Grant Agreement no. 777167

BOUNCE

Predicting Effective Adaptation to Breast Cancer to Help Women to BOUNCE Back

Research and Innovation Action
SC1-PM-17-2017: *Personalised computer models and in-silico systems for well-being*

Deliverable 8.4: Data Management Plan

Due date of deliverable: (10-31-2020)
Actual submission date of the first version: (10-31-2019)

Start date of Project: 01 November 2017 Duration: 48 months

Responsible WP: <insert company name>

The research leading to these results has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 777167		
Dissemination level		
PU	Public	x
PP	Restricted to other programme participants (including the Commission Service	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (excluding the Commission Services)	

0. Document Info

0.1. Author

Author	Company	E-mail
Berta Sousa	CHAMP	berta.sousa@fundacaochampalimaud.pt
Haridimos Kondylakis	FORTH	kondylak@ics.forth.gr
Lefteris Koumakis	FORTH	koumakis@ics.forth.gr
Kostas Marias	FORTH	kmarias@ics.forth.gr
Akis Simos	FORTH	akis.simos@gmail.com
Evangelos Karadimas	FORTH	karademas@uoc.gr
Kostas Perakis	SiLo	kperakis@ep.singularlogic.eu
Gianna Tsakou	SiLo	gtsakou@singularlogic.eu
Poikonen-Saksela Paula	HUS	paula.poikonen-saksela@hus.fi

0.2. Documents history

Document version #	Date	Change
V0.1	01 April 2018	Starting version, template
V0.2	01 April 2018	Definition of ToC
V0.3	30 Sept 2019	First complete draft
V0.4	30 Sept 2019	Integrated version (send to WP members)
V0.5	20 Oct 2019	Updated version (send PCP)
V0.6	20 Oct 2019	Updated version (send to project internal reviewers)
Sign off	30 Oct 2019	Signed off version (for approval to PMT members)
V1.0	31 Oct 2019	Approved Version to be submitted to EU

0.3. Document data

Keywords	
Editor Address data	Name: Haridimos Kondylakis Partner: FORTH Address: N. Plastira 100, Heraklion, Crete, Greece Phone: +302810391449 Fax: E-mail: kondylak@ics.forth.gr
Delivery date	31 Oct 2019

1. Table of Contents

2. Introduction

This report describes the data management life cycle for the datasets collected, analysed, processed and produced by the BOUNCE project consortium. It presents how data are handled during the project, as well as how and what parts of the data sets will be made available following the project's completion.

In an overall view, Section 3 summarizes the data used throughout the BOUNCE project, i.e. the external, the retrospective and the prospective datasets. Then, section 4 elaborates on the FAIRification of the data for making them findable, interoperable and procedures implemented for increasing their value over reuse. Then in Section 5 we present resources allocated for storage and backup and in Section 6 we present data security aspects. In Section 7 we discuss ethical issues and finally section 8 concludes this deliverable and provides directions for the future.

It should be noted that although BOUNCE opted out from participating in the Open Research Data Pilot (ORD pilot), a Data Management Plan was considered as a useful guide for the project's data related operations. The report has followed the "Guidelines on FAIR Data Management in Horizon 2020" as published by the European Commission [1]. Given that most of the related activities and tasks are ongoing and have not yet produced their final results, the data management plan of the project will be revised and finally submitted in M36. The current form of the deliverable presents an intermediate version to be further refined and revised in the following year.

3. Data Summary

3.1. Purpose of the data collection/generation and its relation to the objectives of the project.

Coping with breast cancer more and more becomes a major socio-economic challenge not least due to its constantly increasing incidence in the developing world. There is a growing need for novel strategies to improve understanding and capacity to predict resilience of women to the variety of stressful experiences and practical challenges related to breast cancer. This is a necessary step toward efficient recovery through personalized interventions. BOUNCE will bring together modelling, medical, and social sciences experts to advance current knowledge on the dynamic nature of resilience as it relates to efficient recovery from breast cancer. BOUNCE will take into consideration clinical, cancer-related biological, lifestyle, and psychosocial parameters from relevant data, in order to predict individual resilience trajectories throughout the cancer continuum with the aim to eventually increase resilience in breast cancer survivors, help them remain in the workforce and enjoy a better quality of life. The overreaching goal of BOUNCE is to incorporate elements of a dynamic, predictive model of patient outcomes in building a decision-support system used in routine clinical practice to provide physicians and other health professionals with concrete, personalized recommendations regarding optimal psychosocial support strategies. As such collecting, integrating and processing various types of data is essential for completing the aforementioned project goals.

3.2. Data collected by the project

BOUNCE will deliver a unified clinical model of modifiable factors associated with optimal disease outcomes and will deploy a prospective multi-centre clinical pilot at four major oncology centres (in Italy, Finland, Israel and Portugal), where an estimate of 660 women will be recruited in order to assess its clinical validity against crucial patient outcomes (illness progression, wellbeing, and functionality). The advanced computational tools to be employed will validate indices of patients' capacity to bounce back during the highly stressful treatment and recovery period following diagnosis of breast cancer.

More specifically the data collected by the project include external, retrospective and prospective datasets, whereas the workflow is shown in Figure 1.

