This project has received funding from
The European Union's Horizon 2020 research and innovation programme under
Grant agreement No 101017562

Finding Endometriosis using Machine Learning
FEMaLe

Call/Topic: Digital transformation in Health and Care

Type of action: RIA

Date: 23.06.2023

DELIVERABLE NUMBER	D10.31
DELIVERABLE TITLE	Data Management Plan: Updated
RESPONSIBLE AUTHOR	AU

GRANT AGREEMENT No.	101017562
DOCUMENT TYPE	ORDP
WORKPACKAGE N. | TITLE	10 | GOVERNANCE: Management, Financial, Risk
LEAD CONTRACTOR	AU
AUTHOR(S)	All FEMaLe Beneficiaries
PLANNED DELIVERY DATE	June 30th, 2023
ACTUAL DELIVERY DATE	June 23rd, 2023
DISSEMINATION LEVEL	Public
STATUS	Reviewed and quality checked
VERSION	Final version (1.5)
REVIEWED BY	FEMaLE PMO

Document history

Version	**Date** [1]	**Comment**	**Author**	**Status** [2]
1.1	30-01-2023	First draft created.	AU	Drafted
1.2	22-03-2023	Second draft created, including recommendations from WP participants.	AU	Drafted
1.3	26-04-2023	Third draft prepared for FEMaLe Review Panel.	AU	Drafted
1.4	01-06-2023	Final draft created, based on FEMaLe Review Panel feedback.	AU	Completed
1.5	23-06-2023	Final version ready for submission, quality checked by FEMaLe PMO.	AU	Validated

[1] As per the project's cloud storage or per email date if applicable.
[2] Drafted, completed, or validated as per the project's cloud storage or per email date if applicable.

TABLE OF CONTENTS

Disclaimer

The content of this deliverable does not reflect the official opinion of the European Union. Responsibility for the information and views expressed herein lies entirely with the author(s). All FEMaLe Consortium members are also committed to publish accurate and up to date information and take the greatest care to do so. However, the FEMaLe Consortium members cannot accept liability for any inaccuracies or omissions, nor do they accept liability for any direct, indirect, special, consequential, or other losses or damages of any kind arising out of the use of this information.

Copyright notice

Acknowledgement

The project 'Finding Endometriosis using Machine Learning' (FEMaLe) has received funding from the European Union Horizon 2020 programme under grant number 101000640.

Citation

Be so kind as to cite this work as: Finding Endometriosis using Machine Learning, 2023: Data Management Plan: Updated (Version 2) under the supervision of the Project's Coordinator.

Legislation

Legislation H2020 Framework Programme – Regulation (EU) No 1291/2013 of the European Parliament and of the Council of 11 December 2013 establishing Horizon 2020 - The Framework Programme for Research and Innovation (2014-2020) (OJ 347, 20.12.2013, p. 104).

Euratom Research and Training Programme (2014-2018) – Council Regulation (Euratom) No 1314/2013 of 16 December 2013 on the Research and Training Programme of the European Atomic Energy Community (2014-2018) complementing the Horizon 2020 – The Framework Programme for Research and Innovation (OJ L 347, 20.12.2013, p. 948).

H2020 Specific Programme – Council Decision 2013/743/EU of 3 December 2013 establishing the Specific Programme Implementing Horizon 2020 - The Framework Programme for Research and Innovation (2014-2020) (OJ L 347, 20.12.2013, p. 965).

Rules for Participation (RfP) – Regulation (EU) No 1290/2013 of the European Parliament and of the Council of 11 of December 2013 laying down the rules for the participation and dissemination in Horizon 2020 – the Framework Programme for Research and Innovation (2014-2020) (OJ L 347, 20.12.2013, p.81).

Financial Regulation (FR) – Regulation (EC, Euratom) No 966/2012 of the European Parliament and of the Council of 25 October 2012 on the financial rules applicable to the general budget of the European Union (OJ L 298, 26.10.2012, p.1).

Rules of Application (RAP) – Commission Regulation (EC, Euratom) No 1268/2012 of 29 October 2012 on the rules of application of l Regulation (EC, Euratom) No 966/2012 of the European Parliament and of the Council on the financial rules applicable to the general budget of the Union (OJ L 298, 26.10.2012, p.1).

Finding Endometriosis using Machine Learning: FEMaLe

1. ARONYMS AND ABBREVIATIONS

Abbreviation	Description
DOI	Digital Object Identifier
DMP	Data Management Plan
FAIR	Findable, Accessible, Interoperable and Re-usable
GDPR	General Data Protection Regulation
WP	Work Packages
Abbreviation	**Participant organisation name**
AU	Aarhus University
AAU	Aalborg University
EQUIP	European Society for Quality and Safety in General Practice
SWU	Semmelweis University, Dept. of Gynaecology and Obstetrics
UOXF	University of Oxford
SURGAR	SurgAR
RTU	Riga Technical University
KTH	Kungliga Tekniska Högskolan
IAAD	Istanbul Avrupa Arastimalari Dernegi
PREL	Precision Life ltd.
YCL	Yourcode Lab Kft
UNAB	The University of Aberdeen
CORR	Correlate AS
WBS	Web Bay Solution
UNED	The University of Edinburgh

2. INTRODUCTION

This document provides an updated version of the deliverable D10.30 Data Management Plan: Initial. The purpose of this DMP is to support the data management life cycle of all data that are collected, processed, or generated in the project. The DMP is intended to be a living document where information can be made available at a more detailed level through updates as the implementation of the project progresses or when significant changes occur.

FEMaLe project complies with the FAIR data management concept to develop this DMP. FAIR data management requires the project data to be: Findable, Accessible, Interoperable and Re-usable[3]. These principles precede implementation choices and do not necessarily suggest any specific technology, standard, or implementation-solution.

