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: 3.2 Initial Semantic Model

Due date of deliverable: (31-10-2018)
Actual submission date: (21-11-2018)

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
Galatia Iatraki	FORTH	iatrak@ics.forth.gr
Evangelos Karademas	FORTH	karademas@uoc.gr
Maria Hatzimina	FORTH	hatzimin@ics.forth.gr

0.2. Documents history

Document version #	Date	Change
V0.1	01 Sept 2018	Starting version, template
V0.2	01 Sept 2018	Definition of ToC
V0.3	10 Oct 2018	First complete draft
V0.4	15 Oct 2018	Integrated version (send to WP members)
V0.5	25 Oct 2018	Updated version (send PCP)
V0.6	25 Oct 2018	Updated version (send to project internal reviewers)
Sign off	7 Nov 2018	Signed off version (for approval to PMT members)
V1.0	19 Nov 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 October 2018

1. Table of Contents

2. Introduction

2.1. About task 3.2

T3.2 focuses on extracting for each domain of the BOUNCE project, a well-defined set of domain concepts that sufficiently describe the semantics of the corresponding data sources. Sources include prospective data, retrospective data and external data sources.

Initially the focus of this task is to establish the methodology for developing the semantic model of the project. Following this methodology, a first version of this semantic model is designed based on relevant approaches able to describe both retrospective and prospective data to be used within the project. The final version of the semantic model will be delivered in D3.3 extending the preliminary model presented in this document.

The final semantic model will be defined in a modular, scalable and extensive way and special attention will be given on the temporal aspect of the information.

2.2. Purpose of the document

The purpose of this document is to report on the existing approaches for modelling psycho-emotional data on cancer and to formulate the semantic model for capturing the prospective and retrospective data to be used within the BOUNCE project. The first version of this semantic model is reported here, whereas the final version will be reported on M24 in the D3.3 Final Semantic Model.

2.3. Work methods & main contents of the document

To develop a concrete semantic model a specific methodology should be followed. In this direction, in Section 3 we shortly review existing methodologies and we present the selected one to be followed. Then, following this methodology, we proceed to the purpose and scope specifications in Section 4, identifying the data to be modelled. Next, we focus on knowledge acquisition on the cancer domain collecting other available ontologies for modelling the cancer domain in Section 5. In Section 6 we present our initial conceptualization and the design of the semantic model and its first implementation. Section 7 concludes this deliverable and presents directions for future work.

3. Methodology and Procedure Specification

Currently, there are several methodologies for developing an ontology. Those methodologies give a set of guidelines about how to carry out the activities identified in the ontology development process and what kind of techniques are most appropriate in each activity.

Several of them have been proposed the last few years, with the most well-known of them including the following:

- The state of the art ontology methodology presented in [1], [38], and [4]
- The methodology by Uschold, Gruniger and King [42], [43], and [44]
- The TOronto Virtual Enterprise (TOVE) methodology[1] [8]
- The Sensus Methodology [2] [11]
- The METHONTOLOGY methodology [5]
- The Kactus methodology[3] [37], [38]
- The Dolce methodology[4] [9]

All these approaches have many steps in common. For example, the steps that are described in two of them, namely MENTHONTOLOGY and the one from Uschold, Gruniger and King are presented in Figure 1. The arrows between them show the equivalent phases between these two methodologies. We can see for example that the "Purpose and scope identifications" is the same with the "Build requirements specification document" etc.

Independent of the specific methodology selected for ontology development, the life cycle of an ontology development process is composed of the following iterative processes:

- **Purpose and Scope Specifications:** The goal of this phase is to determine what is expected from the ontology and to define its scope. This includes the set of terms, its distinct characteristics and its granularity. The intended users and their purposes have to be determined. This purpose can be identified by listing typical queries that the ontology has to answer or by describing usage scenarios. In our case we will use the initial use-case scenarios to identify the types of data that the platform needs to store and access.
- **Knowledge acquisition:** This phase begins by gathering all available knowledge resources describing the domain of the ontology. These resources can be:
    - **Other Ontologies**
    - **Terminologies:** A terminology is a collection of terms. It is a broad expression which may refer to any collection of terms. In that sense any semantic resource is, generally speaking, a terminology.
    - **Controlled vocabularies:** It is a simple collection of terms without any other semantic information.
    - **Coding Systems:** Coding Systems are used when codes, usually code numbers, are applied. This is for example done when a diagnosis is referred to a diagnostic code. Coding systems have all advantages of a controlled vocabulary, which uses

[1] http://www.eil.utoronto.ca/enterprise-modelling/

[2] http://www.isi.edu/~hovy/

[3] http://hcs.science.uva.nl/projects/NewKACTUS/

[4] http://www.loa.istc.cnr.it/DOLCE.html

natural language terms, but they further support interoperability since they do not depend on the use of natural language terms but use semantic free identifiers. In that way, they can be internationally used. However, for a coding system to work, documentation and coding keys must be available in different languages.

- **Taxonomies:** The main feature of any taxonomy is a hierarchical structure which is generated by the subsumption (or so called) is-a relation, i.e. the ordering of classes and subclasses. The original and until today most famous taxonomy is the classification of organisms in the Linnaean taxonomy. The term "taxonomy" in information technology may refer to any classification which has the typical structure of the Linnaean taxonomy. It is supposed that every class has only one superclass on the level directly above it. An ontology provides much more and richer relations and connections than a taxonomy.
- **Thesaurus:** A thesaurus is a terminological tool that includes at least a controlled vocabulary. Additionally, thesauri provide definitions. They point out synonymy but also broader and narrower terms. Furthermore, it is possible to mark a related term which is no equivalent, sub- or super-term but otherwise semantically dependent from a given term. One famous example for a thesaurus in information technologies is the Art & Architecture Thesaurus.

  A Thesaurus is a much stronger terminological tool than a controlled vocabulary or a coding system. It is also more sophisticated than a taxonomy insofar as more than the subclass relation is representable. Erroneously, thesauri are sometimes called ontologies, but ontologies provide a more expressive syntactic and semantic description of terms than thesauri.
- **Messaging Standards:** The main goal of messaging standards is interoperability. The language and data types defined by such standards determine the way in which an information is transferred. Medical messaging standards are created e.g. by HL7. For example, HL7 v2.x provides six different message types with segments and fields which contain specific determined information: If a patient is admitted to a hospital one segment contains fields with data on the identity of the patient and another one the data containing the case etc. Like an ontology, a messaging standard advances interoperability but unlike an ontology it does not provide resources for automated reasoning.
- **Dataset repositories:** A dataset repository is a catalogue of datasets. In a dataset repository, datasets can be identified by a code and named. In that way they are easily accessible. Dataset repositories help to organize data. Their aim is not to represent reality or to produce models like it is done in an ontology. Furthermore, they do not give any semantic explanation of terms.
- **Tools/Algorithms:** tools and algorithms used in the domain might be also a good source of information about the domain.
- **Technical documentations:** Usually the tools and the algorithms in the domain have a technical documentation presenting in detail the aforementioned information. This information is usually unstructured text.

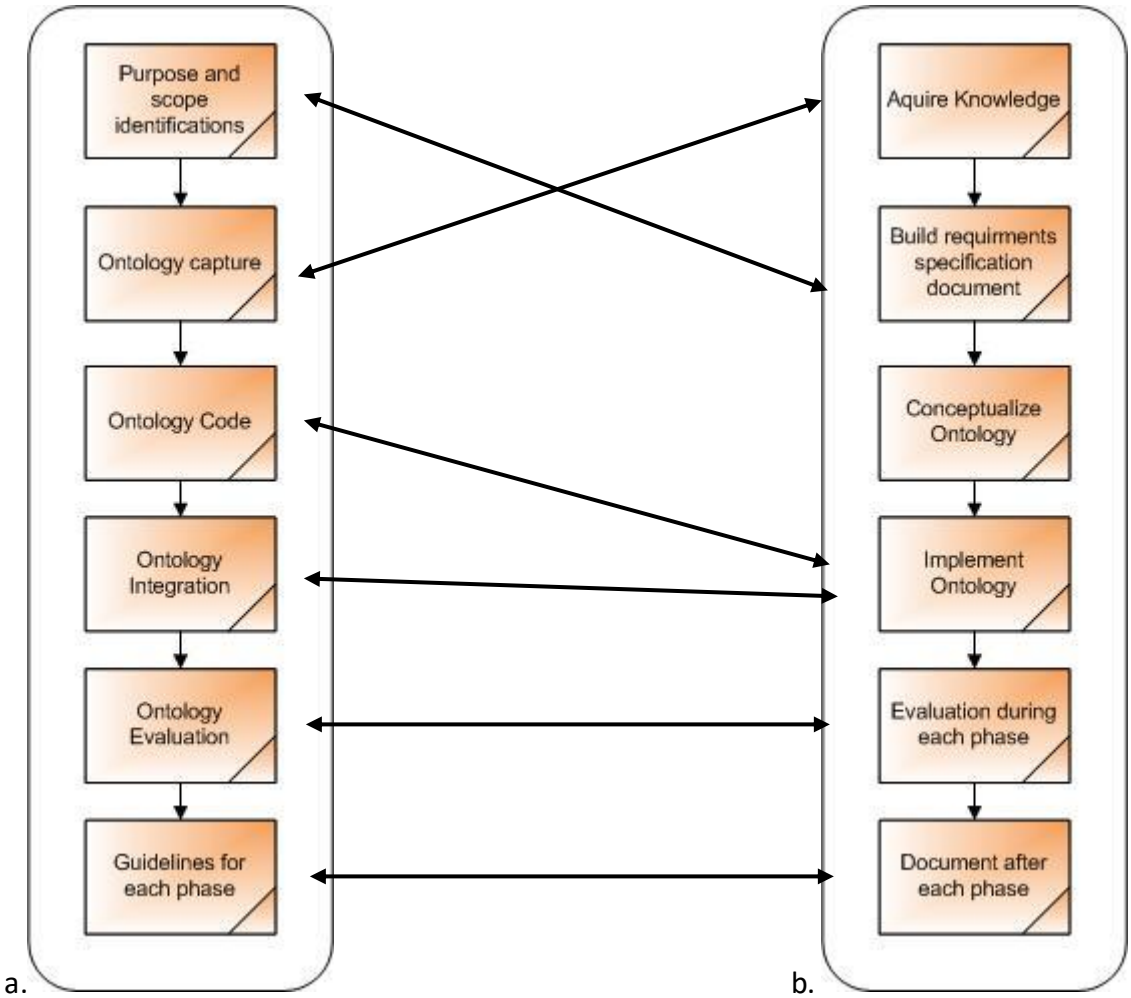


Figure 1 Comparison of two methodologies (a. the approach from Uschold, Gruniger and King and b. Menthontology)

The result of this phase is to identify the most important terms of the domain and define them according to the more consensual definitions

- **Conceptualization:** In this stage, concepts are detected, defined and organized. During this phase, a concept is no longer a term, but a definition. Metadata can be added to those concepts to characterize them.
- **Implementation:** The goal of this phase is to build the formal representation of an ontology. Thus, the ontology engineer has to choose a language to capture the content of the intermediary representation already built. The next stage is to populate the ontology to build the knowledge base.
- **Evaluation:** This phase evaluates the ontology built according to several metrics, including among others the satisfaction of users when testing the ontology, the completeness of the domain representation, or the correctness of the knowledge base and its inference engine.
- **Documentation:** Each choice or problem occurred in the previous phases has to be documented and explained. All the definitions found has to be documented too in order to be precise the source documentation and the authors.

The aforementioned steps will be used for the development of the BOUNCE ontology and the implementation of the first four steps will be reported in the sequel.

4. Purpose, Scope Specifications

The purpose of the BOUNCE semantic model is to effectively represent and model all data that will be collected and analyzed within the BOUNCE platform. This model will be used to integrate, homogenize and semantically uplift the various data available.

The steps to integrate the available data are shown in Figure 2. Initially the ontology is identified or generated based on the data available at the sources. In some cases, it is common to adopt ontologies partially covering the domain and extending them to be able to model the remaining data of interest as well. Then ontologies are used in order to define the mappings, i.e. programmatic correspondences between ontological terms and the various data fields. Based on those mappings, data integration engines can automatically homogenize and semantically uplift available data. The whole workflow is usually an iterative process. When the underlying data is mapped to the sources, the ontology as well as the integrated data might need to be updated.. More details on this process will be provided in D3.4 Solutions for Data Aggregation, Cleaning, Harmonization & Storage.

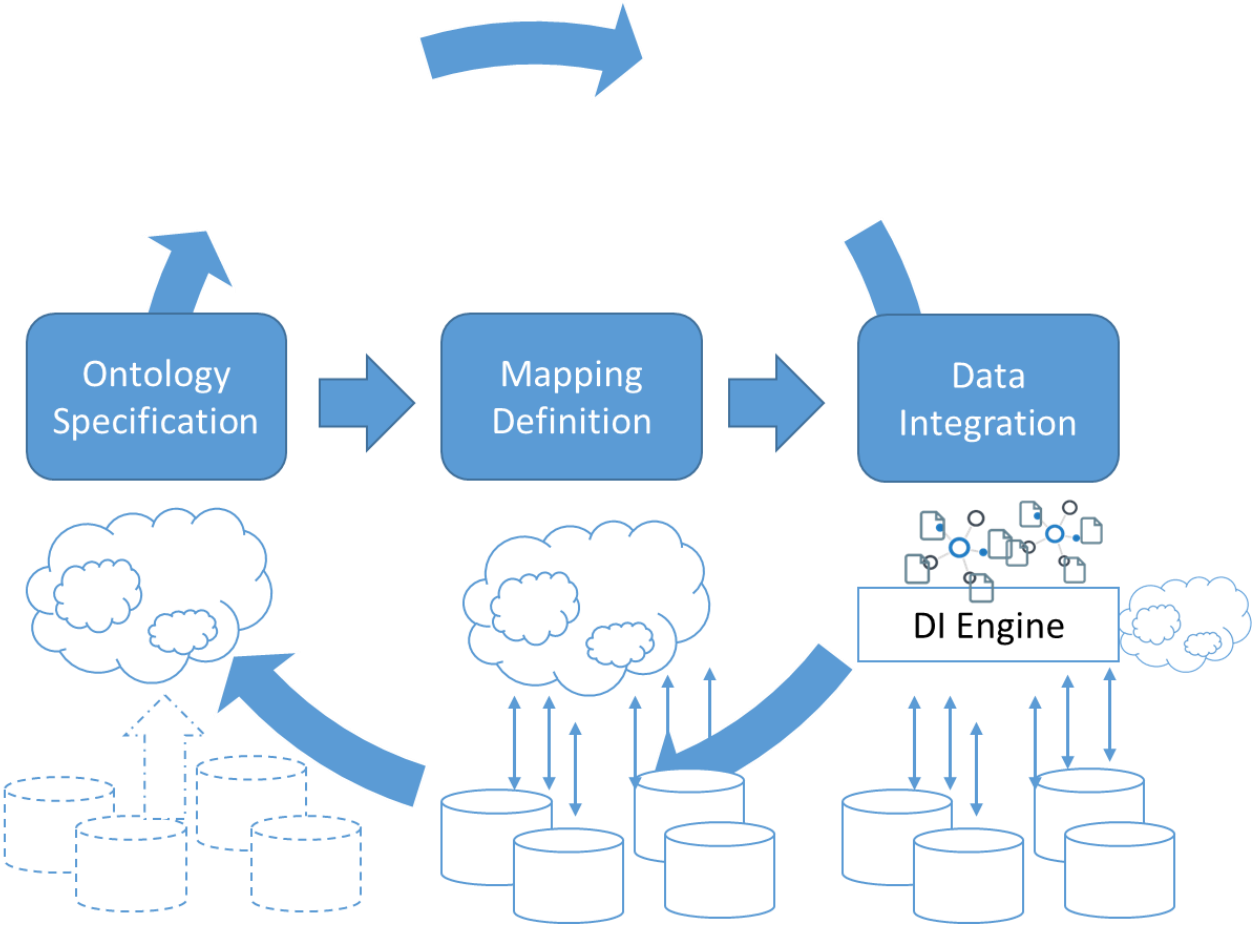


Figure 2. Data Integration & Homogenization through an ontology

In the rest of this chapter we provide a list of the data to be modeled for the retrospective datasets already available, the prospective datasets to be collected during the lifetime of the project and three external datasets identified to be useful that will be also loaded to the BOUNCE central data repository and made available to the project. Then we describe how the ontology will be used for the purposes of the BOUNCE project.

4.1. Retrospective data

In this Section we present the various fields of the datasets, coming from the following four clinical centers:

- European Institute of Oncology (IEO), Milan
- Rabin Medical Center and Shaare Zedek Medical Center – under the coordination of The Hebrew University of Jerusalem (HUJI), Israel
- Helsinki University Hospital (HUS), Finland
- Champalimaud Foundation (CHAMP), Portugal

We have to note that the IEO datasets are not yet available, however we have already a first description of the fields that will be available there.

4.1.1. CHAMP

Table 1 describes the data available from the CHAMP retrospective dataset.

Table 1. The data available from the CHAMP dataset

Data Field	Description	Data Type
SOCIO-DEMOGRAPHIC DATA		
Date of birth	The date of birth.	Time (*dd/mm/yyyy)*
Marital Status	A code that describes the marital status.	Ordinal: 1=married 2=Single 3=Common-law partner 4=Divorced 5=Widow 999= in case of missing data
Education Level (number of years)	A number that indicates years of education.	Ordinal (999=missing data)
DIAGNOSIS DATA		
Date of diagnosis/biopsy	The date of diagnosis.	Time (*dd/mm/yyyy)* (99/99/9999 = Missing data)
Histologic Type	This field describes the histological subtype of the breast cancer. Numbers are used to describe each subtype.	Ordinal: 1=Invasive, NST 2=Invasive, Lobular 3=Mixed, NST and Lobular 4=Histologically special types 5=Ductal carcinoma in situ (DCIS)

		6=Not applicable/Undetermined 999=Missing data
Grade	A description of a tumor based on how abnormal the cancer cells and tissue look under a microscope and how quickly the cancer cells are likely to grow and spread. Low-grade cancer cells look more like normal cells and tend to grow and spread more slowly than high-grade cancer cells. Numbers are used to describe each grade.	Ordinal: 1=Grade 1 2=Grade 2 3=Grade 3 4=Not applicable/Undetermined 999=Missing data
Estrogen Receptor	A protein found inside the cells of the female reproductive tissue, some other types of tissue, and some cancer cells. The hormone estrogen will bind to the receptors inside the cells and may cause the cells to grow. Also called ER.	Ordinal: 1=Negative (Describes cells that do not have a protein to which the hormone estrogen will bind) 2=Positive (Describes cells that have a receptor protein that binds the hormone estrogen.) 3=Not applicable/Undetermined 999=Missing data
Progesterone receptor	A protein found inside the cells of the female reproductive tissue, some other types of tissue, and some cancer cells. The hormone progesterone will bind to the receptors inside the cells and may cause the cells to grow. Also called PR.	Ordinal: 1=Negative (Describes cells that do not have a protein to which the hormone progesterone will bind.) 2=Positive (Describes cells that have a protein to which the hormone progesterone will bind.) 3=Not applicable/Undetermined 999=Missing data
HER- 2 receptor	HER2 (human epidermal growth factor receptor 2) is a gene that can play a role in the development of breast cancer.	Ordinal: 1=Negative 2=Positive 3=Not applicable/Undetermined

		999=Missing data
Ki67	The Ki-67 protein (also known as MKI67) is a cellular marker for proliferation. It is strictly associated with cell proliferation.	Continuous number 999=Missing data
IMAGING DATA		
Date	The date of imaging.	Time (*dd-mm-yyyy*)
Type of imaging	Description of the type of imaging.	Ordinal: 1=Ultrasound + Mammogram 2=Mammogram only 3=Ultrasound only
Tumor Size (cT)	Description of the tumor size.	Ordinal: 1 = TX (Tumor size cannot be assessed) 2=T0 (No tumor can be found) 3=Tis (Carcinoma in situ) 4=T1a (Tumor is larger than 0.1 cm, but no larger than 0.5 cm) 5=T1b (Tumor is larger than 0.5 cm, but no larger than 1 cm) 6=T1c (Tumor is larger than 1 cm, but no larger than 2 cm) 7=T2 (Tumor is larger than 2 cm, but no larger than 5 cm) 8=T3 (Tumor is larger than 5 cm) 9=T4 (Tumor is any size, but has spread beyond the breast tissue to the chest wall and/or skin)
Lymph node involvement (cN)	Before or during surgery to remove an invasive breast cancer, doctor removes one or some of the underarm lymph nodes so they can be examined under a microscope for cancer cells. The	Ordinal: 1=Nx (Regional lymph nodes cannot be assessed)

	presence of cancer cells is known as lymph node involvement.	2=N0 (No regional lymph node metastasis) 3=N1 (Regional lymph node metastasis) 4=N3 (Metastasis in lymph node(s))
Multifocality / Multicentricity	Multifocal breast cancer tends to develop in the same quadrant of the breast. A multicentric tumor describes a situation where there are multiple tumors, occurring in far-separated areas of the breast.	Ordinal (referring to both multifocality and multicentricity): 1=No 2=Yes
Distant metastases (cM)	Refers to cancer that has spread from the original (primary) tumor to distant organs or distant lymph nodes. Also known as distant cancer.	Ordinal: 1=M0 (No distant metastasis) 2=M1 (Distant metastasis (includes metastasis to ipsilateral supraclavicular lymph node(s)) 3=Mx (Presence of distant metastasis cannot be assessed)
GENETIC RISK FACTORS		
Family history	Description of family history of cancer if any.	Ordinal: 1=No known family history of cancer 2=Any family history of breast and/or ovarian cancer 3=Any family history of cancer other than breast and ovarian
Genetic test	BRCA1 and BRCA2 are human genes that produce tumor suppressor proteins. These proteins help repair damaged DNA and, therefore, play a role in ensuring the stability of each cell's genetic material. When either of these genes is mutated, or altered, such that its protein product is not made or does not function correctly, DNA damage may not be repaired properly. As a result, cells are more	Ordinal: 1=Negative test 2=Not available 3=BRCA 1 positive 4=BRCA 2 positive 5=Positive for other tests 6=Positive result of uncertain significance

	likely to develop additional genetic alterations that can lead to cancer.	
PATHOLOGY (Post-surgery)		
pT	pT= primary tumor ** same as IMAGING DATA: Tumor Size (cT) **	
pN	** some values coexist in IMAGING DATA: Lymph node involvement (cN)**	Ordinal: 1=Nx (Regional lymph nodes cannot be assessed) 2=N0 (No regional lymph node metastasis) 3=N1mi 4=N1a 5=N1b 6=N1c 7=N2 8=N3 (Metastasis in lymph node(s))
Histologic Type	** coexist with DIAGNOSIS DATA: Histologic Type **	
Grade	** coexist with DIAGNOSIS DATA: Grade **	
Estrogen receptor	** coexist with DIAGNOSIS DATA: Estrogen receptor **	
Progesteron receptor	** coexist with DIAGNOSIS DATA: Progesteron receptor **	
HER- 2 receptor	** coexist with DIAGNOSIS DATA: HER- 2 receptor **	
Ki67	** coexist with DIAGNOSIS DATA: Ki67 **	
Margins	During or after surgery, a pathologist examines this rim of tissue — called the surgical margin or margin of resection — to be sure it's clear of any cancer cells. If cancer cells are present, this will influence decisions about treatments such as additional surgery and radiation. Margins are checked after surgical biopsy, lumpectomy, and mastectomy.	Ordinal: 1=Free Margins 2=Positive margins with indication for surgery 3=Positive margins with no indication for surgery 4=Not applicable/Undetermined

Lymphovascular invasion	Lymphovascular invasion (LVI or lymphovascular space invasion) is spread of a cancer to the blood vessels and/or lymphatics.	Ordinal: 1=Present 2=Absent 3=Suspected 4=Not applicable/Undetermined
Genomic test	Genomics refers to an organism's entire genetic makeup (DNA), which is called a genome. In cancer patients, genomics addresses all genes and how they're interrelated within the cancer. This can determine more about how the cancer will behave. With genomic testing, the genomic makeup of abnormalities, or mutations, within the cancerous tissue can be identified. This means that genomic testing is used for patients that have been diagnosed with cancer, versus genetic testing which is routinely used as a precaution for someone who has not been diagnosed with cancer.	Ordinal: 1=Not done 2=Luminal low-risk 3=Luminal intermediate or high-risk 4=Not applicable/Undetermined
Molecular classification	**1=Luminal A like** (breast cancer is hormone-receptor positive (estrogen-receptor and/or progesterone-receptor positive), HER2 negative, and has low levels of the protein Ki-67, which helps control how fast cancer cells grow. Luminal A cancers are low-grade, tend to grow slowly and have the best prognosis.) **2=Luminal B like** (breast cancer is hormone-receptor positive (estrogen-receptor and/or progesterone-receptor positive), and either HER2 positive or HER2 negative with high levels of Ki-67. Luminal B cancers generally grow slightly faster than luminal A cancers and their prognosis is slightly worse.) **3=Luminal B, HER 2 enriched** **4=HER 2 enriched** (breast cancer is hormone-receptor negative (estrogen-receptor and progesterone-receptor	Ordinal: 1=Luminal A like 2=Luminal B like 3=Luminal B, HER 2 enriched 4=HER 2 enriched 5=Basal 6=Undetermined

	negative) and HER2 positive. HER2-enriched cancers tend to grow faster than luminal cancers and can have a worse prognosis, but they are often successfully treated with targeted therapies aimed at the HER2 protein, such as Herceptin (chemical name: trastuzumab), Perjeta (chemical name: pertuzumab), Tykerb (chemical name: lapatinib), and Kadcyla (chemical name: T-DM1 or ado-trastuzumab emtansine).) **5=Basal** **6=Undetermined**	
Staging results - AJCC 7th Ed.	The stage of a breast cancer is determined by the cancer's characteristics, such as how large it is and whether or not it has hormone receptors.	Ordinal: 1=0 2=1a 3=1b 4=IIa 5=IIb 6=IIIa 7=IIIb 8=IIIc 9=IV 10=Undetermined
SURGERY		
Date	The date of surgery.	Time (*dd-mm-yyyy*)
Breast Surgery	Type of surgery.	Ordinal: 1=Lumpectomy 2=Mastectomy
Axillary Management	Management of the axilla in breast cancer patients has evolved in the last several decades. With the arrival of the sentinel lymph node biopsy, surgical practice for axillary staging in patients with early breast cancer has become gradually less invasive and formal axillary lymph node dissection has been confined to selected patients.	Ordinal: 1=Sentinel lymph node biopsy (SLNB) 2=Axillary lymph node dissection (ALND) 3=ALND after SLNB
RADIATION THERAPY		

