INITIAL DATA MANAGEMENT PLAN

GALAXY

GUT-AND-LIVER AXIS IN ALCOHOLIC LIVER FIBROSIS

GRANT NUMBER 668031

DELIVERABLE NUMBER: D8.3
DELIVERABLE DUE DATE: JUNE 30TH 2016
COMPLETION DATE OF DELIVERABLE: JUNE 29TH 2016
DISSEMINATION LEVEL: PUBLIC

DOCUMENT HISTORY

Issue date	Version	Reasons for this issue
12 June 2016	v0.1	First draft
	v1.0	Final version

DOCUMENT MAIN AUTHOR: Dr Maja Thiele (OUH), PhD
DOCUMENT SIGNET OFF BY: Project Manager Linda Sevelsted Møller, MD, PhD

TABLE OF CONTENTS

1. INTRODUCTION

The GALAXY project (Gut-and-liver axis in alcoholic fibrosis) is a part of the EU Framework Programme for Research and Innovation Horizon 2020.
The purpose of this data management plan (DMP) is to outline how the GALAXY consortium will manage data generated by the project.
The document reports the initial DMP for GALAXY, which can also be assessed online at the Danish Digital Curation Centre.
The DMP is not a fixed document; it evolves during the lifespan of the project to fine-tune it to the data generated and the uses identified by the consortium. The initial DMP will therefore be supplemented with a Mid-term Review DMP and a Final Review DMP.

A DMP is a key aspect of any research project where large scale, high throughput data are generated. GALAXY is such a project.
The GALAXY project will generate research data from work packages 1-7. In total, we estimate that the GALAXY project will generate around 1 terabyte of data.
The Initial DMP for GALAXY has five sections: (1) Data Set Reference and Name, (2) Data Set Description, (3) Standards and Metadata, (4) Data Sharing, (5) Archiving and Preservation.
The sections are subdivided in general principles regarding the overall GALAXY project that apply to data management across the consortium partners and to the specific principles for clinical study data from Work Package 1. Additionally, we describe data management for raw data, aggregated data and publications.
The DMP may be updated outside the mid-term and final review to fine-tune it to the data generated and the uses identified by the consortium since not all data or potential uses are clear from the start. Therefore, the latest version of the DMP are available online from https://dmponline.dtu.dk.

2. DATA SET REFERENCE AND NAME

GALAXY-IDMP.

3. DATA SET DESCRIPTION

In this DMP, the term ”data” cover raw data generated from investigations in humans and animals, human and animal tissue, aggregated data, result data, data documentation and publications related to the GALAXY project.
The GALAXY project consists of 10 partners (University of Southern Denmark, European Molecular Biology Laboratory, University of Copenhagen, Biomedical Research Foundation Academy of Athens, University of Bonn, Steno Diabetes Center, Nordic Bioscience, Nordic Rebalance, University of Oslo and Odense University Hospital). All partners will generate data to the project.

3.1 DATA SET DESCRIPTION – GENERAL PRINCIPLES

3.1.1 ORIGIN

The data origins from four prospectively conducted human studies at Odense University Hospital (OUH) and animal studies at University of Bonn (UKB).
The data underpins planned publications within the GALAXY consortium.

3.1.2 NATURE

The data consists of the following quantitative, raw data: (1) clinical study logs, (2) patient history, (3) clinical investigations, (4) questionnaires, (5) outcome event data, (6) existing data from the Danish unique personal id registries, (7) biobank logs, (8) data generated from quantitative analyses of human liver tissue, blood, feces, urine, saliva, hair and sigmoid tissue, (9) rodent logs, (10) data generated from quantitative analyses of rodent liver and colonic tissue, blood and feces.
Generated data - except (6) - will be original.
Quantitative data formats are continuous, ordinal, nominal and binary.
Few qualitative and quantitative data, such as signed patient consent forms and questionnaires, exists in paper format. Some result data may also exist in sound, image and video format.

3.1.3 SCALE

The four human studies consists of: (A) 400 patients in the GALA-ALD study, a prospective cohort of alcohol overusing patients; (B) 100 participants in the GALA-HP study, a cross-sectional study of healthy controls, matched for age and gender with GALA-ALD; (C) 136 patients in the GAB-ALD study, recruited from GALA-ALD. GAB-ALD is a randomised controlled trial of the gut-selective antibiotic Rifaximin vs. placebo for alcoholic liver fibrosis; (D) 80 patients in the SYN-ALD study, a randomised controlled trial of the synbiotic Profermin vs. placebo for patients with a dysbiotic microbiome and alcoholic liver fibrosis.
The rodent studies explore two strategies: prevention of fibrosis by intervening concomitantly with gut-selective antibiotic Rifaximine with ethanol diet, and treatment of fibrosis with Rifaximine by shorter use under ethanol diet. Different groups of mice (15 animals per group), animal models and genotypes are used: (A) Lieber-DeCarli diet (wild type germ-free, specific-pathogen-free and Ago2-Flag-HA-k.i. mice); (B) High percentage ethanol diet (cannabinoid receptor-1 (CB1) and receptor-2 (CB2) knockout and wild type mice). (C) Controls (wild type, Ago2-Flag-HA-k.i. and germ-free mice, bile duct ligation, carbon tetrachloride and thioacetamide models as controls for cirrhosis.
We estimate the full dataset to be 1 TB large for which participating centres have adequate storage capacity.

