Tel-Aviv University

Raymond and Beverly Sackler Faculty of Exact Sciences

The Blavatnik School of Computer Science

Early detection of prostate gland and breast cancer risk based on routine check-up data using survival analysis trees for left-truncated and right-censored data

Thesis submitted in partial fulfillment of graduate requirements for

The degree "Master of Sciences" in Tel-Aviv University

School of Computer Science

By

Dan Coster

Prepared under the supervision of

Prof. Ron Shamir

January 2021

Acknowledgments

I would like to extend my sincere thanks to the people who helped me make this thesis become a reality.

Foremost, I would like to express my profound gratitude to my outstanding supervisor Prof. Ron Shamir. Ron encouraged me to explore the fields I was interested in and completely supported me during my journey. I feel fortunate to have been advised by a 'hall of fame' researcher, and great personality, who made me feel inspired on every day of our mutual work from his distinguished thorough persistence for doing excellent science and role model personality. I would also like to express my deep gratitude to him for helping me present my work in domestic venues and abroad and for the financial support I was granted along the way.

Secondly, I would like to thank Eyal Fischer and Shani Shenhar-Tzarfaty for contributing out of their vast experience since the very first day of working on the project.
Thirdly, I would like to thank the great advisors for assisting me and supplying great advice always with patience and true willingness to help - Prof. Shlomo Berliner, Prof. Malka Gorfine and, Prof. Eran Halperin.

I would like to deeply thank my friends from the ACGT group: Ron Z., Dvir, Lianrong, David, Tom, Nimrod, Hagai, Yael, Naama, Omer, Hadar, Yonatan, Dan, Eran Roi, and Maya for being my mates during my scientific way.

A special thanks to Gillit Zohar-Oren for her endless support, always accompanied by kindness and patience.

I would like to thank the agencies that supported my thesis research: Edmond J. Safra Center for Bioinformatics at Tel Aviv University, Israel Science Foundation (ISF) grant No. 1339/18; ISF grant No. 3165/19, within the Israel Precision Medicine Partnership program; grant 2016694 from the US - Israel Binational Science Foundation (BSF), and the US National Science Foundation (NSF), and Google.

Last but not least, I wish to deeply thank my parents Aliza and Boaz that foster my eagerness for inquisitiveness, nurture my meticulousness, and cultivate my confidence. Also to my brother Eran for being my best friend and backer, and to my love Daniel for being my partner for life. I wouldn't have done it without you all.

Abstract

In this study, we aimed to predict breast cancer and prostate gland cancer risk among healthy individuals by analyzing routine laboratory measurements, vital signs, and age. We analyzed electronic medical records of 20,317 healthy individuals who underwent routine checkups, encompassing more than 600 parameters per visit, and identified those who later developed cancer. We developed a novel ensemble method for risk prediction of multivariate time series data using a random forest model of survival trees for left-truncated and right-censored data.

In cross-validation, our method predicted future cancer six months before diagnosis, achieving an area under the ROC curve of 0.62 ± 0.05 for prostate gland cancer and 0.6 ± 0.03 for breast cancer. This performance was better than the standard random forest, Cox-regression model, and a single survival tree. Our method can complement existing screening tests such as clinical breast examination and mammography for breast cancer, and help in detection of subjects that were missed by these tests. We hope that such computational analysis of results of routine checkups of healthy individuals can improve the detection of those at risk of cancer development.

.

Table of Contents

1. Introduction

Early detection of cancer is crucial for providing appropriate care to the patient and can improve both prognosis and survival [1–4]. The current detection strategies use cancer type specific screening tests that require substantial resources and their performance is limited, e.g., serum Prostate-Specific Antigen (PSA) level for Prostate Gland Cancer (PGC), mammography, and clinical breast examination (CBE) for detecting early signs of Breast Cancer (BC) [5].

Machine learning algorithms can improve screening models in two major directions. One approach is utilizing advanced algorithms to improve the performance of the existing tests. A second approach aims to develop new cancer risk prediction tools based on historical medical records of patients, collected as part of routine care in Electronic Medical Records (EMR). Moreover, advanced genetic methods are also employed for screening, mostly using polygenic risk scores [6].

Our objective in this thesis was to develop new models for both BC and PGC based on EMR data collected from healthy individuals in routine periodic checkups, using techniques from machine learning and survival trees.

Survival trees were first introduced by Gordon and Olsen [7] and their objective is to partition the covariate space into smaller and smaller nodes containing observations with homogeneous survival outcomes. Later, different ensembles methods for survival trees analysis were suggested [8].

We considered the problem of predicting survival probability over time. Our objective was to create a model based on subjects' time-dependent covariates obtained in routine laboratory tests and to predict the fully personalized survival function for each subject based on the last available measurement values. We developed a novel method called TVsuRF (Time-Varying SUrvival Random Forest) for this goal. TVsuRF is the first ensemble method based on survival trees for time-dependent covariates that implements the ‘pseudo-object’ concept. Moreover, our method is the first to use the conditional inference trees in that setting. We used the new method to predict future BC and PGC risk in healthy individuals. Our analysis was conducted on EMRs of 20,317 healthy individuals

who underwent routine checkups, encompassing more than 600 parameters per visit. We identified those who later developed cancer using the Israel Cancer Registry. We compared our method to other extant methods and obtained favorable results.

Today, screening tests in the healthy population are used to identify individuals with cancer without symptoms, but these tests are costly, labor-intensive, and suffer from low accuracy. Our method aims to utilize existing clinical measurements of healthy individuals to predict the risk of BC and PGC, the most common cancers among females and males, respectively. To the best of our knowledge, this is the first risk score that is based on routine laboratory measurements proposed for these cancer types.

The thesis is organized as follows: Chapter 2 provides basic background on cancer, and Chapter 3 provides the computational background needed for the thesis. Chapter 4 describes the data sources that we used and the cohort formation process. Chapter 5 describes the new method that we developed. Chapter 6 contains the results. We conclude in Chapter 7 with a discussion.

2. **Basic background on cancer**

Cancer is a complex disease that encompasses different diseases with diverse risk factors and prognosis. It originates in different cell types and organs in the body and caused by genetic alterations that accumulate in a normal cell, turning it into a malignant cell. A fraction of these malignant cells is characterized by extensive proliferation that leads to the formation of tumors that penetrate normal tissues and in some cases form metastases in distant tissues [9].

2.1. **Prostate Gland Cancer**

PGC is the most common cancer among males, with more than 190,000 cases in the US [10] and more than 2,500 cases in Israel annually [11]. Approximately one out of eight men will be diagnosed with prostate cancer at some point during their lifetime, with a median age at diagnosis of 66 years. Several risk factors are positively associated with PGC, such as age, black ethnicity, and family history of prostate cancer [12].