Radiation therapy-type	Types of radiation therapy.	Ordinal: 1=No indication for adjuvant radiotherapy 2=Local therapy (breast) 3=Local-regional therapy (breast + lymph nodes)
Starting date	Radiation therapy starting date.	Time 11/11/1111=Not applicable 99/99/9999=Missing data
End date	Radiation therapy end date.	Time 11/11/1111=Not applicable 99/99/9999=Missing data
Total dose	Number that describes the total dose.	Continuous 1111=Not applicable 999=Missing data
Number of fractions (number of daily sessions)	The number of fractions.	Continuous 1111=Not applicable 999=Missing data
Boost	After the whole breast irradiation treatment sessions are complete, a radiation boost is administered, as a means of preventing a recurrence (the breast cancer coming back).	Continuous 1111=Not applicable 999=Missing data
SYSTEMIC TREATMENT		
Type of Systemic Treatment	Description of systematic treatment.	Ordinal: 1=No indication for systemic treatment 2=Adjuvant/Neoadjuvant Chemotherapy only 3=Adjuvant/Neoadjuvant Chemotherapy plus biologicals 4=Adjuvant/Neoadjuvant Chemotherapy plus biologicals and endocrine therapy (ET) 5=Adjuvant/Neoadjuvant Chemotherapy plus ET 6=ET only 7=Biologicals only
Adjuvant/Neoadjuvant Chemotherapy Start Date	Start date of adjuvant/neoadjuvant chemotherapy	Time 11/11/1111=Not applicable 99/99/9999=Missing data

Adjuvant/Neoadjuvant Chemotherapy End Date	End date of adjuvant/neoadjuvant chemotherapy	Time 11/11/1111=Not applicable 99/99/9999=Missing data
Type of Chemotherapy	Description of the type of chemotherapy.	Ordinal: 1=Anthracyclines and taxanes 2=Taxanes only 3=Anthracyclines only 4=Anthracyclines and taxanes and platinums 5=Not applicable (no indication for Chemotherapy)
Adjuvant/Neoadjuvant Hormone Therapy Start Date	Start date of adjuvant/neoadjuvant hormone therapy.	Time 11/11/1111=Not applicable 99/99/9999=Missing data
Adjuvant/Neoadjuvant Hormone Therapy End Date	End date of adjuvant/neoadjuvant hormone therapy.	Time 11/11/1111=Not applicable 99/99/9999=Missing data
Type of Hormone Therapy	Description of type of hormone therapy.	Ordinal: 1=Tamoxifen 2=Tamoxifen sequential to Aromatase Inhibitors (AI's) 3=AI's 4=Ovarian suppression with aLHRH plus Tamoxifen 5=Ovarian suppression with aLHRH plus AI's 6=Ovarian suppression with aLHRH plus Tamoxifen and AI's 7=Not applicable (no indication for Hormone Therapy)
Biologics ADJ/NEO Start Date	Start date of biologics adj/neo	Time 11/11/1111=Not applicable 99/99/9999=Missing data
Biologics ADJ/NEO End Date	End date of biologics adj/neo	Time 11/11/1111=Not applicable

		99/99/9999=Missing data
Type of Biologicals	Targeted therapy includes drugs that are designed to recognize certain changes in breast cancer cells and to fight the growth and spread of these cells.	Ordinal: 1=Trastuzumab 2=Trastuzumab plus pertuzumab 3=Not applicable (no indication for Biologicals)
Patient options	The option for the patient to choose.	Ordinal: 1=Followed the plan 2=Refused ET 3=Refused CT 4=Refused biologicals
Participation in clinical trials	1=No 2=Yes	Ordinal: 1=No 2=Yes
FOLLOW UP		
Relapse	Relapse patterns.	Ordinal: 1=Without relapse 2=Local/regional relapse 3=Distant relapse 4=Local and distant relapse 999=Missing data
Date of relapse	Date of relapse.	Time 11/11/1111=Not applicable 99/99/9999=Missing data
Current disease status	Description of the current disease status.	Ordinal: 1=Alive and disease free 2=Alive with relapsed disease 3=Dead, not related to relapse 4=Dead, related to relapsed disease 5=Lost to follow up
Date of last follow up	Date of the last follow up.	Time 11/11/1111=Not applicable 99/99/9999=Missing data
PSHYCOLOGICAL VARIABLES		
Distress thermometer (1 - 10)		Continuous 1-10
HADS (0-42)	Cut off value for clinically significant anxious and depressive symptomatology = 11	Continuous

Symbol Search (subtest WAIS-III)	Standardized result (mean value = 10; standard deviation = 3)	Continuous
Digit Span (subtest WAIS-III)	Standardized result (mean value = 10; standard deviation = 3)	Continuous
Trail Making Test A	Results presented in percentile score.	Ordinal
Trail Making Test B	Results presented in percentile score.	Ordinal
Stroop test_Word Task	T score Standardized result (mean value = 50; standard deviation = 10)	Continuous
Stroop test_Color Task	T score Standardized result (mean value = 50; standard deviation = 10)	Continuous
Stroop Test_Color-Word Task	T score Standardized result (mean value = 50; standard deviation = 10)	Continuous
Beck Depression Inventory (BDI-II) (0 - 63)	Standardized result (mean value = 8.8 ; standard deviation = 7.8)	Continuous
STAI_State subscale (20-80)	Standardized result (mean value Mean value (males) = 37.8 ; Standard deviation (males) = 8.9 Mean value (females) = 39.2 ; Standard deviation (females) = 10.2	Continuous
STAI_Trait subscale (20-80)	Standardized result (mean value Mean value (males) = 37.8 ; Standard deviation (males) = 8.9 Mean value (females) = 39.2 ; Standard deviation (females) = 10.2	Continuous
EORTC QLC 30	Results presented in percentage format	Continuous
Mini Mental Status (0 - 30)	Standardized result The normative values, considering the patient's age and schooling, can be found on the table below	Continuous
Addenbrookes Cognitive Examination Revised (ACE-R) (0 - 100)	Standardized result The normative values, considering the patient's age and schooling, can be found on the table below	Continuous

4.1.2. HUJI

Table 2 describes the data available from the HUJI retrospective dataset.

Table 2. Data available from the HUJI dataset.

Data Field	Description	Data Type
DEMOGRAPHIC AND MEDICAL DATA		

Workshop	Whether participated in the intervention workshop	Number
NotFinish	Whether dropped out of the workshop	Numeric 0=Finished the workshop
Israeli	Whether Israeli-born (vs immigrant)	Numeric: 0=No 1=Yes
Age	Age	Scale
Married	Whether married	Numeric: 0=No
Children	Number of children	Numeric
City	Whether lives in a city (vs rural area)	Numeric: 0=Null
EducQue	Education	Numeric
WorkStat	Employment	Numeric: {1, Not Employed}...
RNotWork	Reason for not working	Numeric: {1, Due to the Current Disease}...
IWork	Income from work	Numeric: {0, No}...
IBit	Income from Social Security disability pension	Numeric: {0, No}...
IOth	Income from another pension	Numeric: {0, No}...
Religious	Level of religious faith	Numeric: {1, Religious}...
History	Family history of breast cancer	Numeric: {1, Yes' 2 'No'}...
Genetic	Genetic Testing was Performed	Numeric: {1, Yes' 2 'No'}...
Carrier	If a Genetic Test Performed - are you a Carrier	Numeric: {1, Yes' 2 'No'}...
Stage	Cancer stage	Numeric
Protocol	Treatment protocol (Adria, no Adria, DD)	Numeric
Treatment	Chemo, radiation, both	Numeric
Herceptin	Yes/no	Numeric: {0,No}...
Hormonal	Yes/no	Numeric: {0,No}...
TreatEnd	Date of treatment end (not including Herceptin and Hormonal)	Date
OperDate	Operation Date	Date
SecOpeDa	Second Operation Date (if relevant)	Date
Heat	Heat Waves	Numeric:

		{0, No}...
Mood	Mood Swings	Numeric: {0, No}...
Sleep	Sleep Problems	Numeric: {0, No}...
Fat	Obesity	Numeric: {0, No}...
Body	Decrease in comfort with the body	Numeric: {0, No}...
Sex	Disruption in Sexuality	Numeric: {0, No}...
FemSense	Interference with a sense of femininity	Numeric: {0, No}...
HeatH	How Affected: Heat waves	Numeric: {0, Did not Affect at all}...
MoodH	How Affected: Mood swings	Numeric: {0, Did not Affect at all}...
SleepH	How Affected: Sleep problems	Numeric: {0, Did not Affect at all}...
FatH	How Affected: Obesity	Numeric: {0, Did not Affect at all}...
BodyH	How Affected: Decrease in comfort with the body	Numeric: {0, Did not Affect at all}...
SexH	How Affected: Disruption in sexuality	Numeric: {0, Did not Affect at all}...
FemseneH	How Affected: Interference with a sense of femininity	Numeric: {0, Did not Affect at all}...

Table 3 describes in addition the psychological measures that were used and the corresponding factors that can be calculated based on these measures.

Table 3. The factors/data fields that are available based on the psychosocial measures used for the HUJI retrospective dataset.

PSYCHOSOCIAL MEASURES	MEASURED FACTORS	Data Type
PDS	**PTSD SCALE:** measures posttraumatic stress disorder.	Numeric
FUNCT (Functional impairment items from the Diagnostic Predictive Scales)	1. **Validity** 2. **Precision** 3. **Acceptability**	Numeric
CESD	**Depression:** measure levels of depression.	Numeric
THREE	1. **Stress** 2. **Resilience** 3. **Hope**	Numeric
EGO	**Ego resiliency:** how well people recover from difficult situations and adapt to changes in their environment.	Numeric

CERQPOS / CERQNEG	**1. Self-blame:** referring to thoughts of blaming yourself for what you have experienced **2. Acceptance:** referring to thoughts of resigning to what has happened; **3. Rumination:** referring to thinking all the time about the feelings and thoughts associated with the negative event **4. Positive Refocusing:** which refers to thinking of other, pleasant matters instead of the actual event **5. Refocus on Planning:** or thinking about what steps to take in order to deal with the event **6. Positive Reappraisal:** or thinking of attaching a positive meaning to the event in terms of personal **7. Putting into Perspective:** or thoughts of playing down the seriousness of the event when compared to other events **8. Catastrophizing:** referring to explicitly emphasizing the terror of the **9. Other-blame:** referring to thoughts of putting the blame for what you have experienced on others.	Numeric
FLEX (PACT - The Perceived Ability to Cope with Trauma)	**Perceived Ability to Cope with Trauma:** examines one's perceived capacity of using trauma focus coping strategies.	Numeric
PTGI The Posttraumatic Growth Inventory	**Measures positive psychological changes :** **1. Relating to Others** **2. New Possibilities** **3. Personal Strength** **4. Spiritual Change** **5. Appreciation of Life**	Numeric
DISTR	**Distress**	Numeric
PCL (PCL-5 PTSD Check-List)	**Total Score:** gives a total score of post-traumatic stress disorder.	Numeric

4.1.3. HUS

Table 4 presents the data fields of interest for modelling the psychosocial measures collected from the HUS retrospective dataset.

Table 4. The psychosocial measures used for the HUS retrospective dataset and the corresponding data fields.

PSYCHOSOCIAL MEASURES	MEASURED FACTORS	*DATA TYPE*
Beck depression inventory (BDI)	**Depression Score:** measure levels of depression.	Numeric: 1 , 2, 3, 4, 5, 9
QLQ-C30 EORTC quality of life questionnaire	1. **Global quality of life:** is the general well-being of individuals, outlining negative and positive features of life. 2. **Physical functioning**: is conceptualized as being supported by physical abilities such as walking, reaching, vision, and hearing, as well as by those in the cognitive domain such as spatial orientation, short-term memory, intelligible speech, and alertness. 3. **Role functioning:** assesses a patient's ability to perform daily activities, leisure time activities, and/or work. 4. **Emotional functioning:** refers to the ability to develop and apply self-awareness, self-management and relationship management skills which enable people to understand and manage their own and others' emotions. 5. **Cognitive functioning:** any mental process that involves symbolic operations (i.e. perception, memory, creation of imagery, and thinking). Encompasses awareness and capacity for judgment.	Numeric: 1 , 2, 3, 4, 9

	6. **Social functioning:** defines an individual's interactions with their environment and the ability to fulfill their role within such environments as work, social activities, and relationships with partners and family.	
QLQ-BR23 EORTC quality of life questionnaire breast cancer module	**Functional scales:** 1. **body image:** Body image is how people see themselves when they look in the mirror or when they picture themselves in their mind. 2. **sexual functioning:** This fields records the capability of individuals to experience sexual pleasure and satisfaction when desired. 3. **sexual enjoyment:** This field captures the level of pleasure and satisfaction of sexual experience. 4. **future perspective:** This field captures the individual expectations and hopes for the future. **Symptoms scales:** 1. **systemic therapy side effects:** This field captures the scale of the side-effects of systemic therapy. 2. **breast symptoms:** This fields captures the scale of the symptoms on the breast. 3. **arm symptoms:** This fields captures the scale of the symptoms on the breast. 4. **upset by hair loss:** This fields captures the scale	Numeric: 1, 2, 3, 4, 5, 9

	of the anxiety caused by hair loss.	
Women's Health Questionnaire (WHQ)	1. **Depressed mood** 2. **Somatic symptoms** 3. **Anxiety/fears** 4. **Vasomotor symptoms** 5. **Sleep problems** 6. **Sexual behavior** 7. **Menstrual symptoms** 8. **Memory/concentration** 9. **Attractiveness**	Numeric: 4, 3, 2, 1, 9
FACIT fatigue scores	**Quality of Life:** is the general well-being of individuals, outlining negative and positive features of life.	Numeric: 0 , 1 , 2, 3, 4, 9

4.1.4. IEO

Table 5 presents the data fields of the IEO retrospective dataset.

Table 5. Data fields of the IEO retrospective dataset

Data Field	Description	Data Type
Demographics/ History Questionnaire	1. **Date of Birth:** The birth date of the person. 2. **Education:** The level of education of the individual person 3. **Occupational Status:** The occupational status of the person (self-employed, unemployed, employed) 4. **Child Number**: The number of Childs of a person. 5. **Smoking:** Whether a person is smoking or not. (yes/no) 6. **Alcohol Consumption:** This field indicates whether a person consumes regularly alcohol (yes/no). 7. **CRP:** Serum C-reactive protein level. 8. **PCR:** Real-time polymerase chain reaction level. 9. **Family History of Breast Cancer:** Whether the family has a history of Breast Cancer (yes/no).	1. Date 2. Text 3. Text 4. Numeric 5. YES/NO/PAST 6.YES/NO + TEXT 7. 8. 9.YES/NO 10.YES/NO 11.NUMERIC 12.YES/NO

	10. **Physical activity:** Whether the person performs regularly physical activities (yes/no). 11. **Times of Psychological Counseling:** The number of times a person had psychological counselling. 12. **Psychotropic Medication:** Whether the person receives psychotropic medication (yes/no).	
QLQ-C30 EORTC quality of life questionnaire	1. **Global quality of life:** is the general well-being of individuals, outlining negative and positive features of life. 2. **Physical functioning**: is conceptualized as being supported by physical abilities such as walking, reaching, vision, and hearing, as well as by those in the cognitive domain such as spatial orientation, short-term memory, intelligible speech, and alertness. 3. **Role functioning:** assesses a patient's ability to perform daily activities, leisure time activities, and/or work. 4. **Emotional functioning:** refers to the ability to develop and apply self-awareness, self-management and relationship management skills which enable people to understand and manage their own and others' emotions. 5. **Cognitive functioning:** any mental process that involves symbolic operations (i.e. perception, memory, creation of imagery, and thinking). Encompasses awareness and capacity for judgment. 6. **Social functioning:** defines an individual's interactions with their environment and the ability to fulfill their role within such environments as work, social activities, and relationships with partners and family.	**Scale**: 1-4

QLQ-BR23 EORTC quality of life questionnaire breast cancer module	**Functional scales:** 1. **body image:** Body image is how people see themselves when they look in the mirror or when they picture themselves in their mind. 2. **sexual functioning:** This fields records the capability of individuals to experience sexual pleasure and satisfaction when desired. 3. **sexual enjoyment:** This field captures the level of pleasure and satisfaction of sexual experience. 4. **future perspective:** This field captures the individual expectations and hopes for the future. **Symptoms scales:** 1. **systemic therapy side effects:** This field captures the scale of the side-effects of systemic therapy. 2. **breast symptoms:** This fields captures the scale of the symptoms on the breast. 3. **arm symptoms:** This fields captures the scale of the symptoms on the breast. 7. **upset by hair loss:** This fields captures the scale of the anxiety caused by hair loss.	**Scale**: 1-4
Family Resilience Scale (F.A.R.E)	Patients' and the caregivers' resilience measured by four factors: 1. Communication and Cohesion 2. Perceived Social Support 3. Perceived Family Coping 4. Religiousness and Spirituality	Scale: 1-7
Functional Assessment of Cancer Therapy - Breast Cancer (FACT-B)	**Quality of Life:** is the general well-being of individuals, outlining negative and positive features of life. **Subscales:** 1. Physical well-being 2. Social/family well-being 3. Emotional well-being 4. Functional well-being 5. Additional concerns	Scale: 0-4

IES impact of event scale	Subjective distress caused by traumatic events. The tool assesses intrusive thinking, behavioral avoidance of traumatic event and hyper arousal symptoms.	Scale: 0-4
FACIT Fatigue scale	**Quality of Life:** is the general well-being of individuals, outlining negative and positive features of life.	Numeric: 0 , 1 , 2, 3, 4, 9
Distress Thermometer	The tool measures the level of **psychological distress**	Scale: 0-10
POMS	It evaluates **psychological mood adjustment to illness.** **Subscales:** 1. Anger 2. Confusion 3. Depression 4. Fatigue 5. Tension **6.** Vigour	Scale: 0-4
RSA – Resilience scale for adults	Measures of **Individual resilience.** **Subscales:** **Perception of self**: It measures confidence in their own abilities and judgments, self-efficacy and realistic expectations; **Planned future:** It defines as the ability to plan ahead, have a positive outlook, and be goal oriented; **Social competence**: It measures levels of social warmth and flexibility, ability to establish friendships, and the positive use of humor; **Structured style:** It measures the preference of having and following routines, being organized and the preference of clear goals and strategies; **Family cohesion**: measures whether values are shared or discordant in the family and if family members enjoy spending time with each other; have an optimistic view of the future; have loyalty toward each other, and have the feeling of mutual appreciation and support; **Social resources**: measure availability of social support, if they have a confidante	Scale: 1-7

	outside the family, and if they may turn to someone outside the family for help.	
FACT Cog	**The questionnaire assess perceived cognitive function and impact on quality of life in cancer patients.** **Subscales:** **1. perceived cognitive impairments** (impairments) **2. perceived cognitive abilities** (abilities) **3. comments from others** (noticeability) **4. impact on quality of life** (quality of life)	Scale: 0-4
ILLNESS PERCEPTION QUESTIONNAIRE (IPQ-R)	The measure provides a **quantitative measurement of the components of illness representations.** It is divided into three sections: identity subscale, causal subscale and a third section which contains 7 subscales: **a. consequences** **b. timeline** **c. acute/chronic** **d. cyclical** **e. personal and treatment control/cure** **f. illness coherence** **g. emotional representations**	Scale: 1, 2, 3, 4, 5
mini MAC	**Evaluates cancer patients' responses during their mental adjustment to diagnosis and treatment.** **Five dimensions:** 1. Helplessness–Hopelessness 2. Anxious Preoccupation 3. Fighting Spirit 4. Cognitive Avoidance **5.** Fatalism	Scale: 1-4
MOS social support	A measure of functional social support. It has two subscales covering two domains: emotional and instrumental [tangible]social support	Scale: 1-5
Skindex	**The measure assesses comprehensively the effects of skin disease on Quality of life.**	Scale: 0-7

STAI	**A measure of trait and state anxiety.** It can be used in clinical settings to diagnose anxiety disease and it is also used in research as an indicator of caregiver distress.	Scale: 1-4

4.2. Prospective data

Table 6 and 7 presents the data fields to be collected during the prospective study. Table 7 more specifically presents the measures to be collected based on psychosocial questionnaires.

Table 6. Data fields to be collected from the prospective study

Data Field	Description	Data Type
SOCIO-DEMOGRAPHIC AND LIFESTYLE		
Age	Age	
Highest level of education	Described the highest level of education attained.	Select from : • Primary school • Secondary school • High school • Vocational non-academic diploma • Bachelor degree • Postgraduate education
Marital status	Description of the marital status (i.e. single, widowed).	Select from: • Single • Engaged • Common-law partner • Married • Separated/divorced • Widowed
Number of children	Number of children	Numeric
Employment status and sick days	Patients have to choose one of the following work status descriptions: • Employed full time • Employed part time • Self-employed • Unemployed • Retired • Housewife	Select from: • Employed full time • Employed part time • Self-employed • Unemployed • Retired • Housewife

	Patients answer, using free text, for the workdays they missed because of treatment/illness.	
Flexible arrangements at work	Indicate if there are possible flexible arrangements at work.	Text
Return to work	Indicate if return to work was possible and in what kind of arrangement.	• Yes, full-time. Please specify return date: • Yes, part time. Please specify return date: • No
Income	Income (from work, pension etc.)	Select number: • 0-500 • 501-1,000 • 1,001-1,500 • 1,501-2,000 • 2,001-2,500 • 2,501-3,000 • 3,001-3,500 • 3,501-4,000 • 4,001-4,500 • 4,501 and up
Faith	Description of the level of religious faith.	Select from available descriptions: ALL EXCEPT HUJI: Atheist Practicing believer Non-practicing believer HUJI: secular observing traditions religious ultra-religious/very-religious
Smoking and alcohol consumption	Patients have to answers several questions about the kind of alcohol they consume and the amount.	Smoking: • I only smoked in the past • Yes/No • If you smoke now or in the past: How many cigarettes do/did you smoke during the day? Alcohol

		• Never • Less than once a month • Once or twice a month • About once a week • Several times a week • Every day
Drug use	Indicate if there is drug use	• No • Not medically-prescribed drugs (such as tranquilizers, Ritalin or strong pain-killers) • Medically-prescribed cannabis • Not medically-prescribed cannabis Other drugs (such as MDMA or cocaine)
Weight	Weight	Numeric (g)
Height	Height	Numeric (cm)
Diet	Yes/No	Select from list / Text
Exercise	Yes/No	Text
Number of support sessions	Number of sessions	Numeric
Family reduced work/activities/services/domestic help	Indicate if there is a family reduced at work/activities/services/domestic help.	No/Yes. Please specify
CLINICAL VARIABLES		
Date of diagnosis	Diagnosis date	Date
Cancer stage	The stage of cancer	I, II, III
Chronic illnesses	Whether other chronic illness have occurred	yes/no specification
Genetic Risk factors	Genetic factors that impose cancer risk.	family history positive genetic testing
Menopausal status pretreatment	Status of menopausal pretreatment	premenopausal perimenopausal postmenopausal
Menopausal status posttreatment	Status of menopausal postreatment	premenopausal perimenopausal

		postmenopausal
pT	Primary tumor	mm
pN	The **N category** (N0, N1, N2, or N3) indicates whether the cancer has spread to lymph nodes near the breast and, if so, how many lymph nodes are affected	N1,N2,N3
histological type	This field describes the histological subtype of the breast cancer.	ductal, lobular, other
ER	Estrogen Receptor	%
PR	Progesteron Receptor	%
Grade	Tumour Grade	I,II,III
HER2	HER2 (human epidermal growth factor receptor 2) is a gene that can play a role in the development of breast cancer.	positive (FISH, SISH/CISH) negative
molecular classification	The tumour molecular classification.	
Performance status ECOG	It describes a patient's level of functioning in terms of their ability to care for themselves.	1-5
Psychotropic medications	The psychotropic medications patient received.	Medications name
Hormone replacement therapy pretreatment	If a pretreatment with hormone replacement has been administered	Yes/no
Surgery Date	The date of the surgery	Date
Surgery Type	The type of the surgery	breast conserving, mastectomy
axillary surgery	Axillary dissection is a surgical procedure that incises the axilla to identify, examine, or remove lymph nodes	SNB, evacuation
reoperation date	Date of reoperation	Date
Radiotherapy	Details about the radiotherapy received	start date - end date faraction per day total dose intraoperative radiotherapy Area (breast, ablation area, nodal irradiation)
Chemotherapy	Details about the chemotherapy received	Start date - end date Regimen (anthracyclin-docetaxel based

		anthracyclin -paclitaxel based paclitaxel docetaxel anthracyclin taxane-karboplatin FINXX type with capecitabine cyclophosphamide-docetaxel other)
Endocrine	Details about the endocrine therapy received	Start date and type (tamoxifen, letrotzole, exemestane, anastrotzole, ovarian suppression + tamoxifen, ovarian suppression + AI, oophorectomy)
Anti HER2 treatment	Details about the anti HER2 treatment received	Start date – end date and type (trastutzumab, trastutzumab + pertutzumab)
side effects	Describes the side effects of treatment	Osteoporosis (date of dg), cardiac failure (type asymptomatic decreaser of LVEF, heart failure, coronary sdr), neutropenic infection (date of dg), other sever side effect (type, date of dg)
Patient care path data		
Oncology Clinic	Number and dates of consultations of: • Oncologist • Nurses • Psychiatrists • Other HC professionals	Numbers Dates
	Number and dates of phone consultations of: • Oncologist • Nurses • Psychiatrists • Other HC professionals	Numbers Dates

	Number and dates of treatment visits: • Chemotherapy visits • Radiation therapy visits	Numbers Dates
	Number and dates of inpatient days: Diagnosis/ Reason for stay	Numbers Dates
Other specialized care unit	Number and dates of consultations of: • Oncologist • Nurses • Psychiatrists • Other HC professionals	Numbers Dates
	Number and dates of phone consultations of: • Oncologist • Nurses • Psychiatrists Other HC professionals	Numbers Dates
	Number and dates of treatment visits: • Chemotherapy visits • Radiation therapy visits	Numbers Dates
	Number and dates of inpatient days: Diagnosis/ Reason for stay	Numbers Dates
Primary care/Occupational HC/other	Number and dates of consultations of: • Oncologist • Nurses • Psychologists • Other therapists	Numbers Dates
	Number and dates of phone consultations of: • Oncologist • Nurses • Psychologists • Other therapists	Numbers Dates
	Number and dates of inpatient days: Diagnosis/ Reason for stay	Numbers Dates
Emergency care	Number of visits: Diagnosis/ Reason for visit	Number
Laboratory visits	Number of visits: Test type	Number
Imaging visits	Number of visits: Imaging type	Number
Outpatient medication	List of prescribed medication	Text (list)

Laboratory tests		
Hb	Hemoglobin	grams per deciliter
Leukocytes	A type of blood cell that is made in the bone marrow and found in the blood and lymph tissue.	mg/L
thrombocytes	Platelet count	mg/L
Neutrofiles	Neutrophils, the most numerous and important type of leukocytes in the body's reaction to inflammation, constitute a primary defense against microbial invasion through the process of phagocytosis	mg/L
CRP	C-reactive protein	mg/L

