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:3.1 Identification of Internal and External Data Sources and Registries

Due date of deliverable: (31-07-2018)
Actual submission date: (31-07-2018)

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

Responsible WP: FORTH

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
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
Ketti Mazzocco	IEO	ketti.mazzocco@unimi.it
Flavia Faccio	IEO	Flavia.Faccio@ieo.it
Virginia Sanchini	IEO	virginiasanchini@gmail.com
Pasi Heiskanen	Noona	pasi.heiskanen@noona.com
Paula Poikonen	HUS	paula.poikonen-saksela@hus.fi
Ruth Pat Horenczyk	HUJI	ruth.pat-horenczyk@mail.huji.ac.il
Berta Sousa	CHAMP	berta.sousa@fundacaochampalimaud.pt
Kostas Perakis	SiLo	kperakis@ep.singularlogic.eu
Gianna Tsakou	SiLo	gtsakou@singularlogic.eu
Eleni Kolokotroni	ICCS	ekolok@mail.ntua.gr
Georgios Stamatakos	ICCS	gestam@central.ntua.gr

0.2. Documents history

Document version #	Date	Change
V0.1	01 May 2018	Starting version, template
V0.2	01 May 2018	Definition of ToC
V0.3	14 June 2018	First complete draft
V0.4	20 June 2018	Integrated version (send to WP members)
V0.5	30 June 2018	Updated version (send PCP)
V0.6	20 July 2018	Updated version (send to project internal reviewers)
Sign off	27 July 2018	Signed off version (for approval to PMT members)
V1.0	31 July 2018	Approved Version to be submitted to EU

0.3. Document data

Keywords	Methodology elaboration, state of the art, data workflow
Editor Address data	Name: Haridimos Kondylakis Partner: FORTH Address: N. Plastira 100, Heraklion Phone: +302810 391449 Fax: E-mail: kondylak@ics.forth.gr
Delivery date	31 July 2018

1. Table of Contents

2. Introduction

2.1. About 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 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.

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 a total 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. 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.

2.2. About task 3.1

The Task 3.1 focusses on the identification of both internal and external data sources.

Regarding external data sources, those include but are not limited to a) data sources including biological information, b) medical information, c) clinical information, d) contextual, and psychosocial information related to breast cancer.

In addition, Task 3.1 deals with the identification of internal data sources of the BOUNCE pilot sites both prospective and retrospective. Given that this task was in part accomplished in D1.3, the present deliverable provides an update on the retrospective data where an additional fingerprinting was performed and some details on the data cleaning required for the first dataset that became available (HUS). At this time necessary ethical approvals pertaining to the retrospective data sets from all four BOUNCE clinical have been secured and the data are expected to become available to the FORTH and ICCS teams shortly.

2.3. Purpose of the document

The purpose of this document is to report on the results of T3.1 regarding the identification of internal and external data sources and registries.

2.4. Work methods

Initially an extensive search was performed on the internet to identify potential data sources of interest to the BOUNCE consortium. Out of this, 53 potential resources were identified. Then those resources were carefully examined in order to record their content, their quality, and data access restrictions imposed. Their quality was recorded and finally they were ranked on their relevance to the BOUNCE project.

For internal data, a fingerprinting was performed on the retrospective data and details on the processing of data were provided. As the first of the four datasets is already available, we also report on cleaning actions implemented. Finally, we describe details of the protocol developed for collecting prospective data and elaborate on the process of data collection.

2.5. Main content of the document

The rest of this document has been structured as follows: Section 3 elaborates on external data sources identified and Section 4 on the Internal Data sources, including both the retrospective and the prospective data. Finally, Section 6 concludes this deliverable and presents our next plans.

3. External Data Sources

In order to identify potential data sources of interest to the BOUNCE consortium, we performed an extensive review of datasets available on the web for breast cancer. The results are shown in Table 1 where we shortly describe the content and context of each dataset, its availability and any restrictions on access and further usage, quality, and relevance to the BOUNCE project. Links to these data sources are provided in the Appendix 1.

Source ID	Title	Short Description of the data	Context of the data source (biological/medical/clinical/contextual)	Availability of the data	Restrictions on accessing the data	Quality	Of interest to BOUNCE
1	Breast Cancer Wisconsin (Diagnostic) Data Set	Number of instances: 569, Number of attributes: 32 (ID, diagnosis, 30 real-valued input features)	Medical imaging	Available online (file)	No	No missing data	Imaging data not so relevant
2	PLCO Breast data	Cancer Characteristics, demographics, family history etc.	Medical and Clinical Data	Available online (file)	No	Some missing data	Might be relevant
3	ISPY1 clinical trial	222 patients with cancer characteristics, MRIs, and clinical outcomes	Medical Imaging and Clinical Data	Available online (file)	No	Some cleaning is required, cleaning script is available	Might be relevant
4	NKI Breast Cancer Data	272 patients, 1570 columns including gene expression, patient info, treatment and survival	Gene Expression, medical and clinical data	Available online (file)	No	Some cleaning is required, cleaning script is available	Might be relevant
5	BIRADS evaluation of mammographies	961 instances, 6 attributes	Imaging data	Available online (file)	No	Some missing data	Not relevant
6	Breast Cancer Wisconsin Data Set	699 instances, 10 attributes	Imaging data	Available online (file)	No	Some missing data	Not relevant
7	Breast cancer and cervical cancer screenings	Breast cancer and cervical cancer screenings	Population data	Available online (file)	No	Some missing data	Not relevant
8	Breast cancer risk	Breast cancer risk markedly lower with serum 25-hydroxyvitamin D concentrations ≥60 vs <20 ng/ml (150 vs 50 nmol/L): Pooled analysis of two randomized trials and a prospective cohort	Clinical data	Available online (file)	No	Some cleaning is required	Not relevant
9	Slaoui Meriem	Outcomes of breast cancer in Moroccan young women correlated to clinic-pathological features, risk factors and treatment: a comparative study of 716 cases in a single institution	Medical and clinical data	Available online (file)	No	Some cleaning is required	Might be relevant
10	Nottingham Prognostic Index	Comparison of Nottingham Prognostic Index and Adjuvant Online prognostic tools in young women with breast cancer: review of a single-institution experience	Medical and clinical data	Available online (file)	No	Some cleaning is required	Not relevant
11	Breast cancer in England	provides information on incidence, mortality and survival of breast cancer	Population data	Available online (file)	No	Cleaned data	Might be relevant

12	National Cancer Registry	Age-standardized and age-specific incidence rates of cancer classified by cancer type, behavior, county, and HSE area. Treating and screening information. Relative survival statistics for selected cancers, years and HSE areas are also provided.	Medical and clinical statistical data	Available online (file)	No	Cleaned data	Might be relevant
13	National Cancer Institute	Available data from studies on the NCI Breast Cancer Repository	Gene Expression, medical and clinical data	Available online (file)	No	Cleaned data	Might be relevant
14	TCGA-BRCA	The Cancer Genome Atlas Breast Invasive Carcinoma (TCGA-BRCA) data collection is part of a larger effort to build a research community focused on connecting cancer phenotypes to genotypes by providing clinical images matched to subjects from The Cancer Genome Atlas (TCGA).	Medical imaging	Available online (file)	No	Cleaned data	Not relevant
15	CBIS-DDSM	A database of 2,620 scanned film mammography studies. It contains normal, benign, and malignant cases with verified pathology information. The scale of the database along with ground truth validation makes the DDSM a useful tool in the development and testing of decision support systems.	Medical imaging	Available online (file)	No	Cleaned data	Not relevant
16	ISPY1	Prospective study to test MRI for ability to predict response to treatment and risk-of-recurrence in patients with stage 2 or 3 breast cancer receiving neoadjuvant chemotherapy (NACT).	Medical imaging	Available online (file)	No	Cleaned data	Not relevant, MRI is not done routinely in BOUNCE
17	Breast Screening Programme, England	This report presents information about the NHS Breast Screening Programme in England and includes data on those invited for breast screening, coverage, uptake of invitations, outcomes of screening and cancers detected.	Screening data	Available online (file)	No	Cleaned data	Not relevant
18	NHSOF 1.4.iii	One-year survivors with breast, lung or colorectal cancer.	Clinical Data	Available online (file)	No	Cleaned, summarized data	Might be relevant
19	NHSOF 1.4.iv	Five-year survivors with breast, lung or colorectal cancer.	Clinical Data	Available online (file)	No	Cleaned, summarized data	Might be relevant
20	ICGC Data Portal	International cancer genome consortium data	Gene Expression, medical and clinical data	Available online (file)	No	Cleaned Data	Not Relevant
21	ICGC Data Portal	Breast Cancer Data from ICGC Data Portal	Gene Expression, medical and clinical data	Available online (file)	No	Cleaned Data	Not Relevant
22	NCBI	Multiple datasets available from published reports	Some summarized clinical and medical data	Available online (file)	No	Only the summarized data are mostly available	Not relevant.
23	FigShare	Multiple datasets available from paper figures	Some summarized clinical and medical data	Available online (file)	No	Only the summarized data are mostly available	Not relevant.

24	SEER Incidence Data, 1973-2015	National Cancer Institute initiative and is the premier source for cancer statistics in the United States. It includes, but is not limited to, incidence, mortality, prevalence, survival, lifetime risk, and statistics by race/ethnicity.	Clinical and Medical data	After request	No	Only the summarized data are mostly available	Not relevant
25	APA: Links to Data Sets and Repositories	Data from American Psychological Association. They include Census Data, Child Language Data, Consumer Behavior data, Health and Retirement data, Human Mortality, International Math and Science, Data on Aging, Child abuse and neglect, national elections, alcohol abuse child health, economics for international student assessment, income dynamics, winsonsin longitudinal study etc.	Various types of data	Available online (file)	No	Dependent on the dataset, some values are missing	Might be relevant
26	The Health and Retirement Study	A body of multidisciplinary data that researchers can use to address important questions about the challenges and opportunities of aging.	Biological, medical, clinical, genetic	Available online (file)	Some of the data are public to be downloaded whereas other can be accessed by supplemental agreement signed	Most of the data in the study have already been harmonized and cleaned	Not relevant
27	Breast Cancer Surveillance Consortium	Mammography registries with linkages to tumor and pathology registries	Medical, Clinical, Imaging data	Available online (file)	No	Data have already been cleaned and harmonized	Not relevant
28	Australian Breast Cancer Tissue Bank (ABCTB)	Data contributed by an Australian network of cancer clinicians, researchers, and patients. ABCTB privacy protection policy ensures patients' identities are not revealed and cancer researchers are the only individuals with open access to data.	Clinical, Lifestyle, Pathology and Treatment	Available online (file)	Requires expression of interest and approval	Data have already been cleaned and harmonized	Might be relevant
29	German Centre for Cancer Registry Data	Population-based cancer registries from each German federal state transfered to the German Centre for Cancer Registry Data, as required by the Federal Cancer Registry Data Act. Data are combined, quality-checked, analysed and evaluated, and the results published in collaboration with the public health institutions of the federal states.	Medical, Clinical	Available online (file)	Requires expression of interest and approval	Quality checked data	Might be relevant
30	The Cancer Imaging Archive	Freely accessible repository containing medical images and supporting data from cancer patients, organized by cancer type, imaging modality and research question	Imaging	Available online (file)	No	Quality varies across the datasets	Not relevant, imaging

31	Canadian Epigenetics, Environment and Health Research Consortium Network	Multi-stage funding commitment by the Canadian Institutes of Health Research (CIHR) and multiple Canadian and international partners. The two sites will focus on sequencing human reference epigenomes and developing new technologies and protocols; they will also serve as platforms for other CEEHRC funding initiatives, such as catalyst and team grants. The complementary reference epigenome mapping efforts of the two sites will focus on a range of common human diseases.	Epigenomic data	Available online (file)	No	High quality dataset	Not relevant
32	Inter-university Consortium for Political and Social Research	Social Political and Social Data	Social Political and Social Data	Available online (file)	Depending on the specific dataset	Most of the datasets have high quality	Not relevant
33	UK National Cancer Patient Experience Survey, 2015	A sample of more than 100000 patients including age, ethnicity, cancer type and data related to various stages on cancer journey from diagnosis through to treatment through to aftercare	Medical, Clinical and care experience data	Available online (file)	Registration is required to the UK data service	Preprocessed and cleaned dataset	Might be relevant
34	CDC's Social Vulnerability Index (SVI)	Social vulnerability refers to the resilience of communities when confronted by external stresses on human health, stresses such as natural or human-caused disasters, or disease outbreaks. ATSDR's Social Vulnerability Index uses U.S. census variables at tract level to help local officials identify communities that may need support in preparing for hazards, or recovering from disaster.	Socioeconomic data, minority data & language, housing and transportation	Available online (file)	No	Cleaned data	Not relevant
35	National Institute of Oncology of Rabat	Outcome of breast cancer in Moroccan young women correlated to clinic-pathological features, risk factors and treatment: a comparative study of 716 cases in a single institution diagnosed in 2009	Medical and Clinical data	Available online (file)	No	File with some missing values	Not relevant, BOUNCE patient age is over 40
36	Nottingham Prognostic Index and Adjuvant Online prognostic tools in young women with breast cancer	Data from 92 young women (< 40 years old) who presented with invasive breast cancer between the years 1998-2007 in the North Lincolnshire area in the United Kingdom. The data includes cancer characteristics and follow-up survival figures for each participant up to September 2014.	Medical and Clinical data	Available online (file)	No	File with some missing values	Not relevant, BOUNCE patient age is over 40
37	Health and Medical Care Archive: DS1 -	he primary objective of the CPES was to collect data about the prevalence of mental disorders, impairments associated with these disorders, and their treatment	Prevalence of mental disorders, impairments	Available online (file)	No	Cleaned data	Not relevant

