

D5.8 High-risk markers for patient stratification

30/09/2022

Project title	Patients-centered SurvivorShlp care plan after Cancer treatments based on Big Data and Artificial Intelligence technologies
Grant Agreement number	875406
Call and topic identifier	SC1-DTH-01-2019 - Big data and Artificial Intelligence for monitoring health status and quality of life after the cancer treatment
Funding schema	RIA
Coordinator	FUNDACION CENTRO TECNOLOGICO DE TELECOMUNICACIONES DE GALICIA (GRADIANT)
Website	www.projectpersist.com
Document keywords	EHR, Algorithms, Anonymisation, Data preparation, Ontology
Document Abstract	This deliverable highlights the end-to-end steps required to provide and prepare clinical data for further analysis and processing. Exporting data from clinical repositories and pre-processing data sets to serve as input for AI/ML approaches.



DOCUMENT

Authors	Gaetano MANZO (HES-SO), Yvan PANNATIER (HES-SO), Jean-Paul CALBIMONTE (HES-SO)
Internal reviewers	Thomas LUTZ (SYMP), Stefanie GRUARIN (SYMP), Laura RODRÍGUEZ (SYMP), Umut ARIÖZ(UM), Izidor MLAKAR (UM)
Work package	WP5 – Decision support system at the point of care
Task	T5.3 Cohort and Trajectory Analysis
Nature	Report
Dissemination Level	PU – public

VERSION	DATE	CONTRIBUTOR	DESCRIPTION
0.1	23.11.2021	Jean-Paul Calbimonte (HES-SO), Gaetano MANZO (HES-SO)	Initial draft
0.2	07.12.2021	Gaetano MANZO (HES-SO), Yvan Pannatier (HES-SO)	A detailed description of the trajectory analysis
0.3	14.12.2021	Izidor MLAKAR (UM), Alicia JIMENEZ, Victoria M CAL (GRAD)	Document revision and comments
0.4	22.12.2021	Gaetano MANZO (HES-SO), J-P Calbimonte (HES-SO)	Revised version
0.5	16.09.2022	Gaetano MANZO (HES-SO), Yvan Pannatier (HES-SO)	A detailed description of API and prospective / retrospective data results
0.6	24.09.2022	Umut ARIÖZ (UM), Laura RODRIGUEZ (SYMP)	Document revision and comments
1.0	30.09.2022	Gaetano MANZO (HES-SO), Yvan PANNATIER (HES-SO)	Final version

DISCLAIMER

This document does not represent the opinion of the European Community, and the European Community is not responsible for any use that might be made of its content.

This document may contain material, which is the copyright of certain PERSIST consortium parties, and may not be reproduced or copied without permission. All PERSIST consortium parties have agreed to full publication of this document. The commercial use of any information contained in this document may require a license from the proprietor of that information.

Neither the PERSIST consortium as a whole, nor a certain party of the PERSIST consortium warrant that the information contained in this document is capable of use, nor that use of the information is free from risk and does not accept any liability for loss or damage suffered by any person using this information.

ACKNOWLEDGEMENT

This document is a deliverable of PERSIST project. This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement N° 875406



INDEX

Index of Figures	5
Index of Tables	6
Acronyms and abbreviations	7
Executive Summary	9
Introduction	10
1. PERSIST Project	10
2. Work Package 5: Decision support system at the point of care.....	10
3. Deliverable D5.8 High-risk markers for patient stratification	11
Data ingestion	12
1. Data Flow Architecture.....	12
2. Common data model.....	13
3. Anonymisation and risk for re-identification	18
3.1. Privacy Metrics:.....	18
3.2. Anonymization techniques to address a high re-identification risk	19
Feature extraction	20
Data preparation	23
1. Retrospective data	23
2. Prospective data	24
3. Enriched data.....	27
Patient Trajectory Models	29
1. Prospective Enriched Data Trajectories and Cohorts	34
2. Relapse risk prediction.....	39
2.1. Relapse risk prediction algorithm	39
Trajectory Analysis API	51
Conclusions	53
References	54

Index of Figures

Figure 1 Schematic figure of the data export from clinical partners	12
Figure 2. FHIR resource mappings – FHIR Release 3 (STU)	13
Figure 3. FHIR resource mappings - FHIR Release 3 (STU)	13
Figure 4 Trajectory and cohort data flow	20
Figure 5 Repartition of patients by gender in the whole PERSIST prospective dataset...25	
Figure 6 Repartition of patients by cancer type in the PERSIST dataset	25
Figure 7 Repartition of colon cancer patients by gender in the PERSIST dataset	26
Figure 8 Types of metastasis in colon cancer population.....	27
Figure 9 Colon cancer sample 10% and their respective life statuses over time (diagnosis date set as zero).....	28
Figure 10 Kaplan-Meier breast cancer patient trajectory grouped by treatments.....	30
Figure 11 Kaplan-Meier breast cancer patient trajectory grouped by cancer stage T from TNM.....	30
Figure 12 Kaplan-Meier breast cancer patient trajectory grouped by cancer stage N from TNM.....	31
Figure 13 Kaplan-Meier breast cancer patient trajectory grouped by cancer stage M from TNM.....	31
Figure 14 Kaplan-Meier breast cancer trajectory of ten random patients.....	32
Figure 15 Cox Proportional Hazards for feature importance of breast cancer patients....	33
Figure 16 Survival classification accuracy Logistic Regression, SVM, Decision Tree, and Neural Networks	34
Figure 17 Risk level for patients in prospective data for colon cancer	35
Figure 18 Risk level for patients in prospective data for breast cancer	35
Figure 19 Kaplan-Meier colon cancer patient trajectory grouped by cancer stage M from TNM.....	36
Figure 20 Kaplan-Meier breast cancer patient trajectory grouped by cancer stage M from TNM.....	36
Figure 21 Cox Proportional Hazards survival probability for a colon cancer patient in prospective data	37
Figure 22 Cox Proportional Hazards for feature importance breast cancer patients.....	37
Figure 23 Cox Proportional Hazards for feature importance colon cancer patients	38
Figure 24 Cox Proportional Hazards for feature importance breast cancer patients enriched dataset.....	38
Figure 25 FHIR resource distribution of data from CHU.....	40
Figure 26 Distribution of missing data in breast cancer patients	42
Figure 27 Distribution of missing data in colon cancer patients	42
Figure 28 Components that communicate with the relapse risk prediction service	45
Figure 29 Relapse risk prediction generation process	46
Figure 30 First version of trajectory analysis API documented with swagger	51
Figure 31 List of common input parameters of the API documented with swagger.....	51

Index of Tables

Table 1 FHIR General Resource provided by CHU, SERGAS, UL and UKCM	14
Table 2 FHIR Medical History Resource provided by CHU, SERGAS, UL and UKCM....	14
Table 3 FHIR Diagnosis Resource provided by CHU, SERGAS, UL and UKCM	15
Table 4 FHIR Symptoms Resource provided by CHU, SERGAS, UL and UKCM	16
Table 5 FHIR Genetic Results Resource provided by CHU, SERGAS, UL and UKCM...	16
Table 6 FHIR Tumoral Markers Resource provided by CHU, SERGAS, UL and UKCM .	16
Table 7 FHIR Diagnosis Test Resource provided by CHU, SERGAS, UL and UKCM.....	17
Table 8 FHIR Treatments Resource provided by CHU, SERGAS, UL and UKCM	17
Table 9 FHIR Medication Resource provided by CHU, SERGAS, UL and UKCM.....	18
Table 10 FHIR Procedure code from CHU	22
Table 11 Before OHE, feature treatment with only two values.....	23
Table 12 After OHE, two new columns were added with binary values.	23
Table 13 Before LE, feature T of TNM	24
Table 14 After LE, the values of T are numeric.	24
Table 15 Results obtained for the breast cancer recurrence prediction models	44
Table 16 Results obtained for the colon cancer recurrence prediction models.	44
Table 17 Recovered variables from Observation Resources	48
Table 18 Recovered variables from Condition Resources with patient ID and requested type of cancer	49
Table 19 Recovered variables from Clinical and Pathological cancer staging system TNM	50
Table 20 Recovered variables from Colon cancer comorbidities	50
Table 21 Recovered variables from Breast cancer comorbidities	50

Acronyms and abbreviations

ACRONYM	TITLE
AES	Advanced Encryption Standard
AI	Artificial intelligence
API	Application Programming Interface
API	Application Programming Interface
BMI	Body Mass Index
CDSS	Clinical Decision Support Systems
CEL	CYBERETHICS LAB SRLS
CHU	CENTRE HOSPITALIER UNIVERSITAIRE DE LIEGE
CNN	Convolutional Neural Network
Consortium	Means the consortium created by the execution of the CA
CPH	Cox Proportional Hazard
DMN	Dynamic Memory Network
DoA	Description of Actions
DXC	IT CORPORATE SOLUTIONS SPAIN SL
EC	European Commission
ECOG status	Eastern Cooperative Oncology Group status
EDPB	European Data Protection Board
EHR	Electronic Health Record
EMO	EMODA BILGISAYAR YAZILIM CEVRE DONANIMLARI REKLAMCILIK BEYAZ ESYA IKLIMLENDIRME TEKSTIL SANAYI VE TICARET LIMITED SIRKETI
EMR	Electronic Medical Record
ENISA	European Union Agency for Cybersecurity
EU	European Agency
FHIR	Fast Healthcare Interoperability Resources
GDPR	Regulation of the EU Parliament and Council no. 2016/679
GRAD	FUNDACION CENTRO TECNOLOGICO DE TELECOMUNICACIONES DE GALICIA
HER2	Human Epidermal Growth Factor Receptor-2
HES-SO	HAUTE ECOLE SPECIALISEE DE SUISSE OCCIDENTALE
HL7	Health Level 7
JSON	JavaScript Object Notation
LE	Label Encoder transformation
METABRIC	Molecular Taxonomy of Breast Cancer International Consortium
ML	Machine Learning

NLP	Natural Language Processing
NPO	NATIONAL PATIENTS ORGANISATION
OHC	Open Health Connect
OHE	One-Hot-Encoder
OVH	OVH SAS – cloud provider
Partners	Means the PERSIST partners as indicated within the CA
Project	Means the PERSIST Project
RNN	Recurrent Neural Network
ROC Curve	Receiver Operating Characteristic Curve
RUBY	RUBYNANOMED, UNIPESSOAL LDA
SERGAS	SERVIZO GALEGO DE SAUDE
SSL	Transport Layer Security
SYMP	SYMPTOMA GMBH
TCGA	The Cancer Genome Atlas
TNM classification	Cancer classification based on size of primary tumor (T), lymphatic nodes propagation (N) and metastasis presence (M)
UKCM	Univerzitetni klinicni center Maribor
UL	LATVIJAS UNIVERSITATE
UM	UNIVERZA V MARIBORU
WP	Work Package
WP29	Working Party 29

Executive Summary

This deliverable report is written in the framework of WP5 – Decision support system at the point of care (T5.3 Cohort and trajectory analysis) of PERSIST project under Grant Agreement No. 875406.

Deliverable D5.8 provides the end-to-end steps required for the cohort and patient trajectory analysis. This deliverable includes retrospective, prospective, and enriched data for colon and breast cancer in PERSIST. It consists of data preparation and pre-processing, feature extraction, AI-based models, and the design of the APIs to access these models.

The goal of this deliverable is:

- To describe the data ingestion and preparation process from the retrospective datasets provided by the hospitals in the PERSIST consortium.
- To describe the approach and architecture designed for the patient cohort and trajectory analysis.
- To describe the results produced by the patient cohort and trajectory analysis as an input for the clinical decision support system of PERSIST.
- To describe the data ingestion and preparation process from the prospective datasets provided by the hospitals in the PERSIST consortium.
- To describe the architecture and deployment of the API designed to connect the PERSIST cohort-trajectory system with the CDSS.
- To describe the integration and usage of the enriched dataset provided by T5.2 (EHR data extraction, anonymization, preparation, and enrichment) as an input for the cohort-trajectory system.
- To describe the results of multiple-level cohort analysis using neural networks.
- To describe the result of the analysis of high-risk markers for detrimental treatment effects such as depression and anxiety.

Introduction

1. PERSIST Project

PERSIST aims at developing an open and interoperable ecosystem to improve the care of cancer survivors. The key results to be achieved by partners are:

- Related to patients: increased self-efficacy and satisfaction with care as well as reduced psychological stress for better management of the consequences of the cancer treatment and the disease.
- Related to professionals: increased effectiveness in cancer treatment and follow-up by providing prediction models from Big Data that will support decision-making and contribute to optimal treatment decisions.
- Related to healthcare providers: improved information and evidence to advance the efficacy of management, intervention, and prevention policies. The long-term result will be to reduce the socio-economic burden related to cancer survivors' care.

2. Work Package 5: Decision support system at the point of care

The overall goal of WP5 is to provide and implement a toolkit of basic and cutting-edge analytical methods to build the data mining and knowledge discovery services for guidance and support of the decision-making and personalization of the survivor care plan. This work package will deliver a knowledge-based decision support system. The data-driven tools to be implemented will analyze the information in the databases and give conclusions related to the usage of the self-management platform, reported adverse events and health issues, compliances, health status, quality of life, etc. These tools and services are integrated into the big data platform and support smart data analytics. The specific objectives are:

- To develop a clinical decision support system to support the clinicians for diagnosis, treatment, and follow-up of cancer patients.
- To carry out data exploration, harmonization, visualization, and basic statistics in a way that uncovers new insights into the research questions being asked.
- To evaluate the clinical decision support system in the relevant environment (validation of results of Machine Learning algorithms).

Role of each partner: EMO is the leader of WP5 and will be responsible for software requirements specification, software design, implementation, integration, and evaluation of CDSS. SYMP contributes with its expertise in EHR normalization and processing to define and develop solutions to convert raw exported data records into a conclusive format. HESSO leads the patients' cohort and trajectory analysis that populates the knowledge base for the CDSS. All the involved partners are supported by the clinical partners of the consortium, which will contribute to the development of CDSS and the rule-based inference engine with their clinical expertise and perspectives in an iterative process. The CDSS is one of the main results of the project.

3. Deliverable D5.8 High-risk markers for patient stratification

The scope of this deliverable is to produce the cohort and trajectory analysis, which focuses on identifying high-risk markers for detrimental treatment effects, subsequent cancer disease, and metastatic cancer disease.

This deliverable is the result of the work in Task 5.3 Cohort and Trajectory Analysis. The results of this task are the basis for Task 5.4 CDSS Inference Engine in order to provide the potential leverage points for defining the clinical decision support. The identification of markers that might be changeable (gender, ethnicity, etc.) or unchangeable (treatment combination, lifestyle, etc.) in nature will be achieved by separation into cohorts and sub-cohorts. We perform multiple-level cohort analysis through miscellaneous approaches (supervised/unsupervised) using miscellaneous analysis methods such as vector machine, regression analysis, neural networks, etc. High dimensional unsupervised analysis approaches have proven to identify novel patient cohorts revealing unknown cause-effect relations. We identify similar trajectories of breast and colorectal cancer patients. Trajectories represent the evolution of the patient from the diagnosis of the disease. Two types of trajectories are learned from data. (I) Disease trajectories will associate symptoms to diseases and describe how patients progress from symptoms to disease (and eventually, to death). (II) Event trajectories identify associations between symptoms and events (admission, readmissions, treatments, etc.) and are used to quantify risks. Visually, these trajectories are represented as graphs containing (potentially) several paths. Each path consists of events that may occur sequentially. The transition between one event (e.g., primary tumor detection) and the other (e.g., metastasis) will be associated with a probability of occurrence that will be learned from data.