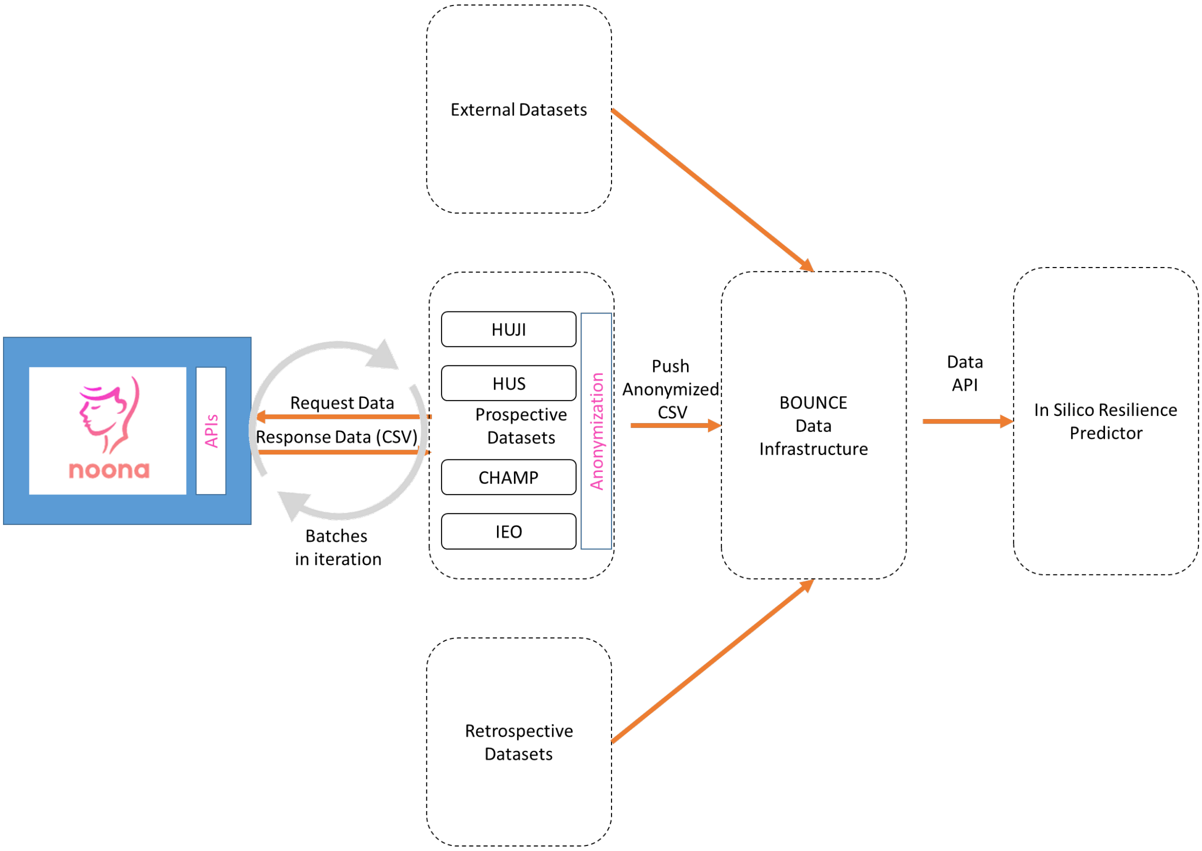

Figure 1. The data management methodology

External datasets have already been collected and stored centrally at the BOUNCE data infrastructure.

Retrospective datasets in addition have already been anonymized by the four individual clinical centres participating in the BOUNCE project and stored centrally in the BOUNCE data infrastructure.

For collecting the **prospective datasets**, patients are using the Noona tool to record their answers to the specified questionnaires at specific time points. In addition, all relevant patient medical and health data are also stored in the Noona system. Each individual clinical site requests data batches to be exported from Noona at frequent intervals. The data are provided to the clinical centers in CSV format and each individual centre, screens exported data and further anonymizes them in order to be pushed to the Data infrastructure. Currently three batches of data have been exported by the Noona tool, including more than 300 patients, they have been anonymized by the clinical centers and pushed to the BOUNCE Data infrastructure.

Within the BOUNCE data infrastructure the raw data are staged, both as CSV files exported from the Noona tool (and subsequently anonymized) and loaded in a PostgreSQL database. In addition, the data are replicated in order to be cleaned, homogenized and integrated. Then, both the raw data and the integrated data can be accessed using the data access API. The data access API communicates with the security layer in order to ensure that the access requests are properly authenticated and authorized. Initially the available data will be used for model development by the technical partners. Eventually the models implemented in the in Silico Resilience Predictor will access the data through the data access API, and the execution of the models will also take place in the central BOUNCE server, ensuring that no data leave the secure premises of the BOUNCE infrastructure. We have to highlight that the data access API offers read-only functionality over the available data and no data modification is possible using it.

However, before delivering the in Silico Resilience predictor tool, all consortium members are already able explore the available data through the *temporary research tool.* Using this tool consortium members will be able to visually explore available data using multiple graphs. We have to note that no export functionality is offered using the temporary research tool. In addition, modelers can upload their model to the model repository and to execute it using the execution engine provided. The models to be executed are monitored from the technical partners for proper usage of the available data- for example models, that only read and export the available data will not be accepted. Further the execution of the models is happening at the BOUNCE infrastructure ensuring that the data are properly accessed and processed.

In the following section, we present a high-level fingerprinting of these datasets. For a detailed description of those please refer to D1.3 [2] and D3.1 [3].

3.3. External Datasets

3.3.1.1. Breast Cancer Dataset

The Breast dataset[1] is a comprehensive dataset that contains nearly all the PLCO study data available for breast cancer incidence and mortality analyses. For many women the trial documents multiple breast cancers, however, this file only has data on the earliest breast cancer diagnosed in the trial. The dataset contains one record for each of the approximately 78,000 women in the PLCO trial.

Dataset Name	PLCO Breast dataset
Owner organization	NIH National Cancer Institute
Dataset description	
Dataset description (informal meta-data)	The Breast dataset is a comprehensive dataset that contains nearly all the PLCO study data available for breast cancer incidence and mortality analyses. For many women the trial documents multiple breast cancers, however, this file only has data on the earliest breast cancer diagnosed in the trial. The dataset contains one record for each of the approximately 78,000 women in the PLCO trial. It is a dataset that was already available when the BOUNCE project started. The data will be explored for their potential to be used in the project for building an in-silico model for resilience prediction.
Formal Meta-data	There is a detailed Data Dictionary available at https://cdas.cancer.gov/files/download/kwftwu0fr9/breast2nd.dictionary.nov18.d070819.pdf
Standards	No specific standards have been followed.

[1] https://biometry.nci.nih.gov/cdas/datasets/plco/19/

Origin	NIH National Cancer Institute
Language	In English language
Size	200 KB
Variety	Fully structured dataset
Type	Numeric
Format	CSV file
Velocity	Data are included in one CSV file
Storage	Data are currently stored in NIH but a copy is also available at the BOUNCE data infrastructure
Quality	High quality data, with not many missing values. In addition a quality and consistency check has already been performed by the owning institution.
Example	https://cdas.cancer.gov/datasets/plco/19/
Data sharing and ownership	
Availability	The whole dataset is available to the public through their download site. However a data request is necessary and a data transfer agreement should be signed. Through the BOUNCE platform, the data will not be made available to external entities.
Availability after the end of the project	The data will be available after the end of the BOUNCE project both internally in the BOUNCE platform for consortium members and through the original dataset's website.
Sharing mechanisms	Data are accessible through the original data web site. Within BOUNCE data are accessible through the Data Access API by all partners for downloading, viewing, and using.
Data Handling	
How the data are going to be used during BOUNCE's lifetime	The data were exploited already for their value for implementing a resilience model. However as they don't include any psychometric measure their value is really limited.
How the data are going to be used after BOUNCE's lifetime	The data are going to be used after the project lifetime to the extent they provide value to the developed models.
Related WPs	WP 3,4,5
Access rights	N/A
Anonymisation / Pseudonumisation	The data are already fully anonymized.

Collection workflow & methodology	Multiple questionnaires were answered at the participating clinical centers in the PLCO trial and then transferred to electronic forms.
Other information	
Ethics & GDPR	BOUNCE is not going to make these data available to members external to the consortium.
Person in charge	Haridimos Kondylakis for the version within the BOUNCE infrastructure.
Additional Cost	No additional cost was required to access the data or to store them.

3.3.1.2. ISPY1 Dataset

This dataset contains the clinical and MRI data from the ISPY1 clinical trial of patients with breast cancer[2]. The goal of this project was to improve the prediction of clinical outcomes to neoadjuvant chemotherapy in patients with breast cancer. Currently, most patients with breast cancer undergo neoadjuvant chemotherapy, which is aimed to reduce the tumor size (burden) before surgery to remove the tumor or the entire breast. Some of the patients response completely to the therapy and the patient does not present any residual tumor at the time of surgery. On the other hand, some patients have residual disease at the time of surgery and further treatment is required.