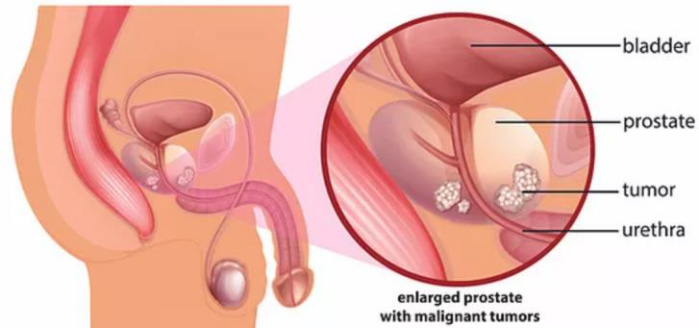


Figure 1: An anatomical illustration of the prostate gland tumor location. (Figure source: www.jamaicamoves.com/single-post/2020/02/04/All-About-Prostate-Cancer-in-Jamaica)

The current early detection strategies are based on simple screening tests that require substantial resources, e.g., digital rectal examination (DRE) and the prostate-specific antigen (PSA) blood test. Given an abnormal finding in one of these tests, an individual may undergo a targeted prostate biopsy [13]. A modern risk score that tries to incorporate several risk factors is the Prostate Cancer Prevention Trial Risk Calculator (PCPTRC) [14]. It is based on race, age, PSA, family history of PGC, rectal examination result, and prior biopsy, but its performance is relatively limited with a median AUC of 0.568 on

discriminating between low-grade tumors and no cancer. Another method aims to improve the PGC risk score based on longitudinal PCPTRC results [9].

Other innovative blood tests aim to assist the physician before the decision of extraction biopsy is taken. 4K [15] is a score based on a combination of free, total, intact PSA, and human kallikrein 2 (hK2) with age, DRE, and prior biopsies information. The PHI test is a formula of pro, free, and total PSA. Other novel urinary based measurements are selectMDx, which is based on mRNA levels of the genes *HOXC6*, *DLX1*, *KLK3,* and MiPS which is based on the combined risk score of serum PSA with the genes *PCA3* and *TMPRSS2: ERG* [16].

2.2. **Breast Cancer**

BC is the most common cancer among females, with more than 250,000 cases in the US [17] and more than 4,500 cases in Israel [11] annually. Approximately one out of eight women will be diagnosed with breast cancer at some point during her lifetime, with a median age at diagnosis of 62 years. Several risk factors are positively associated with BC, such as age, first degree BC family history, early age at menarche, late menopause age, BMI, late age at first live birth and, and presence of mutations in the BRCA gene [18].

The current BC screening tests require substantial resources. The most common tests are mammography, an X-ray modality for detecting early signs of BC, and clinical breast examination (CBE), a physical examination done by a physician to recognize abnormalities in the breast texture [19]. In case of an abnormal finding, usually, another imaging modality is required before a biopsy test. The most common modalities are ultrasound, magnetic resonance imaging (MRI), or another mammography. Following verification of the abnormal finding, a biopsy is taken.

Another approaches to assess BC risk is Gail's model [20,21], which is based on several parameters: age, age at menarche, age at first live birth, number of previous benign breast biopsies, presence of atypical hyperplasia on biopsy, number of affected mother or sisters, and race or ethnicity. More advanced approaches for estimation of BC risk encompass somatic mutations as part of their risk factors such as BRCAPRO [22], IBIS [23], and BOADICEA [24].

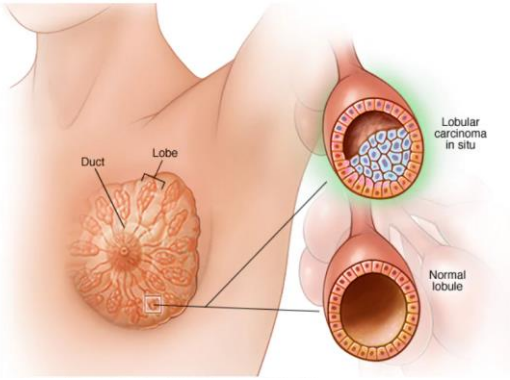


Figure 2: An anatomical illustration of the breast tumor location. (Figure source: www.mayoclinic.org/diseases-conditions/lobular-carcinoma-in-situ/multimedia/lobular-carcinoma-in-situ/img-20008459)

Recently, machine learning algorithms were employed to improve screening models. for example, using deep learning models for analyzing mammography [25–27] or using machine learning models to optimize Gail's model parameters [28] achieving AUC between 0.55 and 0.6.

2.3. **Early detection of Cancer Using EMR data**

EMRs are the digital version of patients' paper charts. EMR systems were designed originally to store data accurately and to enable a longitudinal overview of patient health. The use of EMR has increased dramatically in recent years. The large volume and high-dimensional clinical patient information captured in EMRs may reflect the characteristics of the general population better than those of cohort studies based on a targeted subgroup of limited profiles. Therefore, EMRs provide a unique opportunity to understand the health status at the population level. Different EMR-based models that utilize routine laboratory measurements as part of their input were suggested for cancer risk prediction such as lung cancer [29], colorectal cancer [30], acute myeloid leukemia [31], among others.

A prediction model of the 1-year risk of lung cancer was developed based on EMR data of 873,598 individuals from the Maine Health Information Exchange Network. The model was based on 346 features selected based on correlation in addition to known risk factors (e.g. COPD, previous cancer, age, etc.) and utilized the XGBoost algorithm to achieve AUC of 0.881 [0.873 – 0.889]. The model used EMR data in the preceding six months and

found several risk factors to be most associated with a new incident of lung cancer: age, a history of pulmonary diseases and other chronic diseases, medications for mental disorders, and social characteristics.

A prediction model for colorectal cancer was developed based on the EMR data of the Maccabi database, the second largest HMO in Israel with 2 million insured individuals, and validated on the United Kingdom (UK) Health Improvement network datasets. The cohorts of the study contained 606,403 and 25,613 individuals respectively. The model's input included an individual's demographics (age and sex) as well as the current complete blood count (CBC) and the trends of the various CBC parameters. The output was a combined score of Gradient Boosting and Random Forest models. Using blood counts obtained 3–6 months before diagnosis, the AUC for detecting colorectal cancer was 0.826 ± 0.01 for the Maccabi test set and 0.81 for the UK test set, and found the Hemoglobin and Mean Corpuscular Hemoglobin (MCH) levels as important features of the model.

Another model for acute myeloid leukemia (AML) was developed based on the Clalit database, the biggest HMO in Israel, which contains EMRs of an average of 3.45 million individuals per year and collected over 15 years. In this cohort, 875 AML cases were identified based on ICD-9 codes. The model incorporated only parameters that were routinely documented in EMRs, such as different laboratory measurements, ICD-9 codes of different background disease, age, sex, BMI, and weight. In addition, the trend of the various CBC parameters was added to the model. Utilizing the Gradient Boosting Trees algorithm, the model was able to predict AML 6–12 months before diagnosis with a sensitivity of 25.7% and an overall specificity of 98.2%.