	Collaborative Psychiatric Epidemiology Surveys (CPES), 2001-2003 [United States]	patterns from representative samples of majority and minority adult populations in the United States. Secondary goals were to obtain information about language use and ethnic disparities, support systems, discrimination and assimilation, in order to examine whether and how closely various mental health disorders are linked to social and cultural issues.	and treatment				
38	Australian Data Archive: Australian Survey of Social Attitudes	Records attitudes and behaviours on Religion, Birth Control and Sex Education, Australia's Population, Environment, Crime and Criminal Justice, Religiosity and Spiritual Life, Social Inequality, Old People in Society, Body Image, Elderly Care, Loneliness, Dental Care, Government Services, and Politics and Society. AuSSA 2009 also includes demographic and behavioural categories (Personal Background) that survey: sex, year born, income, education, employment, union membership, languages spoken, birthplace, household composition and religion. There are also questions about the partner of the respondent: employment, highest-level of education and income.	Social attitudes	Available online (file)	Registration required before downlo-ading.	Cleaned data	Not relevant
39	Australian Data Archive: The Australian Longitudinal Study on Male Health	Questions covered socio-demographics, health status, mental health and wellbeing, health behaviours, social determinants, and health knowledge and service use.	health behaviour data	Available online (file)	Registration required before downlo-ading.	Cleaned data	Not relevant
40	Pew research Center	Global Attitudes & Trends, Social & Demographic Trends	Contextual data	Available online (file)	Registration required before downlo-ading.	Cleaned data	Not relevant
41	Global Aging Data: Longitudinal Aging Study in India (LASI)	Education, gender, and state-level disparities in the health of older Indians. Includes also biomarker data	Contextual and health data	Available online (file)	Registration required before downlo-ading.	Data need harmonization	Not relevant, India is very different than BOUNCE population, not directly connected to BC

42	The Irish Longitudinal study on Ageing (TILDA)	Aspects of health, economic and social circumstances from adults > 50 years residing t in Ireland. Waves of data collection take place every two years. TILDA provides a comprehensive and accurate picture of the characteristics, needs and contributions of older persons in Ireland to inform and support improvements in policy and practice; advancements in technology and innovation; tailored education and training through an enhanced ageing research infrastructure; harmonization with leading international research to ensure adoption of best policy and practice and comparability of results.	Contextual and health data	Available online (file)	Registration required before downlo-ading.	Cleaned data	Not relevant
43	The Costa Rican Longevity and Healthy Aging Study	A set of nationally representative longitudinal surveys of health and life course experiences of older Costa Ricans. CRELES is part of the growing set of Health and Retirement Surveys being conducted around the world. CRELES is now composed of multiple waves of data from two birth cohorts.	Contextual and health data	Available online (file)	Registration required before downlo-ading.	Cleaned data	Not relevant, special populatio n not related to BOUNCE
44	Mexican Health and Aging study	The MHAS study aims to Examine the aging processes and its disease and disability burden in a large representative panel of older Mexicans; • Evaluate the effects of individual behaviors, early life circumstances, migration and economic history, community characteristics, and family transfer systems on multiple health outcomes; • Compare the health dynamics of older Mexicans with comparably aged Mexican-born migrants in the U.S. and second generation Mexican-American using similar data from the U.S. population (for example the biennial Health and Retirement Study HRS) to assess the durability of the migrant health advantage; • Assess the health of all components of the population from which migrants are selectively recruited: Mexican-born migrants living in the US, migrants who return to Mexico after various length of time in the U.S., and Mexicans with no residential history in the U.S.	Contextual and health data	Available online (file)	Registration required before downlo-ading.	Cleaned data	Not relevant, especiall y in Finland those problems are not common.
45	Midlife in the United States (MIDUS)	The database provides an opportunity to study the life course, intergenerational transfers and relationships, family functioning, physical and mental health and well-being, and morbidity and mortality from late adolescence through 2011. Genomic data are also E47:H47	Contextual and health data	Available online (file)	Registration required before downlo-ading.	Cleaned data	Not relevant
46	Wisconsin Longitudinal Study	WLS provides an opportunity to study the life course, intergenerational transfers and relationships, family functioning, physical and mental health and	Contextual, health and genomic data	Available online (file)	Registration required before	Cleaned data	Not relevant

		well-being, and morbidity and mortality from late adolescence through 2011. Genomic data are also available.			downlo-ading.		
47	Health Survey for England	a set of core questions, asked each year on general health and psycho-social indicators, smoking, alcohol, demographic and socio-economic indicators, questions about use of health services and prescribed medicines and measurements of height, weight and blood pressure.	Contextual and health data	Available online (file)	Registration required before downlo-ading.	Cleaned data	Not relevant
48	Study on Global AGEing and Adult Health (SAGE)	a longitudinal study collecting data on adults aged 50 years and older, plus a smaller comparison sample of adults aged 18–49 years, from nationally representative samples in China, Ghana, India, Mexico, Russian Federation and South Africa.	Contextual and health data	Available online (file)	Registration required before downlo-ading.	Cleaned data	Not relevant, population is different from target population
49	COURAGE: Ageing survey in Europe	harmonized SAGE-like study covering: health, disability, subjective well-being and health system utilization; social networks; and the built environment, in a sample of approximately 12,000 respondents	Contextual and health data	Available online (file)	Registration required before downlo-ading.	Cleaned data	Not relevant
50	Survey of Health, Ageing and Retirement in Europe	A multidisciplinary and cross-national panel database of micro data on health, socio-economic status and social and family networks of more than 120,000 individuals aged 50 or older (more than 297,000 interviews).	Contextual and health data	Available online (file)	Registration required before downlo-ading.	Cleaned data	Might be relevant
51	USA: National Longitudinal Surveys	A set of surveys designed to gather information at multiple points in time on the labor market activities and other significant life events of several groups of men and women.	Contextual and health data	Available online (file)	Registration required before downlo-ading.	Cleaned data	Not relevant
52	Database of Genotypes and Phenotypes (dbGaP)	Developed to archive and distribute the data and results from studies that have investigated the interaction of genotype and phenotype in Humans.	Biological, medical, clinical	Available online (file)	Registration required before downlo-ading.	Not harmonized data	not relevant t
53	Linked Life Data	Linked Open Databases containing BioGrid, CellMap, Chebi, CHEMBL, ClinicalTrials, DailyMed, DiseaseOntology, Diseasome, DrugBank, Gene Ontology, HapMap, HomoloGene, HPRD, HymanCYC, HumanPhenotypeOntology, NCBI Gene, Reacome, RxNorm, SIDER, SympromOntology, UMLS, Uniprot etc.	Contextual data	Can be accessed using a SPARQL enpoint	No	Harmonized and cleaned data	Not relevant

Table 1. External Data sources with potential interest to the BOUNCE consortium.

As appropriate for further examination we identified databases focussing on women, breast cancer patients, aged > 40 years, consistent with the BOUNCE target population, that recorded socio-demographic and psychological indices, as well as outcome variables (e.g., survival rates, quality of life, mental health).

Databases focussing on (a) the general population (e.g., data on ageing), (b) factors that were not included in BOUNCE (e.g., imaging, assessment of risk factors), or (c) factors or conditions not considered by BOUNCE (e.g., social vulnerability in communities, other diseases and health conditions), were excluded from further consideration.

As such, out of the 53 data sources examined only 14 (26%) were deemed potentially relevant, whereas only one dataset included psychosocial data and two datasets contained lifestyle and contextual data as well. Regarding the quality of the relevant datasets most of them seem to be preprocessed and of adequate quality, and data cleaning is explicitly requested on only 3 of them. Raw data is available in most cases.

The next step, after the identification of the relevant data sources, is to download the available data and to store them within the BOUNCE data management infrastructure making them amenable to further querying and analysis.

4. Internal Data Sources

4.1. Retrospective Data Sources

Bellow we describe the retrospective data sources that will be used during the BOUNCE lifetime. Although the generic descriptions of the retrospective data are also available in D1.3 in this deliverable we proceed to a more structured description of those, whereas in the Appendices a detailed list of the measures to be used by the data provider partners are provided.

4.1.1.HUS

Retrospective data collected by HUS are available at baseline and after 3, 6, 12, 18, 24, 30 and 36 months post diagnosis. The HUS retrospective data include:

- **Clinical data:** age, WHO class, menstruation after chemotherapy, menopausal status, menopause age, BMI, weight, height, bone mineral density, total cholesterol levels, Blood Glucose, Blood Pressure, pulse, any other disease also psychiatric, basic health status, disability status, physical pain
- **Breast and treatment data:** tumor size, pT, pN, histological type, metastatic lymph modes, receptor status (estrogen, progesterone), Her2 expression, type of breast surgery, type of axillary operation, type of treatment (herceptin, chemotherapy, radiotherapy, endocrine treatment)
- **SocioDemographics:** years of education, marital status, number of children, employment status, reason for not working
- **History and Life Style:** competing athlete, smoking, frequency and amount of alcohol consumption, reduced fat in the diet, increased vegetables, increased the amount of exercise etc.
- **Physical performance and activity:** mean figure 8 running, mean 2-km walking test, leisure time physical activity, self-reported physical activity, MET (metabolic equivalent)
- **Survival data:** local, distant relapse free-survival and overall survival
- **Psychosocial self-report questionnaires:**
    - **EORTC QLQ- C30:** A questionnaire of 30 items developed to assess the quality of life of cancer patients. It incorporates five functional scales (physical, role, cognitive, emotional, and social), three symptom scales (fatigue, pain, and nausea and vomiting), a global health status / QoL scale, and a number of single items assessing additional symptoms commonly reported by cancer patients (dyspnoea, loss of appetite, insomnia, constipation and diarrhoea) and perceived financial impact of the disease.
    - **EORTC QLQ- BR23**: It is a breast-specific module of the EORTC QLQ comprising 23 questions to assess body image, sexual functioning, sexual enjoyment, future perspective, systemic therapy side effects, breast symptoms, arm symptoms and upset by hair loss.
    - **WHQ** women's health questionnaire: It contains 37 items distributed among nine domains: depressed mood, somatic symptoms, memory/concentration, vasomotor symptoms, anxiety/fear, sexual behaviour, sleep problems, menstrual symptoms and attractiveness.

- **FACIT-F**-Functional Assessment of Chronic Illness Therapy-Fatigue questionnaire: It is a 13-item compilation of general questions assessing fatigue levels during common daily activities over the past week.
- **BDI** - Beck Depression Inventory short form: Finnish modified version of Beck's 13-item depression scale (**R-BDI**). The short form of Beck Depression Inventory is a screening instrument for assessing depressive symptomatology among the following domains: mood, pessimism, sense of failure, dissatisfaction, guilt, self-hate, suicide, social withdrawal, indecisiveness, body image, work inhibition, fatigue and appetite.

The time points that each type of data were collected are summarized in Table 2.

	T1 Base-line	T2 after 3 mouths	T3 after 6 months	T4 after 12 months	T5 after 18 months	T6 after 24 months	T7 after 30 months	T8 after 36 months
Breast and treatment data	✓							
Clinical data*	✓			✓				✓
Self – report clinical data**			✓	✓	✓	✓	✓	✓
SocioDemographic			✓	✓	✓	✓	✓	✓
History	✓							
Life Style	✓		✓	✓	✓	✓	✓	✓
Physical performance	✓			✓				✓
Physical activity	✓		✓	✓	✓	✓	✓	✓
Survival data								
Psychosocial self-report questionnaires	✓	✓	✓	✓		✓	✓	✓

*Reported by clinical personnel
** Comorbidities (including psychiatric diseases), health status, disability status, physical pain

Table 2. Time availability of HUS retrospective data

Table 4 provides more details on the datasets already exported from the HUS whereas a detailed list of the provided variable can be found in the Appendix 2.

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 is collected before the BOUNCE project in the context of the BREX Exercise Intervention study. Data is essential for the premodelling of resilience. The dataset has long-term 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 in the dataset
Standards	No standard used
Origin	BREX study group File Maker Pro files

Language	English and Finnish (numbered scales and questionnaires are not translated)
Size	3,5 Mt
Variety	semi structured
Type	text
Format	sav
Velocity	static, no plans of updating the file
Storage	BOUNCE data repository
Quality	Data is pre-cleaned but some cleaning mechanisms are needed
Example	Complete dataset already shared with WP3
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 exported already shared with WP3
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 is going to be used according to the guide established in the project, and remain available for one more year. More details will be defined as the project progresses.
Related WPs	WP 3,4,5
Access rights	Everyone within the BOUNCE consortium who signed the data sharing agreement is able to access the data.
Anonymization / Pseudonumization	The data is pseudonymized and all personal identifiers are removed (they are owned by the BREX study group and coded on 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 required. Risk self-evaluation according the GDPR has been done.
Person in charge	Paula Poikonen-Saksela (paula.poikonen-saksela@hus.fi)
Additional Cost	No additional cost is needed.

Table 3. Description of the data exported from HUS.

4.1.2. HUJI

Based on a sample of N=198 women after breast cancer. Data collected at the Davidoff Center, Rabin Medical Center.

Sample: Jewish female breast cancer patients between the ages of 26-72 (M= 50.45, SD=10.85), out of which 143 were born in Israel. Stages of breast cancer: Stage I (n=47) Stage II (N=107), Stage III (N= 37)

Most of the patients (n=177) received both chemotherapy and radiation treatment; the remaining (n=20) received exclusively chemotherapy. Additionally, 69.7% were taking trastuzumab (Herceptin) and 66.3% were receiving complementary hormone therapy.

Assessment time points: 5 times over a two year period with a 3-6-year follow (about 10% of the patients died).