Table 7. The fields that will be available within the prospective dataset

PSYCHOSOCIAL MEASURES	MEASURED FACTORS	Data Type
TIPI Ten Item Personality Measure (brief "Big Five")	1. **Extraversion**: Extraversion is the state of primarily obtaining gratification from outside oneself. People with high levels of extraversion tend to feel more comfortable in social situations. 2. **Neuroticism:** is a long-term tendency to be in a negative or anxious emotional state. It is not a medical condition but a personality trait. 3. **Conscientiousness:** is about how a person controls, regulates, and directs their impulses. 4. **Agreeableness:** measures a person's tendency to be kind, empathetic, trusting, cooperative, and sympathetic. It shows how well she/he harmonizes with society.	Numeric: Select from 1-10

	5. **Openness (to new experience):** A person with a high level of openness to experience in a personality test enjoys trying new things. Individuals who are low in openness to experience would rather not try new things.	
LOT-R Optimism/Pessimism	**Optimism:** refers to an emotional and psychological perspective on life. It is a positive frame of mind and means that a person takes the view of expecting the best outcome from any given situation.	Numeric: Select from 1-10
SOC-13Sense of Coherence	1. **Comprehensibility:** the cognitive dimension, refers to the extent to which one perceives internal and external stimuli as rationally understandable, and as information that is orderly, coherent, clear, structured rather than noise—that is, chaotic, disordered, random, unexpected, and unexplained. 2. **Manageability:** the instrumental or behavioral dimension, defined as the degree to which one feels that there are resources at one's disposal that can be used to meet the requirements of the stimuli one is bombarded by. 3. **Meaningfulness:** the motivational dimension, refers to the extent to which one feels that life has an emotional meaning, that at least some of the problems faced in life a face are worth commitment and dedication, and are seen as challenges rather than only as burdens.	Numeric: Select from 1-10

PCL-5 PTSD Check-List	**Total Score:** gives a total score of post-traumatic stress disorder.	Numeric: Select from 1-20
Recent illness-related events	**Qualitative question for interim measurements**	Text: Write yes or no
Recent negative life events	**Qualitative question for interim measurements**	Text: Write yes or no
PACT The Perceived Ability to Cope With Trauma (Flexibility in coping)	1. **Perceived ability to focus on processing the trauma (trauma focus):** examines not the usage of a given coping strategy, but one's perceived capacity of using trauma focus coping strategies. 2. **Perceived ability to focus on moving beyond the trauma (forward focus):** examines not the usage of a given coping strategy, but one's perceived capacity of using forward focus coping strategies. 3. **Single flexibility score that represented the ability to use both types of coping**	Numeric: Select from 1-20
CERQ Cognitive Emotion Regulation Questionnaire	**1. Self-blame:** referring to thoughts of blaming yourself for what you have experienced **2. Acceptance:** referring to thoughts of resigning to what has happened; **3. Rumination:** referring to thinking all the time about the feelings and thoughts associated with the negative event **4. Positive Refocusing:** which refers to thinking of other, pleasant matters instead of the actual event **5. Refocus on Planning:** or thinking about what steps to take in order to deal with the event **6. Positive Reappraisal:** or thinking of attaching a positive meaning to the event in terms of personal **7. Putting into Perspective:** or thoughts of playing down the	Text: Write yes or no

	seriousness of the event when compared to other events **8. Catastrophizing:** referring to explicitly emphasizing the terror of the **9. Other-blame:** referring to thoughts of putting the blame for what you have experienced on others.	
MAAS – Mindful Attention Awareness Scale	1. **Single score of dispositional mindfulness:** open or receptive awareness of and attention to what is taking place in the present.	Numeric: Select from 1-15
Spirituality coping - a visual bar	1. **A single item with a single score of spirituality coping**	Numeric
mMOS-SS modified Medical Outcomes Study Social Support Survey	1. **Instrumental social support:** measures assistance received from others that is tangible. 2. **Emotional social support:** measures support from others that makes us feel loved.	Text
F.A.R.E. Family Resilience Questionnaire	**1. Communication and cohesion**: corresponds to the ways in which family members inform each other about things that need to be done and to the ways in which family members show MINeach other love and support. **2. Perceived family coping:** Coping is a conscious intentional response to stress. Coping is often invoked to represent competence and resilience.	Numeric: Select from 1-12
Instrumental/emotional perceived social support	1. **Single item with a single score of perceived emotional support**	Numeric
CDRISCConnor-Davidson Resilience Scale	1. **Overall, single score of resilience**	Numeric: Select from 1-10
How much are you back to yourself?	1. **Single item with a single score of resilience:** description, in percentage. To what extent did you	Percentage

	bounce back to your ordinary life (before illness).	
IPQ Illness Perception Questionnaire	1. **Timeline:** the perceived duration of the illness 2. **Timeline-cyclical:** beliefs about the predictability or cyclic nature of illness. 3. **Personal control:** the extent to which an individual has control over illness. 4. **Treatment control:** beliefs about treatment effectiveness. 5. **Illness coherence:** extent to which an individual has a clear understanding of illness. 6. **Consequences:** the expected effects of the illness. 7. **Emotional representations:** the emotional reactions to illness. 8. **Biological:** biological factors that heighten the odds of illness or impede recovery. 9. **Psychological/stress:** psychological factors that heighten the odds of illness or impede recovery. 10. **Environmental:** environmental factors that heighten the odds of illness or impede recovery. 11. **Health behaviors:** actions that heighten the odds of illness or impede recovery.	Numeric
B-IPQ Illness Perception Questionnaire - Brief form	1. **Personal control:** a high personal control score means that the participant perceives having good control of the illness. 2. **Treatment control:** a high treatment control score means that the participant believes the treatment is	Text

	extremely helpful in managing the illness.	
mini-MAC mini-Mental Adjustment to Cancer	1. **Helplessness/hopelessness:** state in which a person feels an irreparable loss, the threat of death, and a lack of over the situation. 2. **Anxious preoccupation:** where the disease presents itself as a threat but where there is some doubt as to the possibility of exercising some control over the situation and its implications. 3. **Fighting spirit:** where the disease is perceived as a challenge and where the patient believes he or she can exert some control over the situation. 4. **Cognitive avoidance:** characterized by minimization of the threat and downplaying the need for personal control. 5. **Fatalism:** characterized by an attitude of passive acceptance of the disease, which the patient considers impossible to control.	Numeric: Select from 1-29
Single item: what has done to cope	1. **Reappraisal.** 2. **Social support.** 3. **Relaxation.** 4. **Distraction.** 5. **Spiritual coping.** 6. **Exercise.** 7. **Emotion expression.**	Numeric: Select from 1-11
CBI-B Cancer Behavior Inventory	**Single overall score of coping self-efficacy**	Numeric: Select from 1-12
A general self-efficacy item	**Single overall score of coping self-efficacy**	
MOS Adherence to medical advice scale	**Single item and a single score of adherence to medical advice**	Text
PTGI The Posttraumatic Growth Inventory - short form	**Measures positive psychological changes :** 1. **Relating to Others**	Numeric: Select from 1-10

	2. **New Possibilities** 3. **Personal Strength** 4. **Spiritual Change** 5. **Appreciation of Life**	
QLQ-C30 EORTC quality of life questionnaire	1. **Global quality of life:** is the general well-being of individuals, outlining negative and positive features of life. 2. **Physical functioning**: is conceptualized as being supported by physical abilities such as walking, reaching, vision, and hearing, as well as by those in the cognitive domain such as spatial orientation, short-term memory, intelligible speech, and alertness. 3. **Role functioning:** assesses a patient's ability to perform daily activities, leisure time activities, and/or work. 4. **Emotional functioning:** refers to the ability to develop and apply self-awareness, self-management and relationship management skills which enable people to understand and manage their own and others' emotions. 5. **Cognitive functioning:** any mental process that involves symbolic operations (i.e. perception, memory, creation of imagery, and thinking). Encompasses awareness and capacity for judgment. 6. **Social functioning:** defines an individual's interactions with their environment and the ability to fulfill their role	Yes/No

	within such environments as work, social activities, and relationships with partners and family.	
QLQ-BR23 EORTC quality of life questionnaire breast cancer module	**Functional scales:** 1. **body image** 2. **sexual functioning** 3. **sexual enjoyment** 4. **future perspective** **Symptoms scales:** 1. **systemic therapy side effects** 2. **breast symptoms** 3. **arm symptoms** 4. **upset by hair loss**	Yes/No
FCRI-SF Fear of Recurrence - short form (severity scale of original FCRI)	**Severity of fear or recurrence.**	Text
HADS Hospital Anxiety and Depression Scale	1. **Anxiety levels:** measure levels of anxiety. 2. **Depression:** measure levels of depression.	Numeric: Select from 1-14
DT NCCN Distress Thermometer	**Single item with a single score of distress.**	Numeric: Select from 1-10
PANASPositive and Negative affectivity - short form	1. **Positive mood:** measure positive feelings. 2. **Negative mood:** measure negative feelings.	Text

4.3. External Datasets

4.3.1. Breast Cancer Dataset

The Breast dataset[5] is a comprehensive dataset that contains nearly all the PLCO study data available for breast cancer incidence and mortality analyses. For many women the trial documents multiple breast cancers, however, this file only has data on the earliest breast cancer diagnosed in the trial. The dataset contains one record for each of the approximately 78,000 women in the PLCO trial. Table 8 presents the various fields of the dataset.

Table 8. Data fields available for the Breast Cancer dataset

Data Field	Description	
	IDENTIFIERS	

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

PLCO ID	PLCO ID	Char
Build	Masterfile build, used to identify the version of the database.	Char,30
TRIAL ENTRY		
Age At Randomization	Age at trial entry, computed from date of birth and randomization date.	Numeric
Age At Randomization	Categorical version of age, created from the derived age variable.	Numeric
Randomization Arm	Randomization group or arm. The intervention (screening) group or the control (usual-care) group.	• 1="Intervention" • 2="Control"
Study Center	The study center at which the participant was randomized.	• 1="University of Colorado" • 2="Georgetown University" • 3="Pacific Health Research and Education Institute (Honolulu)" • 4="Henry Ford Health System" • 5="University of Minnesota" • 6="Washington University in St Louis" • 8="University of Pittsburgh" • 9="University of Utah" • 10="Marshfield Clinic Research Foundation" • 11="University of Alabama at Birmingham"
Year Of Randomization	Calendar year of trial entry, at which point the participant was randomized into an arm.	Numeric
Sex	Gender of the participant.	2="Female"
Personal History of Any Cancer Prior to Trial Entry	Was the participant diagnosed with any cancer prior to trial	• 0="No" • 1="Yes" • 9="Unknown - BQ History Unknown and No Cancer

	entry (randomization)?	History From Another Source"
Personal History of Breast Cancer Prior to Trial Entry	Was the participant diagnosed with breast cancer prior to trial entry (randomization)?	• 0="No" • 1="Yes" • 9="Unknown - BQ History Unknown and No Cancer History From Another Source"
EXIT		
Breast Incidence Exit Age	Age of participant at exit for breast cancer incidence. This is age at diagnosis for participants with breast cancer and age at trial exit otherwise.	Numeric
Days Until Breast Incidence Exit	Days from trial entry (randomization) to cancer diagnosis for participants with breast cancer, or to trial exit otherwise.	Numeric
Breast Incidence Exit Status	Status of the participant at exit for breast cancer incidence.	• -1="Cancer before randomization" • 1="Confirmed cancer" • 3="Confirmed in situ carcinoma/LMP/borderline cancer" • 5="Last Negative ASU prior to cancer report" • 8="Death" • 9="Participant Withdrawal or Lost Contact" • 13="Cutoff for screening center data collection" • 14="13 year cutoff" • 17="Last ASU"
First Cancer Incidence Exit Age	Age of participant at exit for the first cancer incidence. This is age at diagnosis for participants with confirmed cancer and	Numeric

	age at trial exit otherwise.	
Days Until First Cancer Incidence Exit	Days from trial entry (randomization) to first cancer diagnosis for participants with cancer, or to trial exit otherwise.	Numeric
First Cancer Incidence Exit Status	Status of the participant at exit for first cancer incidence.	• 1="Confirmed cancer" • 3="Confirmed in situ carcinoma/LMP/borderline cancer" • 5="Last Negative ASU prior to cancer report" • 8="Death" • 9="Participant Withdrawal or Lost Contact" • 13="Cutoff for screening center data collection" • 14="13 year cutoff" • 17="Last ASU"
Exit Age for Mortality	Age of participant at exit for mortality. This includes age at death for participants who are known to be dead, and age at trial exit for participants not known to be dead.	Numeric
Days Until Exit for Mortality	Days from trial entry (randomization) to death for participants known to be dead, or to trial exit for participants not known to be dead.	Numeric
Exit Status for Mortality	Status of the participant at exit for mortality.	• 1="Dead" • 2="Alive"
Days Until Last Contact	Days from randomization until the last contact with the participant.	Numeric
Status of Last Contact	The status at last contact.	• 1="Last ASU" • 2="Death or NRF"

		• 3="Reached cutoff for screening center data collection" • 4="Reached T13 cutoff"
Has a Non-Response Form	Does the participant have a Non-Response Form (NRF)? This represents an end of participation in the trial, either from refusal to continue with study activities, or from a loss of contact with the participant. The trial generally does not learn about cancers diagnosed after non-response.	• 0="No" • 1="Yes"
Days Until Non-Response Form	Date of the non-response. This represents either the date of the refusal or the last contact prior to loss of contact. It can be used as a date when the trial knew the participant was still alive.	• Numeric • .F="No Form"
Reason For Non-Response	Reason that the participant is no longer participating in study activities. Given on the non-response form.	• .F="No Form" • 1="Lost Contact" • 2="Medical" • 3="Refused"
CANCER DIAGNOSES		
Diagnosed With Breast Cancer?	Was the participant diagnosed with breast cancer. This is set for those with a breast_cstatus_cat of 1 or 11 (invasive or in situ cancer).	• 0="No confirmed cancer" • 1="Confirmed cancer"
Days Until Breast Cancer Diagnosis Date	Days from randomization until	• Numeric • .D="Death Certificate Only" • .N="Not applicable"

	breast cancer diagnosis.	
Breast Cancer Status	The most complete information the trial has about the participant's current breast cancer status.	• -1="Cancer Before Randomization" • 0="No Cancer" • 1="Confirmed Cancer" • 2="Death Certificate Reported Unconfirmable" • 3="Self/Other Reported Unconfirmable" • 4="Erroneous Report of Cancer" • 11="Confirmed In Situ"
Was Breast Cancer the First Diagnosed Cancer?	Among all of a participant's cancers diagnosed during the trial, was breast cancer the earliest?	• 0="No" • 1="Yes"
Count of Breast Cancers Diagnosed	The BCS effort collected not only multiple primaries, but also later recurrences and earlier breast cancers diagnosed prior to the trial.	Numeric
Has A BCS Form	This breast cancer has an additional level of confirmation from the 'Breast Cancer Supplemental' form. Cancers with a BCS form have additional cancer characteristics available.	• 0="False" • 1="True"
Procedure That Diagnosed Breast Cancer	BCS-1.a	• .F="No BCS Form" • 1="FNA" • 2="Excisional biopsy" • 3="Incisional biopsy" • 4="Core biopsy" • 6="Other breast biopsy, yielding tissue" • 7="Other breast biopsy yielding cytology"

		• 8="Other organ (non breast) biopsy yielding tissue" • 10="Lymph node biopsy yielding tissue" • 11="Lymph node biopsy yielding cytology" • 12="other biopsy, yielding tissue (specify)" • 13="other biopsy, yielding cytology (specify)"
Surgical Resection Procedure	BCS-3	• .F="No BCS Form" • 1="Lumpectomy" • 2="Mastectomy" • 3="Biopsy only" • 4="Other, specify"
Reason For Biopsy	BCS-1.b	• .F="No BCS Form" • 1="Screen derived (occult)" 2="Symptomatic" • 9="Other"
CANCER CHARACTERISTICS		
Estrogen Receptor Status	Summary of BCS-12.a	• .F="No BCS Form" • 1="Negative" • 2="Equivocal - positive cells within • range of 1-9%" • 3="Positive" • 4="Indeterminant" • 5="Not Available" • 6="Ordered, No results" • 7="Not Ordered"
Progesterone Receptor Status	Summary of BCS-12.b	• .F="No BCS Form" • 1="Negative" • 2="Equivocal - positive cells within • range of 1-9%" • 3="Positive" • 4="Indeterminant" • 5="Not Available" • 6="Ordered, No results" • 7="Not Ordered"
Quantitative ER: % Positive Cells	BCS-12.a.2	• Numeric • .F="No BCS Form"

		• .M="Missing"
Quantitative PR: % Positive Cells	BCS-12.b.2	• Numeric • .F="No BCS Form" • .M="Missing" • 999="Not Available"
HER2 Status - IHC	BCS-12.c.1	• .F="No BCS Form" • 0="0" • 1="1+" • 2="2+" • 3="3+" • 8="Not Ordered" • 9="Ordered, No Results"
HER2/CEP12 Ratio	BCS-12.c.2.b	• Numeric • .F="No BCS Form" • .M="Missing"
HER2 Summary	A summary of HER2 FISH and HER2 IHC.	• .F="No BCS Form" • 1="Positive" • 2="Equivocal" • 3="Negative" • 4="Indeterminant" • 5="Ordered, No results"
Breast Cancer Stage	Breast cancer pathologic TNM stage. Using the 5th edition AJCC staging manual.	• .F="No BCS Form" • 1="0" • 2="I" • 3="IIA" • 4="IIB" • 5="IIIA" • 6="IIIB" • 7="IV" • 99="Missing Components"
Breast Cancer M Stage Component (Distant Metastases)	BCS-10.3	• .F="No BCS Form" • .M="Missing" • 1="Mx" • 2="M0" • 3="M1"
Breast Cancer N Stage Component (Nodal Involvement)	BCS-10.2	• .F="No BCS Form" • .M="Missing" • 1="Nx" • 2="N0" • 3="N1a" • 4="N1b" • 5="N1bi" • 6="N1bii" • 7="N1biii" • 8="N1biv"

		• 9="N2" • 10="N3"
Breast Cancer T Stage Component (Primary Tumor)	BCS-10.1	• .F="No BCS Form" • .M="Missing" • 1="Tx" • 3="Tis" • 4="T1" • 5="T1mic" • 6="T1a" • 7="T1b" • 8="T1c" • 9="T2" • 10="T3" • 11="T4" • 12="T4a" • 13="T4b" • 14="T4c" • 15="T4d"
Breast Cancer Behavior	Combines BCS-2 (ICDO2 code), BCS-6 (Behavior) and ICDO2 from the corresponding OCF.	• .F="No form" • 2="In situ" • 3="Malignant, primary site" • 4="Malignant, invasive with in situ • components"
Breast Cancer Grade	Combines BCS-2 (ICDO2 code), BCS-7 (Histopathologic Grade) and ICDO2 from the corresponding OCF.	• .F="No Form" • 1="Well differentiated; Grade I" • 2="Moderately differentiated; • Grade II" • 3="Poorly differentiated; Grade III" • 4="Undifferentiated; Grade IV" • 9="Not determined/stated/or • applicable"
Breast Cancer Type	Derived from morphology code.	• .F="No BCS Form" • 1="Lobular" • 2="Tubular" • 3="Ductal, NOS" • 4="Other"
Breast Cancer Morphology	Combines BCS-2 (ICDO2 code), BCS-5 (Histopathologic Type)	• Numeric • .F="No form"

	and ICDO2 from the corresponding OCF.	
Breast Cancer Topography	Breast Cancer Topography (best) .	• " "="Not applicable, no form, or • missing" • "C500"="Nipple" • "C501"="Central portion of breast" • "C502"="Upper inner quadrant of • breast" • "C503"="Lower inner quadrant of • breast" • "C504"="Upper outer quadrant of • breast" • "C505"="Lower outer quadrant of • breast" • "C506"="Axillary tail of breast" • "C508"="Overlapping lesion of • breast" • "C509"="Breast NOS"
Breast Cancer Tumor Size	Derived first from t-stage, or if t-stage is missing, from BCS-8 (Tumor Size).	• .F="No BCS Form" • .M="Missing" • 1="0 cm to less than 2 cm" • 2="2 cm to less than 5 cm" • 3="5 cm or more"
MORTALITY STATUS		
Dead?	Is the participant confirmed dead?	• 0="Not Confirmed Dead" • 1="Dead"
Days Until Death	Days from randomization until date of death.	• Numeric • .N="Not applicable"
Death Status	Death status category. Describes whether or not the participant was confirmed dead, with or without known cause.	• 0="No Report of Death" • 1="Confirmed Dead with Known • Cause" • 2="Confirmed Dead, Cause • Unknown " • 3="Presumed Dead"
CAUSE OF DEATH-DEATH CERTIFICATE		

What is Cause of Death Cancer Code from Death Certificate?	Grouping of ICD-9 codes from the death certificate reported underlying cause of death.	• ="No Icd9code, deathstat=4,5" • .M="Missing ICD9Code" • .N="Not Dead" • 1="Abdomen" • 2="Adrenal glands" • 3="Bladder" • 5="Brain" • 6="Breast" • 7="Cervix" • 10="Digestive system" • 11="Esophagus" • 12="Fallopian tubes" • 13="Female genital" • 14="Hodgkins disease" • 15="Intestine" • 16="Ill-defined site" • 17="Kidney and renal pelvis" • 18="Larynx" • 19="Leukemia" • 20="Liver" • 21="Lung" • 24="Melanoma" • 25="Lip, oral cavity, pharynx" • 26="Non-hodgkins lymphoma" • 28="Ovary" • 29="Pancreas" • 30="Peritoneum" • 33="Skin" • 34="Stomach" • 37="Thyroid" • 38="Uterus" • 39="Vagina" • 40="Anus and anal canal" • 41="Connective, subcutaneous, • and other soft tissues and • peripheral nervous system • (excluding diaphragm)" • 42="Endocrine glands" • 43="_Endometrium_" • 44="Eye" • 45="Gallbladder" • 46="Heart, mediastinum, and

		• pleura" • 47="Hematopoietic and • reticuloendothelial systems • (excluding spleen)" • 49="Meninges" • 50="Multiple myeloma" • 51="Nasopharynx, nasal cavity, • middle ear, sinuses" • 52="Pelvis" • 56="Spinal cord and cranial • nerves" • 58="Thymus" • 60="Ureter, urinary organs" • 80="Colorectal" • 81="Colon Appendix" • 88="Other Cancer" • 99="Not Cancer"
Underlying Cause of Death	####X or E#### ICD-9 code for underlying cause of death from the death certificate.	Char, 5
Death Certificate Cause of Death (From Cancer)	Grouping of ICD-9 codes from the death certificate reported underlying cause of death. This grouping is based on the PLCO trial cancers of interest.	• .F="No Form" • .M="Missing" • .N="Not Dead" • 2="Lung" • 3="Colorectal" • 4="Ovary, Peritoneum and other • Female Genital Organs" • 11="Pancreas" • 12="Melanoma of the Skin" • 13="Bladder" • 14="Breast" • 15="Hematopoietic" • 16="Endometrial" • 17="Glioma" • 18="Renal" • 19="Thyroid" • 20="Head and Neck" • 21="Liver" • 23="Upper-Gastrointestinal" • 24="Biliary" • 98="Other Cancer" • 99="Not Cancer"

Cause of Death from Death Certificate	Grouping of ICD-9 codes from the death certificate reported underlying cause of death. This grouping is based on official trial definitions for PLCO cancers and standard ICD-9 groupings for other causes of death. The PLCO trial assesses the ICD-9 code of 185XX as prostate cancer, 162XX as lung cancer, 153XX-154XX (except 1535X) as colorectal cancer, and 183XX as ovarian cancer.	• .F="No Form" • .M="Missing" • .N="Not applicable" • 2="Lung" • 3="Colorectal" • 4="Ovarian" • 5="Peritoneal" • 6="Fallopian Tube" • 100="Non-PLCO Neoplasms" • 200="Ischemic Heart Disease" • 300="Cerebrovascular Accident" • 400="Other Circulatory Disease" • 500="Respiratory Illness" • 600="Digestive Disease" • 700="Infectious Disease" • 800="Endocrine, Nutritional and • Metabolic Diseases, and Immunity • Disorders" • 900="Diseases of the Nervous • System" • 1000="Accident" • 1100="Other"
What is Final Underlying Cause of Death Cancer Code?	Grouping of ICD-9 codes from the final cause of death.	• .F="No Icd9code, deathstat=4,5" • .M="Missing ICD9Code" • .N="Not Dead" • 1="Abdomen" • 2="Adrenal glands" • 3="Bladder" • 5="Brain" • 6="Breast" • 7="Cervix" • 9="Diaphragm and connective • tissue of thorax" • 10="Digestive system" • 11="Esophagus" • 12="Fallopian tubes" • 13="Female genital" • 14="Hodgkins disease" • 15="Intestine"

		• 16="Ill-defined site" • 17="Kidney and renal pelvis" • 18="Larynx" • 19="Leukemia" • 20="Liver" • 21="Lung" • 24="Melanoma" • 25="Lip, oral cavity, pharynx" • 26="Non-hodgkins lymphoma" • 28="Ovary" • 29="Pancreas" • 30="Peritoneum" • 33="Skin" • 34="Stomach" • 37="Thyroid" • 38="Uterus" • 39="Vagina" • 40="Anus and anal canal" • 41="Connective, subcutaneous, • and other soft tissues and • peripheral nervous system • (excluding diaphragm)" • 42="Endocrine glands" • 43="_Endometrium_" • 44="Eye" • 45="Gallbladder" • 46="Heart, mediastinum, and • pleura" • 47="Hematopoietic and • reticuloendothelial systems • (excluding spleen)" • 49="Meninges" • 50="Multiple myeloma" • 51="Nasopharynx, nasal cavity, • middle ear, sinuses" • 52="Pelvis" • 56="Spinal cord and cranial • nerves" • 58="Thymus" • 60="Ureter, urinary organs" • 80="Colorectal" • 81="Colon Appendix"