3. **Computational Background**

3.1. **Survival Analysis**

Survival analysis is a statistical method for analyzing data on time to a failure event such as death, heart attack, device failure, disease onset, etc. It is widely used in the medical field.

Let X be a non-negative random variable denoting the time-to-event. The survival function $S(x)$ denotes the probability that the event is later than some specified time x, assuming that $S(0) = 1$. Let $F(x)$ be the cumulative distribution function of the event at time x. Then

$$S(x) = \mathbb{P}(X > x) = \int_x^{\infty} f(u)du = 1 - F(x)$$

S must be a non-increasing function, since survival till a later time is possible only if survival was attained for earlier times. An alternative characterization of the distribution of X is by defining the *hazard function*, which describes the instantaneous rate of occurrence of the failure event at time x, given that the subject survived until that time.

$$\begin{aligned}\lambda(x) &= \lim_{\Delta x \to 0} \frac{\mathbb{P}(x \le X < x + \Delta x | X \ge x)}{\Delta x} \\ &= \lim_{\Delta x \to 0} \frac{1}{\Delta x} \cdot \frac{\mathbb{P}(x \le X < x + \Delta x, X \ge x)}{\mathbb{P}(X \ge x)} \\ &= \lim_{\Delta x \to 0} \frac{\mathbb{P}(x \le X < x + \Delta x)}{\Delta x} \cdot \frac{1}{\mathbb{P}(X \ge x)}\end{aligned}$$

In contrast, the survival density function satisfies:

$$f(x) = \lim_{\Delta x \to 0} \frac{\mathbb{P}(x < X < x + \Delta x)}{\Delta x}$$

We can take the derivative of the survival function and get:

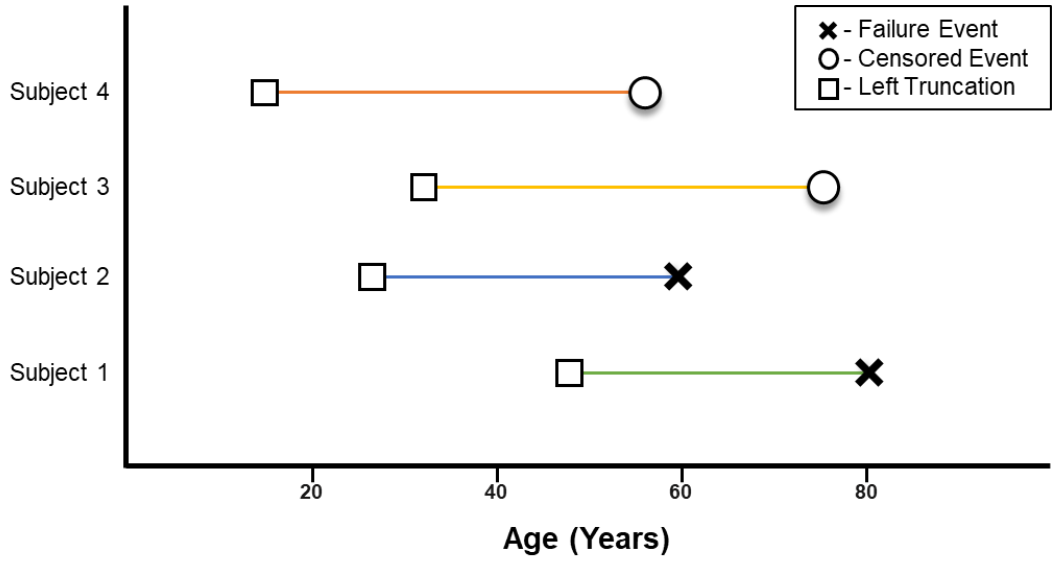
$$S'(x) = \frac{\partial}{\partial x} S(x) = \frac{\partial}{\partial x} \int_x^{\infty} f(u)du = \frac{\partial}{\partial x}\big(1 - F(x)\big) = -f(x)$$

And therefore the two functions are related as:

$$\lambda(x) = \frac{f(x)}{S(x)} = \frac{-S'(x)}{S(x)} = -[\ln S(x)]'$$

A key component of survival analysis is the notion of censoring [32]. Suppose the subject entered the study at time 0 and was followed until time t, at which time follow-up was terminated. At that time, if an event did not occur yet, we say *right censoring* (or simply censoring) happened. Such termination of follow-up can be due to the end of the study, or if the subject dropped out of it. In this case, the exact time the event took place is not known, but we assume that the event will happen at some future time after the censoring time. A second type of censoring is *interval censoring*, which means that the failure event occurred within some known time interval.

Another key concept in survival analysis is left truncation. In studies allowing delayed entry, subjects may enter the study at different time points and not necessarily at time 0. Those subjects entering at t>0 are called *left truncated*. Hence, at time t, only individuals with entry time >t are present in the sample (**Figure 3**).



<u>Figure 3:</u> An illustration of a survival analysis setting for left-truncated and right-censored data. Squares indicate the left truncation times, crosses indicate failure events, and circles indicate right censoring events. Each interval is the time a single subject was part of the study. The x-axis can be age or just time since the study started.

3.2. **The Kaplan-Meier curve**

For a population of individuals, a widely used estimator of the survival function is the Kaplan-Meier curve, which is a non-parametric way to assess both the number of failure events that have occurred as a function of time and the duration of time until a failure event occurs. The survival curve is a step function with jumps at observed failure event times and values held constant for the time between two consecutive observed failure events. **Figure 4** shows a Kaplan-Meier curve. We will present the Kaplan-Meier curve for Left-Truncated and Right-Censored (LTRC) with details in section 5.1.

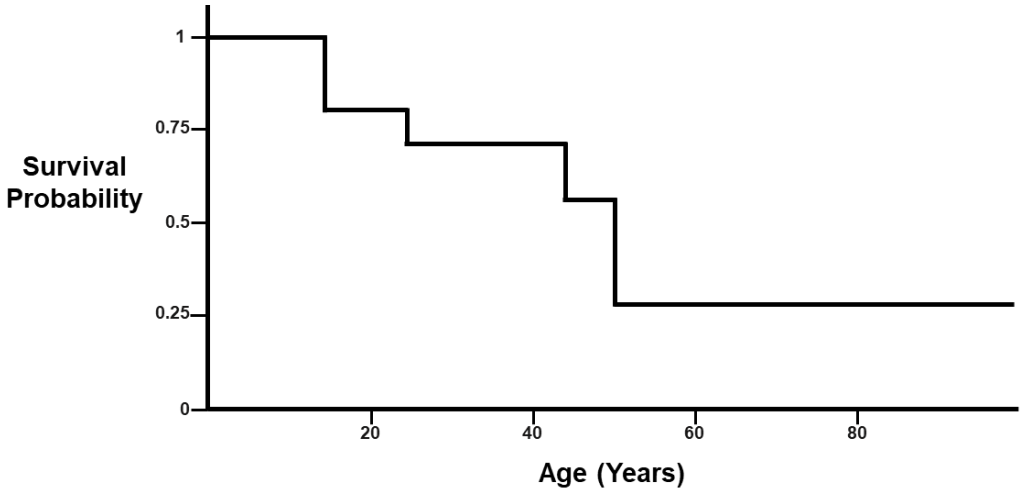


Figure 4: An illustration of the Kaplan-Meier curve. The Y-axis is the survival probability, e.g. the probability to be free of a failure event. X-axis is the age of the subject.