Repeated Measurements	T1 Baseline	T2 after group TX 3 mouths	T3 After 6 months	T4 After 12 months	T5 after 24 months	T6 Follow up
Number of Participants	199	49	112	87	53	139

Table 4. Time availability of HUJI retrospective data

The retrospective data that will be provided include:
T1 Background data:

- **Demographics** information (age, sex etc)
- **Illness parameters (**stage and TX)
- **Physiological data** (sleep problems, obesity etc)
- **Life Style**

T1-T6 psychosocial self-report questionnaires
Posttraumatic stress symptoms. The Posttraumatic Stress Diagnostic Scale [8] was used to assess the severity of posttraumatic distress. The PDS is a commonly used measure of PTSD that assesses the frequency of 17 symptoms and symptom severity.

Functional impairment. Respondents rated their level of impairment in nine domains, including work, relationships with friends or family, or general satisfaction with life, using a scale from 0 (no impairment) to 5 (severe impairment).

Depression. Depressive symptoms were measured using the Center for Epidemiologic Studies Depression Scale [13]. The CES-D is a well-validated 20-item measure based on ratings in four primary symptom areas: (a) depressed affect; (b) lack of positive affect; (c) somatic symptoms; and (d) interpersonal difficulties.

Cognitive and emotion regulation. The Cognitive Emotion Regulation Questionnaire (CERQ) is a multidimensional, 18-item scale that identifies coping strategies used by respondents following stressful or negative life events [5]. Responses are organized into nine subscales, divided into *positive*: acceptance, positive refocusing, refocus on planning, positive reappraisal, putting into perspective and *negative:* self-blame, rumination, catastrophizing, and blaming others.

Coping flexibility. The Perceived Ability to Cope with Trauma (PACT) scale [2]. The PACT is a 20 item-scale that assess ability to cope with potentially traumatic event. The PACT is divided into two subscales: (a) forward focus, comprised of 12 items, and (b) trauma focus, comprised of eight items

Posttraumatic growth. The Posttraumatic Growth Inventory consists of 21 items designed to measure five interrelated subscales that reflect perceived positive outcomes reported after a traumatic event. These include: (a) realization of new possibilities ; (b) an increased sense of personal strength; (c) a greater appreciation of life) ;d) an increased sense of closeness with others ; and (e) spiritual growth

Ego Resilience. Fourteen items measuring the general construct of ego resilience.

Feeling Today. Three items: overall assessment of distress level, level of perceived resilience, and amount of hope for the future – designed for this study.

Distress. The 6-item Kessler psychological distress scale.

PCL-5. PTSD assessment checklist according to DSM-V criteria with 20 items.

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)	Pre-existing numeric data from 198 breast cancer patients who participated in a trial assessing efficacy of a psychological intervention at the Davidoff Center, Rabin Medical Center, Israel, aimed at enhancing the resilience in coping with breast cancer. 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. The data will be used in the project for building an in-silico model for resilience prediction. All data may be used in the future for secondary analysis and derivation of psychological models of coping with illness
Formal Meta-data	An SPSS file 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	2 large files
Storage	In Israel – with Ilan Roziner, they will be stored in the BOUNCE data management infrastructure
Quality	High, as the data have already been cleaned and checked for consistency.
Example	In the Appendix 3 examples from this dataset are provided.
Data sharing and ownership	
Availability	The data are only available to representatives of HUJI, FORTH, and SiLo as described in the data sharing agreements

Availability after the end of the project	For one year as agreed
Sharing mechanisms	A file to be sent to WP3. Then the data will become available through APIs and query endpoints.
Data Handling	
How the data are going to be used during BOUNCE's lifetime	The data are only available to representatives of HUJI, 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 and SiLo for one year as described in the data sharing agreements By HUJI/Davidoff team as specified in IRB permission
Related WPs	WP 3,4,5
Access rights	N/A
Anonymisation / Pseudonumisation	The data are pseudonymized and all the personal identifiers are removed. The identifiers are not available for BOUNCE group (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	N/A

Table 5 Description of the data exported from the HUJI.

4.1.3. IEO

The retrospective data is composed of multiple datasets from psycho-oncology studies, conducted between January 2009 and April 2018. One study (LIVE - "Life Improved by Exercise") investigated the effect of an 8-weeks exercise program on breast cancer survivors in terms of quality of life and global functioning. A second project ("iManageCancer") has two studies; one evaluated the efficacy and usability of e-health platforms developed within this Horizon 2020 project, another study constructed a family resilience questionnaire together with the Resilience Scale for Adults and the ALGA-BC questionnaire, a measure developed and validated in IEO, that investigates the psycho-cognitive profile of patients. Another study ("Rumination study") evaluated social and cognitive aspects that may be related to the degree of rumination, while a different study ("Invernizzi") analyzed the effect of endocrine therapy on cognitive functioning in older patients (60-70 years old). The "Magnifying Glass" study evaluated presence of PTSD symptoms from diagnosis to two years from diagnosis with three time-points. The "OncodermaCare" study evaluated the effect of cosmetic treatment on quality of life for patients undergoing radiotherapy. Finally, the "DIGNICAP" project investigated effectiveness of scalp cooling system DigniCap to prevent alopecia in primary breast cancer patients receiving adjuvant chemotherapy. The subjective perception of the device and quality of life were assessed at

baseline and after each cycle of chemotherapy. The estimated number of breast cancer patients that have participated in these studies is 900.

The above-mentioned datasets are composed of biomedical, psychosocial, functional and demographic data:

Biomedical data: TNM stage, nodal status, surgery type, menopausal status, early age menstruation, family history for tumors, tumor biology, Ki67, and type and duration of treatment (chemotherapy/HT/RT). Moreover, CRP, genetic risk factors and psychotropic medications were included whenever present in the patient's personal health record.

Functional data: frequency and amount of alcohol consumption and frequency and amount of cigarette consumption. Moreover, frequency and type of physical activity, sleep, and fatigue depending on the trial or whenever present in the patient's personal health record.

Psychological data: Depending on the specific clinical trial, data on the following variables can be present: distress, mood and emotional state, resilience, PTSD, psycho-cognitive profile, and counselling or support sessions

The questionnaires used to collect this data include the EORTC QLQ-BR23, EORTC QLQ-C30, FACT-B, IES impact of event scale, FACIT Fatigue scale, Distress Thermometer, POMS, FARE family Resilience Scale Questionnaire, RSA – Resilience scale for adults, MoCA, FACT Cog, Illness Perception Questionnaire (IPQ-R), mini MAC, MOS social support, Skindex, STAI)

Demographic data: age, education, socioeconomic status, marital status and number of children whenever present in the patient's personal health record.

The dataset is described in more detail in Table 6 and Appendix 4.

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.
Variety	Data will be fully structured.
Type	Text/Number
Format	Data will be in an Excel file.
Velocity	Static data. The dataset will be 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	Not currently provided
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 will be extracted from the IEO database and matched with the specific patient. 2. Data will be 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, will be 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

Table 6 Description of the data exported from the IEO.

4.1.4.CHAMP

The data that will be provided to the BOUNCE data infrastructure were obtained in the context of a survey of existing breast cancer patient cohorts focussing on biological, socio-demographic, functional and psychological variables that could influence resilience processes.

Biological variables include cancer type, treatment characteristics, and medical outcomes.

Socio-demographic data includes sociological (e.g. marital status) and demographic information (e.g. age). Psychological variables refer to emotional, cognitive and relational aspects of an individual.

Two databases were used to extract the biological and psychological variables of breast cancer patients that were evaluated in the neuropsychiatry unit. The type of assessment could vary based on the patients' needs at the clinical visit.

The inclusion criteria were female 40-65years of age at the time of diagnosis; Histologically confirmed invasive early or locally advanced operable breast cancer stage I to III.

Concerning psychological variables the following measurements were assessed:

1. The Distress Thermometer [12] - is a distress screening tool used to better identification of oncologic patients on psychological distress and management in the psycho-oncology department. This is a simple, self-report, pencil and paper measure consisting of a line with a 0-10 scale anchored at the zero point with "No distress" and at scale point ten with "Extreme distress". It includes also a problem checklist. The patient is asked to identify those problems from the checklist, which are contributing to their score.
2. Hospital Anxiety and Depression scale (HADs) [17]: It is a fourteen item scale, seven of the items relate to anxiety and seven relate to depression. The anxiety and depressive subscales are also valid measures of severity of the emotional disorder. It was validated also for the Portuguese population in various clinical samples.
3. Mini Mental Status- Examination (MMSE): is a screening tool used to assess objective cognitive function. It consists of a questionnaire with a maximum score of 30 points, grouped in seven categories: orientation to time (5 points); orientation to place (5 points); registration of three words (3 points); attention and calculation (5 points); recall of three words (3 points); language (8 points) and visual construction (1 point).
4. Addenbrookes Cognitive Examination Revised (ACE-R): Is a cognitive screening tool, originally designed by Mioshi et al. [11] to address the lack of MMSE sensitivity in the diagnosis of dementia. The overall result of ACE-R includes an amount equal to the result of the MMSE, and further allows the assessment of multiple domains. The Portuguese experimental version was developed in community and clinical samples geriatric [10].
5. Wechsler Adult Intelligence Scale subtests (WAIS III)- Wechsler Adult Intelligence Scale – III, Digit Span subtest, that comprises two modalities: Forward - repeat number sequences with increasing length, in the same order as presented aurally to access immediate memory; and backward – repeat digit sequences in reverse order, to achieve working memory. Symbol Search subtest: Working within a specific time limit, the examinee scans a search group and indicates whether one of the symbols in the target group matches. This subtest measures processing speed, short-term visual memory, visual-motor coordination, cognitive flexibility, visual discrimination, psychomotor speed, and speed of mental operation. The examinee completes this subtest using a response booklet, and not on his or her digital device.
6. Trail Making Test A and B: Originally created by the US Army psychologists to assess selective attention (Part A), divided attention, the ability to sequence stimuli, cognitive flexibility and the processing speed (Part B). The TMT-A & TMT-B were validated for the Portuguese population with an adult sample by Cavaco et al [3].

7. Stroop test: Assessment tool for executive functions, response inhibition and selective attention, originally developed by Stroop [15] and revised by Golden & Freshwater [9], in an American adult population. The validation for the Portuguese population includes a sample of participants from 15 to 100 years was published by Fernandes [7].
8. Beck Depression Inventory (BDI-II): The BDI is a 21-item, self-report rating inventory that measures characteristic attitudes and symptoms of depression [1]. It is validated for the Portuguese population by Campos & Gonçalves [4] in a community sample, with no cut-off score.
9. State-trait Anxiety Inventory: The STAI is a commonly used measure of trait and state anxiety [16]. It can be used in clinical settings to diagnose anxiety and to distinguish it from depressive syndromes. Form Y, its most popular version, has 20 items for assessing trait anxiety and 20 for state anxiety. It was adapted for the Portuguse population by Santos & Silva [14].
10. EORTC QLC 30: It has been widely used in clinical practice and clinical trials for measuring quality of life (QoL) in patients with cancer. Includes 30 items for 15 dimensions/scales: five functional scales (physical, role, cognitive, social, and emotional functioning), three symptom scales (fatigue, nausea/vomiting, and pain), five single-item symptom scales (dyspnea, sleep disturbances, appetite loss, constipation, and diarrhea), single-item scale for financial impact, and a global health status.

The dataset is described in more detail in Table 6 and Appendix 5.

Dataset Name	CHAMP retrospective
Owner organization	CHAMP
Dataset description	
Dataset description (informal meta-data)	Medical, functional, demographic, and psychometric data collected in CR/CCC databases to examine associations between biological and psychological factors. Includes all breast cancer patients treated with curative intent until 2017 at the Breast Unit of CR/CCC. All the psychological, functional, and biological data that will be extracted from the existing databases are listed in the Appendix.
Formal Meta-data	Meta data will be available describing the schema of the data
Standards	No specific standard have been followed.
Origin	Existing studies in CHAMP
Language	Portuguese, but exported values will be mostly numeric
Size	Not yet fully defined
Variety	Fully structured data
Type	Text/Number
Format	CSV file or Excel file
Velocity	Static data. The dataset will be provided to the consortium partners once-off.
Storage	In CHAMP hospital.
Quality	High quality data. Minor cleaning might be required.
Example	Data have not yet exported and as such no example is currently possible to be provided.
Data sharing and ownership	

Availability	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.
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 that will be shared with WP3
Data Handling	
How the data are going to be used during BOUNCE's lifetime	Data will be 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. More details will be defined as the project progresses.
Related WPs	WP 3,4,5
Access rights	Everyone within the BOUNCE consortium who signed the data sharing agreement is able to access the data.
Anonymisation / Pseudonumisation	The data will be pseudonymized and all the personal identifiers are removed. The identifiers will bot be not available for the BOUNCE group. For CHAMP however, the pseudonyms will be matched to real patients in paper only (not electronically available).
Collection workflow & methodology	Data have already been collected and are available at CHAMP hospital information system.
Other information	
Ethics & GDPR	Internal approval from CHAMP for the usage of data needs to be granted. Since the retrospective nature of the data no consent was required. Risk self-evaluation according the GDPR has been done.
Person in charge	Berta Sousa (berta.sousa@fundacaochampalimaud.pt)
Additional Cost	No additional cost is needed.

Table 7. Description of the data exported from CHAMP.

4.1.5. Current status and cleaning procedures

So far only retrospective data from HUS has been made available to WP3 for further processing. The remaining retrospective datasets have just received ethical committee clearance and will become available shortly.

Below we describe some data cleaning procedures already applied to the acquired data. Data cleaning involves actions to identify incomplete, incorrect, inaccurate or irrelevant parts of the data (screening/diagnostic phase) and then replace, modify, or delete the dirty, coarse or inconsistent data (treatment phase). It is a necessary processing step before any statistical or data mining method within WP4 is applied, so as to avoid inaccurate/biased conclusions, reduced credibility, reduced generalizability, and violation of statistical assumptions.

