Computer Methods and Programs in Biomedicine 231 (2023) 107373

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

Breast cancer survival analysis agents for clinical decision support

Gaetano Manzo[a,e,*], Yvan Pannatier[a], Patrick Duflot[b], Philippe Kolh[b], Marcela Chavez[b], Valérie Bleret[c], Davide Calvaresi[a], Oscar Jimenez-del-Toro[a], Michael Schumacher[a], Jean-Paul Calbimonte[a,d]

[a] University of Applied Sciences and Arts Western Switzerland (HES-SO), Switzerland
[b] CHU of Liege, Department of Information System Management, Belgium
[c] CHU of Liege, Senology Department, Belgium
[d] The Sense Innovation and Research Center, Lausanne and Sion, Switzerland
[e] National Institutes of Health (NIH), Bethesda, MD, USA

ARTICLE INFO

Article history:
Received 6 May 2022
Revised 31 December 2022
Accepted 23 January 2023

Keywords:
Survival analysis
Machine learning
Modular architecture
Decision-system

ABSTRACT

Personalized support and assistance are essential for cancer survivors, given the physical and psychological consequences they have to suffer after all the treatments and conditions associated with this illness. Digital assistive technologies have proved to be effective in enhancing the quality of life of cancer survivors, for instance, through physical exercise monitoring and recommendation or emotional support and prediction. To maximize the efficacy of these techniques, it is challenging to develop accurate models of patient trajectories, which are typically fed with information acquired from retrospective datasets. This paper presents a Machine Learning-based survival model embedded in a clinical decision system architecture for predicting cancer survivors' trajectories. The proposed architecture of the system, named PERSIST, integrates the enrichment and pre-processing of clinical datasets coming from different sources and the development of clinical decision support modules. Moreover, the model includes detecting high-risk markers, which have been evaluated in terms of performance using both a third-party dataset of breast cancer patients and a retrospective dataset collected in the context of the PERSIST clinical study.

1. Introduction

Cancer is a significant public health concern worldwide and the second leading cause of death in the United States and Europe. For women, breast cancer, lung cancer, and colorectal cancers account for 51% of all new diagnoses, with breast cancer alone accounting for almost one-third [1]. According to the American Cancer Society, one in eight women will develop an invasive breast tumor during her life, constituting one of the most common cancers in the women population —second only to skin cancer [2]. Despite the outstanding progress in identifying many risk factors that increase women's chance of developing breast cancer, its detection remains an open challenge caused by a combination of genetic, hormonal, and environmental factors [2,3].

Patient trajectories play a crucial role in detecting high-risk markers of breast cancer patients. On the one hand, patient trajectories describe the probability that a particular event occurs over time (e.g., death or relapse). On the other hand, covariates such as type of treatments, cancer stage, and age affect the trajectories' shape, enabling clustering and risk factors detection. Patient cohort and trajectory analysis are essential supporting tools for patient stratification, identifying risks, and preventing adverse events.

According to the World Health Organization (WHO), improvements in treatment adherence would be more beneficial to the patient's health than the development of new drugs [4]. However, the accurate staging of some cancers and their prognosis is still a challenging task [5], which often leads to insufficient or unnecessary treatments (e.g., for inaccurate staging in oral cancer [6]). Therefore, the development of assistive technologies that *(i)* effectively evaluate the significance of prognostic variables (e.g., death or relapse), *(ii)* facilitate the detection of patient's high-risk markers, *(iii)* support treatment decisions, and *(iv)* improve the patients' treatment adherence, is imperative.

This paper presents the Survival Analysis components for the PERSIST modular architecture, enabling clinical support decisions for breast cancer patients. The PERSIST survival module provides patients stratification based on their trajectory analysis and enables the detection of high-risk markers. Numerical evaluations show the Survival Analysis outputs effectiveness in providing in-

* Corresponding author.
E-mail address: gaetano.manzo@nih.gov (G. Manzo).

https://doi.org/10.1016/j.cmpb.2023.107373
0169-2607/

sightful prediction and classification results in breast cancer survivor patient support the evaluation has been performed using the well-known METABRIC dataset [7], as well as the data collected within the context of the PERSIST clinical trial [8].

The rest of the paper is organized as follows. Section 2 presents the state of the art followed by the open challenges. The multi-agent-modeled architecture of PERSIST is introduced in Section 3, together with its components, interactions, and modules. The PERSIST Survival Analysis module is presented in Section 4, whereas Section 5 provides its evaluation and results. A discussion on the results is provided in Section 6. Finally, Section 7 presents the conclusions of the paper, limitations, and future work.

2. State of the art

2.1. Survival analysis

Survival analysis methods are fundamentally based on the type of disease intended to prevent. In the case of breast cancer, authors in [9] performed patient cohorts based on administered treatment to identify and rank all prognostic biomarkers —genes capable of predicting the expected survival of the patients. In [9], such an analysis was performed using the Cox Proportional Hazard (CPH) model [10], a standard parametric method that assesses patient covariates using linear combinations [11]. Similarly, in [12], authors classified cancer hallmark genes according to their correlation to the survival cohort of distinct cancer types. Although the CPH model presents explanatory output and fast computation for survival analysis, its performance and accuracy drastically drop when dimensionality increases (i.e., the number of covariates growths). Moreover, the CPH model fails to represent non-linearity and time-dependent covariates.

In [13], authors compared the performance of the classical CPH model with several Machine Learning (ML) techniques in predicting breast cancer survival. They used SHapley Additive exPlanation (SHAP) [14] values to exemplify the performance of the classical CPH regression and the best-performing ML techniques, facilitating their interpretation. Machine Learning, the primary technical basis for data mining, provides a methodology for analyzing raw data from medical records. In [15], authors applied ML techniques using survival statistics to predict graft survival. An advanced ML pipeline for survival analysis was assembled by [16]. The authors used seven well-known ML techniques and Cox regression-based survival analysis to identify breast cancer sub-types most significant miRNA biomarkers. The mission of [17] was reducing unplanned and early re-admissions, which burdens limited hospital resources imposing costs on the healthcare system. The authors in [17] proposed applications of survival models to support managerial decision-making and performance measures suitable for assessing the survival models for these applications. Authors in [18] and [19] proposed the architecture of a multi-agent system that enables patients' cohort and trajectory analysis, which inspired our work.

One of the latest survival analysis approaches was conducted in [20] using deep learning. The authors introduced deep survival analysis, a hierarchical generative approach to survival analysis for electronic health records. Compared to the clinically validated risk score, deep survival analysis is superior in stratifying patients according to their risk [20]. In addition, performance was enhanced in [21] and [22] for model ranking and accurate prediction on the overall survival patients. Those ML-based approaches catch highly complex and non-linear relationships between prognostic features and individual risks. However, previous studies have demonstrated mixed results on predicting risk, failing to show improvements beyond the linear Cox model [23,24].

The studies presented lay at the intersection of several disciplines and domains, including patient survival analysis, decision support systems, and eHealth patient support. Challenges and opportunities arising from the combined synergy of these areas can be summarized as follows: *(i)* a system that enables the collection of patient trajectory data for enabling profiling and prediction of trajectory outcomes; *(ii)* The support of persuasive strategies for self-efficacy evaluation; *(iii)* A clinical-decision-support-system that bridges statistical, rule-based, and data-driven approaches.

Breast cancer patients could benefit from such combinations enabling advanced trajectory analysis and identification/detection of markers to support clinician decisions.

2.2. AI-based models

The following briefly describes the artificial intelligence (AI) based models found in the literature, which are used for the survival classification task in Section 5. We first introduce the unsupervised algorithm for the risk evaluation of the health trajectories. Then, we present the supervised algorithms adopted for the survival classification task.

The Gaussian mixture model is a probabilistic model adopted to evaluate patient risk levels. The model assumes that all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters [25]. One can think of mixture models as generalizing k-means clustering to incorporate information about the covariance structure of the data as well as the centers of the latent Gaussians. We adopt such a model to evaluate patients' risk given patient health trajectories.

Logistic Regression (LR) is a statistical model that estimates the probability of an event occurring based on a data set of independent variables. It is the first model that we adopt for the survival classification task (e.g., estimation of patient vital status). The outcome is a probability, so the dependent (or result) variable takes values in the range [0,1]. In LR, a logit transformation is applied to the probability of success divided by the probability of failure [26].