3.3. **Survival Trees**

Survival trees for time-varying covariates: Survival trees were first introduced by Gordon & Olsen [7]. The basic concept is to create a decision tree where each node contains a survival curve of the corresponding subgroup of individuals. The splitting criterion usually aims to maximize the difference in survival between the daughter nodes or the within-node homogeneity. Most of the survival tree methods address right-censored data and time-independent covariates. Incorporating time-varying covariates in survival trees was first introduced by Bacchetti and Segal [33], who suggested the 'pseudo-object' concept. Several methods of constructing survival trees for time-dependent covariates used

this concept [34–37]. Another common approach for analyzing time-dependent covariates is the Cox-regression model [38,39].

An ensemble of survival trees: Several ensemble methods for survival trees analysis were suggested, usually, for time-independent covariates [8,40]. Random survival forests (RSF) were introduced by Ishwaran [41] by combining the concept of RFs of Breiman [42,43], survival trees and the log-rank test as the splitting criteria. An extension of RSF is the utilization of conditional inference trees, which use hypothesis testing to select the splitting covariates and also as a stopping criterion [44]. Further improvements of those methods were demonstrated by Utkin et al. to optimize the weights of each tree [45] and by Steingrimsson et al. using more general weighted bootstrap procedures [46].

Analyzing time-varying covariates to predict an individualized survival curve utilizing an ensemble of survival trees is an emerging field of research [47]. The first method of this kind was for discrete-time data [35]. Another approach utilized martingale equations and ROC-driven splitting criteria [48]. A third one used individualized Bayes estimates of piecewise-constant hazard rates [49].

We considered the problem of predicting survival probability (or equivalently, the probability to be free of a failure event) over time. Our objective was to create a model based on subjects' time-dependent covariates and to predict the fully personalized survival function for each subject based on the last available measurement values. We developed a novel method called TVsuRF (Time-Varying SUrvival Random Forest) that uses longitudinal multidimensional data to predict a personalized survival function.

3.4. **Log-Rank Test**

The log-rank test is a semi-parametric hypothesis test that aims to compare survival functions from two groups with right-censored data. This is an adapted version of the stratified test for 2 X 2 contingency table presented by Mantel [50] and relies on the proportional hazard assumption. Let $S_0(t)$ and $S_1(t)$ be the survival functions of the control and case groups, respectively. Define t_l as the failure time for the l^{th} individual and assume $t_1, \dots t_K$ are distinct failure times. $Y_{i,j}$ denotes the number of individuals who are at risk or who had a failure event at time t_i in group j. $d_{i,j}$ denotes the number of events at time t_i

in group j. d_i and Y_i denote the sums of $d_{i,j}$ and $Y_{i,j}$ among the two groups. Given these variables, for each t_l we can create a 2 X 2 contingency table showing the number of surviving individuals and observed events at that time stratified by group. Assuming that under the null hypothesis $H_0: S_1(t) = S_0(t)$, the distribution of $d_{i,1}$ given the margins of the tables is hypergeometric. Hence the log-rank statistic converges to a normal distribution under H_0 and is calculated as follows:

$$\frac{\sum_{i=1}^{n}\left(d_{i,1} - \frac{Y_{i,1}d_i}{Y_i}\right)}{\sqrt{\sum_{i=1}^{n}\left(\frac{Y_{i,1}}{Y_i}\left(1-\frac{Y_{i,1}}{Y_i}\right)\left(\frac{Y_i - d_i}{Y_i - 1}\right)d_i\right)}} \xrightarrow{D} \mathcal{N}(0,1)$$

3.5. **Random Survival Forest**

The Random Forest (RF) algorithm was introduced by Breiman [42] for classification and regression problems. RF is an ensemble method that trains several decision trees in parallel using bootstrapping. Several individual decision trees are trained in parallel on different subsets of the training dataset utilizing various randomly selected subsets of features. The final predicted outcome is based on an aggregation of the decisions of individual trees (see illustration **Figure 5**).

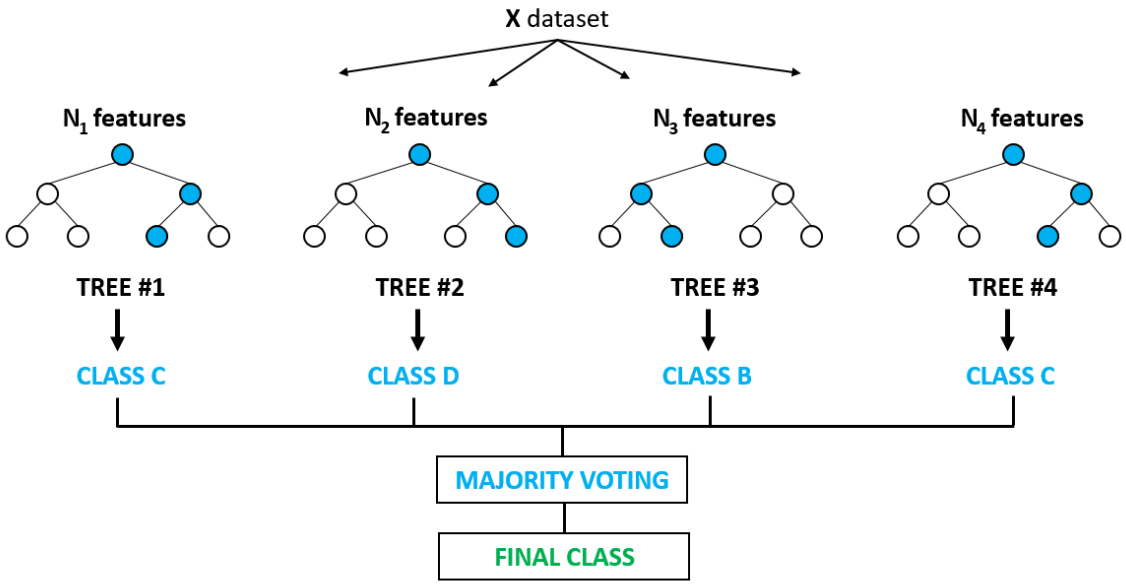


Figure 5: An illustration of the random forest classification algorithm. (Figure source: https://medium.com/towards-artificial-intelligence/use-of-decision-trees-and-random-forest-in-machine-learning-1e35e737b638)