The data cleaning steps performed so far include:

- For every variable, comparison of all values to what is listed in the code/explanation manual provided along with the data (e.g. you may have a rating scale 1-5 but a value 9 appears). In the case of standardized questionnaires, e.g. EORTC QLQ-C30, WHQ, BDI, etc., values were compared against questionnaires' scales and derived overall scores were recalculated.
- Consistency checks between variables to identify erroneous inliers. For example, variables of the same meaning reported by both patients and clinical personnel at the same time point were compared.
- Realization of basic descriptive statistics for every variable of the dataset as well as joint statistics between variables. Descriptive statistics help identify outliers, inconsistencies, strange patterns in (joint) distributions and erroneous inliers (when viewed in relation to other variables). Current analysis has also been focused on identifying variables (a) with too many missing values, (b) with all cases falling in only one category and (c) consisting of very few events (<10) in the less frequent category (for bi-categorical variables). In order to have a sufficient sample size for accurate estimations, variables suffering from the above deficiencies were excluded.
- Understanding causes of missing values (MCAR, MAR, MNAR) and deciding on the best strategy to handle missing values for every variable.
- Reduction of data redundancy, i.e. removal of duplicate variables. Moreover, variables not relevant to the scientific problem addressed were removed.
- Re-formatting (e.g. yes/no variables or missing values have the same coding throughout dataset, translation of variable names, re-formatting of variables names to include temporal features), and re-organization of data (e.g. in long or wide format) for consistency purposes and to facilitate subsequent analysis.

4.2. Prospective Data Sources

In this section, we will present the measures selected to be recorded during the prospective pilot and the workflow captured in D1.3 for transferring data from Noona that will capture these data to the BOUNCE data infrastructure.

Standardized scales and questionnaires to be used in the prospective pilot study are listed below

- Ten item Personality measure (brief Big Five)
- PTSD Checklist
- Life Orientation Test Revised
- Sense of Coherence scale
- The Perceived Ability to Cope with Trauma
- Cognitive Emotion Regulation Questionnaire
- Mindful Attention Awareness Scale
- Modified Medical Outcomes Study Social Support Survey
- Connor Davidson Resilience Scale
- Illness Perception Questionnaire

- Mini-Mental Adjustment to Cancer Scale
- Cancer Behaviour Inventory
- Family Resilience Questionnaire
- MOS Adherence to Medical Advice Scale
- Post-Traumatic Growth Inventory
- EORTC Quality of Life questionnaire – General and Breast Cancer module
- Fear of Recurrence short form
- Hospital Anxiety and Depression Scale
- Positive and Negative Affect short form
- NCCN Distress Thermometer

In addition, one-item questions on relevant psychological constructs have been developed for the purposes of the study.

Socio-demographic and lifestyle variables will also be collected, particularly in reference to:

- Age
- Highest level of education
- Marital status
- Number of children
- Employment status
- Income
- Absence from work
- Smoking and alcohol/drugs consumption
- Weight and height
- Diet
- Exercise
- Number of counselling/support sessions
- Number of visits with physician/nurse/social worker

Questions related to sociodemographic variables will also be embedded in Noona. The following clinical variables will be retrieved from Personal Health Records:

- TNM stage
- ICD-10 Classification
- Tumor biology
- Surgery type and side

- Previous/ongoing oncological therapy
- Side effects due to therapy
- Menopausal status
- Family history of cancer
- Psychotropic medication
- Disease-free survival
- Basic laboratory tests (blood cell counts, CRP)
- Patient pathway data
- Biomarkers

Measurement time points

There will be seven assessment waves, over an 18-month period: 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). During the baseline measurement wave, which will occur within three to four weeks from diagnosis, only non-cancer-specific measures will be delivered (such as personality). Cancer-specific measures will be assessed from M3, when the patient has already had some meaningful experience with the illness.

At baseline and M12 assessments will be collected through Noona during with face-to-face encounters with a site researcher (nurse, psychologist, or social worker). During the first face-to-face encounter the researcher will demonstrate the Noona platform and give a short training, so that in the following time points the patient will be able to use Noona independently.

For those patients who do not want or are unable to use Noona, paper-and-pencil mode will be available.

4.2.1. Data Transfer from Noona

For more descriptions on the Noona healthcare system, please refer to D1.3. In this section we only elaborate on the two approach alternatives that are available for data sharing between the Noona platform and the Bounce infrastructure as shown in Figure 1.

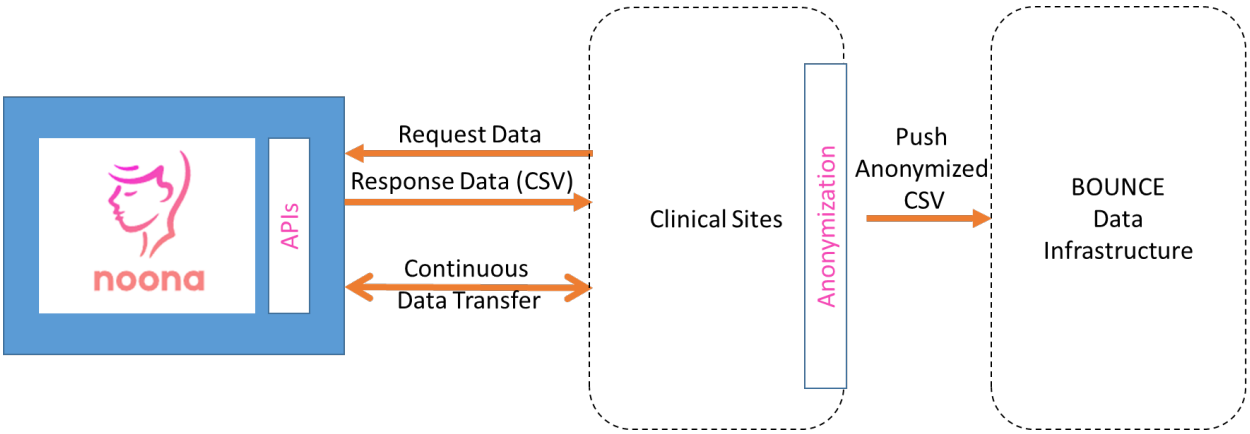

Figure 1. Data Transfer from the Noona platform to BOUNCE data infrastructure.

1. Request-response process

2. Continuous data transfer over API

Request-response functionality will perform a data export for a single clinics data with possible filters of Care team, treatment module and icd-10 code. The data exported contains all symptom, questionnaire, message, diary entry and timeline event data from the patients that fulfil the required filters. After a request has been approved, it will be available to the original requester. The data will be in zip file(s) which will be created when the user clicks Download in UI. CSV files will represent the actual data of the request along with a metadata layer.

Continuous data transfer over API is similar content-wise, however the method is not request based. Instead an API is built between Noona and a research data base, where data extracts can be streamed continuously, e.g. overnight for each day.

The data can be exported at the clinical sites which can then anonymize/pseydonymize them and send them for storage at the BOUNCE Data infrastructure.

Bellow, in **Error! Reference source not found.**, we present the fingerprinting table completed w ith the descriptions for the data to be exported from the Noona healthcare system.

Dataset Name	Noona Patient reported outcomes (PRO) data
Owner organization	Each site
Dataset description	
Dataset description (informal meta-data)	Patient reported symptoms, treatment related side effects, physical, emotional, and social wellbeing, and functionality. Data is collected during the project.
Formal Meta-data	n/a (not retrospective)
Standards	Yes. Standardized scales and single-item, validated questions
Origin	Patient reported via Noona service.
Language	Data is structured, so the raw data will be available in English, despite the language used to enter it.
Size	Depends on patient volume and collection duration. Approximately 10-50 data points per patient per month.
Variety	Fully structured. Some free text provided additionally, but that should not be significant.
Type	Multiselector values, numbers, text.
Format	n/a (not retrospective)
Velocity	Real time collection, batched transfer for study database. Transfer cycle can be agreed freely, e.g. between 1-30 days.
Storage	Amazon AWS loud environment, EU/ Ireland.

Quality	As everything is structured and all data points are mandatory, very little cleaning is anticipated. Cleaning needs would mostly relate to ruling out certain patients/ data sets due to incompleteness.
Example	n/a (not retrospective)
Data sharing and ownership	
Availability	For each study site to define. In principle, no limitations.
Availability after the end of the project	For each study site to define. In principle, no limitations.
Sharing mechanisms	Can be agreed separately, options are API or downloadable files.
Data Handling	
How the data are going to be used during BOUNCE's lifetime	Data could be shared to Bounce data working group. Noona can provide support in curating the use of the data.
How the data are going to be used after BOUNCE's lifetime	Data will be available through the BOUNCE data repository for one year after the completion of the project. More specific details on their usage after the end of the project will be defined the following months.
Related WPs	WP2,3,4,5,6,7
Access rights	Access to data in Noona is limited with access rights controls to only authorized personnel. Once data is extracted to the site personnel/ Bounce centralized team for data processing, the access rights controls should be arranged by site/ centralized teams.
Anonymisation / Pseudonumisation	Data is stored securely in Noona. Each clinical site will handle Pseudonumisation of raw data according to their practices/ jointly agreed practices for Bounce.
Collection workflow & methodology	Patient & clinician reported over an online service.
Other information	
Ethics & GDPR	For each study site to define. We assume anonymization (as required) at a central repository after Noona will make the data available. All use, identified of de-identified should be consented by the patient according to GDPR.
Person in charge	Pasi Heiskanen (pasi.heiskanen@noona.com)
Additional Cost	No additional resources required.

Table 8. Description of the data exported from the Noona healthcare system.

4.2.2. Description of the data to be transferred to BOUNCE data management layer

In Table 9 we present the dataset that will be sent after anonymization to be processed by the BOUNCE data management layer. This layer will include a storage engine and will provide all necessary APIs and query endpoints to make data accessible.

Dataset Name	BOUNCE_prospective (IEO, CHAMP, HUJI, HUS)
Owner organization	The four pilot sites (IEO, CHAMP, HUJI, HUS)
Dataset description	
Dataset description (informal meta-data)	Numeric data from breast cancer patients who will participate in the BOUNCE study. Overall goal is to construct empirically validated predictive models of patient resilience in coping with breast cancer. All data may be used in the future for secondary analysis and derivation of psychological models of coping with illness by consortium members.
Formal Meta-data	Meta data should be available. NOONA will responsible for the metadata of its contents.
Standards	No
Origin	The four pilot sites
Language	English
Size	Expected around 1MB eventually per pilot site
Variety	Fully structured
Type	Numeric and text
Format	CSV files
Velocity	1 large file per clinical site. Data will be provided in cycles according to the methodology of the project.
Storage	N/A
Quality	The quality of the data will be high as they will be exported from a fully-structured database. As all data will be exported by Noona, we expect that they will be consistent with a sufficient quality.
Example	N/A
Data sharing and ownership	
Availability	The whole dataset will be available under specific conditions (only for the WPs and partners who have to process the data according the study protocol, and in order to validate the model. Data sharing agreements for confidentiality and security according to GDPR will be signed.
Availability after the end of the project	The whole dataset will be available under specific conditions (as above) for a period of time to be determined later 1-3 years. This is subject to further discussions within the project.
Sharing mechanisms	Files will be sent to the BOUNCE data management layer. Then they will become available through APIs and SPARQL endpoints.
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, model validation 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 WP2, WP4, WP5, WP6 and WP7.
Access rights	N/A
Anonymisation / Pseudonumisation	The data will be pseudonymized and all the personal identifiers are removed. The identifiers are not available at the analysis layer. They are owned by the individual pilot teams and coded in paper only (not electronically available).
Collection workflow & methodology	Data will be collected during the BOUNCE prospect pilot 1.11.2018-1.11.2020
Other information	
Ethics & GDPR	Ethical Committee approval is needed Only pseudonymized and coded data without personal IDs is sharable between EU-countries Processes are compliant with GDPR Data sharing agreements have to be approved by a/the HUS lawyer
Person in charge	Ilan Roziner (HUJI) Paula Poikonen-Saksela (HUS) Ketti Mazzokko (IEO) Berta Sousa (CHAMP)
Additional Cost	N/A

Table 9. Description of the data sent by the pilots to the BOUNCE data management layer.

5. Conclusions

This document elaborates on all data that will eventually be managed from the BOUNCE data management infrastructure, including external data sources, retrospective data from pilot sites, and the prospective data to be available through the pilots.

The next step is to design a semantic model able to represent and describe all these data (T3.2) and eventually to download, aggregate, and clean the data (T3.3). Finally, through the semantic model designed, the data will be harmonized and stored within the BOUNCE data management layer, providing selective exposure to them through APIs and SPARQL endpoints.