Support-vector machine (SVM) is a machine-learning method for classification problems. Unlike LR, SVM maps the non-linearly input vectors into a very high-dimension feature space. In this feature space, a linear decision surface is constructed. Special properties of the decision surface ensure the high generalization ability of the learning machine [27].

A Decision Tree (DT) is a non-parametric supervised learning algorithm with a hierarchical tree structure consisting of a root node, branches, internal nodes, and leaf nodes [28]. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the characteristics of the input data. Such model provides explainable output based on patients' covariates.

We use Artificial Neural Network (ANN or NN) to finalize the survival classification task. NN is a deep-learning structure inspired by the human brain replicates the way biological neurons communicate with each other [29]. NNs are composed of layers of nodes called neurons. They contain an input layer, one or more hidden layers, and an output layer. These networks can receive several features at their input layer, perform the relevant operations, and provide a prediction at their output layers, offering high performance. Neurons are connected and have an associated weight and threshold, which allow neurons to be activated and transmit information to the next layer only if its output is above the threshold.

3. Architecture

This section presents the agent-based modelization of the PERSIST infrastructure to realize personalized agents leveraging the

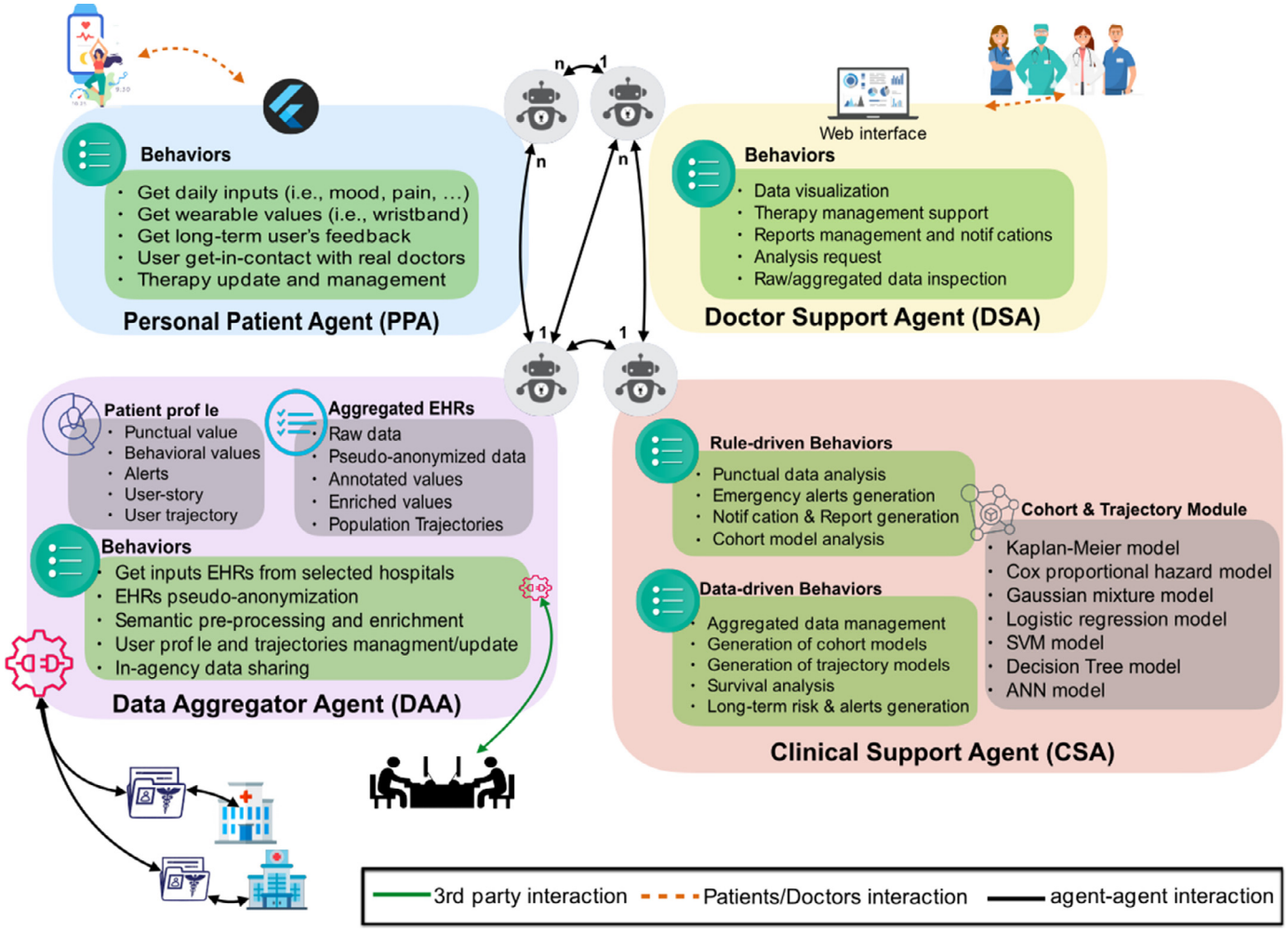


Fig. 1. PERSIST architecture described in terms of multi-agent interactions among its components.

outcome of big data analytics of cancer survivors to support a better therapy definition, patients adherence to the treatment, and follow-up monitoring. The requirements of the PERSIST clinical decision support system have been published in [30], and the system has been used during the PERSIST clinical trial [8].

The overall PERSIST infrastructure helps to (i) improve the management of the several dimensions of the disease and its treatments, promoting better health and well-being; (ii) enhance decision-making support and effectiveness in cancer treatment/follow-up; and (iii) reduce the probability of secondary diseases and fatal events —improving prevention strategies.

The agency is virtualized and interconnected by Docker-Compose, which extends and amends the architecture presented in [19] and [18] due to data (i.e., EHRs) and analysis (i.e., both rule and data-driven) requirements. Figure 1 shows the PERSIST architecture, modeled as decentralized agents with specific behaviors. In particular, this model describes the system components as four types of agents: Personal Patient Agent (PPA), Doctor Support Agent (DSA), Data Aggregator Agent (DAA), and Clinical Support Agent (CSA).

The Personal Patient Agent (PPA) represents a mobile app (multi-platform chatbot app) that enables patients to submit both self-reported values (e.g., patient reported outcomes) and wearable-related data (e.g., heart rate, BMI, and steps). The mobile app enables patients to ask for information (e.g., quality of life), and to get possible updates of their therapy. Moreover, patients can use the PPA (implemented as the app) to obtain personalized medical advice from the clinician(s) in charge, and obtain pertinent alerts and notifications. The details of this implementation have been presented in [31].

The data collected by the PPA are transferred to the Data Aggregator Agent (DAA), which holds the user profile records, and relates them with the EHR obtained from the hospital. The DAA pseudo-anonymizes[1] the multi-source EHRs (i.e., collected from healthcare facilities), enabling a dedicated partner to perform semantic enrichment [32]. Such data are stored on the OVH cloud platform in Health Level Seven (HL7) Fast Healthcare Interoperability Resources (FHIR) standard, which supports data exchange among different hospitals or clinics.

The Doctor Support Agent (DSA) aims to support clinical decisions by providing patient cohorts, risk markers, and trajectories. It is characterized by two main sets of behaviors (i.e., rule-driven and data-driven). The DSA's behaviors enable the agents' interaction and support the doctors through a web-based dashboard. In particular, it (i) enables clinicians to communicate with patients via the PPA, (ii) has access to the patients' profiles and EHRs via the DAA, and (iii) can access the periodic reports composed by the Clinical Support Agent (CSA).

[1] The records are anonymized before being processed by the agency. Nevertheless, the link between the trajectories and aggregated information and the real patient identity is kept (hospital-side) to allow the doctor to identify and monitor given patients.

Each agent plays a crucial role. Nevertheless, the CSA is the knowledge digger, constantly pulling together EHRs, patients' behaviors, models, and predictions (enabling the main contribution of this study). The following section details the CSA, its interactions with the DAA to the Big Data platform, personalization for patient stratification, and output to visualize patients' trajectories.