This deliverable resumes the work provided in D5.4 High Risk Markers for patient stratification. It adds the analysis of both cancer patients (i.e., colon and breast cancers); it includes prospective data from the PERSIST dataset; it enhances trajectory dimensionality by analyzing enriched data; and finally, it provides access to the analysis via APIs integrated into the PERSIST domain.

Data ingestion

1. Data Flow Architecture

In order to export sensitive clinical data from closed clinical environments, several steps need to be taken to ensure GDPR compliance, which includes measures to maximize cybersecurity, traceability, and minimize the risk of re-identification. The PERSIST architecture is managed through the OHC platform, which ingests data through the VIADUCT component (see Figure 1). The data from each hospital is imported in this manner, after a pseudo anonymization process. A PERSIST identifier is generated, which will be used within the context of the project processing tasks, as in this case in T5.3.

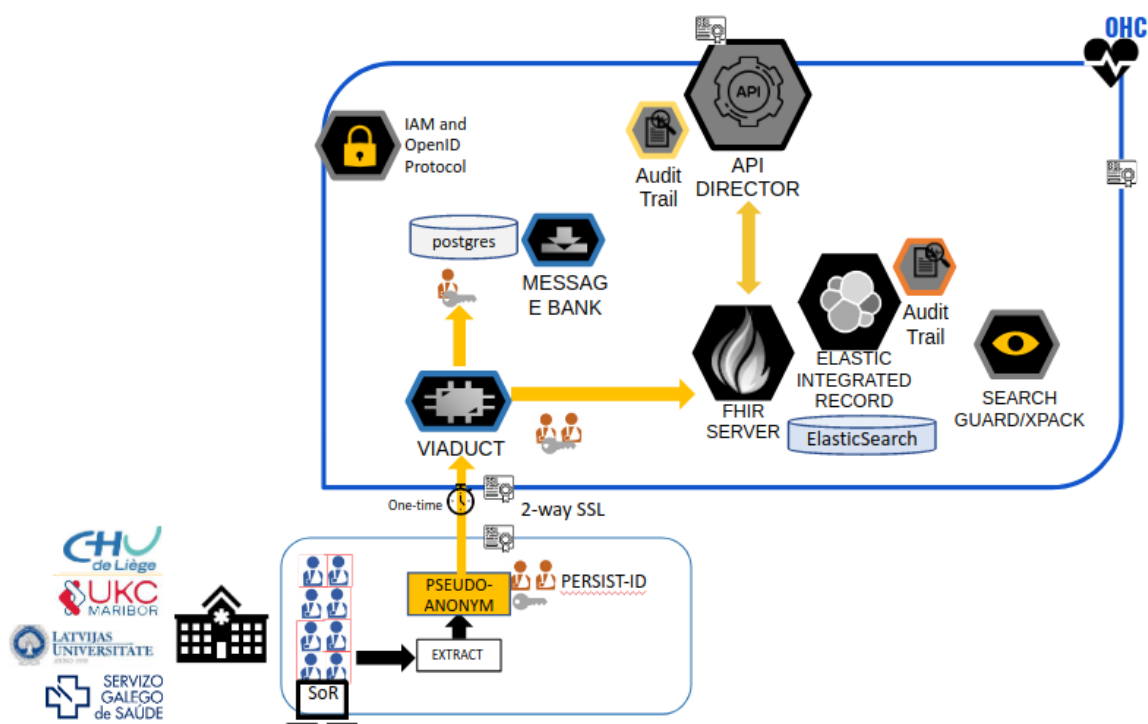


Figure 1 Schematic figure of the data export from clinical partners

For the purpose of the patient cohort and trajectory analysis, the FHIR server in the OHC platform has been queried, so that the required ICD and SNOMED codes are retrieved. Given the heterogeneity of the data from different hospitals, we started with patient data from the CHU hospital (or details in section 3). In particular, FHIR “Patient” resource was queried in order to retrieve basic demographics (identifier, birthdate, gender, deceased). The FHIR “Condition” resource was also queried in order to filter breast and colon cancer conditions, and the “Observation” resource to obtain the tumor stage. Finally, FHIR “Procedure” resource to retrieve patients' treatments.

As shown in Figure 2 and Figure 3 all records comply with the FHIR Release 3 (STU), an international standard that holds the advantage of being highly flexible and interoperable.

JSON-FHIR was chosen as a data format that allows for more flexibility in catering to downstream applications. FHIR allows for individual configuration to match the individual needs of data owners and analyzers, also called "profiles". Therefore, an individualized version for PERSIST is created: "PERSIST-JSON-FHIR".

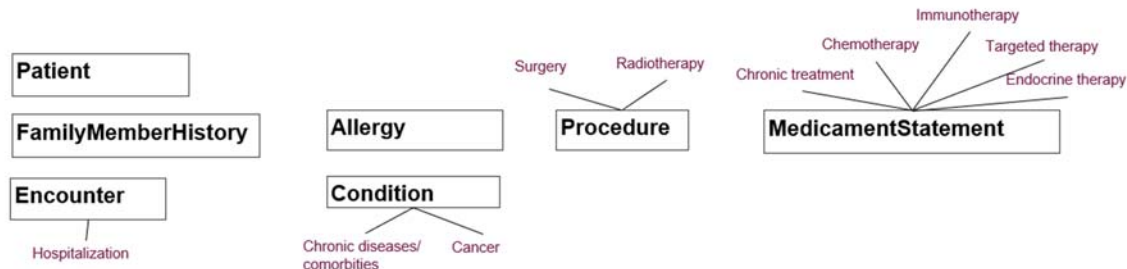


Figure 2. FHIR resource mappings – FHIR Release 3 (STU)

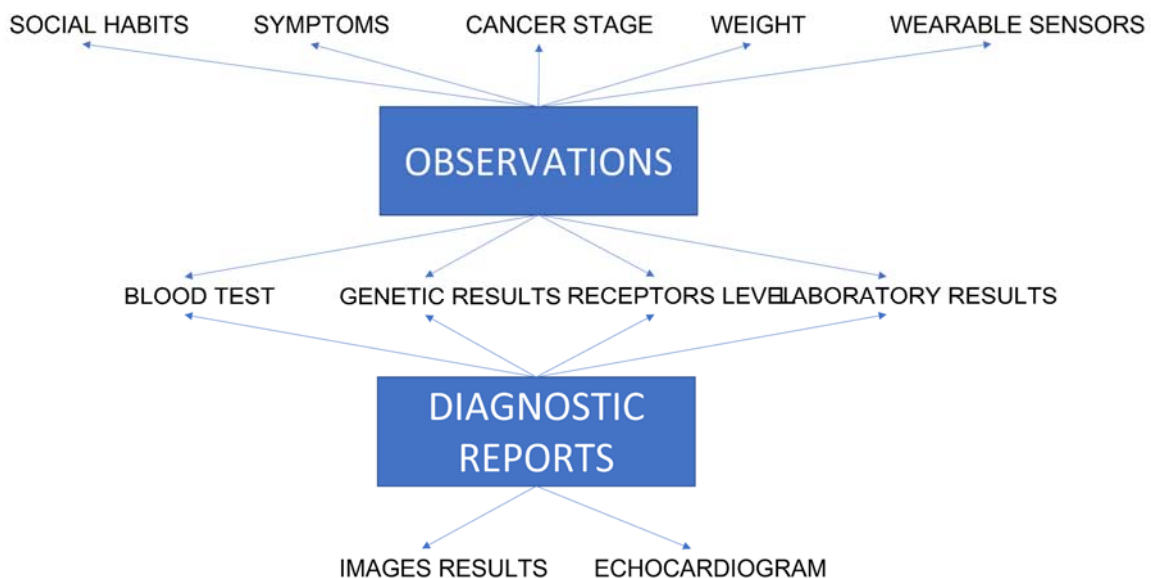


Figure 3. FHIR resource mappings - FHIR Release 3 (STU)

2. Common data model

A common data model between data owners (clinics & hospitals) and technical partners (GRAD, SYMP, HESSO, DH) was established to have a uniform data set for all further analysis steps. For this purpose, data content was evaluated individually (which kind of data (doctors' reports, lab values, etc.), which data formats (pdf, text, csv, etc.)) and a minimal viable set of data defined, in order to match data quality across all participating clinics. For the patient cohort and trajectory analysis, the following data has constituted the basis of the processing tasks (Tables 1 to 9).

FHIR resource	VARIABLES	CHU	SERGAS	UL	UKCM	
General						
male female other unknown	Patient.gender	Gender	X	X	X	X
YYYY-MM-DD	Patient.birthDate	Age	x	X	X	X
Coding system: SNOMED Ethnic group: 364699009	Observation	Ethnicity	?	X	X	
https://www.hl7.org/fhir/valueset-marital-status.html	Patient.maritalStatus	Marital Status	X	X		X
coding system: SNOMED Code: 365539008 Finding of type of job (finding)	Observation	Occupation/Job				
deceasedBoolean (true/false) deceasedDateTime (YYYY-MM-DDThh:mm:ss)	Patient.deceasedBoolean Patient.deceasedDateTime	Deathdate				X
code: SNOMED 184305005 Cause of death (observable entity)	Observation	Cause of death		X		X
Coding system: LOINC Body Weight: 29463-7 unit: kg	Observation	Weight	X	X	X	X
Coding system: LOINC Body height: 8302-2 unit: cm	Observation	Height	X	X	X	

Table 1 FHIR General Resource provided by CHU, SERGAS, UL and UKCM

FHIR resource	VARIABLES	CHU	SERGAS	UL	UKCM	
Medical History						
Coding system SNOMED. Breast Cancer: code- 254837009. Colon Cancer: code- 363406005. (Other diseases are advisable)	FamilyMemberHistory	Family predispositions: 1st degree relative.	X	X	X	
Coding system SNOMED. Code field NOT mandatory. If plain text -> use "text" field	AllergyIntolerance	Allergies	X	X	X	X
Code SNOMED. If other code system used, please describe. If info is unstructured, then use field "note"	Condition	Previous diseases/comorbidities	X	X	X	X
(* go to "Treatment" tab. Row: "Chronic treatment")	MedicationStatement	Medications	X	X	X	X
Coding system SNOMED Alcohol: 228281002	Observation (social history)	Alcohol	X	X	X	
Coding system SNOMED Drugs: 361055000	Observation (social history)	Drugs	X	X	X	X
Coding system SNOMED Tobacco: 77176002	Observation (social history)	Smoking	?	?	X	X
Coding system: LOINC BMI: 39156-5 unit: kg/m2	Observation	BMI	X	X	X	

Table 2 FHIR Medical History Resource provided by CHU, SERGAS, UL and UKCM

FHIR resource	VARIABLES	CHU	SERGAS	UL	UKCM	
Diagnosis						
Condition.category Unknown-category, No-cancer-disease, Cancer-disease, Cancer-disease-C18, Cancer-disease-C50 <u>clinicalStatus</u> : active recurrence relapse inactive remission resolved	Condition	Cancer type				X
Cancer diagnosis, For example: System: SNOMED Code: 254837009 Malignant neoplasm of breast (disorder) Code: 363406005 Malignant neoplasm of colon (disorder) System: ICD-10 Code: C50 Malignant neoplasm of breast	Condition.code	Type of tumor	X	X	X	X

Code: C18 Malignant neoplasm of colon						
Histological grade Observation.code: System: SNOMED Code: 371469007. Available values in SNOMED system are: Grade X - 60815008, Grade 1 - 373375007, Grade 2 - 373377004, Grade 3 - 373373000, Grade 4 - 373374006	Observation	Cancer grade	X	X	X	X
Condition.stage stores the current stage of a disease. Observation.code: Not known if clinical or pathologic: System: SNOMED Code: 399390009. Clinical stage: System: SNOMED Code: 399537006 Pathologic stage: System: SNOMED Code: 399588009 Observation.valueCodeableConcept : stage 0 - 718465002, stage I - 13104003, stage II- 60333009, stage III - 50283003, stage IV -2640006	Observation	Cancer stage	X	X	X	X
Observation.code: Not known if clinical or pathologicTNM : System: http://snomed.info/sct. Code: 399566009. Clinical TNM: System: http://snomed.info/sct. Code: 106248000. Pathologic TNM: System: http://snomed.info/sct. Code: 106249008. Subcomponents: Tumor category - 78873005; Node category - 277206009; Metástasis category - 277208005	Observation	TNM categories	X	X	X	
Observation.code: Code: 371441004 System: http://snomed.info/sct Display: Histologic type (observable entity) Observation.valueCodeableConcept : SNOMED: Code from hierarchy of 367651003 "Malignant neoplasm of primary, secondary, or uncertain origin (morphologic abnormality)	Observation	Type of tumor (morphology)				

Table 3 FHIR Diagnosis Resource provided by CHU, SERGAS, UL and UKCM

FHIR resource	VARIABLES	CHU	SERGAS	UL	UKCM
Symptoms					
SNOMED code: 422587007	Observation	nausea	X	X	X
SNOMED code: 62315008	Observation	diarrhea	X	X	X
SNOMED code: 84229001 SNOMED values: Present - 52101004; Absent - 272519000	Observation	fatigue	X	X	X
SNOMED code: 290070001 - Red breast (finding)	Observation	skin irritation	X	X	X
SNOMED code: 300885006	Observation	Swelling of all or part of the breast	X	X	X
SNOMED code: 53430007	Observation	pain	X	X	X
SNOMED code: 31845005	Observation	Nipple retraction (turning inward)	X	X	X
SNOMED codes:	Observation	The nipple or breast skin	X	X	X

Swelling of nipple - 290101004 Erythema of nipple - 290095000 Eczema of nipple - 237463009 Nipple bleeding - 290103001		appears red, scaly, or thickened.				
SNOMED code: 162164007	Observation	Nipple discharge	X	X		X
SNOMED code: 14760008	Observation	constipation.	X	X		X
SNOMED code: 2761008	Observation	stools that appear narrower than usual.	X	X		X
SNOMED code: 247339003	Observation	feeling that the rectum is not completely empty after having a bowel movement.	X	X		X
SNOMED code: 405729008 - Hematochezia (finding)	Observation	light or very dark red blood in the stool.	X	X		X
SNOMED code: 12063002	Observation	bleeding from the rectum.	X	X		X
SNOMED code: 45979003 - Abdominal wind pain (finding) 249531009 - Bowel spasm (finding) 60728008 - Swollen abdomen (finding)	Observation	gas, abdominal cramps and bloating.	X	X		X
SNOMED code: 77880009	Observation	pain or discomfort in the rectum.	X	X		X