Dataset Name	ISPY1_Trial
Owner organization	This shared data set was provided by David Newitt, PhD and Nola Hylton, PhD from the Breast Imaging Research Program at UCSF, in collaboration with ACRIN, CALGB, the I-SPY TRIAL, and TCIA.
Dataset description	
Dataset description (informal meta-data)	The dataset includes data about 222 patients treated for breast cancer including demographics, clinical measures and outcomes It is a dataset that was already available when the BOUNCE project started. The data will be explored for their potential to be used in the project for building an in-silico model for resilience prediction.
Formal Meta-data	There is a data dictionary available for the various columns used.
Standards	No specific standards have been followed.
Origin	The I-SPY TRIAL
Language	The data are numeric
Size	13 KB
Variety	Fully structured
Type	Numeric
Format	Tabular format
Velocity	1 CSV/XLS file

[2] https://data.world/julio/ispy-1-trial

Storage	A copy of the data is available at the BOUNCE data infrastructure
Quality	High quality data, will not many missing values. In addition a quality and consistency check has already been performed by the owning institution.
Example	https://data.world/julio/ispy-1-trial/workspace/file?filename=clean_data_ISPY-1_clinica.csv
Data sharing and ownership	
Availability	The whole dataset is available to the public through their download site. Through the BOUNCE platform, the data will not be made available to external entities.
Availability after the end of the project	The data will be available after the end of the BOUNCE project both internally in the BOUNCE platform for consortium members and through the original dataset's website.
Sharing mechanisms	Data are accessible through the original data web site. Within BOUNCE data are accessible through the Data Access API by all partners for downloading, viewing, and using.
Data Handling	
How the data are going to be used during BOUNCE's lifetime	The data were exploited already for their value for implementing a resilience model. However as they don't include any psychometric measure their value is really limited.
How the data are going to be used after BOUNCE's lifetime	The data will be accessible through the BOUNCE data access API.
Related WPs	WP 3,4,5
Access rights	N/A
Anonymisation / Pseudonumisation	The data are already fully anonymized.
Collection workflow & methodology	Multiple questionnaires were answered at the participating clinical centers in the trial and then transferred to electronic forms.
Other information	
Ethics & GDPR	BOUNCE is not going to make these data available to members external to the consortium.
Person in charge	Haridimos Kondylakis for the version within the BOUNCE infrastructure.
Additional Cost	No additional cost was required to access the data or to store them.

3.3.2. Retrospective Datasets

3.3.2.1. HUJI

Dataset Name	BOUNCE_Israel_T1_Background.sav

	BOUNCE_Israel_T1_T6_Questionnaires.sav
Owner organization	HUJI (Hebrew University of Jerusalem) and Davidoff center in Beilinson Hospital, Israel
Dataset description	
Dataset description (informal meta-data)	The dataset contains numeric data of a sample of 198 breast cancer patients who participated in a trial assessing efficacy of a psychological intervention in the Davidoff Center, Rabin Medical Center, Israel. The data in the first file are background information, the data in the second file were collected via self-report at six waves of data collection. It includes pre-existing data The data were collected to assess the short- and long-term efficacy of a psychological intervention aimed at enhancing the resilience in coping with breast cancer. The data will be used in the project for building an in-silico model for resilience prediction. All of the data may be used in the future (after the end of the project) for secondary analysis and derivation of psychological models of coping with illness
Formal Meta-data	SPSS data with labels
Standards	No
Origin	HUJI/Davidoff research team
Language	The data are numeric
Size	200 KB
Variety	Fully structured
Type	Numeric
Format	SPSS .sav
Velocity	Two SPSS files
Storage	In Israel – with Ilan Roziner, in the BOUNCE data infrastructure
Quality	High quality, complete dataset with consistency and quality data control already performed.
Example	N/A
Data sharing and ownership	
Availability	The data are only available to representatives of HUJI, ICCS, FORTH and SiLo as described in the data sharing agreements and cannot be moved or transferred outside the BOUNCE data repository.
Availability after the end of the project	After the end of the project the dataset will not be freely available to the public. However, for the duration foreseen (five years) it will continue to be stored at the BOUNCE data infrastructure.
Sharing mechanisms	Data have been provided as two large sav files.

	Within the BOUNCE infrastructure, the data are available through the Data Access API.
Data Handling	
How the data are going to be used during BOUNCE's lifetime	The data are only available to representatives of HUJI, ICCS, FORTH and SiLo as described in the data sharing agreements, by HUJI/Davidoff team as specified in IRB permission.
How the data are going to be used after BOUNCE's lifetime	The data will be available to representatives of FORTH, ICCS and SiLo for one year as described in the data sharing agreements through the BOUNCE data repository available for further exploitation through new models that might be developed by consortium members and according to the exploitation plan developed within the consortium.
Related WPs	WP 3,4,5
Access rights	HUJI, ICCS, FORTH and SiLo have the right to integrate, clean, process and analyse the data.
Anonymisation / Pseudonumisation	The data are pseudonymized and all the personal identifiers are removed. The identifiers are not available for BOUNCE group. They are owned by the HUJI-Davidoff team and coded in paper only (not electronically available)
Collection workflow & methodology	Punched in into SPSS file from paper-and-pencil source
Other information	
Ethics & GDPR	Internal approval from Davidoff and HUJI for the usage of data has been received. Since the retrospective nature of the data no consent was required. Risk self-evaluation according the GDPR was performed.
Person in charge	Ilan Roziner ilanr@post.tau.ac.il
Additional Cost	No additional cost is foreseen required to access the data or to store them.

3.3.2.2. HUS

Dataset Name	BREX (for BOUNCE)
Owner organization	Helsinki University Hospital, Comprehensive Cancer Center
Dataset description	
Dataset description (informal meta-data)	Excel file, Patient (n=573) parameters related to breast cancer biology, treatments, symptoms, patients health, lifestyle, occupation, physical condition and quality of life. The data set is retrospective and it was collected before the BOUNCE project as part of the BREX exercise intervention study. Data is essential for the premodelling of resilience. Whole dataset has longterm value since it describes well a large breast cancer population and it is partly unpublished.
Formal Meta-data	The description of the dataset and files is included to the dataset
Standards	No standards were used
Origin	BREX study group File Maker Pro files

Language	English and Finnish (numbered scales and questionaries are not translated)
Size	3,5 MB
Variety	semi structured
Type	text
Format	sav
Velocity	static, no plans of updating the file
Storage	BOUNCE data infrastructure
Quality	Data is precleaned but some additional cleaning mechanisms are needed
Example	N/A
Data sharing and ownership	
Availability	The data is only available to representatives of ICCS, FORTH and SiLo as described in the data sharing agreements
Availability after the end of the project	The data is only available to representatives of ICCS, FORTH and SiLo as described in the data sharing agreements during the project and one year after it.
Sharing mechanisms	Files already shared with relevant partners Within the BOUNCE infrastructure, the data are available through the Data Access API.
Data Handling	
How the data are going to be used during BOUNCE's lifetime	Data is used according to data sharing agreements by ICCS, FORTH and SiLo.
How the data are going to be used after BOUNCE's lifetime	The data will be available to representatives of FORTH, ICCS and SiLo for one year as described in the data sharing agreements through the BOUNCE data repository available for further exploitation through new models that might be developed by consortium members and according to the exploitation plan developed within the consortium.
Related WPs	WP 3,4,5
Access rights	Data is used according to data sharing agreements by ICCS, FORTH and SiLo.
Anonymisation / Pseudonumisation	The data is pseudonymized and all the personal identifiers are removed. The identifiers are not available for BOUNCE group they are owned by the BREX study group and coded for paper only (not electronically available).
Collection workflow & methodology	Data was collected during the BREX trial to CRFs in Filer Maker Pro.
Other information	

Ethics & GDPR	Internal approval from HUS for the usage of data has been received. Since the retrospective nature of the data no consent was not required. Risk self-evaluation according the GDPR has been done.
Person in charge	Paula Poikonen-Saksela
Additional Cost	No additional cost is foreseen required to access the data or to store them.

