Noname manuscript No.
(will be inserted by the editor)

Cohort and Trajectory Analysis in Multi-Agent Support Systems for Cancer Survivors

Gaetano Manzo[1] · Davide Calvaresi[1] · Jean-Paul Calbimonte[1] · Oscar Jimenez-del-Toro[1] · Michael Schumacher[1]

Received: date / Accepted: date

Abstract In the past decades, the incidence rate of cancer has steadily risen. However, earlier and accurate identifications have increased cancer survival chances, allowing us to consider it as a chronic condition. Nevertheless, post-surgery therapies lack patients' adherence, mostly due to a lack of personalized solutions supporting them daily. Persuading patients to follow the medical indications correctly is crucial to extend and enhance their life quality.

This paper proposes a cohort and trajectory analysis (CTA) module for estimating patient survival probability based on patient health records to be employed within an agent-based personalized chatbot system (named EREBOTS). EREBOTS leverages on the CTA to monitor the progression of the patients' conditions, tailor recommendations to the patient, and support the medical personnel's analysis and treatment adjustment. Moreover, it provides a dedicated interface to fine-tune the chatbot behaviors based on patient trajectory analysis. The development of such assistive tool enables to *(i)* effectively evaluate the significance of prognostic variables (e.g., death or cancer recurrence), *(ii)* detect patient's high-risk markers, *(iii)* support treatment decisions, and *(iv)* improve the patients' treatment adherence.

Keywords Trajectory and cohort analysis · Multi-agent systems

1 Introduction

Breast cancer is the most common cancer in women worldwide and the second leading cause of cancer death in women [1]. From 2013 to 2018, medical advancements reduced the death rate by 1% per year, increasing the cancer survival rate up to 90% after five years from the diagnosis [2].

[1]University of Applied Sciences and Arts Western Switzerland (HES-SO)
Switzerland
name.surname@hevs.ch
Institut Informatique de Gestion, HES-SO Valais-Wallis, 3960 Sierre, Switzerland

According to the World Health Organization (WHO), the improvement of treatment adherence would be more beneficial to the patient's health than the development of new drugs [3]. However, the correct identification of some cancers' stages and evolution is still a challenging task [4]. Unfortunately, inaccurate staging systems often lead to insufficient or unnecessary treatments [5]. Therefore, the development of assistive technologies that *(i)* effectively evaluate the significance of prognostic variables (e.g., death or cancer recurrence), *(ii)* facilitate the detection of patient's high-risk markers, *(iii)* support treatment decisions, and *(iv)* improve the patients' treatment adherence, is imperative.

Scientific and technological research aims to bridge medical and healthcare personnel with the patients. In the eHealth domain, an increasing number of studies deal with patient persuasion leveraging on chat-based interactions [6]. In particular, agent-based applications have been successfully employed in smoking cessation [7], psychology support for patients post-cancer surgery [8], and healthy lifestyle [9]. These approaches have shown that individuals participating in such studies have exchanged an extremely high volume of messages —inconceivable to be manually processed by the medical personnel. However, despite the complexity of existing conversational agent systems (e.g., from menu-based to NLP-base), no existing chatbot-based solutions provide decision support for physicians based on the analysis of patient trajectories. Without these tools, clinicians are eventually incapable of fine-tuning the chatbot behaviors and personalizing the intervention according to individual patients' covariates.

The contributions of this paper are listed as follows:

(i) it introduces a cohort and trajectory analysis (CTA) approach for estimating patient survival probability based on patient health records.

(ii) it proposes the employment of an agent-based personalized chatbot system (named EREBOTS) that leverages on the CTA to monitor the disease progression, support treatment adjustment, and provides a dedicated interface for clinicians to fine-tune the chatbot behaviors.

(iii) it provides empirical evidence of the effectiveness of trajectory analysis for providing insightful prediction and classification results in the context of breast cancer survivor patient support.

The rest of the paper is organized as follows. Section 2 presents the state of the art followed by the open challenges in Section 3. The multi-agents framework EREBOTS is detailed in Section 4, together with its components, behaviors, and interfaces. The EREBOTS cohort and trajectory analysis module is presented in Section 5, whereas Section 6 provides its main results. Finally, Section 7 presents the discussions, whereas Section 8 concludes the paper.

2 State of the art

2.1 Chatbots for cancer survivors

Chatbots have been successfully used for supporting individuals fighting chronic diseases, addiction, or metabolic disorders. This is due to both the immediateness, scalability and availability of chatbot services, as well as their increasingly engaging interaction capabilities. Moreover, thanks to their inherent impersonal nature, patients tend to disclose personal health information (e.g., diet and sexuality) more easily to a chatbot as compared to a human [10].
In the context of cancer survivors, chatbot technology is highly supported [11]. For instance, Belfin et al. [12] designed a basic chatbot with a state machine that answers common questions collected from online forums and organized in a knowledge graph. Such a solution is similar to chatbots employed in customer management handling Frequent Asked Question (FAQ). Chatbots have also been used to upgrade current follow-up procedures. Piau et al. [13] implemented a bot whose primary objective is data collection (i.e., temperature and adherence to the therapy). The system was tested for 7 weeks, recording 52 sessions with 9 patients, registering a general acceptance towards the chatbot. Greer et al. [8] confirmed that technology can be an effective vector to reach young adults with positive psychology stimulation, which proved to reduce psychosocial distress associated with medical conditions. The authors recruited 45 individuals within 5 years of completing active cancer treatment and interfaced them with a chatbot, named Vivibot, for 4 weeks. Such a chatbot shared positive psychology skills, daily emotion ratings, videos, and other sources produced by survivors leveraging on cognitive and behavioral interventions [14]. As a result, the patients reported a sensible reduction of anxiety and depression. However, the authors acknowledged that no AI-based content-personalization mechanism had been employed, which could have boosted the impact of the study. Chaix et al. [11] proposed Vik, a health care chatbot supporting breast cancer survivors. Vik answers the patients' concerns during their convalescence, contributes to medication adherence (via reminders and educational content), and shows medication tutorials and side effects. Yet, Vik does not provide the handover-to-doctors functionality nor real-time monitoring.

Although these works have shown the potential use of chatbots for supporting fragile cancer survivors, the challenge of personalizing interactions and interventions remains open. The usage of AI-powered models built upon patient information and their trajectories have been explored in order to fill this gap as seen in the next section.

2.2 Cancer survivor prediction models

Survival models evaluate the significance of prognostic variables in outcomes such as death or cancer recurrence, informing clinicians and patients of their treatment options [15]. The Kaplan–Meier estimator is one of the most used survival analysis models for cancer patients [16]. It allows establishing an estimation of the survival function from lifetime data, taking into account censored data and estimating lost event occurrence at a patient's follow-up [17]. On the other hand, the Cox Proportional Hazard (CPH) model [18] is a standard method that can be adjusted with patient covariates using linear combinations [19].