FEMaLe has made use of the online tool: https://dmponline.deic.dk/ to generate the initial DMP and the updated DMP. A data management responsible from each work package (refer to Table 1) has been assigned to create an overview of the collected data in the work package, by completing the online template. The DMPonline questionnaire consisted of 40 in-depth questions related to the data collection using the FAIR principles. Initially each data management responsible were asked to address issues related to the purpose, objective, types and formats, size, and utility of the collected data. Then, they were asked to what extent the research data aligned with the four principles of FAIR data.

Table 1: Data management responsible

	Name	Organisation	E-mail
WP1	Liv Juul Nielsen	AU	ljni@ph.au.dk
WP2	Bruno Silva	IAAD	brunoiaad2015@gmail.com
WP3	Dorte Rytter	AU	dr@ph.au.dk
WP4	Nilufer Rahmioglu	UOXF	nilufer@well.ox.ac.uk
WP5	Dóra Balogh	SEMM	dorabiankabalogh@gmail.com
WP6	Dmitrijs Bļizņuks	RTU	dmitrijs.bliznuks@rtu.lv
WP7	Saman Noorzade	SURGAR	saman.noorzadeh@surgar-surgery.com
WP8	Karina Ejgaard Hansen	AU	keh@ph.au.dk
WP9	Nemanja Todic	WBS	nemanja.todic@webbaysulotions.com
WP10	Ulrik Bak Kirk	AU	ubk@ph.au.dk

The primary objective of the FEMaLe management plan is to ensure that the project's generated and gathered data can be preserved, exploited, and shared for verification or re-use in a consistent manner. In line with the goal of making research data findable, accessible, interoperable, and reusable (FAIR), this DMP should provide details on the handling of research data during and after the project. Table 2 illustrates an overview of the dataset collected in the FEMaLe project, categorised by the ten work packages in the project.

[3] European Commission. (2016) *H2020 Programme Guidelines on FAIR Data Management in Horizon 2020.* https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

Table 2: Overview of dataset

Work package	Name of dataset	Responsible partners	Format
WP1	• Impact cases • Pulse checks • Half Double evaluations	EQUIP, AU	Text/numerical in .txt format
WP2	• No data collected	AU, IAAD	-
WP3	• Consensus study • Statistic Denmark • CPRD data • CYKLUS study	AU, UNAB	Text/numerical in .txt format
WP4	• UK Biobank phenotype data • UK Biobank genome-wide array-based genotype data • Danish blood donor study • DBDS genome-wide array-based genotype data • CHB phenotype data • CHB genome-wide array-based genotype data • DBDS Meso-scale human biomarker 54-plex • miENDO/FENOX: deep phenotype data • miENDO/FENOX: geome-wide array-based genotype data • UKBB: OLINK panel with 3000 proteins	AAU, UOXF, PREL	Text/numerical in .txt format Genotype calls in plink format (bed/bim/fam)
WP5	• Lucy App	YCL	JSON
WP6	• Surgical phenotyping using machine learning	RTU	MP4, MKV, MPEG
WP7	• Computer vision tool to detect endometriosis	SURGAR	MP4, MKV, MPEG, JSON
WP8	• Interview data • Questionnaire data • Translation data	YCL, AU	M4A, CSV EXCEL, SPSS text/numerical in .doc format
WP9	• Website, SoMe	EQUIP, WBS	PDF, ODT, TXT, XLS, SLSX, CVS, PDF, PPT, PPTX, JPEG, TIFF, PNG, SVG
WP10	• Project management	AU, CORR	MP4, TXT, PDF, TXT, PPT, PPTX

Publicly available information about the FEMaLe Project is made possible via a dedicated website: https://findingendometriosis.eu. All documents will be archived and stored in FEMaLe's SharePoint site at Aarhus University, which serves as the data repository.

3. DATA COLLECTION

FEMaLe will generate a broad spectrum of data from each of the projects ten work packages. The overall purpose of the data collection is to answer the objective of the FEMaLe project, as described in the proposal. In relation to this each work package has their own purpose of data collection and data generation. Table 3 provides a more detailed description of the data collection from each work package, by answering the following questions:

1. What is the purpose of the data collection/generation and its relation to the objectives of the project?
2. What types and formats of data will the project generate/collect?
3. Will you re-use any existing data and how?
4. What is the origin of the data?
5. What is the expected size of the data?
6. To whom might it be useful ('data utility')?

Table 3: Summary of data utilised in the FEMaLe project.

	Name of data	Purpose of data	Relation to objectives of the project	Data type and format collected	Existing data: re-used	Origin of data	Size of data	Data Utility
WP1	Impact cases	The purpose for all data collected is to provide a guiding star for each work package in relation to creating early impact and to establish a reflective practice about improving impact monitoring and assessment.	To set up indicators to guide the activities in the work packages	Text/numerical in .txt format	No	Workshop	27 impact cases made	The data is useful for the FEMaLe work packages members, to provide a guiding star for each work package in relation to creating early impact and to establish a reflective practice about improving impact monitoring and assessment.
	Pulse checks		To assess stakeholder satisfaction through pulse checks	Text/numerical in .txt format	No	Survey	?	
	Half Double evaluations		To evaluate the use of the half double practices in an innovation and research project	Text/numerical in .txt format	No	Survey and interview	20 evaluations made	
WP3 *Task 3.1*	Consensus study	To determine relevant symptoms in relation to endometrioses	The overall purpose of the data collection is to reach consensus on relevant symptoms related to endometriosis.	Text/numerical in .txt format	No	N/A	-	This data can be used to evaluate how patients and doctors from different countries consider the relevance of different potential symptoms and consequences related to endometriosis.
Task 3.2 3.3	Statistics Denmark	To define disease groups (cases) and disease-free groups (controls) using ICD codes and to identify health seeking behaviour and comorbidity	We use data from statistics Denmark to investigate the health-related consequences of diagnostic delay	Text/numerical in .txt format	Yes	Statistics Denmark	Not relevant as we are not the ones who control/store these data.	This is not our data; it is available via application to Statistics Denmark.