3.1. CSA: Architecture and characterization

The CSA's main aim is to estimate patients' survival and risk levels preventing fatal events (e.g., relapse). To do so, it uses the following behaviors,

- b1 Data retrieving: it gets access to the data that have been preprocessed, filtered, annotated, and enriched by the DAA. In particular, it performs elasticsearch queries to the DAA to fetch data from PERSIST aggregated EHRs [8,32]. Elasticsearch enables retrieving specific covariates efficiently (i.e., data retrieval computational time) by offloading the edge network based on FHIR.
- b2 Data preparation: Once retrieved the data, they are preprocessed (cleaned, encoded, and normalized). The preprocessing strongly depends on the survival model adopted. Therefore, a filter is applied to the dataset enabling the system to handle missing, heterogeneous, and multidimensional data. Finally, CSV-like files containing data are ready to be processed.
- b3 Computation of the survival models: It leverages several models, such as Kaplan-Meier Estimator and Cox Proportional Hazard [10], to estimate patients' trajectories, cohorts, and risk levels. The reasoning engine operates on the output produced by b2 and according to the parameters set by the doctors (via the DSA). Such parameters are used to select time features and events grouped by a given list of covariates. The current engine uses models such as Kaplan-Meier Estimator, Cox Proportional Hazard, and Artificial Neural Networks to compute the patients' trajectories and cohorts (see Section 4 for more details).
- b4 Rule-driven Behavior: it analyzes a tree structure that identifies patients' risks by patients' covariates such as nipple discharge, bilateral mastectomy, or skin retraction features. Each tree leaf is a checkpoint for a specific set of covariates; a risk is flagged if one of the values is above a given threshold. The rule-driven outputs recommended treatments, diagnosis, prognosis, and a related risk score, which matches the outcome score from the Survival Analysis module. This latter relies on patients' EHRs to estimate their survival trajectories. The rule-based analysis is executed punctually and with a short-term aggregation (daily), whereas the data-driven analysis is performed over a more extended period to investigate trajectories and more resilient changes.
- b5 Data summary and aggregation: this task can be triggered periodically by the CSA, or on-demand by the DSA. It elaborates on b3/b4's outputs and organizes the data in a JSON format ready to be displayed by the DSA.

The code of the proposed architecture can be found in the footnote.[2]

4. Methods

Within the Clinical Support Agent, we propose the inclusion of the Cohort and Trajectory Analysis (CTA) models, whose purpose is to provide decision-support information for clinicians regarding risks, symptoms, and disease associations. This section describes the statistical models of the survival analysis architecture previously presented.

[2] https://github.com/tanoManzo/persist

Patient trajectories describe the probability over time that a particular event, such as death or disease recurrence, occurs (i.e., patient's evolution from the diagnosis of the disease). On the other hand, cohorts aim at grouping patients with similar disease progression, pre-and post-treatment, and other covariates. Through the analysis of trajectories and cohorts, it is possible to identify and quantify associations between symptoms and events, enabling the detection of high-risk markers for detrimental treatment effects, subsequent cancer disease, and metastatic cancer disease.

4.1. Survival models

Given t, the time from the beginning of the observation period, we denote $S(t)$ as the survival function. We assume that at the beginning of the study, all the patients are alive, $S(t=0)=1$. We assume that the survival function is a monotonically non-increasing function, $S(u) < S(t), \quad \forall \quad u < t$, therefore, $S(t \to \infty) = 0$. Given T, when an event occurs (e.g., check-up time), the probability that the patient is alive after T is $P(t > T)$, which corresponds to $S(t)$ for $T \in [0, \infty)$. On the other hand, the probability that the patient did not survive after T is $1 - P(t > T)$ or $P(t \leq T)$.

The hazard function, denoted λ, is the event rate at time t conditional on survival until time t. The hazard function describes the probability that the patients will not survive for an additional time dt after surviving at t:

$$\lambda(t) = \lim_{dt \to 0} \frac{Pr(t \leq T < t + dt)}{dt\, S(t)} = -\frac{S'(t)}{S(t)}, \tag{1}$$

where $S'(t)$ is the survival event density function [33].

Given a dataset with patients observing time and event outcome, we can estimate the survival curve through the Kaplan-Meier Estimator [34]:

$$S(t) = \prod_{i=0}^{t} 1 - Pr(T = i, t \geq i) = \prod_{i=0}^{t} 1 - \frac{d_i}{n_i}. \tag{2}$$

With d_i the number of patients that had an event at time i, and n_i the number of patients that survived at time i. Please note that patient data are not collected continuously but at discrete intervals.

The Kaplan-Meier (KM) estimates the survival function from lifetime data. KM estimates patients' trajectories such as cancer relapse, recovery rates, and mental disorders based on the available data. Covariates group population in the KM estimator. For instance, KM outputs trajectories of breast cancer patients grouped by cancer stage or treatments. For survival analysis (i.e., to evaluate the probability of death), KM requires the observation time T and the event —death or alive. The KM formula is a non-parametric statistic obtained with the chain rule for random variables. Indeed, the KM estimator is evaluated considering that the probability is broken up into the product of probabilities during specific intervals. Moreover, KM takes into account *censored data*— any data for which we do not know the exact event time (e.g., patients withdraw from the study or die by other causes). Given its non-parametric nature, KM is limited to estimate survival adjusted for covariates. Indeed, KM considers only the observation time and the event, neglecting other covariates. Parametric models such as Cox Proportional Hazard estimate covariates-adjusted survival.

The Cox Proportional Hazard (CPH) model estimates individual trajectories leveraging on patients' covariates [35]. CPH enables personalized patient treatments evaluating the hazard function previously introduced —the immediate death risk probability for a patient that survived at time t. In this work, we define the factor risk as a linear combination of the patient's features $X = (x_1, x_2, \ldots, x_n)$ and the respective features' weights $\Theta = (\theta_1, \theta_2, \ldots, \theta_n)$, with n the number of patient. In the CPH model,

covariates are multiplicatively related to the hazard, which is assumed to respond exponentially; each unit increase in X results in proportional scaling of the hazard (here the name *Proportional Hazard*):

$$\lambda(t|X) = \lambda_0 e^{(\theta_1 x_1, \theta_2 x_2, \ldots, \theta_n x_n)} = \lambda_0 e^{\left(\sum_{i=0}^{n} \theta_i x_i\right)} = \lambda_0 e^{(\Theta^T X)}, \quad (3)$$

where t is the observation period and λ_0 is the baseline risk (i.e., describing how the risk of event changes over time at baseline levels of covariates). Please notice that the survival function 2 is strictly related to the hazard function, as follows:

$$\lambda(t) = -\frac{S'(t)}{S(t)} \quad (4)$$

and vice versa by integrating both members,

$$S(t) = e^{-\int_0^t \lambda(u)du}, \quad (5)$$

The CPH model assumes that the baseline hazard function λ_0 follows a given function and that the covariates are time-independent. Those limitations are overcome using data-driven AI approaches such as Artificial Neural Networks and Clustering presented below.

4.2. Risk-markers detection

Kaplan-Meier and Cox Proportional Hazard models enable CTA to build comprehensive trajectories to support clinical decisions. Complementary to these features, we propose incorporating supervised and unsupervised machine learning approaches to help understand individual trajectories based on similar healthcare records.

The first task is to classify patients' status based on time-dependent covariates, such as cancer stage, ECOG, and treatments. This task uses binary estimators such as Support Vector Machine (SVM) and Random Forest (RF). Those models handle continuous and categorical features, outliers, and missing data. In particular, SVM and RF provide explanatory outcomes and help identify the decision boundary (i.e., covariates importance). We use deep learning models based on Artificial Neural Networks, which capture complex feature relationships (e.g., cancer type and Nottingham Prognostic Index) and improve classification outcomes (i.e., output the label group to which the patient belongs).

While classification methods investigate event-covariates relationships, clustering approaches identify patients' cohorts and interactions among those groups. Such interactions are the covariates enabling patients to change groups or, in other words, reveal risk-makers. Therefore, detecting the risk-markers means detecting covariates that allow patients to pass from low to high-risk prediction and vice versa (i.e., to change cluster).

To segregate patients with similar traits and assign them into clusters, we use K-Means, Gaussian Mixture Models, and Trajectory-based clustering algorithms. Using clustering, the CTA finds and labels (e.g., low-high risks) similar trajectories without pre-defined assumptions about their patient's characteristics (i.e., unsupervised learning). Such risk-markers are withdrawn by API, which makes them available for the CSA.

