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:1.3 BOUNCE methodology

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

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

Responsible WP: FORTH

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

0. Document Info

0.1. Author

Author	Company	E-mail
Haridimos Kondylakis	FORTH	kondylak@ics.forth.gr
Lefteris Koumakis	FORTH	koumakis@ics.forth.gr
Kostas Marias	FORTH	kmarias@ics.forth.gr
Akis Simos	FORTH	akis.simos@gmail.com
Evangelos Karadimas	FORTH	karademas@uoc.gr
Eleftherios Ouzounoglou	ICCS	elouzou@central.ntua.gr, eleftherios.ouzounoglou@iccs.gr
Georgios Stamatakos	ICCS	gestam@central.ntua.gr
Georgios Manikis	FORTH	gmanikis@gmail.com
Ketti Mazzocco	IEO	ketti.mazzocco@unimi.it
Flavia Faccio	IEO	Flavia.Faccio@ieo.it
Virginia Sanchini	IEO	virginiasanchini@gmail.com
Pasi Heiskanen	Noona	pasi.heiskanen@noona.com
Konstadina Kourou	FORTH	konstadina.kourou@gmail.com
Kostas Perakis	SiLo	kperakis@ep.singularlogic.eu
Gianna Tsakou	SiLo	gtsakou@singularlogic.eu
Poikonen-Saksela Paula	HUS	paula.poikonen-saksela@hus.fi
Ruth Pat Horenczyk	HUJI	ruth.pat-horenczyk@mail.huji.ac.il
Berta Sousa	CHAMP	berta.sousa@fundacaochampalimaud.pt
Eleni Kolokotroni	ICCS	ekolok@mail.ntua.gr
Katerina Argyri	ICCS	kargyri@mail.ntua.gr

0.2. Documents history

Document version #	Date	Change
V0.1	01 May 2018	Starting version, template
V0.2	01 May 2018	Definition of ToC
V0.3	14 June 2018	First complete draft
V0.4	01 July 2018	Integrated version (send to WP members)
V0.5	10 July 2018	Updated version (send PCP)
V0.6	10 July 2018	Updated version (send to project internal reviewers)
Sign off	20 July 2018	Signed off version (for approval to PMT members)
V1.0	29 July 2018	Approved Version to be submitted to EU

0.3. Document data

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

1. Table of Contents

2. Introduction

2.1. About the project

Coping with breast cancer more and more becomes a major socio-economic challenge not least due to its constantly increasing incidence in the developing world. There is a growing need for novel strategies to improve understanding and capacity to predict resilience of women to the variety of stressful experiences and practical challenges related to breast cancer. This is a necessary step toward efficient recovery through personalized interventions. BOUNCE will bring together modelling, medical, and social sciences experts to advance current knowledge on the dynamic nature of resilience as it relates to efficient recovery from breast cancer. BOUNCE will take into consideration clinical, cancer-related biological, lifestyle, and psychosocial parameters in order to predict individual resilience trajectories throughout the cancer continuum with the aim to eventually increase resilience in breast cancer survivors, help them remain in the workforce and enjoy a better quality of life.

BOUNCE will deliver a unified clinical model of modifiable factors associated with optimal disease outcomes and will deploy a prospective multi-centre clinical pilot at four major oncology centres (in Italy, Finland, Israel and Portugal), where an estimate of 660 women will be recruited in order to assess its clinical validity against crucial patient outcomes (illness progression, wellbeing, and functionality). The advanced computational tools to be employed will validate indices of patients' capacity to bounce back during the highly stressful treatment and recovery period following diagnosis of breast cancer. The overreaching goal of BOUNCE is to incorporate elements of a dynamic, predictive model of patient outcomes in building a decision-support system used in routine clinical practice to provide physicians and other health professionals with concrete, personalized recommendations regarding optimal psychosocial support strategies.

2.2. About task 1.3

The Task 1.3 will deal with the elaboration of the methodology to be followed by the BOUNCE consortium partners in order to reach the aspired results. A detailed analysis of the as-is and to-be processes will be performed in workflow diagrams. At the same time the data inputs and structures needed, as well as the expected outputs for every possible process (methodological step) and interactions will be modelled in detail. The task will include elaboration of the methodological steps required, ranging from data source collection, (pre-)processing, computational model definition and fine-tuning, etc. This task will also include a state of the art analysis on existing open source / commercial methods, components and tools that may be relevant for integration into the BOUNCE services and platform infrastructure. This task which will result in deliverable D1.3, will serve as input both to task T3.1 (Data Source identification and collection) and to task T4.2 (Design of the Theoretical In Silico Resilience Trajectory Predictor).

2.3. Purpose of the document

The purpose of this document is to report the results of T1.3 reporting on the BOUNCE methodology, the state of the art tools/services considered for adaptation and integration by the project and to identify the data workflow of the project.

2.4. Work methods

For collecting the state of the art tools and services related to the project all relevant partners contributed according to their expertise. As such a state of the art review was performed on the respective fields reported in this deliverable. Then all pilot sites elaborated on the data workflow of the project and specified the methodology for reaching the project objectives. Finally a summary of a) the protocol for collecting the prospective data was provided, although the corresponding protocol has not yet been finalized and b) of the available retrospective data available by the pilot sites.

2.5. Main content of the document

The rest of this document has been structured as follows: Section 3 elaborates on the state of the art technologies necessary for BOUNCE implementation as well as background work from previous projects that will be exploited to avoid replicate work. Then, Section 4 presents the overall BOUNCE methodology providing more details on the steps documented in the description of work. Next, Section 5 elaborates more on the available data and on the data collection procedures. Finally, Section 6 concludes this deliverable and presents our next plans.

3. State of the art

As within BOUNCE several technologies are going to be used, in this Chapter we present an overview of the current state of the art in technologies relevant to the tools that will be developed and used within the project. As such tools for collecting personal health data, data management, security, anonymization, model collection, curation, validation and integration, temporal data mining technologies and psychoemotional assessment tools and models are being presented, elaborating specifically on the mature technologies that will be exploited within BOUNCE.

3.1. Existing Personal Health Record Platforms

The advancements in healthcare practice, the limitations of the traditional healthcare processes and the need for flexible access to health information, create an ever-growing demand for electronic health systems everywhere. To this direction, Personal Health Record (PHR) systems provide citizens with the ability to become more active in their own care by combining data, knowledge and software tools. The PHR concept is citizen centric, in the sense that its management is the primary responsibility of the citizen. Through a PHR application, the citizen/ patient is able to provide daily life-status information, maintain his/ her own record of medical exams and define the access rights to their personal data, leveraging that access, to improve health and disease management. Over the last twenty years, a large number of PHR-like systems have been developed. Some of them can be found in the Table 1.

911 Medical ID (http://www.911medicalid.com/)	Minerva Health Manager (http://www.myminerva.com/)
CareZone PHR (https://carezone.com/)	MyALERT (http://www.alert-online.com/myalert)
Dossia (http://www.dossia.org/)	myMediConnect PHR (http://www.passportmd.com/)
eclinicalWorks Patient Portal (http://www.eclinicalworks.com/products-patient-portal.htm)	MyOscar (http://myoscar.org/)
Epic MyChart (http://www.epic.com/software-phr.php)	NoMoreClipboard (http://www.nomoreclipboard.com/)
HealtheTracks (http://www.healthetracks.com/)	OpenMRS (http://openmrs.org/)
Indivo-X (http://indivohealth.org/)	Patient Ally (https://www.patientally.com/Main)
KIS PHR (http://kismedicalrecords.com/)	Patient Fusion (http://www.practicefusion.com/pages/phr.html)
LifeLedger (http://www.elderissues.com/)	PatientsLikeMe (http://www.patientslikeme.com/)
MedHelp PHR (http://www.medhelp.org/)	Tolven (http://www.tolven.org/)
MedicAlert (http://www.medicalert.org/)	Web MD Health Manager (http://www.webmd.com/health-manager)
MedicKey PHR (http://medickey.com/)	zweena PHR (http://www.zweenahealth.com/)
Microsoft HealthVault (http://www.microsoft.com/en-gb/healthvault/default.aspx)	

Table 1. PHR product names and websites

Despite the wide variety of potential benefits [52], the uptake of PHRs has been very slow [103]. Recent reviews [24][40] have identified as a problem the fact that only a small subset of the PHR

applications are free, web-based and open-source. In addition, the variety of existing business models, fee-based or commercial, complicate even more the selection of an appropriate PHR. There are also some remaining issues that still need to be resolved, which complicate even further the selection and use of a PHR. These are:

a) **Interoperability**: PHR systems are rarely integrated and interoperable with other electronic services and systems [41][65]. In most cases, end-users need to enter all their health data manually.

b) **Usability/ Adaptability**: The majority of PHR systems follow the approach "one system fits all". However, different persons with different primary diseases have different needs and the PHRs so far fail to adapt to specific needs [69].

c) **Trust**: There are limitations in the methodologies used for sharing information among patients, and their relatives, doctors and researchers. There is a sense of lack of trust as well as inefficient access control and security mechanisms [24][55][66].

d) **Added Value**: PHR systems in their majority are not linked to specific health services. Also, the added-value for citizens to maintain a personal health file through manual input of data has not been adequately demonstrated [24][40][66].

To face these challenges, guidelines and standards are starting to emerge to support the development of quality PHR systems. These include the US Meaningful Use Criteria [66], and the HL7 PHR-S FM [48]. However, the adoption of those is still limited.

3.1.1. The case of iPHR

To tackle the aforementioned challenges, the intelligent Personal Health Record (iPHR) [62] has the objective to provide a cancer-specific self-management platform designed according to the needs of patient groups while focusing on the wellbeing of the cancer patient with special emphasis on avoiding, early detecting and managing adverse events of cancer therapy but also, importantly, on the psycho-emotional evaluation and self-motivated goals. Furthermore, the iPHR adopts a privacy by design approach in its development to ensure the privacy of patients using the application. The platform regularly monitors the psycho-emotional status of the patient and periodically records the everyday life experiences of the cancer patient with respect to the therapy side effects. Different groups of patients and their families can share information through diaries, and clinicians are provided with clinical information.

The iPHR extends Indivo [74] PHR and moves beyond the current state of the art among others in the following directions:

- **GUI**: It provides a nice, user friendly graphical user interface that is platform independent and optimized for mobile devices (laptops, tablets, mobile devices etc.).
- **Sharing & interoperability**: Multiple roles are supported such as patients, health professionals, companions and researchers allowing the secure and seamless sharing of selective information, enhancing the patient-doctor and patient-patient interaction and communication and enabling researches to access statistical information.
- **e-Diary**: If further optimizes interactions of the participants using e-diaries, allowing the patients to enter and to view their activities and behaviors across different period of time

- **Multiple apps**: Besides legacy apps for managing and recording the individual health status (i.e. problems, allergies, medications, procedures, laboratory results etc.), novel apps have been implemented focusing on psycho-emotional monitoring of the cancer patient, providing intelligent services (e.g. drug interactions and recommendations, alerts, patient profiling etc.) and managing medical documents.
- **Advanced data management**: The users are able to connect through their iManageCancer account to external data sources such as activity trackers, sensors, social media and hospital information systems. In this direction a novel big-data infrastructure has been designed and implemented allowing the uninterrupted addition of future data sources.
- **Involve stakeholders in the design and development process:** Right from the beginning of the development process all involved stakeholders were heavily involved in the development process. Besides the requirements elicitation phase where more than 200 possible end-users were contacted, earlier versions of the system had been evaluated by a diverse group of physicians at three different places and time points and the results were used to further improve the system.

The iPHR system provides an innovative ecosystem to support patient self-management through the involvement of all stakeholders participating in the therapeutic process. Added-value services for physicians and researchers (e.g. providing physicians with smart analytical services for cohort statistics, access to patient's medical record and psycho-emotional data etc.) are currently being built on top of these functionalities.

3.1.2. Personal Health Management in BOUNCE: Noona

Central tool for data collection within the BOUNCE project is the Noona Healthcare platform, a PHR system designed for cancer patients.

In the EU region Noona is a CE-marked class 1 medical device, in accordance with standards MEDDEV 2.1/6 January 2012, IEC 62304, and EU directive 93/42/EEC. The manufacturer of the service, Noona Healthcare, has carried out conformity assessment and classification according to class 1 requirements. The GMDN code for the service is 58884. The national regulative authority in Finland is Valvira, National Supervisory Authority for Welfare and Health.

In the US market Noona is a medical device under exemption criteria for mobile medical devices (Mobile Medical Applications - Guidance for Industry and Food and Drug Administration Staff, issued on February 9, 2015) for which the U.S. Food and Drug Administration (FDA) exercises enforcement discretion. Noona mobile service is intended to be used for cancer patient remote monitoring and as a support tool for communication between cancer patients and healthcare professionals. Patients submit information on their symptoms to healthcare professionals in order to receive instructions and guidance during the active treatment phase and in post-treatment recovery phase. Healthcare professionals are able to evaluate the patient symptoms and recovery progress based on the patient reported outcomes data transmitted by the product.

Noona provides a web application-based service to the end user. It is a fully responsive web application usable with web browser and suitable devices including desktop, laptop, pad, and smart phones. All users use Noona with their existing devices. Noona has two distinctive user groups - cancer patients and cancer hospital care personnel (doctors and nurses). Patients are registered to Noona and begin use either at the onset or after completion of basic cancer

treatments (i.e., during the post-treatment recovery phase). During treatment phase, the intended use for Noona is to evaluate symptoms and recovery progress based on patient-reported data transmitted by the product. During the post-treatment recovery phase the intended use is to monitor patient recovery from cancer and related symptoms and provide consultation & advice to support the recovery progress.

Noona is a cloud-based service and can therefore be used on most devices with a supported web-browser and internet connection, limiting technical demands to internet connectivity and user work station.

- For the clinical users requirements are workstations or tablets with a web-browser and internet connection.
- For the patient users requirements are smart phone, tablet, or computer with a web-browser and internet connection.

Noona has two user interfaces, one for each user group. Patient side main functionalities are to report on symptoms using cancer/treatment specific assisted question wizards that cover the clinically relevant questions and measurements per most common and relevant symptom as well as a symptom diary to self-monitor on symptom progress and recovery status. Patients decide when they require clinical assistance regarding their symptoms and react by contacting clinic with Noona. Clinics may also send repetitive scheduled questionnaires to patients. Patients may also contact the clinic regarding other topics than symptoms. This is done via an open question form. Further patient side functionalities include possibility to view notifications intended for all patients as well as patient specific messages from responsible nurses and doctors.

Clinic side main functionality is a work queue to monitor new patients who have requested assistance or who have responded to a scheduled questionnaire, a view on patient information as well as symptom history and the possibility to communicate directly to patient. The possibility for a patient to contact the clinic directly and clinic to message back to patient is optional and configured based on clinic preferences.

Patient identity is encrypted and decryption keys are restricted to authorized users only. Data visibility is limited to individual patient on patient side and clinic level on clinic side. All clinic users see all patients of their own clinic, with the possibility to configure nurse/doctor specific care teams for visibility filtering. Visibility to own clinic patients is enforced by access list mechanism.

Noona Healthcare customer service is responsible for the customer's key users. Other nurse and doctor user accounts are created by customers own key users. All customer nursing staff is responsible and entitled on the patient user account management.

- Invitation link is sent to clinic user via email
- Invitation link is sent to patient via email.
- Patient sees in the system the information they submit and response messages from staff of the clinic.
- Users login to Noona with user name and password from the AWS cloud region specific login website
- According to our experience password authentication offers suitable level of protection for daily use from both perspectives of patient confidentiality and usability.
- Password policy is used to enforce strong passwords

- Patient using the native app from their personal smart phone, a pin code, and when feasible, fingerprint based login is supported.

Noona Healthcare does not offer direct end user support for patients. Instead as part of patient onboarding, the Noona application gives a tutorial for the patient. Intuitive usability is also a design principle followed in Noona's product development process, which is ensured through continued usability testing on cancer patients. As part of the deployment, nursing staff customers are trained to provide patients with end user support. It mainly focuses on support in recovering an expired password or change of password for some other reason as well as giving the initial introduction on what Noona is and how is it used in the customer hospital.

Our plan as part of the Bounce project is to develop a module that focuses on measuring patient coping, mental health and quality of life during and after cancer treatments. The data already collected by Noona will be further enriched with additional data on a comprehensive corpus of resilience-related variables through self-report patient scales and questionnaires designed specifically for BOUNCE. The module will be enriched with automated clinical rules that alert healthcare professionals when significant variations in the patients' psychoemotional status are recorded. The enhanced Noona tool will provide access to all patient reported outcomes throughout the project. The collected data will be accessible in real time through Noona's digital analytics tools.

In the future, predictive models developed as part of the BOUNCE project can be used by health care professionals around the world as an integral part of breast cancer treatments – significantly improving the lives of hundreds of thousands of breast cancer patients.

3.1.2.1. Security features of Noona

Data privacy and security obligations are implemented in Noona Healthcare procedures so that data processing is carried out according to statutory and regulatory requirements. Noona Healthcare has entered in written agreements with sub-processors regarding data privacy and security obligations in accordance with statutory and regulatory requirements. Appropriate process and technical controls are used in processing and storage of customer data according to information security risk assessment (ISO27001).

Security is part of Noona Service system architecture at all levels: software architecture, data center architecture and network architecture. Noona service is secured with application level controls, firewalls, application firewalls and denial of service protection mechanisms. identity and message information is encrypted on application level with rotating and patient specific symmetric encryption keys. Patient data audit log is collected and archived. All data is encrypted in rest and in transit. System components are segregated to dedicated subnets according to component security level. Noona Service is hosted in multitenant architecture. Access to customer data is restricted with authentication, authorization, cryptographic access controls, role access controls and customer user group access controls.

Noona Service system components are deployed in redundant configuration to ensure availability in case of system component failure. System data storages are backed up daily. In the event of storage failure customer data can be recovered from backups. Noona healthcare has disaster recovery procedure to recover from system wide technical failure or data corruption.

The following sub-contractors are used to provide Noona Service:

- Noona mobile service is hosted in Amazon AWS in Europe, Ireland.
- User Analytics are processed in Microsoft Azure in Europe, Ireland.
- Log processing services are provided by Sumologic in Europe

3.2. State of the art on Data management

As within BOUNCE several types of data will be collected for cancer, ranging from medical, clinical, psychosocial etc. in this section we review approaches from other related projects to model, represent, integrate and manage all this information to be collected and eventually processed.

3.2.1. Relevant Project infrastructures

Two relevant project already managing heterogeneous health data are MyHealthAvatar and iManageCancer which will offer the baseline for the development of the BOUNCE data management layer.

3.2.1.1. MyHealthAvatar

The MyHealthAvatar (MHA) EU project [67] was an attempt for the digital representation of patient health status. The goal was to create a *"digital avatar"*, i.e. a graphical representation/manifestation of the user, acting as a mediator between the end-users and health related data collections. It was designed as a lifetime companion for individual citizens in order to facilitate the collection, the access and the sustainability of health status information over the long-term. Among others, key questions that were answered in this context is how to develop optimal frameworks for large-scale data-sharing, how to exploit and curate data from various Electronic and Patient Health Records, assembling them into ontological descriptions relevant to the practice of systems medicine and how to manage the problems of large scale medical data.

The conceptual architecture of the data management platform is shown in Figure 1. In the bottom layer, external sources are pushing data to the original data repository by using a variety of linking services. In addition, there are sources that allow access to the available information (such as the Linked Life Data[1] or the DrugBank[2]) directly from the semantic integration module. The data are semantically linked and integrated using the aforementioned module and stored as triples at the Semantic Data Warehouse to be served. On top of these repositories various APIs allow granular and secure access to the available data either directly from the original data repository or from the semantically integrated data warehouse.

In order to model available data, the MHA Semantic Core Ontology [67], shown in Figure 2, was used as the virtual schema of all data stored within MHA. It is able to semantically describe the different types of data required and processed by the platform.

[1] http://linkedlifedata.com/

[2] http://www.drugbank.ca/

The overall architecture adopts a variation of the *command-query responsibility segregation*[3] principle where a different model is used to update information than what is used to read. Although the mainstream approach people use for interacting with an information system is to treat it as a create, read, update and delete data-store, as the needs become more sophisticated state of the art approaches steadily move away from that model. As such MHA relies on NoSQL technologies to store the original data due to their ability to handle enormous data sets and the "schema-less" nature, which makes, to a large extent, the import of new information to be frictionless. But their limitations in the flexibility of query mechanisms are a real barrier for any application that has not predetermined access use cases. The Semantic Warehouse component in the MHA platform fills these gaps by effectively providing a semantically enriched and search optimized index to the unstructured contents of the Cassandra repository.

Within BOUNCE we will reuse several modules from the MyHealthAvatar infrastructure. For example we intend to have a proprietary data repository for storing all available data. Then depending on the integration needs those data will be cleaned, integrated, harmonized and semantically uplifted. To this end, reusing modules of the MHA Ontology Suite will add value to the project and speed up its implementation.

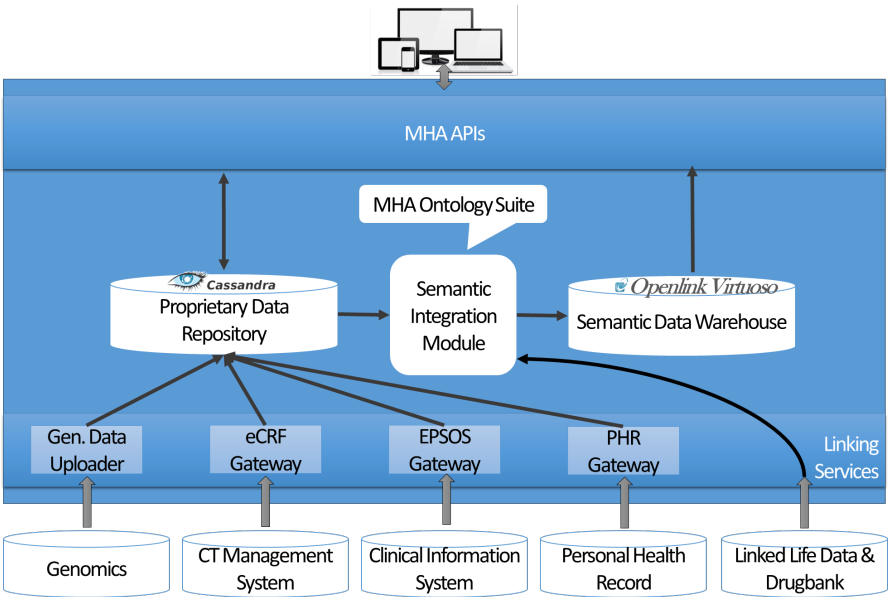

Figure 1. The architecture of the Data Management Approach

[3] http://en.wikipedia.org/wiki/Command%E2%80%93query_separation

Figure 2. The modules of MHA Semantic Core Ontology[4]

3.2.1.2. iManageCancer

The iManageCancer H2020 EU project [106] has the objective to provide a cancer specific self-management platform designed according to the needs of patient groups while in parallel focusing on the wellbeing of the cancer patient with special emphasis on preventing, early detecting and managing adverse events of cancer therapy but also, importantly, on the psycho-emotional evaluation and self-motivated goals.

For managing heterogeneous cancer data, within iManageCancer, a data management architecture has been designed and implemented. Existing ICT systems and tools store and push data to the iManageCancer data lake. Selected subsets out of the data lake are semantically uplifted, integrated and explored, through novel data integration tools. Both the proprietary and the integrated information, is then served through various data access APIs. Bellow we describe each one of the aforementioned modules of the platform.

[4] 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, 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.