In the past few years, researchers have developed non-linear models, based on deep learning architectures, to the problem of survival analysis [20]. In particular, they focused on neural networks (NN) for classification tasks [21], event estimations [22], and risk prediction [23]. Those neural networks learn highly complex and nonlinear relationships between prognostic features and individuals risks. Moreover, the NN learns the relationship without prior feature selection or domain expertise, providing personalized recommendations based on the computed risk of treatment. However, previous studies have demonstrated mixed results on predicting risk, failing to demonstrate improvements beyond the linear Cox model [24, 25].

3 Opportunities and Open Challenges

The studies previously presented intersect several disciplines and domains including patient trajectory analysis, conversational agents, and eHealth patient support. The opportunities arising from the combined synergy of these areas are: *(i)* the dissemination of health information and coaching instructions; *(ii)* the collection of patient data to enable profiling, personalized coaching, monitoring, and adherence boosting interactions; *(iii)* the incentive of positive behavioral change; *(iv)* the support of persuasive strategies for self-efficacy evaluation.

Nevertheless, the following open challenges/issues still need to be addressed:

C1 Dynamic personalization: cancer survivors participate in social campaigns and data collection where informative contents are broadcast. However, the information they receive presents a minimal level of personalization. Generating personalized content relying on the combination of patients' trajectories combined with patients' behavioral information is still an open challenge.

C2 Continuous healthcare supervision: cancer survivors have a labile mental and physical balance. Nevertheless, besides automated messages and monitoring/reporting functionalities, no existing chatbot-based system operating in this context provides interaction means for the healthcare professionals. Having an interface dedicated to the medical personnel that can be used to fine-tune the chatbot behaviors or directly connect with the patient is still an open challenge.

C3 Evolving models & behaviors: agent-based chatbots can model users comprehensively. However, the sociological dynamics and implications can quickly change, and current solutions cannot model evolving behaviors in the complex dynamics of current frameworks.

C4 Dynamic persuasive techniques: current techniques are rather predetermined (i.e., depending on a rule-set). The challenge is to consider patients' conditions with respect to their trajectories and adapt treatments according to the expected model outcome.

C5 Continuous adherence monitoring: when chatbots are employed, the adherence is mainly computed elaborating user surveys. Therefore, the challenge relies on computing and understanding adherence at run-time, identifying the elements possibly responsible for the divergence.

C6 Privacy compliance: current systems propose human-made data management, privacy, and visibility descriptions. The actual compliance of such extracts with the system behavior is not inferrable. Therefore, the challenge is two-folded: *(i)* equipping the system with dynamics, empowering the patient with full control over his data. *(ii)* generating system extracts for mirroring and describing system behaviors that deal with the patients' data.

Cancer survivors could concretely benefit from the accomplishment of such challenges. Equipping chatbots with behaviors bridging patients' trajectories and persuasive techniques can support eHealth systems, which are facing the strain of a significant demand for patient empowerment. Therefore, employing agent-based models and techniques can facilitate achieving the above-mentioned challenges.

4 Architecture of the Agent-based Chatbot Platform

To address the challenges mentioned above, we rely on an agent-based chatbot platform named EREBOTS. The multi-agent-based architecture of the platform allows autonomous execution of personalized behaviors towards patients and isolated management of personal data. The EREBOTS platform comprises four main components: Database management, Communication Server, Multi-agent system back-end for the doctor agents, and Patient agents' back-end and front-end. Each of these components is deployed on a dedicated Docker container and managed via Docker Compose.

Figure 1 schematizes the main functional interactions among the components mentioned above.

- The *Database* component manages two types of information: *(i)* system-related (non-personal) data, managed through *MongoDB*, and *(ii)* user personal data, which includes consent information, chatbot interactions, demographics and other data input by the user. Personal data is managed using *Pryv*[1].

[1] `https://www.pryv.com/`. A GDPR-compliant platform enabling data-stream-based collection and privacy data-visibility management.

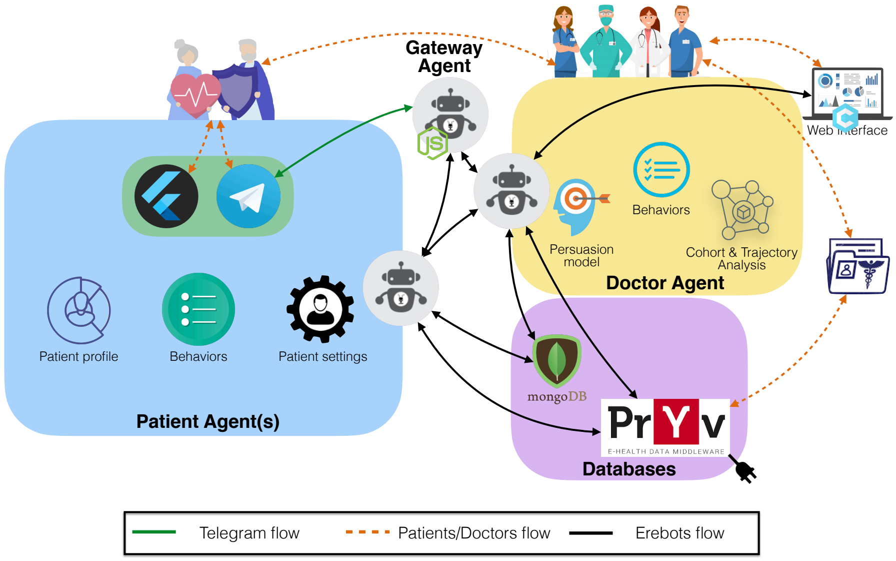

Fig. 1 Main components of the EREBOTS architecture.

- The *Communication server* for inter-agent communication within the Multi-Agent Systems (MAS) uses Prosody[2], an XMPP server instance. Agents can broadcast or unicast messages using this platform.
- The *Doctor agent* is designed to autonomously manage a campaign, including the type of interactions defined for the patients and the monitoring of their activities. It organizes the patients' data (i.e., merges behavioral information and medical examination), elaborates patients trajectories based on machine learning models, updates the forecast trends, and enables further analysis (e.g., patient treatment adherence, results, and persuasive interventions). Overall, the Doctor agent is characterized by three building-blocks: *Persuasion models* to foster the Patient(s) behavioral change, the agent set of *Behaviors*, and the *CTA* module.
- The *Patient agent* manages the patients' connections and their messages from the chat platform(s). It deals with registration and user interactions, such as data-reporting, requests of support, and settings management and proactive interactions. Although extensible to other messaging systems, the framework currently supports the following communication interfaces: (i) *Telegram*[3]*:* a widely used free messaging application for mobile phones released in 2013. (ii) *HemerApp:* a dedicated front-end based on Flutter[4], a framework for native multi-platform development. Moreover, three more building-blocks characterize the Patient Agent: data and models composing the *Patient profile*, the set of *Behaviors*, and the *Patient Settings*.