3.1.4 USEFULLNESS

The dataset may be useful for comparable projects in other liver disease aetiologies (e.g. non-alcoholic fatty liver disease), other fields of medicine (e.g. endocrinology) or for other microbiome and multi-omics studies.

4. STANDARDS AND METADATA

4.1 STANDARDS AND METADATA – GENERAL PRINCIPLES

4.1.1 DATA CAPTURE

Participating partners will capture quantitative data electronically. Files and folders will be named according to local custom in clearly marked GALAXY folders. Datasets will be named with date and version.

4.1.2 DATA QUALITY

Quality and consistency will be controlled through standardised data capture, data entry validation, representation with controlled vocabularies and peer review of data.

4.1.3 DATA FORMAT

Data will be stored according to local preference in accordance with the UK Data Service guidelines https://www.ukdataservice.ac.uk/manage-data/format/recommended-formats. Tabular data are stored in .csv, .tab, .sav, .dta, .mdb, .accdb, .por, command ('setup') files (SPSS, Stata, SAS, etc.), or structured text or mark-up file of metadata information (DDI XML file). Textual data are stored as .rtf, .txt or .doc/.docx files. Image, audio and video data are stored as .tif, .flac and .mp4. Documentation and scripts are stored as .rtf, .pdf, .xhtml, .htm or .odt files.

4.1.4 METADATA

We will follow principles described by CERIF (Common European Research Information Format; developed and maintained by EuroCRIS, http://www.eurocris.org/) for recording metadata. See more here: http://www.dcc.ac.uk/resources/metadata-standards/cerif-common-european-research-information-format.
Converis, Pure or Symplectic Elements may be used as research information system tools for implementing the CERIF metadata standards.

4.2 STANDARDS AND METADATA – SPECIFIC PRINCIPLES FOR CLINICAL STUDY DATA

4.2.1 DATA CAPTURE AND STORAGE OF RAW, ORIGINAL DATA

Four systems will be used for data capture of raw, original data: (1) The REDCap database system hosted by Odense Patient data Explorative Network (OPEN), (2) the

SurveyXact Database hosted by Rambøll and licenced to Region of Southern Denmark, (3) Secure SharePoint drives hosted by the Region of Southern Denmark and (4) the COSMIC electronic patient file system hosted by the Region of Southern Denmark. One database will be generated per clinical study.
Clinical study logs, patient history data, clinical investigations and electronic questionnaires are entered directly into the databases by study investigators, project nurses and study participants. Except two questionnaires (48 hour recall and Food Frequency Questionare) that are stored in paper form to be machine-read at the end of the study. Biochemical data, histological data and outcome event data are stored in the electronic patient files, to be electronically copied to the study database after study completion.
Data files will be named according to GALAXY study ID, type of data and date of creation/revision; e.g. "galaald_complete_yymmdd.dta" or "gabald_biochemistry_yymmdd.dta". Data variables in individual files will be given an abbreviated variable name, together with a longer, explanatory variable label; e.g. "S-bilirubin (umol/L)"
All project data will be stored in accordance with the approval given to the individual projects by the Danish Data Protection Agency under the collective permission given to the Region of Southern Denmark.
Odense Patient data Explorative Network (OPEN) is responsible for secure storage and back up on the databases in REDCap and OPEN Projects (registry of biological material). The Region of Southern Denmark is responsible for the storage and back up on secure SharePoint drives. Rambøll is responsible for the storage and back up of data in SurveyXact.

4.2.2 Data Quality
Database audits by study investigators and OPEN data managers prior to data collection will ensure that the database is complete and consistent.

4.2.3 Meta Data
Metadata are stored in the REDCap system as the "Data Dictionary" and in SurveyXact under analyses/dataset. Metadata can be exported from both programmes as a .csv file.

5. Data sharing

5.1 Data sharing – General principles
Data will be shared through the GALAXY consortium, with participating partners accept and in compliance with applicable legislation.