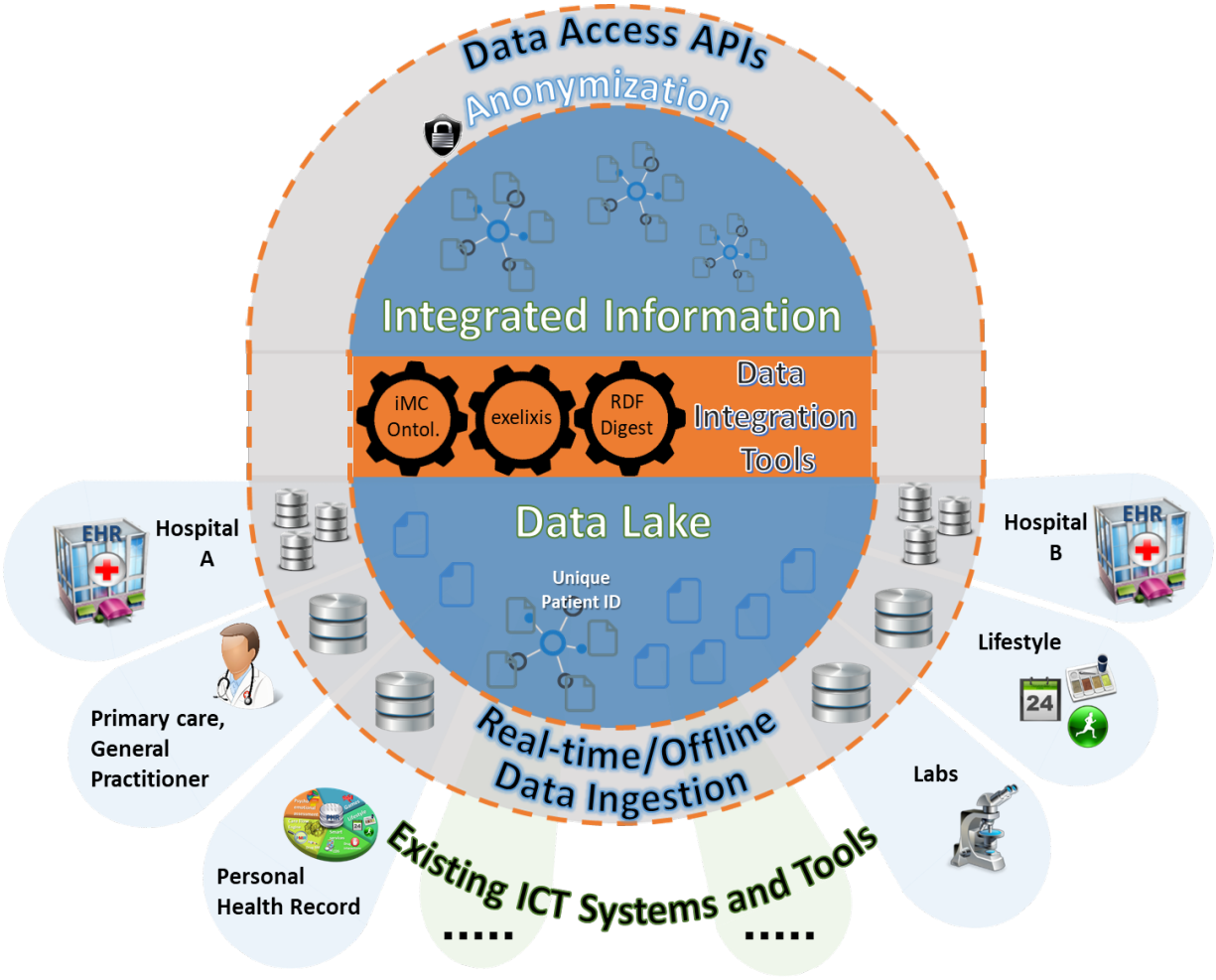

Figure 3. The iManageCancer Data Management Architecture.

The Data Lake. The bottom layer of the data management architecture is an instantiation of the data lake concept. There are multiple, heterogeneous databases using different languages for retrieving data (e.g., SQL, CQL, and APIs), different technologies for storing and serving those data, and different security requirements. For example, a PHR system, using a PostgreSQL database, regularly monitors the psycho-emotional status of the patient and periodically records the everyday life experiences of the cancer patient. At the same time, several apps using Cassandra, MySQL databases and a triple store monitor medications, pain, side-effects of cancer treatment, lifestyle and diet choices, whereas serious games, exposing only data access APIs, try to educate and encourage patients. We argue that this heterogeneity is inhibited in the medical domain and data management architectures should not limit but embrace diversity.

The IMC Semantic Core Ontology. For enabling a common representation of knowledge across the continuum of care and across the different information sources, the iMC Semantic Core ontology was developed. It is a modular ontology including the 48 most widely used subontologies in the health domain, used as the virtual schema of all data stored within the platform, and is able to semantically describe the different types of data required and processed by the platform.

Data Integration Tools. Having a way to model all available information, *exelixis* [64], allows both the real-time access over the integrated information and the offline ETL of the semantically transformed and integrated information to a triple store. The benefit of the first approach is that

the latest information is always accessed, however at a cost of the execution time. On the other hand, accessing the already transformed information is faster. However, the accessed information might be outdated. For the iMC project, we employed both approaches, getting the benefits of both solutions. Besides enabling interoperability and integrating selected available data, a key aspect usually neglected by data management architectures is the continuous evolution of the ontologies/terminologies used. We argue that evolution should not be treated as a side-effect but as a "first-class citizen" in a modern data management infrastructure and this requires the redesign and the restructure of the available solutions and frameworks to reflect this requirement. As such, a unique feature of the *exelixis* data integration engine is that it enables the uninterrupted evolution of the modules of the IMC Semantic Core Ontology. *exelixis*, given multiple ontology versions automatically identifies the changes, reuses past mappings and automatically rewrites input queries among ontology versions. Using this mechanism, mappings to a previous ontology version can co-exist with mappings to a recent ontology version and work uninterrupted.

Data Access APIs. All data available through the aforementioned data management architecture can be accessed using data access APIs. Using those APIs, application requests are transformed to queries that are either targeting individual data sources from the data lake or targeting the integrated information - available using SPARQL queries.

Many parts of the aforementioned architecture can be reused in the context of BOUNCE. For example, the IMC Semantic Core Ontology has extended the MHA Ontology focusing in the cancer disease and can be used for semantic uplifting and annotation. Furthermore the data integration tools for ETL data into a harmonized repository are also ready and available to be used in the context of BOUNCE.

3.2.2. Health vocabularies, ontologies and semantic models

According to Wikipedia, Semantic Integration[5] is the process of interrelating information from diverse sources and has to resolve several heterogeneity problems. These problems can be classified into syntactic and semantic heterogeneities. The former are due to differences in the access interface, query language and database models. The latter are caused by different data representations for schemas or instances. During the last 15 years, numerous systems have been developed, often targeting specific problems or areas. The main approaches, are either centralized – e.g. data warehouses, where data is stored locally – or federated – where data is left at the sources and accessed on demand. The selection of either approach depends on the type of solution to be deployed. Data warehouses might deal with data privacy issues and with outdated data. However, they provide better efficiency and allow tighter control to data managers over what data will be available. Federated approaches always access updated data, allow partial and non-managed data connections, but suffer from efficiency issues. Federated approaches, also known as query translation rely on a virtual schema that represents the space of queries that the user can submit to the system. It is called 'virtual' because no data is stored centrally. Instead, each query is dynamically translated into a set of sub-queries for the databases to integrate, and their single results are merged into a global result, which is presented to the end-users as answer to his initial query.

[5] http://en.wikipedia.org/wiki/Semantic_integration

Either in virtual integration or data warehousing, during the last years, ontologies have been used in order to integrate structured and semi-structured data, obtaining promising results, for example in the fields of biomedicine and bioinformatics[6]. However, there is not a single correct way to model a domain and several ontologies exist. Next, some examples of such ontologies are presented: Symptom Ontology[7] was designed around the guiding concept of a symptom. The Disease Ontology[8] (DO) is trying to link disparate datasets through disease concepts. The Foundational Model of Anatomy[9] has to do with the phenotypic structure of the human body, whereas Adverse Event Ontology [50] tries to model adverse events. The Experimental Factor Ontology focuses on experimental variables in Gene Expression Atlas[10], the Clinical Care Classification System[11] attempts to code health care settings and the Current Procedural Terminology[12] (CPT) is a medical nomenclature used to report medical procedures and services under public and private health insurance programs. UMLS[13], the Unified Medical Language System, is a unifying framework, which integrates different terminologies, which are relevant to medical and biomedical information technologies. The Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) is a clinical terminology, which has been promoted as a reference terminology for electronic health record (EHR) systems. SNOMED CT is used by the College of American Pathologists[14], the UMLS Metathesaurus[15], the European project epSOS[16] and the European project SemanticHealthNet[17]. The Medical Subject Headings[18] (MeSH) are a medical thesaurus published and annually updated by the US National Library of Medicine (NLM). It is used for cataloguing of the library holdings and for indexing of the databases that are produced by the NLM (e.g. MEDLINE). ACGT MO[19] tries to model cancer-related medical knowledge. The International Classification of Diseases[20] is the world's standard tool to capture mortality and morbidity data. LOINC is a database and universal standard for identifying medical laboratory and clinical observations and Medical Dictionary for Regulatory Activities (MEDRA) is a clinically validated international medical terminology for diagnoses, symptoms, surgeries and other medical procedures. The Thesaurus of the National Cancer Institute[21] (NCI) covers vocabulary for clinical care, translational and basic research and public information and administrative activities. Moreover, other ontologies try to model multiscale data such as the Systems Biology

[6] Tom Mitchell 1997 “Machine Learning”, McGraw Hill

[7] http://symptomontologywiki.igs.umaryland.edu/wiki/index.php/Main_Page

[8] http://www.obofoundry.org/cgi-bin/detail.cgi?id=disease_ontology

[9] http://sig.biostr.washington.edu/projects/fm/AboutFM.html

[10] http://www.ebi.ac.uk/gxa/

[11] http://en.wikipedia.org/wiki/Clinical_Care_Classification_System

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

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

[14] http://www.cap.org/apps/cap.portal

[15] http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html

[16] http://www.epsos.eu/

[17] http://www.semantichealthnet.eu/

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

[19] http://bioportal.bioontology.org/ontologies/1126

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

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

Ontology[22] and Gene Ontology[23] (GO), which supports biologically meaningful annotation of genes and their products in different databases.

Besides these ontologies that mostly refer to core medical knowledge, other ontologies try to cover the domain of social entities that are related to health care such as Ontology of Medically Related Social Entities[24] and the BioCaster Ontology (BCO) [17], which tries to describe the terms and relations necessary to detect and risk-assess public health events. The FHHO [84] is representing the family health histories of persons related by biological and/or social family relationships (e.g. step, adoptive) who share genetic, behavioural, and/or environmental risk factors for disease.

Obviously, the amount of information available, the heterogeneity of the information, and the wide range of proposed ontologies dictate the identification of a solution being able to handle all this information available. This is why more recent approaches focus on employing multiple ontologies as target schemata to formulate specific queries [80] that can be answered by the underlying data management solution.

Within BOUNCE, a Semantic Model will be designed in order to allow data harmonization and integration, taking the role of a common information model. This semantic model developed will be used to annotate, integrate, and fuse diverse models. The Semantic Model will be defined in a modular, scalable and extensible way. An initial version of this model, the MyHealthAvatar Ontology Suite developed in the context of MyHealthAvatar project and further refined and expanded into the iManageCancer project, is already available and will be further extended and refined to enable data modelling within the BOUNCE project. The roadmap for implementing this is first to identify all available internal and external data sources that will be reported in D3.1. Then the available data will be examined in detail to identify if the available ontologies/terminologies existing within the IMC Ontology suite can sufficiently describe the semantics of all data sources available, enabling interoperability and integration. This will be reported in D3.2. All available information will then be exposed through SPARQL endpoints offered as RDF/S data. As such our approach tries to offer best of both worlds: efficient persistence and availability of heterogeneous data enabling the exploitation of big-data frameworks like Hadoop, Spark and Flink, and semantic integration for harmonization and searching of the “essence” of the ingested information.

3.2.3. Data Cleaning

Today’s real-world databases are gigantic in size and, therefore, highly inclined to noisy, missing and often inconsistent data [33]. Data cleaning deals exactly with this problem, by detecting and removing errors and inconsistencies from such data in order to improve data quality [90]. Therefore, data cleaning routines are applied to complete missing attributes and values, smooth and level noisy data, identify and remove outliers, as well as to identify and settle inconsistencies [54]. Thus, preparing and cleaning datasets prior to analysis is a perennial challenge in data analytics, whilst failure to do so may result in inaccurate analytics and unreliable decisions. This is why, over the last two decades data cleaning has been a key area of database research [68].

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

[23] http://www.geneontology.org/GO.consortiumlist.shtml

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

Many authors have proposed algorithms for data cleaning, in order to remove inconsistencies and noise from data [98]. Several innovative solutions have been proposed in the literature, to address specific data cleaning problems that may occur. The authors in [20] proposed the NADEEF architecture, an extensible, generalized, and easy-to-deploy data cleaning platform that allows users to specify multiple types of data quality rules, which uniformly define what is wrong with the data and how to repair it through writing code that implements predefined classes. In [19] a method is implemented for managing data duplications, where duplication detection is done either by detecting duplicate records in a single database or by detecting duplicate records in multiple other databases. In the same concept, in [6] a two-step technique that matches different tuples to identify duplicates and merge the duplicate tuples into one is proposed. This technique addresses specific data cleaning problems, such as duplicate detection, record matching, and entity resolution. However, it is unable to identify more complex data expressions. To address the complexity of data expression, many data cleaning methods use heuristic rules and user guidance, such as [8], [14], [32], [75], and [113], which require manual labour for the cleaning process. Hence, in [13] an ontology-based data cleaning solution is implemented, employing existing technologies to understand and differentiate the contents of the data and perform data cleaning without the need of human supervision. Apart from these, in [44] the authors proposed a solution for detecting and repairing dirty data that resolves errors like inconsistency, accuracy, and redundancy by treating multiple types of quality rules holistically, while in [105] a rule-based data cleaning technique is proposed, whereby a set of domain specific rules define how data should be cleaned.

One of the most common inconsistency problems concerns missing data. There are several reasons why data can be missing, including: 1) The data is deemed unimportant and is omitted, 2) In the case of laboratory tests, the tests may simply not be executed as they are irrelevant to the medical risk event, 3) The information is intentionally excluded as it may be an identifier to the corresponding patient. Missing values cause problems like loss of precision due to less data, computational issues due to holes in dataset, and bias due to distortion of the data distribution. In order to formulate a process for handling missing data, one must first classify the type of missing data. The different types of missing data are:

- *Missing-completely-at-random (MCAR)*: If a feature value is MCAR, the probability of the feature being missing in an instance does not depend on the other feature values.
- *Missing-at-random (MAR)*: If a feature value is MAR, the probability of the feature being missing in an instance *does* depend on the other feature values.
- *Missing-not-at-random (MNAR)*: If a feature value is MNAR, the probability of the feature being missing in an instance depends on the other feature values. In practice, to determine whether the missing data is MNAR, one would typically perform a sensitivity analysis to check how the missing-at-random assumption is violated.

To solve the missing data issue, various algorithms have been proposed so far. These algorithms fill the missing values and smooth out noise. Examples of such algorithms include constant substitution, mean attribute value substitution, and random attribute value substitution [98]. Another approach to increase efficiency of a data warehouse is the creation of materialized views (MVs), which improves data warehouse performance by pre-processing and avoiding complex resource intensive calculations [97].

It should be mentioned that apart from the approaches developed by the research community, various commercial tools and frameworks for data cleaning have been released. OpenRefine[25] is one of the most commonly used that deals with messy data by through cleaning, transforming to a different format, and extending the available data through with web services and access to external data. DataWrangler[26] is another interactive tool for data cleaning and interactive transformation of messy, real-world data into the data tables' analysis tools. Moreover, DataCleaner[27] is a data profiling engine for discovering and analyzing the quality of the data by finding patterns, missing values, character sets, and other characteristics of the data values, whilst Drake[28] is a text-based data workflow tool that organizes command execution around data and its dependencies by automatically resolving these dependencies and providing a rich set of options for controlling the workflow.

3.2.3.1. Data Cleaning in BOUNCE

Within BOUNCE data cleaning will consist of discrete modules to (i) identify errors, noise, and inconsistencies of the incoming data that have to conform to specific chosen constraints, (ii) correct/remove all the identified errors, noise, and inconsistencies, and (iii) ensure that the data provided is complete, as well as accurate, and free from erroneous inliers, i.e. data points generated in error but falling within the expected range (erroneous inliers often escape detection).

Retrospective data collected so far, were relatively clean and no data cleaning software was required. The cleaning process for these data is reported in D3.1. However, partners have experience with DataWrangler and DataCleaner and will be used if needed as more data become available.

3.3. Security

Security ensuring protection of personal / sensitive information in BOUNCE will be addressed through a two-fold procedure, which includes: 1) Data Access Control, and 2) Security of data across their whole lifecycle, from storage, to transit and to use. The forthcoming subsections introduce these categories, as well as a description of the state of the art tools that will be evaluated so as to employ the most appropriate ones to safeguard security and privacy of information in the context of the project.

3.3.1. Data Access Control

Access control in general includes authorization, authentication, access approval, and audit. Authentication and access control are often combined into a single operation, so that access is approved based on successful authentication, or based on an access token. Authentication methods and tokens include passwords, biometric scans, physical keys, electronic keys and devices, and other means.

[25] OpenRefine, "OpenRefine,". Available: http://openrefine.org/. Accessed June 2018

[26] Stranford, "Data Wrangler,". Available: http://vis.stanford.edu/wrangler/. Accessed June 2018

[27] DataCleaner, "DataCleaner,". Available: http://datacleaner.org/. Accessed June 2018

[28] Factual, "Drake". Available: https://www.factual.com/blog/introducing-drake-a-kind-of-make-for-data. Accessed June 2018

With regards to access control the following types of fundamental access control models can be described:

Attribute Based Access Control (ABAC)[29]. ABAC defines an access control paradigm whereby access rights are granted to users through the use of policies which combine attributes together. The policies can use any type of attributes (user attributes, resource attributes, object, environment attributes etc.). This model supports Boolean logic, in which rules contain "IF, THEN" statements about who is making the request, the resource, and the action. The key difference with ABAC is the concept of policies that express a complex Boolean rule set that can evaluate many different attributes. Although the concept itself existed for many years, ABAC is considered a "next generation" authorization model because it provides dynamic, context-aware and risk-intelligent access control to resources allowing access control policies that include specific attributes from many different information systems to be defined to resolve an authorization and achieve an efficient regulatory compliance, allowing enterprises flexibility in their implementations based on their existing infrastructures.

Discretionary Access Control (DAC)[30]. DAC is a type of access control defined by the Trusted Computer System Evaluation Criteria "as a means of restricting access to objects based on the identity of subjects and/or groups to which they belong. The controls are discretionary in the sense that a subject with a certain access permission is capable of passing that permission (perhaps indirectly) on to any other subject (unless restrained by mandatory access control)".

Identity Based Access Control (IBAC)[31]. NIST defines identity-based security policies as policies "based on the identities and/or attributes of the object (system resource) being accessed and of the subject (user, group of users, process, or device) requesting access." IBAC is an approach to control access to a digital product or service based on the authenticated identity of an individual. This allows organizations to grant access to specific users to access a variety of digital services using the same credentials, ensuring the accurate match between what users are entitled to and what they actually receive, while also permitting other access constraints such as company, device, location and application type (attributes).

Mandatory Access Control (MAC)[32]. MAC refers to a type of access control by which the operating system constrains the ability of a subject or initiator to access or generally perform some sort of operation on an object or target. In practice, a subject is usually a process or thread; objects are constructs such as files, directories, TCP/UDP ports, shared memory segments, IO devices, etc. Subjects and objects each have a set of security attributes. Whenever a subject attempts to access an object, an authorization rule enforced by the operating system kernel examines these security attributes and decides whether the access can take place. Any operation by any subject on any object is tested against the set of authorization rules (aka policy) to determine if the operation is allowed. MAC-enabled systems allow policy administrators to implement organization-wide security policies. Under MAC (and unlike DAC), users cannot override or modify this policy, either accidentally or intentionally. This allows security administrators to define a central policy that is guaranteed (in principle) to be enforced for all users.

[29] https://en.wikipedia.org/wiki/Attribute-based_access_control

[30] https://en.wikipedia.org/wiki/Discretionary_access_control

[31] https://en.wikipedia.org/wiki/Identity-based_security

[32] https://en.wikipedia.org/wiki/Mandatory_access_control

Organisation Based Access Control (OBAC)[33]. OBAC is an access control model which is based on three main entities: the subject, the action and the object. In order to control access, the policy specifies that a subject is permitted to realize an action on an object. Subjects are abstracted into roles. A role is a set of subjects to which the same security rule applies. An activity is a set of actions to which the same security rule applies. A view is a set of objects to which the same security rule applies. OBAC is context sensitive, so the policy could be expressed dynamically. Furthermore, OBAC owns concepts of hierarchy (organization, role, activity, view, context) and separation constraints.

Role Based Access Control (RBAC)[34]. RBAC is an approach to restricting system access to authorized users. It is used by the majority of enterprises and can implement mandatory access control (MAC) or discretionary access control (DAC). RBAC is a policy neutral access control mechanism defined around roles and privileges. The components of RBAC such as role-permissions, user-role and role-role relationships make it simple to perform user assignments. Although RBAC is different from MAC and DAC access control frameworks, it can enforce these policies without any complication.

3.3.1.1. Access Control in BOUNCE

Within the context of BOUNCE, the Noona system which will be used for data collection at the pilot sites has already established the necessary data access control policies and mechanisms, according to which patient identity information is stored and shared with the research nurse and the treating doctor. Those data will be exported to the pilot sites, properly anonymized and sent to the central BOUNCE data management infrastructure.

Regarding the access to these integrated data, the state of the art technologies safeguarding authorisation and access control examined in the context of work package 5 (T5.3) will be customised appropriately and adopted. For the specification and definition of the policies, regardless of the access control model followed, the consortium will examine the utilisation of the eXtensible Access Control Markup Language (XACML) Protocol[35]. XACML is an OASIS standard that describes both a policy language and an access control decision request/response language (both written in XML). The policy language is used to describe general access control requirements, and has standard extension points for defining new functions, data types, combining logic, etc. The request/response language lets you form a query to ask whether or not a given action should be allowed and interpret the result. The response always includes an answer about whether the request should be allowed using one of four values: Permit, Deny, Indeterminate (an error occurred or some required value was missing, so a decision cannot be made) or Not Applicable (the request can't be answered by this service).

3.3.2. Data Lifecycle Security

With regards to data security throughout the whole lifecycle of the data exploitation, BOUNCE consortium will examine and as appropriate develop and delivere safeguards regarding three main security aspects: 1) Security of data in storage, 2) Security of data in transit, or data in motion, and 3) Security of “Data in Use”. Last but not least, security of technical interfaces (e.g.

[33] https://en.wikipedia.org/wiki/Organisation-based_access_control

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

[35] https://www.oasis-open.org/committees/download.php/2713/Brief_Introduction_to_XACML.html

REST) amongst the various BOUNCE components will also be considered and adopted, minimising the risk associated with the exploitation of the operation of various BOUNCE components from external malicious components.

3.3.2.1. Data-At-Rest/Data-In-Storage Security

Security of data in storage is the first of the three parts of the data lifecycle that will be dealt with in the context of the project and is used as a complement to the terms data in use and data in transit which together define the three states of digital data. Data that falls under this category could include files stored on local or cloud hard drive. Data in storage security refers to the preservation of the security, privacy and integrity of data that is stored physically in any digital form. It deals with any type of security around the storage architecture and the data stored on it. Within the context of BOUNCE, and based upon the architectural solutions adopted for the data storage (storage and cloud options will be examined), within T5.4 and reported in D5.2, we will evaluate five main software solutions and approaches associated with data-in-storage security:

- **Symmetric Encryption Algorithms**[36] - Symmetric-key algorithms are algorithms for cryptography that use the same cryptographic keys for both encryption of plaintext and decryption of cipher-text. The keys may be identical or there may be a simple transformation to go between the two keys. The keys, in practice, represent a shared secret between two or more parties that can be used to maintain a private information link. Some examples of popular symmetric algorithms (symmetric-key algorithms) include AES, Blowfish, DES, IDEA, RC2, RC4 etc. AES is not a tool but a symmetric encryption algorithm which has many reference implementations (i.e. tools/libraries) for many programming languages;
- **Message Authentication Codes (a.k.a. MACs) and Digital Signatures**[37] - In cryptography, a message authentication code (MAC) is a short piece of information used to authenticate a message—in other words, to confirm that the message came from the stated sender (its authenticity) and has not been changed in transit (its integrity). A MAC algorithm, sometimes called a keyed (cryptographic) hash function (which is somewhat misleading, since a cryptographic hash function is only one of the possible ways to generate a MAC), accepts as input a secret key and an arbitrary-length message to be authenticated, and outputs a MAC (sometimes known as a tag). The MAC value protects both a message's data integrity as well as its authenticity, by allowing verifiers (who also possess the secret key) to detect any changes to the message content;
- **Broadcast Encryption (Single Sender)**[38] - Broadcast encryption is the cryptographic problem of delivering encrypted content over a broadcast channel in such a way that only qualified users can decrypt the content. The challenge arises from the requirement that the set of qualified users can change in each broadcast emission, and therefore revocation of individual users or user groups should be possible using broadcast transmissions, only, and without affecting any remaining users;
- **Asymmetric Encryption**[39] - Asymmetric algorithms (public key algorithms) use different keys for encryption and decryption, and the decryption key cannot (practically) be

[36] https://en.wikipedia.org/wiki/Symmetric-key_algorithm

[37] https://en.wikipedia.org/wiki/Message_authentication_code

[38] https://en.wikipedia.org/wiki/Broadcast_encryption

[39] https://en.wikipedia.org/wiki/Public-key_cryptography

derived from the encryption key. Asymmetric algorithms are important because they can be used for transmitting encryption keys or other data securely even when the parties have no opportunity to agree on a secret key in private. Types of Asymmetric algorithms (public key algorithms) include RSA, Diffie-Hellman, Digital Signature Algorithm and more;

- **Attribute-Based Encryption[40] (ABE)** - is a relatively recent approach that reconsiders the concept of public-key (PK) cryptography. In traditional PK, a message is encrypted for a specific receiver using the receiver's PK. Identity-based cryptography and in particular identity-based encryption (IBE) changed the traditional understanding of PK cryptography by allowing the PK to be an arbitrary string, e.g., the email address of the receiver. ABE goes one step further and defines the identity not atomic but as a set of attributes, e.g., roles, and messages can be encrypted with respect to subsets of attributes (key-policy ABE - KP-ABE) or policies defined over a set of attributes (cipher-text-policy ABE - CP-ABE). The key issue is, that someone should only be able to decrypt a cipher-text if the person holds a key for "matching attributes" (more below) where user keys are always issued by some trusted party.