[2] https://prosody.im/
[3] https://telegram.org/
[4] https://flutter.dev/

A dedicated mobile- and web-app is directly connected to the MAS, while Telegram (*passing through* dedicated APIs) requires a *gateway agent*.

5 Model for Cohort and Trajectory Analysis

In our approach, agents deployed in EREBOTS take either the patient or the doctor (healthcare provider) role, autonomously managing the interactions produced and received through the chatbot messages. Within the doctor agent, we propose the inclusion of the cohort and trajectory analysis model, whose purpose is to provide decision-support information for clinicians regarding risks, symptoms, and disease associations.

Figure 2 illustrates the general scheme of the trajectory and cohort analysis process. Specifically, trajectories represent the patient's evolution from the diagnosis of the disease. The CTA requires EHRs data, provided by the doctor-agent, and behavioral data, provided by the patient-agent. Through the analysis of the trajectories, it is possible to identify associations between symptoms and events (e.g., admission, re-admissions, and treatments) and to quantify risks. Moreover, the trajectory analysis enables us to define several paths of events that may occur sequentially representing the transition between one event (e.g., primary tumor detection) and the other (e.g., metastasis) as a probability of occurrence learned from data. Finally, these results allow the identification of high-risk markers for detrimental treatment effects (e.g., depression and anxiety disorder), subsequent cancer disease, and metastatic cancer disease.

In the following sections, we provide details regarding the methods used for the CTA. More specifically, we describe: *(i)* the estimation of survival probability based on observed patient features; *(ii)* Machine Learning-based classification of patients according to their vital or relapse-free status; and *(iii)* clustering methods used to group patients into cohorts based in observed trajectory data.

5.1 Trajectory Estimation

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, which aim to answer the question: "what is the probability that a patient survived 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 $t = \infty$. 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., $S(t_1) <= S(t_2), \forall t_1 >= t_2$). In case the study

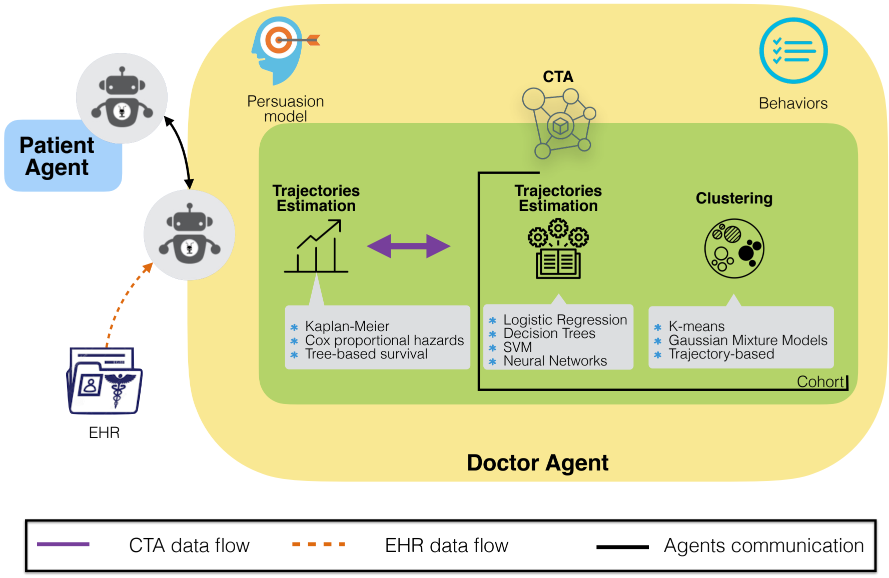

Fig. 2 Trajectory and Cohort Analysis Architecture in EREBOTS

ends or the patient is withdrawn from it, the data is considered *censored*. Given a dataset with patients observing time and event outcome, we are enabled to estimate the survival curve through the Kaplan-Meier Estimator [17]:

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

With d_i and n_i the number of patients that had an event at time i 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. The Kaplan-Meier estimator is calculated considering the notion that the probability can be broken up into the product of probabilities during specific intervals. 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 time t?".

The Cox Proportional Hazard model provides the tool to estimate individual risks as follows:

$$\lambda(t) = \lambda_0 e^{(factor)} \tag{2}$$

where t is the observation period and λ_0 is the baseline risk. Whereas, the $factor$ in 2 identifies the way of modeling patient features (e.g., age, tumor stage, and treatments) to estimate patient risk. In this work, given its performance in the numerical evaluation, we define the factor risk as a linear combination of the patient's features $X = (x_1, x_2, .., x_n)$ and the respective features' weights $\Theta = (\theta_1, \theta_2, .., \theta_n)$, with n the number of patient features.Therefore,

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

Please notice that the survival function 1 is strictly related to the hazard function, as follows:

$$S(t) = e^{-\int_0^t \lambda(u) du}, \tag{4}$$

and vice versa

$$\lambda(t) = -\frac{S'(t)}{S(t)} \tag{5}$$

Finally, to address the non-linearity between the features, we use Tree-based risk models. 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, which provide high-performance levels and interpretations of their results.

5.2 Survival Classification

The survival and risk approaches described above provide the means for doctor agents to build a comprehensive trajectory model that can be used to support clinical decisions. Complementary to these features, we propose incorporating classification prediction capabilities, which may help understanding an individual trajectory based on similar healthcare records.

To identify common patterns of patient data, we define a classification task based on the use of several machine learning and deep learning models. These survival classifiers identify the relationship of the features based on the patient outcome event (e.g., death), while providing interpretable results. They take as input the data of the patients (e.g., cancer type, tumor stage, and NPI) and provide as output the label group to which the patient belongs, such as the patient's vital status or relapse-free status.

Notice that this first step to cluster patients with similar features requires a training phase (i.e., supervised learning), contrary clustering approaches presented later (i.e., unsupervised learning). To prepare the input for the classification task, a data pre-processing is made that includes data retrieval, data cleaning, and data wrangling. Moreover, to prevent under- and over-fitting issues, a study of the learning curves is made followed by 10-fold cross-validation. Technical details and results are included in Section 6.

The classification itself starts first with the task of finding the event probability of the dependent binary variable (outcome), e.g., to be alive or deceased.