Random survival forests (RSF) were introduced by Ishwaran [41] by combining the concepts of RF, survival trees, and the log-rank test as the splitting criterion for time-independent variables. An advantage of this approach is that it is almost fully non-parametric whereas traditional methods assume a distribution for the survival curve. The description of the algorithm is as follows:

1. Draw n bootstrap samples from the original data set.
2. For each bootstrap sample, grow a tree by repeatedly splitting leaf nodes, as follows. At each such node, randomly select m covariates to split on, and find a covariate c and threshold x value such that splitting the samples of the node according to it maximizes the difference between the survival curves of the daughter nodes as measured by the log-rank test. In other words, we seek a pair (c, x) that yields a split with the largest log-rank score, corresponding to the lowest p-value. Stop splitting when no split produces daughter cells with at least d unique failure events.
3. Estimate the predicted survival function by averaging the results of the n trees (e.g. for each time t, calculate the mean survival probabilities at time t over all the trees).
4. Compute the out-of-bag error of the model.

3.6. **Cox Proportional Hazard Model**

The Cox proportional hazards model is a regression model commonly used in medical research for investigating the association between survival time and categorical or continuous predictor (explanatory) variables. Three main assumptions underline the Cox model. First, censoring must be non-informative or statistically independent of the failure times. Second, the observed failure times are distinct, namely, there are no ties. Third, the model assumes a baseline hazard functions called $\lambda_0(t)$ such that all hazard functions satisfy:

$$\lambda_i(X_i) \;=\; \lambda_0(X_i)\exp(\beta' Z_i)$$

In other words, the survival curves must have hazard functions that are proportional over time. This is called the *proportional hazards* assumption. Cox introduced the key idea of using the partial likelihood approach to estimate the model parameters.

Suppose we have n individuals and we observe data (X_i, δ_i, Z_i) for individual i, where X_i is the event time (failure/censoring) random variable, δ_i is the failure/censoring indicator and Z_i represents a set of covariates. Let $\mathcal{R}(t) = \{i: X_i \geq t\}$ denote the set of individuals who are 'at-risk' for failure at time t, called the 'risk set' (we assume here no right truncation). λ and S are the hazard and survival functions, respectively. It can be shown (see REF) that the full likelihood of the model can be described as:

$$L = \prod_{i=1}^{n} \lambda_i(X_i)^{\delta_i} S_i(X_i) = \prod_{i=1}^{n} \left[\frac{\lambda_i(X_i)}{\sum_{j\in\mathcal{R}(X_i)} \lambda_j(X_i)}\right]^{\delta_i} \left[\sum_{j\in\mathcal{R}(X_i)} \lambda_j(X_i)\right]^{\delta_i} S_i(X_i)$$

Under the proportional hazard model, the partial likelihood (the first term of the full likelihood)does not depend on the underlying hazard function:

$$L(\beta) = \prod_{i=1}^{n} \left[\frac{\exp(\beta' Z_i)}{\sum_{j\in\mathcal{R}(X_i)} \exp(\beta' Z_j)}\right]^{\delta_i}$$

Treating the partial likelihood as a regular likelihood function enables estimation of β using the log transformation, where l_i is the log-partial likelihood contributed by subject i:

$$l(\beta) = \log \prod_{i=1}^{n} \left[\frac{\exp(\beta' Z_i)}{\sum_{j\in\mathcal{R}(X_i)} \exp(\beta' Z_j)}\right]^{\delta_i} = \sum_{i=1}^{n} l_i(\beta)$$

See [51] for full derivation of the above formulas.

4. Cohort Description

4.1. **Dataset**

We analyzed data from routine checkups of individuals at the Tel-Aviv Medical Center Inflammation Survey (TAMCIS), Tel-Aviv Sourasky Medical Center, Israel. Participants were men and non-pregnant women with no active malignant or infectious disease who chose to be tested and signed an informed consent form. In each visit, the subject underwent a comprehensive medical history evaluation, a complete physical examination, blood and urine tests, vital signs measurements, an electrocardiogram, an exercise stress test, and a respiratory function test. Data were summarized in structured EMR. Some individuals had multiple visits during several years. We conducted a retrospective analysis of the TAMICS EMR data collected between November 2001 and February 2017. Our study covered 20,271 adults (age ≥ 18). The study was reviewed and approved by the Institutional Review Board (Approval no. 02-049-Tlv).

4.2. **Cancer Registry**

TAMICS participants who later developed cancer were identified (using their national IDs) in the Israeli National Cancer Registry (INCR), which records all cancer cases in Israel. INCR contains for each case the cancer type (ICD9 code) and diagnosis date, and we used all cancer diagnoses until January 1st, 2016. **Supplementary Figure 1** shows the number of patients in the cohort with each cancer type. We focused on the two cancer types with the largest number of cases: BC for females and PGC for males. Patients who had a different type of cancer prior to diagnosis of BC or PGC were excluded.

4.3. **Exclusion & Inclusion Criteria**

Inclusion criteria: All individuals surveyed in TAMICS who had birth and visit dates documented were included (number of individuals n_p= 20,271, number of visits n_v= 50,497). Of those, individuals with cancer diagnosis according to INCR were identified (n_p= 1,547, n_v= 3,999), along with their cancer type (see **Figure 6**).

Cases: Females whose cancer type was BC (n_p= 293, n_v=730) or males whose cancer type was PGC (n_p= 182, n_v=566).

Controls: Individuals who did not have any cancer diagnosis (n_p= 18,724, n_v= 46,498).

Exclusion criteria: Our analysis was based on data from single visits, so exclusion was done per individual and visit.

Cases: Individuals whose cancer diagnosis date was before their first TAMICS visit (BC: n_p= 94, n_v=223, PGC: n_p= 39, n_v=127). Visits that occurred after the cancer diagnosis date (BC: n_v=87, PGC: n_v=107). Visits where more than 50% of the covariates were missing (BC: n_v=44, PGC: n_v=39). Visits that occurred > 730 days before the cancer diagnosis date (BC: n_p= 122, n_v=286, PGC: n_p= 84, n_v=229).

Controls: Visits where more than 50% of the covariates were missing (n_p= 113 individuals and n_v=6,040 visits excluded). Visits that occurred after the last day of reports in INCR (n_p= 934, n_v=4,214). We split the cancer-free group into male (n_p= 11,360, n_v=24,503), and female (n_p= 6,347, n_v=11,741) subgroups.