3.3.2.3. CHAMP

Dataset Name	BOUNCE Breast Database Retrospective
Owner organization	Fundação Champalimaud
Dataset description	
Dataset description (informal meta-data)	The dataset is composed by retrospective data from breast cancer patients followed at the breast unit of the Champalimaud Clinical Center and includes biomedical, psychosocial and functional status data. The retrospective data includes medical, functional, demographic, and psychometric data collected in the Champalimaud databases and examines the correlation between biological and psychological factors. The collection of the data regards all the breast cancer patients treated with curative intent until 2018.
Formal Meta-data	Data description is available
Standards	• The breast unit is accredited by the European Society of Breast Cancer Specialists and follows international guidelines: St. Gallen Interational Expert Consensus Conference on the primary therapy of Early Breast cancer 2017 *(Annals of Oncology 28: 1700–1712, 2017 doi:10.1093/annonc/mdx308)*
Origin	• Retrospective data: ○ Personal health records- breast unit database. ○ Neuropsychiatry unit database.
Language	Portuguese
Size	-
Variety	Fully structured data
Type	Text, Number
Format	CSV or Excel file
Velocity	Static, no plans for updating the file
Storage	BOUNCE data infrastructure
Quality	Data is precleaned but some cleaning mechanisms are needed
Example	N/A
Data sharing and ownership	
Availability	The data are only available to representatives of ICCS, FORTH, and SiLo as described in the data sharing agreements.
Availability after the end of the project	The data is only available to representatives of ICCS, FORTH and SiLo as described in the data sharing agreements during the project and one year after it.

Sharing mechanisms	Files exported by CHAMP. Within the BOUNCE infrastructure, the data are available through the Data Access API.
Data Handling	
How the data are going to be used during BOUNCE's lifetime	Data is used according to data sharing agreements by ICCS, FORTH and SiLo.
How the data are going to be used after BOUNCE's lifetime	After the lifetime of the BOUNCE the data are going to be used according to the guide established by the project, made available for one more year.
Related WPs	WP 3,4,5
Access rights	Within the consortium the sharing will follow the rules described at the data sharing agreement document. Within the Champalimaud clinical center Albino Maia and Berta Sousa have the role to decide who has access to the information.
Anonymisation / Pseudonymisation	The data are pseudonymized and all the personal identifiers are removed. The identifiers are not be available for the BOUNCE group. For Champalimaud however, the pseudonyms can be matched to real patients in paper only (not electronically available).
Collection workflow & methodology	The data were collected by the clinicians during the appointments with the patients and through digital or paper questionnaires answered by the patients.
Other information	
Ethics & GDPR	The study was submitted to the Ethics Committee and to the Data Protection Office (DPO) approval. All the data are anonymized before being shared with the consortium partners and informed consent was signed where necessary.
Person(s) in charge	Berta Sousa, berta.sousa@fundacaochampalimaud.pt; Manuela Seixas, Manuela.seixas@fundacaochampalimaud.pt
Additional Cost	No additional cost is foreseen required to access the data or to store them.

3.3.2.4. IEO

Dataset Name	Dataset Retrospective IEO
Owner organization	European Institute of Oncology, IEO
Dataset description	
Dataset description (informal meta-data)	The dataset contains: sociodemographic variables (e.g., age, education, and marital status), psychological measurements (e.g., Distress, QoL), and biological data (e.g., TNM, lymph node status, previous or current oncological therapy).

	It includes only pre-existing data. Not all patients present exactly the same type of variables especially for the psychological content. Different types of data have been collected depending on the study so there will be missing values for some patients on different measurements. Type of data: Sociodemographic information: real numbers, continuous variables (e.g., age, number of children) and categorical (e.g., gender, marital status) Psychological data: real numbers, categorical data (numbers or labels). Biological data: real numbers, continuous variables (e.g., blood test, tumor proliferation rate) and categorical data (eg, Cytopathology test: alphanumeric data “T-B6000”)
Formal Meta-data	There is no meta-data information available.
Standards	The present study has been devised to comply with both national (i.e., GCPs) and international declarations (i.e. Declaration of Helsinki) regulating proper ethical research involving human subjects.
Origin	Existing breast cancer patient cohorts.
Language	Italian.
Size	The sample size is 900 patiens/rows.
Variety	Data will be fully structured.
Type	Text/Number
Format	Data in an Excel file.
Velocity	Static data. The dataset is provided to the consortium partners once-off.
Storage	Institutional database for breast cancer patients.
Quality	Data cleaning is required. The quality of the data is average.
Example	N/A
Data sharing and ownership	
Availability	The data is made available for the BOUNCE project as confidential according to Article 10.1 of the Consortium Agreement. As such the data can only be used by persons who have signed this agreement for the specific task for which it was made available to any individual consortium member and should not be distributed to others not involved in the task, or to a third party not involved in the project. Furthermore, the data or any result from its analysis shall not be published without the prior written approval of IEO.
Availability after the end of the project	Data is available only during the BOUNCE project and one year after its end for purposes of completing the task and must be destroyed afterwards.
Sharing mechanisms	downloadable files

Data Handling	
How the data are going to be used during BOUNCE's lifetime	The data can only be used by a person who has signed the afore mentioned agreement, and only for the specific task for which it was made available and shall not be distributed to others not involved in the task, or to a third party not involved in the task. Furthermore, the data or any result from its analysis shall not be published without the prior written approval of IEO. Data is available only during the BOUNCE project and one year after its end for purposes of completing the task and must be destroyed afterwards.
How the data are going to be used after BOUNCE's lifetime	Data is available only during the BOUNCE project and one year after its end for purposes of completing the task and must be destroyed afterwards
Related WPs	Data is only available for tasks related to WP3, WP4 and WP5 for persons who have signed a Notice and Agreement of Data Confidentiality and Access Rights.
Access rights	As defined by the data sharing agreements.
Anonymisation / Pseudonumisation	No patient names will be used in any documentation transmitted to the European Institute of Oncology. Items that are used to identify a patient include year of birth and registration number. The local data manager will keep an identification log for all patients entered in this trial including: • Patient's name • Patient's initials • Registration number • Date of birth • Date of registration
Collection workflow & methodology	Data will be retrieved from existing databases and personal health records
Other information	
Ethics & GDPR	1. The aforementioned data are extracted from the IEO database and matched with the specific patient. 2. Data are stored in compliance with GDPR regulation and in line with what declared within the informed consent specifically prepared for these retrospective studies and previously signed by the patient. 3. The data, processed by electronic means, are disseminated only in a strictly anonymous form, for example through scientific publications, statistics and scientific conferences
Person in charge	Massimo Monturano (massimo.monturano@ieo.it)
Additional Cost	No additional cost is foreseen required to access the data or to store them.