Table 4 FHIR Symptoms Resource provided by CHU, SERGAS, UL and UKCM

FHIR resource		VARIABLES	CHU	SERGAS	UL	UKCM
Genetic Results						
Observation.code Coding system LOINC. BRCA1: 21636-6 BRCA2: 38530-2	Observation MAY be nested in DiagnosticReport	BRCA1	X	X	X	
	Observation MAY be nested in DiagnosticReport	BRCA2	X	X	X	
Observation.code Coding system LOINC. HNPCC: 35379-7	Observation MAY be nested in DiagnosticReport	HNPCC	X	X	X	
Observation.code Coding system LOINC. KRAS: 21702-6	Probably HNPCC, KRAS, NRAS, BRAF are nested together. The code for that report hasn't been found.	KRAS				
Observation.code Coding system LOINC. NRAS: 21719-0		NRAS				
Observation.code Coding system LOINC. BRAF: 58483-9		BRAF				

Table 5 FHIR Genetic Results Resource provided by CHU, SERGAS, UL and UKCM

FHIR resource		VARIABLES	CHU	SERGAS	UL	UKCM
Tumoral Markers						
System: LOINC code: 6875-9 - Cancer Ag 15-3 [Units/volume] in Serum or Plasma unit: U/mL	Observation (nested if possible in the same DiagnosticReport)	CA-15-3		X		
System: LOINC code: 24108-3 - Cancer Ag 19-9 [Units/volume] in Serum or Plasma unit: U/mL	Observation (nested if possible in the same DiagnosticReport)	CA 19-9		X		
System: LOINC code: 2039-6 - Carcinoembryonic Ag [Mass/volume] in Serum or Plasma unit: ug/L	Observation (nested if possible in the same DiagnosticReport)	CEA		X		

Table 6 FHIR Tumoral Markers Resource provided by CHU, SERGAS, UL and UKCM

FHIR resource	VARIABLES	CHU	SERGAS	UL	UKCM
Diagnosis tests					
System: LOINC code: 16112-5 - Estrogen receptor [Interpretation] in Tissue values: %	Observation nested in DiagnosticReport	Estrogen receptor level	X	X	
System: LOINC code: 16113-3 - Progesterone receptor [Interpretation] in Tissue	This Observation could be nested inside a DiagnosticReport	Progesterone receptor level	X	X	
Her2 level can be measured with FISH test and with Inmuno Stain test. System: LOINC Code: 31150-6 Display: HER2 [Presence] in Tissue by FISH Values: from SNOMED: Positive: 10828004; Negative: 260385009 To store Inmuno Stain test: System: LOINC Code: 85319-2	Coding system LOINC 74293-2. Oncology plan of care and summary - recommended CDA set	Her2 level	X	X	
System: SNOMED code: 423740007 - ECOG Performance Status SNOMED codes for values: System: SNOMED	Observation	Performance Status	X	X	
Observation code. System: LOINC code: 81695-9 - Microsatellite instability Observation valueCodeable concept: LOINC Stable LA14122-8; MSI-L LA26202-4; MSI-H LA26203-2; Indeterminate LA11884-6	Observation for Microsatellite Instability with components for individual markers.	Microsatellite instability This is a panel.	X	X	

Table 7 FHIR Diagnosis Test Resource provided by CHU, SERGAS, UL and UKCM

FHIR resource	VARIABLES	CHU	SERGAS	UL	UKCM
Treatment					
Encounter.class: https://www.hl7.org/fhir/STU3/v3/ActEncounterCode/vs.html (AMB, IMP,...)	Encounter	Hospitalization		X	X
	Encounter.period	-Duration (days)		X	X
	Procedure	Type of surgery		X	X
SNOMED code: 69031006	Procedure	-mastectomy		X	X
SNOMED code: 370612006	Procedure	-tumorectomy		X	X
SNOMED code: 234262008	Procedure	-axillary lymph node dissection		X	X
SNOMED code: 396487001	Procedure	-sentinel node biopsy		X	X
SNOMED code: 398740003	Procedure	-colostomy		X	X
SNOMED code: 73761001	Procedure	-colonoscopy			
SNOMED code: 33496007	Procedure	-Breast reconstruction		X	
	Procedure	Radiotherapy		X	X
SNOMED codes preferred	Procedure.bodySite	-Zone		X	X
SNOMED code: 108290001- Radiation oncology AND/OR radiotherapy If more specific modality was addressed, it can be specified here.	Procedure.code	- Session modality		X	X

Table 8 FHIR Treatments Resource provided by CHU, SERGAS, UL and UKCM

	FHIR resource	VARIABLES	CHU	SERGAS	UL	UKCM
Medication						
SNOMED code (generic): 76334006	MedicationStatement.medicationCodeableConcept	Immunotherapy		X	X	
SNOMED code (generic): 367336001	MedicationStatement.medicationCodeableConcept	Chemotherapy	X	X	X	
First and second chemotherapy can be distinguished thanks to effectivePeriod	MedicationStatement.effectivePeriod	First Chemotherapy		X	X	
	MedicationStatement.effectivePeriod	Second Chemotherapy		X	X	
No field to indicate that this chemotherapy will be performed before a surgery. Probably only the date can be used to distinguish it.	MedicationStatement.effectivePeriod	Neoadjuvant treatment			X	
SNOMED code (generic): 448288001 - Targeted radionuclide therapy	MedicationStatement	Targeted therapy		X	X	
SNOMED code (generic) 169413002 - Hormone therapy (procedure)	MedicationStatement	Endocrine therapy		X	X	
Since medication is chronic => "status" should be active	MedicationStatement	Chronic treatment		X	X	X

Table 9 FHIR Medication Resource provided by CHU, SERGAS, UL and UKCM

3. Anonymisation and risk for re-identification

In order to export sensitive clinical data, every data set needs to be anonymized and additionally checked for risk of re-identification. For each clinic, an individual assessment was performed and problematic features were identified, potential risks determined. These have been discussed previously in D5.2 for the PERSIST project.

3.1. Privacy Metrics:

Different privacy metrics are calculated to determine the risk of re-identification (or privacy level) of the dataset. These metrics are based on population uniqueness: the re-identification risk is based on the concept of uniqueness in the sample and/or in the population. In general, it is assumed that individuals having rare combinations of variables, or quasi-identifiers, can be more easily identified and therefore have a higher risk of re-identification.

- K: this metric is based on the k-anonymity strategy: computes how many rows on the dataset are equal to other k-1 rows. All the columns in the input dataset are considered during the metric calculation. For example, a value of k = 100 means that the smallest group we can create has 100 individuals, and those individuals within the same group are indistinguishable between them.
- Constrained Attacker K (CAK-N): similar to K metric, but less restrictive since it only considers the specified columns during the metric computation. This metric assesses the probability of an attacker being able to identify an individual if he/she knows the values of the specified features. For example, the attacker could know where they live but not their weight. N represents the number of columns considered during the analysis.

3.2. Anonymization techniques to address a high re-identification risk

Several anonymization techniques that are easy to implement were suggested that can help to reduce the re-identification risk on a dataset, depending on the data type:

- Dates: dates can be easily anonymized by trimming the day, month or year. For example, group the birth dates into decades.
- Numeric values: numeric values can be grouped in ranges (for example, in ranges of 10, a value of 91 and 95 will be mapped to 90). More advanced techniques can use machine learning clustering algorithms to create groups of data automatically like KMeans.
- Categorical values: categories of data can be created, and certain values mapped to a given category. For instance, a set of diseases can be grouped together as a more general one. For example, all codes related to colorectal cancer (C183, C184, C185, etc) can be mapped to the C18 code.
- Identifiers: personal identifiers cannot be anonymized and must be deleted.
- Unused data: this data can be deleted to improve privacy.

For more details on privacy metrics and anonymization techniques, please refer to the related WP.

Feature extraction

Unlocking the potential of data harbored in electronic health records (EHRs) is one of the most promising directions pursued in the digital medicine community. In this section, we describe the extraction for the cohort and trajectory analysis. We start by detailing the data flow and the data retrieval. The image below illustrates the system data flow for cohort and trajectory analysis. In the first version of the deliverable, we focused on the communication between OHC “cohort and trajectory service” and “production server.”

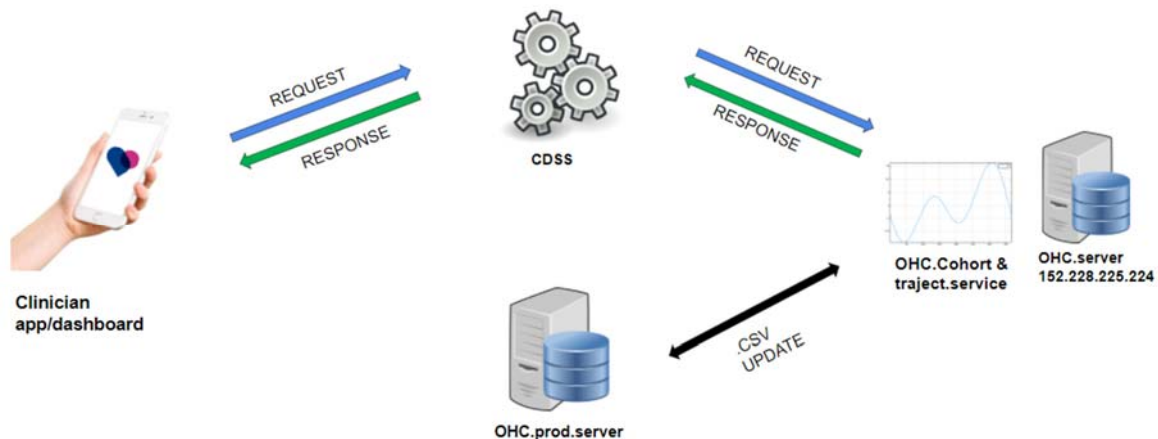


Figure 4 Trajectory and cohort data flow

In order to keep the data flow within the OVH infrastructure of the PERSIST project, we set up the trajectory and cohort analysis on the OVH server (IP: 152.228.225.224). Once we finalized the server installation with the ubuntu2010-server_64 operating system, we accessed the API of the production server to retrieve the PERSIST patient retrospective data. The first API call was made to the “Patient” resource in order to retrieve:

- ➔ patients' ids (text),
- ➔ date of birth (numeric),
- ➔ gender (text),
- ➔ deceased death (numeric).

In D5.4, given the data heterogeneity, we mainly focused on the breast cancer patients from CHU (2085 breast cancer patients, 1753 colorectal cancer patients). In D5.8, we extended our system to other hospitals' prospective data, cancer types, and enriched data. To extract and separate colon and breast cancer patients, we queried the “Condition” resource on the production server checking for patient codes starting with “C50”, which is the breast cancer code according to the ICD coding system (“C18” for colon cancer).

A significant feature of the cohort and trajectory analysis is the TNM classification of malignant tumors. CHU adopts the SNOMED coding system and it provides the clinical TNM and the pathologic TNM (respectively 106248000 and 106249008 SNOMED codes). To retrieve the TNM categories from the production server, we queried the “Observation”

resource by patient id. Given the huge number of observations per patient, we parallelized the queries and stored the results in a python pandas data frame. Here is the data extracted:

- ➔ Clinical tumor stage T category (Text),
- ➔ Clinical tumor stage N category (Text),
- ➔ Clinical tumor stage M category (Text),
- ➔ Pathologic tumor stage T category (Text),
- ➔ Pathologic tumor stage N category (Text),
- ➔ Pathologic tumor stage M category (Text),

Another significant feature collected from the production server is the type of treatments for the breast cancer patient. We queried the “Procedure” resource checking for patient treatments such as mastectomy, tumorectomy, axillary lymph node dissection, and sentinel node biopsy. Moreover, we checked for immunotherapy, chemotherapy, and injection of antibiotics. Please note that even if we tackled only breast cancer from CHU, given the different versions of the coding system (e.g., ICD 9 and ICD 10), the same treatment can have different codes. Here is the table with the most popular treatments in our population and the different codes for the same treatments (Table 10).

Code System	Code	Description	Equivalent Code System	Equivalent Code
ICD-9	85.21	Breast-conserving surgery Local excision of lesion of breast Lumpectomy Removal of area of fibrosis from breast	ICD-10	0HBT0ZZ / 0HBU0ZZ
ICD-9	40.11	Biopsy of lymphatic structure	ICD-10	07B50ZX / 07B60ZX
ICD-9	99.25	chemotherapy Injection or infusion of cancer chemotherapeutic substance Chemoembolization Injection or infusion of antineoplastic agent Use additional code for disruption of blood brain barrier, if performed [BBBD] (00.19)	ICD-10	3E04305 /3E03305
ICD-9	86.07	Insertion of totally implantable vascular access device [VAD] Totally implanted port	ICD-10	0JH60XZ
ICD-9	40.23	Sentinel node biopsy - Excision of axillary lymph node	ICD-10	07B50ZZ / 07B60ZZ
ICD-9	38.93	Venous catheterization, not elsewhere classified	ICD-10	02HV33Z
ICD-9	39.97	Other perfusion NOS Perfusion, local [regional] of: carotid artery coronary artery	ICD-10	3E03305

		head lower limb neck upper limb Code also substance perfused (99.21-99.29)		
ICD-9	85.43	Mastectomy , Unilateral extended simple mastectomy Extended simple mastectomy NOS Modified radical mastectomy Simple mastectomy with excision of regional lymph nodes	ICD-10	0HTU0ZZ
ICD-9	99.21	Injection of antibiotic		
ICD-9	85.22	Resection of quadrant of breast	ICD-10	BP0G0ZZ / BP1G0ZZ

Table 10 FHIR Procedure code from CHU

Finally, the pandas data frame has enough data to be processed in the cohort and trajectory analysis models. Please notice that data such as cancer relapse, family history, and many others are in free text form handled by T5.2. Those data, together with data from other hospitals, colon cancer, and prospective data are included in D5.8. Most importantly, in such a deliverable, we implement and document the connection between CDSS and PERSIST cohort-trajectory system as previously illustrated. In the next section, we describe how we prepared the data for the cohort and trajectory analysis.

Data preparation

1. Retrospective data

In order to prepare the retrospective PERSIST data for the cohort and trajectory analysis, a data wrangling process has been applied. As the first step, we created a data frame containing patient id, treatments, and TNM categories. We separated numerical features (continuous and discrete) and categorical features (independent and dependent categories). For each feature, we analyzed missing data. Whenever the missing data was not statistically relevant (e.g., T category of missing only from the pathology), it was filled with values that do not affect the distribution (e.g., clinical T category or median of the distribution). In case the missing data was relevant (e.g., missing clinical and pathology TNM), we discarded the patients from the model analysis.

For independent categorical values such as the treatments code, we transformed the data using one-hot-encoder or OHE [1]. This encoder creates as many columns as the number of different values in a feature. For each new column, OHE adds “true” if the patient has received that treatment. Here is an example of OHE (Tables 11 and 12):

Patient Id	Treatments	...
id1	9921	...
id2	9945	...
...	9945	...
idn	9921	...

Table 11 Before OHE, feature treatment with only two values.

Patient Id	9921	9945
id1	true	false
id2	false	true
...	false	true
idn	true	false

Table 12 After OHE, two new columns were added with binary values.

Please notice that a patient can have received multiple treatments therefore, the values are not mutually exclusive.