		• 88="Other Cancer" • 99="Not Cancer"
Final Underlying Cause of Death	####X or E#### ICD-9 code for final underlying cause of death, combining the DCF, NDI, and CDQ. Taken from the CDQ if it was completed, and otherwise form the NDI or DCF. If the CDQ was completed but is missing the ICD-9 code, then it will be missing in this variable.	• • Char, 5 • " "="Missing or not applicable"
Cause of Death (From Cancer)	Grouping of ICD-9 codes from the final cause of death. This grouping is based on the PLCO trial cancers of interest.	• .F="No Form" • .M="Missing" • .N="Not Dead" • 2="Lung" • 3="Colorectal" • 4="Ovary, Peritoneum and other • Female Genital Organs" • 11="Pancreas" • 12="Melanoma of the Skin" • 13="Bladder" • 14="Breast" • 15="Hematopoietic" • 16="Endometrial" • 17="Glioma" • 18="Renal" • 19="Thyroid" • 20="Head and Neck" • 21="Liver" • 23="Upper-Gastrointestinal" • 24="Biliary" • 98="Other Cancer" • 99="Not Cancer"
Cause of Death	Grouping of ICD-9 codes from the final cause of death. This grouping is based on official trial definitions for PLCO cancers and standard ICD-9	• .F="No Form" • .M="Missing" • .N="Not applicable" • 2="Lung" • 3="Colorectal" • 4="Ovarian" • 5="Peritoneal"

	groupings for other causes of death. The PLCO trial assesses the ICD-9 code of 185XX as prostate cancer, 162XX as lung cancer, 153XX-154XX (except 1535X) as colorectal cancer, and 183XX as ovarian cancer.	• 6="Fallopian Tube" • 100="Non-PLCO Neoplasms" • 200="Ischemic Heart Disease" • 300="Cerebrovascular Accident" • 400="Other Circulatory Disease" • 500="Respiratory Illness" • 600="Digestive Disease" • 700="Infectious Disease" • 800="Endocrine, Nutritional and • Metabolic Diseases, and Immunity • Disorders" • 900="Diseases of the Nervous • System" • 1000="Accident" • 1100="Other"
BQ COHORT ENTRY		
BQ Entry Age	Age at entering the baseline questionnaire cohort. The days are calculated from the later date of randomization and baseline questionnaire completion.	• Numeric • .F="No Form"
Days Until BQ Entry Date	Days until entering the baseline questionnaire cohort. The days are calculated from the later date of randomization and baseline questionnaire completion.	• Numeric • .F="No Form"
BQ COMPLIANCE		
Did the Participant Return the BQ?	Yes/No	• 0="No" • 1="Yes"
Age at BQ	Calculated from date of baseline questionnaire completion and date of birth.	• Numeric • .F="No Form"
Days Until BQ Completion	Question M48, F63 - "What is the date you	• Numeric • .F="No Form"

	completed this questionnaire?" The number of days between BQ completion and randomization.	
Method of Questionnaire Administration	Part of the section, For Office Use Only, headed "Method of Administration".	• .F="No Form" • .M="Not Answered" • 1="Self" • 2="Self With Assistance" • 3="In-Person Interview By SC Staff" • 4="In-Person Interview By Other" • 7="Telephone(Female Form)"
BQ DEMOGRAPHICS		
Race	BQ Form Versions 1 and 2: Question 2 - "Which of these best describes your race or ethnic background?" BQ Form Version 3: Question 2 - "Which of these groups best describes you?" Question 2a - "Are you of Hispanic origin?" Participants can only be considered white or black when they are not Hispanic. If the participant is white or black and Hispanic, then they are considered Hispanic. If the participant is Asian, Pacific Islander, or American Indian then they are considered that race.	• 1="White, Non-Hispanic" • 2="Black, Non-Hispanic" • 3="Hispanic" • 4="Asian" • 5="Pacific Islander" • 6="American Indian" • 7="Missing"
Are You Of Hispanic Origin?	BQ Form Versions 1 and 2: Question 2. BQ Form Version 3: Question 2a. What is your race or ethnicity?	• .F="No Form" • .M="Not Answered" • 0="Not Hispanic" • 1="Hispanic"
Education	Question 3 - "What is the highest grade or	• .F="No Form" • .M="Not Answered"

	level of schooling you completed?"	• 1="Less Than 8 Years" • 2="8-11 Years" • 3="12 Years Or Completed High • School" • 4="Post High School Training • Other Than College" • 5="Some College" • 6="College Graduate" • 7="Postgraduate"
Marital Status	Question 4 - "What is your current marital status?"	• .F="No Form" • .M="Not Answered" • 1="Married Or Living As Married" • 2="Widowed" • 3="Divorced" • 4="Separated" • 5="Never Married"
Occupation	Question 5 - "Which of these categories best describes your current working situation?"	• .F="No Form" • .M="Not Answered" • 1="Homemaker" • 2="Working" • 3="Unemployed" • 4="Retired" • 5="Extended Sick Leave" • 6="Disabled" • 7="Other"
State of Birth	Question 1 - "In what state or foreign country were you born?" Participants who were born in Canadian provinces and territories were collapsed into Canada. Participants born in the different U.S. Territories were collapsed into a single U.S. territories category.	• .F="No Form" • .M="Missing" • 0="Foreign Country" • 1="Alabama" • 2="Alaska" • 4="Arizona" • 5="Arkansas" • 6="California" • 8="Colorado" • 9="Connecticut" • 10="Delaware" • 11="District Of Columbia" • 12="Florida" • 13="Georgia" • 15="Hawaii" • 16="Idaho" • 17="Illinois" • 18="Indiana"

		• 19="Iowa" • 20="Kansas" • 21="Kentucky" • 22="Louisiana" • 23="Maine" • 24="Maryland" • 25="Massachusetts" • 26="Michigan" • 27="Minnesota" • 28="Mississippi" • 29="Missouri" • 30="Montana" • 31="Nebraska" • 32="Nevada" • 33="New Hampshire" • 34="New Jersey" • 35="New Mexico" • 36="New York" • 37="North Carolina" • 38="North Dakota" • 39="Ohio" • 40="Oklahoma" • 41="Oregon" • 42="Pennsylvania" • 44="Rhode Island" • 45="South Carolina" • 46="South Dakota" • 47="Tennessee" • 48="Texas" • 49="Utah" • 50="Vermont" • 51="Virginia" • 53="Washington" • 54="West Virginia" • 55="Wisconsin" • 56="Wyoming" • 57="Puerto Rico" • 58="U.S. Territories" • 59="Canada"
BQ SMOKING		
Cigarette Smoking Status	Participant's current cigarette smoking status.	• .A="Ambiguous" • .F="No Form" • .M="Not Answered" • 0="Never Smoked Cigarettes" • 1="Current Cigarette Smoker"

		• 2="Former Cigarette Smoker"
# of Years Since Stopped Smoking Cigarettes	The number of years passed since the participant has stopped smoking.	• Numeric • .F="No Form" • .M="Not Answered" • .N="Not Applicable" • 0.5="Six Months"
Duration Smoked Cigarettes	The total number of years the participant smoked.	• Numeric • .F="No Form" • .M="Not Answered" • 0.5="Six Months"
# of Cigarettes Smoked Per Day	Question 14 - "During periods when you smoked, how many cigarettes did or do you usually smoke per day?"	• .F="No Form" • .M="Not Answered" • 0="0" • 1="1-10" • 2="11-20" • 3="21-30" • 4="31-40" • 5="41-60" • 6="61-80" • 7="81+"
Pack Years	Number of packs smoked per day * years smoked.	• Numeric • .F="No Form" • .M="Missing"
Ever Smoked Cigars?	Question 17 - "Do you now or did you ever smoke cigars regularly for a year or longer?"	• .F="No Form" • .M="Not Answered" • 0="Never" • 1="Current Cigar Smoker" • 2="Former Cigar Smoker"
Usually Filtered or Non-Filtered?	Question 15 - "During periods when you smoked, did or do you more often smoke filter or non-filter cigarettes?"	• .F="No Form" • .M="Not Answered" • .N="Not Applicable" • 1="Filter" • 2="Non-Filter" • 3="About Equal"
Ever Smoked a Pipe?	Question 16 - "Do you now or did you ever smoke a pipe regularly for a year or longer?"	• .F="No Form" • .M="Not Answered" • 0="Never" • 1="Current Pipe Smoker" • 2="Former Pipe Smoker"
Smoke Regularly Now?	Question 12 - "Do you smoke cigarettes regularly now?"	• .F="No Form" • .M="Not Answered" • .N="Not Applicable" • 0="No" • 1="Yes"

Age Started Smoking	Question 11 - "At what age did you start smoking cigarettes regularly?"	• Numeric • .F="No Form" • .M="Not Answered Or • Inconsistent Data" • .N="Not Applicable" • .R="Age not in reasonable range."
Ever Smoke Regularly >= 6 Months?	Question 10 - "Have you ever smoked cigarettes regularly for six months or longer?"	• .F="No Form" • .M="Not Answered" • 0="No" • 1="Yes"
Age Stopped Smoking	Question 13 - "At what age did you last stop smoking cigarettes regularly?"	• Numeric • .F="No Form" • .M="Not Answered Or • Inconsistent Data" • .N="Not Applicable" • .R="Age not in reasonable range."
BQ FAMILY HISTORY		
Has Family History of Any Cancer?	Any first-degree relative with cancer. Basal cell skin cancers are not included. First-degree relatives include parents, full-siblings, and children. Half-siblings are not included.	• .F="No Form" • .M="Not Answered" • 0="No" • 1="Yes"
Family History of Female Breast Cancer	Breast cancer family history in first-degree relatives. Includes parents, full-siblings, and children.	• .F="No Form" • .M="Missing" • 0="No" • 1="Yes, Immediate Female Family • Member" • 2="Male Relative Only" • 9="Possibly - Relative Or Cancer • Type Not Clear"
Age of Youngest Relative with Breast Cancer	Diagnosis age of the youngest first-degree relative diagnosed with breast cancer.	• Numeric • .A="Ambiguous" • .F="No Form" • .M="Missing" • .N="Not Applicable"
# of Relatives with Breast Cancer	The number of first-degree relatives with breast cancer.	• Numeric • .F="No Form"

		• .M="Missing"
# of Brothers	Question 19 - "How many full and half-brothers do you have, both living and deceased?" Participants who have more than seven brothers are collapsed into "7 or more."	• .F="No Form" • .M="Not Answered" • 0="None" • 1="One" • 2="Two" • 3="Three" • 4="Four" • 5="Five" • 6="Six" • 7="Seven Or More"
# of Sisters	Question 18 - "How many full and half-sisters do you have, both living and deceased?" Participants with more than seven sisters are collapsed into "7 or more".	• .F="No Form" • .M="Not Answered" • 0="None" • 1="One" • 2="Two" • 3="Three" • 4="Four" • 5="Five" • 6="Six" • 7="Seven Or More"
BQ BODY TYPE		
BMI at Baseline (In kg/m2)	This is the World Health Organization (WHO) standard categorization of BMI. BMI is considered out of range if any of the following occur: - Weight is less than 60 pounds - Height is less than 48 inches - Height is greater than 78 inches for females - Height is greater than 84 inches for males - After BMI is calculated, BMI is less than 15	• .F="No Form" • .M="Not Answered" • .R="Height Or Weight Not In • Reasonable Range" • 1="0-18.5" • 2="18.5-25" • 3="25-30" • 4="30+"
BMI at Baseline (In kg/m2)	BMI is considered out of range if any of the following occur: - Weight is less than 60 pounds - Height is less than 48 inches - Height is greater than	• Numeric • .F="No Form" • .M="Not Answered" • .R="Height Or Weight Not In • Reasonable Range"

	78 inches for females - Height is greater than 84 inches for males - After BMI is calculated, BMI is less than 15.	
Height (inches)	Question 23 - "How tall are you?" Height is considered out of range if any of the following occur: - Height is less than 48 inches - Height is greater than 78 inches for females - Height is greater than 84 inches for males - After BMI is calculated, BMI is less than 15.	• Numeric • .F="No Form" • .M="Missing" • .R="Height Out Of Range"
Weight (lbs) at Baseline	Question 22 - "What is or was your weight at these ages?" Weights less than 60 pounds are out of range.	• Numeric • .F="No Form" • .M="Missing" • .R="Weight Out Of Range"
BMI at Age 20 (In kg/m2)	BMI is considered out of range if any of the following occur: - Weight is less than 60 pounds - Height is less than 48 inches - Height is greater than 78 inches for females - Height is greater than 84 inches for males - After BMI is calculated, BMI is less than 15.	• Numeric • .F="No Form" • .M="Not Answered" • .R="Height Or Weight Not In • Reasonable Range"
Weight at Age 20 (lbs)	Question 22 - "What is or was your weight at these ages?" Weights less than 60 pounds are out of range.	• Numeric • .F="No Form" • .M="Missing" • .R="Weight Out Of Range"
BMI at Age 50 (In kg/m2)	BMI is considered out of range if any of the following occur: - Weight is less than 60 pounds - Height is less	• Numeric • .F="No Form" • .M="Not Answered" • .R="Height Or Weight Not In • Reasonable Range"

	than 48 inches - Height is greater than 78 inches for females - Height is greater than 84 inches for males - After BMI is calculated, BMI is less than 15.	
Weight at Age 50 (lbs)	Question 22 - "What is or was your weight at these ages?" Weights less than 60 pounds are out of range.	• Numeric • .F="No Form" • .M="Missing" • .R="Weight Out Of Range"
BQ NSAIDS		
Use Aspirin Regularly?	Question 24 - "During the last 12 months, have you regularly used aspirin or aspirin-containing products, such as Bayer, Bufferin or Anacin? (Please do not include aspirin-free products such as Tylenol and Panadol.)"	• .F="No Form" • .M="Not Answered" • 0="No" • 1="Yes"
# of Aspirin	Question 25 - "During the last 12 months, how many pills of aspirin or aspirin containing products did you usually take per day, per week or per month?"	• .F="No Form" • .M="Not Answered" • 0="None" • 1="1/Day" • 2="2+/Day" • 3="1/Week" • 4="2/Week" • 5="3-4/Week" • 6="<2/Month" • 7="2-3/Month"
Use Ibuprofen Regularly?	Question 26 - "During the last 12 months, have you regularly used ibuprofen-containing products, such as Advil, Nuprin, or Motrin?"	• .F="No Form" • .M="Not Answered" • 0="No" • 1="Yes"
# of Ibuprofen	Question 27 - "During the last 12 months, how many pills of ibuprofen-containing products did you	• .F="No Form" • .M="Not Answered" • 0="None" • 1="1/Day" • 2="2+/Day"

	usually take per day, per week, or per month?"	• 3="1/Week" • 4="2/Week" • 5="3-4/Week" • 6="<2/Month" • 7="2-3/Month"
BQ DISEASES		
Arthritis	Did the participant ever have arthritis?	• .F="No Form" • .M="Not Answered" • 0="No" • 1="Yes"
Bronchitis	Did the participant ever have chronic bronchitis?	• .F="No Form" • .M="Not Answered" • 0="No" • 1="Yes"
Colon Comorbidities	Did the participant ever have a colon related co-morbidity (ulcerative colitis, Crohn's disease, Gardner's syndrome, or familial polyposis)?	• .F="No Form" • .M="Missing" • 0="No" • 1="Yes"
Diabetes	Did the participant ever have diabetes?	• .F="No Form" • .M="Not Answered" • 0="No" • 1="Yes"
Diverticulitis/Diverticulosis	Did the participant ever have diverticulitis or diverticulosis?	• .F="No Form" • .M="Not Answered" • 0="No" • 1="Yes"
Emphysema	Did the participant ever have emphysema?	• .F="No Form" • .M="Not Answered" • 0="No" • 1="Yes"
Gallbladder Stones or Inflammation	Did the participant ever have gall bladder stones or inflammation?	• .F="No Form" • .M="Not Answered" • 0="No" • 1="Yes"
Heart Attack	Did the participant ever have coronary heart disease or a heart attack?	• .F="No Form" • .M="Not Answered" • 0="No" • 1="Yes"
Hypertension	Did the participant ever have high blood pressure?	• .F="No Form" • .M="Not Answered" • 0="No" • 1="Yes"

Liver Comorbidities	Did the participant ever have a liver related co-morbidity (hepatitis or cirrhosis)?	• .F="No Form" • .M="Missing" • 0="No" • 1="Yes"
Osteoporosis	Did the participant ever have osteoporosis?	• .F="No Form" • .M="Not Answered" • 0="No" • 1="Yes"
Colorectal Polyps	Did the participant ever have colorectal polyps?	• .F="No Form" • .M="Not Answered" • 0="No" • 1="Yes"
Stroke	Did the participant have a stroke?	• .F="No Form" • .M="Not Answered" • 0="No" • 1="Yes"
BQ FEMALE SPECIFIC		
Ever Have a Hysterectomy?	Question F47 - "Have you had a hysterectomy, that is, have you had your uterus or womb removed?" Participants modified to "yes" if an age of hysterectomy is given in question F48	• .F="No Form" • .M="Not Answered" • 0="No" • 1="Yes" • 2="Don't Know"
Age at Hysterectomy	Question F48 - "What was your age when you had your uterus or womb removed?"	• .F="No Form" • .M="Not Answered" • .N="Not Applicable" • 1="<40" • 2="40-44" • 3="45-49" • 4="50-54" • 5="55+"
Removed Ovaries	Question F49 - "Have you ever had one or both of your ovaries removed?" Question F50 - "What exactly was removed?"	• .F="No Form" • .M="Not Answered" • 0="Ovaries Not Removed" • 1="One Ovary - Partial" • 2="One Ovary - Total" • 3="Both Ovaries - Partial" • 4="Both Ovaries - Total" • 5="Don't Know" • 8="Ambiguous"

Ever Tubes Tied?	Question F46 - "Have you had a tubal ligation, that is have you had your tubes tied?"	• .F="No Form" • .M="Not Answered" • 0="No" • 1="Yes" • 2="Don't Know"
Ever Take Birth Control Pills?	Question F43 - "Did you ever take birth control pills for birth control or to regulate menstrual periods?". Participant's answer modified to "yes" if they specified both an age they started taking birth control pills and a total number of years they took them.	• .F="No Form" • .M="Not Answered" • 0="No" • 1="Yes"
Age Started Birth Control Pills?	Question F44 - "How old were you when you first started taking birth control pills?" Participants who were "50-59" or "60+" when they started birth control pills were collapsed into a "50+" category.	• .F="No Form" • .M="Not Answered" • .N="Not Applicable" • 1="<30" • 2="30-39" • 3="40-49" • 4="50+" •
Total Years Took Birth Control Pills?	Question F45 - "For how many total years did you take birth control pills?"	• .F="No Form" • .M="Not Answered" • 0="Not Applicable" • 1="10+ Years" • 2="6-9 Years" • 3="4-5 Years" • 4="2-3 Years" • 5="< 1 Year"
Currently Using Female Hormones?	Question F52 - "Are you currently using female hormones?"	• .F="No Form" • .M="Not Answered" • 0="No" • 1="Yes"
Ever Take Female Hormones?	Question F51 - "Sometimes women take female hormones such as estrogen or progesterone around the time of	• .F="No Form" • .M="Not Answered" • 0="No" • 1="Yes" • 2="Don't Know"

	menopause. Have you ever used female hormones (tablets, pills, or creams) for menopause?" Participant's answers modified to "yes" if they had said "no" but gave an answer for whether they are currently using female hormones and said they used them for greater than 1 year.	
Female Hormone Status	Female hormone status uses ever taken female hormones and currently on hormones to determine the participant's hormone status.	• .F="No Form" • .M="Missing" • 0="Never" • 1="Current" • 2="Former" • 3="Unknown Whether Current Or • Former" • 4="Doesn't Know If She Ever • Took HRT"
# of Years Taking Female Hormones	Question F53 - "For how many total years did you take female hormones?"	• .F="No Form" • .M="Not Answered" • 0="Not Applicable" • 1="10+ Years" • 2="6-9 Years" • 3="4-5 Years" • 4="2-3 Years" • 5="<= 1 Year"
Age at Birth of First Child?	Question F42 - "What was your age at the birth of your first child?"	• .F="No Form" • .M="Not Answered" • .N="Not Applicable" • 1="<16" • 2="16-19" • 3="20-24" • 4="25-29" • 5="30-34" • 6="35-39" • 7="40+"
# of Live Births	Question F41 - "How many of your pregnancies resulted in a live birth?"	• .F="No Form" • .M="Not Answered" • 0="Zero" • 1="One"

	Allowed values are 0-29. Participants with more than five pregnancies are collapsed to "five or more".	• 2="Two" • 3="Three" • 4="Four" • 5="Five Or More"
# of Miscarriages/Abortions	Question F39 - "How many of your pregnancies resulted in miscarriage or an abortion?"	• .F="No Form" • .M="Not Answered" • 0="0" • 1="1" • 2="2+"
Ever Been Pregnant?	Question F35 - "Have you ever been pregnant?" Participant's answer is modified to be "yes" if the participant answers on age of first pregnancy, number of pregnancies, number of still birth pregnancies, number of miscarriages, number of tubal pregnancies, age at birth of first child, or the number of live births implied pregnancy.	• .F="No Form" • .M="Not Answered" • 0="No" • 1="Yes" • 2="Don't Know"
Age When First Became Pregnant?	Question F36 - "How old were you when you first became pregnant?" Participants who were "40-44" or "45+" when they first became pregnant were collapsed into "40+".	• .F="No Form" • .M="Not Answered" • .N="Not Applicable" • 1="<15" • 2="15-19" • 3="20-24" • 4="25-29" • 5="30-34" • 6="35-39" • 7="40+"
# of Pregnancies	Question F37 - "How many times have you been pregnant? Please include stillbirths, miscarriages, abortions, tubal or	• .F="No Form" • .M="Not Answered" • 0="None" • 1="1" • 2="2" • 3="3-4" • 4="5-9"

	ectopic pregnancies, and live births."	• 5="10+"
# of Still Birth Pregnancies	Question F38 - "How many of your pregnancies resulted in a stillbirth?"	• .F="No Form" • .M="Not Answered" • 0="0" • 1="1" • 2="2+"
Ever Tried to Become Pregnant for a Year or More Without Success?	Question F34 - "Have you ever tried to become pregnant for a year or more without success?"	• .F="No Form" • .M="Not Answered" • 0="No" • 1="Yes"
# of Tubal/Ectopic Pregnancies?	Question F40 - "How many of your pregnancies resulted in a pregnancy in one of your tubes, that is, a tubal or ectopic pregnancy?"	• .F="No Form" • .M="Not Answered" • 0="No" • 1="Yes"
Age When Had First Menstrual Period?	Question F31 - "How old were you when you had your first menstrual period?"	• .F="No Form" • .M="Not Answered" • 1="<10" • 2="10-11" • 3="12-13" • 4="14-15" • 5="16+"
Age at Menopause	Question F32 - "How old were you when you had your last period?"	• .F="No Form" • .M="Not Answered" • 1="<40" • 2="40-44" • 3="45-49" • 4="50-54" • 5="55+"
Type of Menopause	Question F33 - "Did your periods stop because of natural menopause, surgery, radiation, or drug therapy?"	• .F="No Form" • .M="Not Answered" • 1="Natural Menopause" • 2="Surgery" • 3="Radiation" • 4="Drug Therapy"
Reason menstrual periods stopped.	Reason the participant's menstrual periods stopped. Because minimal information	• .F="No Form" • 1="Natural postmenopausal" • 2="Bilateral oophorectomy" • 3="Hysterectomy, no bilateral

	was gathered about menopause, the menopause information is supplemented with hysterectomy and oophorectomy information.	• oophorectomy" • 4="Surgical, details unclear" • 5="Drug therapy" • 6="Radiation" • 7="Postmenopausal, reason • unknown" • 8="Menopausal status unknown"
Post-Menopausal Status	Was the participant post-menopausal at trial entry. This question was not asked directly on the BQ, therefore information on menopause has been supplemented with hysterectomy and oophorectomy information.	• .F="No form" • 1="Definitely post-menopausal" • 2="Possibly post-menopausal"
Ever Have Benign or Fibrocystic Breast Disease?	Question F54 - "Have you ever been told by a doctor that you had any of the following conditions?"	• .F="No Form" • .M="Not Answered" • 0="No" • 1="Yes"
Ever Have Benign Ovarian Tumor/Cyst?	Question F54 - "Have you ever been told by a doctor that you had any of the following conditions?"	• .F="No Form" • .M="Not Answered" • 0="No" • 1="Yes"
Ever Have Endometriosis?	Question F54 - "Have you ever been told by a doctor that you had any of the following conditions?"	• .F="No Form" • .M="Not Answered" • 0="No" • 1="Yes"
Ever Have Uterine Fibroid Tumors?	Question F54 - "Have you ever been told by a doctor that you had any of the following conditions?"	• .F="No Form" • .M="Not Answered" • 0="No" • 1="Yes"

4.3.2. ISPY1 Dataset

This dataset contains the clinical and MRI data from the ISPY1 clinical trial of patients with breast cancer[6]. The goal of this project is to improve the prediction of clinical outcomes to neoadjuvant chemotherapy in patients with breast cancer. Currently, most patients with breast cancer undergo neoadjuvant chemotherapy, which is aimed to reduce the tumor size (burden) before surgery to remove the tumor or the entire breast. Some of the patients response completely to the therapy and the patient does not present any residual tumor at the time of surgery. On the other hand, some patients have residual disease at the time of surgery and further treatment is required. Table 9 presents the various fields of the dataset.

Table 9. Data fields available for the ISPY dataset

Data Field	Description	Value
CLINICAL DATA		
PATIENT DEMOGRAPHICS		
Age	Patient Age	Numeric
Race_id	Patient Race	• 1=Caucasian • 3=African American • 4=Asian • 5=Native Hawaiian/Pacific Islander • 6=American Indian/Alaskan Native • 50=Multiple race
ON-STUDY DATA (PRE-TREATMENT)		
ERpos	Estrogen Receptor Status (Allred Score or Community determined), pre-treatment	• 0=Negative • 1=Positive • 2=Indeterminate
PgRpos	Progesterone Receptor Status (Allred Score or Community determined), pre-treatment	• 0=Negative • 1=Positive • 2=Indeterminate
HR Pos	Hormone Receptor Status, pre-treatment	• 0=Negative for both ER and PR • 1=Positive if either ER or PR was Positive • 2=Indeterminate if both ER and PR were Indeterminate
Her2MostPos (replaced Her2CommPos, 4/6/2016)	Her2 Status, pre-treatment, adding in Central Her2 IHC results for missing Community Status	• 0=Negative • 1=Positive • Blank= indeterminate or not done

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

HR_HER2_CATEGORY	3-level HR/Her2 category pre-treatment	• 1=HR Positive, Her2 Negative • 2=Her2 Positive • 3=Triple Negative
HR_HER2_STATUS	3-level HR/Her2 status pre-treatment	• HRposHER2neg = HR Positive, Her2 Negative • HER2pos = Her2 Positive • TripleNeg =Triple Negative
BilateralCa	Does the patient have bilateral breast cancer prior to neoadjuvant therapy?	• 0=No • 1=Yes
Laterality	Index Tumor Laterality: •	• 1=Left • 2=Right
IMAGING DATA		
MRI LD:	LD spans all disease present (inv & DCIS) even if there is intervening normal tissue	Numeric (in mm)
Baseline	Timepoint 1= baseline	Numeric
1-3d AC	Timepoint 2= 1-3days after start of AC	Numeric
InterReg	Timepoint 3= Inter-regimen	Numeric
PreSurg	Timepoint 3= Inter-regimen	Numeric
OUTCOME DATA		
SUBJECTID	I-SPY ID de-identifies a patient's CALGB and ACRIN ID	Numeric
DataExtractDt	Date clinical data was downloaded from the CALGB database	Date format
Sstat	Survival Status •	• 7=Alive • 8= Dead • 9=Lost
SurvDtD	Survival date (time from study entry to death or last follow-up; time unit is days	Numeric
RFS	Recurrence-free survival time – time from neoadjuvant chemotherapy start date until earliest: local or distant progression or death (time unit is days)	Numeric

RFS_ind	Recurrence-free survival indicator •	• 1=event (local or distant progression or death) • 0=censor at last follow-up
pCR	Pathologic Complete Response, post-neoadjuvant (no residual invasive disease in breast or lymph nodes; presence of only in situ disease are considered disease free) •	• 0= No (did not achieve pCR) • 1= Yes • Blank= no surgery
RCBClass	Residual Cancer Burden class •	• 0= 0, RCB index 0 • 1= I, RCB index less than or equal to 1.36 • 2= II, RCB index greater than 1.36 or equal to 3.28 • 3= III, RCB index greater than 3.28 • Blank= unavailable or no surgery

5. Knowledge acquisition

In this section we focus on collecting existing knowledge resources for modeling the cancer domain. We differentiate among the various types of knowledge sources as described in Section 3.