	Name of data	Purpose of data	Relation to objectives of the project	Data type and format collected	Existing data: re-used	Origin of data	Size of data	Data Utility
Task 3.2 3.3	CPRD data (University of Aberdeen)	To determine prevalence and geographical distribution of endometriosis. To develop a prediction model for the diagnosis of surgically confirmed endometriosis. To assess the effect of lifestyle factors on the natural course of endometriosis. To compare resource utilization between women with and without endometriosis.	We will use CPRD data linked with hospital episode data to address the objectives.	Text format	No	CPRD & NHS digital	Data on approximately 2 million women	Data is not owned by us. It is available on application to CPRD who control the data.
Task 3.2	CYKLUS study	To describe health related conditions among women between 15 and 50 years of age and to investigate both causes for and consequences of a number of unspecific and underdiagnosed conditions including endometriosis.	With the CYKLUS study, we collect data to estimate the prevalence and geographical distribution of symptoms related to endometriosis. Also, the data will be used to make a phenotype description of women reporting endometriosis like symptoms.	Text/numerical in .txt format	No	N/A	Includes data on symptoms of 60.000 Danish woman	The collected data will be useful for researcher working with women's health.

	Name of data	Purpose of data	Relation to objectives of the project	Data type and format collected	Existing data: re-used	Origin of data	Size of data	Data Utility
WP4 *Task 4.1*	UK Biobank phenotype data	Definition of disease groups cases and disease-free groups (controls) using ICD codes and self-reported diagnoses	Utilise ICD codes and self-reported diagnoses to group endometriosis case and endometriosis free controls to calculate the high-risk genotype combinations associated with endometriosis subtypes.	Text/numerical in .txt format	Yes	UK Biobank	Not relevant as we are not the ones who control/store these data?	This is not our data; it is available via application to UK Biobank. It is a powerful biomedical database that can be accessed globally to enable biomedical discoveries to improve public health.
Task 4.1	UK Biobank genome-wide array-based genotype data (SNPs)	Define the genome-wide common genetic landscape of study participants	Utilise the genotype data to calculate the high-risk genotype combinations associated with endometriosis subtypes.	Genotype calls in plink format (bed/bim/fam)	Yes	UK Biobank	Not relevant as we are not the ones who control/store these data?	This is not our data; it is available via application to UK Biobank, a biomedical database that can be accessed globally to enable biomedical discoveries to improve public health.
Task 4.1 *4.2*	Danish Blood Donor Study (DBDS) Phenotype data	Definition of disease groups cases and disease-free groups (controls) using ICD codes. Identification of disease sub-types and co-morbidities as recorded by ICD codes.	(1) Utilise ICD codes to define endometriosis group and replicate the genetically determined endometriosis subtypes. (2) Characterise the phenotype (subtypes and co-morbidities) associated with endometriosis subtypes.	Text/numerical in .txt format	Yes	DBDS	Not relevant as we are not the ones who control/store these data?	This is not our data; it is available via application to Danish Blood Donor Study. This is a nation-wide research resource enabling answer health-related research questions.

	Name of data	Purpose of data	Relation to objectives of the project	Data type and format collected	Existing data: re-used	Origin of data	Size of data	Data Utility
Task 4.1 4.2	DBDS genome-wide array-based genotype data (SNPs)	Define the genome-wide common genetic landscape of study participants	Replicate the high-risk genotype combinatory scores in an independent dataset.	Genotype calls in plink format (bed/bim/ fam)	Yes	DBDS	Not relevant as we are not the ones who control/store these data?	This is not our data; it is available via application to Danish Blood Donor Study; a nation-wide research resource enabling answer health-related research questions.
Task 4.1 4.2	Copenhagen Hospital Biobank (CHB) Phenotype data	Definition of disease groups cases and disease-free groups (controls) using ICD codes. Identification of disease sub-types and co-morbidities as recorded by ICD codes.	(1) Utilise the ICD codes to define endometriosis group and replicate the genetically determined endometriosis subtypes. (2) Characterise the phenotype (subtypes and co-morbidities) associated with the endometriosis subtypes.	Text/numerical in .txt format	Yes	CHB	Not relevant as we are not the ones who control/store these data?	This is not our data; it is available via application to Copenhagen Hospital Biobank. This is a national research resource enabling answer health-related research questions.
Task 4.1 4.2	CHB genome-wide array-based genotype data (SNPs)	Define the genome-wide common genetic landscape of study participants	Replicate the high-risk genotype combinatory scores in an independent dataset.	Genotype calls in plink format (bed/bim/fam)	Yes	CHB	Not relevant as we are not the ones who control/store these data?	Same as above.
Task 4.2	DBDS Meso-scale Human Biomarker 54-Plex	Define the plasma levels of 54 proteins.	Association of the 54 proteins with genetically determined endometriosis subtypes.	Text/numerical in .txt format	Yes	DBDS	Not relevant as we are not the ones who control/store these data?	This is not our data; it is available via application to Danish Blood Donor Study. This is a nation-wide research resource enabling answer health-related research questions.