6. References

1. Beck AT, Ward C, Mendelson M, Beck Depression Inventory (BDI). Arch Gen Psychiatry. 1961:4(6): 561–571.
2. Bonanno GA, Galea S, Bucciarelli A, Vlahov D, What predicts psychological resilience after disaster? The role of demographics, resources, and life stress. Journal of Consulting and Clinical Psychology, 2007: 75(5), 671–682.
3. Cavaco S, Gonçalves A, Pinto C, et al. Trail Making Test: Regression-based Norms for the Portuguese Population, Archives of Clinical Neuropsychology, 2013:28(2):189–198, https://doi.org/10.1093/arclin/acs115.
4. Campos RC, Gonçalves B, The Portuguese Version of the Beck Depression Inventory-II (BDI-II), Preliminary Psychometric Data with Two Nonclinical Samples, European Journal of Psychological Assessment 2011:27:258-264.
5. Garnefski N, Kraaij V, Cognitive Emotion Regulation Questionnaire: Development of a short 18-item version (CERQ-short).Personality and Individual Differences 2006:41:1045-1053.
6. Fatima A, Nazir N, Gufran KM, Data Cleaning In Data Warehouse: A Survey of Data Pre-processing Techniques and Tools. International Journal of Information Technology and Computer Science, 9, 50-61, 2017.
7. Fernandes S, Stroop: Teste de cores e palavras: Manual. Lisbon, Portugal: Cegoc, 2013.
8. Foa EB., Cashman L, Jaycox L, Perry K, The validation of a self-report measure of posttraumatic stress disorder: The Posttraumatic Diagnostic Scale. Psychological Assessment 1997:9(4), 445-451.
9. Golden CJ, Freshwater SM, Stroop Color and Word Test: Revised examiner's manual. Wood Dale, IL: Stoelting Co, 2002.
10. Gonçalves C, Pinho MS, Cruz V, et al., The Portuguese version of Addenbrooke's Cognitive Examination–Revised (ACE-R) in the diagnosis of subcortical vascular dementia and Alzheimer's disease, Aging, Neuropsychology, and Cognition, 2015:22:4:473-485.
11. Mioshi E, Dawson K , Mitchell J, Arnold R, Hodges JR, The Addenbrooke's Cognitive Examination Revised (ACE-R): a brief cognitive test battery for dementia screening. Int. J. Geriat. Psychiatry, 2006:21: 1078-1085. doi:10.1002/gps.1610.
12. O'Donnell E, D'Alton P, O'Malley C, Gill F, Canny Á, The distress thermometer: a rapid and effective tool for the oncology social worker. International journal of health care quality assurance, 2013:26(4):353-359.
13. Radloff LS, The CES-D Scale: A Self-Report Depression Scale for Research in the General Population, Applied Psychological Measurement 1977:1(3):385 – 401.
14. Santos SC, Silva DR, Adaptação do State-Trait Anxiety Inventory (STAI) – Form Y para a população portuguesa: Primeiros dados. Revista Portuguesa de Psicologia, 1997:32:85-98.
15. Stroop JR, Studies of interference in serial verbal reactions. Journal of Experimental Psychology, 1935:18(6), 643.
16. Spielberger CD, Jacobs G, Russell S, Crane R, Assessment of Anger: the State-Trait Anger Scale. Manual for the State-Trait Anxiety Inventory. Consulting Psychologists Press, 1983, Inc.; Palo Alto, CA.
17. Zigmond AS, Snaith RP, The hospital anxiety and depression scale. Acta Psychiatrica Scandinavica. 1983:67(6): 361–370.

APPENDIX 1 - Links to the external sources

External Data Source Id	URL
1	https://goo.gl/fKJ4AX
2	https://goo.gl/rKYUCa
3	https://goo.gl/f4VD77 Also at: https://www.cancer.gov/research/resources/data-catalog
4	https://goo.gl/9RXW2i
5	https://data.world/julio/mammographic-masses
6	http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29
7	https://data.europa.eu/euodp/en/data/dataset/75kK9hJE0S7CM2iDHPvvWw
8	https://zenodo.org/record/1219747#.WzSCmNIzaUk
9	https://www.datadryad.org/resource/doi:10.5061/dryad.2pc43
10	https://www.datadryad.org/resource/doi:10.5061/dryad.hc270
11	https://data.gov.uk/dataset/fd41f481-ebc2-4454-a83e-3794a65d0da7/breast-cancer-in-england
12	https://data.gov.ie/dataset/cancer-registry
13	https://portal.gdc.cancer.gov/exploration?filters=%7B%22op%22%3A%22and%22%2C%22content%22%3A%5B%7B%22op%22%3A%22in%22%2C%22content%22%3A%7B%22field%22%3A%22cases.primary_site%22%2C%22value%22%3A%5B%22Breast%22%5D%7D%7D%5D%7D
14	https://wiki.cancerimagingarchive.net/display/Public/TCGA-BRCA
15	https://wiki.cancerimagingarchive.net/display/Public/CBIS-DDSM#b8c12458a3e6438fb0ed9bb28f3ad636
16	https://wiki.cancerimagingarchive.net/display/Public/ISPY1
17	https://www.europeandataportal.eu/data/en/dataset/breast_screening_programme_england
18	https://www.europeandataportal.eu/data/en/dataset/one-year-survival-from-breast-lung-and-colorectal-cancer-nhsof-1-4-iii
19	https://www.europeandataportal.eu/data/en/dataset/five-year-survival-from-breast-lung-and-colorectal-cancer-nhsof-1-4-iv
20	https://dcc.icgc.org/repositories
21	https://dcc.icgc.org/search?filters=%7B%22donor%22:%7B%22primarySite%22:%7B%22is%22:%5B%22Breast%22%5D%7D%7D%7D&donors=%7B%22from%22:1%7D
22	https://www.ncbi.nlm.nih.gov/gds/?term=breast%20cancer and https://www.ncbi.nlm.nih.gov/home/download/
23	https://figshare.com/search?q=breast%20cancer&searchMode=1&types=3
24	http://www.seer.cancer.gov
25	http://www.apa.org/research/responsible/data-links.aspx
26	https://hrs.isr.umich.edu/about?_ga=2.255856012.688852898.1529908782-1068831334.1529908782
27	https://www.re3data.org/search?query=breast+cancer
28	https://www.abctb.org.au/abctbNew2/default.aspx
29	https://www.krebsdaten.de/Krebs/DE/Home/homepage_node.html
30	http://www.cancerimagingarchive.net/
31	http://www.epigenomes.ca/data-release/hg38/
32	https://www.icpsr.umich.edu/icpsrweb/
33	https://discover.ukdataservice.ac.uk/catalogue/?sn=8163&type=Data%20catalogue
34	https://catalog.data.gov/dataset/cdc039s-social-vulnerability-index-svi
35	https://www.datadryad.org/resource/doi:10.5061/dryad.2pc43%20
36	https://www.datadryad.org/resource/doi:10.5061/dryad.hc270
37	https://www.icpsr.umich.edu/icpsrweb/content/HMCA/index.html
38	https://dataverse.ada.edu.au/dataverse/tentomen?q=Australian+Survey+of+Social+Attitudes
39	https://dataverse.ada.edu.au/dataverse/tentomen
40	http://www.pewresearch.org/
41	https://g2aging.org/?section=page&pageid=26
42	http://www.ucd.ie/issda/data/tilda/
43	http://www.creles.berkeley.edu/
44	http://mhasweb.org/DataDocumentationNew.aspx
45	https://www.icpsr.umich.edu/icpsrweb/ICPSR/series/203
46	https://www.ssc.wisc.edu/wlsresearch/
47	https://www.ucl.ac.uk/hssrg/studies/hse
48	http://apps.who.int/healthinfo/systems/surveydata/index.php/catalog/sage/about
49	http://www.who.int/healthinfo/sage/courage/en/
50	http://www.share-project.org/
51	https://www.bls.gov/nls/#overview
52	https://www.ncbi.nlm.nih.gov/gap/
53	http://linkedlifedata.com/sources.html

APPENDIX 2 - HUS Retrospective Data Description and Coding

Note: Variables have been translated to English

Variable	Coding	Available at

		month
Patient number	Number	0
Birthdate	Date	0
Randomisation date	Date	0
Randomisation group	Exercise, Control	0
Menarche age	Age	0
Menopause status before adjuvant therapy	Postmenopausal (amenorrhea >12 months) Premenopausal	0
Menopause age before adjuvant therapy	Age	0
Last menstruation date before adjuvant therapy	Date	0
Hormone replacement therapy	Yes, No	0
Breast surgery	Mastectomy, Breast-conserving, Biopsy	0
Breast re-operation	Mastectomy, Breast-conserving, Other	
Breast re-operation specify	Free text	
Axillary surgery	Dissection, Sentinel node biopsy	0
Axillary re-operation	Dissection, Oher	
Axillary re-operation specify	Free text	
Tumor diameter	Number	0
Investigated lymph nodes	Number	0
Metastatic lymph nodes	Number	0
pT	T1, T2, T3, T4, Tis,Tx	0
pN	N0, N0i+, N1, N1mi, N2, N3	0
Histological type	Lobular, Ductal, Other	0
Histological grade	G1, G2, G3	0
ER	Positive, Negative	0
PR	Positive, Negative	0
Her2 IHC	Negative, +, ++, +++, Not done	0
Her2 FISH	Negative, Positive, Not done	0
Adjuvant CT	Yes, No	0
Adjuvant CT start weight	Number	0
Adjuvant CT start height	Number	0
Adjuvant CT start BSA	Number	0
CT regimen	1.6CEF, 2.3D+3CEF, 3.3DX+3CEX, 4.MUU	0
Neoadjuvant therapy	Yes, No	0
Herceptin	Yes, No	0
ET	Yes, No	0
ET agent	Astrozole, Exemestan, Letrazole, Tamoxifen, Other	0
ET Start date	Date	0
Radiotherapy (RT)	Yes, No	0
RT breast	Residual breast tissue, Scar	0
RT lymph nodes	Yes, No	0
RT total dose	Number	0
RT fraction dose	Number	0
RT booster	Yes, No	0

RT booster total dose	Number	0
Date of start	Date	0
Marital status	1. married or cohabitation 2. not married 3. divorced 4. widow 9. ND 10. other	0, 6, 12, 18, 24, 30, 36
Student years	Number	0
Births	Number	0
First birth	Date (Year)	0
State of health	1. good 2. quite good 3. middle level 4. quite bad 5. bad 9. ND 10. other	0, 6, 12, 18, 24, 30, 36
Disability	1. yes 2. no 9. ND	0, 6, 12, 18, 24, 30, 36
Myocardiac infarction	1. yes 2. no 9. ND	0, 6, 12, 18, 24, 30, 36
Cardiac insufficiency	1. yes 2. no 9. ND	
Arrhytmia	1. yes 2. no 9. ND	0, 6, 12, 18, 24, 30, 36
Other cardiac disease	1. yes 2. no 9. ND	0, 6, 12, 18, 24, 30, 36
Hypertension	1. yes 2. no 9. ND	0, 6, 12, 18, 24, 30, 36
Thrombosis	1. yes 2. no 9. ND	0, 6, 12, 18, 24, 30, 36
Stroke	1. yes 2. no 9. ND	0, 6, 12, 18, 24, 30, 36
Rhematoid arthritis	1. yes 2. no 9. ND	0, 6, 12, 18, 24, 30, 36
Arthrosis	1. yes 2. no 9. ND	0, 6, 12, 18, 24, 30, 36
Other joint disease	1. yes 2. no 9. ND	0, 6, 12, 18, 24, 30, 36
Back disease	1. yes 2. no 9. ND	0, 6, 12, 18, 24, 30, 36
Fracture	1. yes 2. no 9. ND	0, 6, 12, 18, 24, 30, 36
Osteoporosis	1. yes 2. no 9. ND	0, 6, 12, 18, 24, 30, 36
Psychatric disease	1. yes 2. no 9. ND	0, 6, 12, 18, 24, 30, 36
Which psyciatric disease	1. psychosis 2. depression 3. anxiety 4. drug abuse 5. other 9. ND	0, 6, 12, 18, 24, 30, 36
Which psyciatric disease 2	1. psychosis 2. depression 3. anxiety 4. drug abuse 5. other 9. ND	0, 6, 12, 18, 24, 30, 36
Which psyciatric disease 3	1. psychosis 2. depression 3. anxiety 4. drug abuse 5. other 9. ND	0, 6, 12, 18, 24, 30, 36
Diabetes	1. yes 2. no 9. ND	0, 6, 12, 18, 24, 30, 36
Severe headache	1. yes 2. no 9. ND	0, 6, 12, 18, 24, 30, 36
Urinary symptoms	1. yes 2. no 9. ND	0, 6, 12, 18, 24, 30, 36
Degree of disability in work	scale 0 to 10, 11 ND	0, 6, 12, 18, 24, 30, 36
Degree of disability in leasure time	scale 0 to 10, 11 ND	0, 6, 12, 18, 24, 30, 36

Back pain	scale 0 to 10, 11 ND	0, 6, 12, 18, 24, 30, 36
Neck pain	scale 0 to 10, 11 ND	0, 6, 12, 18, 24, 30, 36
Proximal shoulder pain	scale 0 to 10, 11 ND	0, 6, 12, 18, 24, 30, 36
Distal shoulder pain	scale 0 to 10, 11 ND	0, 6, 12, 18, 24, 30, 36
Hip pain	scale 0 to 10, 11 ND	0, 6, 12, 18, 24, 30, 36
Knee pain	scale 0 to 10, 11 ND	0, 6, 12, 18, 24, 30, 36
Reduced amount of fat	1. yes 2. no 9. ND	0, 6, 12, 18, 24, 30, 36
Changed amount of fat	1. yes 2. no 9. ND	0, 6, 12, 18, 24, 30, 36
Increased vegetables	1. yes 2. no 9. ND	0, 6, 12, 18, 24, 30, 36
Reduced sugar	1. yes 2. no 9. ND	0, 6, 12, 18, 24, 30, 36
Reduced salt	1. yes 2. no 9. ND	0, 6, 12, 18, 24, 30, 36
Lost weight	1. yes 2. no 9. ND	0, 6, 12, 18, 24, 30, 36
Increased exercise	1. yes 2. no 9. ND	0, 6, 12, 18, 24, 30, 36
Reduced alcohol	1. yes 2. no 9. ND	0, 6, 12, 18, 24, 30, 36
Reduced smoking	1. yes 2. no 9. ND	0, 6, 12, 18, 24, 30, 36
Alcohol use last 6 m	1. yes 2. no 9. ND	0, 6, 12, 18, 24, 30, 36
Beer	Number	0, 6, 12, 18, 24, 30, 36
Long drink	Number	0, 6, 12, 18, 24, 30, 36
Strong alcvolohol	Number	0, 6, 12, 18, 24, 30, 36
Wine	Number	0, 6, 12, 18, 24, 30, 36
Cider or light wine	Number	0, 6, 12, 18, 24, 30, 36
Frequency of alcohol use	1. never 2. less than once a month 3 once a months 4. once a week 5 daily or almost daily 9. ND 10. other	0, 6, 12, 18, 24, 30, 36
Present smoking	1. yes, daily 2. occasionally 3. never 9. ND	0, 6, 12, 18, 24, 30, 36
Daily number of cigarettes		0, 6, 12, 18, 24, 30, 36
Type of work	1. agricultural 2. factory, mine, construction or similar 3. Office, non-manual work, service 4. study or school 5. housewife 6. retired 7. unemploid 9. ND 10. other	0, 6, 12, 18, 24, 30, 36