5. Results

In this section, we evaluate the performance of the Survival Analysis module previously presented for the personalized classification of trajectory patterns, cohorts, and high-risk markers detection. We first introduce the PERSIST breast cancer dataset stored in the Big Data platform to evaluate our models. Then, we explore the trajectory based on the KM and CPH models. Finally, we depict the high-risk markers (i.e., covariates importance) and cohort the patients in several risk levels based on their trajectories and EHRs. Please notice that, in the following experiments, we define the survivor function $S(t)$ as the survival probability of the patient(s) over time (i.e., years). $S(t)$ can be visualized on a particular sub-population (e.g., breast cancer patients that received specific treatments). Moreover, as shown at the end of this section, $S(t)$ outcomes play a crucial role in detecting the patient risk level and cohorts. The Survival Analysis module generates aggregated information, data visualization, and risk alerts to the DSA crucial for clinicians as decision support for diagnosis, treatments, and prognosis.

5.1. Dataset and data wrangling

We use external and internal breast cancer datasets to evaluate the Survival Analysis module. As a third-party dataset, we use METABRIC (Molecular Taxonomy of Breast Cancer International dataset Consortium [7]). The METABRIC dataset consists of gene expression data and clinical features for 2,498 patients labeled as follows: 33.34% *"Living"*, 25.74% *"Died due to breast cancer"*, 19.80% *"Died due to other causes"*, and the rest *"not observed"*.

As an internal dataset, the Big Data Platform in the PERSIST architecture stores patient data from four European hospitals: Centre Hospitalier Universitaire De Liege (CHU de Liége, Belgium), Univerzitetni Klinicni Center Maribor (UKCM, Slovenia), Latvijas Universitate (LU, Latvia), and National Patients Organisation (NPO, Bulgaria). The heterogeneity of this dataset is an additional challenge, as the coding systems differ between hospitals and within the hospitals themselves. For instance, hospitals may switch from the coding system ICD-9 to a newer version without updating retrospective data. Therefore, we restrict the target cohort to CHU de Liége patients in order to enhance data homogeneity as a first step. To achieve this goal, we design two handlers for our *Dataset Builder module*. The first one selects patients of a chosen hospital, while the second one selects breast cancer patients. The resulting CHU de Liége breast cancer dataset comprises 2085 patients, of which 399 died. All patients were born between 1915 and 1991. The dataset includes 45 different features such as patient age, ECOG status, HER2 levels, tumor stage, and treatments.

Multiple missing values for features such as BMI, ECOG, and HER2 were encountered and addressed as follows. Features with more than 70% of the missing values were erased since drastically reducing the sample population —poor estimator performance. Features with less than 70% of missing values were treated with median based on the feature type. Age and BMI were missing for less than 30% of the patients. Looking at the result of Cox Proportional Hazard, these two features have a very low impact on the survival probability of the patients. Concerning ECOG and HER2, only 8% respectively 17% of the values were missing. Therefore we decided to fill the missing values of these four features with their respective median. Please notice that even if filling missing values with median does not affect the distribution of the feature, it can introduce biases. Therefore, in most cases, we prefer leaving the value to "none", indicating the algorithm of a missing value.

5.2. Survival analysis

We present the main results of our data analysis outputting the survival patients' trajectories. We start analyzing the dataset using the KM estimator. Figure 2 shows the impact of treatments on the survival probability for breast cancer patients. Surgeries such as mastectomy or lymphadenectomy have a higher probability of surviving longer than other treatments. On the other hand, survival probability is dropping faster for patients treated with injections such as antibiotics or chemotherapeutics. According to the estimation, patients following these treatments have around 80% of chances to survive at least ten years. Please notice that such results

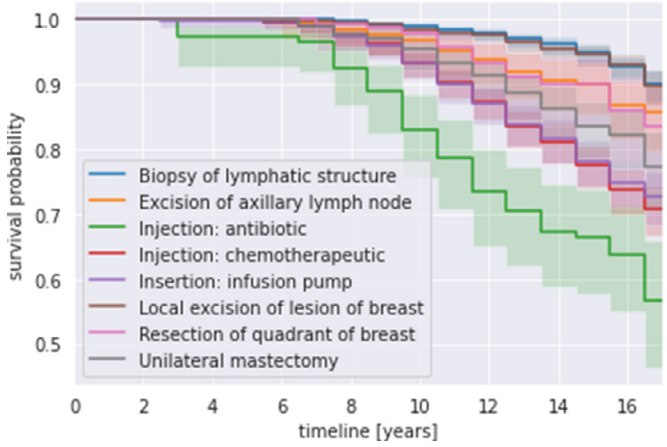


Fig. 2. Kaplan-Meier survival probability estimation of the breast cancer population in PERSIST grouped by treatments.

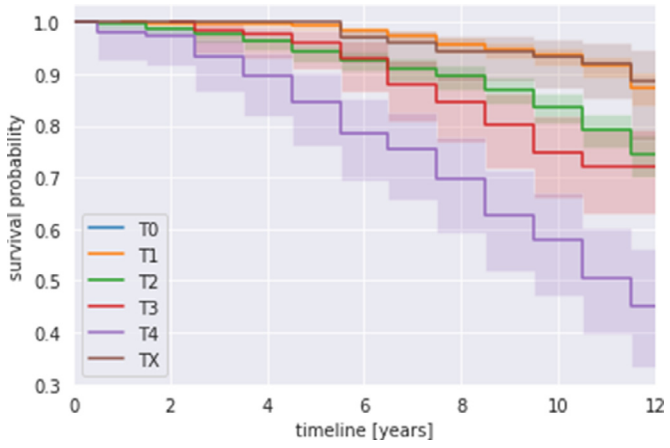


Fig. 3. Kaplan-Meier survival probability estimation of the breast cancer population in PERSIST grouped by tumor size.

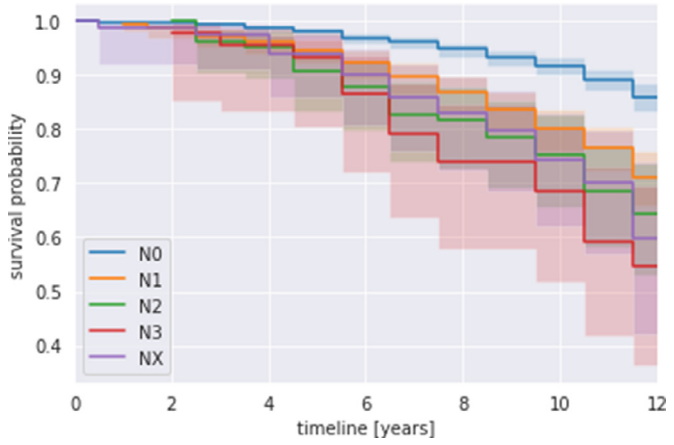


Fig. 4. Kaplan-Meier survival probability estimation of the breast cancer population in PERSIST grouped by cancerous lymph node presence.

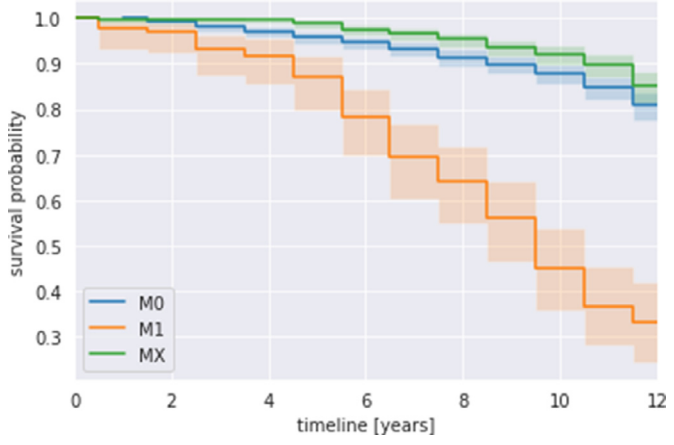


Fig. 5. Kaplan-Meier survival probability estimation of the breast cancer population in PERSIST grouped by metastases presence.

must be consulted with other covariates. Indeed, even if a mutation of the BRCA1 gene increases the risk for breast cancer, not all women carrying the BRCA1 mutation develop cancer [36].

In the PERSIST dataset, the cancer stage is categorized using the TNM staging, both clinical and pathological. TNM is based on tumor size (T), cancer cells spread in lymph nodes (N), and presence of metastases (M). We estimate the impact of these three components on survival probability using KM model. Please notice that we used clinical TNM for missing pathological TNM values.