3.3.2.2. Data-In-Transit Security Schemes

Data in transit, or data in motion, is data actively moving from one location to another such as across the internet or through a private network. Data protection in transit is the protection of this data while it's traveling from network to network or being transferred from a local storage device to a cloud storage device – wherever data is moving, effective data protection measures for in transit data are critical as data is often considered less secure while in motion. Within the context of BOUNCE, and based upon the architectural solutions adopted for the transfer of data amongst components, nodes and devices, within WP5, we will evaluate two main software solutions and approaches associated with data-in-transit security:

- **IPsec (Internet Protocol Security)** [56]- is a protocol suite for secure Internet Protocol (IP) communications, which works by authenticating and encrypting each IP packet of a communication session. IPsec includes protocols for establishing mutual authentication between agents at the beginning of the session and negotiation of cryptographic keys for use during the session. IPsec can protect data flows between a pair of hosts (host-to-host), between a pair of security gateways (network-to-network), or between a security gateway and a host (network-to-host). IPsec uses cryptographic security services to protect communications over IP networks, and supports network-level peer authentication, data-origin authentication, data integrity, data confidentiality (encryption), and replay protection. IPsec is an end-to-end security scheme operating in the Internet Layer of the Internet Protocol Suite, while other Internet security systems in widespread use (such as the Transport Layer Security (TLS) analysed below), operate in the upper layers (for example TLS operates at the Transport Layer). Hence, only IPsec protects all application traffic over an IP network;
- **TLS[41] (Transport Layer Security)** - is a cryptographic protocol that provides communications security over a computer network. Several versions of the protocol find widespread use in applications such as web browsing, email, Internet faxing, instant messaging, and voice-over-IP (VoIP). Websites use TLS to secure all communications

[40] https://en.wikipedia.org/wiki/Attribute-based_encryption

[41] https://tools.ietf.org/html/rfc5246

between their servers and web browsers. TLS protocol aims primarily to provide privacy and data integrity between two communicating computer applications. When secured by TLS, connections between a client and a server have one or more of the following properties: 1) The connection is private (or secure) because symmetric cryptography is used to encrypt the data transmitted, 2) the identity of the communicating parties can be authenticated using public-key cryptography, and 3) the connection ensures integrity because each message transmitted includes a message integrity check using a message authentication code to prevent undetected loss or alteration of the data during transmission.

3.3.2.3. Data-In-Use

"Data in Use" is all data not in an at-rest state, which is kept only one particular node in a network (for example, in resident memory, or swap, or processor cache or disk cache, etc. memory). This data can be regarded as "secure" if and only if (a) access to the memory is rigorously controlled (the process that accessed the data off of the storage media and read the data into memory is the only process that has access to the memory, and no other process can either access the data in memory, or man-in-the-middle the data while it passes through I/O), and (b) regardless of how the process terminates (either by successful completion, or killing of the process, or shutdown of the computer), the data cannot be retrieved from any location other than the original at rest state, requiring re-authorization. Within the context of BOUNCE, and based upon the architectural solutions adopted, we will evaluate three main software solutions and approaches associated with data-in-use security:

- **Homomorphic Encryption**[42] - Homomorphic encryption is a form of encryption that allows computations to be carried out on cipher-text, thus generating an encrypted result which, when decrypted, matches the result of operations performed on the plaintext. Homomorphic encryption allows chaining together different services without exposing the data to each of those services. Homomorphic encryption schemes are malleable by design. This enables their use in cloud computing environment for ensuring the confidentiality of processed data. In addition, the homomorphic property of various cryptosystems can be used to create many other secure systems, for example secure voting systems, collision-resistant hash functions, private information retrieval schemes, etc.;
- **Verifiable Computation**[43] - Verifiable computing (or verified computation or verified computing) is enabling a computer to offload the computation of some function, to other perhaps untrusted clients, while maintaining verifiable results. The other clients evaluate the function and return the result with a proof that the computation of the function was carried out correctly. The introduction of this notion came as a result of the increasingly common phenomenon of "outsourcing" computation to untrusted users in projects such as SETI@home and also to the growing desire of weak clients to outsource computational tasks to a more powerful computation service like in cloud computing;
- **Secure Multi-Party Computation** [87] - Secure multi-party computation consists of a set of cryptographic methods for parties to jointly compute a function over their inputs while keeping those inputs private. The underlying paradigm is that a scheme is secure if whatever a feasible adversary can obtain after attacking it, it is also feasibly attainable

[42] https://en.wikipedia.org/wiki/Homomorphic_encryption

[43] https://en.wikipedia.org/wiki/Verifiable_computing

from scratch. In the case of multi-party computation in specific, we compare the effect of adversaries that participate in the execution of the actual protocol, to the effect of adversaries that participate in an imaginary execution of a trivial protocol for computing the desired functionality with the help of a trusted party.

Data in use security safeguards initially seem out of the scope of the project, but the adoption of such measures will be thoroughly evaluated during the design of the technical architecture to evaluate the corresponding security needs.

3.4. Anonymization Tools

Data anonymization is a data management and de-identification procedure that conceals private data. Data anonymization is classified to anonymization and pseudo anonymization. Pseudo-anonymization provides the possibility to reconstruct the initial data unlike pure anonymization. Bellow we provide a short overview of the existing open source tools that can be used to this purpose.

ARX[44] is an open source software for anonymizing sensitive personal data. The tool transforms datasets into syntactic privacy models that mitigate attacks leading to privacy breaches. ARX removes direct identifiers such as names from datasets and adds further constraints on indirect identifiers, such as email addresses or phone numbers. The tools also provides built-in data import facilities for relational databases (MS SQL, DB2, SQLite, MySQL), MS Excel and CSV files.

CX-Mask[45] is a paid tool which removes the sensitive data hindering test, outsourcing, and analytics. The tool is using data masking, also referred to as data de-identification, pseudo anonymization, anonymization or obfuscation, which is a method of protecting sensitive data by replacing original data with fictitious but realistic data. The tool de-identifies sensitive data, and retains the realism and functionality of the original data set. The data categories this tools can work with include names, addresses, credit cards, SSN/SIN, phone, and more.

NLM-Scrubber[46] is a new, free clinical text de-identification tool. The software is currently in its initial beta stage. NLM-Scrubber is to be mainly used for de-identifying medical documents.

The University of Nottingham has created an Open Source standalone windows desktop application called OpenPseudonymiser[47]. The application allows users to pseudonymise datasets by creating a digest of one or more columns of a CSV file.

MAT[48] is a toolbox composed of a GUI application, a CLI application and a library, to anonymize/remove metadata. The software is currently on beta stage and the development is on hold. The current version of MAT is not compatible with Python3.

Oracle Data Masking and Subsetting Pack[49] is a free tool which replaces sensitive information such as credit card or social security numbers with realistic values. It provides downloadable masking templates which simplify the task of defining masking rules.

[44] https://arx.deidentifier.org/

[45] https://www.imperva.com/data-security/data-security-101/data-masking/

[46] https://scrubber.nlm.nih.gov/How_to_run_NLM-Scrubber.html

[47] https://www.openpseudonymiser.org/

[48] https://mat.boum.org/

[49] https://www.oracle.com/technetwork/database/options/data-masking-subsetting/overview/index.html.

Jumble DB[50] is a paid software and provides a complete data scrambling solution for the following databases:

- Oracle 9i and up
- SQL Server 2005 and up
- MySQL 5.1 and up
- DB2 LUW 9 and up
- DB2/400 (DB2 system i) V5 and up

DataVantage Global[51] is paid tool which masks and de-identifies data safely and efficiently to help prevent data breaches while providing facilities for data browsing, viewing, editing, sub-setting and more.

Data Masker[52] is a paid software which removes sensitive data from test databases and replaces it with realistic looking false information.

Dgmasker[53] is a paid tool which masks data and supports Oracle, DB2 and MS SQL Server databases, multiple advanced masking algorithms, and the tool is capable of performing data anonymization.

Tricryption[54] anonymization secures sensitive submitted ID Data through the power of cryptography, deriving an encrypted alias/pseudonym appropriate for mining and other database correlation without privacy exposure risks.

Within BOUNCE, it has already been decided, following a comparison of the pros and cons of the above tools, that the ARX tool will be installed and deployed at the local pilot sites, where it will be configured to appropriately anonymize or pseudonymize the available data as needed.

3.5. Model collection, curation, validation and integration tools

3.5.1. Model Collection, Curation and Validation tools

The last decades have shown a great advance both in the development of models of the mathematical biology field but also of statistical and machine learning models. The emergence of the fields of Bioinformatics[55], Systems Biology[56], In Silico modeling for Oncology and Medicine[57] as well as the significant focus given recently in Artificial Intelligence and Big Data analysis, resulted to the need for the development computational models as well as tools for collection, curation and validation of the outcomes of the related research. In order to ensure that the derived models will be safely stored and accessible for reuse, a number of repositories have been emerged for the needs of model collection purposes. In addition to the storage of the models, the re-usability of the models would be significantly strengthened by the use of common

[50] http://www.orbiumsoftware.com/data-masking-tool.php
[51] https://datavantage.com/products/distributed-data-management-and-masking/datavantage-global
[52] http://www.datamasker.com/.
[53] https://www.dataguise.com/protect/
[54] www.eruces.com
[55] https://en.wikipedia.org/wiki/Bioinformatics
[56] https://en.wikipedia.org/wiki/Systems_biology
[57] https://en.wikipedia.org/wiki/In_silico_medicine

protocols for their description. This led to the creation of human- as well as machine-readable languages that assist the sharing process of models and facilitates the curation of them by related tools and processes. The latter includes the interpretation of the model to commonly accepted descriptive languages and the annotation of the model and its details at the semantic level. Moreover, it facilitates the continuous integration and the consistent update of the models. Finally, the continuous increase in the adoption of models creates also the need for the validation of their representation and their results' reproducibility. In this section, paradigms extracted from the aforementioned fields are presented which are considered suitable to be adopted or at least inspire the related BOUNCE tasks.

Starting from the mathematical biology sector, in the Systems Biology field two xml-based markup languages, the Systems Biology Markup Language[58] and the CellML[59] are the basic means used for communicating and storing computational models of biological processes. SBML can represent many different classes of biological phenomena, including metabolic networks, cell signaling pathways, regulatory networks, infectious diseases, and many others. Regarding CellML, it is language that could describe any mathematical model. However it was originally created with the Physiome Project[60] in mind, and hence used primarily to describe models relevant to the field of biology. CellML is similar to Systems Biology Markup Language SBML but provides greater scope for model modularity and reuse and is not specific to descriptions of biochemistry. Regarding model curation both languages adopt and are supported in the MIRIAM (Minimum Information Required In The Annotation of Models) [82], which consists of a set of guidelines suitable for use with any structured format, allowing different groups to collaborate and share resulting models. Adherence to these guidelines also facilitates the sharing of software and service infrastructures built upon modeling activities. MIRIAM is a registered project of the MIBBI[61] (minimum information for biological and biomedical investigations) and its guidelines are composed of three parts, reference correspondence, attribution annotation, and external resource annotation, dealing with different aspects of information that should be included within a model. Moreover, MIRIAM Guidelines lead to the creation of a by-product, the MIRIAM Registry[62], which is a database of namespaces and associated information that is used in the creation of uniform resource identifiers. It contains the set of community-approved namespaces for databases and resources serving, primarily, the biological sciences domain and is used for the proper annotation of models during the curation process. A really significant number of tools that support the SBML and CellML languages has been developed allowing the development, the simulation, and the curation of models. These can be found in web-published lists[63] [64]. The acceptance of the aforementioned languages led to the creation of Database targeting to facilitate the collection of models and easier and better communication of them. The most well-known is the BioModels database[65] which is a part of the international initiative BioModels.net. The resource provides access to published, peer-reviewed, quantitative models of biochemical and cellular systems. Each model is carefully (basically manually) curated to verify that it

[58] sbml.org
[59] http://www.cellml.org/
[60] http://physiomeproject.org/
[61] https://fairsharing.org/collection/MIBBI
[62] https://identifiers.org/
[63] http://sbml.org/SBML_Software_Guide/SBML_Software_Matrix
[64] https://www.cellml.org/tools
[65] https://www.ebi.ac.uk/biomodels

corresponds to the reference publication and give the proper numerical results. Curators also annotate the components of the models with terms from controlled vocabularies and links to other relevant data resources. This allows the users to search accurately for the models they need. The models can currently be retrieved in the SBML, CellML, SciLab XPP-Aut, and BioPAX formats. Recently, additionally to the SBML and CellML languages, the Biomodels Database started to host models encoded in formats other than SBM. The original source for these model files, such as Matlab, Mathematica, and R code are made available for download from the model page. For the SBML and CellML cases however, the models stored in BioModels' curated branch are compliant with MIRIAM, the standard of model curation and annotation. Other well-known tool collection repositories for this field include the CellML repository[66] and the JWS Online database[67].

In order for the reproduction of the simulation results of these models to be guaranteed, the SED-ML(Simulation Experiment Description Markup Language)[68] language has been proposed which is a representation format, based on XML, for the encoding and exchange of simulation descriptions on computational models of biological systems. The SED-ML format is built of five major blocks. The Model entity is used to reference the models used in the simulation experiment and to define pre-processing procedures on these models before simulation. Models must be in standard representation formats (e.g., SBML, CellML, NeuroML). Examples for pre-processing are, e.g., changing the value of an observable, computing the change of a value using mathematics, or general changes on any XML element of the model representation. The Simulation entity contains all information about the simulation settings and the steps taken during simulation, e.g., the particular type of simulation and the algorithm used for the execution of the simulation. The simulation algorithm is specified with a Kinetic Simulation Algorithm Ontology[69] term. The Task entity applies one of the defined simulations with one of the referenced models at a time. The DataGenerator entity encodes post-processing procedures which need to be applied to the simulation result before output, e.g., normalisation of data. Finally, the Output entity specifies the simulation output, e.g., the particular plots to be shown.

Regarding the validation of the models, what it is mainly available is a set of tools that check the representation validity of the models taking into account the target descriptive language. Since the vast majority of these languages are xml-based, the basic methodology to validate these models is based on the validation of the model representation against the XML Schema Definition (XSD). This logic is followed both for CellML[70] and in SBML (for example using the Online SBML Validator[71]). This approach maybe followed for any XML based representation language using XML validation tools available as standalone applications or as libraries for the well-known programming languages (e.g. in Java the javax xml validation package[72]). It should be mentioned, however, that the language specific validation tools may offer additional consistency and correctness checking procedures. For example, in SBML Validator there are additional options and rules checking consistency of measurement units associated with

[66] http://models.cellml.org/cellml

[67] https://fair-dom.org/jws-online/

[68] https://sed-ml.github.io/

[69] http://co.mbine.org/standards/kisao

[70] https://www.cellml.org/tools/validation

[71] http://sbml.org/validator/

[72] https://docs.oracle.com/javase/8/docs/api/javax/xml/validation/package-summary.html

quantities[73], correctness and consistency of identifiers used for model entities[74], proper syntax of MathML mathematical expressions[75], validity of SBO identifiers (if any) used in the model[76], static analysis of whether the model is overdetermined, additional checks for recommended good modeling practices[77], and other general SBML consistency checks[78].

To continue with, the In Silico modeling for Oncology and Medicine sector showed significant advancements in the creation, implementation and testing of models for the simulation of human diseases, such as cancer, also in the framework of previous EC funded projects such as ACGT, Contra Cancrum, TUMOR, P-medicine and CHIC. Especially the TUMOR and CHIC projects particularly focused on the creation of model collection and model combination procedures. Especially for the former, in TUMOR project a repository to store and curate cancer-related models, developed in diverse programming or descriptive languages, has been provided. The model repository consisted of five entities: Tools/Models, References, Licenses, Associated Files, and Parameters. The basic principles of TUMOR model repository were: every tool/model has some information that defines it (id, title, description, etc.), every tool/model can have some information that classifies it (mathematical type, biocomplexity direction etc.), every tool/model can have some information that that defines cancer type and therapy considered (cancer type, treatment included or not etc.), every tool/model can be associated with a set of references, every tool/model can be associated with a set of iles (these files can be documents, source code files, executable files, parameters file etc. and if the associated file is an implementation (source code file or executable file) then it can be associated with a set of parameters which could be can be numbers, strings or files). Additionally, in seeking a standardized way of describing mathematical and computational models and to enable interoperation between systems, repositories, and between the models themselves a new markup language, TumorML (Tumor model repositories Markup Language) [23], for describing computational models that fall within the domain of cancer has been defined in TUMOR. TumorML is an XML-based markup language that wraps existing cancer model implementations with metadata for model curation, parametric interface description, implementation description, and compound model linking. For example, the aforementioned TUMOR repository exposed part of its content in TumorML in order to be used by the workflow environment for the execution of models developed in TUMOR. TumorML was developed to overcome the limitations of existing markup languages, not as a competitor to either of CellML and SBML, but to deal with storing and transmitting existing cancer models among research communities. As an XML-based language, TumorML instances can be validated against the XML schema (markup specification) as discussed previously.

A follow-up of the TUMOR project was the CHIC project[79] which among others had as central objectives the creation of model and tool repositories and procedures to combine models in hyper-models and execute them in synergy. The CHIC model and tool repository permanently hosts multiscale cancer models that have been developed in the context of the CHIC project. It also hosts tools such as linkers and data transformation tools, which are needed for the

[73] http://sbml.org/Facilities/Documentation/Error_Categories#Units
[74] http://sbml.org/Facilities/Documentation/Error_Categories#Identifier
[75] http://sbml.org/Facilities/Documentation/Error_Categories#MathML
[76] http://sbml.org/Facilities/Documentation/Error_Categories#SBO
[77] http://sbml.org/Facilities/Documentation/Error_Categories#Practice
[78] http://sbml.org/Facilities/Documentation/Error_Categories#General
[79] https://www.chic-vph.eu/

construction of hypermodels. For each model, the model repository contains all the related information, including descriptive information (abstract and detailed description, references, etc.), input and output parameters (for proper linking with other models and tools), source files, documentation and executables of the models. Moreover, information about model authorship, ownership, and access permissions are also stored in the model repository database. In order for the user to be able to interact with the Repository, a web-based interface has been designed and implemented. Apart from the aforementioned graphical interface, many web services have been developed so as to be able to expose the contents of the repository to other tools developed in the CHIC project, such as the hypermodelling Editor, the CRAF (Clinical Research Application Framework) and the hypermodelling Framework. The user is able to store in an elegant and user-friendly way new models in the Model Repository through a five-step wizard, or even browse, view, change and delete the content of the Repository.

The CHIC model and tool repository consists of four main entities, the models/tools, the properties, the parameters and the files.

The basic principles of the model repository are:

- Each model/tool has basic descriptive information, stored in the entity “mr_tool”. This information uniquely defines the model/tool and differentiates it from the other models/tools.
- Each model is categorized based on the perspective from which it is viewed in the basic science context. This metamodeling description of each hypomodel based on the CHIC 13 perspective approach facilitates its technology mediated linking.
- The descriptive information of the perspectives is stored in the entity “mr_property”. This entity does not contain the value of the perspective (related to a specific model/tool), but only the description of the perspective. The value that a perspective takes in case of a specific model/tool is stored in the entity “mr_tool_property”.
- The models are treated as generic stubs which have entry and exit points. Consequently, each model/tool has various parameters, serving as input parameters or output parameters, which are stored in entity “mr_parameter”. This entity facilitates the transition from an abstract representation to a concrete one. Logical compatibilities between connected parameters must be taken into account along with the aspect of units, in order to avoid inconsistencies between the connected models/tools.
- Each model/tool may be associated with a set of references, stored in the entity “mr_reference”, which provides direct or indirect links to additional material, extending in this way the knowledge base related to the specific model/tool.
- Every model/tool can be accompanied by a set of files. The information concerning the aforementioned files is stored in the entity “mr_file”. The entity “mr_file” only holds the metadata of the file and not its data. The data of the files are stored internally in a file based repository. If a file is an implementation or a computational representation of a model/tool, then a suitable engine is specified for running the file.

According to the aforementioned principles of the model repository, the Entity-Relationship (ER) diagram of model/tool repository is presented in Figure 4.

The schema of the relational database of the model repository has been designed in order to be able to efficiently store within the CHIC platform all the persistent data that are related to the models. Both the metadata description of the models (parameters, perspective values, references, basic descriptive information) and the files that are related to the models

(executables, documentation, configuration files, source code) are stored in the MySQL database of the Repository. Since the MySQL database server is the component which is responsible for the persistent storage of the models, it is considered the most sensitive, critical, and vital part of the model repository. Nevertheless, the model repository consists of many other components, such as the Apache Application Server, the Django Web Framework, and some back-end and front-end libraries and dependencies.