Duration of working day	Number	0, 6, 12, 18, 24, 30, 36
competitive sport	1. yes 2. no 9. ND	0, 6, 12, 18, 24, 30, 36
Competitive sport age	Age (from to)	0, 6, 12, 18, 24, 30, 36
Exercise work duration	Number	0, 6, 12, 18, 24, 30, 36
Physical strain at work	1. mainly sitting 2. walking quite a lot 3 walking and lifting a lot 4. heavy physical work 9. ND 10. other	0, 6, 12, 18, 24, 30, 36
Erexcise on way to work before breast cancer	1. not working or working at homezwater exercise 6. other 9. ND	0, 6, 12, 18, 24, 30, 36
Exercise at leasure time before breast cancer	1. watching television 2. walking bicycling 3. proper exercise 4. competetitve exercise 9. not done 10. other answer	0, 6, 12, 18, 24, 30, 36
Exercise at leasure time before breast cancer, other	1. watching television 2. walking bicycling 3. proper exercise 4. competetitve exercise 9. not done 10. other answer	0, 6, 12, 18, 24, 30, 36
Type of exercise, mostly 1	1. ball game 2. gym 3. other gymnastics 4. running walking 5. swimming, water exercise 6. other 9. ND	0, 6, 12, 18, 24, 30, 36
Type of exercise, mostly 2	1. ball game 2. gym 3. other gymnastics 4. running walking 5. swimming, water exercise 6. other 9. ND	0, 6, 12, 18, 24, 30, 36
CRF Visits Menstrual cycle after therapy	Amenorrhea >12 months (postmenopausal) Unknown, Amenorrhea 6-12 months, Irregular, Regular (every 3-4 weeks)	0
CRF Visits Menopause status cause specify	Free text	0
Brex FULRFS years	Number	Follow-up 10-12 years
Brex FULRFS state	0= no, 1= event	Follow-up 10-12 years
Brex FUDDFS years	Number	Follow-up 10-12 years
Brex FUDDFS state	0= no, 1= event	Follow-up 10-12 years
Brex FUOS years	Number	Follow-up 10-12 years
Brex FUOS state	0= no, 1= event	Follow-up 10-12 years
Age	Number	0
Hospital in patient	1. yes 2. no 9. ND	6, 12, 18, 24, 30, 36
Hospital in patient, times	Number	6, 12, 18, 24, 30, 36
Hospital in patient, days	Number	6, 12, 18, 24, 30, 36
Doctor's appointment	1. yes 2. no 9. ND	6, 12, 18, 24, 30, 36

Number of doctor's appointments	Number	6, 12, 18, 24, 30, 36
Treatment due to mental problems	1. yes 2. no 9. ND	6, 12, 18, 24, 30, 36
Physiotherapy	1. yes 2. no 9. ND	6, 12, 18, 24, 30, 36
Physiotherapy, times	Number	6, 12, 18, 24, 30, 36
Duration of working day hours	Number	6, 12, 18, 24, 30, 36
Duration of working day-minutes	Number	6, 12, 18, 24, 30, 36
Change of work due to disease	1. yes 2. no 9. ND	6, 12, 18, 24, 30, 36
Amount of leisure time exercise	1. None 2. Some times per year 3. 1-3 times per month 4. Once a week 5. 2-3 times per week 6. 4-5 times per week 7. more than 5 times per week 9. ND (no data) 10. other	6, 12, 18, 24, 30, 36
Type of exercise, mostly	1. ball game 2. gym 3. other gymnastics 4. running walking 5. swimming, water exercise 6. other 9. ND	6, 12, 18, 24, 30, 36
At work last 6 m	1. yes 2. no 9. ND	6, 12, 18, 24, 30, 36
Work situation now	1 at work, sick leave ended 2 at sick leave, which started 3 presently not at work 9 ND 10 other	6, 12, 18, 24, 30, 36
Physical strain at work	1. mainly sitting 2. walking quite a lot 3 walking and lifting a lot 4. heavy physical work 9. ND 10. other	0, 6, 12, 18, 24, 30, 36
Physical strain at work, other	1. mainly sitting 2. walking quite a lot 3 walking and lifting a lot 4. heavy physical work 9. ND 10. other	6, 12, 18, 24, 30, 36
Exercise on way to work before breast cancer	1. not working or working at home 2. I do not walk or bicycle daily 3. less than 15 min daily 4. 15-29 min daily 5. 30-44 min daily 6. 45-59 min daily 7. more than 1 hour daily 9. ND 10. other	0
Exercise on way to work	1. not working or working at home 2. I do not walk or bicycle daily 3. less than 15 min daily 4. 15-29 min daily 5. 30-44 min daily 6. 45-59 min daily 7. more than 1 hour daily 9. ND 10. other	6, 12, 18, 24, 30, 36
Exercise on way to work, other	1. not working or working at home 2. I do not walk or bicycle daily 3. less than 15 min daily 4. 15-29 min daily 5. 30-44 min daily 6. 45-59 min daily 7. more than 1 hour daily 9. ND 10. other	6, 12, 18, 24, 30, 36
CRF VisitsWHO	Number	0, 12, 36
CRF Visits Height	Number	0, 12, 36
CRF Visits Weight	Number	0, 12, 36
CRF Visits Pulse	Number	0, 12, 36
CRF Visits BP systolic	Number	0, 12, 36

CRF VisitsBP diastolic	Number	0, 12, 36
CRF Visits Menopause status changed	Yes, No, Unknown	12, 36
CRF Visits Menopause status unknown reason	Free text	12, 36
CRF Visits Menopause age	Number	12, 36
CRF Visits Menopause status cause	Chemical, Surgical, Natural, Other	12, 36
CRF Visits Menstrual cycle	Amenorrhea >12 months (postmenopausal) Unknown, Amenorrhea 6-12 months, Irregular, Regular (every 3-4 weeks)	12, 36
CRF Visits ET changed	No changes, Changed	12, 36
CRF Visits Fracture region	Free text	0, 12, 36
CRF Visits Diabetes	Yes, No	0, 12, 36
CRF Visits Cardiovascular disease	Yes, No	0, 12, 36
CRF Visits Coronary heart disease	Yes, No	0, 12, 36
CRF Visits Coronary stroke	Yes, No	0, 12, 36
CRF Visits Hypertension	Yes, No	0, 12, 36
CRF Visits Musculoskeletal morbidity	Yes, No	0, 12, 36
CRF Visits Total cholesterol	Number	0, 12, 36
CRF Visits Glucose	Number	0, 12, 36
Fyys akt kyselySeurantakerta kk		
Light exetcise total min	Number	0, 6, 12, 18, 24, 30, 36
Moderately heavy exercise total min	Number	0, 6, 12, 18, 24, 30, 36
Heavy exercise total min	Number	0, 6, 12, 18, 24, 30, 36
Very heavy exercise total min	Number	0, 6, 12, 18, 24, 30, 36
Figure of eight run 1 time	Number	0, 12, 36
Figure of eight run 2 time	Number	0, 12, 36
Figure of eight number of cycles	Number	0, 12, 36
Walking test result	Number	0, 12, 36
Waist circumference	Number	0, 12, 36
C30Strenuous activities	1. Not at all 2. A little 3. Quite a bit 4. Very much 9. ND	0, 3, 6, 12, 18, 24, 30, 36
C30Long walk	1. Not at all 2. A little 3. Quite a bit 4. Very much 9. ND	0, 3, 6, 12, 18, 24, 30, 36
C30Short walk	1. Not at all 2. A little 3. Quite a bit 4. Very much 9. ND	0, 3, 6, 12, 18, 24, 30, 36
C30Rest or sitting	1. Not at all 2. A little 3. Quite a bit 4. Very much 9. ND	0, 3, 6, 12, 18, 24, 30, 36
C30Help with eating dressing washing	1. Not at all 2. A little 3. Quite a bit 4. Very much 9. ND	0, 3, 6, 12, 18, 24, 30, 36
C30Difficulties in daily activities	1. Not at all 2. A little 3. Quite a bit 4. Very much 9. ND	0, 3, 6, 12, 18, 24, 30, 36
C30Difficulties in leisure time activities	1. Not at all 2. A little 3. Quite a bit 4. Very much 9. ND	0, 3, 6, 12, 18, 24, 30, 36
C30Short of breath 1	1. Not at all 2. A little 3. Quite a bit 4. Very much 9. ND	0, 3, 6, 12, 18, 24, 30, 36
C30Pain	1. Not at all 2. A little 3. Quite a bit 4. Very much 9. ND	0, 3, 6, 12, 18, 24, 30, 36
C30Need to rest	1. Not at all 2. A little 3. Quite a bit 4. Very much 9. ND	0, 3, 6, 12, 18, 24, 30, 36

C30Insomnia	1. Not at all 2. A little 3. Quite a bit 4. Very much 9. ND	0, 3, 6, 12, 18, 24, 30, 36
C30Weakness	1. Not at all 2. A little 3. Quite a bit 4. Very much 9. ND	0, 3, 6, 12, 18, 24, 30, 36
C30Appetite loss	1. Not at all 2. A little 3. Quite a bit 4. Very much 9. ND	0, 3, 6, 12, 18, 24, 30, 36
C30Nausea	1. Not at all 2. A little 3. Quite a bit 4. Very much 9. ND	0, 3, 6, 12, 18, 24, 30, 36
C30Vomitting	1. Not at all 2. A little 3. Quite a bit 4. Very much 9. ND	0, 3, 6, 12, 18, 24, 30, 36
C30Constipation	1. Not at all 2. A little 3. Quite a bit 4. Very much 9. ND	0, 3, 6, 12, 18, 24, 30, 36
C30Diarrhea	1. Not at all 2. A little 3. Quite a bit 4. Very much 9. ND	0, 3, 6, 12, 18, 24, 30, 36
C30Fatigue	1. Not at all 2. A little 3. Quite a bit 4. Very much 9. ND	0, 3, 6, 12, 18, 24, 30, 36
C30Distracting pain	1. Not at all 2. A little 3. Quite a bit 4. Very much 9. ND	0, 3, 6, 12, 18, 24, 30, 36
C30Difficulty concentrating	1. Not at all 2. A little 3. Quite a bit 4. Very much 9. ND	0, 3, 6, 12, 18, 24, 30, 36
C30Tense	1. Not at all 2. A little 3. Quite a bit 4. Very much 9. ND	0, 3, 6, 12, 18, 24, 30, 36
C30Worry	1. Not at all 2. A little 3. Quite a bit 4. Very much 9. ND	0, 3, 6, 12, 18, 24, 30, 36
C30Irritable	1. Not at all 2. A little 3. Quite a bit 4. Very much 9. ND	0, 3, 6, 12, 18, 24, 30, 36
C30Depressed	1. Not at all 2. A little 3. Quite a bit 4. Very much 9. ND	0, 3, 6, 12, 18, 24, 30, 36
C30Difficulty remembering	1. Not at all 2. A little 3. Quite a bit 4. Very much 9. ND	0, 3, 6, 12, 18, 24, 30, 36
C30Disturbance in family life	1. Not at all 2. A little 3. Quite a bit 4. Very much 9. ND	0, 3, 6, 12, 18, 24, 30, 36
C30Disturbance in social activities	1. Not at all 2. A little 3. Quite a bit 4. Very much 9. ND	0, 3, 6, 12, 18, 24, 30, 36
C30Financial difficulties	1. Not at all 2. A little 3. Quite a bit 4. Very much 9. ND	0, 3, 6, 12, 18, 24, 30, 36
C30Health status	1. Very poor - 7. Excellent 9. ND	0, 3, 6, 12, 18, 24, 30, 36
C30Quality of life	1. Very poor - 7. Excellent 9. ND	0, 3, 6, 12, 18, 24, 30, 36
BR23Seurantakerta kk	1. Not at all 2. A little 3. Quite a bit 4. Very much 9. ND	0, 3, 6, 12, 18, 24, 30, 36
BR23Dry mouth	1. Not at all 2. A little 3. Quite a bit 4. Very much 9. ND	0, 3, 6, 12, 18, 24, 30, 36
BR23Taste different	1. Not at all 2. A little 3. Quite a bit 4. Very much 9. ND	0, 3, 6, 12, 18, 24, 30, 36
BR23Irritated eyes	1. Not at all 2. A little 3. Quite a bit 4. Very much 9. ND	0, 3, 6, 12, 18, 24, 30, 36
BR23Hair loss	1. Not at all 2. A little 3. Quite a bit 4. Very much 9. ND	0, 3, 6, 12, 18, 24, 30, 36
BR23Upset by hair loss	1. Not at all 2. A little 3. Quite a bit 4. Very much 9. ND	0, 3, 6, 12, 18, 24, 30, 36
BR23Ill or unwell	1. Not at all 2. A little 3. Quite a bit 4. Very much 9. ND	0, 3, 6, 12, 18, 24, 30, 36
BR23Hot flushes	1. Not at all 2. A little 3. Quite a bit 4. Very much 9. ND	0, 3, 6, 12, 18, 24, 30, 36
BR23Headaches	1. Not at all 2. A little 3. Quite a bit 4. Very much 9. ND	0, 3, 6, 12, 18, 24, 30, 36