Figure 3 shows the survival probability for patients grouped by tumor size, which strongly affects the survival probability. Indeed, patients with small tumor size (T1) or not assessable tumors (T0) are more likely to survive longer given the small size of the mass. As the tumor size increases (e.g. T4) the survival probability decreases rapidly. According to the estimations, middle size tumor (T2-T3) has almost the same probability of surviving 6 years. Please notice that patients were given additional cancer treatment to lower the relapse risk —adjuvant.

The presence of cancer in lymph nodes highlights the relation between the number of lymph node-positive breast cancer and the decreasing survival probability. Figure 4 shows this relationship by plotting the KM trajectories grouped by node-positive breast cancer (N). Patients with node-negative breast cancer (N0) are likely to survive longer than those with positive nodes. On the one hand, the survival probability decreases quickly after 6 years for patients presenting several node-positive breast cancer (e.g N2).

Finally, Fig. 5 shows how metastasis strongly decreases the patient survival probability. In the PERSIST datasets, patients at diagnosis without metastases (M0) or missing metastases information (MX) have about 90% chance of survival for ten years and more, against 45% of patients with one metastasis (M1).

Kaplan-Meier enables data exploration in order to gain insights into the breast cancer population and their estimated trajectories. However, given its non-parameter statistical property, covariates only indirectly affect the shown trajectories through the observed event (i.e., patient alive or death). To overcome such limitations, we use Cox proportional hazard for highlighting feature contributions to the estimated survival probability.

5.3. High-risk markers detection

We describe the detection of high-risk markers for the given career population. Such markers are used as decision support for clinicians. To detect and capture the non-linear correlation of such high-risk markers, we use AI-based models. In particular, unsupervised machine learning models for detecting high-risk trajectories and supervised machine learning models for the survival classification task. In the following, we use the CPH model to estimate the feature importance. Figure 6 shows the influence of the different features. Positive values mean a negative impact on the survival probability. In contrast, negative values mean a positive impact on survival probability. Each covariate is within the confidence interval denoting the accuracy of the system for specific features. For

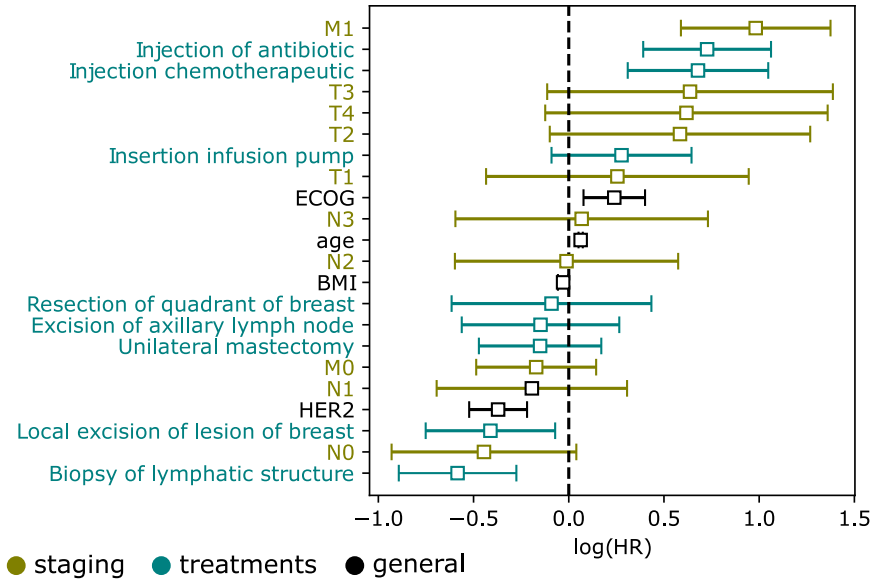
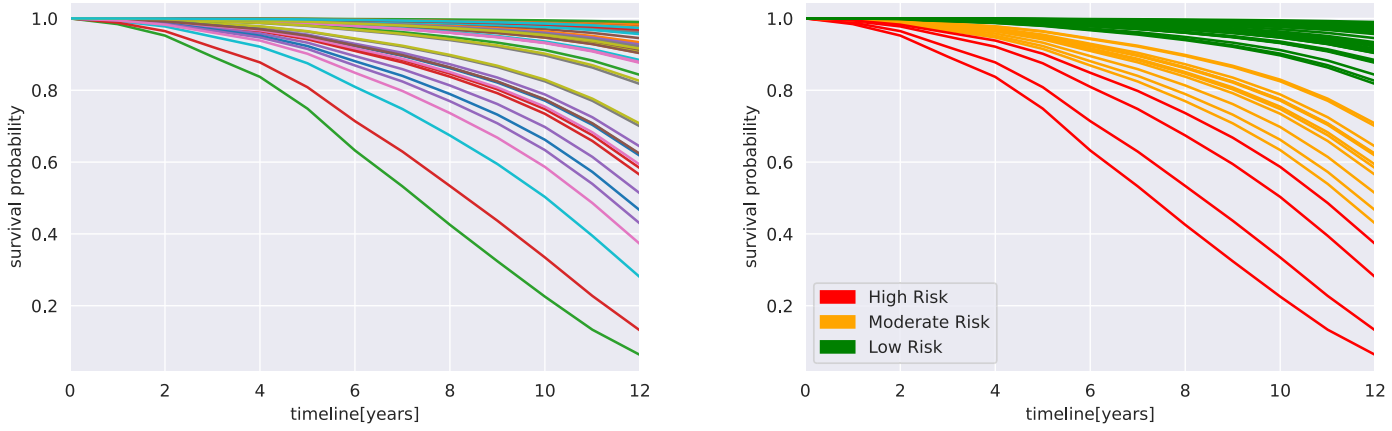


Fig. 6. Cox Proportional Hazard based on the survival analysis of the breast cancer population in PERSIST.



(a) Trajectories (b) Risk clusters

Fig. 7. CPH based on a 50 patients in PERSIST.

instance, given the slight interval, the system is confident to highlight the patient's age as a negative contribution to the patient's survival probability —the older the patients, the greater the impact on their survival trajectory. The metastases presence is the feature with the most negative impact on the survival probability. The tumor size is the second feature that has the greatest impact. On the contrary, surgery such as unilateral mastectomy has a positive impact on the survival probability. We also notice that the absence of node-positive breast cancer enhances the survival probability. The CPH shows the contribution of each covariate to the patient survival probability and helps to estimate covariates-adjusted trajectories. Such trajectories can be grouped denoting patients' risk levels. Therefore, the high-risk markers can be defined as the covariates that move the patients' trajectories from a low-risk to a higher-risk cluster.

We aim to enable decision support by categorizing patients according to their risk level. To this end, Fig. 7 shows the K-means model performing the cohort analysis based on the trajectory of patients from the PERSIST dataset. This allowed us to generate Fig. 7a, which shows the results obtained for the 50 randomly selected patients.

Each survival function in Fig. 7a represents the probability that a patient will survive a given number of years. We applied a k-means clustering algorithm on the sample population to obtain three distinct clusters: high-risk, medium-risk, and low-risk. On the one hand, patients in the low-risk area have shown a high survival probability. On the other hand, patients in the high-risk area have shown a low survival probability. The K-means model clusters the patient trajectory in one of the above-mentioned areas defining the patient risk. The trajectory, based on the Cox Proportional Hazard model, takes into account the relationship among the patient covariates and the relationship between patients. Please notice that covariates' importance and divergence are both significantly affecting patients' cluster membership.