3.3.3. Prospective datasets

3.3.3.1. HUJI

Dataset Name	BOUNCE prospective
Owner organization	HUJI (Hebrew University of Jerusalem); Davidoff center in Beilinson Hospital, Israel; probably additional medical centers
Dataset description	
Dataset description (informal meta-data)	Numeric data of a sample of 200 breast cancer patients who participated in the BOUNCE study. The data are collected during the BOUNCE project lifetime. The data will be used in the project for model development and for validating the in-silico model for resilience prediction. All of the data will be used after the end of the project for secondary analysis and derivation of psychological models of coping with illness
Formal Meta-data	Data dictionary is available
Standards	No specific standards are used.
Origin	HUJI/Davidoff/other hospitals research teams
Language	The data are numeric
Size	Expected to be around 400 KB
Variety	Fully structured
Type	Numeric
Format	CSV/XLS/SAV
Velocity	1 large file gradually updated (each update will overwrite the old file)
Storage	with Ilan Roziner at HUJI, at BOUNCE data infrastructure
Quality	High quality data are currently collected, with almost no missing values. Consistency and quality checks are performed as each batch of data is exported.
Example	N/A
Data sharing and ownership	
Availability	The data are only available to representatives of HUJI, ICCS, FORTH and SiLo.
Availability after the end of the project	After the end of the project the data will be available and stored at the BOUNCE data infrastructure for five years.
Sharing mechanisms	Sent to WP3 by uploading the CSV/XLS file in the secure file repository. Then stored in the BOUNCE data infrastructure. Further accessible only through the data access APIs and through the temporary research tool
Data Handling	
How the data are going to be used during BOUNCE's lifetime	Data is shared through secured link to partners who are involved in WPs regarded to processing of the data , modelling, model computational implementation and integration and performance evaluation and analysis. The usage of data is described in the data sharing agreements.

How the data are going to be used after BOUNCE's lifetime	As will be agreed
Related WPs	Data is generated in WP6 and will be used in WP4, WP5, WP6 and WP7.
Access rights	As described in the data sharing agreements.
Anonymisation / Pseudonumisation	The data are pseudonymized and all the personal identifiers are removed. The identifiers are not available for BOUNCE group. They are owned by the HUJI-Davidoff tream and coded in paper only (not electronically available)
Collection workflow & methodology	Punched in into electronic form from paper-and-pencil source.
Other information	
Ethics & GDPR	Internal approval from Davidoff and HUJI / other hospitals for the usage of data will be received. Risk self-evaluation according the GDPR is already be performed.
Person in charge	Ilan Roziner ilanr@post.tau.ac.il
Additional Cost	No additional cost is foreseen required to access the data or to store them.

3.3.3.2. HUS

Dataset Name	PROSHUS Prospective pilot data set of HUS
Owner organization	Helsinki University Hospital, Comprehensive Cancer Center
Dataset description	
Dataset description (informal meta-data)	BOUNCE prospective pilot data of 250 patients treated in HUS including the information collected to the NOONA platform. This information includes psychosocial and quality of life questionnaires and side-effects reported by the patient and medical and treatment data collected from the hospital registry by the trial nurse. Data type: csv Data will be collected during the project M12-38 Data is needed for the modelling and validation of resilience predictor. Whole data has long term value
Formal Meta-data	Data dictionary is already available.
Standards	No specific standard is used
Origin	NOONA, HUS Patient record registry Uranus
Language	Finnish, English
Size	At most 1MB of data
Variety	Fully structured
Type	Numeric mostly in CSVs
Format	csv
Velocity	Data are being produced in certain intervals during the project.
Storage	BOUNCE data infrastructure

Quality	The quality of the data is already high. Additional cleaning mechanisms (e.g. missing values handling) will be implemented by the cleaning mechanism of the BOUNCE data infrastructure.
Example	N/A
Data sharing and ownership	
Availability	The whole dataset available under specific conditions (only for the WPs and partners who have to process the data according the study protocol, data sharing agreements for confidentiality and security according to GDPR))
Availability after the end of the project	The whole dataset available under specific conditions (as above) for five years after the end of the project
Sharing mechanisms	Sent to WP3 by uploading the CSV/XLS file in the secure file repository. Then stored in the BOUNCE data infrastructure. Further accessible only through the data access APIs.
Data Handling	
How the data are going to be used during BOUNCE's lifetime	Data is shared through secured link to partners who are involved in WPs regarded to processing of the data, modelling, model computational implementation and integration and performance evaluation and analysis. The usage of data is described in the data sharing agreements.
How the data are going to be used after BOUNCE's lifetime	As will be agreed
Related WPs	Data is generated in WP6 and will be used in WP4, WP5, WP6 and WP7.
Access rights	As described in the data sharing agreements.
Anonymisation / Pseudonumisation	Data are pseudonymised and coded before the delivery and the codes will remain in HUS.
Collection workflow & methodology	Data will be collected during the BOUNCE prospect pilot 1.11.2017-1.11.2019
Other information	
Ethics & GDPR	Ethical Committee approval was obtained before the start of the data collection Only pseudonymized and coded data without personal IDs is sharable between EU-countries Processes are compliant with GDPR Data sharing agreements has to be approved by HUS lawyer
Person in charge	Paula Poikonen-Saksela
Additional Cost	No additional cost is foreseen required to access the data or to store them.

3.3.3.3. CHAMP

Dataset Name	BOUNCE Breast Database Prospective
Owner organization	Fundação Champalimaud
Dataset description	
Dataset description (informal meta-data)	The dataset is composed by prospective data from breast cancer patients followed at the breast unit of the Champalimaud Clinical Center and includes biomedical, psychosocial and functional status data. The prospective data will have additional information about sociodemographics, lifestyle, phsycological & psychosocial and cognitive function to capture resilience, well-being and disease status. The information is gathered from 3 sources: ● **Personal health records:** From the records data related with patient's personal history, biological tumor characteristics, staging tests and treatment information such as radiotherapy, surgery and systemic treatment will be collected (The set of fields are described in the Appendix 1 - Table 1, 2 and 3). ● **Noona platform:** Will collect information related with the patients socio-demographic, psychosocial, well-being and performance status during the period of 18 months, through questionnaires (Appendix 2 & Appendix B of the document *"Predicting Effective Adaptation to Breast Cancer to Help Women to BOUNCE Back: a multicentre clinical pilot study"* describes the variables collected in the Noona platform). ● **The cognition sub-study:** The study aims to evaluate cognitive function longitudinally in breast cancer patients receiving chemotherapy or only endocrine treatment. To this aim several tests will be used for neuropsychology assessment (Appendix 3) at baseline, 6 months and 1 year. All the data has long-term value with the exception of data related with patients to which the state of their disease changed during the period of study.
Formal Meta-data	Data dictionaries are already available
Standards	The breast unit is accredited by the European Society of Breast Cancer Specialists and follows international guidelines: St. Gallen Interational Expert Consensus Conference on the primary therapy of Early Breast cancer 2017 (Annals of Oncology 28: 1700–1712, 2017 doi:10.1093/annonc/mdx308). Noona is a CE-marked class 1 medical device, in accordance with standards MEDDEV 2.1/6 January 2012, IEC 62304, and EU directive 93/42/EEC.
Origin	● Personal health records. ● Noona. ● Cognition sub-study.