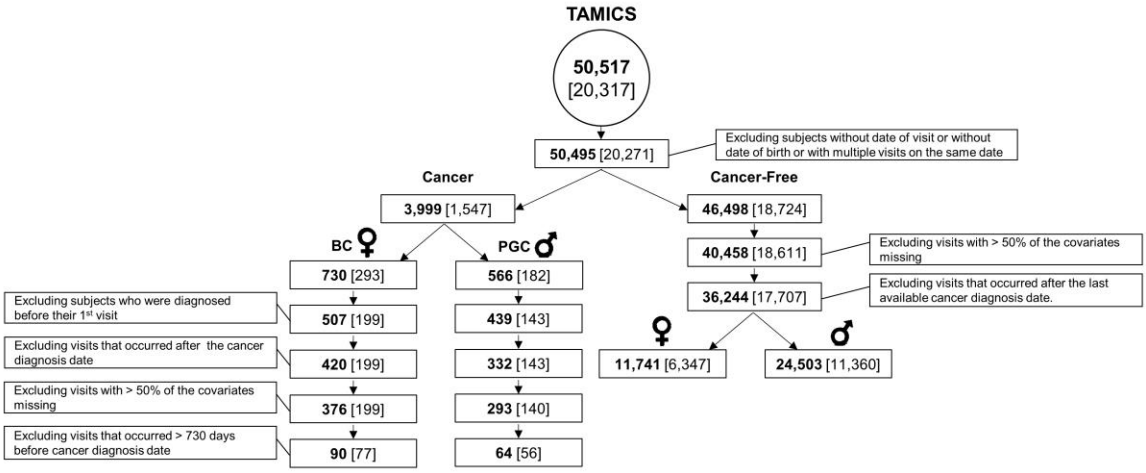


Figure 6: Study design. The bold number is the number of TAMICS visits; the number of individuals appears in parentheses.

4.4. **Data Extraction and Feature Choices**

We used only features that were available for more than 80% of the individuals. The missing values were imputed using Predictive-Mean-Matching on age [52] using the *mice* package [53].

For BC risk prediction we used 20 covariates (**Table 1**) that include demographic parameters such as age and BMI, along with Complete Blood Count (CBC), since BC is a systemic disease that affects the immune system and its progression is expected to be

reflected in the CBC results. For PGC risk prediction, we added 28 covariates that include the Basic Metabolic Panel (BMP), Lipids, Vital Signs, and more. (**Table 2**)

	BC			**BC-Free**			**Matched BC-Free**			**BC vs. BC-Free P-value**		**BC vs. Matched BC-Free P-value**	
Parameter	**Visits**	**Subjects**	**Mean±STD**	**Visits**	**Subjects**	**Mean±STD**	**Visits**	**Subjects**	**Mean±STD**	**T-test**	**MW**	**T-test**	**MW**
Baso (%)	90	77	0.63±0.33	11,739	6,347	0.58±0.29	5,883	3,635	0.59±0.3	1	1	1	1
Eos (%)	90	77	2.61±1.73	11,738	6,347	2.5±1.84	5,882	3,635	2.54±1.78	1	1	1	1
Hmt (%)	90	77	39.06±2.62	11,741	6,347	38.59±2.81	5,884	3,635	38.88±2.86	1	1	1	1
Hgb (g/dL)	90	77	13.2±0.96	11,740	6,347	13.15±0.96	5,883	3,635	13.24±0.96	1	1	1	1
Lym (%)	90	77	30.71±8.26	11,739	6,347	30.75±7.17	5,883	3,635	30.99±7.2	1	1	1	1
Lym (K/μL)	90	77	2.13±0.76	11,734	6,347	2.04±0.57	5,880	3,635	2.01±0.56	1	1	1	1
MCH (pg)	90	77	29.8±2.27	11,740	6,347	29.95±2.04	5,884	3,635	30.04±2.06	1	1	1	1
MCHC(g/dL)	90	77	33.85±0.86	11,740	6,347	34.11±0.98	5,884	3,635	34.08±1.05	0.114	**0.049**	0.344	0.159
MCV (fl)	90	77	87.99±5.62	11,741	6,347	87.75±5.06	5,884	3,635	88.1±5.09	1	1	1	1
Mono (%)	90	77	6.88±1.45	11,739	6,347	6.97±1.91	5,883	3,635	7.12±1.71	1	1	1	1
Mono (K/μL)	90	77	0.48±0.16	11,734	6,347	0.46±0.15	5,880	3,635	0.46±0.13	1	1	1	1
MPV (fl)	87	74	9.19±0.97	11,312	6,234	9.01±1.07	5,688	3,559	9.01±1.08	1	1	1	1
Neu (K/μL)	90	77	4.23±1.42	11,734	6,347	4.06±1.37	5,880	3,635	3.95±1.33	1	1	1	0.739
RBC (M/μL)	90	77	4.45±0.35	11,740	6,347	4.4±0.34	5,883	3,635	4.42±0.35	1	1	1	1
Neu (%)	90	77	59.16±8.63	11,739	6,347	59.21±8.17	5,883	3,635	58.75±8.16	1	1	1	1
PLT (K/μL)	90	77	262.67±52.95	11,740	6,347	263.17±61.56	5,884	3,635	261.35±61.31	1	1	1	1
RDW (%)	90	77	13.42±1.26	11,741	6,347	13.25±1.06	5,884	3,635	13.29±1.02	1	1	1	1
WBC (K/μL)	90	77	7.07±1.84	11,741	6,347	6.77±1.7	5,884	3,635	6.63±1.66	1	1	0.538	0.379
BMI (kg/m^2)	83	71	25.9±4.74	11,273	6,057	25.45±4.72	5,574	3,445	26.23±4.63	1	1	1	1
Age (Years)	90	77	53.46±7.97	11,741	6,347	47.16±10.56	5,884	3,635	53.2±7.66	**< 0.0001**	**< 0.0001**	1	1

Table 1 Characteristics of the BC, BC-free, and Matched BC-free groups. Values are mean ± SD. MW: p-value of the Mann–Whitney test, T-test: p-value of the Student t-test. All p-values were Bonferroni corrected for multiple hypotheses. Baso – basophils; EOS – eosinophils; Hmt – hematocrit, Hgb – hemoglobin; Lym – lymphocytes; MCH- mean corpuscular hemoglobin; MCHC- mean corpuscular hemoglobin concentration; MCV – mean corpuscular volume; Mono-monocytes; MPV- mean platelet volume; Neu – neutrophils; RBC – red blood cells; PLT – platelets; RDW - red cell distribution width; WBC – white blood Cells; BMI – body mass index.