For dependent categorical features, such as TNM values ($T_0 < T_1 < T_2$, etc), we applied the label encoder transformation or LE [2]. This encoder transforms categorical features into numerical features keeping their distance relationship. Here is an example of LE (Tables 13 and 14):

Patient Id	T	...
id1	T2	...
id2	T1	...
...	T3	...
idn	T0	...

Table 13 Before LE, feature T of TNM .

Patient Id	T	...
id1	2	...
id2	1	...
...	3	...
idn	0	...

Table 14 After LE, the values of T are numeric.

Please notice that, unlike OHE, LE keeps the feature dependency (e.g., $T_0 < T_1 < T_2 < T_3$, etc.).

After transformation, both categorical and numerical features are normalized with values between $[0, 1]$ using a min-max normalization. This normalization allowed the model to be trained fairly.

The last step of data preparation is feature engineering. For instance, in order to shape the patient survival probability, we need to have the event and the related time. Therefore, as the event, we chose the deceased event (if the patient is dead “true”, otherwise “false”), and as event date, we selected the deceased date or the last patient visit. Those features were extracted by the “Observation” resource on the OHC server in FHIR format. Given that feature, engineering is based on the model selected. More details are provided in the next section.

2. Prospective data

The prospective dataset contains a total of 182 patients. Among them, 46 are from UL, 44 from UKCM, 41 from Sergas, and 51 from CHUL.

Looking at the repartition of the patients by cancer type, we notice that both populations have the same number of patients (see Figures 5 and 6). However, given the nature of breast cancer, the whole population is female. The issue does not persist for colon cancer, which presents a gender-balanced population (see Figure 7).

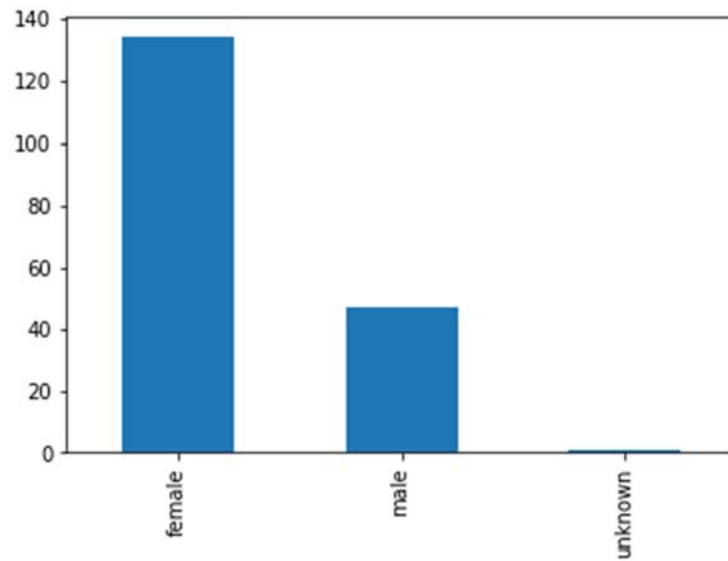


Figure 5 Repartition of patients by gender in the whole PERSIST prospective dataset

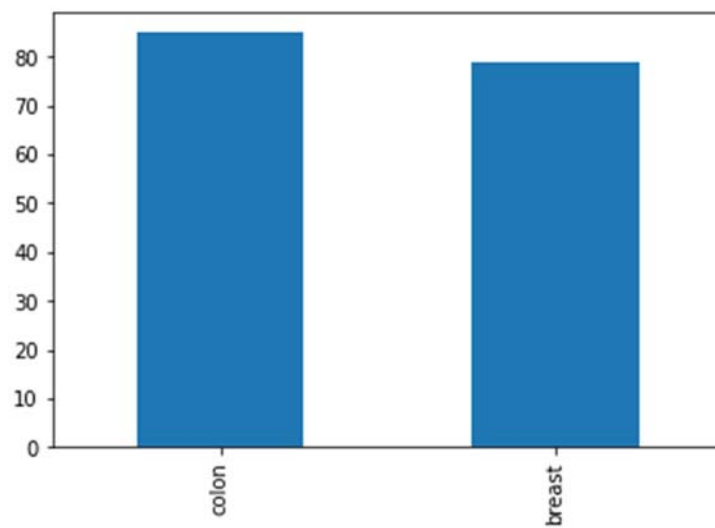


Figure 6 Repartition of patients by cancer type in the PERSIST dataset

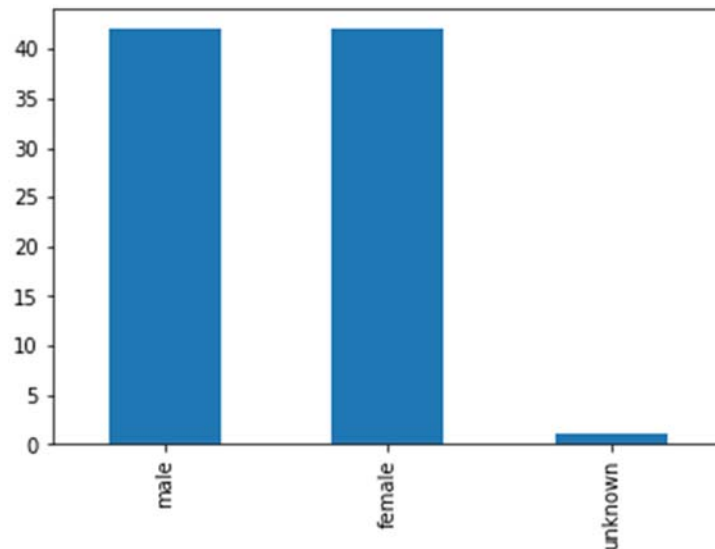


Figure 7 Repartition of colon cancer patients by gender in the PERSIST dataset

A data-driven model requires much data to enhance its accuracy and reproducibility. Unfortunately, the PERSIST prospective dataset suffers from a low sample population, which involves bias (e.g., 95% treatment A, 5% treatment B, or gender representation). To solve this issue, we merge retrospective and prospective data. However, this solution leads us to another challenge: when a feature exists only in one of the two datasets.

Indeed, some treatments in the retrospective dataset are not found in the prospective data. For instance, procedures with icd-10 codes 9925 (injection of chemotherapeutic) and 4011 (biopsy of lymphatic structure) cannot be found in the prospective dataset. We experience the same issue with some procedures from the prospective datasets, such as 108290001 (radiotherapy) and 61938004 (breast reconstruction), which cannot be found in the retrospective dataset.

To face this challenge, we decided to apply the following methodology. We created a new dataset, which corresponds to the inner join of retrospective and prospective dataset features, only keeping the common features for these datasets. This allows us to quickly be able to compute the first trajectories for the prospective patients. Due to the small number of common features, some loss in accuracy was expected. However, the model is able to provide accurate results since pre-trained on the whole dataset. This approach makes it possible to lay the foundations of the system, which can then be easily updated in order to enhance the trajectory results. An additional advantage of this approach is that we can quickly provide the basic APIs to enable system integration within the CDSS. The system update then requires little to no change to be available.

3. Enriched data

A comprehensive list of the enriched data available, also called concepts, can be found in D5.3. Such information is extracted from non-structure data such as medical reports and surveys. For the cohort and trajectory analysis, we collect data such as metastasis, pain, and family history suggested by the clinician in the PERSIST consortium. Each feature requested includes a set of sub-features. For instance, metastasis includes pulmonary metastasis, bone metastasis, liver metastasis, spinal metastasis, and peritoneal metastasis.

Figure 8 illustrates the types of metastasis and the number of colon cancer patients involved. Among 1000 colon cancer patients, about 700 developed metastasis (mostly liver metastasis).

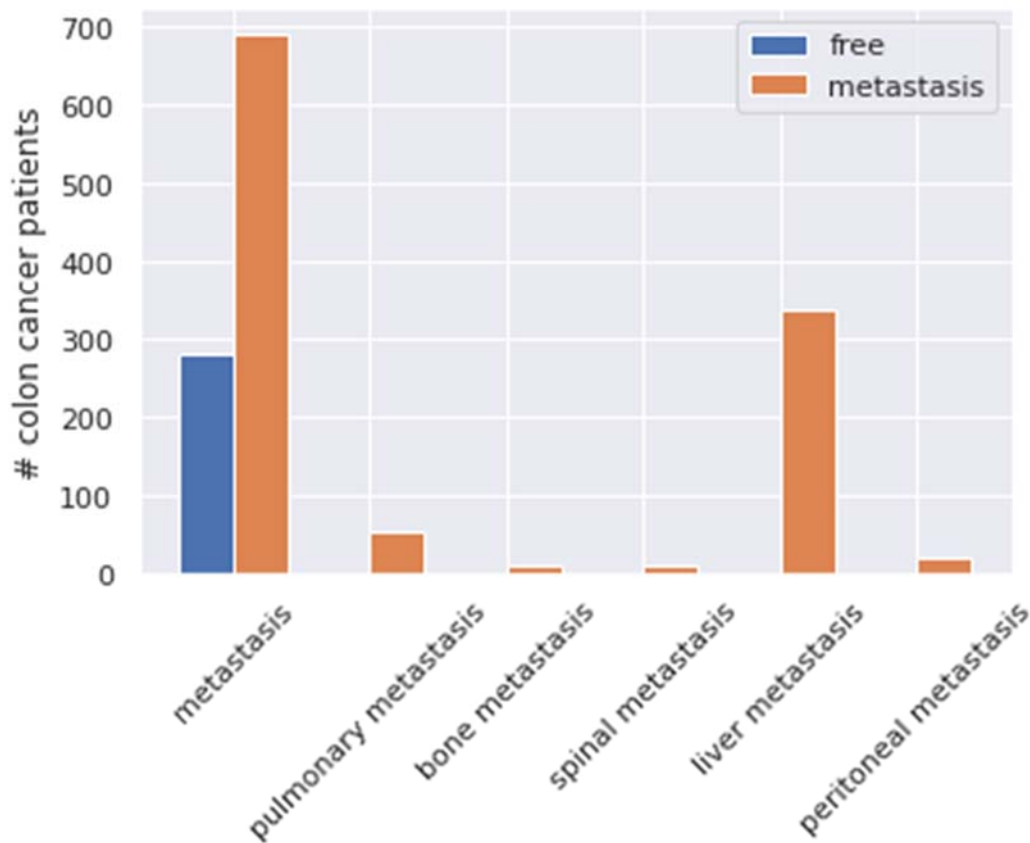


Figure 8 Types of metastasis in colon cancer population

To evaluate the impact of metastasis dissemination versus metastasis-free patients, we sample the colon cancer population and plot it in Figure 9. The figure shows a random sample of 100 patients (with metastasis in orange and metastasis-free in blue) and their respective life statuses (filled circle in case the patient is dead) over time. In Figure 9, patients with metastasis present a shorter life expectancy than metastasis-free patients.

However, other dimensions such as the patient's age, cancer stage, and treatments must be taken into account to evaluate Figure 9 results.

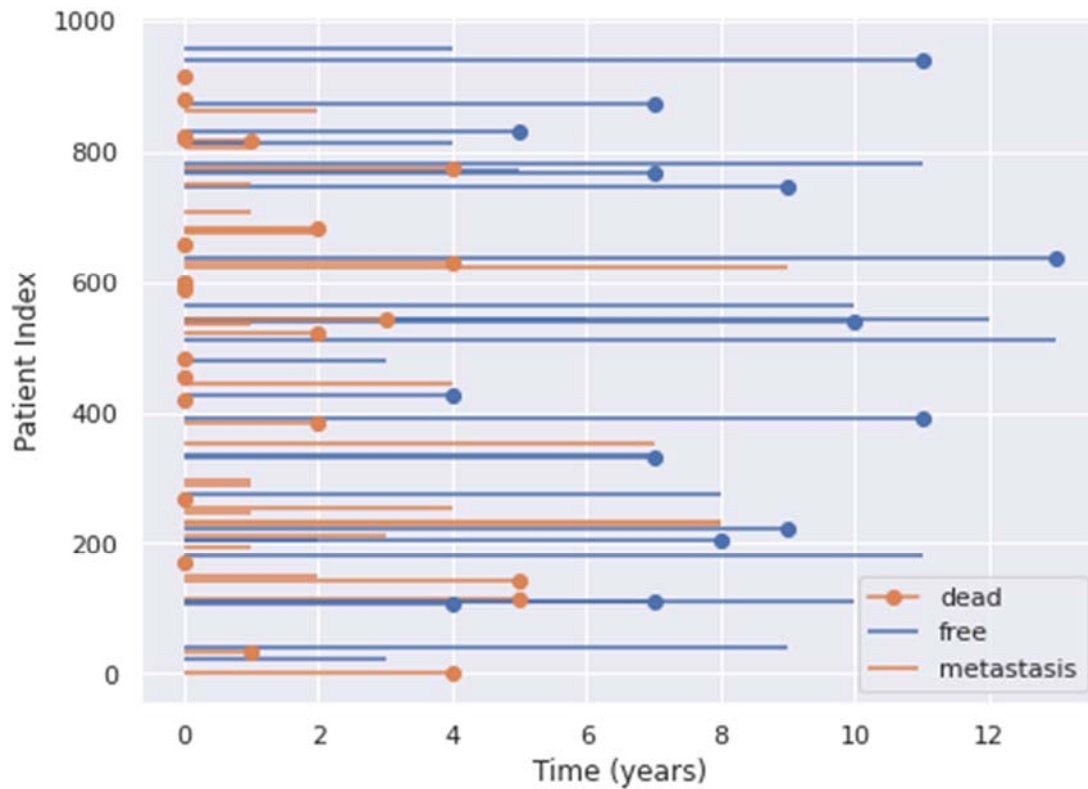


Figure 9 Colon cancer sample 10% and their respective life statuses over time (diagnosis date set as zero)

Patient Trajectory Models

Patient trajectories can be analyzed for different goals. In the context of cancer survivorship, one key aspect is the prediction of life expectancy, related to the probability of cancer relapse. To address this challenge, we introduce survival models [3][4], which aim to answer the question: what is the probability that a patient survives any time t ?

We denote the survival function as:

$$S(t) = Pr(T < t),$$

where T is the time of an event (e.g., death, relapse, or recovery), and t is the time from the beginning of an observation period (e.g., surgery or treatment) to an event. Please notice that $S(t) = 1$ when $t = 0$, whereas $S(t) = 0$ when $t = \inf$. In other words, with probability 1 the patient is alive at the beginning of the observation time t , and the probability tends to 0 when the observation time increases (i.e., $S(t_1) \leq S(t_2)$, for all $t_1 \geq t_2$). In case the study ends or the patient is withdrawn from it, the data is considered as “censored”. Given a dataset with patients observing time and event outcome, we are able to estimate the survival curve through the Kaplan-Meier Estimator[5]:

$$S(t) = \prod_{i: t_i \leq t} \left(1 - \frac{d_i}{n_i} \right)$$

With d and n the number of patients that had an event and the number of patients that survived at time i , respectively. Please note that the Kaplan-Meier estimator formula is obtained by using the chain rule for random variables. Indeed, the Kaplan-Meier estimator is calculated considering the notion that the probability can be broken up into the product of probabilities during specific intervals.