Using Logistic Regression, a linear relationship is learnt from the given dataset and then introduces non-linearity through the activation function, such as ReLu or Sigmoid. To understand the decision boundary of the classification task (i.e., the range values in which a patient belongs to a group), we use Support Vector Machine, also known as SVM or Large Margin Classifier. Our SVM model catches the non-linear relationship among the patient's features by using several kernel functions and by applying a regularization term to overcome high-bias and high-variance issues. Decision trees, random forests, and stochastic gradient boosting are the tree-based models used for the classification as well for the trajectory analysis tasks. Those models handle continuous and categorical features, outliers, and missing data. Ensemble learning methods such as Random Forest combines numerous decision trees providing as output the mean from every individual tree. Random Forest models bring about a dissimilarity measure among the observations and unlabelled data, providing a more accurate output. To conclude our analysis, we employ the Stochastic Gradient Boosting model, which creates numerous decision trees in an incremental error-correcting process. Finally, as the last classifier, we use a deep learning model —Neural Networks, which represent the state of the art of survival analysis and show high accuracy for datasets with a large population and features.

5.3 Clustering

While in the previous section the classification methods targeted prediction on predefined trajectory outcomes, in this section we investigate different clustering approaches for the patient cohort analysis. Such unsupervised approaches benefit from the trajectory analysis and overcome the limitations of the classification task, which is based on the outcome of events such as the vital or the relapse status of the patient. Using clustering, a doctor agent may be able to find similar trajectories without pre-defined assumptions about their individual characteristics.

To fine-tune groups of patients with similar outcomes but different features (e.g., patient with tumor stage 1 and patient with tumor stage 3, both alive), we use K-Means, Gaussian Mixture Models, and Trajectory-based clustering algorithms. The K-means algorithm randomly finds k clustering centers (or centroids), and then iteratively groups the data point to the nearest clustering center according to the deviation until the change of all clustering centers converges. K-means clustering minimizes within-cluster variances, but not regular Euclidean distances. However, k-means clustering tends to find clusters of comparable spatial extent, while the expectation-maximization mechanism allows clusters to have different shapes. For this reason, we use the Gaussian Mixture model, representing the presence of sub-populations within an overall population. The Gaussian Mixture model corresponds to the mixture distribution that represents the probability distribution of observations in the overall population. Finally, trajectory-based clustering is the measurement of trajectory similarity (or distance) is one of the key points in defining trajectory clustering, grouping similar trajectories into the same cluster, and finding the most common patterns.

6 Model Evaluations

To evaluate our cohort and trajectory analysis approach, we have applied it in the context of breast cancer support. The fundamental principle is that these models can autonomously provide personalized prediction of risks probabilities, as well as classification of trajectory patterns, so that they can be incorporated into an EREBOTS doctor agent. In the following, we describe first the dataset used during the evaluation, as well as the pre-processing steps, followed by the trajectory and cohort analysis results.

6.1 Dataset & Pre-processing

In order to train and evaluate our models, we use the METABRIC dataset (Molecular Taxonomy of Breast Cancer International Consortium [26]). The METABRIC dataset consists of gene expression data and clinical features for 2498 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"*. METABRIC includes 32 features such as age at diagnosis, type of breast surgery, and ER status. Moreover, the dataset presents the number of patient's months of relapse-free status and overall survival status, respectively with a median relapse time of 99 months and survival time of 116 months. The mean age at diagnosis is 60, with the youngest patient at the age of 21 and the oldest at the age of 96.

In order to prepare the data for the patient trajectory and cohort analysis, we performed a data wrangling process. The first step is to select the features to be analyzed and to drop the insignificant ones. In our models, we consider all the features present in the dataset, and we drop only those with single values, which do not add any extra information to our analysis. In the second step, we handle the different feature types such as continuous and categorical. For the missing values of continuous features, we impute the median values for each feature to keep steady their statistical properties. Then, we transform features by scaling each of them to the range $[0, 1]$ by using a min-max scaler. For the categorical features, in order not to add bias by imputing extra information, we consider an extra class "None" for the missing data. Finally, we perform one-hot encoding for transforming the categorical feature into numerical. Please notice that the one-hot encoding could drastically slow the process by including extra features. However, given the low number of features in the METABRIC dataset, this encoding process does not affect our system performance.

6.2 Trajectory Analysis

In this section, we illustrate the main results of our trajectory analysis. We start analyzing the METABRIC dataset using the Kaplan-Meier estimator (KM).

In Figure 3, we report the KM estimation of the overall METABRIC breast cancer population for the event of patient's status (Alive or Dead). The plot shows the overall survival probability of the breast cancer population over time (months) and the related Greenwood confident interval of 95%. We see that 80% of the population survived at least 50 months, whereas 60% for at least 10 years.

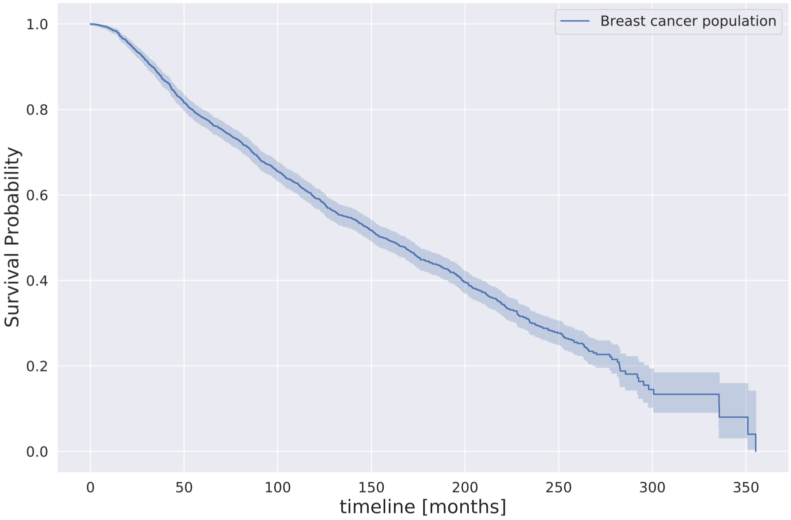

Fig. 3 Kaplan-Meier survival probability estimation of the overall breast cancer population in METABRIC.

Enhancing the granularity of the KM estimator, Figure 4 shows the KM estimation of the breast cancer population grouped by tumor stage. We can see the impact of the tumor stage on the survival probability. Indeed, while a stage 2 tumor patient has a similar trend to the overall KM estimation above presented, patients with a stage 4 tumor have a huge drop in the survival probability estimation. According to our estimation, 80% of patients with a tumor stage 2 survived at least 50 months, whereas only 40% of patients with tumor stage 4 survived at least 50 months and just a few percentages survived more than 10 years.