5.1.1 Data with unique patient identifiers
Patient identifiable data connected to the participants in the four human studies will be stored at Odense University Hospital in accordance with the approval given to the

individual project by the Danish Data Protection Agency under the collective permission given to the Region of Southern Denmark.
We will not share unique identifiers with project partners outside the Odense University Hospital, unless specific approval are granted from Danish Data Protection Agency under the collective permission given to the Region of Southern Denmark.

5.1.2 Pseudo-anonymised data

Pseudo-anonymised data can be shared between project partners if a Data Processing Agreement is signed by both partners and if sharing is in accordance with the approval given to the individual project by the Danish Data Protection Agency under the collective permission given to the Region of Southern Denmark.
Transfer of pseudo-anonymised data between GALAXY partner institutions will follow local protocols in accordance with the Data Processing Agreement. The giving and receiving partners guarantee that adequate security, hardware, software and bandwidth are available for any large volume data movement.

5.1.3 Anonymised data and animal data

We consider patient data fully anonymised when – by the data controller or by any other person – it is not possible to single out an individual from the data, it is not possible to link records relating to an individual, and information concerning an individual cannot be inferred from the data.
In accordance with the Opinion 05/2014 on "Anonymisation Techniques" (European Union article 29 on the General Data Protection Regulation), we will use a combination of anonymisation techniques (e.g. noise addition, permutation, differential privacy, aggregation, k-anonymity, l-diversity and t-closeness) to ensure that data are anonymised.
Human anonymised data and data from the animal studies are shared within the GALAXY consortium in agreement with the Consortium Agreement.
Raw anonymised data and related metadata, provided they are not subject to terms of confidentiality and/or IP protection, will be made available to the public by three methods:

- Registration of study protocols with OPEN. By request, OPEN can make study data available to outside researchers, if agreed upon by the GALAXY Steering Committee.
- Sharing of datasets and metadata to the Zenodo website (http://zenodo.org/) or similar data repositories.
- Long-term storage of datasets, related metadata and study protocols by the Danish National Archived after completion of the project period (see section 6)

Anonymised raw datasets will only be shared with the public when GALAXY has published results from the data (or pending an accepted publication, in accordance with regulations by some scientific journals).

5.1.4 Study protocols, result data and publications

Data documentation, data descriptions and protocols for clinical studies, which are not subject to terms of confidentiality and/or IP protection, will be published in three places to enhance further research collaboration:

- The OPEN website will publish project summaries for each individual clinical study, with the possibility of sharing detailed protocols, data documentation and data sets by request, after acceptance from the GALAXY Steering Committee.
- The GALAXY website will publish a general project description of the GALAXY project and published results with the possibility of sharing detailed protocols, data documentation and data sets by request, after acceptance from the GALAXY Steering Committee.
- Institutional repositories such as the University of Southern Denmark's Publication & REsearch platform (PURE) will be used to publish project descriptions, links to shared datasets and published results.

Study protocols for the randomised studies (GAB-ALD and SYN-ALD) will be registered with EUDRA-CT, clinicaltrials.gov or a similar internationally recognised study registration site.

Publications (peer reviewed journal articles, non-peer reviewed articles, books and book chapters, conference papers etc.) will be shared with the public in accordance with copyright agreements with the publisher.
We will make published results available for download from institutional repositories (e.g. PURE) and the GALAXY website using Green Open Access.

6. Archiving and Preservation

6.1 Archiving and Preservation – General Principles

Data from the GALAXY project will be stored for long-term preservation at the earliest five year after project completion, in accordance with the Grant Agreement. Data will only be entered into long-term storage if it is in agreement with the GALAXY Steering Committee, or when the relevant Data Protection Agency approval expires.

6.1.1 Data volumnes, preservation and discovery

The Danish Data Repository hosted by The Danish National Archives (Rigsarkivet) will be used for long-term archiving of human study data. Archiving is without an end-date and is free of charge to the GALAXY consortium. The National Archives performs data curation, does regular backup of data and guarantees adequate storage facilities.
Data can be searched and discovered by other researchers using metadata http://dda.dk/simple-search. Any publication using data from the Danish Data Repository require permission from the donor.

6.1.2 Data Security

Data security will conform to national Data Protection guidelines and the signed Data Processing Agreements.

The security of electronic data described is guaranteed by the host institution and will follow national and European Data Protection Acts. All data in paper format will be kept in accordance with Data Protection guidelines (e.g. https://www.ukdataservice.ac.uk/manage-data/store/security).

6.1.3 Documentation and Preservation Plan

Data and metadata will be prepared for preservation in accordance with standards described in 4.1 and 4.2. Data and metadata files will be stored together with study protocols, publications relating to the project and standard operating procedures.