Below, we find the retrospective cancer patients from CHU trajectory analysis using Kaplan Meier estimator, grouped by treatments, tumor size (T), lymphatic nodes propagation (N), and metastasis presence (M) respectively. Such event trajectories identify associations between symptoms and events by cohort and sub-cohorts.

Concerning the treatments, we have chosen the most popular in the PERSIST population and from the third-party dataset previously analyzed (i.e., Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) and The Cancer Genome Atlas (TCGA) datasets).

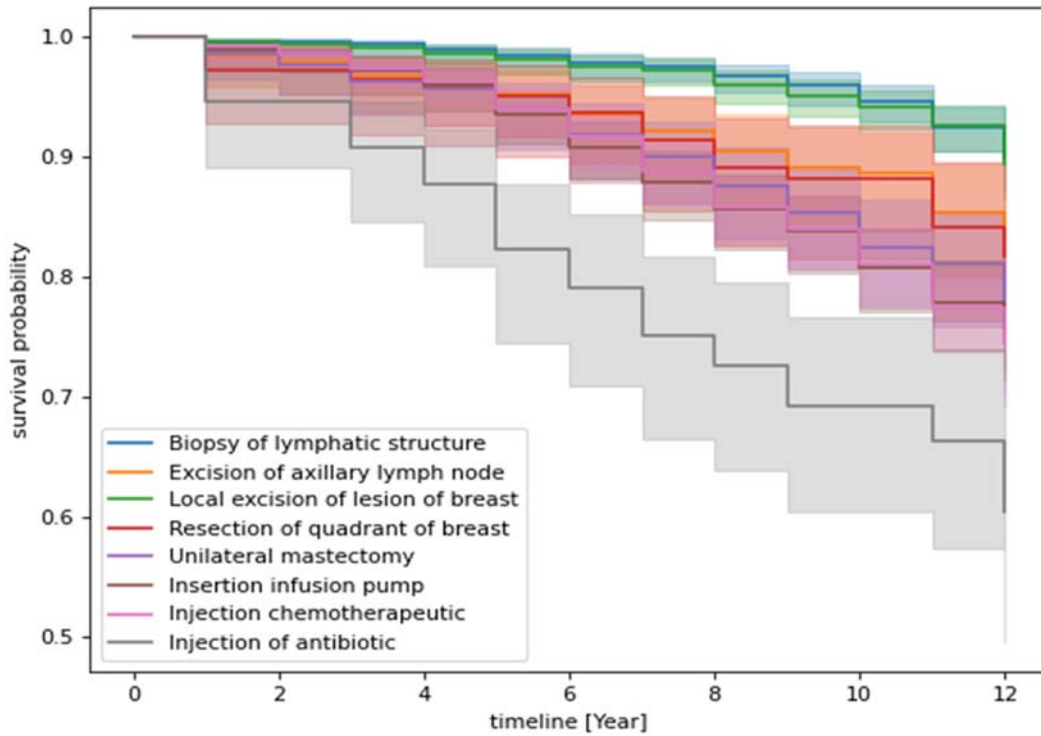


Figure 10 Kaplan-Meier breast cancer patient trajectory grouped by treatments

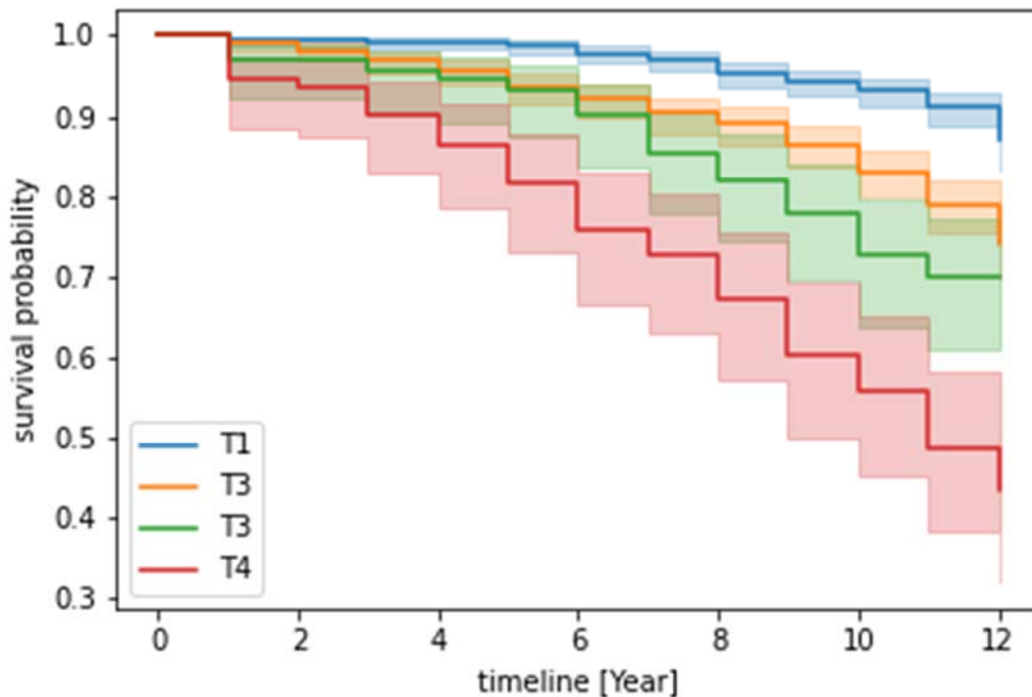


Figure 11 Kaplan-Meier breast cancer patient trajectory grouped by cancer stage T from TNM

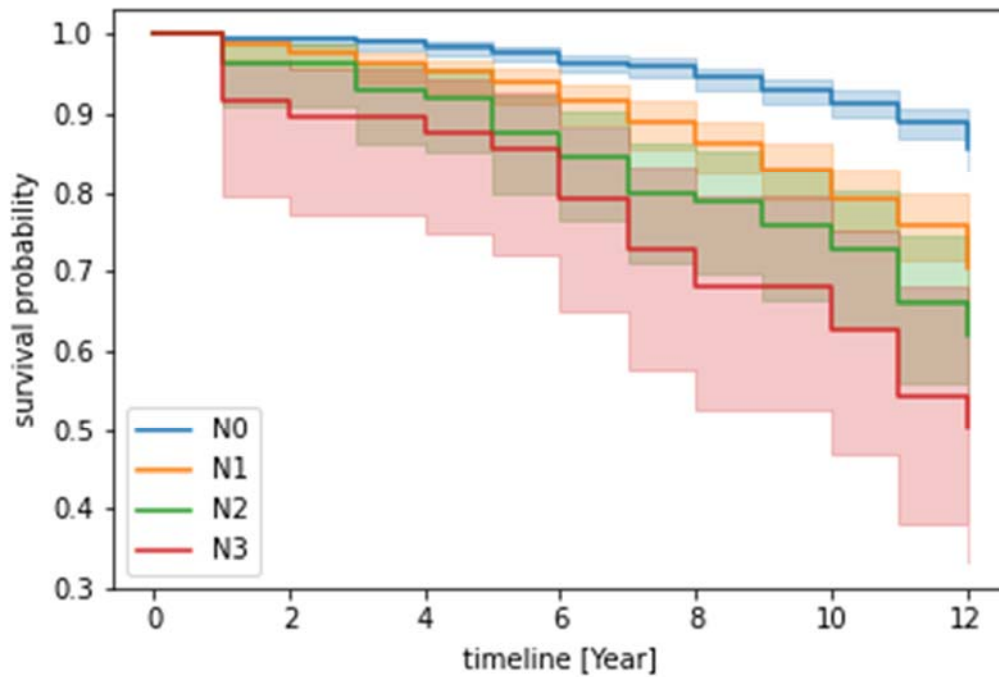


Figure 12 Kaplan-Meier breast cancer patient trajectory grouped by cancer stage N from TNM

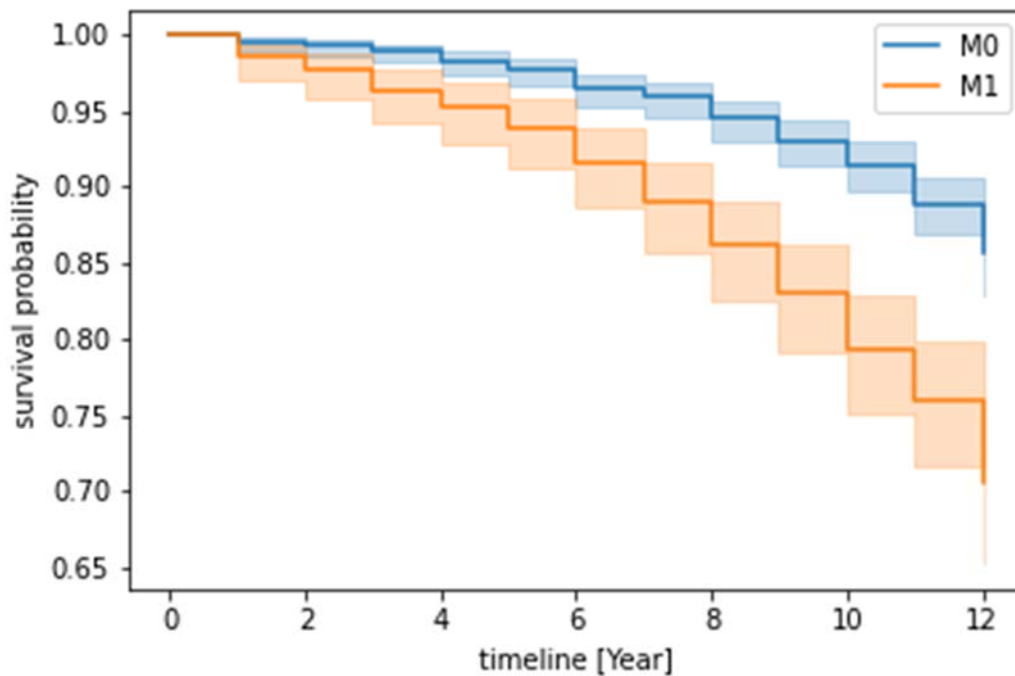


Figure 13 Kaplan-Meier breast cancer patient trajectory grouped by cancer stage M from TNM

Although this is useful to compare different survival groups and establish the basis for a prediction model, it does not indicate risk levels for individual trajectories. In order to provide personalized patient treatments, we need to evaluate the hazard function that

analyzes individual risks answering the question: What is the immediate death risk for a patient that survived at the time “t”?

The Cox Proportional Hazard or CPH model provides the tool to estimate individual risks as follows [6]:

$$\lambda(t) = \lambda_0 e^{\theta\omega}$$

where t is the observation period and “ λ_0 ” is the baseline risk. Whereas, the “ $\theta\omega$ ” identifies the way of modeling patient features (e.g., age, tumor stage, and treatments) to estimate patient risk. We defined the factor risk as a linear combination of the patient's features:

$$\omega = (\omega_1, \omega_2, \dots, \omega_n)$$

and the respective features' weights “ $\theta = (\theta_1, \theta_2, \dots, \theta_n)$ ”, with n the number of patient features.

Following, we present the trajectories of 10 random patients in the PERSIST retrospective breast cancer population from CHU.

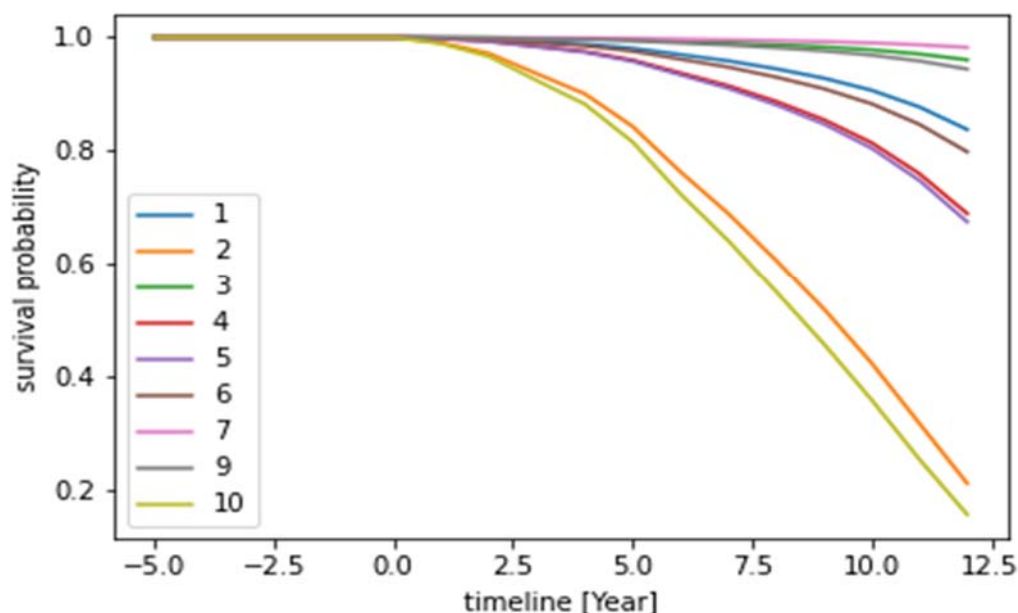


Figure 14 Kaplan-Meier breast cancer trajectory of ten random patients

Following, we find the feature importance of our breast cancer population. Values below zero positively affect patient survival probability. On the contrary, values above zero have a negative effect on the patient's survival probability. Please notice that the accuracy of such results is affected by the size of the dataset (e.g., data from other hospitals and prospective data).

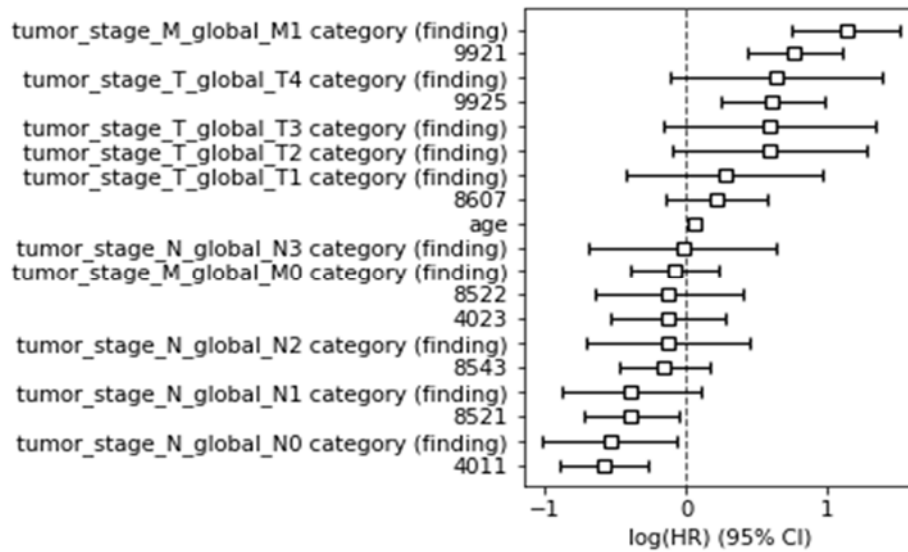


Figure 15 Cox Proportional Hazards for feature importance of breast cancer patients

Finally, to address the non-linearity relationship between the features, we use Tree-based risk models [7]. Besides the non-linear relationship, Tree-based models handle both continuous and categorical data in a time-efficient manner. In particular, after dealing with missing data, we adopt Decision Trees and Random Forests models [8], which provide high-performance levels and interpretations of their results. Figure 16 shows the survivor classification accuracy in terms of F-score over several AI-based models. Logistic Regression and Decision Tree outperform Neural Network and SVM due to the low sample population and feature dimensionality.