In Figure 5, we report the KM estimation of the breast cancer population grouped by cancer type. In particular, we arranged 4 types of cancers: Invasive Ductal, Invasive Lobular, Mixed Ductal-Lobular, and other types that represent less than 5% of our dataset. From our KM estimation, we notice that all

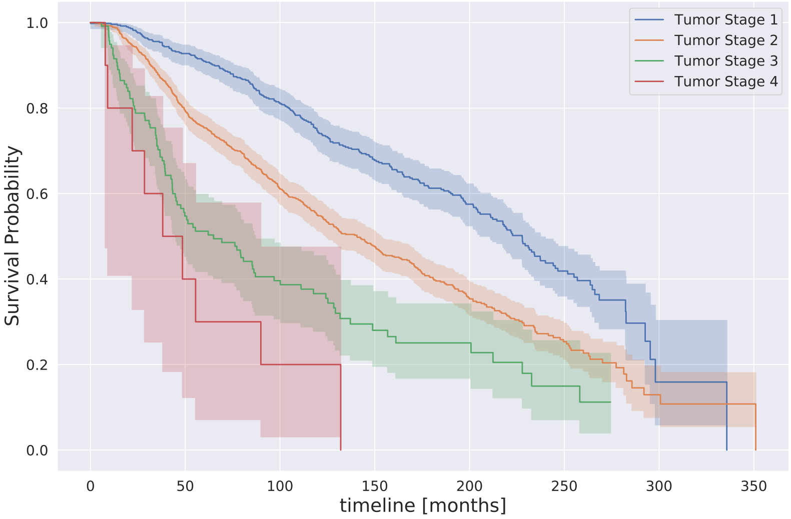

Fig. 4 Kaplan-Meier survival probability estimation of the breast cancer population in METABRIC grouped by tumor stage.

types of breast cancer in our dataset report the same KM estimation except for the group "Others", which reports the best survival probably. However, given the small population and the numerous missing data, the KM estimation for the group "Others" presents a huge confident interval and a survival probability drop at 250 months.

A significant difference in the survival probability concerns the type of surgery. Figure 6 shows the KM estimation for the breast cancer population grouped by two surgeries: mastectomy and breast-conserving. We can see that breast-conserving surgeries provide a higher survival probability than mastectomy surgeries even for patients with similar tumor stages. Indeed, in Figure 6, we report the KM estimation of both surgeries for patients with stage 2 tumors (given its similarity with the overall breast cancer population KM estimation). We notice that tumor stages do not affect the survival trajectory estimation based on the type of surgery.

A significant discrepancy in the survival estimation probability concerns the menopause status of the patients, which is strongly related to the patient's age. Figure 7 shows the KM survival estimation of the breast cancer patients grouped by menopause status: pre-menopause and post-menopause. As Figure 7 shows, patients post-menopause present higher risks than patients pre-menopause due particularly to the elder breast cancer population in the post-menopause group.

As a support to the relationship between patient's age at diagnosis and patient's risk, in Figure 8, we present the Cox Proportional Hazard based on the survival analysis of the breast cancer population in METABRIC. Figure 8 illustrates the relationship of the features with the *log* of the hazard function

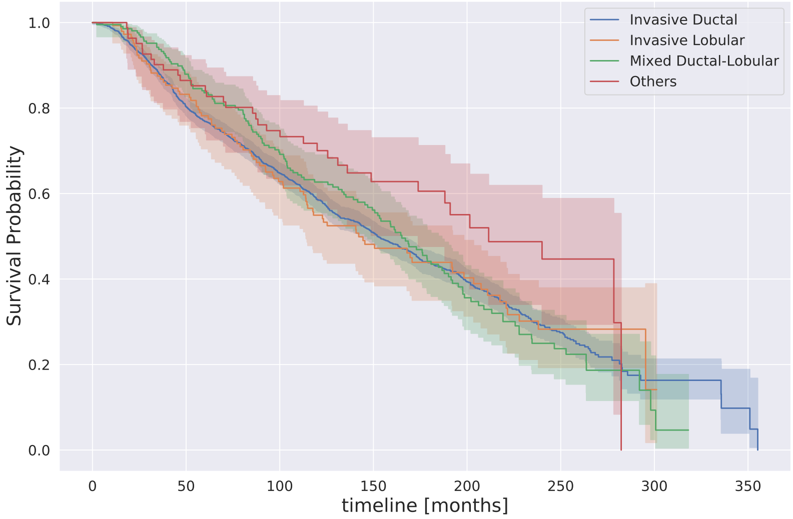

Fig. 5 Kaplan-Meier survival probability estimation of the breast cancer population in METABRIC grouped by cancer type.

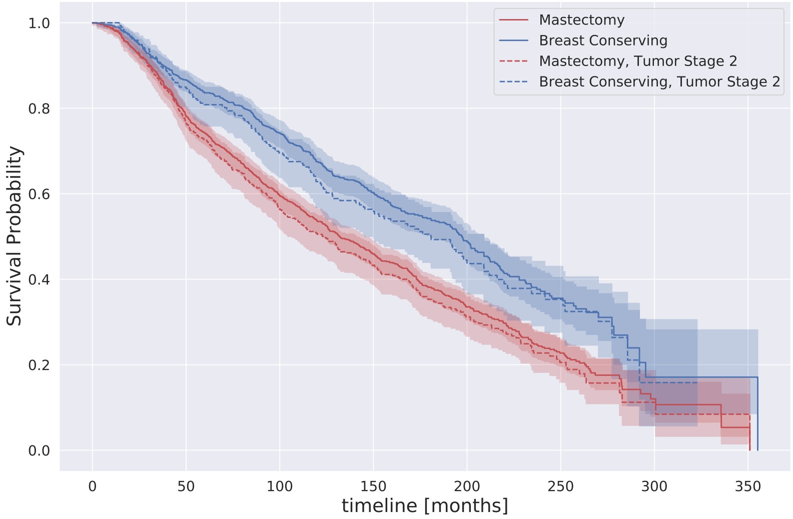

Fig. 6 Kaplan-Meier survival probability estimation of the breast cancer population in METABRIC grouped by surgery type and tumor stage.

presented in Section 8 and the respective confident interval of 95%. Please note that the defined hazard function is exponential (see equation 3), therefore, the relationship between the features and the *log* of the hazard function is linear. In such a plot, a positive relationship means higher risk (e.g., Tumor Stage, Lymph nodes examined positive, and PR Status). On the other hand, a

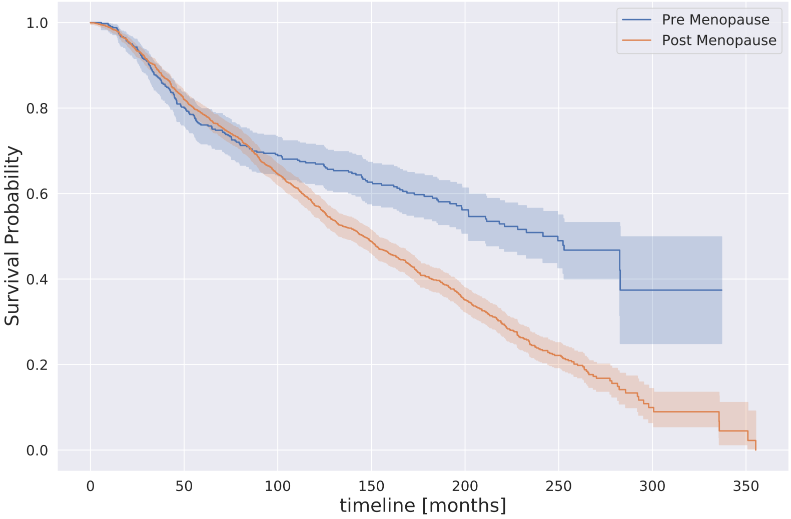

Fig. 7 Kaplan-Meier survival probability estimation of the breast cancer population in METABRIC grouped by menopause status.

negative relationship means lower risk (e.g., Integrative Cluster, HER2 Status, and ER by IHC). As shown in Figure 8, the patient's age at the diagnosis is strongly related to higher risk (i.e., the associated weight to the feature is the highest). On the other hand, Relapse Free Status is strongly related to lower risk (i.e., the associated weight to the feature is the lowest).