	Name of data	Purpose of data	Relation to objectives of the project	Data type and format collected	Existing data: re-used	Origin of data	Size of data	Data Utility
Task 4.3	miENDO/FENOX: OLINK panel with 3000 proteins	Define plasma levels of 3000 proteins	Association of 3K proteins with genetically determined endometriosis subtypes	Text/numerical in .txt format	No	miENDO and FENOX	~50GB	miENDO is an endometriosis patient cohort that enabling investigation of underlying causal factors for endometriosis. FENOX is an endometriosis and uterine fibroids study enabling investigation of factors underlying these two conditions.
Task 4.3	miENDO/FENOX: Deep phenotype data	Define the endometriosis cases with deep phenotypic	To further characterize the genetically determined endometriosis subtypes	Text/numerical in .txt format	No	miENDO and FENOX	~1GB	Same as above.
Task 4.3	miENDO/FENOX: genome-wide array-based genotype data (SNPs)	Define the genome-wide common genetic landscape of women.	Replicate the genetically determined endometriosis subtypes in these more deeply phenotype datasets.	Genotype calls in plink format (bed/bim/fam)	Yes	miENDO and FENOX	~30GB	Same as above.
Task 4.3	UKBB: OLINK panel with 3000 proteins	Define plasma levels of 3000 proteins	Replicate the association of proteins with endometriosis cases and subtypes.	Text/numerical in .txt format	Yes	UKBB	Not relevant as we are not the ones who control/store these data?	This is not our data; it is available via application to UK Biobank. It is a powerful biomedical database that can be accessed globally to enable biomedical discoveries to improve public health.

	Name of data	Purpose of data	Relation to objectives of the project	Data type and format collected	Existing data: re-used	Origin of data	Size of data	Data Utility
WP5	Lucy App	To collect patient-shared data by using a new generation of period tracker, Lucy App, to increase both health literacy and patient empowerment. We will identify distinctive clinical cohorts, based on known digital footprints, symptoms, patient journeys, comorbidities, clinical severity, and lifestyle patterns.	To raise awareness in general practice and produce shared decision-making materials. The main output of this WP will be the development of patient profiles and structured clinical data.	JSON formatted data of questionnaire answers, lifestyle, and dietary data. Please use following link to enter the types and formats of the collected data: https://www.ietf.org/rfc/rfc4122.txt	Not decided yet.	Self-reported data.	1000 completed questionnaires.	It might be useful for researchers who want to validate questions on a similar topic.
WP6	Surgical phenotyping using machine learning.	Data (images) will be used to train a machine learning system for the automatic evaluation of endometriosis.	Automatic endometriosis lesion detection will help surgeons to improve treatment.	Laparoscopic surgery videos (RGB under white illumination.	No, all data is new.	Semmelweis University (Hungary) SurgAR (France)	50,000 – 100,000 lesion images	Data scientists, surgeons

	Name of data	Purpose of data	Relation to objectives of the project	Data type and format collected	Existing data: re-used	Origin of data	Size of data	Data Utility
WP7	Computer vision tool to detect endometriosis.	To automatically detect and classify endometriosis lesions using machine learning and assists surgeons for resection while operation.	The data collection is done for the objective of the enrichment of the data so that it can be used by artificial intelligence (AI) approaches so that the result can help surgeons to have a better surgery.	Videos: mp4, mkv, mpeg, etc. Annotation: JSON	SURGAR will re-use anonymized data to generate other neural networks to assist the surgeon and therefore benefit the patient.	The health centres with which SURGAR has legal contracts: University hospital of Clermont-Ferrand (France) University hospital of Semmelweis (Hungary) Beneficência Portugues Hospital (Brazil) Euroclinic Group of Hospitals - Athens (Greece) University hospital of Bologna (Italy)	> 100 surgical videos	The data will be beneficial for scientific outputs whether it is research or a scientific product. The data is useful to scientists to be used as big data for deep learning (or more generally machine learning) algorithms. The collected surgical videos will indirectly benefit patients through new tools (AI based) for better surgery. SURGAR develops augmented reality software in which the classification of endometriosis lesions and also the suggestion of division planes could be integrated. It aims at improving surgeries and therefore the health of the patients.

	Name of data	Purpose of data	Relation to objectives of the project	Data type and format collected	Existing data: re-used	Origin of data	Size of data	Data Utility
WP8 *Task 8.1*	Interview data.	Bring insight into qualitative experiences with the programme	Evaluate the feasibility and preliminary examine effects of the MY-ENDO programme	Audio files in m4a format and text/numerical in .doc format	No	Originate from study participant	-	These are sensitive personal *not publicly available due to restrictions pertaining to the Danish Data Protection Act and the General Data Protection Regulation.*
Task 8.2	Questionnaire data.	Gain quantitative and comparable data on the effects of the programme	Evaluation of the effects of MY-ENDO programme in a 3-armed RCT	Text/numerical in CSV, .doc, and excel format as well as SPSS syntax format	No	Originate from study participants	-	These are sensitive personal *not publicly available due to restrictions pertaining to the Danish Data Protection Act and the General Data Protection Regulation.*
Task 8.3	Translation data.	To describe the content of the MY-ENDO programme	To create an English and a Hungarian version of the MY-ENDO programme	Text/numerical in .doc format	No	Originate from the Danish version of MY-ENDO programme	-	These data will be publicly available on: www.endometriosis.org and in the LUCY app.