Figure 16 Survival classification accuracy Logistic Regression, SVM, Decision Tree, and Neural Networks

The rest of the section presents the cohort and trajectory analysis using prospective and enriched data from the PERSIST dataset.

1. Prospective Enriched Data Trajectories and Cohorts

The following section contains the trajectories and the respective risk level obtained from the analysis of the prospective and enriched data.

Figures 17 and 18 represent the trajectories of colon cancer and breast cancer patients, respectively. The color indicates the level of risk for each patient, which has been evaluated by applying a k-means clustering. As shown in the figures, the level of risk depends on the survival probability in a specific year. For instance, Figure 17 shows that the classification is strongly affected by small changes in the survival probability during the first years. In fact, a patient is considered at low risk if he has a survival probability of about 1 for year 3. For the same year, a patient is at moderate risk if his survival probability is around 0.95 and is already at high risk if his survival probability is about 0.9. In year 12, a patient is at low risk if his survival probability is over 0.8, he is at moderate risk for values between 0.3 and 0.8 and is at high risk for values below 0.3.

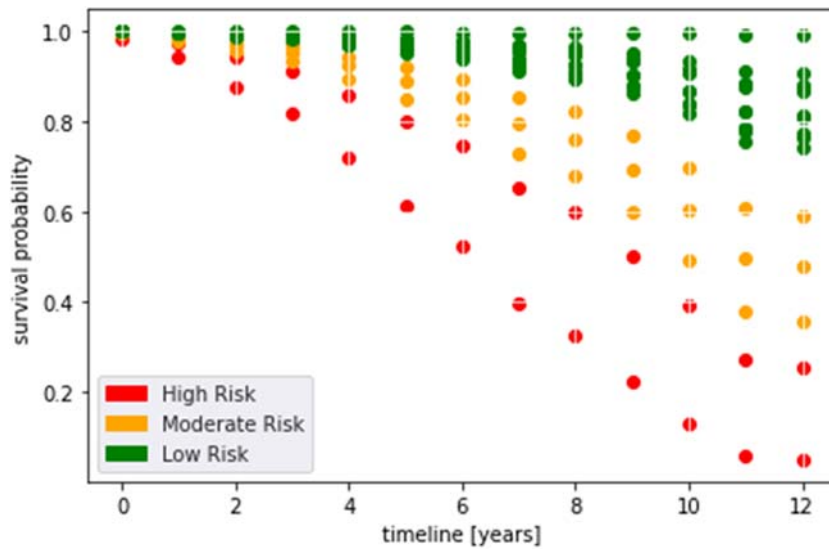


Figure 17 Risk level for patients in prospective data for colon cancer

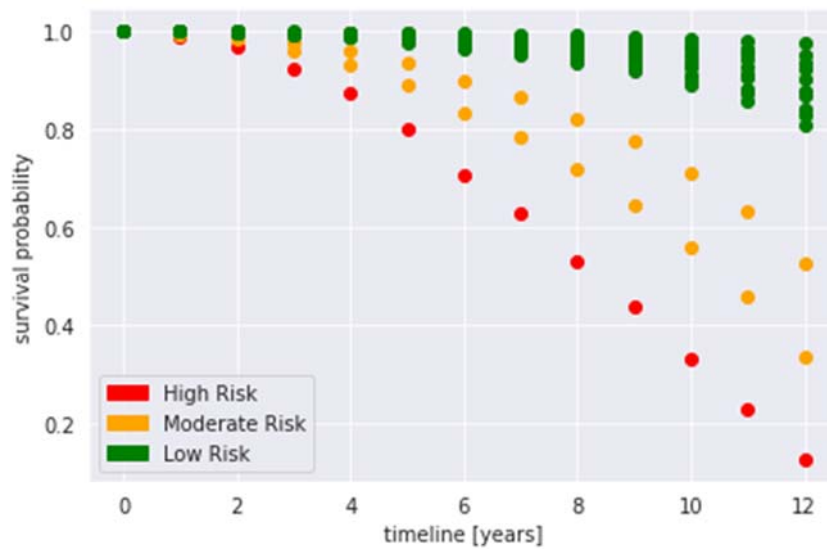


Figure 18 Risk level for patients in prospective data for breast cancer

Figures 19 and 20 show the survival probability calculated using Kaplan-Meier for colon and breast cancer grouped by the value M from the TNM staging system. As shown, the patients diagnosed with metastasis (M1) have the lowest survival probability.

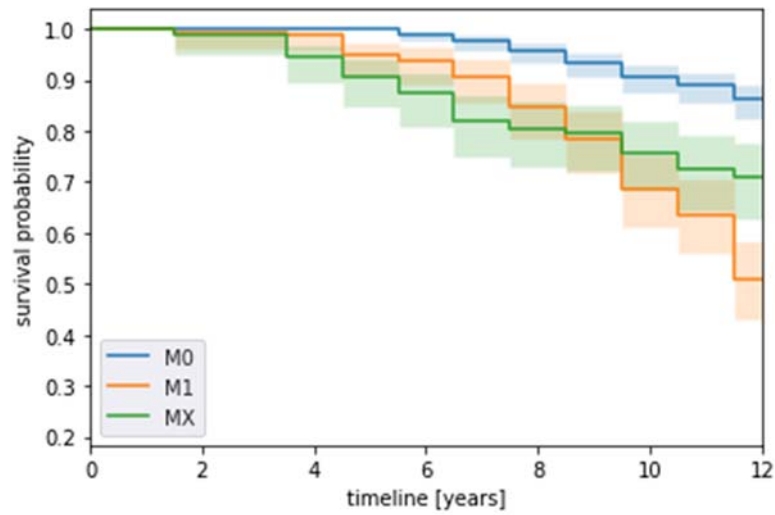


Figure 19 Kaplan-Meier colon cancer patient trajectory grouped by cancer stage M from TNM

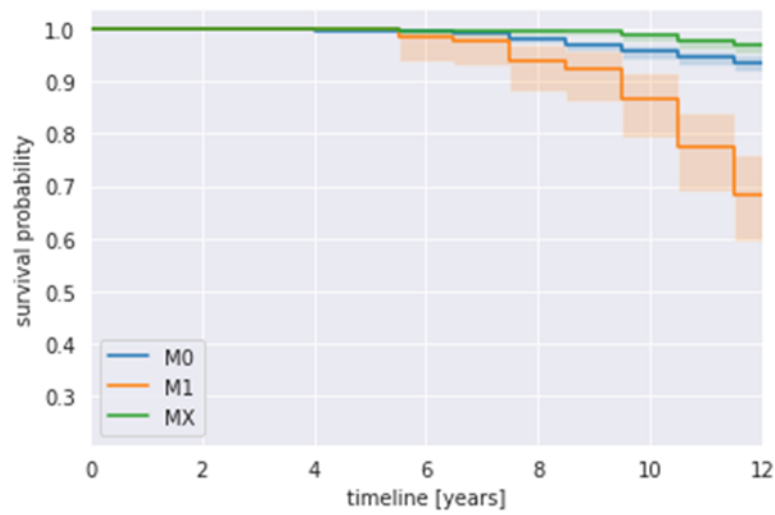


Figure 20 Kaplan-Meier breast cancer patient trajectory grouped by cancer stage M from TNM

Figure 21 illustrates the survival probability computed using the Cox Proportional Hazard model.

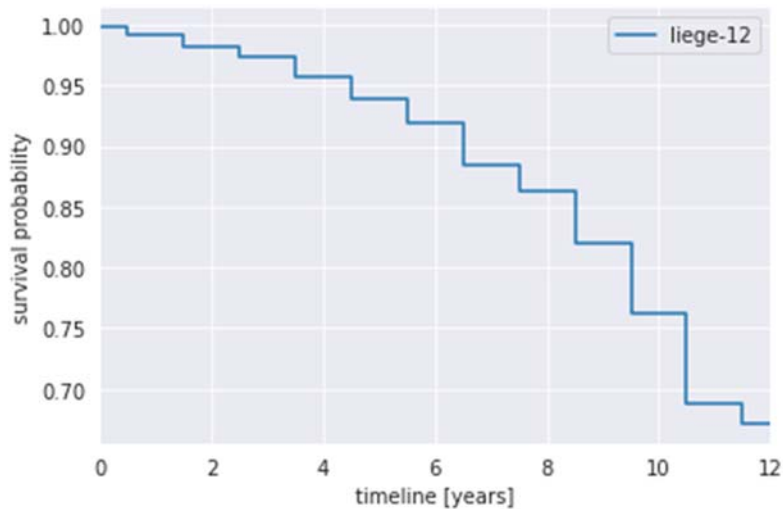


Figure 21 Cox Proportional Hazards survival probability for a colon cancer patient in prospective data

Figures 22 and 23 represent the feature importance for breast cancer patients, and colon cancer patients using the Cox Proportional Hazards model.

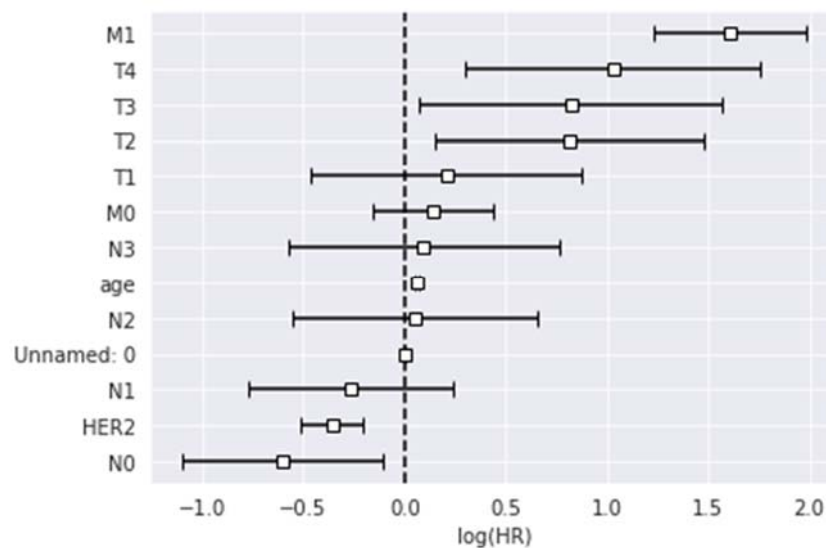


Figure 22 Cox Proportional Hazards for feature importance breast cancer patients

In the case of breast cancer, the T feature of TNM follows a logical order ($T4 > T3 > T2 > T1$). It means that a patient with breast cancer with T4 has a lower chance to survive than a patient diagnosed with T1. However, in the figure concerning colon cancer, we can see that the order is $T2 > T4 > T1 > T3$. This result can be explained by the lack of data about colon cancer or by other features.

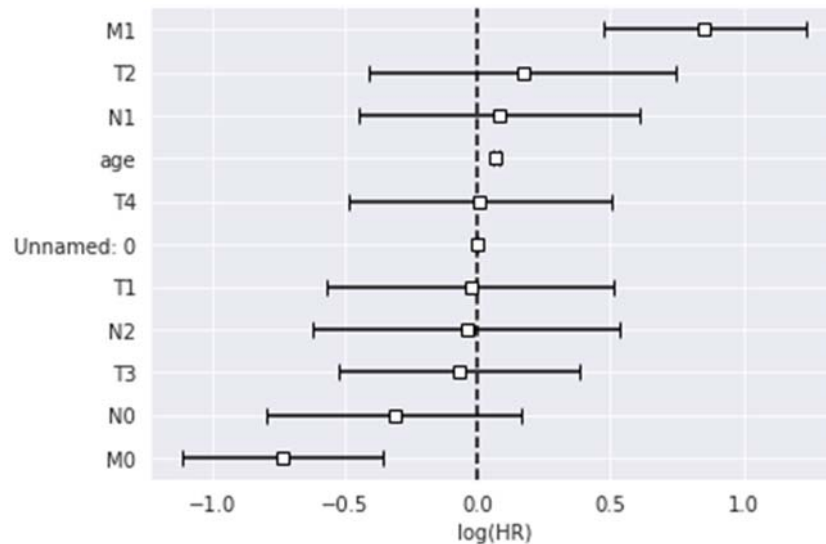


Figure 23 Cox Proportional Hazards for feature importance colon cancer patients

Finally, Figure 24 shows how the enriched data such as the Eastern Cooperative Oncology Group status (ECOG), the Body Mass Index (BMI), and the Human Epidermal Growth Factor Receptor-2 (HER2) enhances the accuracy of Cox Proportional Hazard model results.

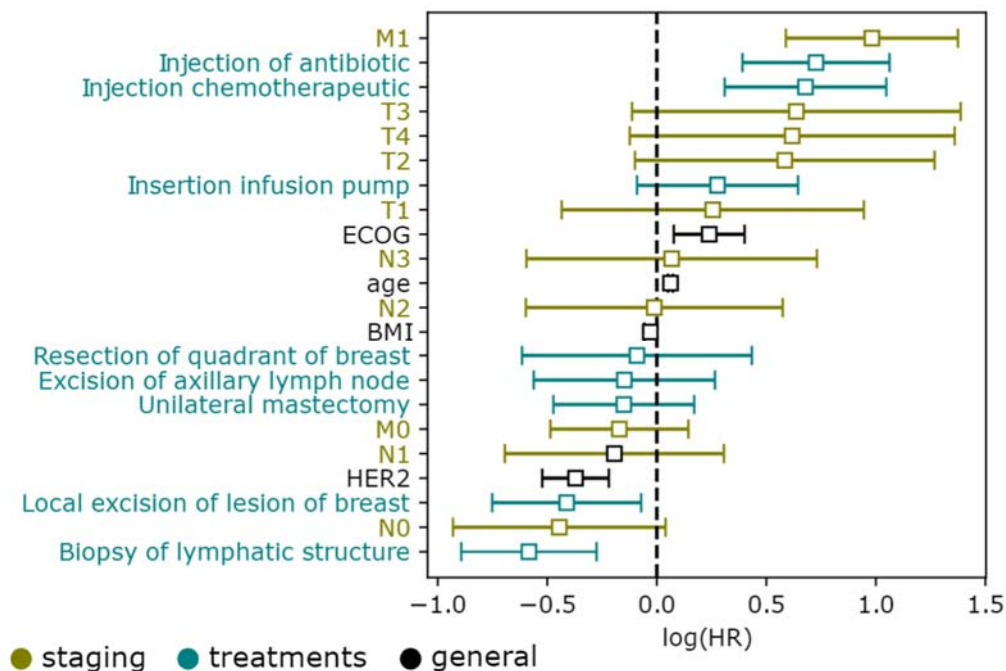


Figure 24 Cox Proportional Hazards for feature importance breast cancer patients enriched dataset

2. Relapse risk prediction

Although cancer may be in remission thanks to early detection and improvements in treatment, some patients may experience a relapse of cancer. Recurrence is a fundamental clinical manifestation and is one of the main causes of death related to cancer, so its early detection is of great help in improving the prognosis of patients.