5.1. Ontologies

5.1.1. Symptom Ontology (SO)

The Symptom Ontology (SO) [41] was designed around the guiding concept of a symptom being: "A perceived change in function, sensation or appearance reported by a patient indicative of a disease". The Symptom Ontology captures and documents the semantics of two sets of terms, the term “Sign” and the term “Symptom”. The ontology is open source.

It was developed as part of Gemina project, starting in 2005 at the Institute Genome Sciences (IGS) at the University of Maryland. Work ended on 2009. In July 2008 the Symptoms Ontology was submitted for inclusion and review to the OBO Foundry and today the standardization body for the Symptom Ontology is the OBO Foundry. The ontology also provides human readable text together with computer readable format. The available computer readable formats of the ontology are in OBO and OWL. The semantics of the Ontology are coherent, consistent and there is a rigid domain specification. Last but not least the symptom ontology reaches high level of interoperability.

5.1.2. Human Disease Ontology

The Disease Ontology (DO)[34] was initially developed as part of the NUgene project, starting in 2003 at Northwestern. It is an open source ontology that is designed to link disparate datasets through disease concepts. The aim is to facilitate the connection of genetic data, clinical data, and symptoms through the lens of human disease.

The DO enables the cross-walk between disease concepts, genes contributing to disease, and the 'cloud' of associated symptoms, findings and signs.

DO is a formally valid, it encapsulates a comprehensive theory of disease, and has a general domain, the health domain. The standardization body of the DO is the OBO Foundry. Terms in DO use standard references such as SNOMED, ICD-10, MeSH, and UMLS. The ontology also provides human readable text together with computer readable format and thus shows syntactic and semantic interoperability. The available computer readable formats of the ontology are in OBO and OWL. The semantics of the Ontology are coherent, consistent and there is a rigid domain specification.

As mentioned before this ontology has a broad domain, the health domain, so it can be used to model general information to model a disease.

5.1.3. The Foundational Model of Anatomy (FMA)

The Foundational Model of Anatomy[6] is a computer-based, open source ontology available for general use. It is created for biomedical informatics and it has to do with the representation of classes, types and relationships necessary for the symbolic representation of the phenotypic

structure of the human body in a form that is understandable to humans and is also navigable, parse-able and interpretable to machine-based systems. It is a domain ontology that represents a coherent body of explicit declarative knowledge about human anatomy. It is composed of the following components:

1. Anatomy taxonomy (At): classifies anatomical entities according to the characteristics they share (genus) and by which they can be distinguished from one another (differentia)

2. Anatomical Structural Abstraction (ASA): specifies the part-whole and spatial relationships that exist between the entities represented in At;

3. Anatomical Transformation Abstraction (ATA): specifies the morphological transformation of the entities represented in At during prenatal development and the postnatal life cycle;

4. Metaknowledge (Mk): specifies the principles, rules and definitions according to which classes and relationships in the other three components of FMA are represented.

It contains approximately 75,000 classes, over 120,000 terms, over 2.1 million relationship instances and over 168 relationship types. As the FMA should serve as an ontology, its classes are defined in structural terms and grouped into classes on the basis of the structural properties that they share. In such a way, it is possible to aggregate such data in a taxonomy format.

The FMA is the best candidate for serving as a foundation and reference for the correlation of other ontologies in biomedical informatics. Clearly, the FMA is not an application ontology as it is not intended as an end-user application and does not target the needs of particular user groups. Due to the diverse and implied meanings associated with the term “ontology”, the FMA is described (Rosse & Mejino 2003) as a symbolic model rather than an ontology.

5.1.4. Ontology of Adverse Events (AEO)

The Adverse Event Ontology (AEO)[10], shown in Figure 3, is a realism-based biomedical ontology for adverse events. Currently AEO has 484 representational units annotated by means of terms including 369 AEO-specific terms and 115 terms from existing feeder-ontologies. In AEO, the term “adverse event” is used exclusively to denote pathological bodily processes that are induced by a medical intervention.

The development of AEO follows the OBO Foundry principles such as openness, collaboration, and use of a common shared syntax. AEO is thus aligned with the Basic Formal Ontology (BFO) and the Relation Ontology (RO). The AEO is up-to-date since the last version release was version 1.1.64 in July 2012. The available computer readable format of the ontology is OWL and it is being used in 2 projects: the OntoCAT project and the OntoMaton project.

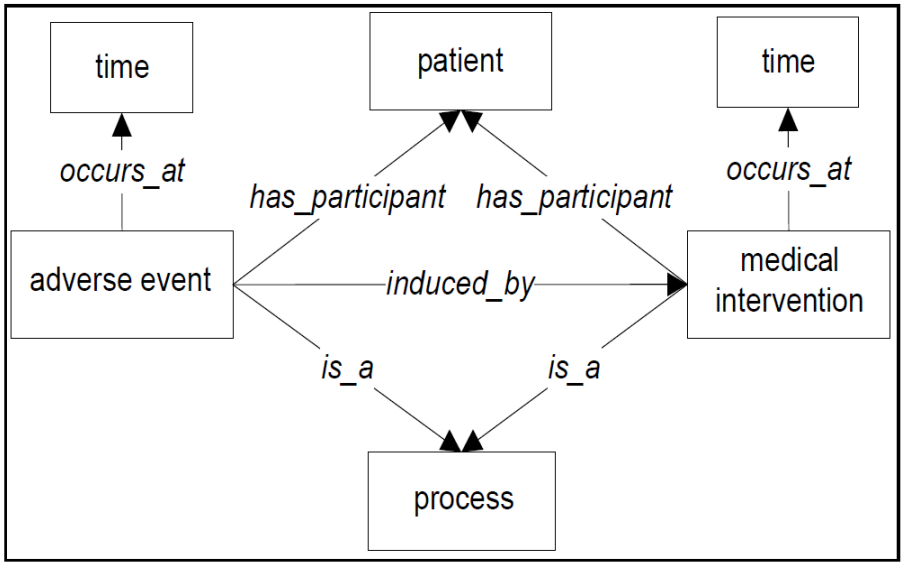


Figure 3. Basic AEO adverse event design pattern.

5.1.5. Experimental Factor Ontology

The Experimental Factor Ontology (EFO)[33] is an application focused ontology modeling the experimental variables in the Gene Expression Atlas. The ontology describes cross-product classes from reference ontologies in area such as disease, cell line, cell type and anatomy. EFO combines parts of several biological ontologies, such as anatomy, disease and chemical compounds. The scope of EFO is to support the annotation, analysis and visualization of data handled by the EBI Functional Genomics Team. The EFO is application ontology – an ontology engineered for domain specific use or application focus and whose scope is specified through testable use cases and which maps to reference or canonical ontologies. The ontology is kept up-to-date since the last updated version of the EFO ontology is version 2.30 on November 2012.

The EFO methodology reuses reference ontologies (full list available at http://www.ebi.ac.uk/efo/metadata), where they exist, and where they describe classes that are in scope for EFO. To promote interoperability with the OBO Foundry ontologies, EFO is using the BFO as an upper ontology. Furthermore, the EFO Ontology is used in the following projects:

- ISA software suite
- NCBO Annotator
- NCBO Resource Index
- OntoCAT
- MeRy-B
- Neural ElectroMagnetic Ontologies
- OntoMaton

5.1.6. UMLS

UMLS [15], the Unified Medical Language System, is a unifying framework which integrates different terminologies which are relevant to medicine and biomedical information technologies. It consists mainly of three parts. The Metathesaurus and the Semantic Network

are the most important ones. The third part, the SPECIALIST lexicon, is a source of lexical information and language processing programs.

The Metathesaurus is currently distributed in two versions, the Rich Release Format (RRF) is provided since 2004. The Original Release Format (ORF) is older. Since RRF is more accurate and precise than ORF it is the preferable option. The Metathesaurus is the core of UMLS. With over five million names for over one million concepts and about 12 million relations between these concepts it is a very broad scoped but also detailed resource for the domain of biomedicine. The purpose of the Metathesaurus is not to give a new terminology but to give an extensive lexicon of existing vocabularies and coding systems. According to a ranking of source vocabularies one of the different terms which belong to the same concept is designated as a preferred term. Whatever is contained in the Metathesaurus has a unique identifier. For example, concepts are attached to a CUI (Concept Unique Identifier), terms get a LUI (Lexical Unique Identifier) and relationships are named by a RUI (Relationship Unique Identifier).

The second part of UMLS is the Semantic Network. Its aim is "to provide a consistent categorization of all concepts represented in the UMLS". The network is a system of abstract categories and provides the foundation for the categorization of the concepts in the Metathesaurus. Every concept in the Metathesaurus is associated to at least one of the categories, usually to the most specific available category. Currently, the Semantic Network has 133 broad categories and 54 relationships. The is-a relation, i.e. subsumption, is essential for the hierarchical structure. The Semantic Network does not aim to be a complete characterization of the world but it is rather limited to medical purposes. This becomes obvious with respect to granularity. Narrow classes are only provided for the domain of biomedicine. Further relations are "physically related to", "spatially related to", "temporally related to", "functionally related to", "conceptually related to" and relation which are subtypes of these five relations. Relations between entities are usually inherited to the terms which are subsumed.

The UMLS has been under development by the US National Library of Medicine (NLM) since the eighties. As an integrating framework its goal is to unite the knowledge expressed in currently over 100 source terminologies for diseases, procedures, supplies and diagnoses, including for example the ICD terminologies and SNOMED, and, thereby, to support interoperability. All parts of UMLS are machine readable. Using UMLS is free of charge but a license agreement is necessary. The UMLS is a global and comprehensive source for manifold medical terminologies and it is hardly possible to ignore it when working on interoperability.

5.1.7. Ontology of medically related Social Entities

This ontology covers the domain of social such as demographic information (social entities for recording gender (but not sex) and marital status, for example) and the roles of various individuals and organizations (patient, caregiver, hospital, etc.) A subset of this ontology may be helpful to implement shared decision making between end-users and automated reasoning.

The last version of the Ontology was released in June 2012 and it consists of 119 classes and only 6 ObjectProperties. It is a rather small ontology with little semantic impact. It is available in OWL format [18].

5.1.8. Neuroscience Information Framework Standardized Ontology (NIFSTD)

The Neuroscience Information Framework Project (NIF) [19] has been developing tools and strategies for creating resources that can be integrated across neuroscience domains. The end product is a semantic search engine and a knowledge discovery portal for describing neuroscience resources and provides access to multiple types of information organized by relevant categories. Through its resource catalog and data federation, NIF represents the source of neuroscience information available on the web.

The semantic framework through which these diverse resources are accessed is provided by the NIF Standardized Ontologies (NIFSTD). NIFSTD represents a collection of terms and concepts from the domains of neuroscience.

The NIFSTD ontologies are built in a modular fashion, where each module covers a distinct, orthogonal domain of neuroscience. Modules covered in NIFSTD include anatomy, cell types, experimental techniques, nervous system function, small molecules, and so forth. The upper-level classes in NIFSTD modules are carefully normalized under the classes of Basic Formal Ontology (BFO). The Ontology is open source available online.

5.1.9. Biocaster Ontology (BCO)

The BioCaster Ontology (BCO) [2] aims to (a) describe the terms and relations necessary to detect and risk assess public health events in the grey literature at an early stage; (b) bridge the gap between the (multilingual) grey literature and existing standards in biomedicine; (c) to be open source and freely available for general usage.

In contrast to other ontologies that describe infectious diseases, the BCO focuses on the usage of terms and relations within informal unstructured reports which are often made at a pre-diagnostic stage of a disease outbreak by non-medically trained reporters. This is done to provide monitoring and early warning about public health hazards from online media reports. An example of its usage can be seen in the BioCaster Global Health Monitor.

The BCO is maintained by Dr. Nigel Collier's group at the National Institute of Informatics in Tokyo with the collaboration of partners in the international life science and computational linguistics communities. This ontology was developed to cover the need of a specific project, the Biocaster project and thus has a very broad and vague domain. It contains models for the analysis of Internet news and research literature for public health workers, clinicians and researchers interested in communicable diseases.

5.1.10. Family Health History Ontology (FHHO)

The FHHO [35] is representing the family health histories of persons related by biological and/or social family relationships (e.g. step, adoptive) who share genetic, behavioral, and/or environmental risk factors for disease. Projects that are linked with this Ontology, as well as other ontologies mapped to FHHO can be found online.

5.1.11. Advancing Clinico-Genomic Trials Master Ontology (ACGT MO)

Advancing Clinico-Genomic Trials on Cancer (ACGT)[21] was a project financed by the European Union within the 6th Framework Program, which aimed at enabling the rapid sharing of data

gained in both clinical trials and associated genomic studies. In order to meet such a goal, ACGT provided a grid-based infrastructure, designed to transmit the data between different groups of users in real time according to their needs with data integration being achieved by means of an ontology-based mediator. The system had been designed to enable the smooth and prompt transfer of laboratory findings to the clinical management and treatment of patients.

The ACGT consortium developed its own Master Ontology (MO) in order to address the goal of data integration for the domains of clinical studies, genomic research and clinical cancer management and care. The MO has been grounded on the Basic Formal Ontology (BFO), which is the Open Biomedical Ontologies (OBO) Foundry's upper level ontology. BFO assured to MO's classes a high level of semantic specification.

The ACGT MO was the core of the ACGT Semantic Mediation Layer (ACGT-SM) which comprised a set of tools and resources working together to serve processes of Database Integration and Semantic Mediation. The ACGT-SM followed a Local-as-View Query Translation approach in order to cope with the problem of database integration. In such a way, the data is not actually integrated but it is made accessible to users via a virtual repository. This repository represents the integration of the underlying databases and ACGT-MO acts as database schema, providing resources for formulation of possible queries.

The MO was constructed in modular fashion with Clinical Trial and Patient Management Ontology modules designed to be reused for different clinical domains. However, although this ontology is already a well-established ontology it has not been widely used.

5.1.12. Systems Biology Ontology (SBO)

The Systems Biology Ontology[22] is a collaborative effort led my Biomodels.Net and it is a set of controlled, relational vocabularies of terms commonly used in Systems Biology, and in particular in computational modelling. There are several orthogonal vocabularies in the ontology defining the following:

- Reaction participants roles (e.g. substrate)
- Quantitative parameters (e.g. pi)
- Classification of mathematical expressions describing the system (e.g. mass action rate law)
- Modelling framework used (e.g. logical framework)
- The nature of the entity (e.g. molecule)
- The type of the interaction (e.g. process)
- The different types of metadata present in a model

The ontology is defined in various formats such as OBO, OWL and XML. Moreover, to allow programmatic access to the resources Web Services have been implemented and libraries, documentation and samples as well.

5.1.13. Psychological Ontology for Breast Cancer Patients (POBC)

The Psychological Ontology for Breast Cancer Patients (POBC) [7] represents the main aspects that have been analyzed by clinical experts for assessing the psychological state of breast cancer patients.

The POBC ontology specifies a variety of psychological categories and each of them is connected to a set of questions that are used by clinical experts in daily practice to assess the patients' psychological state. Although all categories are useful for the psychological analysis, clinical experts selected eight of them as more significant that can also be used computationally to identify a critical situation, empower the patients for self-monitoring, and reduce the need for clinical visits. Figure 4 essentially provides an abstract overview of the subtree in POBC that is relevant in the computational psychological analysis.

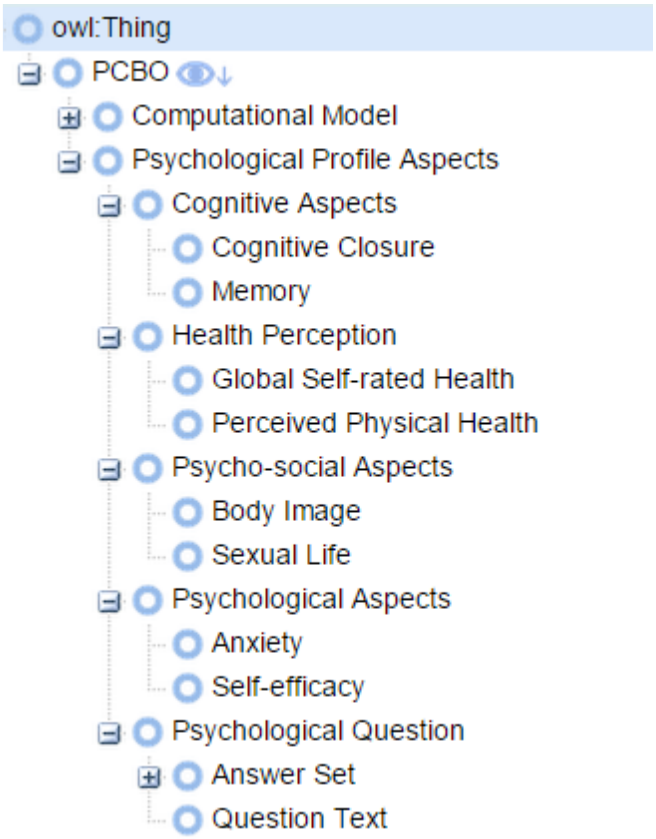


Figure 4 POBC psychological categories

5.1.14. Mental Functioning Ontology (MF)

The Mental Functioning Ontology (MF) is a modular domain ontology for mental functioning, including mental processes such as cognition and traits such as intelligence.

MF has been based on the Basic Formal Ontology (BFO) and has been developed in the context of the OBO Foundry library of interrelated modular domain ontologies.

Figure 5 illustrates the upper levels of the ontology, based on the framework laid out in, together with the alignment to BFO. At the top level, BFO introduces a distinction between continuants and occurrents. Occurrents are processes and other entities that unfold in time, i.e. entities that have temporal parts. Continuants, on the other hand, are those things that exist in full at all

times that they exist, have no temporal parts, and continue to exist over an extended period of time.

Within continuants, BFO further distinguishes between those entities that are independent and those that are dependent. Independent continuants can exist by themselves, while dependent continuants are those sorts of things that need a "bearer" in order to exist, such as colours, social roles, or behavioural dispositions that are realized in behaviour, a concurrent entity.

The illustrated upper levels of MF show several important distinctions in the framework to annotate and describe mental functioning allowing interrelationships across a wide variety of different levels of description. The organism is the fundamental independent continuant in which mental functioning takes place.

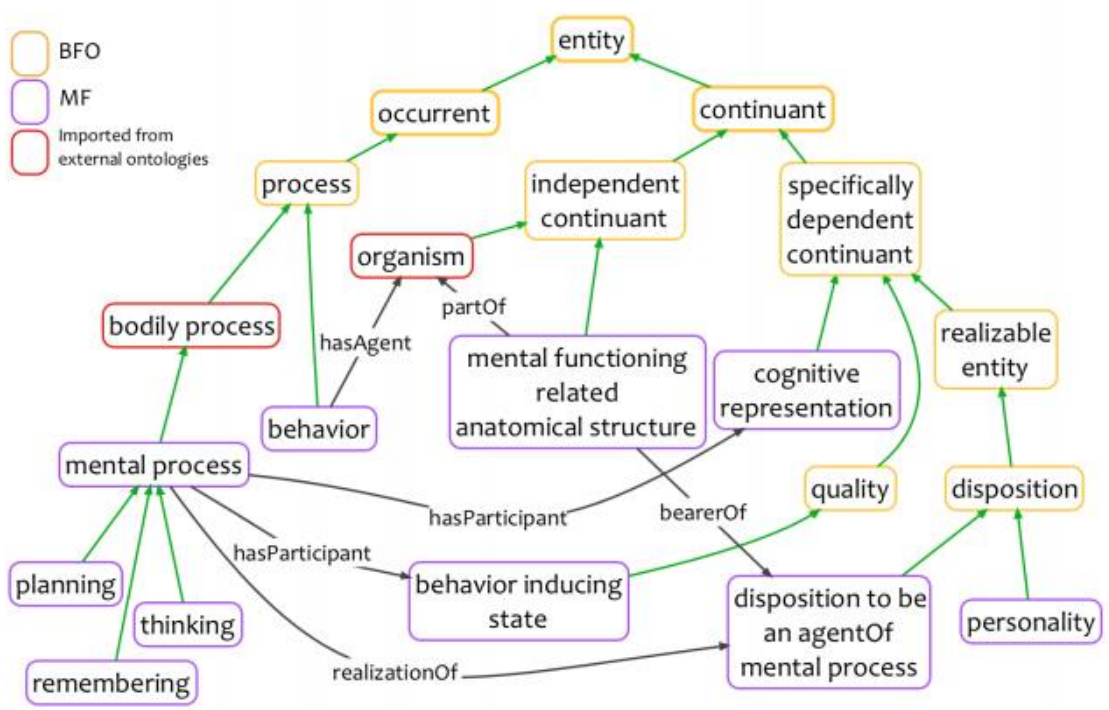


Figure 5 The Mental Functioning Ontology upper level aligned to BFO.

MF is being developed modularly, allowing different teams with different core areas of expertise to focus on the extension of the overall ontology to describe the entities relevant to their scientific area. One such extension is the Emotion Ontology (see bellow), describing entities of relevance to all aspects of affective science.

5.1.15. Emotion Ontology (MFOEM)

The MFOEM [28] is an ontology of affective phenomena such as emotions, moods, appraisals and subjective feelings, designed to support interdisciplinary research by providing unified annotations. The ontology is a domain specialization of the broader Mental Functioning Ontology.

The ontology aims to include all relevant aspects of affective phenomena including their bearers, the different types of emotions, moods, etc., their different parts and dimensions of variation,

their facial and vocal expressions, and the role of emotions and affective phenomena in general in influencing human behavior.

An overview of the organising upper levels of EMO, aligned with the BasicFormal Ontology (BFO) and the Ontology of Mental Disease (OMD) is illustrated in Figure 6.

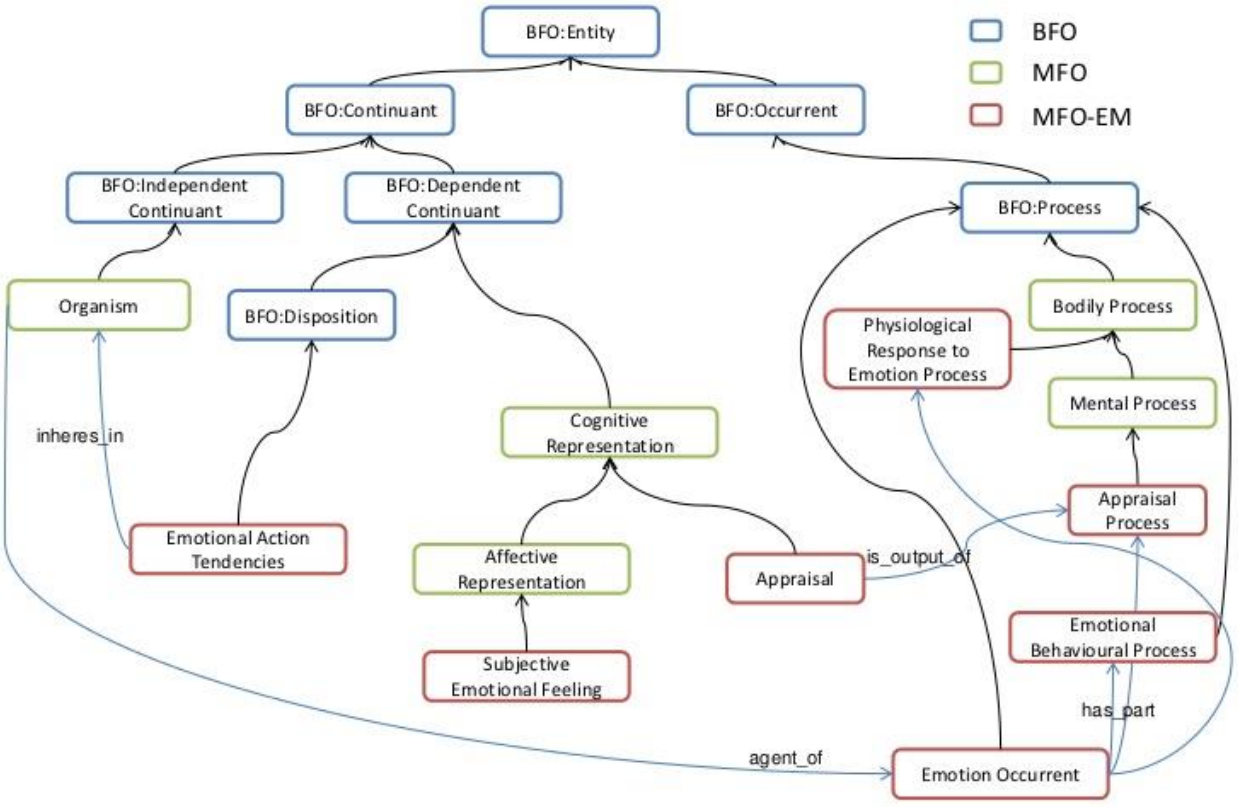


Figure 6 An overview of the Emotion Ontology. Unlabelled arrows represent 'is a' relations

5.1.16. The Health Data Ontology Trunk (HDOT)

The Health Data Ontology Trunk (HDOT) [36] is being developed by the Institute for Formal Ontology and Medical Information Science (IFOMIS) of the University of Saarland and it is conceived as a modular middle-layer ontology. HDOT is being designed, maintained and extended using the ontology editor Protégé, and is released in OWL-DL under the following web address: http.//code.googlc.com/p/hdot/. HDOT integrates under the same semantic umbrella the first version of the Basic Formal Ontology, the Relational Ontology (RO), the Information Artifact Ontology (IAO), the Middle Layer Ontology for Clinical Care (MLOCC), parts of the Phenotypic Quality Ontology (PATO), and parts of the Ontology for General Medical Science (OGMS). Most of them are part of the OBOFoundry initiative and are widely used in the biomedical domain for data annotation and integration.

