in line with GDPR subject privacy and other rights. In the prospective studies, we plan to have relevant research data with informed consent of the research participant openly available after the research questions of the SLEEP REVOLUTION project have been answered. The exact research data, process and level of this open data access is still under development by the SR beneficiaries. This information will be included in future versions of this document after a review of all the contractual obligations of the SR project as well as different national restrictions on research data access.

Secure access, monitoring, and physical facilities are provided by Reykjavik University. Arrangements are in place for secure controlled access, and inclusion of the cluster in the University's secure interior, with controlled virtual private network (VPN) access for authorised researchers. Data that is gathered by the project is typically provided in a manufacturer proprietary format or open-source formats. All proprietary formats will be converted to an open-source data format, suitable for long term storage. For example, the proprietary Noxturnal software (Nox Medical, Reykjavik, Iceland) will be used for access to data collected by their equipment, but this data will also be converted to EDF, a non-proprietary format that is accessible by open-source software.

Methods and software tools needed to access the data

To access the cluster and the data on it, the researchers must connect to the RU VPN, using e.g., Cisco AnyConnect, and then log on the cluster using Secure Shell (SSH) or PuTTy (free and open-source terminal emulator, serial console and network file transfer application). We have installed Python, as well as other open source programs, that can be used to view and process the data. For sleep technologists that use Noxturnal to score sleep studies, we plan to provide a remote desktop connection with Windows-based virtual machines.

The documentation about the software needed to access the data

We are creating documentation for the cluster and all relevant processes (e.g. requesting access, connecting to the cluster, basic operations, how to use the provided software to access the data) in the form of an online wiki. All researchers working with the data will be given access to the wiki when they get access to the cluster.

Inclusion of the relevant software (e.g. in open source code)?

Open-source software will be used where possible and be accessible within the cluster. All proprietary file formats will be converted to open source when possible. We plan to support access to third-party software provided by consortium members, such as Noxturnal, within the cluster, provided that all security and privacy guarantees are supported by this access.

Data and associated metadata, documentation, and code deposition?

Data, metadata and code will be stored in the cluster provided by the project at Reykjavik University. However, code developed in the project will be made publicly available i.e. through GitHub. Documentation will be in the wiki (wiki.sleep.ru.is) or in the project's Teams channel. Open access via the EOSC is provided as detailed above.

Ascertainment of the identity of persons accessing the data

All requests for access must come from the Principal Investigator of each beneficiary and will be processed in line with the Sleep Revolution Joint Controller Agreement for access rights. In exceptional cases, access requests will be discussed in the Executive Board.

We have constructed an authorisation system to give researchers access to the shared GPU cluster. For this purpose, we developed processes for granting access to the cluster as follows:

- The Principal Investigator of a beneficiary of the SLEEP REVOLUTION project fills in an Access Request form (see Appendix) and sends it to the cluster system administrator at Reykjavik University for researcher(s) at their organization who need access to the cluster.
- The researcher(s) contact information including the phone number, which datasets the specific researcher has permission to work on and for purposes of what work package in line with the Joint Controller Agreement of the project as well as any relevant ethical permits. The leader thereby validates their identity, authenticates the researcher and issues a work permission within the cluster.
- The cluster system administrator separately authenticates that the relevant beneficiary has allowed access to the relevant dataset in line with the Joint Controller Agreement and for which purposes. When needed, the Executive Board will be asked to authenticate a specific request by the administrator (when it is unclear if a specific beneficiary is allowed access to the relevant dataset and/or for the specific purpose stated).
- Next, the system administrator creates a user account for the researcher and sends them the login details, excluding the password.
- The password is then shared via a phone call between the system administrator and the researcher, who immediately must change the password.
- Once an individual has been granted access to the cluster, they will be assigned a trackable user ID with their authorised privileges.

Similarly, the Principal Investigator for a specific beneficiary can request access to additional datasets by sending an email with information to the system administrator for the relevant researcher(s). Such access requests will be handled in line with the Joint Controller Agreement procedures. We ensure that this structure is in compliance with the Data Management Plan and public data sharing as planned in WP1 (Deliverable 1.5 and 1.6).

2.3. Making data interoperable

Interoperable data strategy

Our goal is to have the data interoperable as much as possible. We are collecting medical and research data from over 25 beneficiaries across Europe into a common data store where the data will be organized and documented with the goal of allowing reuse of this data by other researchers and institutions within EOSC when appropriate. This reuse is subject to the appropriate ethical permissions and under the condition that the data will not leave the secure data store. We will adhere to suitable standards for formats, such as the European Data Format (EDF) that are compliant with open software applications. Metadata about the sleep studies, including age, gender, body mass index, and relevant sleep indexes, will be stored in a database that can queried to find suitable subjects for each study. This will allow for re-combination with different datasets.