Figure 7b, shows that the algorithm identified three clusters of patients. In red, the high-risk patients with the lowest 12-year survival probability. In orange, patients with a medium 12-year survival probability of survival and a moderate risk. In green, patients with a very high probability of 12-year survival and therefore present a low risk. However, we believe that this model is only marginally representative of the level of risk incurred by the patient over the years since only one value of the time line (in this

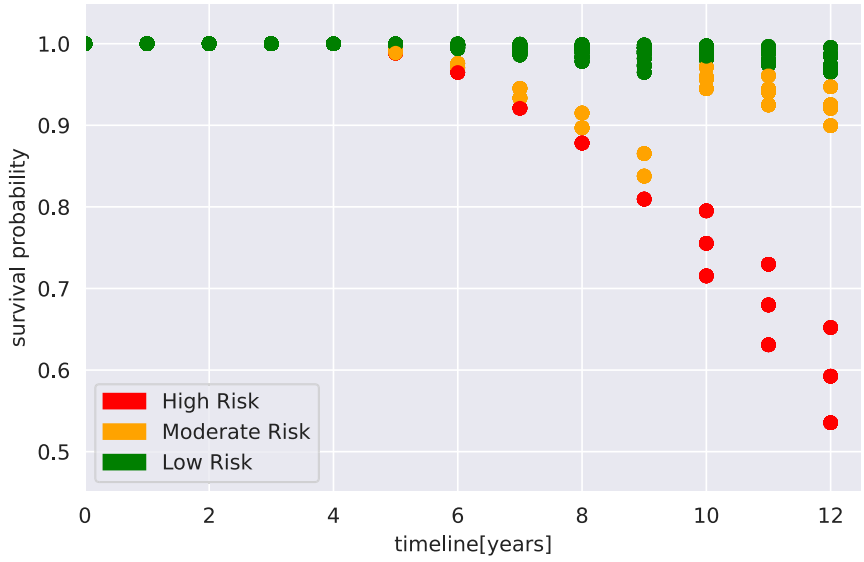


Fig. 8. Cox Proportional Hazard survival function clustered in 3 risk groups for each years based on a 15 patients sample of the breast cancer population in PERSIST.

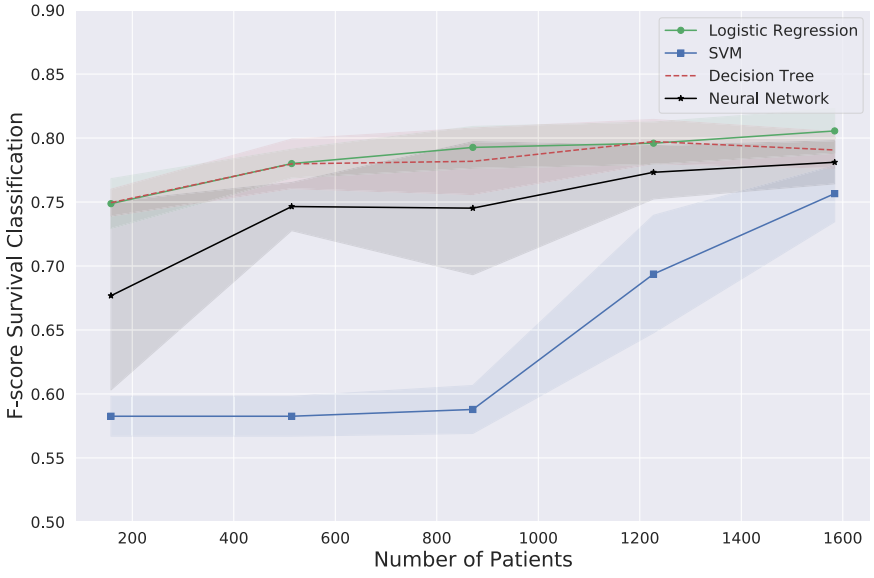


Fig. 9. F-score for survival classification using Logistic Regression, SVM, Decision Tree, and Neural Networks.

case, the last one) is taken into account. If we look at all the points at, for instance, 2-year survival, most of the curves are very close making it difficult to identify which patient is at risk and which is not.

Figure 8 shows how the algorithm identifies the three clusters (i.e., risk-level) based on patients' trajectories. Please notice that for the sake of clarity, in Fig. 8 we reduced the number of trajectories (from 50 to 15) and removed the interpolation (from lines to points). To evaluate patient trajectory risk, we adopt the unsupervised Gaussian model introduced in 2. The model takes the patients trajectories as input and shapes the risk level based on the magnitude of the derivatives. We train the model with three components, which correspond to the three risk levels, and with full covariance (i.e., Cholesky decomposition of the precision matrices of each mixture component).

In red, the high-risk patients with the lowest 12-year survival probability. In orange, patients with a medium 12-year survival probability and a moderate risk. In green, patients with a very high probability of 12-year survival present a low risk.

In Fig. 8, from 1 to 4 years of prediction, the system fails to split the risk level given the high correspondence between the trajectories. However, as soon as the trajectories approach low probabilities (see years from 5 to 12), the system identifies the three main clusters (i.e., low, moderate, and high risks).

We adopted the supervised machine learning algorithms presented in Section 2 for the survival classification task. The goal is to classify patients' trajectories as a function of patient vital status (alive/dead). All the models have been trained with 60%, validated with 20% (10-fold cross-validation), and tested with 20% of the datasets. Please notice that for the Logistic Regression, no penalty is added; for the SVM a support vector classifier, a linear kernel is added; for the Decision Tree, no max depth is applied; finally, for NN, a multi-layer perceptron with two hidden layers is added.

Figure 9 depicts the accuracy metric F-score for the survival classification task, defining patients' risk over the number of patients used for the training phase. We see that the classifiers enhance their accuracy when the number of patients in training set increases. Particularly for the Neural Network model, which re-

quires many examples to train its neurons. Models such as Decision Tree and Logistic Regression provide the best accuracy level even for a few train examples.

6. Discussion

6.1. PERSIST architecture and CSA

The presented architecture aims at creating an open and interoperable ecosystem to improve the care of cancer survivors. The impact of such architecture can (i) boost self-efficacy and satisfaction with care, reducing psychological stress for better management of the consequences of the cancer treatment; (ii) increase effectiveness in cancer treatment and follow-up by providing prediction models from Big Data that support optimal treatment decisions; and (iii) advance the efficacy of management, intervention, and prevention in order to timely treat side effects and secondary diseases.

In the presented architecture, the Clinical Support Agent, with its primary module *Cohort and Trajectory*, plays a crucial role as a support decision tool for clinicians. Predictions are exploited in diagnosis, treatments, and prognosis. For the presented scenario and beyond, the trajectories help stratify patients, highlighting peculiar covariates. The predictions support treatment decisions; for instance, we noticed that chemotherapy could reduce patient survival probability over time for patients with an advanced cancer stage. Finally, for prognosis, patients' trajectories help prevent side effects and secondary diseases (e.g., stress, burnout, and depression). The Clinical Support Agent must be part of the PERSIST architecture to leverage the other modules for data collection, management, and visualization. But the Clinical Support Agent provides meaning to such data, using state-of-the-art models to estimate and predict trajectories, which could enhance patient care. The trajectory analysis results performed on the cancer survivors show different pathways for feeding and enriching the knowledge-bases of clinical decision support systems. The prediction models built on retrospective data gathered from patients provide precious information, orienting the most appropriate recommendations for survivors.

6.2. Survivor analysis via tajectories

The probabilistic foundations of the estimators provide sufficient certainty about the potential outcomes of a given patient. Specifically, using the analysis with the Kaplan-Meier approach, it is possible to estimate survival probability over time with respect to features such as the type of treatment. There are apparent trajectory differences, such as axillary lymph node excision or chemotherapeutic injection. These differences may lead to different clinical and lifestyle recommendations to maintain or improve the quality of life after these treatments. Depending on the survivorship projections (e.g., number of years) and the type of treatment, the patient's expectations and the possibilities of post-therapeutic care may significantly vary.

We can summarize the results of the survivor analysis as follows:

- *Treatments.* We have shown how the survival probability is impacted by previous treatments for breast cancer. Higher survival probability is linked to mastectomy, and lymphadenectomy, among other surgeries, compared to treatments like chemotherapeutics. These results, however, need to be analyzed carefully, as covariates reveal additional information for better informing physicians(e.g., genetic mutations).
- *Tumor size.* Survival probability is strongly impacted by tumor size. As expected, this probability decreases for larger sizes of the tumor mass, although for middle-sized ones, the number of survival years is rather similar.
- *Lymph nodes.* Presence of cancer in lymph nodes is related to a decrease in survival probability over time, especially after six years. However, before this period, it is generally too early to discriminate among multiple positive cancerous lymph nodes (N1-N3).
- *Metastases.* Compared to no metastases or missing information, the detection of metastases has a clear negative impact on survival probability over time, especially after four years.