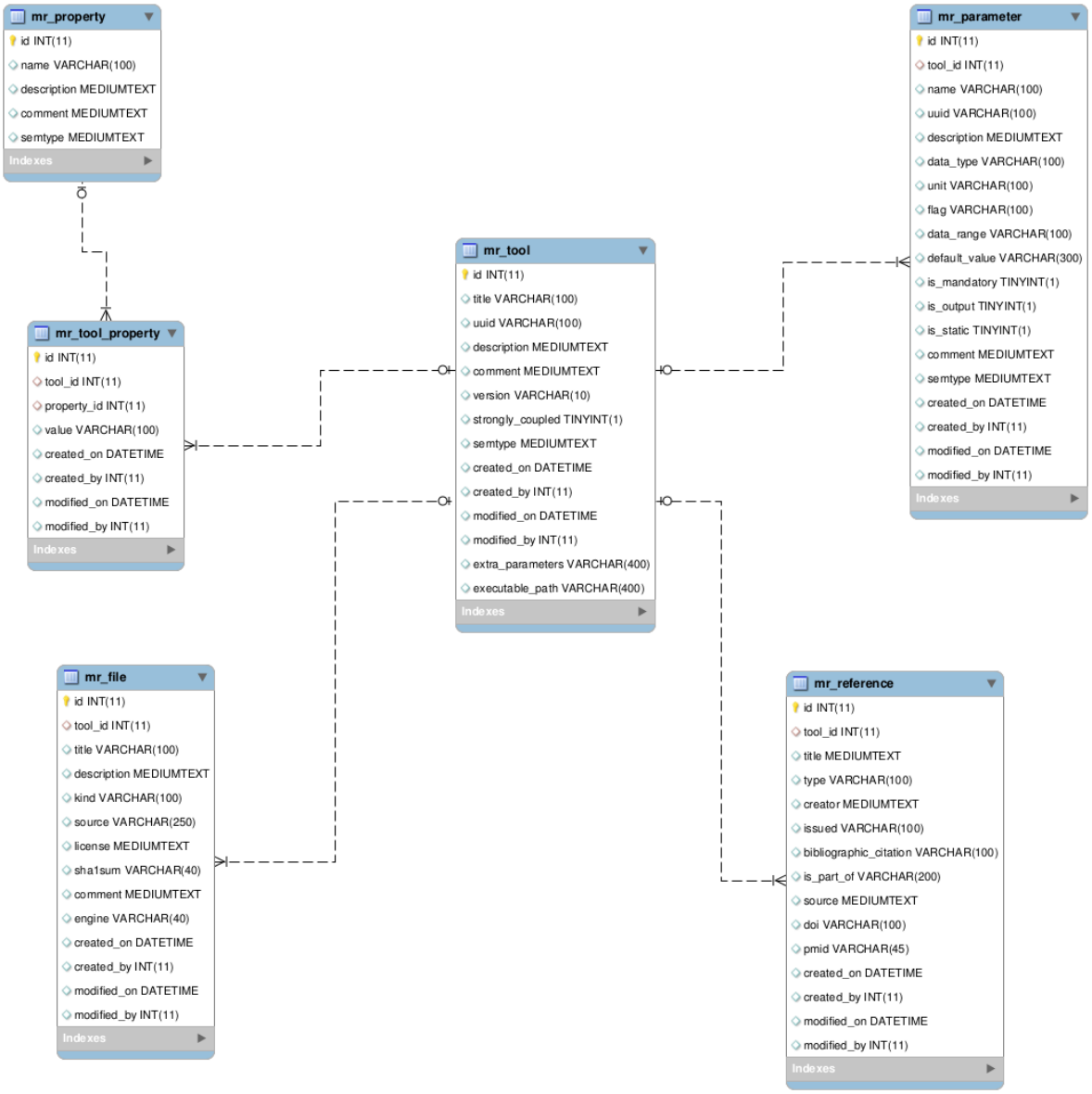

Figure 4. Entity Relationship (ER) diagram of CHIC Model and Tool Repository

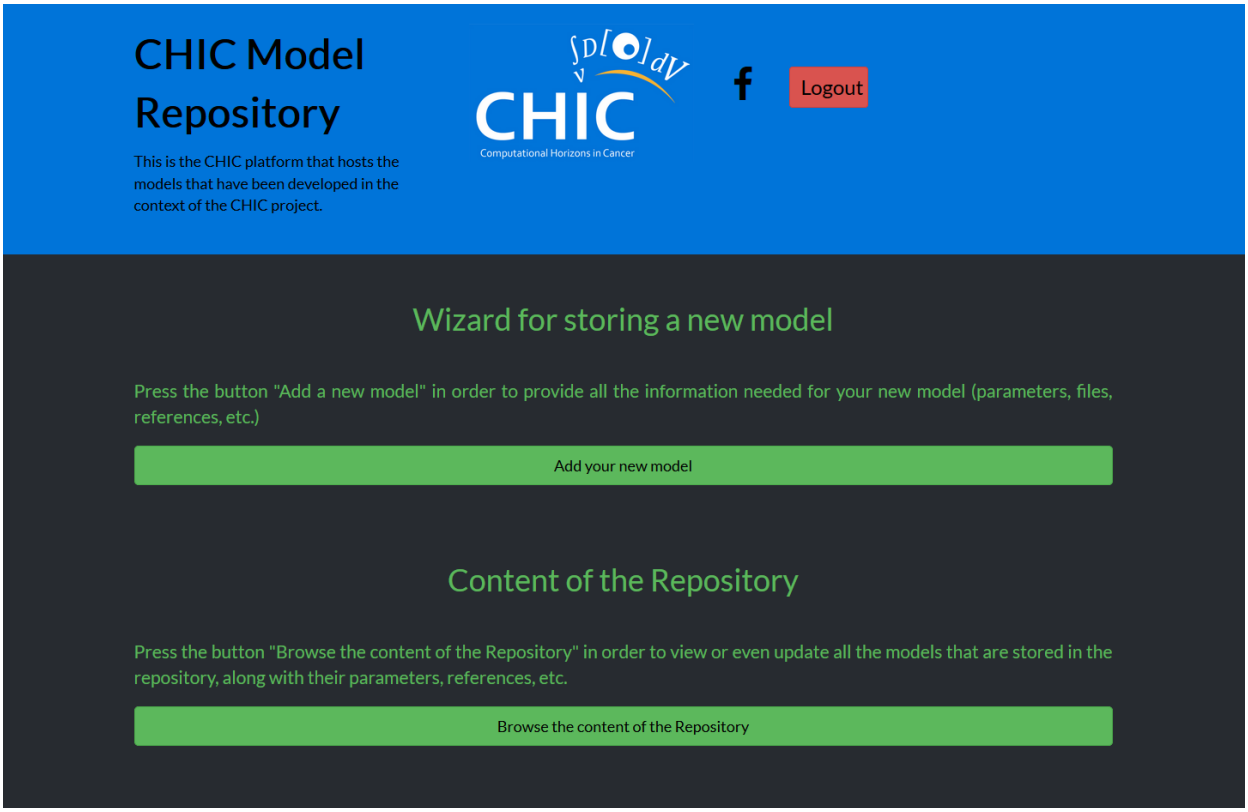

Figure 5. The main page of the CHIC Model Repository

The model repository, makes use of user interface design principles in order to produce a user interface which makes the interaction with the user (researcher, clinician, modeller) self-explanatory, efficient, enjoyable and user-friendly. It has been given special emphasis during the development of the Model Repository to provide a user interface where the user will need to provide minimal input to achieve the desired output and where the Repository will minimize undesired outputs to the user.

Figure 5 displays the main page of the model repository. As shown right after the authentication and authorization processes, the user is able to store a new model through a wizard, or browse the content of the Repository in order to view or even update the models that have been stored. There are also specific workflows for the storage of a new model and the browsing of the content of the Repository.

Since the BOUNCE project will not only focus on mechanistic models, but also on statistical and machine learning methods for the development of the In Silico Resilience Trajectory Predictor in WP4, tools and methods that may be used or the paradigm followed for this category of models has been searched in literature. Generally, in this field researchers typically use an environment of choice (like R, MATLAB, WEKA, RapidMiner etc.) and the sharing of machine learning models is limited to the exchange of tools' specific files. However, this practice makes model sharing and production usually a difficult task. Among the most cited solutions for machine learning models' curation and exchange is the Predictive Model Markup Language (PMML)[80]. The PMML is an XML-based predictive model interchange format providing a way for analytic applications to describe and exchange predictive models produced by data mining and machine learning algorithms. It supports common models such as logistic regression and feedforward neural

[80] http://dmg.org/pmml/v4-3/GeneralStructure.html

networks. It is now developed by the Data Mining Group[81] and supported by a really significant number of machine learning related tools and products[82] including R, apache Spark, WEKA, KNIME, RapidMiner, SAS etc.

Since PMML is an XML-based standard, the specification comes in the form of an XML schema. PMML itself is a mature standard with over 30 organizations having announced products supporting PMML.

PMML is the de facto standard language used to represent data mining models. Predictive analytic models and data mining models are terms used to refer to mathematical models that use statistical techniques to learn patterns hidden in large volumes of historical data. Predictive analytic models use the knowledge acquired during training to predict the existence of known patterns in new data. PMML allows to easily share predictive analytic models between different applications. Therefore, you can train a model in one system, express it in PMML, and move it to another system where you can use it to predict, for example, the likelihood of machine failure.

PMML can represent not only the statistical techniques used to learn patterns from data, such as artificial neural networks and decision trees, but also pre-processing of raw input data and post-processing of model output (see Figure 6).

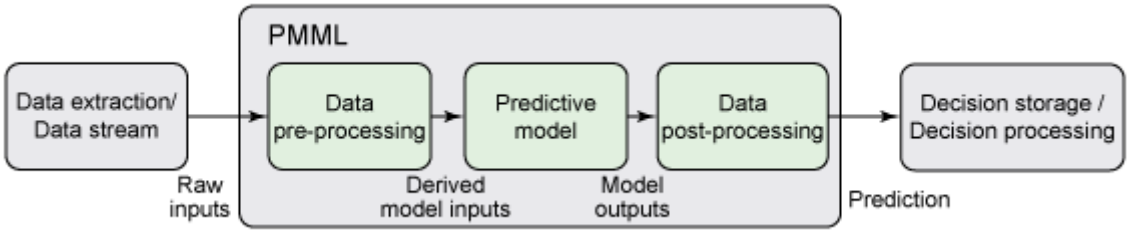

Figure 6. PMML incorporates data pre-processing and data post-processing as well as the predictive model itself

The structure of a PMML file follows the steps commonly used to build a predictive solution:

- Data Dictionary is a product of the data analysis phase that identifies and defines which input data fields are the most useful for solving the problem at hand. These can include numerical, ordinal, and categorical fields.
- Mining Schema defines the strategies for handling missing and outlier values. This is extremely useful since more often than not, whenever models are put to work, required input data fields may be empty or misrepresented.
- Data Transformations define the computations required for pre-processing the raw input data into derived fields. Derived fields (sometimes referred to as feature detectors) combine or modify input fields in order to obtain more relevant information.
- Model Definition defines the structure and the parameters used to build the model. PMML covers a variety of statistical techniques.
- Outputs define the expected model outputs. For a classification task, outputs can include the predicted class as well as the probabilities associated with all possible classes.
- Targets define the post-processing steps to be applied to the model output. For a regression task, this step allows for outputs to be transformed into scores (the prediction results) which humans can interpret easily.

[81] http://dmg.org/

[82] http://dmg.org/pmml/products.html

- Model Explanation defines the performance metrics obtained when passing test data through the model (as opposed to training data). These include field correlations, confusion matrix, gain and lift charts, and receiver operating characteristics (ROC) graphs.
- Model Verification defines a sample set of input data records together with expected model outputs. This ensures that the new system produces the same outputs as the old when presented with the same inputs. Whenever this is the case, a model is considered to be verified and ready to be put to work.

Given that PMML allows for predictive solutions to be expressed in their entirety (including data pre-processing, data post-processing, and modeling technique) its structure and main elements are a reflection of the eight steps outlined above.

It is proposed that PMML Validation requires two steps. The first step is based on XSD Validation as discussed previously. The second step requires that the PMML elements in combination are not only syntactically correct but are also understandable to a properly implemented model consumer. For this a different XML technology is proposed based on XSLT (Extensible Stylesheet Language Transformations). This technology is typically used to translate XML from one form to another. More importantly, it can look across more than one XML element at a time and can be used ensure that key features of PMML are implemented properly. Similarly to the SBML case discussed previously, a set of rules is created that cover particular requirements of the PMML specification. These rules are embodied into XSL transformations and are applied to a particular PMML using an XSLT processor. The result is another document which contains any rules that were violated which can be classified in three categories:

- Rules that validate adherence to the XSD.
- Rules that verify PMML features that are common to all model types are implemented properly.
- Rules that verify PMML features unique to a particular type of model.

To the best of our knowledge there is currently at least one online repository hosting PMML based model. However, the DMG has on its website a PMML Examples[83] page that includes dozens of models based on a handful of public domain datasets.

Complementary to the PMML and also developed by the Data Mining Group, The Portable Format for Analytics (PFA)[84] is a common language for representing statistical scoring engines, the "predict" method of a model. A PFA scoring engine is a JSON file containing model parameters and a scoring procedure. The scoring procedure transforms inputs to outputs by composing functions that range in complexity from addition to neural nets. If a method can be expressed in terms of common data science primitives (arithmetic, special functions, matrices, list/map manipulations, decision trees, nearest cluster/neighbor, and "lapply"-like functional programming) it can be written in PFA (JSON format).

PFA has an open specification developed by the not-for-profit Data Mining Group with implementations for Java, Python, and R. PFA can be compared to the PMML, however, PFA adds the flexibility of arbitrary function composition, rather than choosing from a set of established models.

[83] http://www.dmg.org/pmml_examples/index.html

[84] http://dmg.org/pfa/

Regarding machine learning model collection tools, one of the few contributions found is the ModelDB open source project[85] [86]. ModelDB consists of three key components: native client libraries for different machine learning environments, a backend that defines key abstractions and brokers access to the storage layer, and a web-based visualization interface. ModelDB client libraries are currently available for scikit-learn spark.ml. Data scientists can perform experimentation and model building in their favorite ML environment as usual while, in the background, the client library automatically extracts relevant information and passes it to the ModelDB backend. The ModelDB backend exposes a thrift interface to allow clients to communicate in different languages with the ModelDB backend. ModelDB stores models and pipelines as a sequence of actions (as opposed to states) and uses a branching model of history to track the changes in models over time [4]. The backend uses a relational database to store pipelines while a custom storage engine is used to store and index models. The third component of ModelDB, the visual interface, provides an easy-to navigate layer on top of the database that permits visual exploration and analyses of models and pipelines. Native logging libraries for different ML environments are offered, capturing models built by data scientists along with pre-processing operations performed on the data (e.g., one-hot-encoding, scaling). The data scientist imports the library and initializes the ModelDB syncer. Then by using "sync"-variants of pre-processing or modeling functions the relevant operations and associated data are logged to the ModelDB. In addition to logging pipelines and models, ModelDB also allows data scientists to log annotations or insights about models (e.g., "pipeline with no normalization.")

Based on what was presented in previous paragraphs, related to model collection, curation, and validation, BOUNCE tasks may acquire the solutions or the basic principles of the tools presented. The roots of usage are depicted in Figure 7.

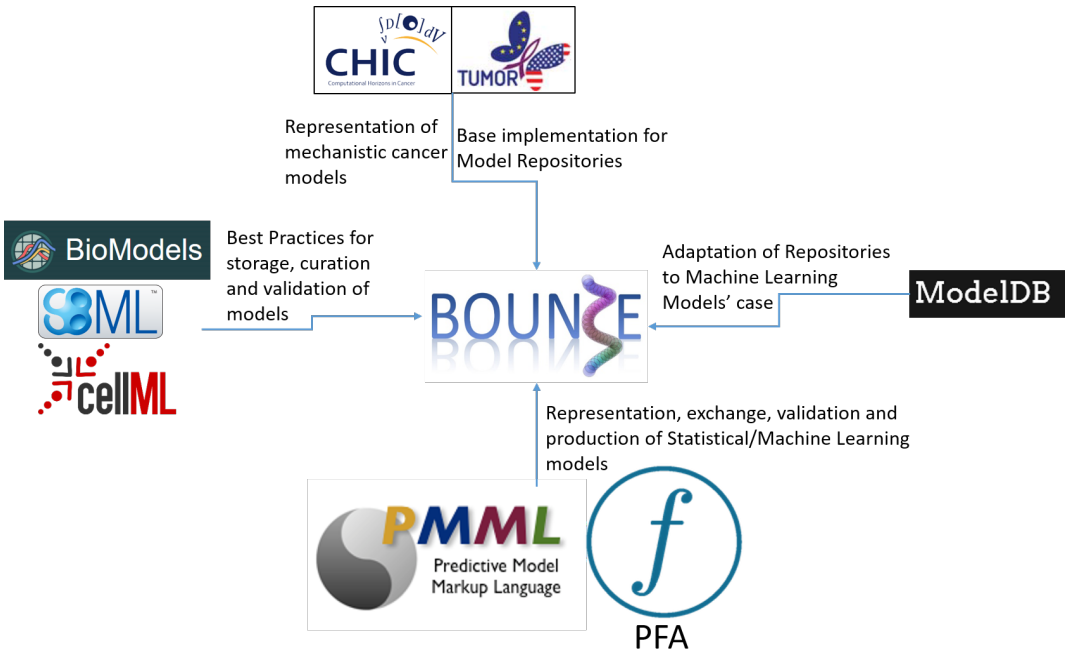

Figure 7. Possible usage approaches of the presented tools in BOUNCE.

[85] https://mitdbg.github.io/modeldb/

[86] https://github.com/mitdbg/modeldb

3.5.2. Statistical and machine learning models

BOUNCE major scope is to develop a comprehensive analysis framework for predicting individual resilience trajectories at different discrete timepoints, as defined in the longitudinal pilot study. Multimodal multiscale data comprised of biological, social, environmental, lifestyle, sociodemographic, and psychosocial data values at different time points will be retrieved for analysis using BOUNCE semantic layer mechanisms. Concerning the primary endpoint of the project, state-of-the-art machine learning techniques will be implemented for predicting and forecasting individual resilience trajectories at a specific time point and across time respectively. Machine learning techniques will be coupled with statistical analysis tools when appropriate, aiming to improve predictive efficacy and to select the most important patient characteristics with respect to the clinical and well-being endpoints. The overall analysis will be developed using freely usable and distributable software libraries written in Python and R and exposed to BOUNCE users via secured web services.

Regarding *single-timepoint* resilience trajectory prediction, a plethora of linear and nonlinear classifiers comprising generalized linear models (GLM), random forests (RF), extreme gradient boosting (XGB), support vector machines (SVM), etc. will be tested under a generalized iterative cross-validation framework using the same input data across all iterations. Preprocessed data, using data cleaning and imputation preprocessing techniques, from all clinical sites will be repeatedly split into independent training and test sets and validated quantitatively using performance metrics such as the area under the receiver operating characteristic curve (AUC), accuracy, sensitivity, specificity, and F1-score. Imbalanced distribution across the examined groups/classes of patients will be handled with respect to the predictive accuracy though proper selection of classifiers (i.e. using tree-based classifiers which are less sensitive to imbalanced datasets), stratified data splitting, and/or using widely used oversampling methods like synthetic minority over-sampling technique (SMOTE). Feature selection techniques relying on filtering (i.e. ReliefF, Mutual Information, non-parametric statistical tests, Gini Index, etc.) and wrapper (i.e. recursive feature elimination) approaches will be followed within BOUNCE predictive analysis framework to reduce “curve of dimensionality” across the examined populations and select the most significant biological, social, environmental, lifestyle, and psychosocial features that contribute to accurate resilience trajectory prediction.

Time series resilience trajectory forecasting will be a major component of BOUNCE methodology, assisting physicians and other health professionals with time-dependent trajectory predictions to account for diverse illness endpoints. Two main methodologies will be examined and tested including: a) conventional classification methods and data reformulation techniques applied to BOUNCE multicenter pilot data over time, and b) deep learning (DL) architectures combining convolutional and recurrent neural networks (RNN). In case of a), similar classification techniques as mentioned in the single timestamp resilience trajectory prediction scenario will be applied to the pilot data. Data restructuring, using sliding window techniques to formulate time series forecasting as machine learning prediction, will be followed. This will be performed prior to proper classification. In this context, autoregressive integrated moving average (ARIMA) and hidden Markov models (HMM) will be used in constructing BOUNCE patient specific time series resilience trajectory predictors using data from all clinical centers. Deep learning architectures have recently shown great success in mental health care problems[87]. Therefore,

[87] Yoshihiko Suhara, Yinzhan Xu, and Alex 'Sandy' Pentland. 2017. DeepMood: Forecasting Depressed Mood Based on Self-Reported Histories via Recurrent Neural Networks. In Proceedings of the 26th International Conference on

BOUNCE will develop a convolutional long-short term memory (ConvLSTM) network to handle the sequential nature of the provided data and incorporate potential dependencies from previous statuses of resilience as recorded throughout the cancer continuum. Techniques such as dropout will be applied to the network to surpass/eliminate overfitting issues.

The overall predictive analysis framework will be applied both to a unified integrated dataset from all clinical pilots and to each clinical pilot independently to assess potential differences in the predictive and output variables across all clinical sites of the BOUNCE pilot study. The selected resilience trajectory predictors, both in a single timestamp and longitudinally, will further provide input to task 4.2.3 (?) Models fusion and integration and the design of the BOUNCE decision system.

3.5.3. Mechanistic Cancer models

Cancer is characterized by invasive, uncontrolled cell growth. The initial cause of abnormal growth of cells are mutations. Therefore, the main area of focus in cancer research has been for many years on cancer genetics, the investigation of the genes involved, and the intrinsic cellular processes they affect and regulate throughout tumorgenesis [57]. However, recent research has shown that cancer cells not only influence the microenvironment around them for their benefit, but also impact the stroma and other non-cancer cells [109], [110]. This is a complex, interactive process, which cannot be easily studied by using conventional lab experiments alone, whether *in vitro* or *in vivo* or both. Mathematical models and computer simulations can help overcome these limitations by offering the ability to simulate and monitor tumour growth, cellular distribution and movement in real time, and also to observe the genetic mutations that lead to aggressive growth and metastasis [30], [109].

Current computational cancer modelling approaches can be divided into three categories: discrete, continuum, and hybrid, i.e., the combination of both [25], [71], [106], [91], [95], [53].

Briefly, discrete models employ experimentally derived, computationally coded rules to define the step-wise or discrete interactions between individual cells and provide insight on tumour microstructure, cell proliferation, death rates, and cell densities.

Continuum models represent the tumour as a continuum and give information about the overall tumour morphology and nutrient distribution while neglecting the influences of individual cells in the environment.

Since discrete and continuum domains are often inescapably linked, directly influencing one another from the viewpoint of an *in vivo* system, the hybrid modelling approach has emerged. It combines aspects of both discrete and continuum modelling to provide a more complete description of the tumour environment.

Decades of cancer modelling have produced established models representing all the key phases of solid tumour growth i.e. avascular growth, tumour-induced angiogenesis, immune response to cancer, invasion and metastasis and vascular growth. New areas are also now being investigated concerning the spatio-temporal modelling of intracellular pathways associated with cancer such as p53-Mdmd2 [86]. A comprehensive overview of the field may be found in the review article of Lowengrub et al. [71]. Especially in the past few years, multiscale models of

World Wide Web (WWW '17). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 715-724. DOI: https://doi.org/10.1145/3038912.3052676

solid tumour growth have been developed in order to account for the different spatial and temporal scales (from genes to tissues) that occur not only in cancer but in all biological systems [86]. A review of recent models in this area may be found in [25]. There has also been a concerted effort to integrate mathematical models of cancer with real data in an attempt to develop quantitative, predictive models of cancer progression and its treatment using chemotherapy and radiotherapy [61], [99], [111], [42], [43], [100], [60], [28].

3.5.3.1. Breast Cancer Modelling

Complex interactions of the various cancer cell populations with the local microenvironment have been simulated in breast cancer via mechanistic modelling. In Boghaert *et al.* 2014 [7] a 2D multi-cell lattice-based model of ductal carcinoma in situ (DCIS) that incorporates cell proliferation, apoptosis, necrosis, adhesion, and contractility is described. It considers four morphologies, namely micropapillary, cribriform, solid and comedo. In Frieboes *et al.* 2009 [37] a mathematical model of tumour drug response that hypothesizes specific functional relationships linking tumour growth and regression to the underlying phenotype is implemented. The model, continuous in nature, incorporates the effects of local drug, oxygen, and nutrient concentrations within the three-dimensional tumour volume, and includes the experimentally observed resistant phenotypes of individual cells. In the previous studies, the estimation of mathematical model parameters has primarily been based on *in vitro* experimental data.

A number of mechanistic models of breast cancer progression and/or response to treatment, designed to exploit clinical data in a patient specific context, have been proposed. In Ubezio and Cameron (2008) [108] a model of tumour progression during treatment is developed. The model is applied to a database of the time course of volumes of breast tumours in patients undergoing preoperative chemotherapy. Ki67 measures, were used to reduce uncertainties in estimates of parameter values in a personalized context. In Macklin et al 2012 [73] a mechanistic, agent-based cell model is applied to simulate growth and calcification of ductal carcinoma *in situ*. This study introduces a calibration methodology that estimates the population dynamic and mechanical parameters of the model based upon IHC for proliferation (Ki-67), apoptosis (cleaved Caspase-3), and morphometric measurements from haematoxylin and eosin (H&E) histopathology images *at a single time point*, i.e. from a single patient biopsy.

The *In Silico* Oncology-*In Silico* Medicine Group of ICCS is specialized in the development of clinically oriented models aiming at supporting patient specific treatment optimization. Two different modelling approaches simulating different treatment modalities applied in breast cancer have been proposed:

- *Breast Cancer Oncosimulator*: A predominantly discrete model addressing the early breast cancer treated by epirubicin according to a branch of an actual clinical trial (the Trial of Principle, TOP trial) is described in [58], [59], [60], [99], [100].The model has been developed to support and incorporate individualized clinical data such as imaging data (e.g., CT, MRI, PET slices, possibly fused), including the definition of the tumour contour and internal tumour regions (proliferating, necrotic), histopathologic (e.g., type of tumour) and genetic data (e.g., p53 status, if available).

- *Vascular tumour growth under antiangiogenic treatment*: A continuum model of vascular tumour growth simulating the time evolution of a breast tumour under bevacizumab mono-therapy, [1]. Although this is not within the context of BOUNCE as it includes metastasized

cancer, it is worth exploring this dimension with the potential available data that will be collected. This approach extends relevant literature [49], [85] by incorporating bevacizumab pharmacokinetic properties and by correcting the asymptotic behaviour of tumour volume in the theoretical case of a total treatment-induced destruction of tumour neovasculature.