	Name of data	Purpose of data	Relation to objectives of the project	Data type and format collected	Existing data: re-used	Origin of data	Size of data	Data Utility
WP9	Website and Social Media.	Data collected is used for making meaningful reports regarding performance of FEMaLe project's communication and dissemination activities in digital and offline world. This will help the Consortium decide on the course of dissemination during the later months and influence the uptake and exploitation of the project results.		Text documents (.pdf, .odt), text file (.txt), spreadsheets (.xls, .xlsx, .csv), Presentations (.pdf, .ppt, .pptx), Images (raster) (JPEG, TIFF, PNG), Images (vector) (.svg)	Software used is Open-Source, thus allows the widest re-use. Data and software generated are owned by the beneficiary that generates them and is under open license to use for other beneficiaries in the project.	Interviews.	The datasets do not exceed a size of 100kb per user account	
WP10	Project management.	To facilitate the decision making and problem solving, and to coordinate the project via fine-tuning the calendar of activities and report efficiently to all partners.	The data collected in this WP are mainly from the remaining work packages in the project. Further, WP10 is responsible for the project practicing the principles of FAIR data.	Audio, videos (mp4), transcriptions (.txt, .pdf), Presentations (.pdf, .ppt, .pptx)	No	Online meetings (Executive Board meetings and the FEMaLe Consortium meetings). DMPonline from WP1-WP10.	N/A	The data is useful for the FEMaLe WP members to provide guidance for each of the WPs in relation to coordinate the project activities.

4. FAIR PRINCIPLES

FEMaLe project follows the FAIR data management concept to the extent that is possible. FAIR data management requires project data to be: **F**indable, **A**ccessible, **I**nteroperable and **R**e-usable.

4.1 Findable

To make data findable the project office had made platforms such as SharePoint and Google drive available for storing data in a safe and GDPR compliant manner. However, each work package is responsible to ensure that the collected data is easily discoverable. Since some work packages (WP3, WP7, WP8) collect patient data, the data is anonymised to protect patient privacy. The metadata collected should not identify or reveal the patient's information. Any raw data in the project is openly accessible. Currently, the data is not shared with the public, but in the end of the project, the collected data will be publicly available via http://www.findingendometrisis.eu/.

Table 4 illustrated an overview of the discoverability of data internally in the FEMaLe project.

Table 4: Discoverability of data in FEMaLe project

Platform	**Work packages**
SharePoint	WP1 WP2 WP8 (task 8.1) WP10
Miro	WP1 WP10
Google drive	WP1 WP2 WP10
SurveyXact	WP1 WP8 (task 8.2)
Zoom recordings	WP1 WP10
Personal drives that follow GDPR regulations (ex. drive server at Aarhus University, anonymized centre at SurgAR)	WP3 WP5 WP6 WP7 WP8 (task 8.2, task 8.3) WP9

The data analysis is shared with the European Commission, using the format D1.1, D1.2 D1.X etc.

The data collection and reporting follow a consistently style and structure, based on a template, that include information about: Title and ID of the project, Logos of the project and of the H2020 Program, Title of the document, Related Work package(s), Related task(s), Author(s), Dissemination level, Due submission date. and Actual submission date.

We provide clear versioning though naming the file “V1”, “V2”, “V3”, “Vx”, and the version history is saved in the software used. GIT control system is used in WP6 to trach changes in code or other files over time. Also, version numbers will be included in the file names and metadata for each dataset to ensure that the most recent version of the data is being used for analysis. This will help to ensure that the data is accurate, reproducible, and can be easily tracked over time.

There is no specific naming convention or specific keywords that we follow. Naming conventions are self-explanatory in the datasets, only for internal purposes. For example, in WP7 they use specific nomenclature to name the data by coding the three following criteria:

- Surgical specialty (here gynaecology).
- The data collection health centre.
- The data collection date.

Further, WP3 uses code books describing the variables and labels, which will be made accessible on the Departmental webpage when the data collection has ended. Scientific papers will be published with a persistent identifier (PID), such as Digital Object identifier (DOI), so that the research data and outputs are easier to find and cite.

The metadata provision will be stored in a repository available only to partners, and password protected. High online hosting security measures are in place to ensure that the data collected is safe from any breaches and potential misuse. Anonymized data will be stored using descriptive metadata.

For example, WP7 collects metadata for each surgical video, including a temporal indication of the surgery, surgical information like type of surgery, whether it contains suturing or bleeding, etc., pathology and the surgical procedure, the name of the healthcare centre who recorded the video. Existing standards are used, notably for the pathology referring to the ICD10 (international) classification as well as for the coding of the surgical act referring to the CCAM (French classification) classification.

4.2 Accessible

The FEMaLe project is committed to making its produced data openly accessible, as far as possible. All data generated by the project will undergo quality control according to the Grant agreement and will be accessible to the partners. However, data that includes sensitive information or general personal data affected by GDPR, or that raise any ethical concerns, will not be shared with others than relevant partners and stakeholders.

To ensure easy accessibility and findability, the projects data will be deposited in the AU repository at FEMaLe's SharePoint server in relevant fil formats, such as TXT, PDF, Excel, XLS, JSON, CSV. Please refer to Table 3 for further information on the formats used. The accessibility of the data depends on the size of the dataset, so it is in accordance with the GDPR-compliant regulations. As future findings are made, they will be published as scientific material or papers with a DOI number to enable other to find them. Table 5 illustrate the accessibility for data produced and/or used in the project.

Table 5: Accessibility for data produced and/or used in the project.