Language	Portuguese
Size	-
Variety	Fully structured data
Type	Text, Number
Format	CSV or Excel file
Velocity	The biological and treatments information will be taken during the patients visits to the oncologists. The timeframe of these collection is dependent on the disease and treatment profile. The data collected by Noona will have seven assessment waves, for a duration of 18 months: baseline, which will occur after the first visit with the oncologist, Month 3 (M3), Month 6 (M6), Month 9 (M9), Month 12 (M12), Month 15 (M15), and Month 18 (M18). The information collected for the cognition sub-study will be at the baseline and M3 and M6.
Storage	The data will be stored in the Champalimaud clinical system and in Noona platform. Those datasets will be exported and sent to the BOUNCE data infrastructure.
Quality	The quality of the data is considered high, as it will be collected by the Champalimaud clinicians following the international standards for best clinical practices. To assure the quality of the data, quarterly assessments are going to be performed by Diana Frasquilho and Berta Sousa to validate the collected biological and treatments data. For Noona data, monthly assessments are performed to find incomplete questionnaires or incoherent answers. If such errors are found patients are contacted by phone to correct the mistakes. To reduce the mistakes patients can require a trained assistant to help them fill in the forms in Noona platform.
Example	N/A
Data sharing and ownership	
Availability	The data are only available to representatives of HUJI, ICCS, FORTH, and SiLo as described in the data sharing agreements.
Availability after the end of the project	The publication policy will be in accordance with the Consortium Agreement of the Horizon 2020 BOUNCE project.
Sharing mechanisms	Sent to WP3 by uploading the CSV/XLS file in the secure file repository. Then stored in the BOUNCE data infrastructure. Further accessible only through the data access APIs.
Data Handling	
How the data are going to be used during BOUNCE's lifetime	As it is described in the study protocol.
How the data are going to be used	After the lifetime of the BOUNCE the data are going to be used according to the guide established by the project, made available for five years.

after BOUNCE's lifetime	
Related WPs	Working packages 3, 4 and 5
Access rights	Within the consortium the sharing will follow the rules described at the data sharing agreement document. Within the Champalimaud clinical center Albino Maia and Berta Sousa have the role to decide who has access to the information. - Champalimaud clinical system: - Doctors and nurses have access to all the patients data within the system. - Noona: - Diana Frasquilho, Berta Sousa, Silvia Almeida, Luzia Travado have read access to all the data within the system. - Patients are going to have access only to their own data. - Cognition data: - Diana Frasquilho, Berta Sousa, Silvia Almeida, Raquel Lemos are going to have read and write access to all the data.
Anonymisation / Pseudonymisation	The data will be pseudonymized and all the personal identifiers will be removed. The identifiers will not be available for the BOUNCE group. For Champalimaud however, the pseudonyms will be matched to real patients in paper only (not electronically available).
Collection workflow & methodology	The data will be collected by the clinicians during the appointments with the patients and through digital or paper questionnaires answered by the patients for the period of the study.
Other information	
Ethics & GDPR	The study was submitted to the Ethics Committee and to the Data Protection Office (DPO) approval. All the data should be anonymized before being shared with the consortium partners and an informed consent form will be signed by the patients authorizing the use and share of the data.
Person(s) in charge	Berta Sousa, berta.sousa@fundacaochampalimaud.pt; Manuela Seixas, Manuela.seixas@fundacaochampalimaud.pt
Additional Cost	The acquisition of tablet devices for assist patients answering the questionnaires will be required with the total cost of 1000 € for the acquisition of 1 device.

3.3.3.4. IEO

Dataset Name	Dataset Prospective IEO
Owner organization	European Institute of Oncology, IEO
Dataset description	
Dataset description (informal meta-data)	BOUNCE prospective pilot data of patients treated in IEO including the information collected to the NOONA platform. This information includes psychosocial and quality of life questionaries and side-effects

	reported by the patient and medical and treatment data collected from the hospital registry by the trial nurse.
Formal Meta-data	There is no meta-data information available.
Standards	The present study has been devised to comply with both national (i.e., GCPs) and international declarations (i.e. Declaration of Helsinki) regulating proper ethical research involving human subjects.
Origin	Noona.
Language	Italian and English
Size	Almost 1MB of data
Variety	Fully structured
Type	Numeric mostly in CSVs
Format	csv
Velocity	Data are being produced in certain intervals during the project.
Storage	BOUNCE data infrastructure and IEO Secure repositories
Quality	The quality of the data is already high. Additional cleaning mechanisms (e.g. missing values handling) will be implemented by the cleaning mechanism of the BOUNCE data infrastructure.
Example	N/A
Data sharing and ownership	
Availability	The whole dataset available under specific conditions (only for the WPs and partners who have to process the data according the study protocol, data sharing agreements for confidentiality and security according to GDPR)).
Availability after the end of the project	The whole dataset available under specific conditions (as above) for five years after the end of the project
Sharing mechanisms	Sent to WP3 by uploading the CSV/XLS file in the secure file repository. Then stored in the BOUNCE data infrastructure. Further accessible only through the data access APIs.
Data Handling	
How the data are going to be used during BOUNCE's lifetime	Data is shared through secured link to partners who are involved in WPs regarded to processing of the data , modelling, model computational implementation and integration and performance evaluation and analysis. The usage of data is described in the data sharing agreements.
How the data are going to be used after BOUNCE's lifetime	As will be agreed
Related WPs	Data is generated in WP6 and will be used in WP4, WP5, WP6 and WP7.
Access rights	As described in the data sharing agreements.
Anonymisation / Pseudonumisation	The data are pseudonymized and all the personal identifiers are removed. The identifiers are not available for BOUNCE group. They are owned by the IEO team.
Collection workflow & methodology	Punched in into electronic form from paper-and-pencil source.

Other information	
Ethics & GDPR	Internal approval from IEO for the usage of data is received. Risk self-evaluation according the GDPR is already be performed.
Person in charge	Ketti Mazzocco (ketti.mazzocco@ieo.it)
Additional Cost	No additional cost is foreseen required to access the data or to store them.

3.4. Project deliverables

Besides data collected and processed through the lifetime of the project, the project is also producing output in the means of project deliverables and source code/models.

Regarding document deliverables. All intermediate and final versions of the deliverables are stored in an online collaboration platform CBMLBox, hosted by FORTH, offering to each partner independent access to important documents, code, meeting agendas, supporting materials, individual to-do lists and other miscellaneous project information. To ensure adequate quality of the deliverables, each deliverable is circulated among work package members and modified according to their comments. After the work package members have approved the deliverable it is reviewed by two internal reviewers (assigned per deliverable) who are reviewing the document using the template provided in D9.2, Annex B. Upon receipt of the reviewers' feedback, the partner responsible for the document is requested to do the required modifications and/improvements (if any), in an iterative process, until the reviewers approve the document. After the internal review, the document is sent to the STC which has three working days to make further comments. The partner responsible for the document modifies the document according to STC's comments (if any). Upon completion of this review phase, the Deliverable is treated as final, and is uploaded on the CBMLBox with a particular label (FINAL VERSION) and uploaded to the EU server by the Project Coordinator. Finally public project deliverables are announced at the web page, where the files are available for download.
Regarding source code/models. The technical partners develop their tools at individual source code repositories. As soon as a stable version is available the code is sent to FORTH and deployed at FORTH's premises. In case of models, those models are loaded to the model repository from where they can be executed. After the end of the project, source code may become available, published through partner's own git repository, providing the necessary links from the BOUNCE web page.

4. FAIR data

4.1. *Making data findable, including provisions for metadata*

Metadata about both the external datasets and the retrospective and the prospective datasets are already available and documented within D1.3 [2], D3.1 [3] and D3.2 [3]. In addition a semantic model, i.e. the BOUNCE ontology, is being currently constructed and the final version of it will be reported in D3.3. Using this model all available data will be mapped to the ontology terms ensuring that all data are persistent and unique ontology identifiers (URLs) will be assigned to each data item. Both the ontology and the datasets have a clear version number.

4.2. *Making data openly accessible*

It is not in the intentions of the BOUNCE project to make data openly accessible.