BR23Physically less attractive	1. Not at all 2. A little 3. Quite a bit 4. Very much 9. ND	0, 3, 6, 12, 18, 24, 30, 36
BR23Less feminine	1. Not at all 2. A little 3. Quite a bit 4. Very much 9. ND	0, 3, 6, 12, 18, 24, 30, 36
BR23Hard to look at yourself naked	1. Not at all 2. A little 3. Quite a bit 4. Very much 9. ND	0, 3, 6, 12, 18, 24, 30, 36
BR23Dissatisfied with your body	1. Not at all 2. A little 3. Quite a bit 4. Very much 9. ND	0, 3, 6, 12, 18, 24, 30, 36
BR23Worried about future health	1. Not at all 2. A little 3. Quite a bit 4. Very much 9. ND	0, 3, 6, 12, 18, 24, 30, 36
BR23Sexual interest	1. Not at all 2. A little 3. Quite a bit 4. Very much 9. ND	0, 3, 6, 12, 18, 24, 30, 36
BR23Sexual activity	1. Not at all 2. A little 3. Quite a bit 4. Very much 9. ND	0, 3, 6, 12, 18, 24, 30, 36
BR23Sexual enjoyment	1. Not at all 2. A little 3. Quite a bit 4. Very much 9. ND	0, 3, 6, 12, 18, 24, 30, 36
BR23Pain in arm shoulder	1. Not at all 2. A little 3. Quite a bit 4. Very much 9. ND	0, 3, 6, 12, 18, 24, 30, 36
BR23Swollen arm or hand	1. Not at all 2. A little 3. Quite a bit 4. Very much 9. ND	0, 3, 6, 12, 18, 24, 30, 36
BR23Difficulty raising arm	1. Not at all 2. A little 3. Quite a bit 4. Very much 9. ND	0, 3, 6, 12, 18, 24, 30, 36
BR23Pain in affected breast	1. Not at all 2. A little 3. Quite a bit 4. Very much 9. ND	0, 3, 6, 12, 18, 24, 30, 36
BR23Swollen affected breast	1. Not at all 2. A little 3. Quite a bit 4. Very much 9. ND	0, 3, 6, 12, 18, 24, 30, 36
BR23Oversensitive affected breast	1. Not at all 2. A little 3. Quite a bit 4. Very much 9. ND	0, 3, 6, 12, 18, 24, 30, 36
BR23Skin problems in affected breast	1. Not at all 2. A little 3. Quite a bit 4. Very much 9. ND	0, 3, 6, 12, 18, 24, 30, 36
WHQ Wake up at night	1. Yes definitely 2. Yes sometimes 3 No not much 4. No not at all 9. ND	0, 3, 6, 12, 18, 24, 30, 36
WHQ Frightened or panic	1. Yes definitely 2. Yes sometimes 3 No not much 4. No not at all 9. ND	0, 3, 6, 12, 18, 24, 30, 36
WHQ Miserable or sad	1. Yes definitely 2. Yes sometimes 3 No not much 4. No not at all 9. ND	0, 3, 6, 12, 18, 24, 30, 36
WHQ Anxious outside home	1. Yes definitely 2. Yes sometimes 3 No not much 4. No not at all 9. ND	0, 3, 6, 12, 18, 24, 30, 36
WHQ Lost interest	1. Yes definitely 2. Yes sometimes 3 No not much 4. No not at all 9. ND	0, 3, 6, 12, 18, 24, 30, 36
WHQ Get palpitations	1. Yes definitely 2. Yes sometimes 3 No not much 4. No not at all 9. ND	0, 3, 6, 12, 18, 24, 30, 36
WHQ Still enjoy the same things	1. Yes definitely 2. Yes sometimes 3 No not much 4. No not at all 9. ND	0, 3, 6, 12, 18, 24, 30, 36
WHQ Life is not worth living	1. Yes definitely 2. Yes sometimes 3 No not much 4. No not at all 9. ND	0, 3, 6, 12, 18, 24, 30, 36
WHQ Tense	1. Yes definitely 2. Yes sometimes 3 No not much 4. No not at all 9. ND	0, 3, 6, 12, 18, 24, 30, 36
WHQ Good appetite	1. Yes definitely 2. Yes sometimes 3 No not much 4. No not at all 9. ND	0, 3, 6, 12, 18, 24, 30, 36
WHQ Restless	1. Yes definitely 2. Yes sometimes 3 No not much 4. No not at all 9. ND	0, 3, 6, 12, 18, 24, 30, 36
WHQ More irritable	1. Yes definitely 2. Yes sometimes 3 No not much 4. No not at all 9. ND	0, 3, 6, 12, 18, 24, 30, 36
WH Worry about growing	1. Yes definitely 2. Yes sometimes 3 No not much 4. No not at all 9. ND	0, 3, 6, 12, 18, 24, 30, 36
WHQ Headaches	1. Yes definitely 2. Yes sometimes 3 No not much 4. No not at all 9. ND	0, 3, 6, 12, 18, 24, 30, 36

WHQ More tired	1. Yes definitely 2. Yes sometimes 3 No not much 4. No not at all 9. ND	0, 3, 6, 12, 18, 24, 30, 36
WHQ Dizzy spells	1. Yes definitely 2. Yes sometimes 3 No not much 4. No not at all 9. ND	0, 3, 6, 12, 18, 24, 30, 36
WHQ Breasts tender or uncomfortable	1. Yes definitely 2. Yes sometimes 3 No not much 4. No not at all 9. ND	0, 3, 6, 12, 18, 24, 30, 36
WHQ Pain in back or limps	1. Yes definitely 2. Yes sometimes 3 No not much 4. No not at all 9. ND	0, 3, 6, 12, 18, 24, 30, 36
WHQ Hot flushes	1. Yes definitely 2. Yes sometimes 3 No not much 4. No not at all 9. ND	0, 3, 6, 12, 18, 24, 30, 36
WHQ More clumsy	1. Yes definitely 2. Yes sometimes 3 No not much 4. No not at all 9. ND	0, 3, 6, 12, 18, 24, 30, 36
WHQ Lively and excitable	1. Yes definitely 2. Yes sometimes 3 No not much 4. No not at all 9. ND	0, 3, 6, 12, 18, 24, 30, 36
WHQ Abdominal cramps or discomfort	1. Yes definitely 2. Yes sometimes 3 No not much 4. No not at all 9. ND	0, 3, 6, 12, 18, 24, 30, 36
WHQ Sick or nauseous	1. Yes definitely 2. Yes sometimes 3 No not much 4. No not at all 9. ND	0, 3, 6, 12, 18, 24, 30, 36
WHQ Lost interest in sex	1. Yes definitely 2. Yes sometimes 3 No not much 4. No not at all 9. ND	0, 3, 6, 12, 18, 24, 30, 36
WHQ Feelings of well-being	1. Yes definitely 2. Yes sometimes 3 No not much 4. No not at all 9. ND	0, 3, 6, 12, 18, 24, 30, 36
WHQ Heavy periods	1. Yes definitely 2. Yes sometimes 3 No not much 4. No not at all 9. ND	0, 3, 6, 12, 18, 24, 30, 36
WHQ Night sweats	1. Yes definitely 2. Yes sometimes 3 No not much 4. No not at all 9. ND	0, 3, 6, 12, 18, 24, 30, 36
WHQ Bloated stomach	1. Yes definitely 2. Yes sometimes 3 No not much 4. No not at all 9. ND	0, 3, 6, 12, 18, 24, 30, 36
WHQ Difficult to sleep	1. Yes definitely 2. Yes sometimes 3 No not much 4. No not at all 9. ND	0, 3, 6, 12, 18, 24, 30, 36
WHQ Feel pins in hands or feet	1. Yes definitely 2. Yes sometimes 3 No not much 4. No not at all 9. ND	0, 3, 6, 12, 18, 24, 30, 36
WHQ Satisfied with my sexual relationship	1. Yes definitely 2. Yes sometimes 3 No not much 4. No not at all 9. ND	0, 3, 6, 12, 18, 24, 30, 36
WHQ Physically attractive	1. Yes definitely 2. Yes sometimes 3 No not much 4. No not at all 9. ND	0, 3, 6, 12, 18, 24, 30, 36
WHQ Difficult to concentrate	1. Yes definitely 2. Yes sometimes 3 No not much 4. No not at all 9. ND	0, 3, 6, 12, 18, 24, 30, 36
WHQ Uncomfortable sex due to vaginal dryness	1. Yes definitely 2. Yes sometimes 3 No not much 4. No not at all 9. ND	0, 3, 6, 12, 18, 24, 30, 36
WHQ More frequent urination	1. Yes definitely 2. Yes sometimes 3 No not much 4. No not at all 9. ND	0, 3, 6, 12, 18, 24, 30, 36
WHQ Poor memory	1. Yes definitely 2. Yes sometimes 3 No not much 4. No not at all 9. ND	0, 3, 6, 12, 18, 24, 30, 36
WHQ Living with some symptoms is difficult	1. Yes 0. No 9. ND	0, 3, 6, 12, 18, 24, 30, 36
FACIT Fatigued	0. Not at all 1. A little bit 2. Somewhat 3. Quite a bit 4. Very much 9. ND	0, 3, 6, 12, 18, 24, 30, 36
FACIT Weak all over	0. Not at all 1. A little bit 2. Somewhat 3. Quite a bit 4. Very much 9. ND	0, 3, 6, 12, 18, 24, 30, 36
FACIT Listless	0. Not at all 1. A little bit 2. Somewhat 3. Quite a bit 4. Very much 9. ND	0, 3, 6, 12, 18, 24, 30, 36
FACIT Tired	0. Not at all 1. A little bit 2. Somewhat 3. Quite a bit 4. Very much 9. ND	0, 3, 6, 12, 18, 24, 30, 36
FACIT Trouble starting things	0. Not at all 1. A little bit 2. Somewhat 3. Quite a bit 4. Very much 9. ND	0, 3, 6, 12, 18, 24, 30, 36
FACIT Trouble finishing things	0. Not at all 1. A little bit 2. Somewhat 3. Quite a bit 4. Very much 9. ND	0, 3, 6, 12, 18, 24, 30, 36

FACIT Energy	0. Not at all 1. A little bit 2. Somewhat 3. Quite a bit 4. Very much 9. ND	0, 3, 6, 12, 18, 24, 30, 36
FACIT Able to do usual activities	0. Not at all 1. A little bit 2. Somewhat 3. Quite a bit 4. Very much 9. ND	0, 3, 6, 12, 18, 24, 30, 36
FACIT Need sleep during day	0. Not at all 1. A little bit 2. Somewhat 3. Quite a bit 4. Very much 9. ND	0, 3, 6, 12, 18, 24, 30, 36
FACIT Too tired to eat	0. Not at all 1. A little bit 2. Somewhat 3. Quite a bit 4. Very much 9. ND	0, 3, 6, 12, 18, 24, 30, 36
FACIT Need help for usual activities	0. Not at all 1. A little bit 2. Somewhat 3. Quite a bit 4. Very much 9. ND	0, 3, 6, 12, 18, 24, 30, 36
FACIT Frustrated about being tired to do things	0. Not at all 1. A little bit 2. Somewhat 3. Quite a bit 4. Very much 9. ND	0, 3, 6, 12, 18, 24, 30, 36
FACIT Limit social activity because tired	0. Not at all 1. A little bit 2. Somewhat 3. Quite a bit 4. Very much 9. ND	0, 3, 6, 12, 18, 24, 30, 36
BDI Mood sadness	1- 5 scale 9. ND	0, 3, 6, 12, 18, 24, 30, 36
BDI Future pessimism	1- 5 scale 9. ND	0, 3, 6, 12, 18, 24, 30, 36
BDI Past failure	1- 5 scale 9. ND	0, 3, 6, 12, 18, 24, 30, 36
BDI Dissatisfaction	1- 5 scale 9. ND	0, 3, 6, 12, 18, 24, 30, 36
BDI How do you like yourself	1- 5 scale 9. ND	0, 3, 6, 12, 18, 24, 30, 36
BDI Disappointment	1- 5 scale 9. ND	0, 3, 6, 12, 18, 24, 30, 36
BDI Suicidal thoughts	1- 5 scale 9. ND	0, 3, 6, 12, 18, 24, 30, 36
BDI Social withdrawal	1- 5 scale 9. ND	0, 3, 6, 12, 18, 24, 30, 36
BDI Indecisiveness	1- 5 scale 9. ND	0, 3, 6, 12, 18, 24, 30, 36
BDI Body image	1- 5 scale 9. ND	0, 3, 6, 12, 18, 24, 30, 36
BDI Changes in sleep	1- 5 scale 9. ND	0, 3, 6, 12, 18, 24, 30, 36
BDI Tiredness or fatigue	1- 5 scale 9. ND	0, 3, 6, 12, 18, 24, 30, 36
BDI Changes in appetite	1- 5 scale 9. ND	0, 3, 6, 12, 18, 24, 30, 36
BDI Anxious or tense	1- 5 scale 9. ND	0, 3, 6, 12, 18, 24, 30, 36
C30 Global QoL	0-100 score A high score represents a high QoL	0, 3, 6, 12, 18, 24, 30, 36
C30 Physical functioning	0-100 score A high score represents a healthy level of functioning	0, 3, 6, 12, 18, 24, 30, 36
C30 Role functioning		
C30 Emotional functioning		
C30 Cognitive functioning		
C30 Social functioning		
C30 Fatigue	0-100 score A high score represents a high level of symptomatology	0, 3, 6, 12, 18, 24, 30, 36
C30 Nausea and vomiting		
C30 Pain		
C30 Dyspnea		
C30 Insomnia		
C30 Appetite loss		