WP	Accessibility	Methods/tools needed to access the data
WP1	The data is not openly accessible, but available for internal project members only.	TBD
WP2	No data collected.	-
WP3	The raw data will not be openly accessible at any time due to GDPR. However, codebooks for the survey data, including category of questions, variable names and labels will be uploaded at the CYKLUS webpage once the data has been collected. Also, scientific papers will be published with a DOI number so that others can find it.	Data are stored at an encrypted drive at the Department of Public Health, Aarhus University. Only employees at Aarhus University with an affiliation with the project, has access to the data. Some data will be uploaded to Statistics Denmark and linked to register data. Only researchers employed at Aarhus University who has an affiliation with the project can get access to the data.
WP4	Task 4.3 miENDO/FENOX: genome-wide array-based genotype data (SNPs) - No meta-data will be created. - Controlled access to raw genotype data. - Data can be access via contacting the corresponding author/PI of each respective study. - Standard protocols for analysis of genotype data are widely published upon. Software, documentation, and code to be utilised depends on the objective of the analysis. Hence no particular standards will be provided.	TBD

	Task 4.3 miENDO/FENOX: Deep phenotype data - Summary of the utilized phenotype data will be published as part of the peer-review research article from this study. - Controlled access to raw phenotype data. - Data can be access via contacting the corresponding author/PI of each respective study. - Software, documentation, and code to be utilised depends on the objective of the analysis. Hence no standards will be provided.	
WP5	Openly available data: baseline questionnaire, lifestyle, and dietary module, which was developed in the framework of the FEMaLe project. Certain datasets which are collected from the Lucy app will be shared with the researchers only. The final data will be the patient profiles.	- Text editor, SQL Database Software. - Original data are not available. It is protected and stored on a private server. Data export will be made and uploaded to the FEMaLe share point. - Documents will be publicly available. - There will be no restrictions on who can access it.
WP6	Some parts of endometriosis lesion will be opened to validate other algorithms. The project coordinator is checking all legal aspects of sharing all dataset.	GIT and public custom-made metadata sharing platform.
WP7	SURGAR establishes collaboration contracts independently and directly with healthcare institutions. This does not allow free access to the data (according to consortium agreement which is based on REGULATION (EU) No 1290/2013 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 11 December 2013 laying down the rules for the participation and dissemination in “Horizon 2020 – the Framework Programme for Research and Innovation (2014-2020)”). Moreover, initially before processing, the data is characterized as sensitive, the process of free access to the data without agglomeration of information is complex to achieve.	The data and the associated metadata are stored in secured storage (HDS) provided by SURGAR on Google storage. This storage is not accessible by the public outside SURGAR company, and is accessible by limited people at the company, however it gives the possibility of database structures for future use. In the future, SURGAR will allow surgeons to access their anonymized videos of the operations they have performed through a dedicated secure platform. This access will be provided for teaching purposes and scientific publications only. Access to data is currently not allowed to the public outside of SURGAR
WP8	Interview data (task 8.1) and questionnaire data (task 8.2) are sensitive personal data and not publicly available due to GDPR restrictions. Translation data (task 8.3) will be publicly available on www.endometriosis.org and in the LUCY app.	TBD
WP9	The consortium will store the collected data in a format which is suited for long-time preservation and accessibility. To prevent file format obsolescence, some precautions will be taken. One such measure is to select file formats which have a high chance of remaining usable in the far future.	TBD

	As the data collected is restricted only to internal viewing and processing, the data will be made available through online document sharing systems (Correlate, Google Drive, Drobox) and password protected to prevent misuse. All software used to collect and preserve data will fall under the Common usage license (MS Office) or Open-Source (Google Drive). The access to data will be provided under the agreement with the whole Consortium and under supervision of the Project Coordinator.	
WP10	The collected data is confidential and not accessible for others.	TBD

In accordance with the Grant Agreement all research related data, excluding personal and sensitive data, will be stored for at least five years after the end of the research project (in case there is a high interest in the datasets or due to different national legislation, data may be stored for a longer period, which will be transparently discussed and approved within the consortium and relevant parties). Data that are used for publication will be stored at least five years after publication.

The Head of Department at the Department of Public Health at Aarhus University figures as contact person for access to data.

4.3 Interoperable

Ensuring data can be shared and used across different systems and platforms is a key to making data interoperable. To achieve this, the project is utilizing open standards such as JSON, SQL, CVL, and XML, as outlined in Table 3. However, due to data protection concerns, we are currently unable to exchange and re-use our data with researchers, institutions, organisations, and countries outside the project.

To enhance the interoperability and reusability of our data, we in include codes and scripts when publishing scientific outputs in the future. Additionally, the project partner is developing a metadata sharing platform that will share the best interoperability practices. The project coordinator is checking all legal aspects of licensing and re-using dataset.

To enable inter-disciplinary interoperability, some of the work packages is utilizing vocabularies, standards, or methodologies for the data set, as detailed in Table 6 below:

Table 6: Data and metadata vocabularies, standards, or methodologies

Work package	Data and metadata vocabularies, standards, or methodologies
WP3	Standard vocabularies will be used for all data types present in the dataset to allow inter-disciplinary interoperability. Codebooks including information on variables and labels will be uploaded at the CYKLUS homepage. The data will be stored in Excel and STATA.
WP7	The collected video data is standardized as much as possible. The following points can make the source video data interoperable: - The format of the videos are standard formats readable on normal machines. - The data is named according to certain coded metadata (which will not identify the patient) The traceability of data is well documented. The following points can make the annotations interoperable: - The annotations are made according to a documented ontology. - The annotations are provided by JSON formatted files, which provides clear, standard, readable, and interpretable information for any machine. The associated metadata follow a common classification, allowing very few free fields. Moreover, as already mentioned, we also use standardized classifications such as ICD10 or CCAM. Some interoperability standards are not directly applicable to the data we are studying. The annotations are finally converted to JSON format. This contains data in the form of key/value pairs, though making it easily interpretable and usable by any machine.
WP8	Text/numerical files are logically named by its content and saved in continuous numbered versions in folders.
WP9	To support the interoperability of the project data, a list of datasets is defined between partners to define the needed type of data and their relevance. The datasets defined should allow interdisciplinary interoperability (for example: website visits count connected to the self-management portal usage data which should allow further insights into the users' interaction with the research data publicized). If necessary, mapping to more commonly ontologies will be made available to further define the datasets. The defined datasets between beneficiaries are still in place and are being collected per agreed usage and importance. Most of the datasets defined allow the interdisciplinary interoperability. No need for mapping to more commonly ontologies arose during the initial period of the project.