3.5.3.2. Breast Cancer Oncosimulator

The Breast Cancer Oncosimulator is a multiscale, four-dimensional modelling approach, comprising a predictive mathematical model capable of simulating existing treatment protocols. Current implementation addresses the case of neoadjuvant chemotherapy with epirubicin but it can be easily adapted to other types of treatment or chemotherapeutic agents. This specific model addresses primary tumours during their clinical course of life, well beyond their initiation phase. It has been designed to incorporate patient-specific data, such as imaging-based, histological, molecular, and treatment data. In the modelling approach, the high complexity of cancer is reflected by key hallmarks of cancer that drive cancer progression that span from intracellular to super-cellular length scales. Hypoxia due to an insufficient tumour neovasculature, reversible dormancy, active proliferation, and spontaneous, starvation-induced or treatment-induced cell loss are among the biological processes considered. Intra-tumour heterogeneity has been implemented via the cancer stem cell (CSC) hypothesis. Different resistance profiles of cancer cells have been taken into account. An essential feature of this model is its capability to simulate the effect of tumour microenvironment on cell cycle dynamics. Different proliferation patterns based on the assumed conditions of tumour microenvironment can be produced. Eventually tumour progression is regulated by the interplay of the above considered biological mechanisms. The modelling approach can also be applied either to primary, metastatic or recurrent tumors to simulate their clinical course of life, well beyond their initiation phase, or to investigate the dynamics and timing of local/distant recurrences after treatment taking into consideration the effect of cancer dormancy.

The Breast Cancer Oncosimulator model is based on the consideration of a discrete time and space stochastic cellular automaton, representing the tumour region. A cubic discretizing mesh is superimposed over the anatomic region of interest. The mesh element is called geometric cell (GC). Each GC corresponds to a cluster of heterogeneous cells found in various states. At each time step, i.e. every hour, the discretizing mesh covering the anatomical region of interest is virtually scanned in order to apply the basic metabolic, cytokinetic, pharmacokinetic/pharmacodynamic, and mechanical rules that govern the spatiotemporal evolution of the tumour system. For practical reasons each complete virtual scan can be viewed as consisting of four mesh scans. The first one simulates primarily cell kinetic model as described below (Figure 7). The second one simulates unloading of the overloaded GC due to mitoses and the creation of new tumour GCs (leading to differentiated tumour expansion), and the third one leads to achievement of the normal density of cells by freeing GCs with few biological cells (leading to differentiated tumour shrinkage). The fourth one aims at preserving a continuous tumour in case that fragmentation takes place. Random directions for differentiated growth expansion and shrinkage are used. The aim of the last three scans is a realistic, conformal to the initial shape of the tumour, simulation of expansion and shrinkage, in the cases of untreated tumour growth and chemotherapy treatment, respectively. Figure 8 depicts the generic cell kinetic model proposed for the case of tumour growth and response to chemotherapy:

- *Free growth:* Tumour propagation is modeled based on the ‘cancer stem cell’ hypothesis and is regulated by the balance between active cell cycle, quiescence, differentiation, and death.

The tumour population comprises the following five cancer cell categories: a. cancer stem cells, which possess an unlimited mitotic capacity, b. cells of limited mitotic potential (LIMP), c. terminally differentiated cells (DIFF) that have lost the ability to divide, d. cells that have died through apoptosis, and e. cells that have died through necrosis. In addition, stem and LIMP cells can be either cycling, distributed in the four phases of the cell cycle (G1, S, G2, M), or quiescent (G0). The rules that describe the transition between the various categories/phases of the cancer cells are depicted in Figure 7. On the top of the developmental hierarchy lie the stem cells that have the ability of self-renewal and differentiation. Two types of stem cell division are allowed: symmetric that gives rise to two stem cells, and asymmetric that gives rise to one stem and one LIMP cell. LIMP cells follow a type of aberrant differentiation pathway. After N_{LIMP} divisions, LIMP cells are assumed to generate the population of the DIFF cells. Cycling cells (stem or LIMP) require a time equal to cell cycle duration in order to progress throughout the active cell cycle. Cellular dormancy is considered to be due to both nutrient deprivation (hypoxia) and lack of growth-promoting stimuli. Proliferating cells found under either one of the aforementioned conditions are assumed to withdraw from the active cycle into a common G0 state upon completion of mitosis. Under conditions of insufficient nutrient supply and oxygenation, quiescent cells can survive for a limited period. Subsequently, they die through necrosis unless the local metabolic conditions stimulate the entry into the cell cycle. Cycling and quiescent cells may die through spontaneous apoptosis. Differentiated cells may die through apoptosis or necrosis. Apoptotic and necrotic cells are assumed to be present in the tumor bulk for a limited period of timetime length, before their final elimination.

- *Treatment:* In the specific modelling approach, chemotherapy is assumed to affect only cancer cells with proliferative capacity, either cycling or quiescent, depending on the cell cycle specificity of the drug. The activation of apoptosis is regarded as the major mode of action of chemotherapeutic drugs against cancer cells at clinical relevant doses. Chemotherapy-induced apoptosis is implemented through the parameter 'cell kill rate' (CKR) that expresses the fraction of cancer cells that sustain lethal damage by the drug(s) and are destined to die. These cells enter a rudimentary cell cycle before the triggering of the apoptotic pathway. The exact time point within the cell cycle when lethally hit cells enter the apoptotic compartment depends on the mechanism of action of the specific drug. As far as the value of the CKR parameter is concerned, two methodologies have been used depending on the nature and the goals of each particular simulation study [27], [61], [100]. The "forward method", uses a priori calculated values of CKR based on values of pharmacokinetic quantities of interest and pharmacodynamic data derived from pertinent literature. For example, an open three-compartment model with elimination from the central compartment can represent epirubicin pharmacokinetics [21], whereas experimental data from cytotoxicity studies can be used to compute pharmacodynamic parameters. Processed molecular data (e.g. biopsy material and/or blood) can also be exploited to perturb the radiobiological or pharmacodynamic cell kill parameters about their population-based mean values. In the "inverse method", the value of the CKR is suggested by the clinical data and the simulation itself; it is the value that -after having selected the values of the remaining model parameters results in good agreement between the evolution of the simulated tumour and that of the real tumour according to the clinical data (the "apparent" CKR). The "apparent CKR" for each particular clinical case, can be thought of as summarizing important

genetic determinants influencing the tumour's response to therapy.

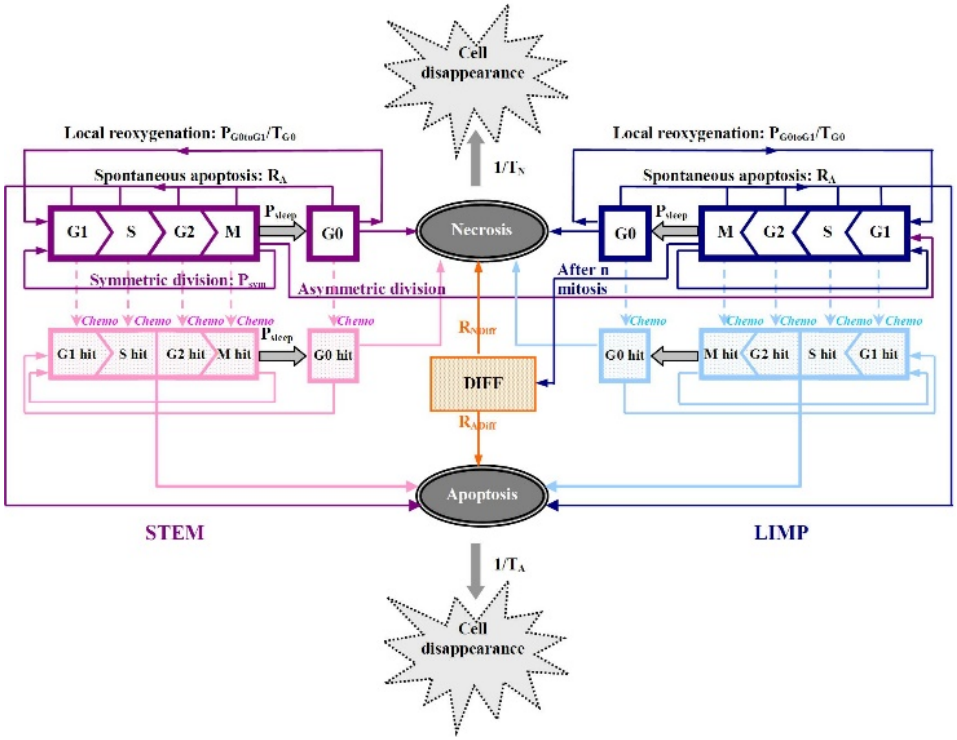

Figure 8. General cell kinetic model for tumour response to chemotherapy. STEM: stem cells. LIMP: Limited proliferative potential cells. DIFF: terminally differentiated cells. G1: Gap 1 phase. S: DNA synthesis phase. G2: Gap 2 phase. M: Mitosis phase. G0: Dormant, resting phase. Chemo: Chemotherapeutic treatment. Hit: Cells lethally hit by the drug.P_{G0toG1}: fraction of dormant cells that reenter the cell cycle, R_A: spontaneous apoptosis rate, P_{sleep}: fraction of newborn cells that enter G0, P_{sym}: fraction of stem cells that perform symmetric division, T_{G0}: duration of dormant phase, T_A: duration of apoptosis, T_N : duration of necrosis, R_{NDiff} : necrosis rate of differentiated cells, R_{ADiff}: apoptosis rate of differentiated cells.

3.5.3.3. Vascular tumour growth under antiangiogenic treatment

All major biological phenomena of cancer cell population dynamics are incorporated into the model i.e. cancer cell proliferation, cancer cell apoptosis, post-vascular dormancy (state where tumour growth ceases due to the balance achieved between pro-angiogenic and antiangiogenic factors), endothelial cell death, spontaneous loss of functional vasculature, excretion of endogenous proangiogenic factors (such as vascular endothelial growth factor, fibroblast growth factors, platelet-derived growth factor, angiopoietin-1 etc.), excretion of endogenous anti-angiogenic factors (angiostatin, endostatin, angiopoietin-2 etc.) and anti-angiogenic treatment – induced endothelial cell death as well as the resulting cancer cell death.

The implicit assumptions on which the basic framework of this specific model is based are that the tumour is a three dimensional spheroid, the diffusion process is in a quasi-stationary state i.e. the tumour growth rate as well as the rate of change of drug concentration are relatively small compared to the rate of distribution of angiogenesis stimulators and the concentration of the stimulator is a radially symmetric function.

The model of [85] makes use of the concept of a variable carrying capacity i.e. the maximal tumour volume that can be supported by the given vasculature which was originally introduced in [49]. The dynamical system described in [85] consists of a pair of ordinary differential

equations (ODEs) which reflect the interplay between tumour volume (*V*) and carrying capacity (*K*).

Regarding the pharmacokinetic properties of bevacizumab, two-compartmental models assuming first-order elimination appear to give the best description of bevacizumab pharmacokinetic data [104], [72], [39]. Taking into account that the specific antiangiogenic agent is administered to human patients via the intravenous route, the case of intravenous infusion has been addressed by applying zero-order absorption, reflecting steady drug delivery into the patient's systemic circulation.

Summarizing, the model of vascular tumour growth under anti-angiogenic treatment consists of three – components: a tumour compartment monitoring the rate of change of the tumour volume *V* under the assumption of Gompertzian growth, a vascular compartment keeping track of the temporal evolution of the carrying capacity of the cancer cell population and finally an anti-angiogenic treatment compartment monitoring the time-course of bevacizumab concentration in plasma based on a two-compartmental pharmacokinetic model. The specific modelling approach has successfully reproduced the results of a series of *in vivo* experiments in mice bearing diverse types of tumours (breast, lung, head and neck, colon) and treated with bevacizumab.

3.5.3.4. Mechanistic modelling in BOUNCE

In BOUNCE tumor progression, response to pre-operative treatment and relapse following surgery and adjuvant treatment will not be directly attempted in the absence of an imageable tumor at two time points, given that the majority of BOUNCE patients will receive adjuvant therapy. However, it may be possible to exploit tumor data from post-surgery residual tumor tissue. These modelling approaches can also be applied in the case of recurrent cases for the simulation of appearance, progression and response to treatment of the recurrent tumor. However, the incidence of tumor relapse during BOUNCE prospective pilot is expected to be very low, because BOUNCE will consider only stages I, II and III and the follow–up period is short (12-18 months). Furthermore, no imaging or volumetric data is planned to be gathered or to be made available for BOUNCE purposes.

In the context of BOUNCE, the developed mechanistic models will serve as a learning and educational environment that will help disseminate the importance of this type of models to the public, to medical stakeholders, industry and funding agencies. The aim is multifold; to explain the biological mechanisms that govern disease progression and response to different types of treatment (health literacy), to demonstrate how mechanistic models can contribute to the battle against cancer and to promote familiarity with the vocabulary that characterize *in silico* cancer modelling. The mechanistic models will operate on synthetic data acquired by literature and will produce a simple set of results.

3.5.4. Models fusion and integration

A fusion strategy will be followed for the combination of the outcomes of each prediction model (i.e. machine learning and statistical models as well as mechanistic models) towards resilience estimation in the context of BOUNCE In Silico Resilience Trajectory Predictor. The overall Trajectory Predictor will effectively combine the psychological, lifestyle and socioeconomic

trajectory (Trajectory I) and the biological and clinical trajectory (Trajectory II). The BOUNCE model aggregation strategy will be applied at the decision level aiming to boost the models' accuracy in both Trajectories. Certain approaches designed to perform model-based fusion will be applied to the most prominent parameters identified for assessing resilience.

The combination of multiple predictive classifiers for more accurate results and the integration of information from multiple modalities to make overall predictions have been studied widely in the literature [10], [107], [11], [3]. In the context of BOUNCE resilience estimation and assessment, different predictive models will be developed based on the heterogeneous data sources (i.e. (a) clinical/biological/genetics, (b) psychosocial data (c) socio-demographic data, (v) lifestyle data and (vi) patient reported information). The generated output of each predictor will be further combined by utilizing weighted average schemes (i.e. linear combination [38] and/or Bayesian Model Averaging (BMA) [36]) as well as machine learning techniques (i.e. Multiple Kernel Learning methods (MKL)) [4], [46]. The utilization of MKL methods within BOUNCE methodology will enable the effective data-dependent kernel combination and the application of kernel-based predictive algorithm for estimating resilience and patient quality of life. In a similar manner, linear and Bayesian combinations of the models' decisions will permit the efficient computation of the resilience score based on different data sources and parameters among the two main Predictors.

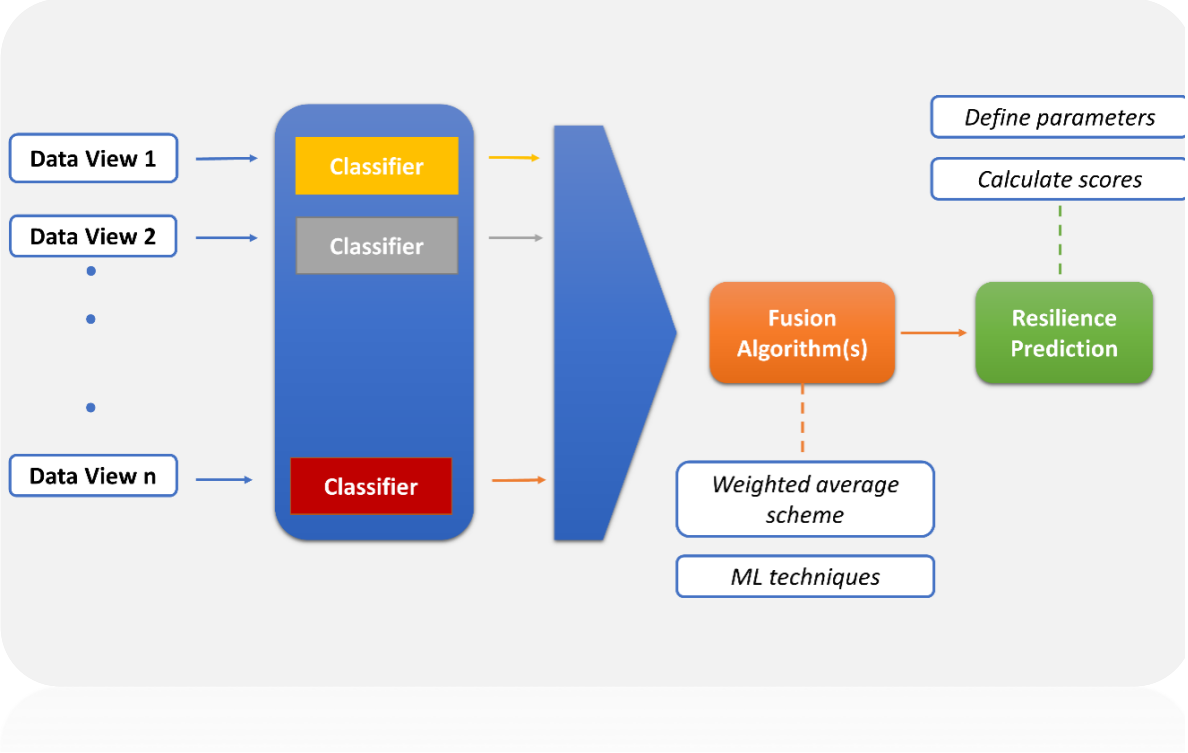

Figure 9. The fusion methodology that will be followed in BOUNCE

Figure 9 illustrates the fusion methodology that will be followed in BOUNCE to combine model decisions and achieve robust outcome predictions related to well-being, functionality, and clinical relapse. Heterogeneous datasets, described in the next chapter, will be exploited by machine learning techniques for predictive modeling purposes. Each predictive model will be based on the exploitation of a certain dataset and specific parameters will be calculated. Subsequently, the decision output of each model will be further combined and an overall

estimation of the resilience score will be extracted. In the case of linear methods, combination rules for classifiers' output fusion will be applied. An integrated predictive model for the overall resilience estimation may also be applied in the results of each developed model. Moreover, in the case of multiple kernel learning method, a kernel-based classifier will be further applied after the effective kernel selection in each data source.

The overall fusion strategy in BOUNCE is considered as the integration of models' outcomes through the utilization of machine learning techniques and/or weighted average techniques. The process and integration of decisions from different data resources when conducting predictive modeling is of utmost importance within the BOUNCE aggregation strategy.

3.6. Temporal Data Mining

The discovery of hidden information in datasets has mainly been focused on association rule mining, data classification, and data clustering. One major problem that arises during the mining process is treating data with temporal feature, i.e. the attributes related with the temporal information present in the database. Traditional data mining techniques would treat temporal data as an unordered collection of events, ignoring its temporal information. However, the temporal attribute requires a different procedure from the other kinds of attributes. The aim of the present section is to present a quick overview of techniques that deal with temporal data mining. They comprise a pool of candidate techniques to be applied in the framework of BOUNCE for both retrospective and prospective datasets.

3.6.1. Prediction

The task of time-series prediction has to do with forecasting (typically) future values of the time series based on its past samples [96]. For this purpose, we need to build a predictive model for the data. The autoregressive family of models can be used to predict a future value as a linear combination of earlier sample values, provided the time series is stationary. Linear non-stationary models like ARMA models have also been found useful in many economic and industrial applications where some suitable variant of the process can be assumed to be stationary. Another popular work-around for non-stationarity is to assume that the time series is piece-wise stationary. The series is then broken down into smaller pieces called "frames", within each of which the stationary condition can be assumed to hold, and then separate models are learnt for each frame. In addition to this standard ARMA family of models, there are many nonlinear models for time series prediction, e. g. neural networks. The prediction problem for symbolic sequences has been addressed in Artificial Intelligence research, regarding various rule models, such as disjunctive normal form model, periodic rule model etc. Based on these models sequence-generating rules are obtained that state some properties that constrain which symbol can appear next in the sequence.

In many cases, prediction may be formulated as classification, association rule finding or clustering problems. Generative models can also be used effectively to predict the evolution of time series.

3.6.2. Classification of Temporal Data

The basic goal of temporal classification is to predict temporally related fields in a temporal database based on other fields. The problem in general is cast as determining the most likely

value of the temporal variable being predicted given the other fields, the training data in which the target variable is given for each observation, and a set of assumptions representing one's prior knowledge of the problem [70].

The sequence classification methods can be divided into three large categories [112].

- The first category is feature based classification, which transforms a sequence into a feature vector and then applies conventional classification methods. Feature selection plays an important role in this kind of methods.

- The second category is sequence distance based classification. The distance function which measures the similarity between sequences determines the quality of the classification significantly.

- The third category is model based classification, such as using Hidden Markov Model (HMM) and other statistical models to classify sequences.

There are three major challenges in sequence classification. First, most of the classifiers, such as decision trees and neural networks, can only take input data as a vector of features. However, there may be no explicit features in sequence data. Second, even though with various feature selection methods we can transform a sequence into a set of features, the feature selection is far from trivial. The dimensionality of the feature space for the sequence data can be very high, and the computation can be costly. Third, besides accurate classification results, in some applications, we may also want to get an interpretable classifier. Building an interpretable sequence classifier is difficult since there are no explicit features.

Over the years, sequence classification applications have seen the use of both pattern based as well as model based methods [96]. In a typical pattern based method, prototype feature sequences are available for each class. The classifier then searches over the space of all prototypes, for the one that is closest or most similar to the feature sequence of the new pattern. Typically, the prototypes and the given features vector sequences are of different lengths. Thus, in order to score each prototype sequence against the given pattern, sequence aligning methods like Dynamic Time Warping are needed. Another popular class of sequence recognition techniques is a model based method that use Hidden Markov Models (HMMs).

Since traditional classification algorithms are difficult to apply to sequential examples, mostly because there is a vast number of potentially useful features for describing each example, an interesting improvement consists of applying a preprocessing mechanism to extract relevant features [2]. One approach to implement this idea consists on discovering frequent subsequences and then using them as the relevant features to classify sequences with traditional methods, like Naive Bayes or Winnow.

Classification is relatively straightforward if generative models are employed to model the temporal data [2]. Deterministic and probabilistic models can be applied in a straightforward way to perform classification since they give a clear answer to the question of whether a sequence matches a given model.

Indicative examples of time series classification involves the use of semi-supervised learning. Semi-supervised learning is an appealing method in areas where labeled data is hard to collect.

Another approach is a Dynamic Bayesian Network (DBN), a Bayesian network which relates variables to each other over adjacent time steps. This is often called a *Two-Timeslice* BN (2TBN) because it says that at any point in time T, the value of a variable can be calculated from the internal regressors and the immediate prior value (time T-1).

3.6.3. Temporal Cluster Analysis

Temporal clustering targets separating the temporal data into subsets that are similar to each other and are able to represent the different sequences. There are two fundamental problems of temporal clustering: to define a meaningful similarity measure between sequences, and, to choose the number of temporal clusters (if we do not know the cluster numbers).

Considering that *K* is known, if a sequence is viewed as being generated according to some probabilistic model, for example by a Markov model, clustering may be viewed as modeling the data sequences as a finite group of *K* sequences in the form of a finite mixture model. Through the EM (Expectation Maximization) algorithm their parameters could be estimated and each *K* group would correspond to a cluster [2]. Learning the value of *K*, if it is unknown, may be accomplished by a Monte-Carlo cross validation approach.

A different approach proposes to use a hierarchical clustering method to cluster temporal sequences databases [2]. The algorithm used is the COBWEB, and it works on two steps: first grouping the elements of the sequences, and then grouping the sequences themselves. Considering a simple time series, the first step is accomplished without difficulties, but to group the sequences is necessary to define a generalization mechanism for sequences. Such mechanism has to be able to choose the most specific description for what is common to different sequences.

Other method proposed in literature for clustering time series data utilize fuzzy logic. Fuzzy clusters provide the flexibility of allowing an object or changes in time series variables to participate in multiple clusters.

3.6.4. Temporal pattern discovery - Association Rules

Temporal pattern discovery deals with the discovery of *temporal patterns of interest in time series or temporal sequences*, where the interest is determined by the domain and the application. For example: patients who are on drug X for over a month sometimes start suffering from severe headaches after a month. This is a temporal association rule, but also a potentially *causal* rule [81].

The discovery of relevant *association rules* is one of the most important methods used to perform data mining on transactional databases [96]. An effective algorithm to discover association rules is the *apriori*. Association rule discovery is an important task in data mining in which we extract the relation among the attribute on the basis of support and confidence. The association rule discovery can be extended to temporal association. However, the manipulation of temporal sequences requires that some adaptations are made to the *apriori* algorithm.

The presence of a temporal association rule may suggest a number of interpretations [92].

- The earlier event plays some role in causing the later event.
- There is a third (set of) events that cause both other events,
- The confluence of events is coincidental.

The first interpretation is associated with the concept of causal rule, i.e. a relationship in which changes in one part of the modeled reality cause subsequent changes in other parts of the domain. Causal rules are common targets of scientific investigation within the medical domain, where the search for factors that may cause or aggravate particular medical conditions is a fundamental objective. In this domain, KDD (Knowledge Discovery in Database) tools can be applied at a preliminary stage, namely, to discover associations that can be observed as candidate causal rules. The tests for causality follow in a subsequent stage, involving expert guidance and extensive statistical tests [92].