Only project public deliverables will be made publically available and accessible for download through the web page of the project.

4.3. *Making data interoperable*

As already identified within D3.2 [3] although multiple ontologies/terminologies are available for modelling medical and clinical information, there is a limited set of resources available for modelling psychological resources. Nevertheless, the BOUNCE project has contributed to a novel ontology to this domain, i.e. the BOUNCE psychological ontology. The ontology is able to model all scales that are currently used within the BOUNCE project, promoting as such the interoperability and the reuse of the available data and is presented in D3.3 [5]. In addition, the ontology includes mappings to the more commonly used ontologies further promoting data reuse.

4.4. *Increase data re-use (through clarifying licences)*

As already stated all available data will be used through the lifetime of the project and will remain in the BOUNCE data infrastructure for five more years after the end of the project, except retrospective datasets which will be available for one more year after the end of the project. Project's exploitation policy will define in the last year of the project the licencing scheme for using the available data (if any).

5. Allocation of resources

All costs for data processing have already been calculated and allocated within project's budget. All partners have already confirmed the availability of the prospective datasets for five years after the project completion for the prospective datasets and for one year for the retrospective, and have allocated the appropriate resources to this purpose.

In addition, a Data Control Committee (DCC) has been formulated, consisting of three coordinating partners which has been established to take over the role of data controller for the project. This committee will be in charge of implementing the DMP and assigning the relevant roles and responsibilities (from data capture and data quality assurance to metadata production and data archiving and sharing) as well as ensure that it is reviewed and revised, if necessary.
The Data Control Committee (DCC) is formed by:

- **Project coordinator:** Poikonen-Saksela Paula (HUS)
- **Scientific & Technical Manager**: Akis Simos (FORTH);
- **Ethical Manager**: Gabriella Pravettoni (IEO)

5.1. Storage and Back up

All data are stored in FORTH's private cloud, located in Heraklion, Crete, Greece. The private cloud consists of high-end servers that utilize the Openstack[3] cloud technology and all the facilities of the private cloud are hosted in an enterprise grade data centre that provides all the necessary technologies against common physical hazards, i.e. air-conditioning, fire protection, physical security access, redundancy in power supply etc, as well as firewall based network data access security and restrictions.

The private cloud provides both Virtual Machines (VMs) for hosting services and executing computational tasks as well as resources for storage services (file storage, object storage, and block storage). One out of many advantages of the cloud technology is its elastic and on demand scalability, which can expand or shrink to provide the necessary resources without under-utilizing the available infrastructure. So, even though for the needs of BOUNCE project there is currently allocated a storage space of 150GB, should the need arise this can easily expand to a scale of Terabytes (TBs) that are available in the pool of resources of the private cloud without any disruption or migration to its deployed technologies.

The storage facilities of the private cloud provide mechanisms to ensure data safety and integrity against software and hardware based hazards. The low-level storage is performed over a pool of disks linked with RAID[4] 60 hardware based technology, which ensures a high availability and fault tolerance of two hard drives, i.e. there has to be a simultaneous failure of three hard disks to result in data loss. On top of RAID technology lies the logical volume management (LVM) system which provides block storage for the Virtual Machines, such as the VM that hosts the BOUNCE services and data. A semi-automatic backup mechanism is available in the private cloud with the provision of additional storage pool of many TBs, where archive backups are taken when necessary to provide time snapshots –not incremental backups.

[3] https://www.openstack.org/

[4] https://en.wikipedia.org/wiki/Standard_RAID_levels

6. Data security

Data security concerns the protection of data from any potential accidental or intentional but unauthorised abuse, harm, disclosure or destruction with the utilisation of sophisticated and complex security controls and safeguards. In principle, data security is related with identification of the subject that attempts to access the data, authentication for the verification of the subject's identify and authorisation that controls the access to the underlying data based on the subject's identity.

In order to protect the underlying data in BOUNCE particular attention is paid into security and privacy issues by designing and deploying the necessary data security and privacy protection mechanisms. Hence, the protection of the personal or sensitive information is ensured by the adoption of the security by design principle incorporated within the BOUNCE that is applied to all the framework's operations covering on the one hand the data access control and on the other hand the security of the data during the whole lifecycle of the data exploitation (data in storage and data in transit), as well as the security of the technical interfaces.

In accordance with the security by design principle, data access control is applied in order to perform effective and efficient access control over the various resources of the framework and to formulate the access control decision that either grants or denies the access to a resource from a subject when a request is received. Within BOUNCE, the Role Based Access Control (RBAC)[5] paradigm is adopted for the implementation of the access control mechanism. RBAC is an approach restricting access to resources (objects) based on the role of the subject. In RBAC, a set of pre-defined roles that hold a set of privileges is employed and the subjects are assigned to these roles. Subjects assigned to different roles have access to different set of resources. The access is based on the person that assigns the roles to the subjects and the resources' owners that determine the privileges associated with a role for their resources. The access control mechanism evaluates a request based on the role assigned in the subject performing the request and the privileges of this role authorised to perform on the resource in order to permit or deny the access.

Within the context of BOUNCE, distinct roles are defined based on the list of identified actors of the framework and the appropriate set of privileges are set for each role. Moreover, the access control mechanism is a key component of the BOUNCE framework ensuring the appropriate access control is applied across all the operations of the BOUNCE framework. The authorisation, authentication and access control decision formulation is based on a token-based mechanism in which the state-of-the-art JSON Web Token (JWT)[6] open standard (RFC 7519)[7] is exploited. In a nutshell, JWT is a compact, self-contained secure way to exchange information between two communicating parties in the form of a JSON Object and contains the valuable information, such as the identity and the role of the subject, that is utilised in the access control decision formulation.

Security for data in storage is referred to any kind of security mechanisms employed for data stored physically in any digital form within the BOUNCE infrastructure. The scope of these mechanisms is to preserve the security, privacy and integrity aspects of these data and is built around the storage architecture of the BOUNCE framework by exploiting the security

[5] https://en.wikipedia.org/wiki/Role-based_access_control

[6] https://jwt.io/

[7] https://tools.ietf.org/html/rfc7519

mechanisms offered from the utilised storage solutions and their respective vendors, as well as the adoption of well-established approaches in this field. Hence, in the utilised storage solutions the appropriate configuration is set in order to ensure the hardening of the applied security for both the deployment of these solutions and the access to the data stored within these solutions. Furthermore, to safeguard to the integrity of the stored data the usage of checksum is exploited. Checksum is utilised to ensure that the underlying data have not been altered or corrupted in any way by applying a cryptographic hash function or checksum algorithm on a block of data or file during data ingestion that produces a small-size datum. The derived checksum is checked every time the specific block of data or file is accessed and any modification will produce a different outcome indicating the compromise of the integrity of the specific block of data or file.

Security of data in transit is referred to the security of data moving from one location to another, such as across the internet or through a private network and between the services and clients of the framework. Within the BOUNCE framework, the well-established and widely used Hypertext Transfer Protocol Secure (HTTPS)[8] is utilised in order to ensure the secure data transfer and communication providing data encryption via Transport Layer Security (TLS) at the RPC layer. The HTTPS is utilised in all communications between the framework's services and across the network ensuring the required security level within the framework.