	PGC			PGC-Free			Matched PGC-Free			PGC vs. PGC-Free P-value		PGC vs. Matched PGC-Free P-value	
Parameter	**Visits**	**Subjects**	**Mean+STD**	**Visits**	**Subjects**	**Mean+STD**	**Visits**	**Subjects**	**Mean+STD**	**T-test**	**MW**	**T-test**	**MW**
Baso (%)	64	56	0.57±0.26	24,382	11,344	0.54±0.27	6,080	3,320	0.54±0.27	1	1	1	1
Eos (%)	64	56	2.51±1.32	24,382	11,344	2.86±1.87	6,080	3,320	2.92±1.86	1	1	0.809	1
Hmt (%)	64	56	43.65±2.8	24,390	11,344	43.73±2.7	6,083	3,320	43.71±2.87	1	1	1	1
Hgb (g/dL**)**	64	56	14.93±0.97	24,390	11,344	14.94±0.94	6,083	3,320	14.9±1	1	1	1	1
Lym (%)	64	56	27.52±6.89	24,382	11,344	29.79±6.74	6,080	3,320	28.58±6.78	0.537	1	1	1
Lym (K/μL**)**	63	55	1.8±0.53	24,369	11,269	1.98±0.56	6,079	3,290	1.93±0.59	0.597	0.748	1	1
MCH (pg**)**	64	56	30.33±1.67	24,389	11,344	30.17±1.66	6,083	3,320	30.46±1.76	1	1	1	1
MCHC (g/dL**)**	64	56	34.23±0.79	24,389	11,344	34.21±0.89	6,083	3,320	34.13±0.92	1	1	1	1
MCV (fl**)**	64	56	88.57±4.14	24,390	11,344	88.18±4.28	6,083	3,320	89.25±4.46	1	1	1	1
Mono (%)	64	56	8.06±1.97	24,382	11,344	7.99±1.8	6,080	3,320	8.21±1.86	1	1	1	1
Mono (K/μL**)**	63	55	0.54±0.17	24,370	11,269	0.53±0.16	6,079	3,290	0.56±0.16	1	1	1	1
MPV (fl**)**	63	55	8.87±1.22	23,498	11,257	8.85±1.02	5,899	3,289	8.84±1.05	1	1	1	1
Neu (K/μL**)**	63	55	4.18±1.37	24,368	11,269	4.01±1.28	6,078	3,290	4.14±1.29	1	1	1	1
RBC (M/μL**)**	64	56	4.93±0.38	24,387	11,344	4.97±0.36	6,083	3,320	4.9±0.38	1	1	1	1
Neu (%)	64	56	61.34±8.05	24,382	11,344	58.82±7.52	6,080	3,320	59.75±7.52	0.767	1	1	1
PLT (K/μL**)**	64	56	244.08±80.53	24,389	11,344	238.68±55.85	6,083	3,320	233.5±55.56	1	1	1	1
RDW (%)	64	56	13.34±0.86	24,389	11,344	13.01±0.79	6,083	3,320	13.2±0.84	0.190	0.138	1	1
WBC (K/μL**)**	64	56	6.71±1.66	24,390	11,344	6.75±1.64	6,083	3,320	6.87±1.67	1	1	1	1
Pulse (bpm**)**	59	53	69.95±14.05	23,053	10,896	68.68±11.86	5,591	3,155	68.14±11.7	1	1	1	1
DBP (mmHg)	59	53	81.05±8.26	23,331	10,896	78.66±8.63	5,672	3,155	80.71±8.55	1	1	1	1
SBP (mmHg)	59	53	131.44±15.59	23,326	10,896	125.1±14.32	5,671	3,155	131.08±15.48	0.142	0.099	1	1
Spirometry (Score**)**	56	50	0.34±0.48	22,563	10,716	0.39±0.49	5,435	3,080	0.4±0.49	1	1	1	1
Temp. (C°**)**	59	53	36.34±0.33	22,104	10,947	36.35±0.34	5,397	3,184	36.33±0.33	1	1	1	1
BUN (mg/dL**)**	61	55	16.34±3.75	24,056	11,003	15.36±3.67	6,027	3,195	16.37±4.15	1	1	1	1
Chloride (mmol/L**)**	60	54	104.05±2.53	24,015	10,920	103.52±2.42	6,023	3,160	103.64±2.56	1	1	1	1
Creatinine(mg/dL**)**	60	54	1.15±0.12	24,019	10,920	1.14±0.15	6,026	3,160	1.16±0.16	1	1	1	1
GGT (U/L**)**	60	54	27.57±23.54	23,993	10,920	25.07±22.42	6,018	3,160	26.36±22.21	1	1	1	1
Glucose (mg/dL**)**	61	55	100.18±21.96	24,059	11,003	92.58±16.83	6,030	3,195	97.51±19.7	0.457	0.002	1	1
Potassium(mmol/L**)**	60	54	4.45±0.35	24,019	10,920	4.35±0.37	6,025	3,160	4.37±0.38	1	0.511	1	1
Albumin (g/L**)**	60	54	44.8±2.13	24,014	10,920	45.52±2.32	6,022	3,160	44.82±2.27	0.599	1	1	1
Globulin (g/L**)**	60	54	27.12±3.67	23,995	10,920	28.12±3.2	6,017	3,160	27.98±3.25	1	1	1	1
Phosphorus(mg/dL**)**	60	54	3.16±0.39	24,012	10,920	3.23±0.44	6,022	3,160	3.16±0.43	1	1	1	1
Calcium(mg/dL**)**	60	54	9.35±0.43	24,011	10,920	9.32±0.42	6,021	3,160	9.27±0.43	1	1	1	1
Uric Acid (mg/dL**)**	60	54	6.19±1.12	23,995	10,920	6.09±1.1	6,016	3,160	6.17±1.14	1	1	1	1
Sodium (mmol/L**)**	60	54	141.82±2.91	24,019	10,920	141.19±2.53	6,025	3,160	141.09±2.58	1	1	1	1
Protein (g/L**)**	60	54	71.92±4.18	24,005	10,920	73.64±3.91	6,020	3,160	72.8±3.89	0.118	0.049	1	1
Bilirubin (μmol/L**)**	60	54	0.81±0.37	24,014	10,920	0.83±0.37	6,023	3,160	0.81±0.33	1	1	1	1
ALP (U/L**)**	59	53	63.85±17.3	23,214	10,840	64.64±17.54	5,850	3,131	64.48±17.57	1	1	1	1
LDH (U/L**)**	60	54	323.6±44.04	24,013	10,920	317.76±55.91	6,022	3,160	324.77±55.11	1	1	1	1
Triglycerides(mg/dL**)**	63	56	126.63±56.12	24,207	11,260	123.48±73.12	6,044	3,289	127.33±70.01	1	1	1	1
HDL (mg/dL**)**	63	56	47.42±11.16	24,182	11,260	49.81±10.67	6,036	3,289	50.63±11.54	1	1	1	0.810
LDL (mg/dL**)**	63	56	114.54±28.54	24,095	11,260	115.78±29.83	6,023	3,289	113.03±30.3	1	1	1	1
Cholesterol (mg/dL**)**	63	56	188.27±35.1	24,204	11,260	190.14±34.74	6,043	3,289	189.01±35.08	1	1	1	1
Troponin (ng/dL**)**	63	56	4.11±1.04	24,141	11,260	3.94±0.97	6,026	3,289	3.86±0.9	1	1	1	1
Urine PH	64	56	6.14±0.89	24,134	11,344	6.13±0.82	6,014	3,320	6.1±0.81	1	1	1	1
Urine SG	64	56	1.01±0.01	24,112	11,344	1.01±0.05	6,005	3,320	1.01±0.05	1	1	1	1
BMI (kg/m^2**)**	62	54	27.34±3.29	23,543	11,177	26.88±3.74	5,729	3,266	27.74±3.65	1	1	1	1
Age (Years**)**	64	56	59.61±6.33	24,471	11,344	47.13±10.78	6,102	3,320	59.24±5.77	**< 0.0001**	**< 0.0001**	1	1