While the concept of association rule discovery is the same for temporal and non-temporal rules, algorithms designed for conventional rules cannot directly be applied to extract temporal rules [92]. The reason is that classical association rules have no notion of order, while time implies an ordering. This ordering affects the statistical properties of the data and the semantics of the rules being extracted from them. Moreover, patients are associated with both static properties, such as gender, and temporal properties, such as age or current medical treatments, any or all of which may be taken into account during mining.

Fuzzy temporal association rules arise from the use of fuzzy sets to describe quantitative temporal and/or not temporal attributes of items in a database, and/or to introduce fuzzy temporal specifications [15].

3.7. Psychoemotional assessment tools and models

There have been several attempts to model disease progression using advanced computational tools. In view of the extremely high complexity of the phenomenon that is modelled (encompassing clinical, physiological, molecular, lifestyle and psychological components), the majority of studies to-date have intentionally restricted their scope to a limited set of predictors and outcomes with a special focus on survival rate and risk of metastasis [77].

Another set of studies have primarily focused on patient well-being. The NIH Toolbox Emotion Battery [51] is considered as state of the art for emotion measures and consists of surveys assessing Positive Affect, General Life Satisfaction, Emotional Support, Friendship, Loneliness, Perceived Rejection, Perceived Hostility, Self-Efficacy, Sadness, Perceived Stress, Fear, Anger, Meaning and Purpose and Instrumental Support. Disease-specific surveys are also provisioned, such as the Neuro-QoL (assessing physical, mental, and social effects experienced by adults and children living with neurological conditions) and the PROMIS [31] (Patient-Reported Outcomes Measurement Information System, comprising person-centered measures of physical, mental, and social health).

Models focusing on the well-being and functionality of breast cancer patients are limited. Some prospective studies have examined potential correlates of clinical recovery as a function of well-being. In Table 2 the most prominent studies focusing on breast cancer patients and highlighting elements of resilience are summarized. It is obvious from all these studies that social support, family resilience, stress-levels, diet and exercise are correlated to resilience and quality of life.

BOUNCE will incorporate the most acceptable/mature measures and outcomes in order to define the optimal prediction measures for understanding illness adaptation in breast cancer. Members of the consortium have the expertise and can provide mature predictive models for well-being, survival, risk of metastasis, persistent pain and cognitive profiling. HUS, the

coordinator of the BOUNCE project, has already a clinically validated and mature toolbox for predicting patient-specific risk of sentinel node and nonsentinel node metastases in breast cancer. The specific models of HUS are also available as an application in the app store Predictive Tools for Breast Cancer [88]. The ALGA-C and ALGA-BC (breast cancer psycho-emotional profiling) have been designed and clinically validated by the European Institute of Oncology in Milan and developed by FORTH as a predictive model for cognitive outcomes.

Tool name	Characteristics	Measures	Outcomes
HUS Predictive tools in Breast Cancer [78]	Predictors of patient-specific risk of sentinel node and nonsentinel node metastases in breast cancer.	Clinical, Histoathological, Biomarkers	risk of axillary metastases, nonsentinel & sentinel node metastases
HUS persistent pain		Clinical	persistent pain
ALGA-C questionnaire [63]	Psycho-emotional evaluation	Psycho-emotional, cognitive	Cognitive capacity
EORTC QLQ-C30 [94]		Demographics, Clinical, QoL	Health-related quality of life
Schlegel *et al.* [94]	Predictors of psychological health among breast cancer patients	Demographics	Depressive symptoms
Donoval *et al.* [29]	Post- cancer treatment longitudinal changes in fatigue	Demographics, Clinical, Psychosocial	Fatigue Following Breast Cancer Treatment
Dirksen, Shannon Ruff [26]	Predicting well-being among breast cancer survivors	Health care orientation, Uncertainty, Social support, Resourcefulness, Self-esteem	well-being
McDonough *et al.* [76]	Social support in relation to posttraumatic growth (PTG) and subjective well-being	Demographics, Social, stress, Cancer worry, Posttraumatic growth, Subjective well-being	posttraumatic growth and subjective well-being

Table 2. Predictive tools for well-being of breast cancer patients.

3.8. Evaluating resilience in existing medical practice

In this section we briefly describe ongoing procedures in the clinical management of breast cancer patients at each of the four BOUNCE pilot sites.

3.8.1. HUS

In HUS routine psychosocial assessments are not routinely conducted. Resilience estimation is based on the oncologist´s or nurse´s clinical experience. If they recognize the need for psychosocial support, the patient is referred to the psychosocial unit consisting of psychiatrists, psychologists, and a nurse trained on basic counselling techniques. Normal patient records are used to report patient´s psychosocial wellbeing, and psychiatrist uses Warwick-Edenburg Mental Well-Being-Scale for some patients.

Patient support is based on personal contact and supportive discussions. The number and timing of these sessions varies depending on individual needs. In an acute psychiatric crisis or illness patient is referred to the HUS Psychiatry Department. If patients use Noona they can report all their symptoms (including cognitive and psychological complaints) to Noona. Otherwise the possible information is only in the patients records and not easily available.

3.8.2. HUJI

The Israeli health system is predominantly a public service. Israel is a multiethnic society and hospitals are hiring social workers from various backgrounds. The guiding principles of the oncology social work model are to provide a continuum of therapies including individual, family, and group, as well as to provide ad hoc responses to urgent problems. Social work interventions within the contact of oncology are usually brief and include aspects of practical problem solving and supportive therapy, and tend to maintain professional contact with patients and family members over long periods of time.

The therapeutic goals of oncology social work are:

- To encourage emotional reflection and emotional growth
- To set achievable goals
- To assist in the rehabilitation process and to help the patient find meaning
- To address the family needs and particularly the needs of the more vulnerable ones (e.g., the patient's children)
- To facilitate the interpersonal communication and the continual dialogue with the medical staff and, if needed, to translate medical language into the patient's narrative.
- To assist the patient and family in coping with advanced illness and expected death.
- To help solve day-to-day problems brought about by the illness and treatment.

Social workers also develop and lead group interventions for women recovering from breast cancer.

As currently there are no available group interventions aimed at building resilience, previous research on the outcomes and effectiveness of this group intervention showed an improvement in the patients' quality of life. Most patients feel more confident they can learn to live with early-

stage breast cancer better. However, the transition back to normative life is often more challenging than they expect. The intervention included psychoeducation, resilience-building techniques, general empowerment, help in maintaining social relationships and developing an effective support network, and cognitive behavioral therapy to challenge negative thoughts and to encourage behavioral activation. A translated version of the intervention was provided for Arabic-speaking women. The intervention was provided free of charge and was offered at least two months after completion of chemotherapy. All women were invited to participate; about one-third of the women treated choose to attend the group.

In addition, In Israel there is an active psycho-oncology society that works cooperatively with social workers and psychiatrists, nurses as well as spiritual counsellors. All these health professionals work cooperatively to maximize the options for supporting both patient and family. The local hospitals have their own system of how to organize and utilize these various options for helping patients.

3.8.3. CHAMP

At the Champalimaud Clinicial Center breast cancer patients are initially assessed by a nurse specialist who conducts a brief psychosocial evaluation. Therapeutic education and information about the patient supportive group *mamahelp* is provided.

Any issue, mainly related with psychological distress will be reported to the oncologist and psycho-oncologist. A referral to a neuropsychiatry consultation will be considered after the psycho-oncology assessment. The neuropsychiatry unit will provide assessment by a psychologist, psychotherapist or psychiatrist. No comprehensive assessment of resilience is yet performed at the clinical center.

3.8.4. IEO

No direct measure of resilience is currently used in daily practice at IEO. A direct measure of resilience has been collected only in a specific clinical trial (iManageCancer). In order to understand the patient's psychological state and how she is facing with illness and treatments, the Distress Thermometer (DT) is administered at the admission to and at the discharge from the hospital, and whenever a salient event occurs that motivates the distress assessment. Relatively to outpatients undergoing Chemotherapy, the DT is administered at the admission at the department, before starting the therapy. The DT is an integral part of the Oncology Nursing Minimum Data Set (ONMDS), a nursing record used in daily practice for the assessment of the patient's outcomes that used for planning the intervention coherent with the actual patient's situation. The DT informs nurses on the general distress level and on problematic areas (practical, emotional, family-related, spiritual and physical problems) that may impair the patient's quality of life and her way to face with therapy. Depending on the patient's distress level, the nurse decides whether or not to propose the patient with the psychological support and, if accepted, to activate it. The psychological intervention will be included in the nursing plan as taken action to solve the critical psychological outcome.

The activation process of the psychological support

- If patient's distress level indicated on the DT is moderate to severe (score equal or above 5) the patient is informed about the possibility to receive psychological support by psychologists/psychotherapists of Psycho-Oncology Division of the Institution. If the patient

accepts, a detailed psychological assessment is performed, in order to identify the patient's needs and the possible intervention to respond to them.

- Patients can ask directly for psychological support as well in any moment independently on the DT outcome.
- Patients who undergo major surgery (e.g., patients whose possible treatment could be a risk reducing mastectomy) are systematically reported to psychologists who will ensure that patients have received the main and necessary information related to the care and surgery plan and to support the decision making process.

With regards to patients with: personal history of mental health problems, anxiety and depressive symptoms, behaviour disorders, somatic symptoms, psychotic conditions, delirium, Substance use disorder and mental disorder due to a general medical condition, they must be reported to the Psycho-oncology unit in order to decide, in cooperation with the multidisciplinary team, the best clinical intervention (psychological interview, psychiatric and/or neurological counseling, etc.).

Due to the multidisciplinary approach, the main information is shared with nurses and physician even if no regular meetings are scheduled, except for risk reduction interventions, for which a weekly meeting is arranged in order to report patients who will undergo prophylactic surgery.

4. The BOUNCE Methodology

In this chapter we present the BOUNCE methodology, as it has been updated based on its initial description available in the DOW. We expect that this methodology will be further refined and updated as the project progresses, however the current deliverable reflects its current state. The workflow and the methodology is depicted in Figure 11 and is composed of the following six steps:

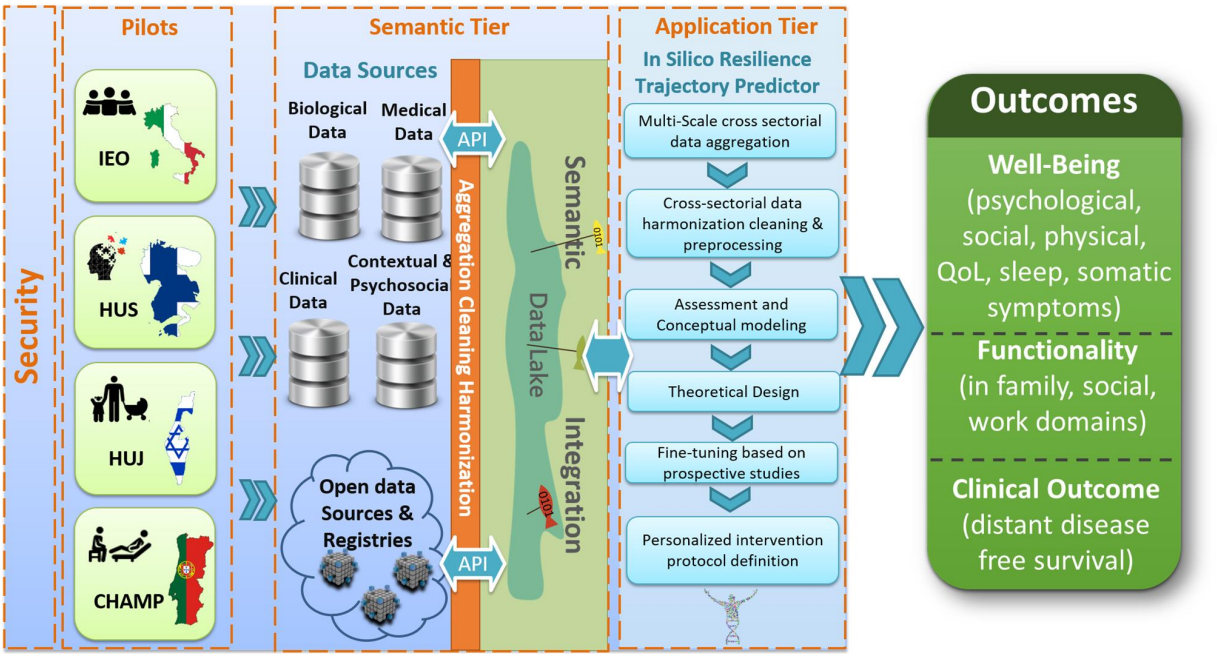

Figure 11. The BOUNCE methodology

4.1. Step #1 Multi-scale, cross-sectional data aggregation

Retrospective sociodemographic, clinical (pertaining to cancer type, treatment characteristics, and medical outcomes) and essential well-being data from existing registries accumulated by clinical partners over several years of clinical practice on large numbers of breast cancer patients will be aggregated during the first 12 months of the project. This data will be used for identifying the most prominent variables affecting resilience as described in step #3 below. This, together with the theoretical work (transtheoretical model definition) will define the initial set of parameters for modelling resilience and conducting the prospective, multicentre pilot study.

Vast amounts of heterogeneous multi-scale data are already available to all four BOUNCE clinical partners (HUS, IEO, JUJ, CHAMP), including

a) clinical/biological/genetic data: genetic risk factors (i.e. BRCA 1/2, triple negative), epidemiological factors (early age menstruation and late age menopause, no pregnancy or first pregnancy after the age of thirty, little or no breastfeeding, obesity), imaging data, type and timing of treatment and medication (chemo, hormonal, antibody, radiotherapy, different agents and radiotherapy doses), patient reported symptoms, tumour biology and type, age, basic laboratory tests, no. of visits to various carers and emergency units, survival, psychotropic drugs received etc.

b) Psychosocial data: Life events, and stressors, health-related Quality of Life, perceived social support, counselling and support sessions received, PTSD, Depression, Coping (CERQ) Flexibility and Posttraumatic Growth, Distress Thermometer, etc.

c) socio-demographic data: age, gender, family history, family status, working status, level of education, insurance status, absence from work, no. of disability pensions, etc.

d) lifestyle data: alcohol consumption, smoking (past or current), physical exercise.

Additional data collected from HUS patients as of January 2017 through the current version of the Noona tool will also be available to BOUNCE, mainly focusing on patient-reported information on pain, fatigue and weakness, changes in mood or emotions, stomach and bowel symptoms, respiratory symptoms, reduced muscle strength or numbness in legs, mental performance, changes in general state of health and many more.

4.2. Step #2 Cross -sectional data variable harmonization, cleaning / pre-processing of the available retrospective data

In cancer research the heterogeneity of relevant data and knowledge sources, the diversity of stakeholders, as well as the disparity in the required predictive models, pose a series of obstacles. For that reason, the BOUNCE architecture will devote significant effort toward harmonization of data sources and their alignment in a single environment that efficiently and effectively streamlines heterogeneous data. The semantic interoperability layer will provide access to a rich range of data with the aid of widely adopted ontologies in order to enable software components of BOUNCE to share meaning with their data. BOUNCE will follow an iterative and incremental process of software development. Short iterations will help keep quality under control by driving to a releasable state frequently, which will prevent the project from collecting a large backlog of defect correction work. Refinements of the architecture will take place during the entire lifetime of BOUNCE as a result of iterative feedbacks from stakeholders.

While these issues are particularly acute when it comes to analysing retrospective data from existing registries, main pilot data will require integration across sites, but there will be minimal need for harmonization given that pilot data will consist of a common set of variables (resulting from a common, agreed-upon in advance, set of measures and the set of clinical/biologic variables in each practice that is common across sites).

Privacy and data protection constitute core values of individuals and of democratic societies. In BOUNCE a privacy and security layer will wrap the whole architecture, data flow and interactions. Security will be considered from the very beginning of the system development based on the “Privacy by Design” [16], which ensures that privacy will be taken into account throughout the entire engineering process from the earliest design stages to the operation of the productive system.

4.3. Step #3 Assessment and conceptual modelling of resilience

Analyses on the harmonized cross-sectional data will help refine the assessment protocol with respect to the types of variables and, when available, the types of measures used to assess core outcomes (i.e., wellbeing, functionality, clinical relapse). Significant correlates of these outcomes, identified through analyses of the cross-sectional data, will be complemented by

several additional measures prompted by our transtheoretical model of resilience and illness adaptation. These measures include a series of *brief self-report scales and patient diaries* to assess distress levels, daily life stressors, illness representations, coping strategies, and personal / lifestyle factors. Indices of *biological processes* that have been identified in recent research as potential contributors to medical breast cancer outcomes (such as stress hormones, plasma neuropeptide Y (NPY), dehydroepiandrosterone (DHEA) levels, lymphocytes), and clinical indices and events (i.e., treatment types, psychotropic drugs received and schedules) will also serve in computational models as predictors of illness outcomes.

BIOMEDICAL STATUS (BMS)	PSYCHOSOCIAL STATUS (PSS)	FUNCTIONAL STATUS (FUS)	(TENTATIVE) RESILIENCE IN RESILIENCE DEGREES (RD)
GOOD	GOOD	GOOD	10
GOOD	GOOD	INTERMEDIATE	9
GOOD	GOOD	BAD	8
GOOD	INTERMEDIATE	GOOD	9
GOOD	INTERMEDIATE	INTERMEDIATE	8
GOOD	INTERMEDIATE	BAD	6
GOOD	BAD	GOOD	7
GOOD	BAD	INTERMEDIATE	6
GOOD	BAD	BAD	5
INTERMEDIATE	GOOD	GOOD	7
INTERMEDIATE	GOOD	INTERMEDIATE	7
INTERMEDIATE	GOOD	BAD	6
INTERMEDIATE	INTERMEDIATE	GOOD	7
INTERMEDIATE	INTERMEDIATE	INTERMEDIATE	5
INTERMEDIATE	INTERMEDIATE	BAD	5
INTERMEDIATE	BAD	GOOD	5
INTERMEDIATE	BAD	INTERMEDIATE	5
INTERMEDIATE	BAD	BAD	4
BAD	GOOD	GOOD	4
BAD	GOOD	INTERMEDIATE	4
BAD	GOOD	BAD	3
BAD	INTERMEDIATE	GOOD	4
BAD	INTERMEDIATE	INTERMEDIATE	3
BAD	INTERMEDIATE	BAD	2
BAD	BAD	GOOD	3
BAD	BAD	INTERMEDIATE	2
BAD	BAD	BAD	1

Table 3. A tentative and hypothetical numerical quantification of resilience for the various combinations of the BMS, PSS and FUS statuses. The precise values of resilience in Resilience Degrees (RD) - in a scale of 1 RD to 10 RD - for each BMS, PSS and FUS combination will be one of the outcomes of the implementation of the BOUNCE project. It is noted that the values or characterizations of the three statuses of the patient can generally refer to the same and/or different time points. More refined gradings of the three statuses can be adopted.

An brief overview of the proposed generic model toward quantification of resilience is presented below (details on the conceptual modelling of resilience can be found in Deliverable D4.1). Three broad categories (clusters) of patient's data (biomedical, psychosocial, functional) are considered as key determinants of resilience. An abstract conceptual approach to the quantification of resilience as a function of the biomedical, the psychosocial and the functional statuses of the patient is proposed through the use of a simple diagram in Table 3. The values or characterizations of the three statuses of the patient can generally refer to the same and/or different time points. Table 3 also outlines a tentative and hypothetical preliminary numerical quantification of resilience for the various combinations of the BMS, PSS and FUS statuses. The precise values of resilience in "Resilience Degrees (RD)" – in a scale of 1 RD to 10 RD - for each BMS, PSS and FUS combination are to be determined by the BOUNCE project by applying a host of statistical and machine learning methods on the available data and especially on the data to be generated by the prospective BOUNCE pilot study.

If the resilience value is above a threshold (which might be e.g. 7 RD) no clinical action might be required. If it lies between say 5 RD and 7 RD, then a light action might be required (e.g. more physical exercise, more strict diet etc.). If it is say below 5 RD then "emergency" measures might need to take place (e.g. psychological interventions, psychiatric examinations, further biomedical examinations and tests etc.).

Moreover, BOUNCE adopts a comprehensive approach to addressing the complex and inconsistently-defined concept of resilience in the extant literature. Specifically, based on the results of computational modeling of the (prospective) pilot data, the relative clinical value of two complementary operational definitions and corresponding measures of resilience will be evaluated. In the psychometric approach, a widely used brief scale will provide a subjective measure of each patient's self-perception of personal qualities that have been shown in previous research to be crucial for illness adaptation. In the data-driven approach of BOUNCE resilience will be identified from our data as a latent variable encompassing the degree of change in several variables indexing emotional and physiological distress. As mentioned, the composite resilience trajectory will be allowed to vary in relation to each of the predictor trajectories outlined in the previous paragraph. A comprehensive approach to the definition of resilience is available in deliverables D2.1 and D2.1 of the BOUNCE project.

4.4. Step #4 Explicit modelling of the resilience trajectory as a function of multi scale data and fine-tuning of the prediction models

A continuous interaction between the modelled psychological, lifestyle and sociodemographic trajectory (Trajectory I) and the modelled biological and clinical trajectory including long term treatment (Trajectory II) will take place, as illustrated in Figure 12. This interaction will involve both the retrospective and the prospective data.

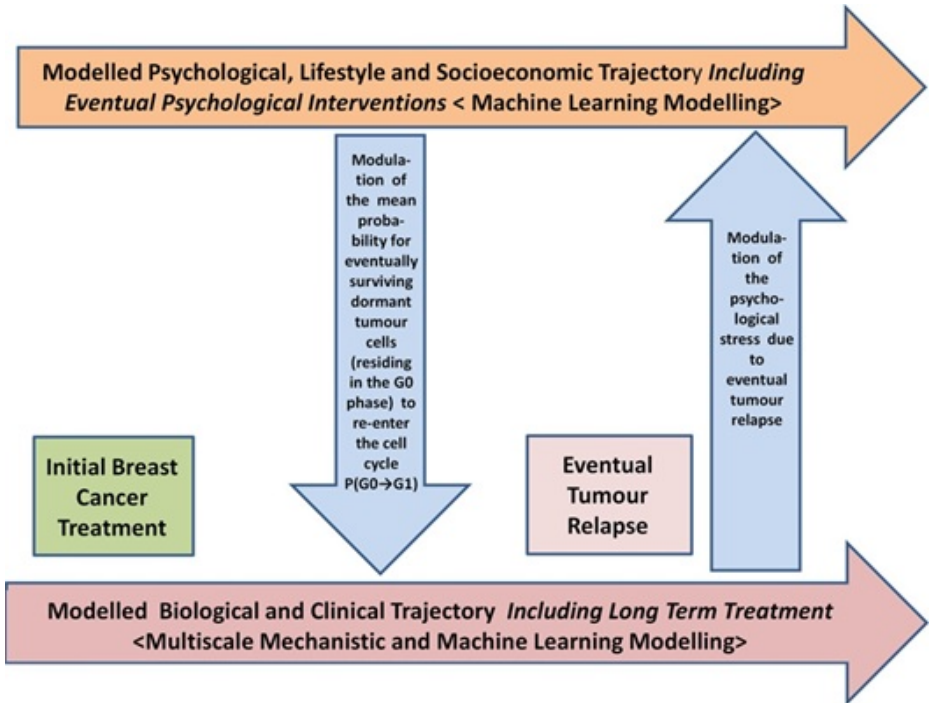

Figure 12. Continuous interaction between the modelled psychological, lifestyle and socioedemographic trajectory (Trajectory I) and the modelled biological and clinical trajectory including long term treatment (Trajectory II).

Trajectory I will be modelled primarily through machine learning techniques whereas Trajectory II will be primarily modelled through multiscale mechanistic modelling in conjunction with machine learning techniques.

Model fusion will take place at the decision level using supervised learning methods. Furthermore, the interaction between the two trajectories will take place through particular parameters. The value of such a parameter generated by one trajectory will modulate the course of the other trajectory. For example Trajectory I will modulate the mean probability for eventually surviving dormant tumour cells (residing in the G0 phase) to re-enter the cell cycle P(G0→G1) and therefore will act as a modulator of Trajectory II. Conversely, an eventual tumour relapse generated by Trajectory II will modulate Trajectory I in leading to increased psychological distress.

Multiscale and multifactor data to be generated by the BOUNCE multicenter clinical pilots will be used in order to adapt both the trajectory models and their interaction parameters. To this end extensive use of machine learning techniques such as Bayesian networks, Random Forests and Support Vector Machines will be exploited. Both trajectories will update the resilience trajectory. The final model will be refined to optimize prediction of clinical, well-being, and functionality endpoints. Importantly, the *weights* of resilience predictors will be further adjusted based on the capacity of separate resilience variables to predict actual (observed) illness outcomes.