Another critical aspect covered by the security by design principle that is adopted by the BOUNCE framework is the security of the technical interfaces of the framework. As the various components and services of the BOUNCE framework expose a number of technical interfaces in order to facilitate the flow of the required information with the rest components and services the enforcement of a security mechanism is necessary. Hence, for all technical interfaces the token-based mechanism with JWT that is utilised for the access control mechanism of the framework is also applied covering the appropriate authentication, authorisation and access approval aspects of the technical interfaces.

[8] https://tools.ietf.org/html/rfc2818

7. Ethical aspects

The BOUNCE project makes its best efforts to ensure that all data related processes, from collection and sharing to data research and sustainability shall take place in compliance with the legal requirements established by GDPR (General Data Protection Regulation).

7.1. Data Protection

The ethical, data protection and IPR issues surrounding the data research in BOUNCE will be extensively analysed in D.9.4 *Ethical Considerations and Compliance* (M48). In particular, the legal requirements for data sharing (i.e. informed consent and/or Ethics Approval, the data de-identification measures (pseudonymisation vs anonymization), the security obligations on part of platform administrator and processing partners. Compliance with the ethics requirements is to be ensured by WP9, which i.a. proves that data is collected and used for BOUNCE only after the receipt of written approval by the Ethics Committees; that data protection officers and/or authorities are duly involved.

Regarding data privacy and security by pseudonymisation, the two options, namely: anonymisation and pseudonymisation, were evaluated. With regard to the safeguards for research, Article 89 (1) GDPR advocates research on anonymous data “*where the research can be fulfilled by further processing which does not permit or no longer permits the identification of data subjects*”; however, the GDPR allows pseudonymisation, if anonymization would render the fulfilment of research goals impossible. The BOUNCE Consortium decided to apply pseudonymisation for the management of incidental findings, dictated by the ethics of clinical research, also operates on the possibility of linking any research findings back to the individuals. The implemented solution is thoroughly described in D1.3.

As already mentioned in Section 5.1, all data will be stored in the storage devices of the FORTH’s private cloud. FORTH has already encrypted all communication to the aforementioned infrastructure ensuring that only the technical partners and data researchers (acting as data processors) are able to access the technical Cloud resources. Each data processor can perform strictly its own individual tasks as assigned in the BOUNCE project workplan. Any data access will only be available on the pseudonymised data through the necessary security mechanisms, including secure REST services in the Virtual Private Network (VPN) of the Cloud infrastructure. No public access to the Virtual Machines (VMs) will be allowed, while the clinical partners will be able to connect to the BOUNCE user services only to specifically dedicated Virtual Machines. The tools and services which will be developed within the Cloud infrastructure will prohibit any direct access to the BOUNCE Integrated Cohort, while no delete operation will be available. Public IP addresses will be disabled to the development of the VMs, a limitation which will ensure that no unauthorized transmission or data breach in any BOUNCE data storage will be performed.

7.1.1. Data protection by default

Data protection by default is applied to BOUNCE project from the very beginning to provide extended safeguards related with the protection of personal data. By taking into account the nature, scope and context of processing, the data controller is implementing data protection by design including the purposes of processing regarding also potential risks that can affect the rights and freedoms of patients that are participating in the processing. All principles of article 5 of the GDPR are taken into account for the data protection by design. Security requirements of data are presented in section 6 while the rights of data subjects are protected with technical processing security measures as described in Article 32 of the GDPR. To support the data protection by design and by default in the BOUNCE project the following table is presented.

GDPR - Article 25 Obligations	**Data Phases**		
	Collection (M1-M42)	**Processing (M15-M42)**	**Harmonization (M15-M42)**
"the controller shall, both at the time of the determination of the means for processing and at the time of the processing itself, implement appropriate technical and organisational measures, such as **pseudonymisation**, which are designed **to implement data-protection principles, such as data minimisation**, in an effective manner and to integrate the necessary safeguards into the processing in order to meet the requirements of this Regulation and protect the rights of data subjects."	Article 5 obligations including consent forms & Protection of Personal Rights	Pseudonymization & Data Minimization	Pseudonymization & Data Minimization & Data Retention Policy & Safeguards of Human Rights
"The controller shall implement appropriate technical and organisational measures for ensuring that, **by default**, only personal data which are necessary for each specific purpose of the processing are processed. That obligation applies to the amount of personal data collected, the extent of their processing, the period of their storage and their accessibility."	Amount of Personal Data= Retrospective data + data about 660 patients & Extend of Processing=28 months & Period of Storage=one year after project ends[9]		
"In particular, such measures shall ensure that by default personal data are not made accessible without the individual's intervention to an indefinite number of natural persons."	Users= Clinicians	Users=Clinicians, Technical Experts	Users= Clinicians, Technical Experts

Table 1 Data protection by design and by default in BOUNCE

[9] (depending on the BOUNCE sustainability plan this might be subject to changes)

7.2. IPR and ownership

The issues of IPR and ownership with respect to the Integrative Cohort and project infrastructure are to be clarified by contractual arrangement between the project parties, namely by conclusion of the D8.5 IPR Management & Innovation to be delivered in M48.The separation between rights in data and rights in results produced from data will be taken under consideration. Pursuant to Article 24.1 EC-GA, the Parties, who contribute clinical data into the Project, as entered in the agreement on background, hold rights in the data. These Parties retain the rights in the raw data they contribute throughout the duration of the Project and allow such data to be processed in the Project under the terms as specified in the agreement on background and/or terms for the grant of Access Rights, as provided in Section 9, Article 9.2.6 and 9.3 CA, in particular.

As regards ownership in the BOUNCE integrated dataset, qualified as research result, it would fall under the rules of composite ownership:

"If Project Results are generated by two or more Parties and if contributions of the Parties are separately identifiable and/or constitute separate and/or independent works in themselves, but are assembled into a collective whole as inter-dependent parts (albeit without an intent to be merged to the point of being used as a whole only), the contributing Parties agree that such Results constitute composite work and shall be owned by the contributing Parties according to the contribution of each."

Following this definition, parties contributing datasets into the project would hold rights in the integrated dataset according to the contribution of each.

It should be noted, however, that given the complexity of the project's activities and the accompanying legal analysis, the aforementioned issues are still under discussion and analysis. The main issues associated with data protection obligations and management of IPR are expected to be resolved by M48.

8. Conclusions

This document presented the data management plan for the BOUNCE project, covering external, retrospective data and prospective data. The document presents in detail the data collection process, the metadata and accompanying documentation, the data preservation, sharing and access methods as well as the ethical and legal compliance of the DMP. Finally, the specific responsibilities regarding the DMP implementation and review (if required) have been described. As already explained this is only a preliminary data management plan. As more data are being accumulated through the clinical trial running in the clinical sites and results become available from the models this document will be further updated. In addition, as models of the project become available and produce their own data as a result, they will be included in the following versions of this deliverable. Furthermore the next version of the deliverable will include a detailed exploitation plan about the collected datasets and the licencing scheme that will be eventually implemented and finalized through the remaining months of the project.

9. References

[1] http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

[2] The BOUNCE Consortium, Deliverable:1.3 BOUNCE methodology, July 2018

[3] The BOUNCE Consortium, Deliverable:3.1 Identification of Internal and External Data Sources and Registries, July 2018

[4] The BOUNCE Consortium, Deliverable:3.2 Initial Semantic Model, November 2018

[5] The BOUNCE Consortium, Deliverable:3.4 Final Version Semantic Model, October 2019

[6] The BOUNCE Consortium, Deliverable:9.2 Quality Plan, November 2017