4.4 Re-useable

The open and standard data formats used in the project, will help that the data can be easily read and used by others. The re-use of data also included explicit licenses to promote reuse and sharing of data, which in the project has not been decided yet how the research can be used by others. Most of the data is also underling to restrictions pertaining to the Danish Data Protection Act and/or the General Data Protection Regulation, which makes re-use of the data difficult. The following work packages are examples of how we reuse data collected in the project:

- **WP1:** The data is expected to become available for re-use at the end of the project period, when the dataset is larger and therefore usable for other projects. For example, we continuous collect data on the flow, impact and leadership of all project work packages related to the project management methodology. This data will be useful for other related research and innovation projects.

- **WP3:** The data collected in the survey, may be disclosed to research projects of major importance to society, provided that the relevant permits are obtained, and the applicable legislation is complied with. Such disclosure may be made pursuant to Section 10 of the Data Protection Act. As part of a research-based collaboration, IFS may transfer the personal data to recipients from other countries, both within and outside the EU. This applies only to the data collected via the questionnaire. Due to restrictions in Danish law for protecting patient privacy, the data from the national registers can only be made available through Statistics Denmark. Danish scientific organisations which are university-based can apply for authorization to work with deidentified data within Statistics Denmark, and such organisations can provide access to individual scientists inside and outside of Denmark. The data will be available for reuse until 14. December 2028.

- **WP9:** Software used is Open-Source, thus allows the widest re-use. Data and software generated are owned by the beneficiary that generates them and is under open license to use for other beneficiaries in the project.

Currently, the data is not useable by third parties and the length of time for which the data will remain re-usable are not specified yet. The data quality assurance processes are described in the initial data management plan.

4.5 Project overview: FAIR principles

The following Table 7 summarises the previous four sections about FAIR principles and provides an overview of the extent to which the FAIR principles are followed in each individual WP. In this way, it is possible to continuously track the development of the WP towards conducting more FAIR research. However, it should be noted that not all WPs could fully comply with the principles. Therefore, the principles are seen as a guide and not a goal of the project. Nevertheless, the FAIR principles still contributed to increasing the transparency of the FEMaLe Project and improve the quality of our output in the future.

Table 7: Heat map of FAIR principles

	Findable	Accessible	Interoperable	Re-useable
WP1	🟡	🟢	🟡	🟡
WP2	🟢	🟢	🟢	🟢
WP3	🟡	🟢	🟡	🟡
WP4	🟢	🟠	🟢	🟢
WP5	🟢	🟢	🟡	🟢
WP6	🟢	🟢	🟢	🟢
WP7	🟡	🟡	🟠	🟢
WP8	🟢	🟢	🟡	🟢
WP9	🟠	🟠	🟡	🟡
WP10	🟢	🟡	🟢	🟢

🟢 = 0%-25% 🟡 = 26%-50% 🟠 = 51%-75% 🔴 = 76%-100%

5. DATA SERCURITY

Data security and data storage includes the provisions made for the data created, including data recovery, as well as the secure storage and transfer of sensitive data. Each WP is responsible for ensuring that the data is stored securely in certified repositories, preferably for long-term preservation and curation. However, Aarhus University, as the project's lead organization, has the main responsibility for data storage. Table 8 illustrates how the individual WPs store and secure:

Table 8: Data security

WP	**Provisions for data security**	**Data storage**
WP1	The storage location will vary according to the partner who is working with the specific data. Identifiable information (participant e-mails, names etc.) will be kept only for work package leader. AU will be responsible for data collection. EQuiP will use the data to conduct online, descriptive data analysis, based on the collected data. The results will be shared with the FEMaLe Beneficiaries.	Data will be stored in the AU institutional servers at SharePoint, which complies with the overall conditions of data storage. All anonymised data will be stored for 5 years. The main data of the monthly FEMaLe Pulse Checks will be kept online in SurveyXact. The backup data of the Pulse Checks will be stored at Aarhus University, Bartholins Allé 2, DK-8000 Aarhus, Denmark on an encrypted drive.
WP2	No data collected	-
WP3	The consensus data are stored at an encrypted drive at the Department of Public Health, Aarhus University. Only employees at Aarhus University with an affiliation with the project, has access to the data. Once collected, the survey data will be stored at an encrypted drive at the Department of Public Health, Aarhus University. Only employees at Aarhus University with an affiliation with the project, has access to the data. Also, this data will be uploaded to Statistics Denmark and linked to register data. Only researchers employed at Aarhus University who has an affiliation with the project can get access to the data.	Consensus study: The consensus data includes information collected from patients, researchers, and healthcare workers on their view on the relevance of specific symptoms related to endometriosis. This information was collected using REDCAP and is stored in Excel and STATA format. Project CYKLUS: Self-reported information on symptoms related to endometriosis as well as health in general and lifestyle will collect from a random sample of Danish Women (N=60.000) using SurveyXact and stored as Excel and STATA files. Only researchers employed at Aarhus University can get access to data at this point.
WP4	-	UK Biobank Data: Two independent copies of the UK Biobank phenotype and genotype data are stored in PrecisionLife secure server space and Wellcome Centre for Human Genetics secure server space. Both institutions have their own agreements with UK Biobank. No transfer of data. CHB and DBDS Data: All CHB and DBDS phenotypic, genotypic, and proteomic data are stored in Computerome secure server space in Denmark. No transfer of data.