4.5. Step #5 Risk predictor and in-silico decision-support system

The BOUNCE Model Repository is the web-based component that will permanently host the models developed in the context of BOUNCE. More specifically, both the psychosocial/behavioral trajectory models and the biological/medical trajectory models will be

stored in the BOUNCE Model Repository. Moreover, the in silico resilience trajectory predictor (RTP) which will serve as decision support system for predicting resilience evolution in women with breast cancer throughout cancer treatments and recovery, will also be stored in the BOUNCE Model Repository. The design and development of the repositories will follow closely the development of the models.

For each model the BOUNCE Model Repository will contain all the related information including:

- descriptive information for each model (abstract and detailed description, references, etc.)
- information related to the model input parameters needed for the execution of the model (data type, units, description etc.)
- information related to the output data of the model (description, type, etc.).
- several versions of binaries

A *web-based user interface* will be developed by using cutting edge front-end technologies in order to allow users to interact with the Repository. Web services are also going to be developed for the integration of the Model Repository with the other components. The aforementioned integration will facilitate the retrieval of the Repository information (model executables, descriptive information of the models, etc.).

Appropriate *authentication and authorization mech*anisms will be implemented in order to ensure that only authorized persons have access to the content of the repository.

The BOUNCE In Silico Trial and Prediction Repository will be a web-based application, capable of persistently storing the predictions of the models developed within the BOUNCE project. Since the in silico predictions may require many computational resources, especially when the simulations involve multiscale data, the development of an In Silico Trial and Prediction Repository in the context of the BOUNCE project is of utmost importance. The input data of each simulation (biological status, medical information sets, clinical information sets, contextual and psychosocial information sets, etc.), the model used in the simulation, and the output data will be stored persistently after the completion of the simulation scenario. Moreover, since the two trajectories (psychosocial/behavioral and biological/medical) interact with each other through particular parameters, the values of those parameters in each simulation will be stored in the In Silico Trial and Prediction Repository along with the corresponding input data. Information related to the input (biological markers, medical imaging, lifestyle, psychological status, etc.) and the output (predicted psychological status, biological status or level of resilience of women with breast cancer) of all the simulations conducted using the in silico resilience trajectory predictor (RTP) will be readily available through the In Silico Trial and Prediction Repository for evaluation, comparison and validation. Consequently, since all predictions will be stored in the Repository, there will be no need for executing the same simulation again.

Just like the BOUNCE Model Repository, a user-friendly web interface and appropriate web services will be developed for the In Silico Trial and Prediction Repository in order to expose its content to the users or other software components. Furthermore, pertinent authorization and authentication mechanisms will deny any unauthorized access.

4.6. Step #6 Decision support for personalized intervention

The end-product of BOUNCE will be a set of clinically validated, in silico resilience trajectory prediction algorithms, one for each of the three crucial outcomes assessed in the prospective, multicenter clinical pilot (clinical relapse, physical/psychological well-being, functionality). These algorithms will provide reliable indications to health professionals to enable the translation of the conclusions of the resilience trajectory models, which were refined using the clinical pilot data, into the initial planning and, when necessary, ongoing adaptations of personalized interventions (psychological, medical, lifestyle, etc.) ultimately aiming to enhance the capacity of individual breast cancer patients to efficiently adapt and resume a full life.

The BOUNCE Pilot Study aims at extending this scope by identifying, not only the potentially beneficial or detrimental impact of several bio-behavioural factors, but also the time point across the resilience trajectory at which each of these factors becomes important for a particular person. The realization of the role and the time of impact of each factor on each patient's resilience trajectory will enhance the effectiveness of the already existing and/or the development of new, more efficient personalized interventions that will place a greater emphasis on the modifiable nature of resilience. These interventions will focus in addition, in proper timing as a crucial aspect of their intervention goals and plans.

5. Data Source Identification

Below we describe the current status of the protocol for collecting prospective data and the available retrospective data from the four BOUNCE clinical sites. The complete protocol will be submitted with the relevant deliverable (D6.1) at month 12.

5.1. Prospective data collection protocol

The broad and general objective of the BOUNCE project is to build a quantitative mathematical model of factors associated with optimal adjustment capacity to cancer. We will collect data concerning the biomedical status (BMS), the psychosocial status (PSS) and the functional status (FUS) of breast cancer patients. Selection of variables to be measured was based from the relevant literature and extensive clinical experience of the partners to cover constructs most likely to predict patients' capacity to bounce back during the highly stressful treatment and recovery period following diagnosis of breast cancer. Refinement of the set of measures is expected to take place based on the results of the ongoing analysis of retrospective data provided by the four clinical sites, derived from existing breast cancer patient cohorts. The overreaching goal of the model is to understand the optimal time period to intervene through personalised psychological support.

Estimating resilience is considered as a critical prerequisite in making decisions regarding the need of psychosocial interventions for a given patient. In this context grading of a person's resilience level, which indirectly expresses the risk of adverse future psychological outcomes, is clinically important. If this value is above a certain threshold no action might be required, in another range a light action might be required (e.g. more physical exercise, more strict diet etc.), if it is below a certain threshold then "emergency" measures might need to take place (e.g. psychological support, psychiatric examinations, further biomedical examinations and tests etc.).

5.1.1. Primary endpoint

The main goal of the prospective pilot study is to identify factors, processes and health care infrastructures that may predict at different points in time and in the long-term the patient's resilience to the physically and psychologically stressful their physical wellbeing and the psychological outcomes of the cancer event.

5.1.2. Secondary endpoints

- To cross-validate the prediction models in order to assess the accuracy of their performance in clinical practice and to test their generalizability. This will take place through bootstrapping methods and, mainly, through data splitting. Specifically, through sophisticated sampling methods, the dataset will be split into two: one part of the dataset will be used so as to develop and train the model; the other part will be used to validate it. Proposed prediction models are detailed below.
- To identify processes and interactions that can more accurately predict final (i.e., at 18 months) and intermediate (i.e., at 3, 6, 9, 12 and 15 months) psychological outcomes.

- To develop a multi-dimensional index of resilience as a function of the biomedical status (BMS), the psychosocial status (PSS) and the functional status (FUS) of the patient.
- To deliver an advanced, empirically validated and more inclusive definition of resilience.
- To perform a series of moderation, multiple mediation and moderated mediation analyses (e.g., from personality traits to health outcomes, through health-related beliefs and behaviour, with socio-demographic variables as moderating conditions) in order to gain an enhanced understanding of the dynamic process of adaptation to breast cancer, and resilience-as-a-process.
- To conduct cost-benefit analysis in order to assess the strengths and limitations of the project outcomes and also determine the best approach to achieve the maximum benefits.
- To examine potential differences in the predictive and outcome variables across the four clinical sites of the BOUNCE Pilot Study considering also health care infrastructures and patient flow/support/culture.

5.1.3. Methods and study design

Primary and secondary endpoints will be reached through a large scale multi-center study at four clinical centres:

- European Institute of Oncology (IEO),
- Rabin Medical Center/Shaare Zedek Medical Center/Kaplan Medical Center [coordinated by HUJI]
- Helsinki University Hospital – HUS (Finland, Coordinator of The Project),
- Champalimaud Foundation (CHAMP, Portugal).

The processes of defining the instruments for the BOUNCE prospective pilot study started with a list of about 50 relevant concepts and their measures. This initial pool was determined basing on two sources: (1) The results of literature search; and (2) Preliminary proposals based on the research experience of the five national BOUNCE clinical teams, comprised of psycho-oncologists, health psychologists, social workers, and psychometricians.

The following criteria were proposed for choosing the measures:

- Sound psychometric properties (reliability and construct validity).
- Divergent validity in context of the present research (low overlap with other measures).
- Proven usefulness in research on breast cancer patients.
- Ability to predict important outcomes in RCT's or in longitudinal studies (controlling for initial levels of the outcome measures).
- Preferably short.

5.1.4. Measures

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

- Ten item Personality measure (brief Big Five)

- PTSD Checklist
- Life Orientation Test Revised
- Sense of Coherence scale
- The Perceived Ability to Cope with Trauma
- Cognitive Emotion Regulation Questionnaire
- Mindful Attention Awareness Scale
- Modified Medical Outcomes Study Social Support Survey
- Connor Davidson Resilience Scale
- Illness Perception Questionnaire
- Mini-Mental Adjustment to Cancer Scale
- Cancer Behaviour Inventory
- Family Resilience Questionnaire
- MOS Adherence to Medical Advice Scale
- Post-Traumatic Growth Inventory
- EORTC Quality of Life questionnaire – General and Breast Cancer module
- Fear of Recurrence short form
- Hospital Anxiety and Depression Scale
- Positive and Negative Affect short form
- NCCN Distress Thermometer

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

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

- Age
- Highest level of education
- Marital status
- Number of children
- Employment status
- Income
- Absence from work
- Smoking and alcohol/drugs consumption
- Weight and height
- Diet
- Exercise

- Number of counselling/support sessions
- Number of visits with physician/nurse/social worker

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

- TNM stage
- ICD-10 Classification
- Tumor biology
- Surgery type and side
- Previous/ongoing oncological therapy
- Side effects due to therapy
- Menopausal status
- Family history of cancer
- Psychotropic medication
- Disease-free survival
- Basic laboratory tests (blood cell counts, CRP)
- Patient pathway data
- Biomarkers

5.1.5. Measurement time points

There will be seven assessment waves, over an 18-month period: baseline, which will occur after the first visit with the oncologist, Month 3 (M3), Month 6 (M6), Month 9 (M9), Month 12 (M12), Month 15 (M15), and Month 18 (M18). During the baseline measurement wave, which will occur within three to four weeks from diagnosis, only non-cancer-specific measures will be delivered (such as personality). Cancer-specific measures will be assessed from M3, when the patient has already had some meaningful experience with the illness.

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

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

5.1.6. Patient selection: criteria for patient eligibility/ineligibility

5.1.6.1. Participant population

Participants will be breast cancer patients with stage I-III histologically confirmed diagnosis.

5.1.6.2. Total number of patients

The targeted number of patients is 660. Each clinical site (IEO, Rabin Medical Center/Shaare Zedek Medical Center/Kaplan Medical Center [coordinated by HUJI], and HUS) is expected to recruit 200 breast cancer patients. An additional 60 patients will be recruited by CHAMP.

To be eligible for inclusion in the study, each patient must fulfil the following inclusion criteria:

- Provide signed informed consent
- Female patients, 40-70 years of age at the time of recruitment
- Histologically confirmed invasive early or locally advanced operable breast cancer
- Tumour stage I, II or III
- Patient received surgery as part of the local treatment
- In addition to local treatment, patient received radiotherapy and or/systemic treatment

Patients who meet any of the following criteria will be excluded:

- Refusal to sign informed consent
- Presence of distant metastases
- History of another malignancy or contralateral invasive breast cancer within the last five years except cured basal cell carcinoma of skin or carcinoma in situ of the uterine cervix
- History of early onset (i.e., before 40 years of age) mental disorder (i.e., schizophrenia, psychosis, bipolar disorder, major depression) or severe neurologic disorder (i.e., neurodegenerative disorder, dementia)
- Serious other uncontrolled concomitant diseases such as clinically significant (i.e. active) cardiac disease (e.g. congestive heart failure, symptomatic coronary artery disease or cardiac arrhythmia not well controlled with medication) or myocardial infarction within the last 12 months.
- Major surgery for a severe disease or trauma which could affect patients psychosocial wellbeing Treatment for invasive cancer
- Treatment for any major illness in the last half year

5.1.7. Procedures to register a patient

Specific informed consent form will be prepared for the study. As Promoter of the multi-centre pilot, IEO is in charge of the preparation of informed consent sheet, which will be then adopted by the other clinical centre, once translated into their native languages.

5.1.8. Data sharing between the Noona plartform and the BOUNCE data infrastructure

Two main approach alternatives are available for data sharing between the Noona platform and the Bounce infrastructure as shown in Figure 13:

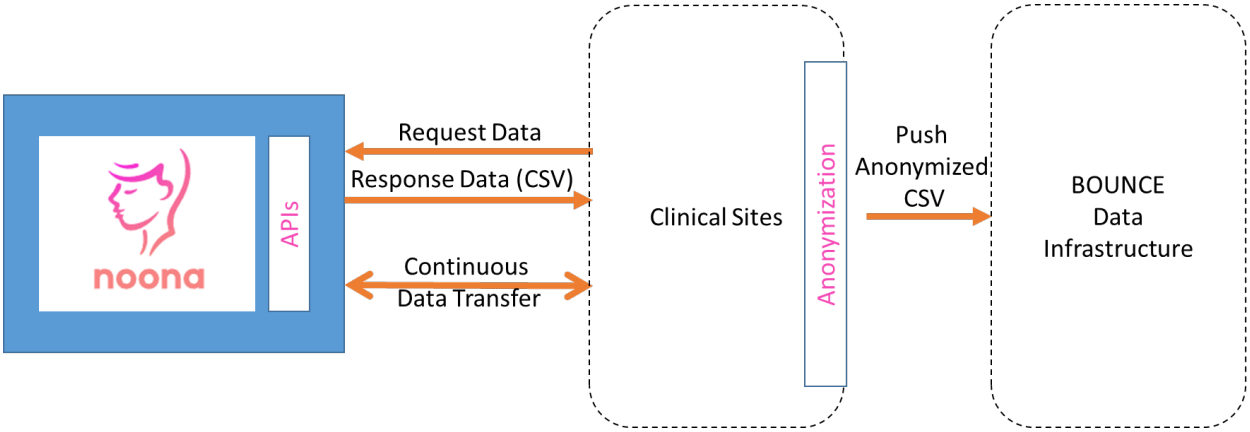

Figure 13. Data Transfer from the Noona platform to BOUNCE data infrastructure.

1. Request-response process
2. Continuous data transfer over API

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

Continuous data transfer over API is similar content-wise, however the method is not request based. Instead an API is built between Noona and a research data base, where data extracts can be streamed continuously, e.g. overnight for each day. The whole workflow will be respecting all security guidelines for transferring sensitive data.

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

5.2. Description of the retrospective data

Bellow we describe the retrospective data sources that will be used during the BOUNCE lifetime. In this deliverable only a high level description is provided. A detailed list of the measures included in the retrospective BOUNCE data can be found in D3.1 – Identification of Internal and External Data Sources and Registries.

5.2.1.HUS

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

- **Clinical data:** age, WHO class, menstruation after chemotherapy, menopausal status, menopause age, BMI, weight, height, bone mineral density, total cholesterol levels, Blood Glucose, Blood Pressure, pulse, any other disease also psychiatric, basic health status, disability status, physical pain
- **Breast and treatment data:** tumor size, pT, pN, histological type, metastatic lymph modes, receptor status (estrogen, progesterone), Her2 expression, type of breast

surgery, type of axillary operation, type of treatment (herceptin, chemotherapy, radiotherapy, endocrine treatment)

- **SocioDemographics:** years of education, marital status, number of children, employment status, reason for not working
- **History and Life Style:** competing athlete, smoking, frequency and amount of alcohol consumption, reduced fat in the diet, increased vegetables, increased the amount of exercise etc.
- **Physical performance and activity:** mean figure 8 running, mean 2-km walking test, leisure time physical activity, self-reported physical activity, MET (metabolic equivalent)
- **Survival data:** local, distant relapse free-survival and overall survival
- **Psychosocial self-report questionnaires:**
    - **EORTC QLQ- C30:** A questionnaire of 30 items developed to assess the quality of life of cancer patients. It incorporates five functional scales (physical, role, cognitive, emotional, and social), three symptom scales (fatigue, pain, and nausea and vomiting), a global health status / QoL scale, and a number of single items assessing additional symptoms commonly reported by cancer patients (dyspnoea, loss of appetite, insomnia, constipation and diarrhoea) and perceived financial impact of the disease.
    - **EORTC QLQ- BR23**: It is a breast-specific module of the EORTC QLQ comprising 23 questions to assess body image, sexual functioning, sexual enjoyment, future perspective, systemic therapy side effects, breast symptoms, arm symptoms and upset by hair loss.
    - **WHQ** women's health questionnaire: It contains 37 items distributed among nine domains: depressed mood, somatic symptoms, memory/concentration, vasomotor symptoms, anxiety/fear, sexual behaviour, sleep problems, menstrual symptoms and attractiveness.
    - **FACIT-F**-Functional Assessment of Chronic Illness Therapy-Fatigue questionnaire: It is a 13-item compilation of general questions assessing fatigue levels during common daily activities over the past week.
    - **BDI** - Beck Depression Inventory short form: Finnish modified version of Beck's 13-item depression scale (**R-BDI**). The short form of Beck Depression Inventory is a screening instrument for assessing depressive symptomatology among the following domains: mood, pessimism, sense of failure, dissatisfaction, guilt, self-hate, suicide, social withdrawal, indecisiveness, body image, work inhibition, fatigue and appetite.

The time points that each type of data were collected are summarized in Table 4.

	T1 Base-line	T2 after 3 mouths	T3 after 6 months	T4 after 12 months	T5 after 18 months	T6 after 24 months	T7 after 30 months	T8 after 36 months
Breast and treatment data	✓							
Clinical data*	✓			✓				✓
Self – report clinical data**			✓	✓	✓	✓	✓	✓
SocioDemographic			✓	✓	✓	✓	✓	✓
History	✓							

Self-Reported Life Style	✓		✓	✓	✓	✓	✓	✓
Physical performance	✓			✓				✓
Physical activity	✓		✓	✓	✓	✓	✓	✓
Survival data								
Psychosocial self-report questionnaires	✓	✓	✓	✓		✓	✓	✓

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

Table 4. Time availability of HUS retrospective data

5.2.2. HUJI

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

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

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

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

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

Table 5. Time availability of HUJI retrospective data

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

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

T1-T6 psychosocial self-report questionnaires

Posttraumatic stress symptoms. The Posttraumatic Stress Diagnostic Scale [35] was used to assess the severity of posttraumatic distress. The PDS is a commonly used measure of PTSD that assesses the frequency of 17 symptoms and symptom severity.

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

Depression. Depressive symptoms were measured using the Center for Epidemiologic Studies Depression Scale [89]. The CES-D is a well-validated 20-item measure based on ratings in four primary symptom areas: (a) depressed affect; (b) lack of positive affect; (c) somatic symptoms; and (d) interpersonal difficulties.

Cognitive and emotion regulation. The Cognitive Emotion Regulation Questionnaire (CERQ) is a multidimensional, 18-item scale that identifies coping strategies used by respondents following stressful or negative life events [22]. Responses are organized into nine subscales, divided into *positive*: acceptance, positive refocusing, refocus on planning, positive reappraisal, putting into perspective and *negative:* self-blame, rumination, catastrophizing, and blaming others.

Coping flexibility. The Perceived Ability to Cope with Trauma (PACT) scale [9]. The PACT is a 20 item-scale that assess ability to cope with potentially traumatic event. The PACT is divided into two subscales: (a) forward focus, comprised of 12 items, and (b) trauma focus, comprised of eight items

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

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

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

Distress. The 6-item Kessler psychological distress scale.

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

5.2.3.IEO

The retrospective data is composed of multiple datasets from psycho-oncology studies, conducted between January 2009 and April 2018. One study (LIVE - “Life Improved by Exercise”) investigated the effect of an 8-weeks exercise program on breast cancer survivors in terms of quality of life and global functioning. A second project (“iManageCancer”) has two studies; one evaluated the efficacy and usability of e-health platforms developed within this Horizon 2020 project, another study constructed a family resilience questionnaire together with the Resilience Scale for Adults and the ALGA-BC questionnaire, a measure developed and validated in IEO, that investigates the psycho-cognitive profile of patients. Another study (“Rumination study”) evaluated social and cognitive aspects that may be related to the degree of rumination, while a different study (“Invernizzi”) analyzed the effect of endocrine therapy on cognitive functioning in older patients (60-70 years old). The “Magnifying Glass” study evaluated presence of PTSD symptoms from diagnosis to two years from diagnosis with three time-points. The “OncodermaCare” study evaluated the effect of cosmetic treatment on quality of life for patients undergoing radiotherapy. Finally, the “DIGNICAP” project investigated effectiveness of scalp cooling system DigniCap to prevent alopecia in primary breast cancer patients receiving adjuvant chemotherapy. The subjective perception of the device and quality of life were assessed at baseline and after each cycle of chemotherapy. The estimated number of breast cancer patients that have participated in these studies is 900.

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

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

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

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

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

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

5.2.4. CHAMP

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

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

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

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

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

Concerning psychological variables the following measurements were assessed:

1. The Distress Thermometer [83] - is a distress screening tool used to better identification of oncologic patients on psychological distress and management in the psycho-oncology department. This is a simple, self-report, pencil and paper measure consisting of a line with a 0-10 scale anchored at the zero point with "No distress" and at scale point ten with "Extreme distress". It includes also a problem checklist. The patient is asked to identify those problems from the checklist, which are contributing to their score.
2. Hospital Anxiety and Depression scale (HADs) [114] It is a fourteen item scale, seven of the items relate to anxiety and seven relate to depression. The anxiety and depressive

subscales are also valid measures of severity of the emotional disorder. It was validated also for the Portuguese population in various clinical samples.

3. Mini Mental Status- Examination (MMSE): is a screening tool used to assess objective cognitive function. It consists of a questionnaire with a maximum score of 30 points, grouped in seven categories: orientation to time (5 points); orientation to place (5 points); registration of three words (3 points); attention and calculation (5 points); recall of three words (3 points); language (8 points) and visual construction (1 point).
4. Addenbrookes Cognitive Examination Revised (ACE-R): Is a cognitive screening tool, originally designed by Mioshi et al. [79] to address the lack of MMSE sensitivity in the diagnosis of dementia. The overall result of ACE-R includes an amount equal to the result of the MMSE, and further allows the assessment of multiple domains. The Portuguese experimental version was developed in community and clinical samples geriatric [47].
5. Wechsler Adult Intelligence Scale subtests (WAIS III)- Wechsler Adult Intelligence Scale – III, Digit Span subtest, that comprises two modalities: Forward - repeat number sequences with increasing length, in the same order as presented aurally to access immediate memory; and backward – repeat digit sequences in reverse order, to achieve working memory. Symbol Search subtest: Working within a specific time limit, the examinee scans a search group and indicates whether one of the symbols in the target group matches. This subtest measures processing speed, short-term visual memory, visual-motor coordination, cognitive flexibility, visual discrimination, psychomotor speed, and speed of mental operation. The examinee completes this subtest using a response booklet, and not on his or her digital device.
6. Trail Making Test A and B: Originally created by the US Army psychologists to assess selective attention (Part A), divided attention, the ability to sequence stimuli, cognitive flexibility and the processing speed (Part B). The TMT-A & TMT-B were validated for the Portuguese population with an adult sample by Cavaco et al [12].
7. Stroop test: Assessment tool for executive functions, response inhibition and selective attention, originally developed by Stroop [101] and revised by Golden & Freshwater [45], in an American adult population. The validation for the Portuguese population includes a sample of participants from 15 to 100 years was published by Fernandes [34].
8. Beck Depression Inventory (BDI-II): The BDI is a 21-item, self-report rating inventory that measures characteristic attitudes and symptoms of depression [5]. It is validated for the Portuguese population by Campos & Gonçalves [17] in a community sample, with no cut-off score.
9. State-trait Anxiety Inventory: The STAI is a commonly used measure of trait and state anxiety [102]. It can be used in clinical settings to diagnose anxiety and to distinguish it from depressive syndromes. Form Y, its most popular version, has 20 items for assessing trait anxiety and 20 for state anxiety. It was adapted for the Portuguse population by Santos & Silva [93].
10. EORTC QLC 30 has been widely used in clinical practice and clinical trials for measuring quality of life (QoL) in patients with cancer. Includes 30 items for 15 dimensions/scales: five functional scales (physical, role, cognitive, social, and emotional functioning), three symptom scales (fatigue, nausea/vomiting, and pain), five single-item symptom scales

(dyspnea, sleep disturbances, appetite loss, constipation, and diarrhea), single-item scale for financial impact, and a global health status.

A more complete description of the data to be provided by CHAMP is presented in D3.1.

6. Conclusions

This deliverable presented the state of the art on key technological areas related to BOUNCE objectives as well as background work from previous projects that can be reused. To this end, key technologies have been selected for storing personal health data for data management, semantic modelling, data cleaning, anonymization, security, model collection, curation, validation and integration, and temporal data mining. Second, existing clinical practice has been reviewed regarding evaluating resilience, identifying that limited procedures exist in all participating clinical centres. Third, the BOUNCE methodology and the steps required for multi-scale, cross-sectional data aggregation, harmonization, assessment and conceptual and explicit modelling of resilience, risk prediction and decision support, describing also the multicentre clinical pilot study, has been presented. Finally, the first version of the protocol for collecting prospective data and the retrospective data already available from the clinical sites has been also presented.