6.3 Cohort Analysis

In this section, we illustrate the main results of our cohort analysis using the METABRIC dataset. As mentioned, we use supervised and unsupervised machine learning algorithms such as Logistic Regression, SVM, Decision Tree, and Neural Networks.

We start analyzing the learning curves of the different classifiers shown in Figure 9. The figure illustrates the accuracy-test 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 the training set increases. In particular for the Neural Network model, which requires many examples to train its neurons. However, models such as Decision Tree and Logistic Regression provide the best accuracy level even for a small number of train examples.

Another important characteristic is the time to perform the cohort analysis. Indeed, while the trajectory analysis does not require any training phase, the cohort could imply some delay in the EREBOTS framework.

Figure 10 shows the training time required for each classifier based on the trainset size (i.e., number of patients used for training). As shown in the figure,

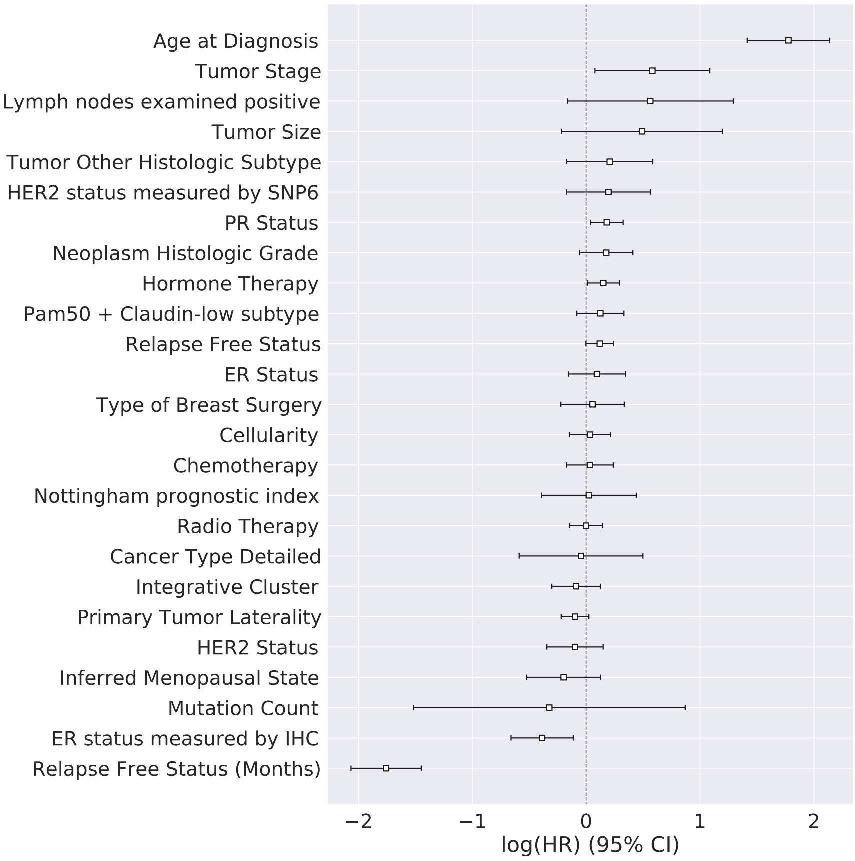

Fig. 8 Cox Proportional Hazard based on the survival analysis of the breast cancer population in METABRIC.

Decision Tree outperforms other classifiers for the specific task of survival classification. Logistic Regression provides the best accuracy but not the best fitting time. Finally, the Neural Network classifier, given our hardware, needs nearly half-second for the training 80% of the METABRIC dataset (about 1600 patients).

In Figure 11, we show the Gaussian Mixture model performing the cohort analysis based on 10 random patient trajectories. We selected three main areas: high-risk, medium-risk, and low-risk. On the one hand, patients in the low-risk area have shown high survival probability. On the other hand, patients in the high-risk area have shown low survival probability. The Gaussian Mixture 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.

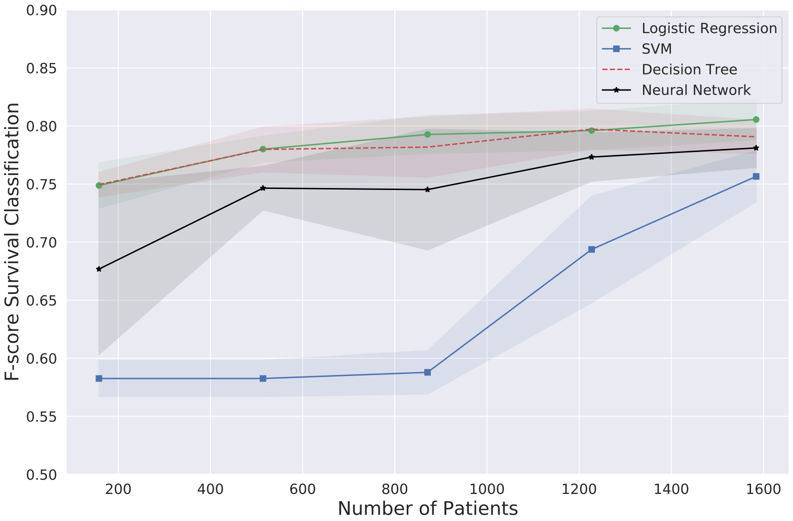

Fig. 9 F-score for survival classification using Logistic Regression, SVM, Decision Tree, and Neural Networks.