These trajectory prediction outcomes can allow the clinical decision support system to base its action plans on actual data. For example, in the specific case of the PERSIST project, the patient App allows monitoring survivors while considering relevant information, such as cancer stage (TNM: tumor size, cancer cells spread in lymph nodes, and metastases). Furthermore, these analyses can be used separately or combined to elaborate (remote) care plans and identify groups of patients with similar conditions and, thus, similar needs. An advantage of the PERSIST approach is the combination of different features at scale, allowing patients to be monitored.

6.3. High-risk markers detection

We have also shown that detecting high-risk markers is of primal importance, as it allows presenting the impact of individual covariates on survival probability. It helps identify risk levels informed on real data. This has proven to be a scalable model that can be applied to large-size patient cohorts serving as the cornerstone for clinical digital solutions. Action plans for high, moderate, and low-risk patients, combined with knowledge about their evolution and patient-reported outcomes, can substantially help contribute to better support for cancer survivorship. However, as seen in the clustering results, the risk levels are harder to identify with sufficient accuracy early in the timeline. In this respect, the data collection behaviors described in the architecture of the PERSIST ecosystem play a fundamental role in acquiring relevant information for the training and enhancement of risk classification. The risk marker analysis results can be summarised as follows:

- *Feature influence.* Results show that age has a strong negative contribution to survival probability. Metastases and tumor size have been shown to have the largest impact on survival, while the absence of node-positive breast cancer, or surgeries such as mastectomies, have a positive impact.
- *Risk level clustering.* We have shown how patients can be grouped into three risk levels related to the probability of survival time and covariates. The results also show the difficulty of separating these clusters for the first four years after diagnosis, given that at that point, the trajectories overlap to a considerable degree.

In sum, the results of this study provide detailed information about patient trajectories and prediction of survival with respect to several treatments, cancer characteristics, and other relevant aspects for survivorship support. Given the importance of having these outcomes distilled for clinicians, it will be necessary to conduct further studies to show the clinical appropriateness of these results. This would then be translated into personalized recommendations and adapted treatments for patients.

7. Conclusions

Modeling patient trajectories has the potential to radically improve personalized and data-centric support for clinical decisions, especially in highly-prevalent diseases such as breast cancer. Moreover, as we have shown, these trajectories help describe survival

probability with respect to different patient treatments, tumor characteristics, or lymph node presence, among others. More precisely, we have provided a detailed description of our survival models, relying on estimates such as Kaplan-Meier and Cox Proportional Hazard, combined with supervised and unsupervised machine learning approaches. The evaluation of these models, using the METABRIC and the PERSIST datasets, has shown the effectiveness and appropriateness of our techniques and their potential for clinical decision support.

7.1. Limitations

Although the patient trajectory analysis presented in this work provides several indicators with a clear potential for enhancing clinical decision support, certain limitations must be considered. First, the algorithms may be sensitive to the type of patients and conditions in the training datasets. Algorithmic bias should then be taken into account, given that different populations may have different outcomes. These differences are not only due to the prevalence of cancer-specific characteristics but also to demographics (e.g., age differences among countries), region-specific lifestyles, quality of healthcare services, etc. In this study, we focused on the existing METABRIC dataset and our newly collected PERSIST dataset, which includes data from 4 different countries. However, it was limited to CHU de Liége for this work. Expanding the study to other retrospective databases would enhance the generalizability of results and reduce the impact of certain types of bias. Regarding classifying patients according to risk groups, we have reported limited accuracy in short periods. Clusters can be identified only after about 5–6 years, which would need to be analyzed in terms of clinical usefulness. Further studies would be required to determine how early this classification would need to be to impact clinical decisions or treatment changes. Another limitation is linked to survival as the main prediction target. However, other indicators have special relevance for survivorship support, such as adherence to post-operatory treatment or other home-based treatments known to have a high risk of low compliance. Additional data would be required to train models on trajectories that include this type of information and may lead to discovering new patient clusters based on lack-of-adherence risk. Moreover, this study does not yet include emotional and mental-health aspects, which also profoundly impact the quality of life of cancer survivors. In the context of the PERSIST project, the patient App described in the architecture currently collects some of this information via questionnaires and chatbot interactions. However, the data gathered stills needs to be fully acquired and processed.

7.2. Future work

Although we have focused mainly on survival analysis and high-risk markers detection in the evaluation performed in this work, it also provides the foundation for generating further data-driven insights. In future work, we plan to continue exploring prediction models related to survival and adherence to cancer survivorship treatments, emotional outcomes, mental-health & depression, and relapse episodes, among others. Another important line of research includes the extension of the modular architecture to incorporate semi-automatic interactions through the patient module to foster and potentially induce positive behavior tailored to the predicted trajectory of the patient. Moreover, we intend to include trajectory-based persuasion strategies so that clinicians can use data-driven insights to support future interventions on patient cohorts having similar characteristics.

Finally, the PERSIST clinical trial will acquire prospective data from cancer survivors in four European countries in addition to the previous retrospective data. The complementarity of these new datasets will further enrich the algorithms presented in this work and constitute more robust models for trajectory analysis, clustering, and prediction.

Declaration of Competing Interest

All authors have participated in (a) conception and design, or analysis and interpretation of the data; (b) drafting the article or revising it critically for important intellectual content; and (c) approval of the final version.

This manuscript has not been submitted to, nor is under review at, another journal or other publishing venue.

The authors have no affiliation with any organization with a direct or indirect financial interest in the subject matter discussed in the manuscript.

Acknowledgments

This work was partially supported by the H2020 Project PERSIST under grant agreement No. 875406. The authors gratefully acknowledge the helpful remarks from Dina Demner-Fushman (National Institute of Health).

References

[1] K.D. Miller, L. Nogueira, A.B. Mariotto, J.H. Rowland, K.R. Yabroff, C.M. Alfano, A. Jemal, J.L. Kramer, R.L. Siegel, Cancer treatment and survivorship statistics, 2019, CA Cancer J. Clin. 69 (5) (2019) 363–385.

[2] C.D. Runowicz, C.R. Leach, N.L. Henry, K.S. Henry, H.T. Mackey, R.L. Cowens-Alvarado, R.S. Cannady, M.L. Pratt-Chapman, S.B. Edge, L.A. Jacobs, A. Hurria, L.B. Marks, S.J. LaMonte, E. Warner, G.H. Lyman, P.A. Ganz, American cancer society/American society of clinical oncology breast cancer survivorship care guideline, CA Cancer J. Clin. 66 (1) (2016) 43–73.

[3] A. Cheville, M. Lee, T. Moynihan, K.H. Schmitz, M. Lynch, F.R. De Choudens, L. Dean, J. Basford, T. Therneau, The impact of arm lymphedema on healthcare utilization during long-term breast cancer survivorship: a population-based cohort study, J. Cancer Surviv. 14 (3) (2020) 347–355.

[4] E. Sabaté, E. Sabaté, et al., Adherence to Long-Term Therapies: Evidence for Action, World Health Organization, 2003.

[5] D.N. Louis, A. Perry, G. Reifenberger, A. Von Deimling, D. Figarella-Branger, W.K. Cavenee, H. Ohgaki, O.D. Wiestler, P. Kleihues, D.W. Ellison, The 2016 world health organization classification of tumors of the central nervous system: a summary, Acta Neuropathol. (2016).

[6] D.W. Kim, S. Lee, S. Kwon, W. Nam, I.-H. Cha, H.J. Kim, Deep learning-based survival prediction of oral cancer patients, Sci. Rep. (2019).

[7] C. Curtis, S.P. Shah, Chinriadis, The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups, Nature (2012).

[8] I. Mlakar, S. Lin, I. Aleksandraviča, K. Arcimoviča, J. Eglītis, M. Leja, Á. Salgado Barreira, J.G. Gómez, M. Salgado, J.G. Mata, D. Batorek, M. Horvat, M. Molan, M. Ravnik, J.-F. Kaux, V. Bleret, C. Loly, D. Maquet, E. Sartini, U. Smrke, Patients-centered survivorship care plan after cancer treatments based on big data and artificial intelligence technologies (persist): a multicenter study protocol to evaluate efficacy of digital tools supporting cancer survivors, BMC Med. Inform. Decis. Mak. 21 (1) (2021) 243, doi:10.1186/s12911-021-01603-w.

[9] B. Győrffy, Survival analysis across the entire transcriptome identifies biomarkers with the highest prognostic power in breast cancer, Comput. Struct. Biotechnol. J. 19 (2021) 4101–4109, doi:10.1016/j.csbj.2021.07.014.