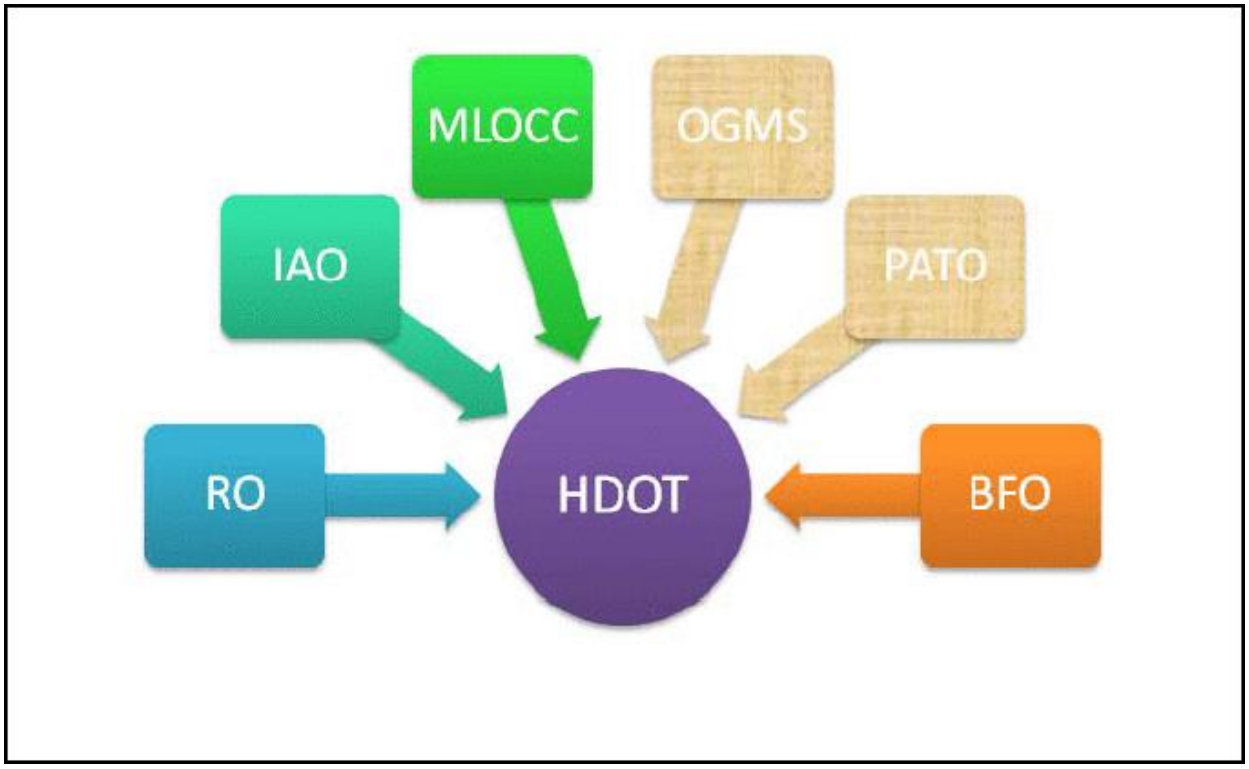


Figure 7 HDOT's modular structure

HDOT is designed in a modular fashion (Figure 7) as a middle-layer ontology in the sense that it specifies upper-level domain independent classes down to the biomedical domain while maintaining at the same time a very general semantic and axiomatic structure that can be further developed and specialized in different modules for different purposes and applications. A part of the ontology depicting the pathological formation class in HDOT is shown in Figure 8.

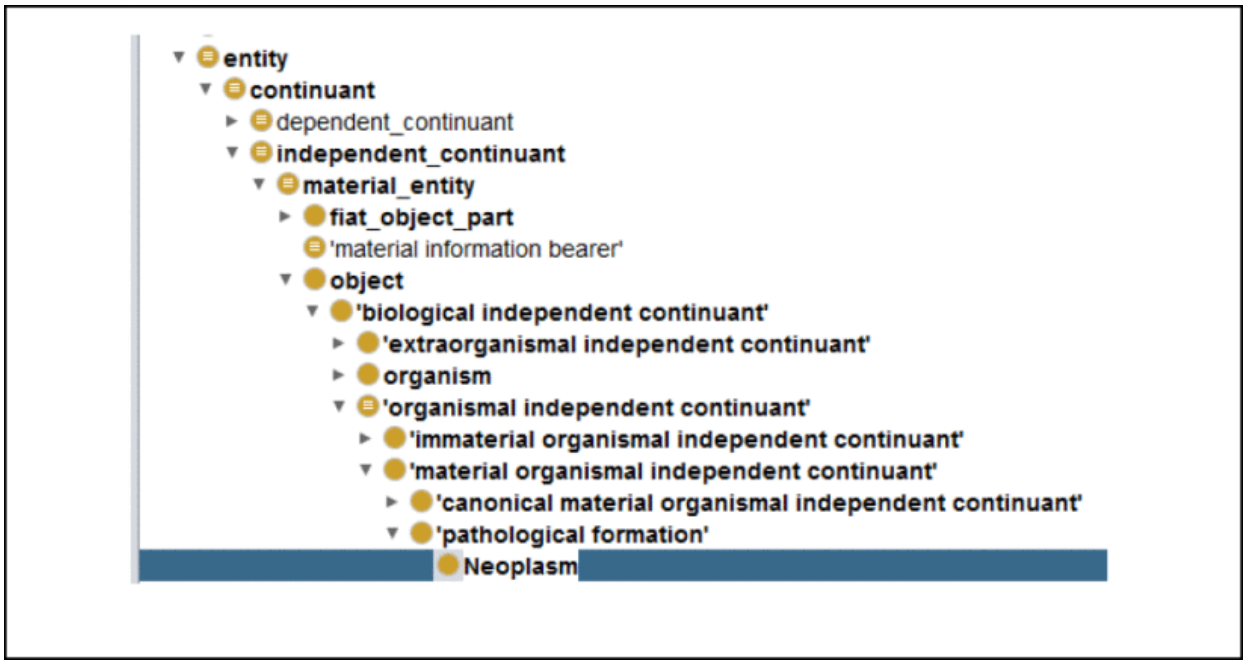


Figure 8 Pathological formation class in HDOT

5.2. Terminologies

5.2.1. Clinical Care Classification System (CCC)

The Clinical Care Classification System [1] , is a standardized framework and a coding structure for assessing, documenting, and classifying patient care by nurses and other clinical professionals in any health care setting. The CCC system consists of two interrelated terminologies, the CCC of Nursing Diagnoses and the CCC of Nursing Interventions/Actions. The two terminologies are both classified by 21 Care Components that represent the Functional, Health Behavioural, Physiological, and Psychological Patterns of Patient Care (Table 1: CCC Care Components).

Table 10. CCC Care components

Care Components		
Activity	Medication	Self-Care
Bowel/Gastric	Metabolic	Self-Concept
Cardiac	Nutritional	Sensory
Cognitive/Neuro	Physical Regulation	Skin Integrity
Coping	Respiratory	Tissue Perfusion
Fluid Volume	Role Relationship	Urinary Elimination
Health Behavior	Safety	Life Cycle

The CCC System is being used to document nursing care in the electronic health record (EHR) computer-based patient record (CPR) and Personal Health Record (PHR) Systems. It serves as a language for nursing and other health care providers such as physical, occupational, and speech therapists, medical social workers, etc. The CCC System is used to:

- Document integrated patient care processes
- Classify and track clinical care
- Develop evidence-based practice models
- Analyze patient profiles and populations
- Predict care needs, resources, and costs

In 2007, the CCC was accepted by the US Department of Health and Human Services as the first national nursing terminology. The coding structures for the terminologies are based on the ICD-10 consisting of five alphanumeric characters for information exchange among health care terminologies promoting interoperability. They are used to track and measure patient/client care holistically over time, across settings, population groups, and geographic locations. The CCC has open architecture and is specially designed for computer-based systems – EHR, CIS and PHR. Furthermore, CCC was tested as an international nursing standard based on the An Integrated Reference Terminology Model for Nursing, approved by the International Organization for Standardization (ISO/TC 215: Health Informatics) in October 2003. The computable structure of the CCC is protected under copyright permission.

5.2.2. American Medical Association's Current Procedural Terminology Codes (AMA CPT)

The Current Procedural Terminology (CPT®) [12] is a medical nomenclature used to report medical procedures and services under public and private health insurance programs. It describes medical, surgical, and diagnostic services and is designed to communicate uniform information about medical services and procedures among physicians, coders, patients, accreditation organizations, and payers for administrative, financial, and analytical purposes.

There are three types of CPT codes: Category I, Category II, and Category III. There are six main sections for Category I:

- Codes for Evaluation and Management
- Codes for Anaesthesia
- Codes for Surgery
- Codes for Radiology
- Codes for Pathology & Laboratory
- Codes for Medicine

Category II and III contain optional Codes for Performance Measurement and Emerging Technology respectively. Last, but not least it is necessary for users of the CPT code to pay license fees for access to the code.

5.2.3. Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT)

The Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) [13] is a clinical terminology, which has been promoted as a reference terminology for electronic health record (EHR) systems. Its purpose is to serve as a standardized terminology in healthcare software applications, as it enables clinicians, researchers and patients to share comparable data. SNOMED CT is owned, maintained and distributed by the International Health Terminology Standard Development Organization. It is open source and its current version was released in July 2011. SNOMED CT is used by the College of American Pathologists, the UMLS Metathesaurus, the European project epSOS and the European project SemanticHealthNet. SNOMED CT is designed to support translation. This multi-lingual resource is used in more than 50 countries. Available mapping to SNOMED CT exist with ICD-9-CM, ICD-03 and ICD-10.

SNOMED CT is the result of the combination of SNOMED Reference Terminology (SNOMED RT), developed by the College of American Pathologist, with the Clinical Terms Version 3 (CTV3), developed by the National Health Service of the United Kingdom. It consists of concepts, descriptions and relationships between concepts:

Concepts

- SNOMED CT concepts represent clinical ideas, ranging from abscess to zygote.
- Every concept has a unique numeric code known as the "concept identifier".
- Concepts are organized in hierarchies, from the general to the specific. This allows detailed clinical data to be recorded and later accessed or aggregated at a more general level.

Descriptions

- SNOMED CT descriptions link appropriate human-readable terms to concepts. A concept can have several associated descriptions, each representing a synonym that describes the same clinical idea.
- Each translation of SNOMED CT includes an additional set of descriptions, which link terms in another language to the same SNOMED CT concepts.

Relationships

- SNOMED CT relationships link each concept to other concepts that have a related meaning. These relationships provide formal definitions and other characteristics of the concept.
- One type of link is the "is a" relationship which relates a concept to its more general concepts. For example, the concept "viral pneumonia" has an "is a" relationship to the more general concept "pneumonia". These "is a" relationships define the hierarchy of SNOMED CT concepts.
- Other types of relationship represent other aspects of the definition of a concept. For example, the concept "viral pneumonia" has a "causative agent" relationship to the concept "virus" and a "finding site" relationship to the concept "lung".
- There are well over a million relationships in SNOMED CT.

Although SNOMED is widely used it is also criticized as having a vague domain [112].

5.2.4. Anatomical Therapeutic Chemical Classification System (ATC/DDD)

The ATC/DDD system[14] is an instrument for presenting active ingredient utilization statistics with the aim of improving drug use. The system is suitable for international comparisons of active ingredient utilization, for the evaluation of long term trends in drug use, for assessing the impact of certain events on drug use and for providing denominator data in investigations of drug safety.

- ATC - Anatomical Therapeutic Chemical (ATC) classification system

Within the ATC system active substances are divided into different groups according to the organ or system on which they act and their therapeutic, pharmacological and chemical properties. The drugs are classified in groups (five different levels).

- DDD - Defined Daily Dose

'The DDD is the assumed average maintenance dose per day for a drug used for its main indication in adults.'

The DDD will only be assigned for drugs that already have an ATC code.

Latest versions are obtainable at http://wido.de/amtl_atc-code.html in the file format of Excel (xls). Linkage to commercial drugs with brand or generic names can be realized using mapping tables.

The classification does not fulfil criteria for a semantically meaningful classification since the ATC classification is rather vague (Group V for example contains various entities i.e. allergens for hypersensitisation but also surgical dressings). Furthermore, the DDD is actually given without

consideration of personal background (age, gender, etc.). The purpose of this classifying is simply to determine a letter for a code. Hence, ATC with DDDs is best described as a coding system. So, it cannot be taken into consideration as a resource of knowledge representation or a foundation for automated reasoning.

5.2.5. MeSH

The Medical Subject Headings (MeSH) [16] are a medical thesaurus published and annually updated by the US National Library of Medicine (NLM). It is used for cataloging of the library holdings and for indexing of the databases that are produced by the NLM (e.g. MEDLINE).

It consists of sets of terms naming descriptors in a hierarchical structure that permits searching at various levels of specificity. MeSH descriptors are arranged in both an alphabetic and a hierarchical structure. At the most general levels of the hierarchical structure are very broad headings such as "Anatomy" or "Mental Disorders." More specific headings are found at more narrow levels of the twelve-level hierarchy, such as "Ankle" and "Conduct Disorder." There are 26,853 descriptors in 2013 MeSH. There are also over 199,000 entry terms that assist in finding the most appropriate MeSH Heading, for example, "Vitamin C" is an entry term to "Ascorbic Acid." In addition to these headings, there are more than 205,000 headings called Supplementary Concept Records (formerly Supplementary Chemical Records) within a separate thesaurus.

The MeSH thesaurus is used by NLM for indexing articles from 5,400 of the world's leading biomedical journals for the MEDLINE®/PubMED® database. It is also used for the NLM-produced database that includes cataloging of books, documents, and audiovisuals acquired by the Library. Each bibliographic reference is associated with a set of MeSH terms that describe the content of the item. Similarly, search queries use MeSH vocabulary to find items on a desired topic.

5.2.6. International Classification of Functioning, Disability and Health (ICF)

The International Classification of Functioning, Disability and Health, known more commonly as ICF [17], is a classification of health and health-related domains. These domains are classified from body, individual and societal perspectives by means of two lists: a list of body functions and structure, and a list of domains of activity and participation. Since an individual's functioning and disability occurs in a context, the ICF also includes a list of environmental factors. ICF is a WHO framework to measure health and disability at both individual and population levels.

ICF puts the notions of ‘health’ and ‘disability’ in a common understanding in acknowledging that every human being may experience a decrement in health and thereby experience some degree of disability. By shifting the focus from cause to impact it places all health conditions on an equal footing allowing them to be compared using a common metric – the ruler of health and disability. Furthermore, ICF takes into account the social aspects of disability and does not see disability only as a 'medical' or 'biological' dysfunction. By including Contextual Factors, in which environmental factors are listed, ICF allows to records the impact of the environment on the person's functioning.

However, the whole model suffers from shortcomings [31]. The classification is not coherent, as the criteria are sometimes based on the anatomic structure which has a function and sometimes on the process which is supported by the function. Another critical point is the “overemphasis on subsumptions”, i.e. the restriction to the is-a relation. Though the categories from ICF are

useful, one should put more effort in the definition the relations which hold between them and add more ontological power and expressivity.

ICF can be used online or file-based.

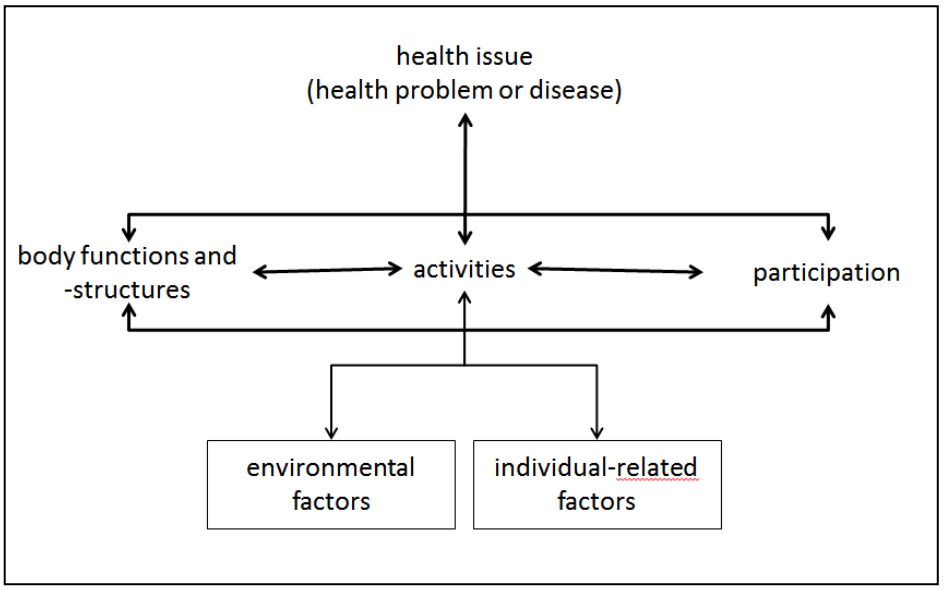


Figure 9 ICF as classification of components for health issues

Figure 10 View of ICF components

5.2.7. ICD-10

The International Classification of Diseases [23] is the world's standard tool to capture mortality and morbidity data. Generally speaking, the ICD contains information related to diagnoses, symptoms, abnormal laboratory findings, injuries and poisonings, external causes of morbidity and mortality, factors influencing health status from all the different branches of medicine: Oncology, Dentistry and Stomatology, Dermatology, Psychiatry, Neurology and so on.

The basic ICD-10 is a single coded list of three-character categories (from A00 to Z99), each of which can be further divided into up to ten four-character subcategories (for example, A00.0, A02.2, B51.9 and so on). Thus, a disease is related to three main data: namely Chapter, Block and Title. An example concept is shown in Figure 11.

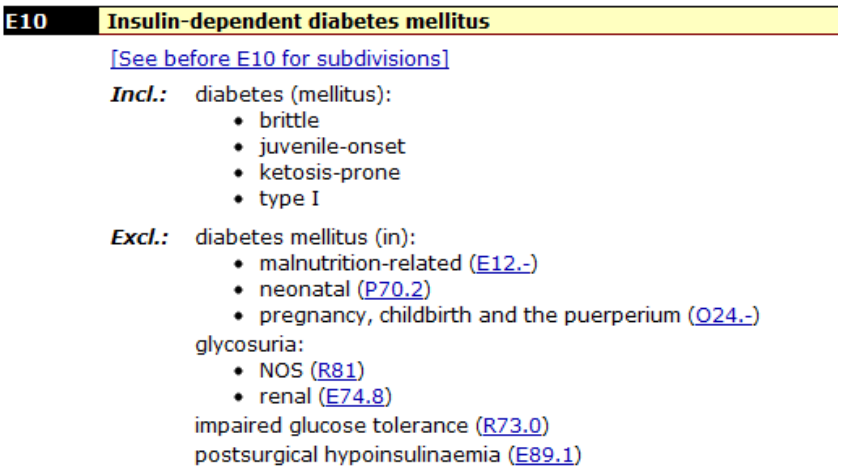


Figure 11 ICD-10 Concept E10

ICD-10 treats diseases as health problems which have been recorded, for example, on health records, or death certificates. However, there are several cases where the ontology is not coherent with wrong human labels.

5.2.8. Medical Directory for Regulatory Activities (MedDRA)

Medical Dictionary for Regulatory Activities [25] is a clinically validated international medical terminology for diagnoses, symptoms, surgeries, and other medical procedures. It is used by regulatory authorities and the regulated biopharmaceutical industry during the regulatory process, from pre-marketing to post-marketing activities, and for data entry, retrieval, evaluation, and presentation. In addition, it is the adverse event classification dictionary endorsed by the International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH). It is translated into several languages and according to the European Union, MedDra is the mandatory tool for coding and transmitting information on product characteristics and side-effects.

Since terms of MedDra are mandatory in the regulatory process, it has an important impact. However, it is not a useful tool for automated reasoning since clarification is missing for the relations which hold between the links to upper and lower term. It is not a real hierarchy and it is not a simple vocabulary with some links and connections which carry no semantic content at all. In that way MedDRA can be a useful controlled vocabulary but doesn't provide semantic relations.

5.3. Vocabularies and Thesauri

5.3.1. Glossary of Terms for Community Health Care and Services for Older Persons

The Ageing and Health Program of the WHO[20] Kobe Centre, in collaboration with the WHO Collaborating Centre for Population Ageing: Research, Education and Policy in Adelaide, Australia, initiated a project to develop an international glossary of terms applying to community health care and services for older persons through consultation with global experts, both via the Internet and in face-to-face meetings.

It aims to define and standardize the basic concepts and functions of community health care for older persons and organize them into a glossary, utilizing existing WHO definitions where appropriate, promote a common language for cross-program description and information dissemination.

The limitation of this glossary is that it just focuses on older persons. Furthermore, the semantics of this glossary are very poor.

5.3.2. Logical Observation Identifiers Names and Codes (LOINC)

LOINC [24] is a database and a universal standard for identifying medical laboratory and clinical observations. It was developed and maintained by the Regenstrief Institute. The LOINC vocabulary provides a set of universal names and ID codes for identifying laboratory and clinical test results in the context of existing HL7, ASTM E1238, and CEN TC251 observation report messages. The LOINC codes are mainly intended to identify test results and clinical observation. Other fields in the LOINC message can transmit, for example, the identity of the source laboratory or other special details about the sample. A formal, distinct and unique name (composed by six parts) is given to each LOINC component term.

The LOINC codes were released in April 1996 and, to date, thirteen revisions of LOINC, now including over 30,000 observation concepts, were released. LOINC contains fields for each of the six parts of the name, synonyms and comments for all observations in order to facilitate searches for individual laboratory test and clinical observation results. The database is dived into four categories: Lab, Clinical, Attachments, and Surveys. Such categories are not rigidly fixed and users can freely sort the database by whatever class is convenient in their application.

LOINC uses HL7 codes (see paragraph 6.2.5) for clinical documents aiming at avoiding the development of a new terminology. According to the LOINC Guide 2011, the component terms used in the creation of the names of document type codes will be mapped to either the UMLS Metathesaurus, or SNOMED CT as soon as possible.

5.3.3. Thesaurus of the National Cancer Institute (NCIT)

The Thesaurus of the National Cancer Institute (NCI) [26] covers vocabulary for clinical care, translational and basic research and public information and administrative activities. It can be browsed online and downloaded in OWL-DL or OBO format and currently contains over 34,000 concepts, structured into 20 taxonomic trees. The NCI Thesaurus provides concept history tables to record changes in the vocabulary over time as the science changes.

NCIT is published under an open content license. It covers a broad domain of entities which are related to cancer, e.g. in genetics, anatomy and medication. The vocabulary is related to some other terminologies. For example, the semantic type of the concepts from the UMLS Semantic Network is given. There exists also an NCI Metathesaurus which integrates terms from over 70 terminologies.

Although it lacks many qualities of a good ontology design, i.e. objective, descriptive definitions and a high level of formal exactness, it is easy to understand for human domain experts. According to the OBO foundry, it provides the most granular and consistent terminology available today.

6. Conceptualization & Implementation

In this section, we focus on conceptualization of the semantic model and on the corresponding implementation.

The development of the BOUNCE Semantic model will be based on the following three principles:

- **Reuse**: Avoid "reinventing the wheel" and reuse already established high quality ontologies.
- **Granularity**: Annotations or mappings cannot be extracted from a single ontological resource. So, multiple ontologies should be used.
- **Modularity**: Create a framework where different ontologies would be able to integrate many modules through mappings between ontologies.

Ontology construction is deemed to be a labour-intensive and a time-consuming process [29]. In addition, the development of new ontologies does not necessarily tap the full potential of existing domain-relevant knowledge sources. Due to these problems the latest years the tendency is not to create new ontologies from scratch but to try to integrate high quality, domain-specific ontologies that have already proven their value.

Projects like INTEGRATE[7] relied on a Common Information Model (CIM) based on HL7 and SNOMED-CT to represent the information on cancer domain, p-Medicine[8] used an ontology called HDOT to integrate already existing and well found ontologies, VPH NoE[9] identified mappings between different ontologies to enable interoperability and eHealthMonitor[10], MyHealthAvatar and iManageCancer extended an ontology named TMO to allow integrating several well-known ontologies the latest version of which is known as the iManageCancer Semantic Core Ontology. As consortium members have already extensive experiences in reusing and extending the specific ontology, in the sequel we will first describe the various modules available in the iManageCancer Semantic Core Ontology that will be reused for modeling demographics, clinical, biological and lifestyle data. As psychological data that the BOUNCE project will collect cannot be covered by existing ontologies, we report them on a unique module developed for modeling psychological data, the BOUNCE psychological ontology.

6.1. The iManageCancer Semantic Core Ontology

For enabling a common representation of knowledge across the continuum of care and across the different information sources, the iManageCancer project[11] developed the iManageCancer Semantic Core ontology. It is used as the virtual schema of all data stored within the platform of the project, and is able to semantically describe the different types of data required and processed by the platform.

The development of the iMC Semantic Core Ontology was based besides the principles of *reuse, granularity,* and *modularity* on the principle of *multilinguality* as well since the data management

[7] http://www.fp7-integrate.eu/

[8] http://www.p-medicine.eu/

[9] http://www.vph-noe.eu/

[10] http://www.ehealthmonitor.eu/

[11] http://imanagecancer.eu/

layer of the project is used to store medical data in three European countries (United Kingdom, Italy and Germany). Although more than one country is also participating in the BOUNCE consortium, according to the data currently available and examined, multilingualism in the ontology level is not required.

Figure 12. The modules of iMC Semantic Core Ontology[12] and the BOUNCE Psychological Ontology Module (BPO)

The ontology contains 36 sub-ontologies integrated using an extension of the Translational Medicine Ontology [32] which is used as an upper layer ontology. An overview of the different modules of the iMC Semantic Core Ontology is shown in Figure 12 (the blue and the green ones).

The Extended TMO is an OWL compliant ontology and consists of about 10000 triples. In those triples we have 329 classes and 38 properties that represent the following entities relevant to biomedical studies:

- Materials: e.g. molecule, protein, cell lines, pharmaceutical preparations
- Processes: e.g. diagnosis, study, intervention
- Roles: e.g. subject target, active ingredient
- Informational Entities: e.g. dosage, mechanism of action, sign/symptom, family history

[12] ACGT: ACGT Master Ontology, BFO: Basic Formal Ontology, CHEBI: Chemical Entities of Biological Interest, CIDOC-CRM: CIDOC Conceptual Reference Model, CTO: Clinical Trial Ontology, DO: Human Disease Ontology, DTO: Disease Treatment Ontology, FHHO: Family Health History Ontology, FMA: Foundation Model of Anatomy, FOAF: Friend of a Friend Ontology, GALEN: Galen Ontology, GO: Gene Ontology, GRO: Gene Regulation Ontology, HDOT: Health Trunk Ontology, IAO: Information Artifact Ontology, ICD: International Classification of Diseases, ICO: Informed Consent Ontology, LOINC: Logical Observation Identifier Names and Codes, MESH: Medical Subject Headings, NCI-T: NCI theraurus, NIFSTD: Neuroscience Information Framework Standardized ontology, NNEW: New Weather Ontology, OBI: Ontology for Biomedical Investigation, OCRE: Ontology for Clinical Research, OMRSE: Ontology of Medically Related Social Entities, PATO: Phenotypic Quality Ontology, PLACE: Place Ontology, PRO: Protein Ontology, RO: Relation Ontology, SBO: Systems Biology Ontology, SNOMED-CT: SNOMED clinical terms, SO:Sequence Ontology, SYMP: Symptom Ontology, TIME: Time Ontology, UMLS: Unified Modeling Language System, HDOT: Health Data Ontolog Trung.