C30 Constipation		
C30 Diarrhea		
C30 Financial impact	0-100 score A high score represents a high level of problem	0, 3, 6, 12, 18, 24, 30, 36
BR23 Body image	0-100 score A high score represents a healthy level of functioning	0, 3, 6, 12, 18, 24, 30, 36
BR23 Sexual functioning		
BR23 Sexual enjoyment		
BR23 Future perspective		
BR23 Systemic therapy side effects	0-100 score A high score represents a high level of symptomatology	0, 3, 6, 12, 18, 24, 30, 36
BR23 Breast symptoms		
BR23 Arm symptoms		
BR23 Upset by hair loss		
WHQ Depressed mood	0-1 score	0, 3, 6, 12, 18, 24, 30, 36
WHQ Somatic symptoms		
WHQ Memory/concentration		
WHQ Vasomotor Symptoms		
WHQ Anxiety/fears		
WHQ Sexual behaviour		
WHQ Sleep Problems		
WHQ Menstrual symptoms		
WHQ Attractiveness		
BDI Depression	0-39 score The higher the score, the higher the depression	0, 3, 6, 12, 18, 24, 30, 36
FACIT score	0-52 score The higher the score, the better the QoL	0, 3, 6, 12, 18, 24, 30, 36
FACIT score prorated for missing items		
BMI	Number	0, 12, 36
Figure 8 time	Number	0, 12, 36
MetH	Number	0, 6, 12, 18, 24, 30, 36
MetHheavy	Number	0, 6, 12, 18, 24, 30, 36

APPENDIX 3 - HUJI Retrospective Data Description & Examples

The demographic and medical data (collected at T1):

Variable name	Meaning
Workshop	Whether participated in the intervention workshop
NotFinish	Whether dropped out of the workshop
Israeli	Whether Israeli-born (vs immigrant)
Age	Age
Married	Whether married
Children	Number of children
city	Whether lives in a city (vs rural area)
EducQue	Education
WorkStat	Employment
RNotWork	Reason for not working
IWork	Income from work
IBit	Income from Social Security disability pension
IOth	Income from another pension

Religious	Level of religious faith
History	Family history of breast cancer
Genetic	Genetic Testing was Performed
Carrier	If a Genetic Test Performed - are you a Carrier
Stage	Cancer stage
Protocol	Treatment protocol (Adria, no Adria, DD)
Treatment	Chemo, radiation, both
Herceptin	Yes/no
Hormonal	Yes/no
TreatEnd	Date of treatment end (not including Herceptin and Hormonal)
OperDate	Operation Date
SecOpeDa	Second Operation Date (if relevant)
Heat	Heat Waves
Mood	Mood Swings
Sleep	Sleep Problems
Fat	Obesity
Body	Decrease in comfort with the body
Sex	Disruption in Sexuality
FemSense	Interference with a sense of femininity
HeatH	How Affected: Heat waves
MoodH	How Affected: Mood swings
SleepH	How Affected: Sleep problems
FatH	How Affected: Obesity
BodyH	How Affected: Decrease in comfort with the body
SexH	How Affected: Disruption in sexuality
FemseneH	How Affected: Interference with a sense of femininity

Psychosocial measures

Variable name	The scale	No of items	Resp. Scale	Collected at
PDS	PTSD - The Posttraumatic Stress Diagnostic Scale (Foa, Cashman, Jaycox, & Perry, 1997)	17	0-3	T1-T5
FUNCT	Functional impairment items from the Diagnostic Predictive Scales (Lucas, Zhang, Fisher et al, 2001)	9	1-5	T1-T5
CESD	CES-D – depression (Radloff, 1977)	20	0-3	T1-T5
THREE	Overall self-report of stress, resilience, and hope "today"	3	1-10	T1-T5
EGO	Ego resilience Scale (Block and Kremen, 1996)	14	1-4	T1-T5
CERQPOS	CERQ – Cognitive Emotion Regulation Questionnaire – Positive coping strategies (Garnefski & Kraaij, 2006)	9	1-5	T1-T6
CERQNEG	CERQ – Cognitive Emotion Regulation Questionnaire – Negative coping strategies (Garnefski & Kraaij, 2006)	9	1-5	T1-T6
FLEX	PACT - The Perceived Ability to Cope with Trauma (Bonanno, Pat-Horenczyk, & Noll, 2011)	20	1-7	T1-T6
PTG	PTGI - The Posttraumatic Growth Inventory (Tedeschi & Calhoun, 1996)	21	0-5	T1-T6
DISTR	K6 – Kessler Psychological Distress Scale (Kessler, Andrews, Colpe, et al, 2002)	6	1-5	T6
PCL	PCL – PTSD Check-List (Weathers, Litz, Herman, Huska & Keane, 1993)	21	1-5	T6

Example Dataset: T1 – T6 Questionnaire answers

Time	1	pds2	0	pds5	3
Number	2	pds3	0	pds6	3
pds1	1	pds4	3	pds7	3

pds8	0	cerq8	5	ptg16	4
pds9	0	cerq9	2	ptg17	5
pds10	0	cerq10	5	ptg18	0
pds11	0	cerq11	5	ptg19	4
pds12	0	cerq12	5	ptg20	0
pds13	0	cerq13	5	ptg21	0
pds14	3	cerq14	3	ego1	4
pds15	3	cerq15	5	ego2	4
pds16	3	cerq16	5	ego3	4
pds17	1	cerq17	3	ego4	3
func1	0	cerq18	3	ego5	4
func2	0	flex1	7	ego6	4
func3	0	flex2	7	ego7	4
func4	0	flex3	7	ego8	4
func5	0	flex4	7	ego9	4
func6	0	flex5	7	ego10	1
func7	0	flex6	6	ego11	4
func8	0	flex7	7	ego12	4
func9	0	flex8	7	ego13	4
cesd1	3	flex9	7	ego14	4
cesd2	0	flex10	7	SressTod	3.70
cesd3	1	flex11	7	ResTod	3.50
cesd4	0	flex12	7	HopeTod	6.10
cesd5	0	flex13	7	distress1	#NULL!
cesd6	0	flex14	7	distress2	#NULL!
cesd7	0	flex15	7	distress3	#NULL!
cesd8	0	flex16	7	distress4	#NULL!
cesd9	0	flex17	7	distress5	#NULL!
cesd10	0	flex18	6	distress6	#NULL!
cesd11	0	flex19	7	pcl1	#NULL!
cesd12	0	flex20	7	pcl2	#NULL!
cesd13	0	ptg1	0	pcl3	#NULL!
cesd14	3	ptg2	4	pcl4	#NULL!
cesd15	2	ptg3	5	pcl5	#NULL!
cesd16	1	ptg4	5	pcl6	#NULL!
cesd17	0	ptg5	5	pcl7	#NULL!
cesd18	2	ptg6	2	pcl8	#NULL!
cesd19	0	ptg7	3	pcl9	#NULL!
cesd20	1	ptg8	5	pcl10	#NULL!
cerq1	5	ptg9	5	pcl11	#NULL!
cerq2	4	ptg10	5	pcl12	#NULL!
cerq3	5	ptg11	5	pcl13	#NULL!
cerq4	4	ptg12	5	pcl14	#NULL!
cerq5	5	ptg13	5	pcl15	#NULL!
cerq6	4	ptg14	0	pcl16	#NULL!
cerq7	5	ptg15	0	pcl17	#NULL!

pcl18	#NULL!	PTG	3.19	CERQPOS	5.00
pcl19	#NULL!	EGO	3.71	CERQNEG	3.50
pcl20	#NULL!	PDS	1.35	TODAY	4.43
CERQ	4.33	FUNCT	0.00	DISTR	#NULL!
FLEX	6.90	CESD	0.65	PCL	#NULL!

Example Dataset: Background Sample

Variable	Description	Type	Width	Demicals
Number	Participant number	Numeric	3	0
WORKSHOP	Participated in the intervention	Numeric	1	0
NotFinish	Did Not Finish the Workshop	Numeric	2	0
H_T2	Has T2 data	Numeric	1	0
H_T3	Has T3 data	Numeric	1	0
H_T4	Has T4 data	Numeric	1	0
H_T5	Has T5 data	Numeric	1	0
H_T6	Has T6 data	Numeric	1	0
Age	Age at Diagnosis	Numeric	2	0
Children	Number of Children	Numeric	2	0
DiagnDate	Date of Diagnosis	Date	11	0
Stage	Disease stage	Numeric	1	0
Protocol	Treatment Protocol Encoding	Numeric	1	0
Treatment	Treatment Type	Numeric	1	0
Herceptin	Herceptin	Numeric	1	0
Hormonal	Hormonal	Numeric	1	0
TreatEnd	End of Treatment (not including Herceptin and Hormonal)	Date	11	0
City	Lives in a city (vs. rural)	Numeric	1	0
Married	Whether married (vs. single)	Numeric	1	0
ISRAELI	Born in Israel (vs. immigrant)	Numeric	1	0
child	Has Children	Numeric	1	0
EducQue	EducQue	Numeric	1	0
WorkStat	Work Status	Numeric	1	0
RNotWork	Reason for not Work	Numeric	2	0
IWork	Income from Work	Numeric	1	0
IBit	Income Frome Social Security Disability	Numeric	1	0
IOth	Income from pension or from another source	Numeric	1	0
Religious	Definition of religious faith	Numeric	1	0
SressTod	Today distress level	Numeric	4	2
ResTod	Level of Perceived Resilience Today	Numeric	4	2
HopeTod	Amount of hope for the future	Numeric	4	2
OperDate	Operation Date	Date	11	0
Genetic	Genetic Testing was Performed	Numeric	1	0
Carrier	If a Genetic Test Performed - are you a Carrier	Numeric	1	0
History	Family history of breast cancer	Numeric	1	0
Heat	Heat Waves	Numeric	1	0

Mood	Mood Swings	Numeric	1	0
Sleep	Sleep Problems	Numeric	1	0
Fat	Obesity	Numeric	1	0
Body	Decrease in comfort with the body	Numeric	1	0
Sex	Disruption in Sexuality	Numeric	1	0
FemSense	Interference with a sense of femininity	Numeric	1	0
HeatH	How Affected Heat Waves	Numeric	1	0
MoodH	How Affected Mood Swings	Numeric	1	0
SleepH	How Affected Sleep Problems	Numeric	1	0
FatH	How Affected Obesity	Numeric	1	0
BodyH	How Affected Decrease in comfort with the body	Numeric	1	0
SexH	How Affected Disruption in Sexuality	Numeric	1	0
FemseneH	How Affected Interference with a sense of femininity	Numeric	1	0

APPENDIX 4 - IEO Retrospective Data Description

	PHR	Database
Age	X	
Height	X	
Weight (BMI)	X	
Education	X	
Socioeconomic status	X	
nulliparity or pregnancy	X	
occupational status	X	
past/current smoking	X	
Frequency and amount of alcohol consumption	X	
Frequency and type of physical activity	not regularly	
TNM stage	X	X
nodal status	X	X
date of first diagnostic sampling	X	X
surgery type and side	X	X
menopausal status	X	X
early age menstruation	X	X
breastfeeding	X	
family history	X	
tumor biology (estrogen, progesterone and HER2 receptor expression, grade and state, vascular invasion, margins)	X	X
ki67	X	X
basic laboratory tests (CBC, Hb, creatinine, bilirubine CRP, ALT)	X	
imaging results (mammography, CT, ultrasound)	X	
genetic risk factors	X	
RMI, mammograph, ecography in BRCA cases	X	
amount of counselling (and support sessions) received during cancer treatment	X	
psychotropic medication	X	

disease free survival	X	X
Type of treatment (chemotherapy/HT/RT)	X	X
Psychological dimensions/measures		
Distress levels (Distress thermometer)	X	X
emotion regulation (Emotion Thermometers)	not to all patients	
life events and stressors	only for patients receiving psy support	
quality of life (EORTC QLQ-C30 or FACT-B)	not to all patients	
impact of cancer-event (IES, impact event scale)	not to all patients	
positive and negative mood (POMS)	not to all patients	
Patient Reported symptoms (IBCSG patient reported symptomps form)	not to all patients	
FACIT Fatigue scale	not to all patients	
F.A.R.E Family Resilience	not to all patients	
HADS (Hospital Anxiety and Depression Scale)	not to all patients	
MoCA	not to all patients (for patients >70)	
instrumental activities of daily living	not to all patients (for patients >70)	
Activity daily living	not to all patients (for patients >70)	
care-giver's reaction assessment instrument	not to all patients (for patients >70)	

APPENDIX 5 - CHAMP Retrospective Data Description

DOB	
Marital status	
Date of diagnosis	
Imaging data	Date
	Type of Imaging
	Tumor Size (cT)
	Lymph node involvement (cN)
	Multifocality and multi-centricity
	Staging results (M)
Genetic risk factors	Family history
	Genetic test
Pathology	pT
	pN

	Histological type
	Grade
	Estrogen receptor
	Progesteron receptor
	HER- 2 receptor (IHC, FISH/SISH)
	Margins
	Multifocality and multi-centricity
	Lymphovascular invasion
	Molecular classification
Surgery	Date
	Type of surgery
	Complications
Radiation therapy	Dates (starting date/ end date)
	Duration
	Irradiation volumes
	Total dose, number of fractions, fraction dose
Complications	Boost
	Complications
Systemic treatment	Type of treatment (chemotherapy, endocrine therapy, biologics)
	Dates (starting date/ end date)
	Patient options
	Complications
	Participation in clinical trials
Follow-up	Disease status (no relapse local and/or regional relapse, distant disease relapse)
	Imaging tests
	Aesthetic outcome
Psychological variables	
When referral to neuropsychiatry or psycho-oncology visit	Hospital Anxiety and Depression Scale
	Distress thermomether
	Mini Mental Status

	Addenbrookes Cognitive Examination Revised (ACE-R)
	Wechsler Adult Intelligence Scale (WAIS III)
	Trail Making Test A and B
	Stroop test
	Beck Depression Inventory (BDI-II)
	State-trait Anxiety Inventory
	EORTC QLC 30