[10] D.R. Cox, Regression models and life-tables, J. R. Stat. Soc. Ser. B (Methodological) (1972).

[11] A. Xiang, P. Lapuerta, A. Ryutov, J. Buckley, S. Azen, Comparison of the performance of neural network methods and Cox regression for censored survival data, Comput. Stat. Data Anal. (2000).

[12] Á. Nagy, G. Munkácsy, B. Győrffy, Pancancer survival analysis of cancer hallmark genes, Sci. Rep. 11 (1) (2021) 6047, doi:10.1038/s41598-021-84787-5.

[13] A. Moncada-Torres, M.C. van Maaren, M.P. Hendriks, S. Siesling, G. Geleijnse, Explainable machine learning can outperform Cox regression predictions and provide insights in breast cancer survival, Sci. Rep. 11 (1) (2021) 6968, doi:10.1038/s41598-021-86327-7.

[14] S. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, 2017. 1705.07874

[15] K.D. Yoo, J. Noh, H. Lee, D.K. Kim, C.S. Lim, Y.H. Kim, J.P. Lee, G. Kim, Y.S. Kim, A machine learning approach using survival statistics to predict graft survival in kidney transplant recipients: a multicenter cohort study, Sci. Rep. 7 (1) (2017) 8904, doi:10.1038/s41598-017-08008-8.

[16] J.P. Sarkar, I. Saha, A. Sarkar, U. Maulik, Machine learning integrated ensemble of feature selection methods followed by survival analysis for predicting breast cancer subtype specific miRNA biomarkers, Comput. Biol. Med. 131 (2021) 104244, doi:10.1016/j.compbiomed.2021.104244.

[17] J. Todd, A. Gepp, S. Stern, B.J. Vanstone, Improving decision making in the management of hospital readmissions using modern survival analysis techniques, Decis. Support Syst. (2022) 113747, doi:10.1016/j.dss.2022.113747.
[18] G. Manzo, D. Calvaresi, O. Jimenez-Del-Toro, J.P. Calbimonte, M. Schumacher, Cohort and trajectory analysis in multi-agent support systems for cancer survivors, J. Med. Syst. 45 (12) (2021) 109.
[19] D. Calvaresi, J.-P. Calbimonte, E. Siboni, S. Eggenschwiler, G. Manzo, R. Hilfiker, M. Schumacher, EREBOTS: privacy-compliant agent-based platform for multi--scenario personalized health-assistant chatbots, Electronics (2021).
[20] R. Ranganath, A. Perotte, N. Elhadad, D. Blei, Deep survival analysis, in: F. Doshi-Velez, J. Fackler, D. Kale, B. Wallace, J. Wiens (Eds.), Proceedings of the 1st Machine Learning for Healthcare Conference, Proceedings of Machine Learning Research, Vol. 56, PMLR, Northeastern University, Boston, MA, USA, 2016, pp. 101–114. https://proceedings.mlr.press/v56/Ranganath16.html
[21] E. Bilal, J. Dutkowski, J. Guinney, I.S. Jang, B.A. Logsdon, G. Pandey, B.A. Sauerwine, Y. Shimoni, H.K. Moen Vollan, B.H. Mecham, O.M. Rueda, J. Tost, C. Curtis, M.J. Alvarez, V.N. Kristensen, S. Aparicio, A.-L. Børresen-Dale, C. Caldas, A. Califano, S.H. Friend, T. Ideker, E.E. Schadt, G.A. Stolovitzky, A.A. Margolin, Improving breast cancer survival analysis through competition-based multidimensional modeling, PLoS Comput. Biol. (2013).
[22] H. Shimizu, K.I. Nakayama, A 23 gene-based molecular prognostic score precisely predicts overall survival of breast cancer patients, EBioMedicine (2019).
[23] S. Bussy, R. Veil, V. Looten, A. Burgun, S. Gaïffas, A. Guilloux, B. Ranque, A.S. Jannot, Comparison of methods for early-readmission prediction in a high--dimensional heterogeneous covariates and time-to-event outcome framework, BMC Med. Res. Methodol. (2019).
[24] L. Mariani, D. Coradini, E. Biganzoli, P. Boracchi, E. Marubini, S. Pilotti, B. Salvadori, R. Silvestrini, U. Veronesi, R. Zucali, F. Rilke, Prognostic factors for metachronous contralateral breast cancer: a comparison of the linear Cox regression model and its artificial neural network extension, Breast Cancer Res. Treat. (1997).
[25] S.P. Chatzis, D.I. Kosmopoulos, T.A. Varvarigou, Signal modeling and classification using a robust latent space model based on distributions, IEEE Trans. Signal Process. 56 (3) (2008) 949–963, doi:10.1109/TSP.2007.907912.
[26] J. Tolles, W.J. Meurer, Logistic regression: relating patient characteristics to outcomes, JAMA 316 (5) (2016) 533–534, doi:10.1001/jama.2016.7653.
[27] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (3) (1995) 273–297, doi:10.1007/BF00994018.
[28] B. Kamiński, M. Jakubczyk, P. Szufel, A framework for sensitivity analysis of decision trees, Cent. Eur. J. Oper. Res. 26 (1) (2018) 135–159, doi:10.1007/s10100-017-0479-6.
[29] J.J. Hopfield, Neural networks and physical systems with emergent collective computational abilities, Proc. Natl. Acad. Sci. 79 (8) (1982) 2554–2558, doi:10.1073/pnas.79.8.2554.
[30] U. Arioz, B. Yildiz, M. Yilmaz, V.M.C. Gonzalez, D. Caldy, S. Lin, Persist deliverable 5.1 CDSS requirements, 2020, (https://projectpersist.com/wp-content/uploads/2022/05/D5.1.-CDSS-requirements_compressed.pdf).
[31] I. Mlakar, V. Săfran, D. Hari, M. Rojc, G. Alankuş, R. Pérez Luna, U. Ariöz, Multilingual conversational systems to drive the collection of patient-reported outcomes and integration into clinical workflows, Symmetry 13 (7) (2021), doi:10.3390/sym13071187.
[32] L. González-Castro, V.M. Cal-González, G. Del Fiol, M. López-Nores, CASIDE: a data model for interoperable cancer survivorship information based on FHIR, J. Biomed. Inform. 124 (2021) 103953, doi:10.1016/j.jbi.2021.103953.
[33] E.T. Lee, O.T. Go, Survival analysis in public health research, Annu. Rev. Public Health 18 (1) (1997) 105–134, doi:10.1146/annurev.publhealth.18.1.105. PMID: 9143714
[34] E.L. Kaplan, P. Meier, Nonparametric estimation from incomplete observations, J. Am. Stat. Assoc. (1958).
[35] S.V. Deo, V. Deo, V. Sundaram, Survival analysis-Part 2: Cox proportional hazards model, Indian J. Thorac. Cardiovasc. Surg. 37 (2) (2021) 229–233.
[36] K.B. Kuchenbaecker, J.L. Hopper, D.R. Barnes, K.A. Phillips, T.M. Mooij, M.J. Roos-Blom, S. Jervis, F.E. van Leeuwen, R.L. Milne, N. Andrieu, D.E. Goldgar, M.B. Terry, M.A. Rookus, D.F. Easton, A.C. Antoniou, L. McGuffog, D.G. Evans, D. Barrowdale, D. Frost, J. Adlard, K.R. Ong, L. Izatt, M. Tischkowitz, R. Eeles, R. Davidson, S. Hodgson, S. Ellis, C. Nogues, C. Lasset, D. Stoppa-Lyonnet, J.P. Fricker, L. Faivre, P. Berthet, M.J. Hooning, L.E. van der Kolk, C.M. Kets, M.A. Adank, E.M. John, W.K. Chung, I.L. Andrulis, M. Southey, M.B. Daly, S.S. Buys, A. Osorio, C. Engel, K. Kast, R.K. Schmutzler, T. Caldes, A. Jakubowska, J. Simard, M.L. Friedlander, S.A. McLachlan, E. Machackova, L. Foretova, Y.Y. Tan, C.F. Singer, E. Olah, A.M. Gerdes, B. Arver, H. Olsson, Risks of breast, ovarian, and contralateral breast cancer for BRCA1 and BRCA2 mutation carriers, JAMA 317 (23) (2017) 2402–2416.