By contrast, particulars (e.g. "a patient with a given name" and "a blister package of a pharmaceutical product with a particular identifying code on it") refer to individuals and are represented as instances of classes in the ontology. Consequently, a particular consultation at a given time and day, the particular patient role in that consultation, and the physician role in that consultation can be represented as instances of classes in the ontology.

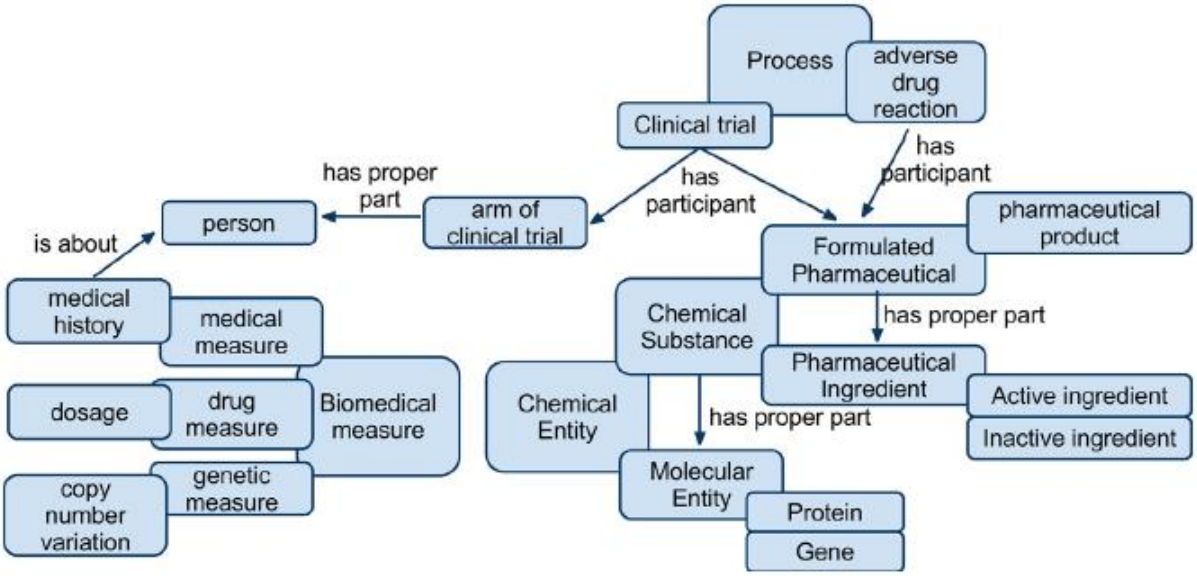


Figure 13. Overview of selected types, subtypes (overlap) and existential restrictions (arrows) in the TMO as presented in [32]

Figure 13 demonstrates a portion of the TMO and illustrates selected types, subtypes, and existential restrictions that hold between types. The TMO extends the basic types defined in the Basic Formal Ontology[13] (BFO) and uses relations from the Relation Ontology[14] (RO). Moreover, it uses the Information Artifact Ontology[15] (IAO) as well.

All other ontologies are integrated using the TMO ontology on top. Among the added ontologies are HDOT, a cancer specific ontology developed within the p-Medicine project. In addition, we foresee that in the data level (not the ontological level) the multilingual versions of the ICD-10 will be used for capturing the classification of diseases in the countries that participate in the BOUNCE project. The integration is achieved by introducing terms from these sub-ontologies to the TMO ontology and via relations of equivalence and subsumption from eTMO to the various ontology modules. These relations (~400) were manually identified and verified using the NCBO BioPortal[16].

In the sequel we will provide some examples of modelling socio-demographic information and medical clinical using existing ontologies and schemas from the iMC Semantic Core Ontology and we focus on the BOUNCE Psychological Ontology a module created especially for modelling the psychological data available within the BOUNCE project.

[13] http://www.ifomis.org/bfo

[14] http://www.obofoundry.org/ro/

[15] http://code.google.com/p/information-artifact-ontology/

[16] http://bioportal.bioontology.org/

6.2. Socio-Demographics and Medical/Clinical model

As multiple ontologies are available within the iManageCancer Semantic Core ontology in this subsection we present some examples on how socio-demographic and medical/clinical information can be modeled.
For example, within the HDOT ontology there are the appropriate structure for modeling disease phenotypes (Figure 14) and physical object qualities such as laboratory results, weight, mass, height etc. (Figure 15). Details of clinical trials are detailed from the CTO ontology (Figure 16) whereas family history can be recorded using the FHHO ontology (Figure 17).

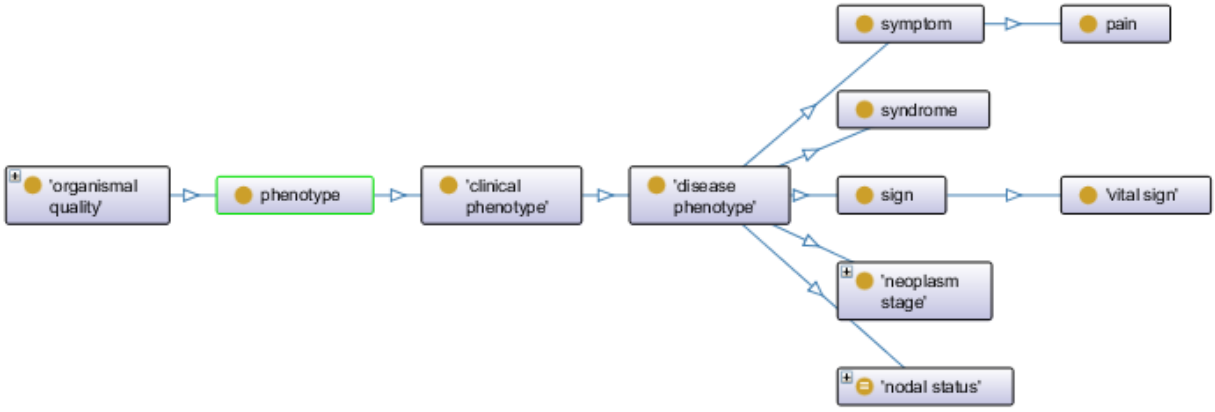


Figure 14. Example hierarchy for disease phenotype from HDOT ontology

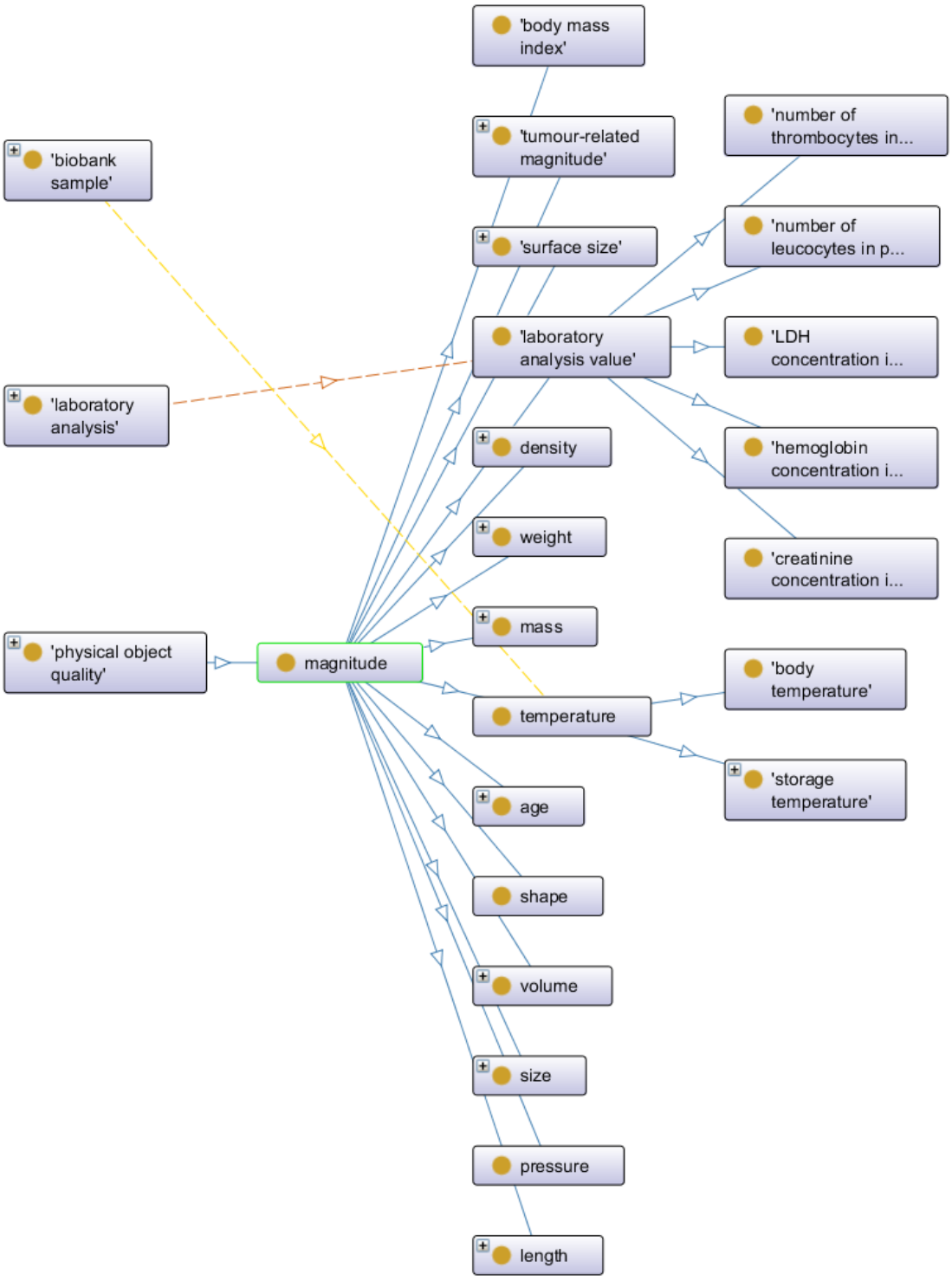


Figure 15. Physical object qualities

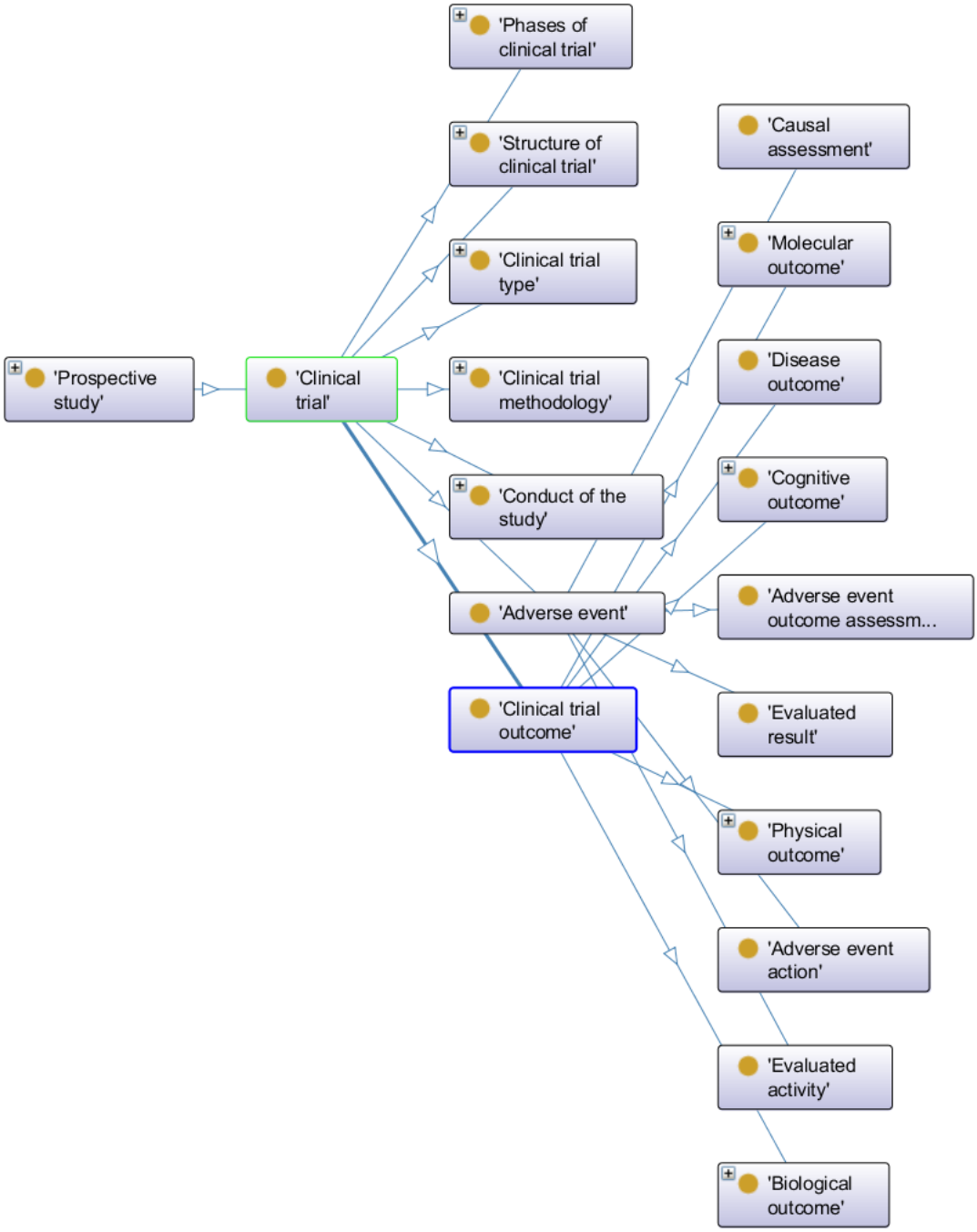


Figure 16. Clinical trial description within the CTO ontology.

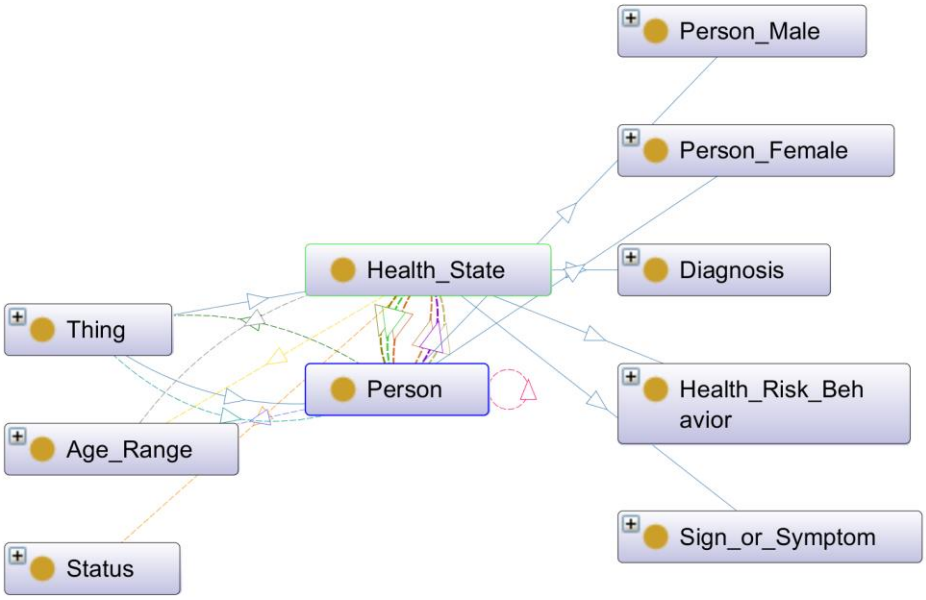


Figure 17. Snapshot of classes around the Health State class in FHHO

6.3. The BOUNCE psychological ontology

The high level of the entities available within the BOUNCE psychological ontology is shown in Figure 18. According to the diagram, *Trait* relates to *Coherence*, *Coherence* can predict *Trauma* and *Trauma* on the other hand affects *Coherence. Illness related events* affect both *Trauma* and *Coherence* and the same applies for *Negative life events*. *Negative life events* affect *cognitive and emotional representation of illness* on the other hand and determine *Coping with illness*. *Self-efficacy for coping with illness* predicts *coping with illness* and is determined by *emotional and cognitive representation of illness*. *Resilience* regulates the relationship between *Distress*, *Anxiety* and *Fear of Recurrence* and *Fear of recurrence* predicts *Depression*, *Anxiety* and *Distress*.

We have to note that we treat the relation "affects" and "predicts" interchangeable although in reality they are not the exactly the same ("predicts" does not imply cause and effect, whereas "affects" does). We expect that as soon as we have results from the models we will be able to make the correct distinction among these terms in the final version of the model.

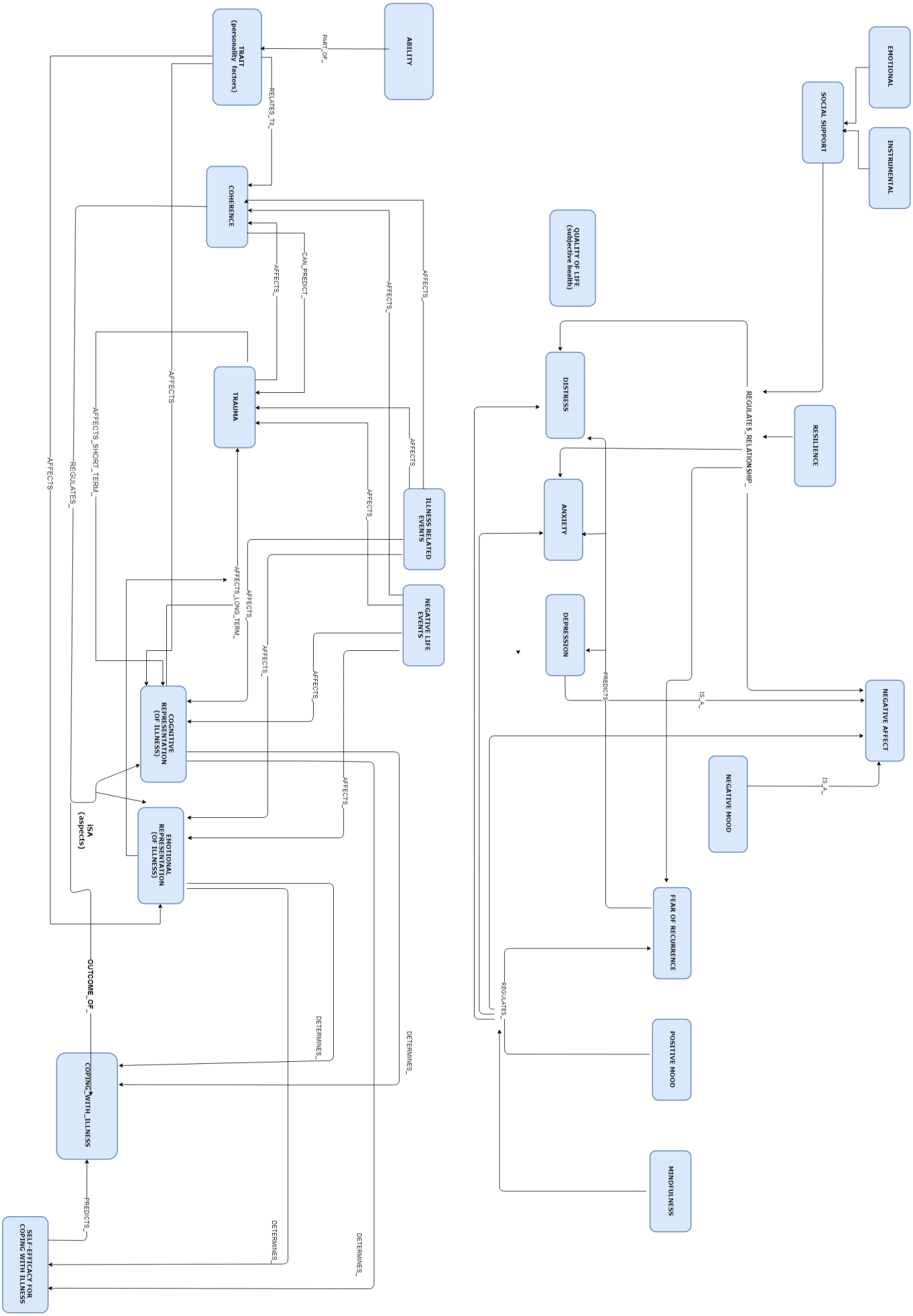

ABILITY
TRAIT (personality factors)
PART_OF
RELATES_TO
COHERENCE
CAN_PREDICT
AFFECTS
TRAUMA
AFFECTS_SHORT_TERM
AFFECTS_LONG_TERM
ILLNESS RELATED EVENTS
NEGATIVE LIFE EVENTS
REGULATES
COGNITIVE REPRESENTATION (OF ILLNESS)
EMOTIONAL REPRESENTATION (OF ILLNESS)
ISA (aspects)
OUTCOME_OF
COPING_WITH_ILLNESS
PREDICTS
SELF-EFFICACY FOR COPING WITH ILLNESS
DETERMINES
EMOTIONAL
INSTRUMENTAL
SOCIAL SUPPORT
QUALITY OF LIFE (subjective health)
DISTRESS
RESILIENCE
REGULATES_RELATIONSHIP
ANXIETY
DEPRESSION
PREDICTS
ISA
NEGATIVE AFFECT
NEGATIVE MOOD
FEAR OF RECURRENCE
POSITIVE MOOD
MINDFULNESS
REGULATES

Figure 18. The top level of the entities available in BOUNCE

In the following images, the high level of the entities available within the BOUNCE psychological ontology will be described in more details.

Figure 19 shows that social support entity has two parts: emotional and instrumental support. The emotional support can be measured by: 1) emotional social support and 2) a single item with a single score of perceived emotional support. The instrumental support is measured by instrumental social support entity.

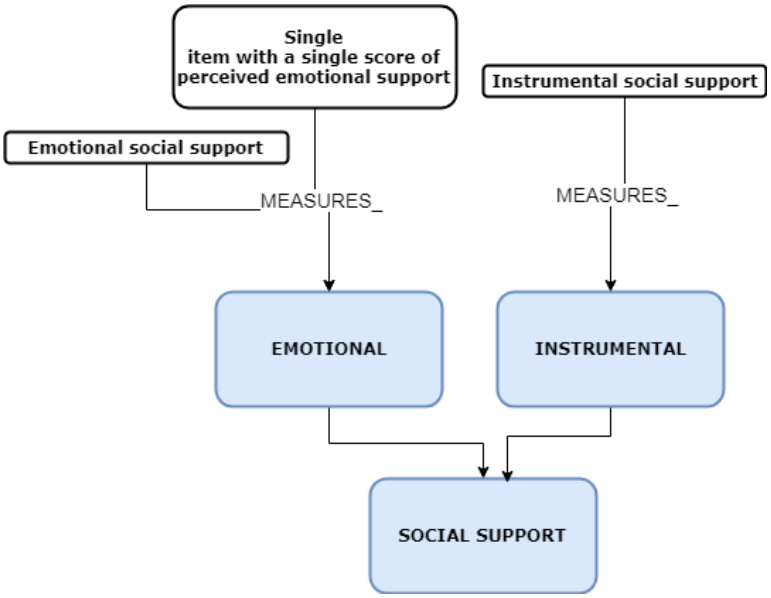


Figure 19 Description of SOCIAL SUPPORT entity and it's relevant entities.

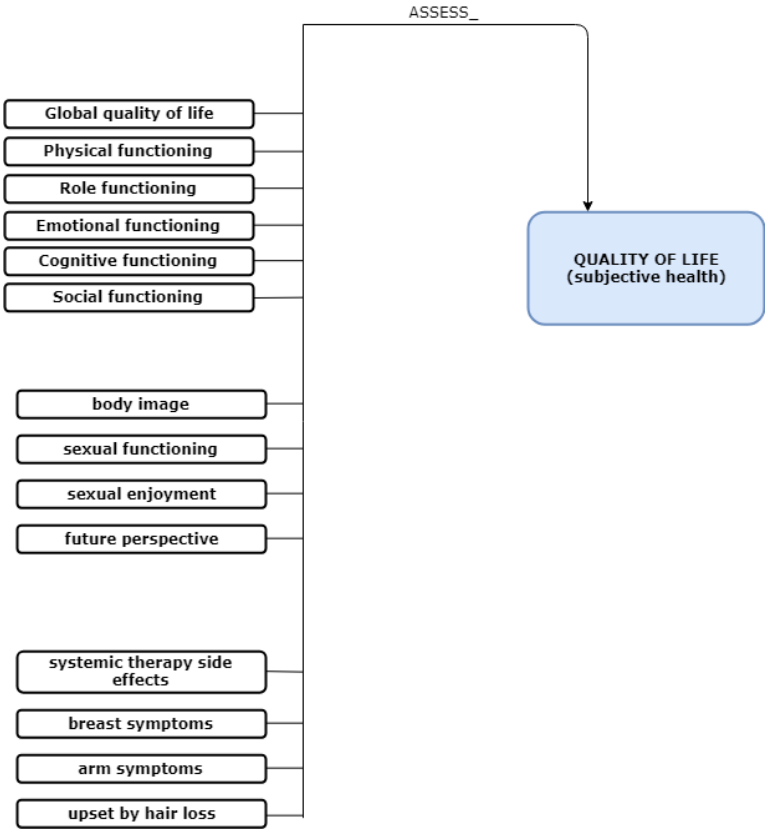


Figure 20 Description of QUALITY OF LIFE(subjective health) entity and it's relevant entities.

Several entities described in Figure 20 can assess the quality of life like the global quality of life, the physical functioning, the role functioning, the emotional functioning, the cognitive functioning, the social functioning, the body image, the sexual functioning, the sexual enjoyment, the future perspective, the systemic therapy side effects, the breast symptoms, the arm symptoms and the upset caused by hair loss.

Next in Figure 21 we show that Illness representation has two parts: the cognitive and the emotional. Each part of the representation has several aspects as shown in Figure 21.