We believe that following this well-defined methodology, BOUNCE will be able to accomplish its objectives to collect clinical, biological, psychological, and social parameters which will enable the description and preliminary modelling of the resilience trajectory. However, we do not expect this methodology to be static, but rather subject to continuous elaboration, refinement and change, as we will be able to gradually adapt and optimise the project's methodology based on the collected retrospective and prospective data.

7. References

1. Argyri KD, Dionysiou DD, Stamatakos GS, Modelling the interplay between pathological angiogenesis and solid tumour growth: The anti-angiogenic treatment effect, Advanced Research Workshop on In Silico Oncology and Cancer Investigation - The TUMOUR Project Workshop (IARWISOCI), 2012 5th International , vol., no., pp.1,4, 22-23 Oct. 2012.
2. Antunes C, Oliveira A, Temporal data mining: An overview. In Proceedings of the KDD Workshop on Temporal Data Mining. 1—13, 2001.
3. Baltrušaitis T, Chaitanya A, Louis-Philippe M, Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018.
4. Baltrušaitis T, Chaitanya A, and Louis-Philippe M, Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence 2018.
5. Beck AT, Ward C, Mendelson M, Beck Depression Inventory (BDI). Arch Gen Psychiatry. 1961:4(6): 561–571.
6. Benjelloun O, Garcia-Molina H, Menestrina D, Su Q, Whang SE, Widom J, Swoosh: A Generic Approach to Entity Resolution. Stanford: Stanford InfoLab, 2005.
7. Boghaert E, Radisky DC, Nelson CM. Lattice-based model of ductal carcinoma in situ suggests rules for breast cancer progression to an invasive state. PLoS Comput Biol. 2014 Dec 4;10(12):e1003997. doi: 10.1371/journal.pcbi.1003997. eCollection 2014.
8. Bohannon P, Fan W, Flaster M, Rastogi R, A Cost-based Model and Effective Heuristic for Repairing Constraints by Value Modification. Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data. New York.
9. Bonanno GA, Galea S, Bucciarelli A, Vlahov D, What predicts psychological resilience after disaster? The role of demographics, resources, and life stress. Journal of Consulting and Clinical Psychology, 2007: 75(5), 671–682.
10. Bonissone PP, Feng X, Raj S, Fast meta-models for local fusion of multiple predictive models. Applied Soft Computing 11.2: 1529-1539, 2011.
11. Bostrom H. Feature vs. classifier fusion for predictive data mining a case study in pesticide classification. Information Fusion, 2007 10th International Conference on. IEEE, 2007.
12. Cavaco S, Gonçalves A, Pinto C, et al. Trail Making Test: Regression-based Norms for the Portuguese Population, Archives of Clinical Neuropsychology, 2013:28(2):189–198, https://doi.org/10.1093/arclin/acs115.
13. Cheng KS, Lang JH, A Novel Data Cleaning with Data Matching, Advanced Science and Technology Letters, pp. 161-16, 2016.
14. Cong G, Fan W, Geerts F, Jia X, Ma S, Improving Data Quality: Consistency and Accuracy. Proceedings of the 33rd International Conference on Very Large Data Bases. Vienna, 2007.
15. Purificacion C, Fuzzy temporal association rules: combining temporal and quantitative data to increase rule expressiveness, WIREs Data Mining Knowl Discov 2014, 4:64–70. doi: 10.1002/widm.1116.
16. Cavoukian A. Former Information and Privacy Commissioner of Ontario, Canada; for PbD see http://www.privacybydesign.ca/

17. Campos RC, Gonçalves B, The Portuguese Version of the Beck Depression Inventory-II (BDI-II), Preliminary Psychometric Data with Two Nonclinical Samples, European Journal of Psychological Assessment 2011:27:258-264.
18. Collier N., et al. An ontology-driven system for detecting global health events, Int. Conf. on Computational Linguistics (COLING), 215-222, 2010.
19. Dagade AA, Mali MP, Pathak, NP, Survey of Data Duplication Detection and Elimination in Domain Dependent and Domain-Independent Databases. International Journal of Advance Research in Computer Science and Management Studies, 4(5), 2016.
20. Dallachiesa M, Ebaid A, Eldawy A, Elmagarmid A, Ilyas IF, Ouzzani M, Tang N, NADEEF: A Commodity Data Cleaning System. Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data. New York.
21. Danesi R, Innocenti F, Fogli S et al. Pharmacokinetics and pharmacodynamics of combination chemotherapy with paclitaxel and epirubicin in breast cancer patients, Journal of Clinical Pharmacology, vol. 53, pp 508-518, 2002.
22. Garnefski N, Kraaij V, Cognitive Emotion Regulation Questionnaire: Development of a short 18-item version (CERQ-short).Personality and Individual Differences 2006:41:1045-1053.
23. Johnson D, et al. Connecting digital cancer model repositories with markup: introducing TumorML version 1.0. ACM SIGBioinformatics Record 3.3, 5-11, 2013.
24. Detmer D, Bloomrosen M, Raymond B, Tang T. Integrated personal health records: transformative tools for consumer-centric care. BMC Medical Informatics and Decision Making 2008;8:45.
25. Deisboeck TS, Wang Z, Macklin P, Cristini V. Multiscale cancer modelling. AnnuRev Biomed Eng 2011;13:127–55.
26. Dionysiou DD, et al. 2004. J. Theor. Biol., 230, 1–20.
27. Dionysiou D, The ISOG, NTUA tumour response to treatment discrete simulation models: a review of basic concepts and algorithms" in G. Stamatakos and D. Dionysiou (Eds): Proc. 4th Int. Adv. Res. Workshop on In Silico Oncology and Cancer Investigation (4th IARWISOCI) - The ContraCancrum Workshop, Athens, Greece, Sept. 8-9, 2010 (www.4th-iarwisoci.iccs.ntua.gr), pp.49-53. 2010
28. Dionysiou DD, Stamatakos GS, Gintides D, Uzunoglu N, Kyriaki K, Critical Parameters Determining Standard Radiotherapy Treatment Outcome for Glioblastoma Multiforme: A Computer Simulation The Open Biomedical Engineering Journal 2, 43-51, 2008.
29. Donovan KA, Small BJ, Andrykowski MA, Munster P, Jacobsen PB, Utility of a Cognitive–Behavioral Model to Predict Fatigue Following Breast Cancer Treatment, Health Psychol. 2007 Jul; 26(4): 464–472.
30. Edelman LB, Eddy JA, Price ND. In silico models of cancer. Wiley Interdiscip RevSyst Biol Med 2010;2:438–59.
31. Emotion Measures: http://www.healthmeasures.net/explore-measurement-systems/promis
32. Fan W, Li J, Ma S, Tang N, Yu, W, Towards Certain Fixes with Editing Rules and Master Data. The VLDB Journal, 21(2), 213-238, 2012.
33. Fatima A, Nazir N, Gufran KM, Data Cleaning In Data Warehouse: A Survey of Data Pre-processing Techniques and Tools. International Journal of Information Technology and Computer Science, 9, 50-61, 2017.
34. Fernandes S, Stroop: Teste de cores e palavras: Manual. Lisbon, Portugal: Cegoc, 2013.

35. Foa EB., Cashman L, Jaycox L, Perry K, The validation of a self-report measure of posttraumatic stress disorder: The Posttraumatic Diagnostic Scale. Psychological Assessment 1997:9(4), 445-451.
36. Fragoso TM, Wesley B, Francisco L, Bayesian model averaging: A systematic review and conceptual classification. International Statistical Review 86.1 (2018): 1-28.
37. Frieboes HB, Edgerton ME, Fruehauf JP, Rose FR, Worrall LK, Gatenby RA, Ferrari M, Cristini V. Prediction of drug response in breast cancer using integrative experimental/computational modelling. Cancer Res. 2009 May 15;69(10):4484-92. doi: 10.1158/0008-5472.CAN-08-3740. Epub 2009 Apr 14.
38. Fumera G, Fabio R. Performance analysis and comparison of linear combiners for classifier fusion. Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR). Springer, Berlin, Heidelberg, 2002.
39. Gaudreault J, Greig G, Cosson V, Gupta M, Jumbe N, Hooker AC. Population pharmacokinetics of bevacizumab. ASCO Meeting Abstracts. 2008;26:14570
40. Genitsaridi I, Kondylakis H, Koumakis L, Marias K, Tsiknakis M. Evaluation of Personal Health Record Systems through the Lenses of EC Research Projects, Computers in Biology and Medicine 2015;59:175-185. [cited 2017 June 11]. Available from: http://www.sciencedirect.com/science/article/pii/S0010482513003223
41. Genitsaridi I, Kondylakis H, Koumakis L, Marias K, Tsiknakis M. Towards Intelligent Personal Health Record Systems: Review, Criteria and Extensions, Procedia Computer Science 2013;21:327-334. [cited 2017 June 11]. Available from: http://www.sciencedirect.com/science/article/pii/S1877050913008363
42. Georgiadi EC, Dionysiou DD, Graf N, Stamatakos G, Towards In Silico Oncology: Adapting a Four Dimensional Nephroblastoma Treatment Model to a Clinical Trial Case Based on Multi-Method Sensitivity Analysis. Computers in Biology and Medicine, in press, doi: 10.1016/j.compbiomed.2012.08.008. Epub 2012 Oct 10. vol.42 (11) pp. 1064-1078 2012.
43. Giatili SG, Stamatakos GS, A detailed numerical treatment of the boundary conditions imposed by the skull on a diffusion-reaction model of glioma tumour growth. Clinical validation aspects,, Applied Mathematics and Computation, 218 (2012), 8779.
44. Gohel AC, Patil AV, Vadhwana PP, Patel HS, A Commodity Data Cleaning System. International Research Journal of Engineering and Technology , 4(5), 2017.
45. Golden CJ, Freshwater SM, Stroop Color and Word Test: Revised examiner's manual. Wood Dale, IL: Stoelting Co, 2002.
46. Gönen M, Ethem A, Multiple kernel learning algorithms. Journal of machine learning research 12.Jul (2011): 2211-2268.
47. Gonçalves C, Pinho MS, Cruz V, et al., The Portuguese version of Addenbrooke's Cognitive Examination–Revised (ACE-R) in the diagnosis of subcortical vascular dementia and Alzheimer's disease, Aging, Neuropsychology, and Cognition, 2015:22:4:473-485.
48. Gunter TD, Terry NP. The Emergence of National Electronic Health Record Architectures in the United States and Australia: Models, Costs, and Questions. J Med Internet Res 2005;7(1):e3.
49. Hahnfeldt P, Panigrahy D, Folkman J, Hlatky L. Tumour development under angiogenic signaling: a dynamical theory of tumour growth, treatment response, and postvascular dormancy. Cancer Res. 1999;59:4770–5.

50. He Y, Xiang Z, Sarntivijai S, Toldo L, Ceusters W, AEO: A Realism-Based Biomedical Ontology for the Representation of Adverse Events, Int. Conf. on Biomedical Ontology, Representing Adverse Events Workshop, July 26, 2011.
51. HealthMeasures.Net:http://www.healthmeasures.net/explore-measurement-systems/nih-toolbox/intro-to-nih-toolbox/emotion
52. Iakovidis I. Towards personal health record: current situation, obstacles and trends in implementation of electronic healthcare record in Europe, International Journal of Medical Informatics, Volume 52, Issues 1–3, 1 October 1998, Pages 105–115.
53. Johnson D, McKeever S, Stamatakos G, Dionysiou D, Graf N, Sakkalis V, et al.Dealing with diversity in computational cancer modelling. Cancer Inform2013;12:115–24.
54. Jony RI, Mohammed N, Habib A, Momen S, Rony RI, An Evaluation of Data Processing Solutions Considering Preprocessing and "Special" Feature. 11th International Conference on Signal-Image Technology & Internet-Based Systems, 2015.
55. Jung SY, Lee K, Hwang H, Yoo S, Baek H Y, Kim J, Support for Sustainable Use of Personal Health Records: Understanding the Needs of Users as a First Step Towards Patient-Driven Mobile Health. JMIR Mhealth Uhealth 2017;5(2):e19.
56. Kent S, Atkinson R, IP Encapsulating Security Payload (ESP). IETF. RFC 2406, 1998.
57. Knudson AG, Cancer genetics, Am J Med Genet. 2002 Jul 22;111(1):96-102.
58. Kolokotroni EA, Stamatakos GS, Dionysiou DD, et al., Translating Multiscale Cancer Models into Clinical Trials: Simulating Breast Cancer Tumour Dynamics within the Framework of the "Trial of Principle" Clinical Trial and the ACGT Project.In: Proc. 8th IEEE International Conference on Bioinformatics and Bioengineering (BIBE), 2008.
59. Kolokotroni EA, Dionysiou DD, Georgiadi E et al., Breast cancer modelling in the clinical context: parametric studies, In G. Stamatakos and D. Dionysiou (Eds): Proc. 4th Int. Adv. Res. Workshop on In Silico Oncology and Cancer Investigation (4th IARWISOCI) - The ContraCancrum Workshop, Athens, Greece, Sept. 8-9, 2010 (www.4th-iarwisoci.iccs.ntua.gr), pp.58-61, 2010.
60. Kolokotroni EA, Dionysiou DD, Uzunoglu NK, Stamatakos GS, Studying the growth kinetics of untreated clinical tumours by using an advanced discrete simulation model, Mathematical and Computer Modelling, 54 1989-2006 2011.
61. Kolokotroni EA, Dionysiou DD, Veith C, et al., In Silico Oncology: Quantification of the In Vivo Antitumor Efficacy of Cisplatin-Based Doublet Therapy in Non-Small Cell Lung Cancer (NSCLC) through a Multiscale Mechanistic Model. PLoS Computational Biology, 12(9), 2016.
62. Kondylakis H, Bucur A, Dong F, Renzi C, Manfrinati A, Graf N, Hoffman S, Koumakis L, Pravettoni G, Marias K, Tsiknakis M, Kiefer S. iManageCancer: Developing a platform for Empowering patients and strengthening self-management in cancer diseases, 30th IEEE International Symposium on Computer-Based Medical Systems - IEEE CBMS, 2017.
63. Kondylakis H, et al. Development of Interactive Empowerment services in support of personalized medicine, Ecancermedicalscience, 8: 400, 2014.
64. Kondylakis H, Plexousakis D, Ontology evolution without tears, Journal of Web Semantics 2013:19:42-58.
65. Kondylakis H, Plexousakis D, Hrgovcic V, Woitsch R, Premm M, Schuele M. Personal eHealth knowledge spaces though models, agents and semantics. International Conference on Conceptual Modeling (ER) 2014;293-297.

66. Kondylakis H, Flouris G, Fundulaki I, Papakonstantinou V, Tsiknakis M. Flexible access to patient data through e-Consent. International Conference on Wireless Mobile Communication and Healthcare (MobiHealth) 2015.
67. Kondylakis H, Spanakis EG, Sfakianakis S et al. Digital patient: Personalized and translational data management through the MyHealthAvatar EU project. EMBC 2015: 1397-1400
68. Krishnan S, Haas D, Franklin MJ, Wu E, Towards Reliable Interactive Data Cleaning: A User Survey and Recommendations. Proceedings of the Workshop on Human-In-the-Loop Data Analytics HILDA '16. 2016.
69. Krukowski A, Barca C C, Rodríguez J M, Vogiatzaki E. Personal Health Record, Cyberphysical Systems for Epilepsy and Related Brain Disorders, pp 205-238
70. Lin W, Orgun MA, Williams GJ, An overview of temporal data mining, Proc. 1st Australian Data Mining Workshop (ADM02), eds. S. J. Simoff, G. J. Williams and M. Hegland (2002) pp. 83–90.
71. Lowengrub JS, Frieboes HB, Jin F, Chuang YL, Li X, Macklin P, et al. Nonlinearmodelling of cancer: bridging the gap between cells and tumours. Nonlinearity 2010;23:R1–91.
72. Lu J-F, Bruno R, Eppler S, Novotny W, Lum B, Gaudreault J. Clinical pharmacokinetics of bevacizumab in patients with solid tumors. Cancer Chemother Pharmacol. 2008;62:779–86.
73. Macklin P, Edgerton ME, Thompson AM, Cristini V, Patient-calibrated agent-based modelling of ductal carcinoma in situ (DCIS): From microscopic measurements to macroscopic predictions of clinical progression. J Theor Biol. 2012 May 21; 301: 122–140.
74. Mandl KD, Simons WW, Crawford WCR, Abbett JM. Indivo: a personally controlled health record for health information exchange and communication. BMC Med. Inf. & Decision Making 2007;7:25.
75. Mayfield C, Neville J, Prabhakar S, ERACER: A Database Approach for Statistical Inference and Data Cleaning. Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data. Indianapolis.
76. McDonough M.H., et al. 2014. Psycho-Oncology 23.1: 114-120.
77. MdAnderson: http://www3.mdanderson.org/app/medcalc/index.cfm?pagename=bcnt
78. Meretoja TJ, et al. 2014. Annals of surgical oncology, 21(7), 2229-2236.
79. Mioshi E, Dawson K , Mitchell J, Arnold R, Hodges JR, The Addenbrooke's Cognitive Examination Revised (ACE-R): a brief cognitive test battery for dementia screening. Int. J. Geriat. Psychiatry, 2006:21: 1078-1085. doi:10.1002/gps.1610.
80. Mitchell T, Machine Learning, McGraw Hill.
81. Mitsa T. Temporal Data Mining, 1st Edition, Chapman & Hall, CRC 2010.
82. Novère NL, Finney A, Hucka M, et al., Minimum information requested in the annotation of biochemical models (MIRIAM). Nature Biotechnology. 23 (12): 1509–15. doi:10.1038/nbt1156. PMID 16333295, 2005.
83. O'Donnell E, D'Alton P, O'Malley C, Gill F, Canny Á, The distress thermometer: a rapid and effective tool for the oncology social worker. International journal of health care quality assurance, 2013:26(4):353-359.
84. Peace, J, Brennan, PF, Ontological representation of family and family history, at AMIA Annu Symp Proc. 2007.
85. Poleszczuk J, Bodnar M, Foryś U. New approach to modelling of antiangiogenic treatment on the basis of Hahnfeldt et al. model. Math Biosci Eng. 2011;8:591–603.

86. Powathil GG, Swat M, Chaplain MA, Systems oncology: towards patient-specific treatment regimes informed by multiscale mathematical modelling, Semin Cancer Biol. 2015 Feb;30:13-20. doi: 10.1016/j.semcancer.2014.02.003. Epub 2014 Mar 4.
87. Prabhakaran MM, Sahai A, Secure Multi-Party Computation, IOS Press, 2013, ISBN 978-1-61499-169-4.
88. Predictive Tools for Breast Cancer: https://itunes.apple.com/us/app/predictive-tools-for-breast-cancer/id578648407?mt=8
89. Radloff LS, The CES-D Scale: A Self-Report Depression Scale for Research in the General Population, Applied Psychological Measurement 1977:1(3):385 – 401.
90. Rahm E, Do H-H, Data Cleaning: Problems and Current Approaches. IEEE Bulletin of the Technical Committee on Data Engineering, 23(4), 2000-2012.
91. Rejniak KA, Anderson AR. Hybrid models of tumour growth. Wiley InterdiscipRev Syst Biol Med 2011;3:115–25.
92. Roddick JF, Spiliopoulou M, A survey of temporal knowledge discovery paradigms and methods, IEEE Trans. Knowledge Data Eng., 14 (4) (2002), pp. 750-767.
93. Santos SC, Silva DR, Adaptação do State-Trait Anxiety Inventory (STAI) – Form Y para a população portuguesa: Primeiros dados. Revista Portuguesa de Psicologia, 1997:32:85-98.
94. Schlegel RJ et al. 2012. Psychology & health 27.3:277-293.
95. Schnell S, Grima R, Maini PK. Multiscale modelling in biology – new insights intocancer illustrate how mathematical tools are enhancing the understanding oflife from the smallest scale to the grandest. Am Sci 2007;95:134–42.
96. Shahnawaz M, Ranjan A, Danish M, Temporal Data Mining: An Overview‖(IJEAT) ISSN: 2249 – 8958, vol - 1, Issue -1, Oct 2011.
97. Saqib M, Arshad M, Ali M, Rehman NU, Ullah Z, Improve Data Warehouse Performance by Preprocessing and Avoidance of Complex Resource Intensive Calculations. International Journal of Computer Science Issues, 9(2), 2012.
98. Somasundaram, RS, Nedunchezhian R, Evaluation of three Simple Imputation Methods for Enhancing Preprocessing of Data with Missing Values. International Journal of Computer Applications, 21(10), 2011.
99. Stamatakos G, Dionysiou D, Lunzer A, et al. The Technologically Integrated Oncosimulator: Combining Multiscale Cancer Modelling with Information Technology in the In Silico Oncology Context, DOI:10.1109/JBHI.2013.2284276 IEEE J Biomedical and Health Informatics vol.18, No. 3, pp.840-854 2014.
100. Stamatakos GS, Kolokotroni EA, Dionysiou DD, Georgiadi EC, Desmedt C, An advanced discrete state - discrete event multiscale simulation model of the response of a solid tumour to chemotherapy: Mimicking a clinical study. J Theor Biol. 2010, 266(1):124-139. DOI:10.1016/j.jtbi.2010.05.019.
101. Stroop JR, Studies of interference in serial verbal reactions. Journal of Experimental Psychology, 1935:18(6), 643.
102. Spielberger CD, Jacobs G, Russell S, Crane R, Assessment of Anger: the State-Trait Anger Scale. Manual for the State-Trait Anxiety Inventory. Consulting Psychologists Press, 1983, Inc.; Palo Alto, CA.
103. Tang PC, Ash JS, Bates DW, Overhage JM, Sands DZ, Personal Health Records: Definitions, Benefits, and Strategies for Overcoming Barriers to Adoption. Journal of the American Medical Informatics Association (JAMIA) 2006;13(2):121-126.

104. Ternant D, Cézé N, Lecomte T, Degenne D, Duveau A-C, Watier H, Dorval E, Paintaud G, An enzyme-linked immunosorbent assay to study bevacizumab pharmacokinetics. Ther Drug Monit. 2010;32:647–52.
105. Tian Y, Michiardi P, Vukolic, M, Bleach: A Distributed Stream Data Cleaning System. Computer Research Repository (CoRR), 2016.
106. Tracqui P, Biophysical models of tumour growth. Rep Prog Phys 2009;72:056701.
107. Tulyakov S, et al., Review of classifier combination methods. Machine learning in document analysis and recognition. Springer, Berlin, Heidelberg, 2008. 361-386.
108. Ubezio P, Cameron D, Cell killing and resistance in pre-operative breast cancer chemotherapy. BMC Cancer. 2008 Jul 21;8:201. doi: 10.1186/1471-2407-8-201.
109. Wang Z, Butner JD, Kerketta R, Cristini V, Deisboeck TS, Simulating cancer growth with multiscale agent-based modelling, Semin Cancer Biol. 2015 Feb;30:70-8. doi: 10.1016/j.semcancer.2014.04.001. Epub 2014 May 2.
110. Werfel J, Krause S, Bischof AG, Mannix RJ, Tobin H, Bar-Yam Y, Bellin RM, Ingber DE, How changes in extracellular matrix mechanics and gene expression variability might combine to drive cancer progression, PLoS One. 2013 Oct 3;8(10):e76122. doi: 10.1371/journal.pone.0076122. eCollection 2013.
111. Wolkenhauer O, Auffray C, Brass O, Clairambault J, Deutsch A, Drasdo D, Gervasio F, Preziosi L, Maini P, Marciniak-Czochra A, Kossow C, Kuepfer L, Rateitschak K, Ramis-Conde I, Ribba B, Schuppert A, Smallwood R, Stamatakos G, Winter F, Byrne H, Enabling multiscale modelling in systems medicine, Genome Med. 2014 Mar 21;6(3):21. doi: 10.1186/gm538. eCollection 2014.
112. Xing Z, Pei J, Keogh E. A brief survey on sequence classification. ACM SIGKDD Explorations,Volume 12, Issue 1, pages 40-48, June 2010, ACM Press.
113. Yakout M, Elmagarmid AK, Neville J, Ouzzani M, Ilyas IF, Guided Data Repair. Proceedings of the VLDB Endowment.
114. Zigmond AS, Snaith RP, The hospital anxiety and depression scale. Acta Psychiatrica Scandinavica. 1983:67(6): 361–370.