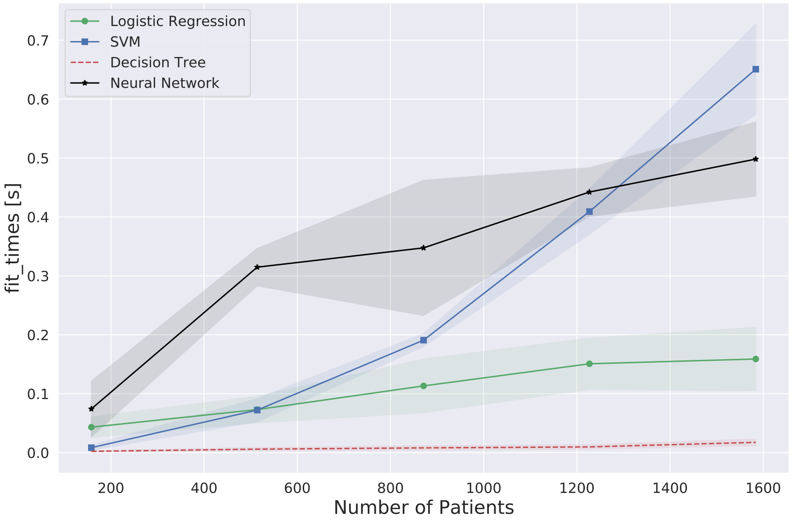

Fig. 10 Scalability for survival classification using Logistic Regression, SVM, Decision Tree, and Neural Networks.

7 Discussion

The CTA enables EREBOTS to provide dynamic personalization of the interactions and story-line addressing the challenge C1. Indeed, the CTA generates personalized content relying on the combination of patients' aggregated data

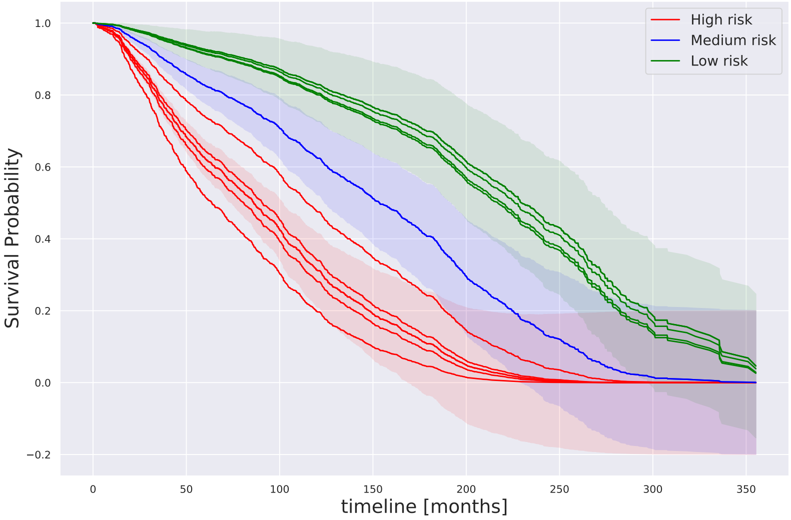

Fig. 11 Gaussian Mixture model Patients cohort based on their trajectory.

(i.e., trajectories) combined with patients' behavioral information (e.g., conduct in the social campaigns). However, the medical personnel needs functionalities that go beyond the in-chat personalization/differentiation. Therefore, as ongoing work, we are investigating how to dynamically integrate user-groups dedicated to enriching the chatbot interface (HemerApp) and its interactions.

By monitoring and reporting high-risk markers, the CTA provides support for the medical personnel addressing the challenge of continuous healthcare supervision (C2). Besides automated messages and reporting functionalities, EREBOTS provides an interaction means to the doctors. In particular, the medical personnel can use EREBOTS to fine-tune the chatbot behaviors and retrieve personalized patient information and risks. Moreover, EREBOTS provides an initial set of tools to monitor in real-time the running campaign. We plan to extend our mechanisms with logic-based triggers to involve medical personnel proactively when needed.

In EREBOTS, the CTA can model the users quite comprehensively, addressing the challenge of evolving models & behaviors (C3). Indeed, the CTA enriches the patient's dataset by including the sociological dynamics, evolving behaviors, and continuous risk estimation in EREBOTS. Moreover, the user modeling and knowledge representation can be dynamically reshaped to satisfy possibly different investigations/campaigns. Currently, it is possible to execute diverse campaigns in parallel. Nevertheless, the integration of contextually diverse knowledge is an ongoing work.

The CTA can trigger an adjustment of the patient's therapy, addressing the challenge of dynamic persuasive techniques (C4). Indeed, the CTA provides support to the medical personnel, who can evaluate and perform changes in the therapy/story-line from a personalized chatbot. Moreover, if the action is not life-threatening, it can be suggested autonomously by the chatbot without clinicians' direct intervention (ongoing work).

By collecting users' feedback related to the tasks conducted within the application, EREBOTS aims at tracking/enhancing the user quality of experience (QoE), addressing the challenge of continuous adherence monitoring (C5). Indeed, by identifying features responsible for the user divergence, the CTA enables EREBOTS to understand user adherence at run-time. As ongoing work, we are studying the automation of such feedback classifications and placing autonomous logic triggers for sensitive feedback raising the attention of the personnel managing a given campaign.

Finally, concerning **C6** (privacy compliance), EREBOTS employs Pryv as a privacy-compliant stream-based database. Pryv can expose its data using semantically rich representations [27] and use standard vocabularies (e.g., HL7 FHIR). Moreover, when the platform is deployed, an automated behavior composes an informative scrutinizing all the agents' behaviors within the system and collects *which data* is used for *which purpose* and visible to *who*. If a new behavior is added into EREBOTS or an existing one is modified, the information is updated accordingly.

8 Conclusion

This work coped with the challenge of personalized agent-based chatbots as virtual assistants for breast cancer survivals and clinicians' support. In such context, we presented EREBOTS framework and its Cohort and Trajectory Analysis module for continuous healthcare personnel supervision, evolving models and behaviors, and multi-stakeholder personalized therapy. Such a module has been tested using METABRIC dataset, which allowed us to discuss the extent of satisfaction of the above-mentioned challenges.

Overall, *(i)* the CTA enables EREBOTS to personalize mainstream interaction story-lines for dynamic personalization. *(ii)* By monitoring and reporting high-risk markers, the CTA provides support for the medical personnel for continuous healthcare supervision and prognosis. *(iii)* Our agent-based chatbots by using the CTA can model the users comprehensively for evolving patient models and behaviors. *(iv)* The CTA can trigger an adjustment of the patient's treatments for dynamic persuasive techniques. *(v)* By collecting users' feedback related to the tasks conducted within the application, EREBOTS enhances user quality of experience (QoE) for continuous adherence monitoring. *(vi)* Finally, EREBOTS employs Pryv as a privacy-compliant stream-based database for privacy compliance.

Conflict of interest

The authors declare that they have no conflict of interest.