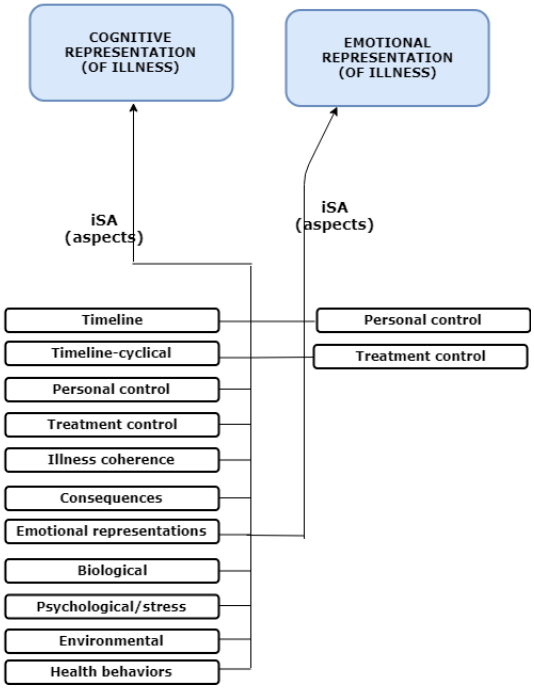


Figure 21 Description of ILLNESS REPRESENTATION entities and their relevant entities.

As described in Figure 22, self-efficacy for coping with illness can predict coping with illness and can be measured by 1) cbi-b cancer behaviour inventory score and 2) a general self-efficacy item. Entities such as self-blame, acceptance, rumination etc. are general behaviours, which can identify the strategy of coping with illness. These general behaviours can predict the specific behaviours, such as helplessness/hopelessness, anxious preoccupation etc. The specific behaviours are strategies of coping with illness.

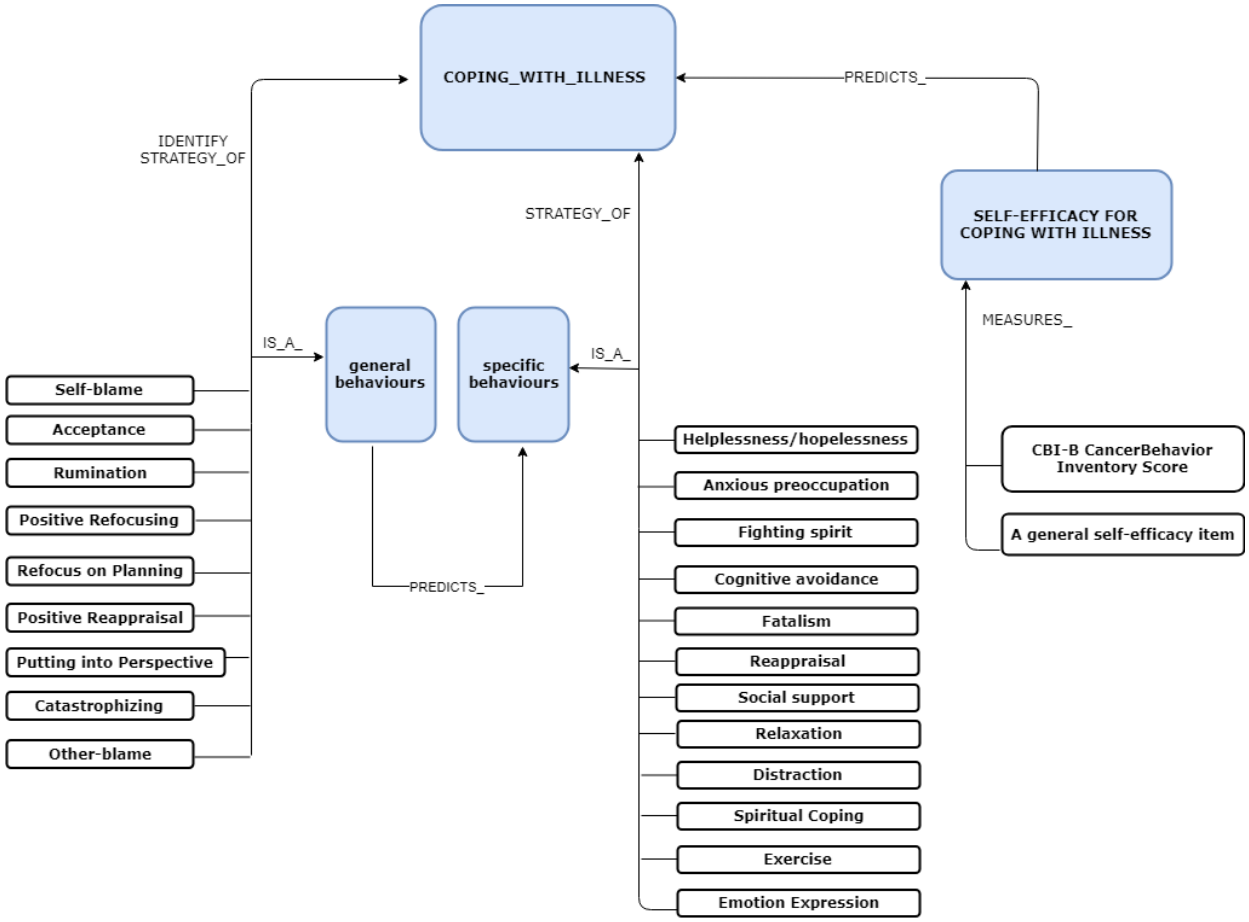


Figure 22 Description of the enity COPING WITH ILLNESS and the relevant entities.

Trait, shown in Figure 23, is an entity that is related to coherence and can be described by several entities like, conscientiousness, agreeableness etc. Coherence is measured by meaningfulness, comprehensibility and manageability.

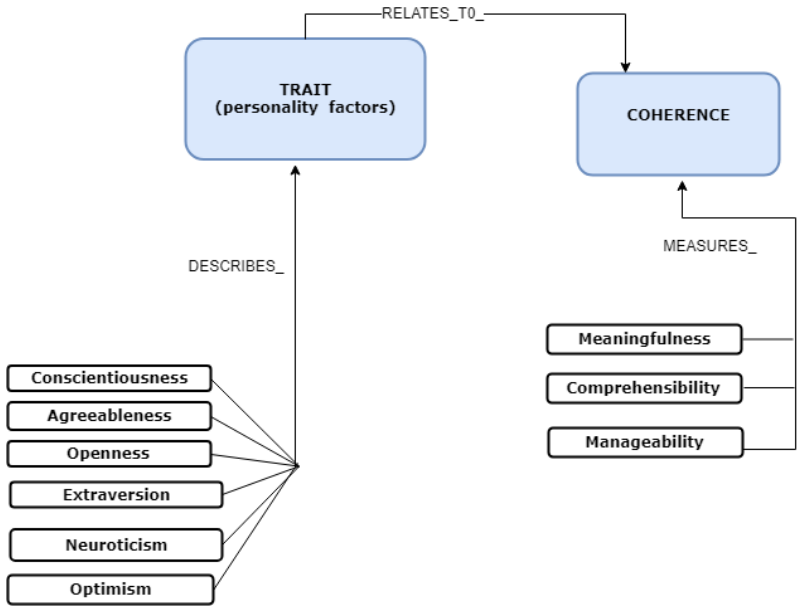


Figure 23 Description of TRAIT and relevant entities.

Next, qualitative questions measure the illness related events and the negative life events. Those are shown in Figure 24.

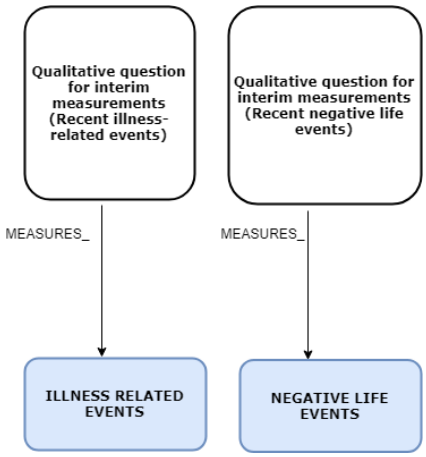


Figure 24 Description of EVENTS :Illness related and negative.

Distress, anxiety and depression are measured by various scores as shown in Figure 25.

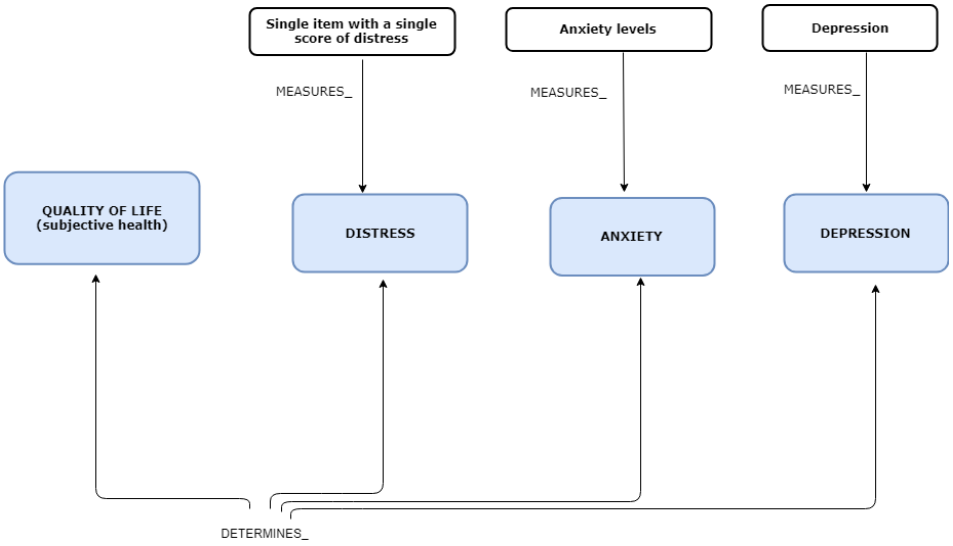


Figure 25 Description of DISTRESS, ANXIETY and DEPRESSION.

Other relevant psychological measures evaluate negative mood, fear of occurrence and positive mood as shown in Figure 19. Mindfulness within BOUNCE will be assessed by two different scores as shown.

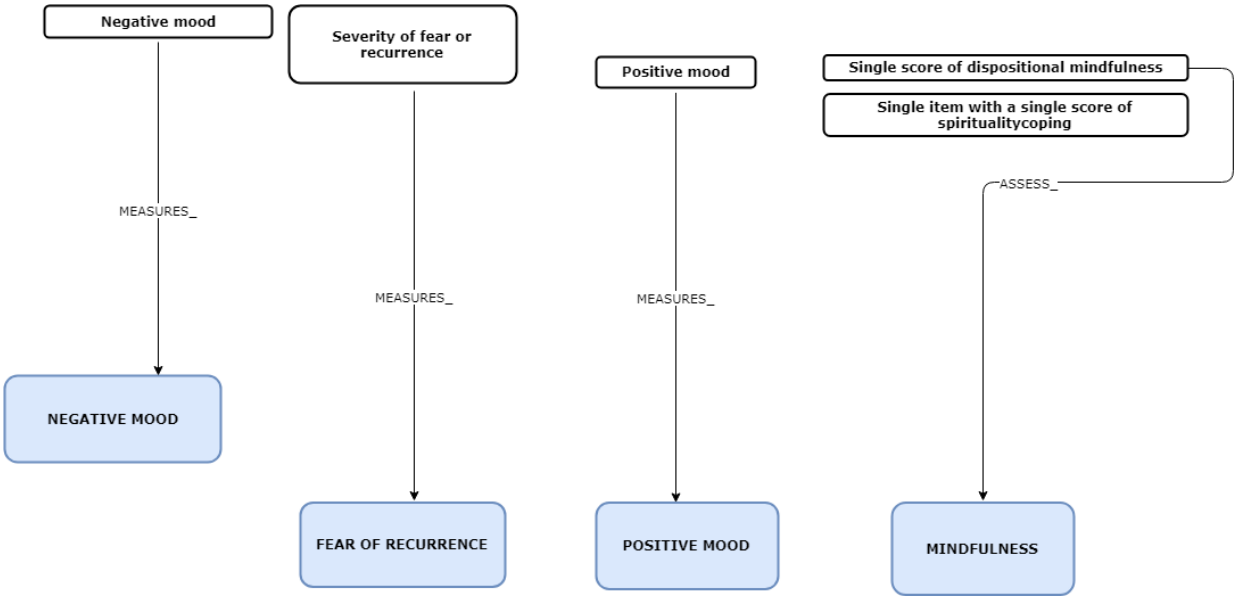


Figure 26 Description of NEGATIVE MOOD, FEAR OF OCCURRENCE, POSITIVE MOOD and MINDFULNESS.

We have to note that the entire process modelled using BPO is based on continuous feedback loops. That is, even though we can assert, for instance, that Cognitive and Emotional representation of illness "predict" Coping which in turn "predicts" Outcomes (e.g., quality of life), which is accurate. But in the long term, outcomes also predict representations and coping behavior: a better outcome reinforces existing representations and coping behaviors, whereas a worse outcome may lead to a change in representations and coping behaviors.

The full picture of the BPO is shown in the Appendix 1. The various entities presented in this section have been implemented in a self-contained ontology, the BOUNCE Psychological Ontology formulating a novel module of the iManageCancer Semantic Core Ontology. The BPO module has been developed using PROTÉGÉ and is an RDF ontology. Soon it will be released in an github repository.

7. Conclusions

In this deliverable, we focused on describing the methodology followed for developing the first version of the BOUNCE semantic model. First, we reported the available data fields to be modelled. Then, in the knowledge acquisition phase, we collected, presented and reviewed relevant ontologies from the cancer domain. Although we were able to identify multiple ontologies covering the sociodemographic and the clinical data of the BOUNCE project, we did not identify an appropriate ontology for psychological data. In the conceptualization phase we selected a modular ontology, the iManageCancer Semantic Core Ontology, for modelling sociodemographic and clinical data, and we progressed further in implementing a novel module for modelling the psychological data available within the project.

The ontology developed will be used to generate mappings to data sources (prospective, retrospective and external data) in order to be subsequently homogenized, integrated and semantically uplifted. The results of this process will be described in D3.4 Solutions for Data Aggregation, Cleaning, Harmonization & Storage. However, ontologies are not static artefacts, but subject to continuous change. As such we expect that new terms will be identified while trying to generate the corresponding mappings or that the existing modelling constructs might have to be updated. The final version of the overall BOUNCE semantic model will be provided in M24 by D3.3 Final Semantic Model.

8. References

[1] Clinical Care Classification System. Available Online:http://www.sabacare.com/

[2] Collier, N., Matsuda Goodwin, R., McCrae, J., Doan, S., Kawazoe, A., Conway, M., Kawtrakul, A., Takeuchi, K. and Dien, D.: An ontology-driven system for detecting global health events, Proc. 23rd International Conference on Computational Linguistics (COLING), pp.215-222, 2010.

[3] Corcho, O., Fernandez-Lopez, M., Gomez-Perez, A.: Methodologies, tools and languages for building ontologies: Where is their meeting point? Data Knowl. Eng., 46, pp 41–64, July 2003.

[4] Fernández-López, M.: Overview of Methodologies for Building Ontologies. In: The IJCAI Workshop on Ontologies and Problem-Solving Methods, Stockholm, Sweden, 1999.

[5] Fernandez-Lopez M., Gomez-Perez A., Juristo N.: METHONTOLOGY: from Ontological Art towards Ontological Engineering, Proceedings of the AAAI97 Spring Symposium, Stanford, USA , pp. 33 - 40, 1997.

[6] Foundational model of Anatomy. Available Online: http://sig.biostr.washington.edu/projects/fm/AboutFM.html

[7] Genitsaridi, I., Marias, K., Tsiknakis, M.: An ontological approach towards psychological profiling of breast cancer patients in pervasive computing environments Proceedings of the 8th ACM International Conference, 2015

[8] Grüninger M., Fox, M. S.: Methodology for the Design and Evaluation of Ontologies, Workshop on Basic Ontological Issues in Knowledge Sharing, IJCAI-95, Montreal, 1995.

[9] Gomez-Perez, A., Fernandez, M. and De Vicente, A.J.: Towards a Method to Conceptualize Domain Ontologies. In: ECAI-96 Workshop on Ontological engineering, Budapest (1996).

[10] He, Y., Xiang, Z., Sarntivijai, S., Toldo, L., Ceusters W.: AEO: A Realism-Based Biomedical Ontology for the Representation of Adverse Events, International Conference on Biomedical Ontology • Buffalo, NY, USA Representing Adverse Events Workshop, July 26, 2011.

[11] Hovy, E.: Methodologies for the Reliable Construction of Ontological Knowledge, In Proceedings of ICCS, pp. 91-106,Springer 2005.

[12] http://www.ama-assn.org/ama/pub/physician-resources/solutions-managing-your-practice/coding-billing-insurance/cpt.page

[13] https://www.snomed.org/

[14] http://www.dimdi.de/static/de/klassi/atcddd/index.htm

[15] http://www.nlm.nih.gov/research/umls/

[16] http://www.ncbi.nlm.nih.gov/mesh

[17] http://www.who.int/classifications/icf/en/

[18] http://omrse.googlecode.com/svn/trunk/omrse/omrse.owl

[19] http://neuinfo.org

[20] http://www.who.int/

[21] www.eu-acgt.org/

[22] http://www.ebi.ac.uk/sbo/main/

[23] http://www.who.int/classifications/icd/en/

[24] http://en.wikipedia.org/wiki/LOINC

[25] http://www.meddra.org/

[26] http://ncit.nci.nih.gov/

[27] https://bioportal.bioontology.org/ontologies/MF

[28] https://bioportal.bioontology.org/ontologies/MFOEM

[29] Kondylakis, H., Dimitris, P.: Exelixis: Evolving Ontology-Based Data Integration System. SIGMOD, pp. 1283-1286 (2011)

[30] Kondylakis, H., Spanakis, E.G., Sfakianakis S., et al.: Digital Patient: Personalized and Translational Data Management through the MyHealthAvatar EU Project, EMBC 2015.

[31] Kumar, A., Smith, B.: The Ontology of Processes and Functions. A Study of the International Classification of Functioning, Disability and Health, http://ontology.buffalo.edu/medo/ICF.pdf

[32] Luciano, J.S., Joanne S., et al.:The Translational Medicine Ontology and Knowledge Base: Driving Personalized Medicine by Bridging the Gap between Bench and Bedside, Journal of Biomedical Semantics 2.Suppl 2 (2011): S1. 2015.

[33] Malone, J., Holloway, E., Adamusiak, T., Kapushesky, M., Zheng, J., Kolesnikov, N., Zhukova, A., Brazma, A., Parkinson, H.: Modeling sample variables with an Experimental Factor Ontology, Bioinformatics, 26(8), pp. 1112–1118, 2010.

[34] Obo Foundry: Available Onlinehttp://www.obofoundry.org/cgi-bin/detail.cgi?id=disease_ontology

[35] Peace, J, Brennan, P.F.: Ontological representation of family and family history, at AMIA Annu Symp Proc. 2007.

[36] Sanfilippo E. M., Schwarz U., Schneider L., The Health Data Ontology Trunk (HDOT). Towards an ontological representation of cancer-related knowledge.

[37] Schreiber, A. Th., Terpstra P.: Sisyohus-VT: A CommonKADS solution. Technical Report, ESPRIT Project 8145 KACTUS, University of Amsterdam, 1995. Submitted for publication.

[38] Schreiber, A. Th., Wielinga, B. J., Jansweijer, W. H.: The KACTUS view on the 'O' word. Technical Report, ESPRIT Project 8145 KACTUS, University of Amsterdam, 1995.

[39] SemanticHEALTH Report, 2009, Available online: http://www.eurorec.org/files/filesPublic/2009semantic-health-report.pdf

[40] Staab, S., Studer, R.: Handbook on Ontologies, Springer-Verlag 2004, http://books.google.gr/books?id=0Elgz95mM8QC.

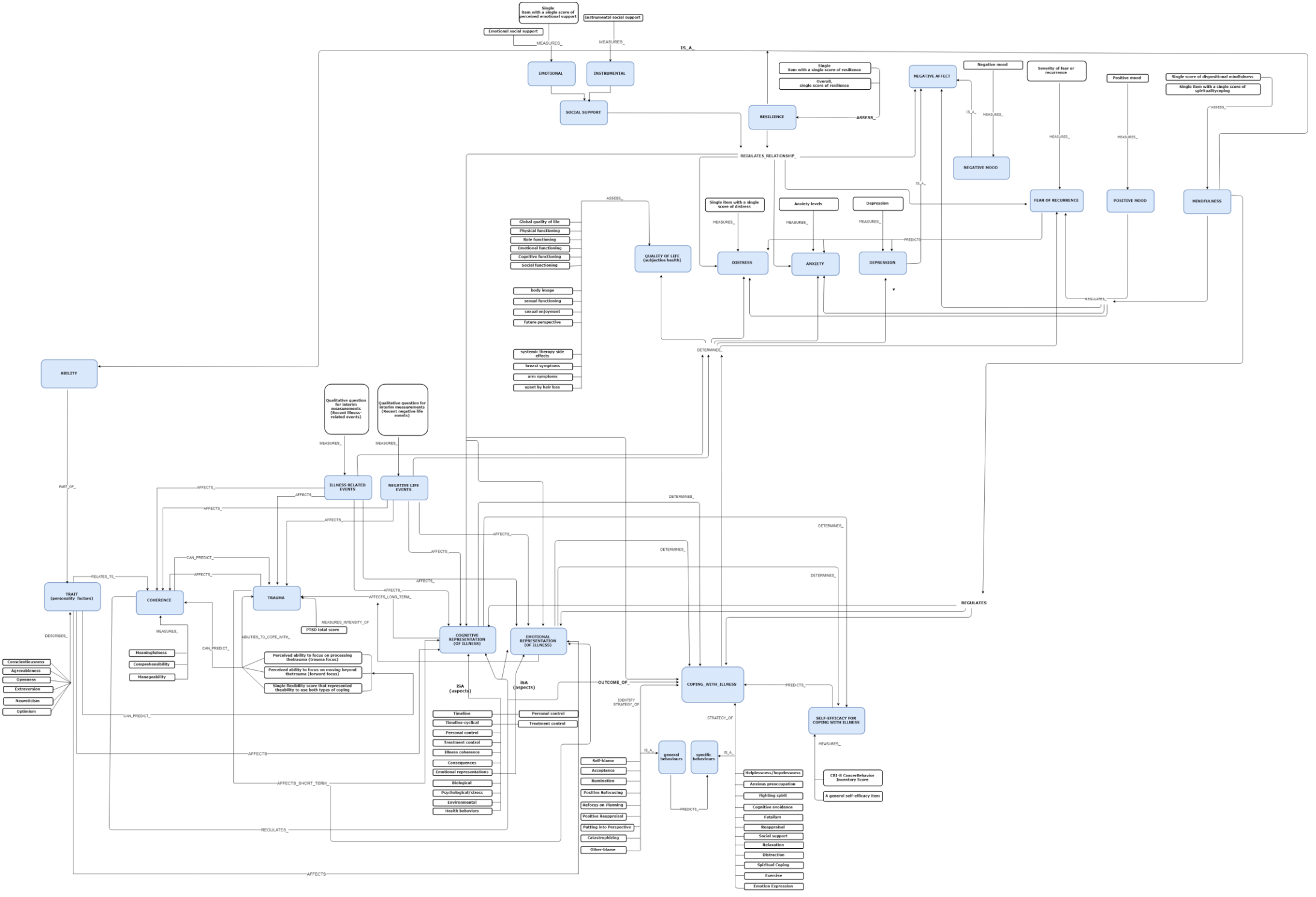
[41] Symptom Ontology Wiki: Available Online: http://symptomontologywiki.igs.umaryland.edu/wiki/index.php/Main_Page

[42] Uschold, M.: Building Ontologies: Towards a unified methodology. InI. Watson (Ed.), 16th Annual Conference of the British Computer Society Specialist Group on Expert Systems, Cambridge, UK,1996.

[43] Uschold, M., Gruninger, M.: Ontologies: Principles, methods and applications. The Knowledge Engineering Review, 11(2) ,1996.

[44] Uschold M., King, M.: Towards a Methodology for Building Ontologies, Workshop on Basic Ontological Issues in Knowledge Sharing, held in conduction with IJCAI-95,1995.

Appendix 1


Single item with a single score of perceived emotional support
Instrumental social support
Emotional social support
MEASURES_
IS_A_
EMOTIONAL
INSTRUMENTAL
SOCIAL SUPPORT
RESILIENCE
Single item with a single score of resilience
Overall, single score of resilience
ASSESS_
NEGATIVE AFFECT
Negative mood
Severity of fear or recurrence
Positive mood
Single score of dispositional mindfulness
Single item with a single score of spiritualitycoping
NEGATIVE MOOD
REGULATES_RELATIONSHIP_
FEAR OF RECURRENCE
POSITIVE MOOD
MINDFULNESS
Single item with a single score of distress
Anxiety levels
Depression
PREDICTS_
Global quality of life
Physical functioning
Role functioning
Emotional functioning
Cognitive functioning
Social functioning
QUALITY OF LIFE (subjective health)
DISTRESS
ANXIETY
DEPRESSION
body image
sexual functioning
sexual enjoyment
future perspective
systemic therapy side effects
breast symptoms
arm symptoms
upset by hair loss
REGULATES_
DETERMINES_
ABILITY
Qualitative question for interim measurements (Recent illness-related events)
Qualitative question for interim measurements (Recent negative life events)
ILLNESS RELATED EVENTS
NEGATIVE LIFE EVENTS
PART_OF_
AFFECTS_
CAN_PREDICT_
RELATES_TO_
TRAIT (personality factors)
COHERENCE
TRAUMA
AFFECTS_LONG_TERM_
REGULATES
DESCRIBES_
MEASURES_
MEASURES_INTENSITY_OF
PTSD total score
ABILITIES_TO_COPE_WITH_
Meaningfulness
Comprehensibility
Manageability
Perceived ability to focus on processing thetrauma (trauma focus)
Perceived ability to focus on moving beyond thetrauma (forward focus)
Single flexibility score that represented theability to use both types of coping
Conscientiousness
Agreeableness
Openness
Extraversion
Neuroticism
Optimism
COGNITIVE REPRESENTATION (OF ILLNESS)
EMOTIONAL REPRESENTATION (OF ILLNESS)
isA (aspects)
Timeline
Timeline-cyclical
Personal control
Treatment control
Illness coherence
Consequences
Emotional representations
Biological
Psychological/stress
Environmental
Health behaviors
Personal control
Treatment control
OUTCOME_OF_
IDENTIFY STRATEGY_OF
COPING_WITH_ILLNESS
PREDICTS_
STRATEGY_OF
SELF-EFFICACY FOR COPING WITH ILLNESS
general behaviours
specific behaviours
Self-blame
Acceptance
Rumination
Positive Refocusing
Refocus on Planning
Positive Reappraisal
Putting into Perspective
Catastrophizing
Other-blame
Helplessness/hopelessness
Anxious preoccupation
Fighting spirit
Cognitive avoidance
Fatalism
Reappraisal
Social support
Relaxation
Distraction
Spiritual Coping
Exercise
Emotion Expression
CBI-B CancerBehavior Inventory Score
A general self-efficacy item