Vocabularies, standards, and methodologies

We are currently determining which are the appropriate medical metadata standards and methodologies for the project and assessing the existing metadata standards being used in the medical data already collected that will be included in the SLEEP REVOLUTION data store. These will be documented in a future version of this document. Our goal is to allow for inter-disciplinary interoperability as much as possible. When this is not possible, we will provide mappings to more commonly used ontologies.

2.4. Increase data re-use (through clarifying licenses)

Licencing

Access to the data is made available to authorised researchers as detailed above. Where possible, data reuse will be supported for data created directly by the project, and appropriate permissions requested as detailed above with regards to the EOSC. The data will be made available for reuse at the end of the project or sooner if the intended reuse does not compromise the outcomes of the project, meaning that all the relevant research questions of the Grant Agreement have been answered.

Time span for re-use

As specified in the SLEEP REVOLUTION project description, we aim to preserve data for at least five years or more after the project whenever possible as per local beneficiary ethical and data protection rules for data sharing within SLEEP REVOLUTION. As the procedure and ethical permits of the project's partners are currently in progress the exact time span for re-use will be defined in future versions of this document, hand in hand with the framework set out by the ethical permissions from each country and partner. This will either be done by continued provision of research facilities at Reykjavik University, or by using the National Archives of Iceland (https://skjalasafn.is/national_archives_of_iceland).

Data quality assurance processes described

A procedure for automatic data quality assurance and integrity checking is being defined, and will be applied to all data collected by the project. Note that some elements of quality control with respect to collected OSA data are currently open research questions, in particular problems associated with sensors moving or being temporarily disrupted during sleep. Development of processes to detect this is an important part of the project, and our goal is to maintain updated information against collected datasets of issues being reproducibly reported by software and other sources.

Costs of FAIR data

An HPC cluster is being installed for the project at Reykjavik University, the capital cost of which has been provided by the Horizon 2020 grant for the project, along with accompanying costs for administration as well as to provide support and documentation for cluster users. This state-of-the-art cluster will provide data storage and the ability for researchers to access and run algorithms on the data. After the project period, Reykjavik University has committed to keeping the secure data store operational. In addition, we will apply for other funds to ensure that the data remains FAIR.

Responsibility of data management

The Executive Board members which are then individual work package leaders will be responsible for data management within their work packages, and WP1 will provide overall responsibility with the ultimate responsibility being that of the Principal Investigator of Reykjavik University, the coordinator of SLEEP REVOLUTION.

Long-term preservation

Once assembled, the entire data created by this project will be of significant value to researchers in OSA and other conditions. It will be one of the largest OSA data collections in the world, and will also provide a library of professionally scored data, which can be used for algorithmic development for machine learning. Reykjavik University has committed to long-term support

for the cluster (as permitted by different beneficiaries' ethical and data protection authorities), beyond the four-year grant period for SLEEP REVOLUTION.

In the absence of other funding, decisions over long term support for research is controlled by the Research Council, and the Dean of the School of Technology at Reykjavik University. If the data can no longer be hosted at Reykjavik University, a copy will be lodged with the National Archives of Iceland (https://skjalasafn.is/national_archives_of_iceland).

Data security

Provisions for data security

Data security is an integral part of the project. The equipment procured for the project provides advanced security monitoring, data duplication and backup within the cluster, and disk-level encryption. This equipment is located in a protected zone in the data centre of Reykjavik University. Data is stored in an encrypted format, and access to the data is controlled and monitored, with weekly reporting. Researchers will only be allowed access to the data they have received appropriate authorisation for. Additional security will be provided around the cluster. All transfers will be performed using encryption, it will not be possible to transfer data off the cluster, and monitoring will be performed on data uploads. Backups will be performed to tape and held in a secure offsite archive. Regular data integrity checks will be held, and the backup facility tested periodically with random recovery and checking of files. We will have an accredited certification body do an audit of the system to ensure it meets the requirements of security standards, such as ISO27101, as well as requirements for long-term preservation and curation.

One of the first stages of the project is to transfer retrospective data from our partners to the cluster. This includes over 30,000 sleep studies. The transfer is done either via a secure encrypted transfer across the internet, from the beneficiaries' computational resources into the secure data store or by copying the sleep studies to an encrypted Universal Serial Bus (USB) drive which is sent via courier to Reykjavik University.

Ethical aspects

Ethical or legal issues that can have an impact on data sharing

Ethical issues are an integral part of the SLEEP REVOLUTION, because of the medical and research data that will be collected and processed. Hence, subject data rights including informed consent for prospective studies, and the right to withdraw consent, request data removal, and other GDPR rights are supported throughout the whole project. Furthermore, the project involves many participants, including the 24 participant institutions of the European Sleep Apnea Database (ESADA) network. Therefore, an ethical guardian has been appointed by each partner to apply for ethical approval in each of the beneficiaries involved in the data collection. An open dialog will be ensured to discuss ethical issues as they come up and apply for additional

ethical approvals when needed. In addition, the goal of Work Package 2 is to build a dedicated storage infrastructure for secure and traceable storage for all research data gathered in the project. The data store itself complies with the GDPR requirements for "privacy and security by design" and allows data to be removed on request and permanently erased, both from the store and its backups. Additionally, a two person External Ethics Advisory Board (EEAB) has been created and approved via the Consortium Agreement; giving feedback at the beginning of the project, at the end of the development phase/beginning of the testing phase and a final report at the end of the project.