		FENOX Data: The FENOX phenotype data is currently saved in secure REDCap database hosted at University of Oxford. Once the proteomic data are generated, they will be secure transferred from the outsourced company to secure server space. miENDO Data: The miENDO phenotype data is currently saved in secure REDCap database hosted at the Capital Region of Denmark. Once the proteomic data are generated, they will be secure transferred from the outsourced company to secure server space.
WP5	No sensitive data collected	-
WP6	Project partner 'SurgAR' has experience in all aspects of medical data acquisition, storage, and transfer.	-
WP7	Google Cloud which is used to store the data is an HDS-certified host, meaning that companies that work with and in the French healthcare industry and that comply with France's General Security Policy for Health Information Systems (PGSSI-S) can confidently exchange, store data, and run workloads pertaining to French PHI on Google Cloud.	The anonymised data are stored on an HDS certified storage provider.
WP8	-	Sensitive data are stored on a secure network drive at the servers at Aarhus University and at SurveyXact. Personally identifiable information will be pseudonymized immediately. No sensitive data will be transferred.
WP9	Various types of data will be generated. The raw data will be stored by each partner according to their own standard procedures minimum for 5 years after ending of the project. Smaller, internal datasets will be stored on network drives and local drives, that are backed-up on external hard drives. Furthermore, smaller data sets and postprocessed data will be stored on cloud services, e.g., Google Drive, that is convenient for sharing data in a secure manner.	The processed data will become available in the form of project reports and open access publications. This data will be further exploited by the beneficiaries to produce meaningful reports. The data produced from for communication, dissemination and exploitation will be stored also on internal communication platform Correlate. This internal platform is only accessible for the project partners. Access to the data which is not marked as confidential will be granted via a repository.
WP10	The storage location will vary according to the partner who is working with the specific data. Identifiable information (i.e., participant e-mails, names) will be kept only for work package leader. All anonymised data will be stored for 5 years.	Data will be stored in the AU institutional servers at SharePoint, which complies with the overall conditions of data storage. AU will be responsible for data collection. The results will be shared with the FEMaLe Beneficiaries.

6. ETHICAL ASPECTS

As described in the sections above, the FEMaLe project is responsible for that all personal data collected complies with current GDPR rules or Ethical committee allowance. To avoid ethical issues, the project emphasizes research integrity, which involves clarification of confidentiality, anonymizing data, and obtaining written and verbal informed consent from all participants.

For the interview data in WP8 the research was conducted, based on the principles of the American Anthropological Association Guidelines for Ethics, specifically the *Principles of Professional Responsibility* (American Anthropological Association, 2012). The study was pre-registered at the Danish Data Protection Agency via Aarhus University´s internal registration (journal number: 2016-051-000001, running number: 2332) and all participants gave written informed consent regarding data collection and sound recording.

Also, for the Questionnaire data in WP8 the research will be conducted based on the *Ethical Principles of Psychologists and Code of Conduct* by the American Psychological Association (APA, 2002 including 2010 and 2016 amendments). In compliance with the rules of the General Data Protection Regulation *and the Danish Data Protection Act, t*he study is pre-registered at Aarhus University´s internal record (Journal number: 2016-051-000001, running number: 2332), approved by the Central Denmark Region Committees on Health Research Ethics (journal number: 1-10-72-178-22) and will also be pre-registered at clinicaltrials.gov. Further, WP4 has collected data from various databases, based on the following ethical approvals:

<u>UK Biobank ethics and data approval</u>
UK Biobank has been approved by the North West Multi-Centre Research Ethics Committee (MREC) as a Research Tissue Bank (RTB) approval. This approval means that researchers do not require separate ethical clearance and can operate under the RTB approval. Part our research is being conducted using the UK Biobank Resource under Application Number 44288.

<u>CHB and DBDS ethics and data approvals</u>
Copenhagen Hospital Biobank (CHB) is classified as a ‘biobank for future research’. It is part of the Danish National Biobank and has been approved by the Danish Data Protection Agency (general approval number 2012-58-0004, and local number: RH-2007-30–4129/I-suite 00678).

The CHB data utilized in the FEMaLe project WP4 is a sub-cohort of CHB with phenotypes related to reproductive health (CHB Repro). The study of genetics in CHB Repro has been approved by the Danish National Committee on Health Research Ethics (1805807) and the Danish Data Protection Agency (2019-49). Patients included in CHB were informed about their right to refuse the use of their samples for research via the Danish Tissue Utilisation Registry. The Danish Blood Donor Study (DBDS) was approved by the Central Denmark (1-10-72-95-13) and Zealand (SJ-740) Regional Committees on Health Research Ethics and the Data Protection Agency (P-2019–99). The analysis of genetic data (DBDS GWA study) was approved by the Danish National Committee on Health Research Ethics (1700407).

FENOX ethics approval

FENOX study has been approved by National Research Ethics Service (NRES) Committee South Central Oxford (09/H0604/58b).

miENDO ethics approval

Ethics approval for the miENDO study has been obtained from the Regional Ethics Committee of the Capital Region of Denmark (Protocol No. H-17017580). The study has furthermore been approved by the Danish Data Protection Agency (Protocol No. 2012-58-0004).

Aarhus University, CVR no. 31119103, is the data controller for the processing of all personal data. The FEMaLe group is responsible for the project (hereinafter referred to as the 'Project Group'). The project group is headed by Ulrik Bak Kirk and can be contacted at Bartholins Allé 2, building 1260, room 123, by email (ubk@ph.au.dk) or phone (+45 28 86 438 63).

7. OTHER

In the project we will continue to collect data and continuously discuss how and in what way our data can follow the FAIR principles, this to help other future projects and increase the impact of the FEMaLe project.

The DMP will be a living document, reviewed periodically and updated as necessary.