In recent years, many researchers have tried to find a particular pattern that predicts cancer recurrence. For example, in breast cancer, when characterizing the presence of certain receptors (such as ER, PR, or HER2), each subtype may carry a higher risk of recurrence than others in a specific situation or over certain years. However, these patterns require considerable cost and are time-consuming.

For this reason, at PERSIST, we propose the development of a non-invasive computational system to predict the risk of relapse of breast and colon cancer based on the clinical and treatment information of the patients available in the electronic medical record (EHR). The increasing availability of access to EHR data offers a timely and low-cost alternative to traditional cohort studies, with the potential to include large and wide real-world populations. However, the heterogeneity and diversity of EHR data also introduce difficulties in obtaining quality data for cancer-related research. Our main objective is to take advantage of the data that is routinely collected in the EHR to build models that allow us to predict the risk of suffering a relapse after the first diagnosis of breast or colon cancer. To this end, we have combined structured and semi-structured data from the EHR to extract clinical variables and compose two datasets (one for breast and the other for colon cancer) that we have used to build the models. We have trained several ML models and compared their performance in predicting the 5-year probability of recurrence for breast cancer and colon cancer. These models can be a great aid for decision-making for clinical professionals, who can take the prediction results as a reference to optimize the follow-up and treatment provided to patients.

In addition to the models, the relapse risk prediction system developed in PERSIST has a web service that runs the models and provides the desired predictions. This web service is easily consumable, allowing easy integration of the relapse risk prediction system with the other components of the overall PERSIST system.

The following is a description of the relapse risk prediction system divided into the two points mentioned above: the models and the service that consumes them.

2.1. Relapse risk prediction algorithm

This section describes the work done to develop the recurrence risk prediction algorithm. It includes the description of the dataset and its pre-processing, the different models evaluated, and the results obtained.

Dataset

To train recurrence risk prediction models, we have created two datasets, one for colon cancer and the other for breast cancer.

The initial cohort contains a total of 4,282 patients who have been diagnosed with colon or breast cancer. These patient records have been extracted from the CHU medical record, mapped to the FHIR format defined in PERSIST, and located on the project platform's FHIR server (OHC). The data in these registries is heterogeneous and contains the patient's clinical history, including diagnoses, laboratory values, medical tests, procedures, medications, and clinical reports.

The distribution of FHIR resources generated for this cohort is shown in Figure 25.

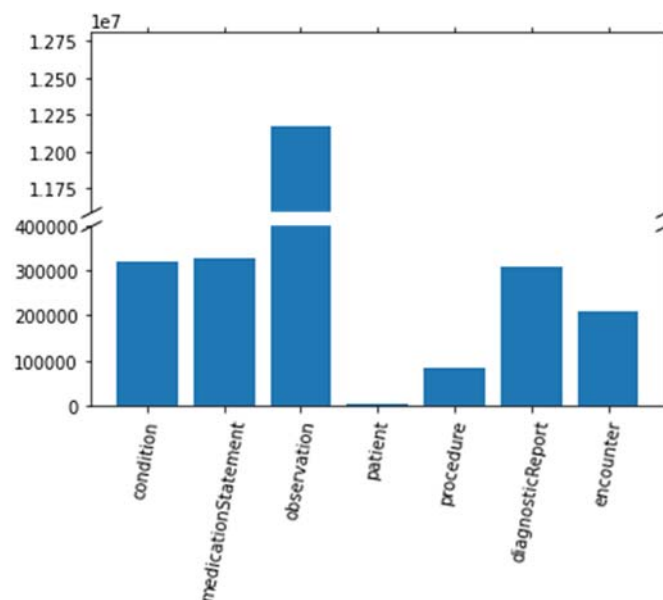


Figure 25 FHIR resource distribution of data from CHU

The first step has been to convert the FHIR data into a more easily computable format for ML algorithms. The AI-ready adapter tool available in the OHC has been used. In addition, several mappers have been developed that have made it possible to convert data from FHIR resources to .csv files.

Data Preprocessing for relapse risk

Once the data has been exported to csv format, they have been subjected to a data cleaning and transformation process to obtain the final datasets that will be used to train the algorithms. This process is described below:

Get a single row per patient. Originally, the data of a patient is distributed along several rows and spread in several csv files. The first step is to analyze all the information in the patient record, select the variables and values that will be retained to make the prediction, and generate a file with a single row per patient. The

selected variables are based on an analysis of the state of the art and the consultation with the clinical experts participating in the project. They are summarized below:

- Primary site of the tumor
- Comorbidities (other diagnoses of the patient)
- Gender
- Age at diagnosis of cancer
- TNM staging (clinical and pathological)
- Weight
- Height
- BMI
- Social habits of the patient: smoke, drink, drug consumption
- Performance status (using ECOG scale)
- Tumor grade
- Tumor morphology
- Lab values and tumor markers: ER, PR, HER2, Ki67, FSH, LH, estradiol, CEA, C19, C15
- Treatments: surgery, radiotherapy, chemotherapy

Inclusion criteria. Inclusion criteria are then applied to eliminate records that do not contain valid information or sufficient information to develop the 5-year recurrence prediction algorithm. These criteria include:

- a. Patients must have at least one diagnosis of colon cancer and/or breast cancer.
- b. The patient record must contain information about the TNM staging.
- c. The patient record must contain information on at least one type of treatment (surgery, radiotherapy, and/or chemotherapy)
- d. Patients who do not present recurrence must have survived at least 5 years after diagnosis.

Patient records that do not meet any of these criteria are excluded.

Missing values. Once the patients have been filtered, exploratory analysis is performed to examine the completeness of the dataset. In Figures 26 and 27, we see the distribution of missing values for each set of patients (breast and colon).

- a. Combine data into a single variable, for example, calculate BMI from weight and height.
- b. Group the values of a variable into broader categories, in order to avoid values with low representation. For example, diagnostic codes associated with comorbidities have been grouped according to Elixhauser categories¹.
- c. Prepare data for ingestion by ML algorithms. The categorical variables have been converted into dummies (a variable that takes only the value 0 or 1 to indicate the absence or presence) and the ordinal ones into integer numerical values.

As output after this preprocessing, we obtain two datasets, one for breast cancer and one for colon cancer, with 823 and 333 samples, respectively.

Recurrence prediction models

The datasets obtained in the previous section have been used to build ML models for the prediction of 5-year recurrence of cancer. To this end, several state-of-the-art algorithms have been evaluated and compared, in particular: Logistic Regression (LR), Decision Tree (DT), Gradient Boost (GB), eXtreme Gradient Boost (XGB), and Deep Neural Networks (DNN).

First, we have performed a hyperparameter tuning for each of the algorithms using the grid-search technique. For this, we have applied the cross-validation method with repetition, and we have reserved 10% of the dataset for the final validation (a set of data that has not been used during the training of the models or for the adjustment of the hyperparameters).

In each pass of the cross-validation, we have applied the following pipeline:

- Oversampling. The datasets obtained are highly unbalanced (87%/13% for breast cancer and 77%/23% for colon cancer). In general, such a large imbalance tends to have a negative effect on the performance of the classification algorithms. In order to alleviate this problem, the SMOTE² technique has been applied to oversample the training dataset and ensure that the classes are balanced. SMOTE allows new synthetic data to be generated from existing data using the k-NN algorithm.
- Scaling. All values have been normalized to the range [0,1] prior to model training.
- Classifier training.

¹ Metcalfe, David, et al. "Coding algorithms for defining Charlson and Elixhauser co-morbidities in Read-coded databases." *BMC medical research methodology* 19.1 (2019): 1-9.

² Chawla, Nitesh V., et al. "SMOTE: synthetic minority over-sampling technique." *Journal of artificial intelligence research* 16 (2002): 321-357.

The selection of the best estimator has been made based on the F1 metric, since it allows obtaining a good balance between precision and recall. The final performance of the models has been evaluated in terms of the following metrics: Precision, Recall, F1, and AUC - ROC (Area Under the Receiver Operating Characteristic (ROC) Curve).

The results obtained for the best estimator of each of the models evaluated for breast cancer and colon cancer, respectively, are shown in Tables 15 and 16.

Classifier	Precision	Recall	F1	AUC - ROC
LR	0.86	0.80	0.82	0.72
DT	0.87	0.86	0.86	0.70
GB	0.91	0.90	0.91	0.80
XGB	0.92	0.93	0.92	0.84
DNN	0.89	0.86	0.87	0.80

Table 15 Results obtained for the breast cancer recurrence prediction models

Classifier	Precision	Recall	F1	AUC - ROC
LR	0.86	0.59	0.66	0.57
DT	0.76	0.71	0.73	0.53
GB	0.84	0.74	0.77	0.53
XGB	0.83	0.68	0.73	0.43
DNN	0.82	0.65	0.71	0.68

Table 16 Results obtained for the colon cancer recurrence prediction models.

The most notable aspect that we can observe is that the general performance of the models built for colon cancer is quite lower. This is because the size of the dataset is much smaller, so we can assume that by increasing the number of samples we could obtain significant improvements in the models.

In the case of breast cancer recurrence prediction, the model that offers the best performance on all metrics is XGB, with 92% accuracy, 93% recall, 92% F1 and 84% roc_auc. This is not surprising, as XGB is one of the most widely used ML algorithms nowadays, having demonstrated state-of-the-art results in a wide variety of ML benchmarks [9][10]. It is characterized by being an algorithm that allows obtaining good prediction results with relatively little effort, particularly for problems with heterogeneous data.

In the case of the prediction of recurrence for colon cancer, the selection of the model with the best performance is not so evident. GB offers better recall and F1 values, with 74% and 77% respectively. The LR model is the clear winner in terms of accuracy, but this is at the cost of lowering the recall. DNN, for its part, provides a significantly higher ROC_AUC value than the rest. Considering all the metrics in general, we have selected the GB model as the most appropriate to use in the context of PERSIST, since it is the one that improves the F1, and presents a better compromise between precision and recall.

Relapse risk prediction service API

In Figure 28, we present the relapse risk prediction service alongside the modules with whom it communicates.

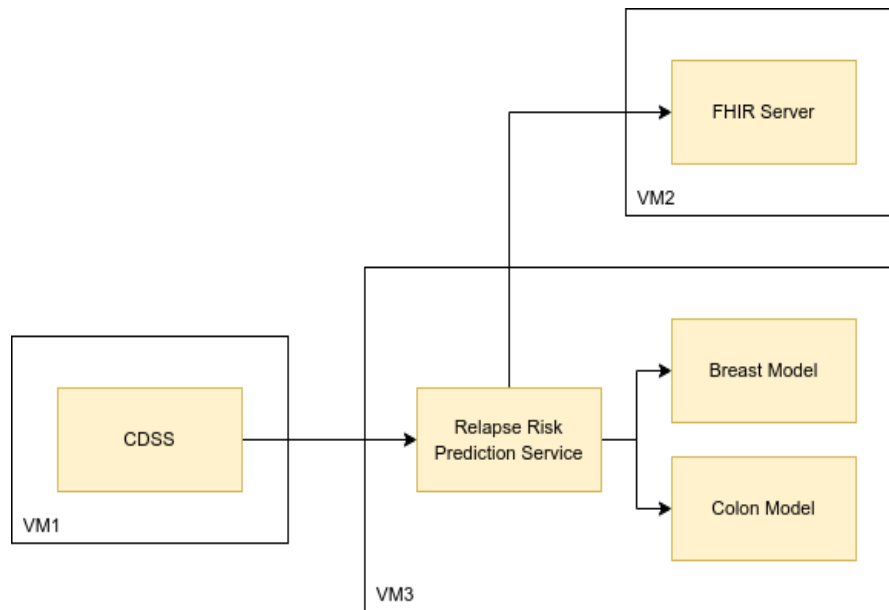


Figure 28 Components that communicate with the relapse risk prediction service

The service is called by the CDSS. With the objective of providing this information to the final user. The service itself is a web service written in the python programming language using the FastAPI framework. It orchestrates the process that obtains the predictions for a particular patient.

The FHIR server hosts information about clinical trials, pathologies, lab test results and such, which is standardized according to the FHIR standard. This infrastructure is managed by Dedalus and is used by the service to retrieve the data from the patient from which it will extract the variables needed by the models.

The service hosts a trained model for colon cancer and another model for breast cancer. These models are used by the prediction service to calculate the probability of recurrence for patients.

The diagram shows that the service and the models are deployed on an individual virtual machine as part of the distributed PERSIST ecosystem. Incoming communications are limited by IP address as a way of securing the application.

Now, the process by which a recurrence prediction for a patient is elaborated is as follows:

1. The CDSS requests a prediction for a patient to the relapse risk prediction service. It specifies whether this prediction is for colorectal or breast cancer.
2. The service parses the patient identifier and the requested cancer type.
3. Depending on the cancer type, the service makes several calls to the FHIR server to retrieve patient data.

4. The service then parses this data to extract the variables needed for the models to work (either the colorectal or the breast cancer model). If the patient doesn't have any of the required variables, the service will return a warning.
5. Here, the service calls the required model passing all the extracted variables.
6. The model will then process the variables and produce a prediction.
7. Finally, the service will return the prediction given by the model.

The diagram shown in Figure 29 shows these interactions.

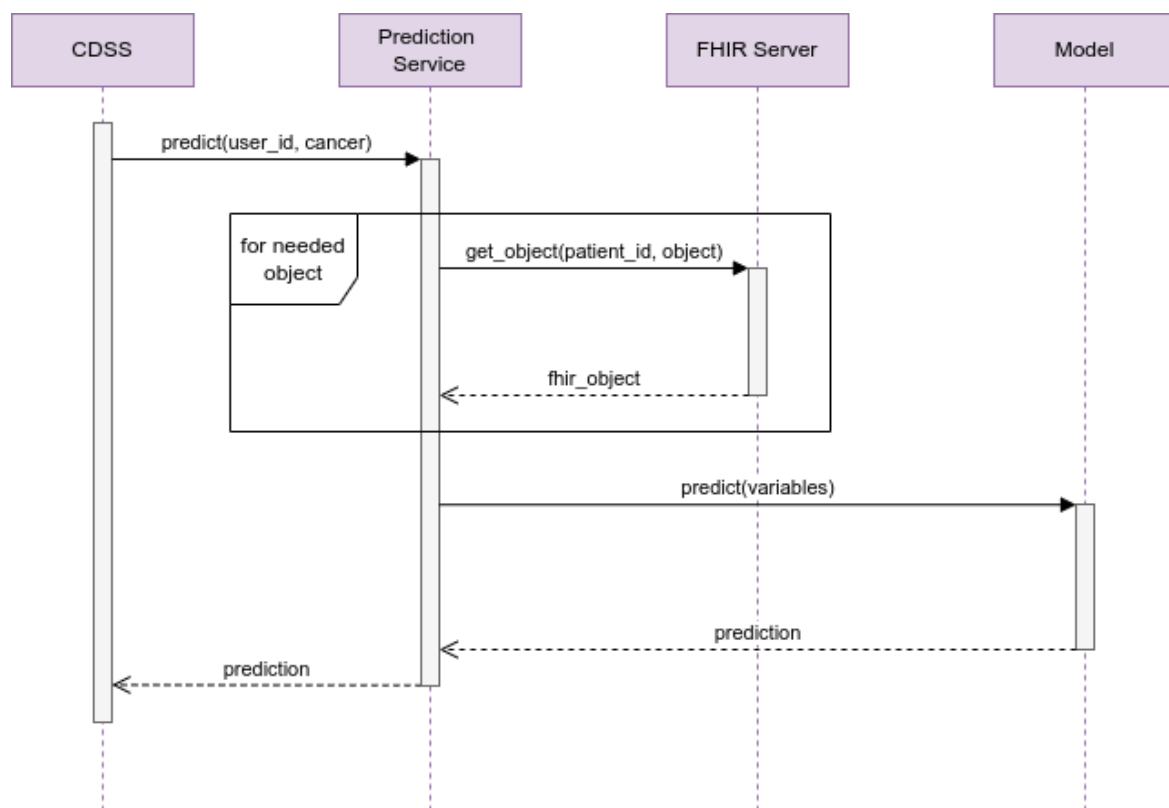


Figure 29 Relapse risk prediction generation process

A prediction request must be accompanied by a JSON body, which must include the following fields:

- "event": can have one of two values:
 - ✓ "RECURRENCE_EVENT_PROBABILITY_CC": for colon cancer recurrence probability requests.
 - ✓ "RECURRENCE_EVENT_PROBABILITY_BC": for breast cancer recurrence probability requests.