The project is divided into two main phases, a development phase, and a testing phase:

The development phase: Consists of the creation of an infrastructure to manage all data and analyses as well as ensure data security throughout the project. The development phase also includes retrospective data analysis and mining of around 30,000 existing sleep studies with processing of personal data (pseudonymized) with beneficiaries both inside and outside the EU.

The testing phase: Includes all prospective data collection in SLEEP REVOLUTION. This phase consists of pilot studies for validating software and newly developed methodology, validation of different devices, a generation of digital management platform as well as three prospective intervention studies. The outcomes of this project include a validated device for clinical sleep studies, wearables for research and consumer use, as well as smartphone applications, a cloud-based software and a digital management platform. Each of beneficiaries will comply with local legal, ethical and data protection requirements as well as European ones including the GDPR. In this phase, informed consent will be acquired for data sharing, data reuse, and long-term preservation for data collected in this phase of the project.

Each relevant beneficiary is in the process of applying for ethical approval according to local and national regulations (Deliverable 13.1). The documentation will be held at each participant site, as well as centrally by the project coordination at RU to ensure that such documentation is available for audit at any time.
All devices, treatments, methods, and algorithms developed in the project will, once operational, need to comply with GDPR and related national laws as far as it handles personal data, and with "The Directive on security of network and information systems (NIS Directive)" of the EU in terms of the wider security requirements (https://digital-strategy.ec.europa.eu/en/policies/nis-directive).

Privacy Policy

A Privacy Policy has been adopted which ensures that the six-core principle of processing of personal data as set out in Article 5 of GDPR are respected. In particular, data will be processed in accordance with appropriate security. The primary tool for such security will be that data are maintained in the location where they are collected and will not be transferred to other

locations. Others can access data based on their clearance levels, but not store it. As part of Work Package 2, compute resources are provided with the data storage, obviating the need to move copies around. The full detail on data security by design and default will be a core development issue of the tool development undertaken in the project.

Privacy Information Notices

Generic privacy information notices are prepared centrally in the project which will be adapted for each prospective study site. The notices will be drafted in accordance with Article 13 GDPR to ensure that consent provided is based on adequate information.

Data Pseudonymization

All data collected in the project will be pseudonymized for data processing. For the retrospective study in the development phase, the personal data that goes into the secure cloud-store data base will possess only a study ID and no identifiable information. However, as the analysis of the sleep recordings may result in new diagnosis for the participants involved, a connection between the research ID and the identity of the person whom the data belongs to will be kept offline in a secure location in each participant institute, and not shared with the SLEEP REVOLUTION explicitly. This is done so that the subjects can be informed of important findings regarding their health. In the prospective studies in the testing phase, we will collect research data. In this case, the data will not be completely pseudonymized, because the study co-ordinators will have to know the identity of the subjects in order to assist and guide them through the study. Furthermore, a couple of principal investigators at RU will have access to the key, so they can intervene if necessary. This key will be kept in a secure location in the cluster, with controlled access of only authorized users. In some cases, such as EOSC, the data may need further anonymisation, e.g. to scramble the sound of sleep recordings.

Data Privacy Impact Assessments

We will conduct a Data Protection Impact Assessment (DIPA) as is required according to Article 35 of the GD

Appendix

Access Request to the SLEEP REVOLUTION Cluster

- Sleeping Beauty-

This form must be completed and signed by the researcher who seeks access to the SLEEP REVOLUTION cluster **and** a Principal Investigator (PI) of a beneficiary of the SLEEP REVOLUTION. The PI thereby authenticates the researcher and validates their identity and work permission within the cluster. The form should be sent by the PI to the system administrator (SA) of the cluster (FILL EMAIL). The SA will create a user account for the researcher and send them the login details, excluding the password. The password is then shared via a phone call between the SA and the researcher, who immediately must change the password.

Researcher Information

Name:

Email:

Position:

Organization:

Phone number:

WP:

Duration of access:

To which datasets do you have access:

Will you apply for access to other datasets? If yes, which ones (Please provide the necessary documentation):

Short description of the project you will work on the cluster:

Principal Investigator Information

Name:

Email:

Position:

Organization:

Relation to the researcher:

Relation to SLEEP REVOLUTION:

Date and location_____

Researcher's signature_____

Principal Investigator's signature_____