References

1. C Fitzmaurice, C Allen, RM Barber, L Barregard, ZA Bhutta, H Brenner, DJ Dicker, O Chimed-Orchir, R Dandona, L Dandona, et al. Global burden of disease cancer collaboration global, regional, and national cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life-years for 32 cancer groups, 1990 to 2015: a systematic analysis for the global burden of disease study. *JAMA Oncol*, 3(4): 524–548, 2017.
2. Bayu Setiaji and Ferry Wahyu Wibowo. Chatbot using a knowledge in database: human-to-machine conversation modeling. In *2016 7th International Conference on Intelligent Systems, Modelling and Simulation (ISMS)*, pages 72–77. IEEE, 2016.
3. Eduardo Sabaté, Eduardo Sabaté, et al. *Adherence to long-term therapies: evidence for action.* World Health Organization, 2003.
4. David N Louis, Arie Perry, Guido Reifenberger, Andreas Von Deimling, Dominique Figarella-Branger, Webster K Cavenee, Hiroko Ohgaki, Otmar D Wiestler, Paul Kleihues, and David W Ellison. The 2016 world health organization classification of tumors of the central nervous system: a summary. *Acta neuropathologica*, 131(6):803–820, 2016.
5. Dong Wook Kim, Sanghoon Lee, Sunmo Kwon, Woong Nam, In-Ho Cha, and Hyung Jun Kim. Deep learning-based survival prediction of oral cancer patients. *Scientific reports*, 9(1):1–10, 2019.
6. Fabien Dubosson, Roger Schaer, Roland Savioz, and Michael Schumacher. Going beyond the relapse peak on social network smoking cessation programmes: Chatbot opportunities. *Swiss medical informatics*, 33(00), 2017.
7. Davide Calvaresi, Jean-Paul Calbimonte, Fabien Dubosson, Amro Najjar, and Michael Schumacher. Social network chatbots for smoking cessation: agent and multi-agent frameworks. In *2019 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pages 286–292. IEEE, 2019.
8. Stephanie Greer, Danielle Ramo, Yin-Juei Chang, Michael Fu, Judith Moskowitz, and Jana Haritatos. Use of the chatbot “vivibot” to deliver positive psychology skills and promote well-being among young people after cancer treatment: randomized controlled feasibility trial. *JMIR mHealth and uHealth*, 7(10):e15018, 2019.
9. Ahmed Fadhil and Silvia Gabrielli. Addressing challenges in promoting healthy lifestyles: the al-chatbot approach. In *Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare*, pages 261–265. ACM, 2017.
10. Gale M Lucas, Jonathan Gratch, Aisha King, and Louis-Philippe Morency. It’s only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior*, 37:94–100, 2014.
11. Benjamin Chaix, Jean-Emmanuel Bibault, Arthur Pienkowski, Guillaume Delamon, Arthur Guillemassé, Pierre Nectoux, and Benoît Brouard. When chatbots meet patients: one-year prospective study of conversations between patients with breast cancer and a chatbot. *JMIR cancer*, 5(1):

e12856, 2019.

12. RV Belfin, AJ Shobana, Megha Manilal, Ashly Ann Mathew, and Blessy Babu. A graph based chatbot for cancer patients. In *2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS)*, pages 717–721. IEEE, 2019.
13. Antoine Piau, Rachel Crissey, Delphine Brechemier, Laurent Balardy, and Fati Nourhashemi. A smartphone chatbot application to optimize monitoring of older patients with cancer. *International journal of medical informatics*, 128:18–23, 2019.
14. Judith T Moskowitz, Adam W Carrico, Larissa G Duncan, Michael A Cohn, Elaine O Cheung, Abigail Batchelder, Lizet Martinez, Eisuke Segawa, Michael Acree, and Susan Folkman. Randomized controlled trial of a positive affect intervention for people newly diagnosed with hiv. *Journal of consulting and clinical psychology*, 85(5):409, 2017.
15. Robert W. Yeh, Eric A. Secemsky, Dean J. Kereiakes, Sharon-Lise T. Normand, Anthony H. Gershlick, David J. Cohen, John A. Spertus, Philippe Gabriel Steg, Donald E. Cutlip, Michael J. Rinaldi, Edoardo Camenzind, William Wijns, Patricia K. Apruzzese, Yang Song, Joseph M. Massaro, Laura Mauri, and for the DAPT Study Investigators. Development and Validation of a Prediction Rule for Benefit and Harm of Dual Antiplatelet Therapy Beyond 1 Year After Percutaneous Coronary Intervention. *JAMA*, 315(16):1735–1749, 04 2016. ISSN 0098-7484. doi: 10.1001/jama.2016.3775. URL `https://doi.org/10.1001/jama.2016.3775`.
16. Mevlut Ture, Fusun Tokatli, and Imran Kurt. Using kaplan–meier analysis together with decision tree methods (c&rt, chaid, quest, c4. 5 and id3) in determining recurrence-free survival of breast cancer patients. *Expert Systems with Applications*, 36(2):2017–2026, 2009.
17. E. L. Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282): 457–481, 1958. doi: 10.1080/01621459.1958.10501452.
18. David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.
19. Anny Xiang, Pablo Lapuerta, Alex Ryutov, Jonathan Buckley, and Stanley Azen. Comparison of the performance of neural network methods and cox regression for censored survival data. *Computational statistics & data analysis*, 34(2):243–257, 2000.
20. K. Liestøl, P. K. Andersen, and U. Andersen. Survival analysis and neural nets. *Stat Med*, 13(12):1189–1200, Jun 1994.
21. P. Andersson, J. Johnsson, O. Björnsson, T. Cronberg, C. Hassager, H. Zetterberg, P. Stammet, J. Undén, J. Kjaergaard, H. Friberg, K. Blennow, G. Lilja, M. P. Wise, J. Dankiewicz, N. Nielsen, and A. Frigyesi. Predicting neurological outcome after out-of-hospital cardiac arrest with cumulative information; development and internal validation of an artificial neural network algorithm. *Crit Care*, 25(1):83, Feb 2021.

22. J. M. Jerez, L. Franco, E. Alba, A. Llombart-Cussac, A. Lluch, N. Ribelles, B. Munárriz, and M. Martín. Improvement of breast cancer relapse prediction in high risk intervals using artificial neural networks. *Breast Cancer Res Treat*, 94(3):265–272, Dec 2005.
23. E. Biganzoli, P. Boracchi, and E. Marubini. A general framework for neural network models on censored survival data. *Neural Netw*, 15(2):209–218, Mar 2002.
24. S. Bussy, R. Veil, V. Looten, A. Burgun, S. Gaïffas, A. Guilloux, B. Ranque, and 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*, 19(1):50, 03 2019.
25. L. Mariani, D. Coradini, E. Biganzoli, P. Boracchi, E. Marubini, S. Pilotti, B. Salvadori, R. Silvestrini, U. Veronesi, R. Zucali, and 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*, 44(2):167–178, Jun 1997.
26. C. Curtis, S. P. Shah, and riadis Chin. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486 (7403):346–352, Apr 2012.
27. Jean-Paul Calbimonte, Fabien Dubosson, Ilia Kebets, Pierre-Mikael Legris, and Michael Ignaz Schumacher. Semi-automatic semantic enrichment of personal data streams. 2019.