- “timeline”: must be "RECURRENCE_EVENT_TIME".
- “patient_id”: the desired patient ID.

The service will respond with a JSON object containing the requested recurrence probability.

The structure of this response JSON mirrors the one returned by the trajectories prediction service developed by HESSO. This was done to ease the integration of this service with the rest of the PERSIST distributed ecosystem.

Example of service response:

```

{
  "meta": {
    "version": "0.0.1"
  },
  "plot": {
    "functions": [{
      "label": "RECURRENCE_PROBABILITY",
      "data": [{
        "timeline": 0.0,
        "value": 0.2752279424371614,
        "confidence_interval": {
          "upper": 0.2752279424371614,
          "lower": 0.2752279424371614
        }
      }
    ]
  },
  "x-axis": {
    "label": "RECURRENCE_EVENT_TIME",
    "min": 0.0,
    "max": 0.0
  },
  "y-axis": {
    "label": "RECURRENCE_EVENT_PROBABILITY_CC",
    "min": 0.2752279424371614,
    "max": 0.2752279424371614
  }
}

```

Data retrieval from FHIR servers

The recurrence prediction for a patient is made using the same variables mentioned during training. These variables are calculated from information obtained from different resources recovered from the FHIR servers.

The EHR servers host information about clinical trials, pathologies, lab test results and such, and are standardized according to the FHIR standard. This infrastructure is managed by Dedalus and is divided into two components: one with development data, and another one with production data. The relevant variables are retrieved by accessing the “persist” tenant, which contains the prospective data. For a solicited patient, the service obtains the resources of type Observation according to the relevant pathology, and all the patient’s resources of type Condition and Procedure.

For each patient recurrence request, a collection of patient's resources is obtained from FHIR servers. Resources of type Patient, Condition, and Observation are obtained.

Variable	Resource code field	Resource display field	Code
Gender	Patient.id	Patient.gender	Patient ID
Weight	Observation.code, Observation.patient	Observation. valueQuantity. value	29463-7
Height	Observation.code, Observation.patient	Observation. valueQuantity. value	8302-2
BMI	Observation.code, Observation.patient	Observation. valueQuantity. value	39156-5
Smoking_behaviour	Observation.code, Observation.patient	Observation. valuableCodeableConcept .coding.code	365981007
Drinking_behaviour	Observation.code, Observation.patient	Observation. valuableCodeableConcept .coding.code	228273003
Drug_misuse_behaviour	Observation.code, Observation.patient	Observation. valuableCodeableConcept .coding.code	228366006
Ecog_performance_status	Observation.code, Observation.patient	Observation. valuableCodeableConcept .coding.code	423740007
Histologic_grade	Observation.code, Observation.patient	Observation. valuableCodeableConcept .coding.code	371469007
Ki67	Observation.code, Observation.patient	Observation .valueQuantity .value	74489-6
CA_15_3	Observation.code, Observation.patient	Observation .valueQuantity .value	6875-9
CA_19_9	Observation.code, Observation.patient	Observation .valueQuantity .value	24108-3
CEA	Observation.code, Observation.patient	Observation .valueQuantity .value	2039-6
Estrogen_receptor	Observation.code, Observation.patient	Observation .valueQuantity .value	16112-5
Progesterone_receptor	Observation.code, Observation.patient	Observation .valueQuantity .value	16113-3
HER2_by_IHC	Observation.code, Observation.patient	Observation. valuableCodeableConcept .coding.code	85319-2
HER2_by_FISH	Observation.code, Observation.patient	Observation. valuableCodeableConcept .coding.code	31150-6

Table 17 Recovered variables from Observation Resources

Variable	Resource code field	Resource display field	Code
Body_site_codes	Condition. bodySite.coding.code	Condition. bodySite.coding.code	All bodySite objects inside all the retrieved conditions

Oldest_cancer_on_set_datetime	Condition.resource.onsetDateTime	Condition.resource.onsetDateTime	Selected among the retrieved conditions
Newest_cancer_on_set_datetime	Condition.resource.onsetDateTime	Condition.resource.onsetDateTime	Selected among the retrieved conditions
Morphology_code	Condition.Evidence.code.coding.code	Condition.Evidence.code.coding.code	371441004
Number_of_surgeries	Procedures.code.coding.code	Procedures.code.coding.code	List of codes for colon: "82035006", "33507007", "174171002", "771568007", "425851003", "23968004", "738552004", "16564004", "448143009", "265414003", "387713003", List of codes for breast cancer: "392021009", "69031006", "70183006", "428564008", "33496007", "443611007", "172061002", "303445008", "61938004", "396487001", "234254000", "234262008", "387713003"
Number_of_radiotherapies	Procedures.code.coding.code	Procedures.code.coding.code	108290001
Number_of_chemotherapies	Procedures.code.coding.code	Procedures.code.coding.code	367336001

Table 18 Recovered variables from Condition Resources with patient ID and requested type of cancer

The T, N, and M categories are obtained by recovering the Observation resources contained in the field “evidence” from the patient’s Condition resources. The algorithm will recover the most recent clinical, pathological and unspecified TNM from the FHIR server, and will later discern which one will be served to the ML model.

Variable	Resource code field	Resource display field	Code
Clinical_TNM_codes	Observation.code.coding.code	Observation.valuableCodeableConcept.coding.code	106248000
Pathologic_TNM_codes	Observation.code.coding.code	Observation.valuableCodeableConcept.coding.code	106249008
Unspecified_TNM_codes	Observation.code.coding.code	Observation.valuableCodeableConcept.coding.code	399566009

Table 19 Recovered variables from Clinical and Pathological cancer staging system TNM

Finally, a set of variables called "comorbidities" is retrieved, which depend on the requested cancer type. These variables are stored in the patient's Condition resources.

Variable	Resource code field	Resource display field	Code
Cardiac_arrhythmia	Condition.code.coding.code	Condition.code.coding.code	386089
Valvular_disease	Condition.code.coding.code	Condition.code.coding.code	415855, 390073 or 390075
Hypertension_uncomplicated	Condition.code.coding.display	Condition.code.coding.display	386155
Chronic_pulmonary_disease	Condition.code.coding.code	Condition.code.coding.code	391995
Diabetes_uncomplicated	Condition.code.coding.display	Condition.code.coding.display	"Diabetes"
Hypothyroidism	Condition.code.coding.code	Condition.code.coding.code	390182
Renal_failure	Condition.code.coding.display	Condition.code.coding.display	"Renal failure"
Lymphoma	Condition.code.coding.display	Condition.code.coding.display	"Lymphoma"
Obesity	Condition.code.coding.code	Condition.code.coding.code	388028
Weight_loss	Condition.code.coding.code	Condition.code.coding.code	385688
Fluid_and_electrolyte_disorders	Condition.code.coding.code	Condition.code.coding.code	386696
Deficiency_anemia	Condition.code.coding.code	Condition.code.coding.code	386560, 388686 or 388057

Table 20 Recovered variables from Colon cancer comorbidities

Variable	Resource code field	Resource display field	Code
Hypertension	Condition.code.coding.code	Condition.code.coding.code	386155
Chronic_pulmonary_disease	Condition.code.coding.code	Condition.code.coding.code	391995
Diabetes	Condition.code.coding.display	Condition.code.coding.display	"Diabetes"
Hypothyroidism	Condition.code.coding.code	Condition.code.coding.code	390182
Obesity	Condition.code.coding.code	Condition.code.coding.code	388028

Table 21 Recovered variables from Breast cancer comorbidities

Trajectory Analysis API

The API is composed of different endpoints. Each endpoint is in charge of providing information related to a certain type of analysis. For instance, one endpoint is used to compute the influence of the features on the survival probability, while another one is responsible for providing the survival probability. Figure 30 shows the swagger documentation for the list of endpoints available in the first version of the API.

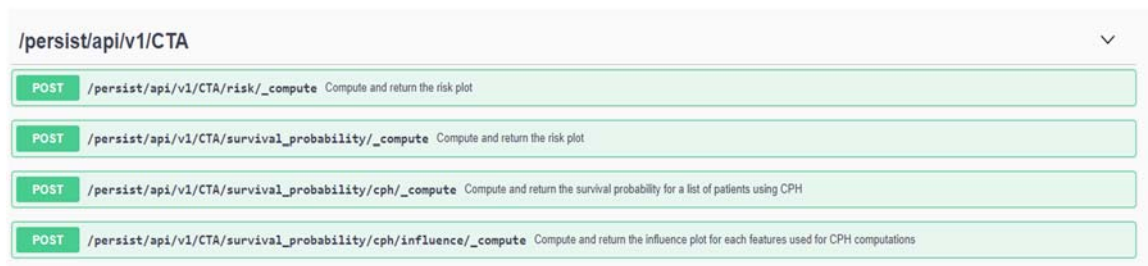


Figure 30 First version of trajectory analysis API documented with swagger

Although each endpoint has a different task, we facilitate the integration between each endpoint. Therefore, we decided to POST a request with a JSON body to define the different input parameters. Figure 31 shows the documentation of the list of common input parameters that can be used with each endpoint.



Figure 31 List of common input parameters of the API documented with swagger

This solution can be easily extended by adding a specific attribute related to a specific endpoint.

The API responds with a JSON providing all the information necessary to compute the graph resulting from the processing of the input data. This way, the integrator can create a

graph displaying exactly the desired information. This solution implies a small amount of data transfer and increases the possibility of customization on the client side. Another possibility, it was to respond directly with images. However, images are heavier than JSON text and are less customizable. For example, it's difficult for the integrator to implement filters with images.

To facilitate API integration with CDSS, we decided to deliver the API in the form of a Docker container. This container includes all the dependencies and datasets required to run the API. From a security perspective, the client needs a token to query the API. Since some steps, such as data cleaning, are highly time-consuming, we decided to split the API delivery into three different packages. With this distributed approach, we are enabled to test the integration of the API during the development and not only at the end.

The first version contained all the endpoints, but the analysis was performed only on the retrospective data for breast cancer in PERSIST. The second package, based on the first one, added the support of retrospective and prospective data for both colon and breast cancer. Finally, the third version delivered added the support to enriched data coming from task T5.2.

As each version of the API is an enhancement of the previous one, only the latest version is available inside the CDSS.

Conclusions

We presented deliverable 5.8 based on the continuation of D5.4. Deliverable 5.8 includes the pipeline for the cohort and trajectory analysis of PERSIST patients, the risk analysis and module, and the API integration. Among the main points, we introduced the FHIR data structure and the respective OHC server to store the data. Given the heterogeneity of the data and coding systems, we started working on a small set of data from CHU. We illustrated the data pre-processing, which includes data retrieval, data cleaning, and feature preparations. We selected several AI-based models, such as Kaplan-Meier, to estimate the survival probability of breast and colon cancer patients grouped by treatments and cancer stages. Moreover, we highlighted the most relevant features using Cox Proportional Hazard using prospective and enriched data from the PERSIST dataset. In order to provide high-performance levels and interpretations of their results, we adopted Decision Trees and Random Forests models. We add the risk of cancer relapse, describing the dataset, pre-processing, models, and metrics adopted for the analysis.

Among the subtasks included in this deliverable, data retrieval is certainly one of the most challenging. Indeed, data from different hospitals are coded with different coding systems, such as ICD and SNOMED. Moreover, the same hospital can use different versions of the coding system (e.g., ICD-9 and ICD-10), making data retrieval and harmonization time-consuming and complex. This issue has been also detected for the case of colon cancer patients' data retrieval. Nevertheless, a parallel task (T5.2 - with related deliverable D5.3) provided data harmonization to create a universal ontology for diagnosis and procedures in the PERSIST datasets. Therefore, the issue was addressed for enriched and prospective data. Particularly, the enriched data enable us to include into the trajectory and cohort analysis unstructured data, such as family history and cancer occurrences based on the clinical location (e.g., French for CHU and Latvian for UL).

Furthermore, multiple-level cohort analysis through miscellaneous approaches (supervised and unsupervised) using support vector machines, regression analysis, and neural networks have been included to enhance model accuracy.

We identified high-risk markers for detrimental treatment effects, subsequent cancer disease, and metastatic cancer disease. However, given the lack of cancer behavioral data, the trajectory and cohort analysis to prevent anxiety, depression, and other psychological issues remain an open challenge.

Finally, we implemented and documented the API integrated into the CDSS as support to T5.4.

References

1. Brownlee, Jason. "Why One-Hot Encode Data in Machine Learning?". Machinelearningmastery, 2017.
2. Brownlee, Jason. "Ordinal and One-Hot Encodings for Categorical Data". Machinelearningmastery, 2020.
3. Miller, Rupert G. "*Survival analysis*", John Wiley & Sons, 1997
4. Singh, R.; Mukhopadhyay, K. "Survival analysis in clinical trials: Basics and must know areas". *Perspect Clin Res*, 2011
5. E. L. Kaplan and Paul Meier. "Nonparametric estimation from incomplete observations". Journal of the American Statistical Association, 1958.
6. David R Cox. "Regression models and life-tables". Journal of the Royal Statistical Society: Series B (Methodological), 1972.
7. Leblanc, Michael; Crowley, John. "Survival Trees by Goodness of Split". *Journal of the American Statistical Association*. 88, 1993
8. Ishwaran, Hemant; Kogalur, Udaya B.; Blackstone, Eugene H.; Lauer, Michael S. "Random survival forests". *The Annals of Applied Statistics*, 2008
9. Qiu, Y., Zhou, J., Khandelwal, M. *et al.* "Performance evaluation of hybrid WOA-XGBoost, GWO-XGBoost and BO-XGBoost models to predict blast-induced ground vibration". *Engineering with Computers*, 2021
10. Tianqi Chen and Carlos Guestrin. "XGBoost: A Scalable Tree Boosting System". In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016