Table 2. Characteristics of the PGC, PGC-free, and Matched PGC-free groups. Values are mean ± SD. MW: p-value of the Mann–Whitney test, T-test: p-value of the Student t-test. P-values are Bonferroni corrected for multiple hypotheses. DBP – diastolic blood pressure; SBP – systolic blood pressure; Temp – body temperature; BUN – blood urea nitrogen; GGT – gamma-glutamyl transferase; ALP – alkaline phosphatase; LDH – lactate dehydrogenase; Urine SG – urine specific gravity; Urine PH – PH stick for a urine test.

5. The TVsuRF Method

5.1. **Preliminaries**

Consider a dataset of N subjects, where for each of them data from one or more visits were recorded. Subject i had M^i visits at times $t_1^i < \ ... < t_{M^i}^i$. The d covariates measured at time t_j^i are denoted by the vector $x^i(t_j^i)$ (For simplicity, we assume that all covariates were recorded in every visit). Note that covariates can be either time-dependent or time-independent (static). Hence, $\mathcal{X}^i = (x^i\,(t_1^i), ..., x^i\,(t_{M_i}^i))$ summarizes the longitudinal data of subject i. The last time point subject i was at risk, which can be either failure or censoring time, is $\tau^i > t_{M^i}^i$. $\delta^i \in \{0,1\}$ denotes if the subject experienced a censoring ($\delta^i = 0$) or failure event ($\delta^i = 1$) at time τ^i. Hence, the full data can be summarized by the set of triplets $\mathcal{D} = \left\{\left(\mathcal{X}^i,\ \tau^i, \delta^i\right)\right\}_{i=1}^N$ (**Supplementary Figure 2A**). $\mathcal{X}^i(t)$ denotes the data of subject i that were measured until time t, i.e., $\mathcal{X}^i(t) = \left\{x^i\left(t_j^i\right): 0 \le t_j^i \le t\right\}$. We assume time homogeneity so that w.l.o.g. we can shift times per subject to set $\forall i: t_1^i = 0$, i.e., all first visits were at time 0 (**Supplementary Figure 2B**). We also assume that the age of the subject at each visit is one of the covariates.

Our model aims to estimate the probability for being free of the failure event (the cancer diagnosis) at least until time t based on the patient's covariates at the latest visit before that time. That is, let $t_*^i = \max\left\{t_j^i < t \middle| j\right\}$. We wish to estimate the survival function:

$$S\left(t\middle|x^i\left(t_*^i\right)\right) = \mathbb{P}(\tau^i > t | x^i\left(t_*^i\right), \tau^i > t_*^i)$$

In order to model the time-dependent covariates, we transform the data following [33]. We split the data of each subject into disjoint intervals $[t_j^i, t_{j+1}^i)$ and we assume that the covariates $x^i\left(t_j\right)$ are constant in the interval (**Supplementary Figure 2C)**. In that manner, we consider t_j as the left truncation time. If $[t_j^i, t_{j+1}^i)$ is not the last interval of subject i then we view time t_{j+1}^i as censoring time. We denote the pseudo-object of the j^{th} interval of subject i as $[L_j^i, R_j^i)$ where:

$$L_j^i = t_j^i; R_j^i = \begin{cases} \mathrm{t}_{\mathrm{j+1}}^{\mathrm{i}} & , \quad \text{if } 1 \le \mathrm{j} < \mathrm{M_i} \\ \tau^{\mathrm{i}} & , \quad \text{othetwise} \end{cases}; \; \delta_j^i = \begin{cases} 0 & , \quad \text{if } 1 \le \mathrm{j} < \mathrm{M_i} \\ \delta^i & , \quad \text{othetwise} \end{cases}$$

Hence, the transformation is:

$$\begin{aligned} (\mathcal{X}^i,\ \tau^i, \delta^i) &\to \left\{\left(t_1^i, t_2^i, \delta_1^i, x^i\left(t_1^i\right)\right), \left(t_2^i, t_3^i, \delta_2^i, x^i\left(t_2^i\right)\right), \dots, \left(t_{M^i}^i, \tau^i, \delta^i, x^i\left(t_{M^i}^i\right)\right)\right\} \\ &\equiv \left\{\left(L_1^i, R_1^i, \delta_1^i, x^i\left(t_1^i\right)\right), \left(L_2^i, R_2^i, \delta_2^i, x^i\left(t_2^i\right)\right), \dots, \left(L_{M^i}^i, R_{M^i}^i, \delta^i, x^i\left(t_{M^i}^i\right)\right)\right\} \end{aligned}$$

Each pseudo-interval is therefore possibly left-truncated and/or censored.

The standard Kaplan-Meier (KM) estimator of the survival function can now be generalized for left truncation right-censored (LTRC) data [32], as follows. Assume that there were D failure events and they occurred at distinct times $t_1 < \cdots < t_D$. We denote by Y_j the number of pseudo-objects at risk at time t_j, $Y_j = \sum_{i=1}^{N} \sum_{k=1}^{M_i} \mathbb{I}(L_i^k \le t_j \le R_i^k)$ i.e., the number of individuals who entered the study before time t_j and did not experience a failure or censoring event until t_j. d_j is defined as the number of patients that experienced a failure event at time t_j and due to our prior assumption $d_j = 1$. The KM estimator is defined as a step function with jumps at observed failure times:

$$\hat{S}(t) = \begin{cases} 1 & , \quad if\ t_1 > t \\ \prod_{t_j \le t}[1 - \frac{d_j}{Y_j}], & Otherwise \end{cases}$$

The survival probability will be calculated in a step ahead prediction manner - we calculate the probability of a patient in time t to experience failure in the next time window Δt given its covariates at time t, namely $\mathbb{P}\left(\tau^i < t + \Delta t, \delta_i = 1 \middle| \tau^i > t, x_i(t)\right)$.

5.2. **Survival Tree Construction**

We now describe the construction of the survival tree for pseudo-objects data. For simplicity, we will just call them objects (**Figure 7A**). Suppose we have the set of samples along with their covariates as described above, and we wish to use the survival information to build a decision tree. We use the framework of conditional inference trees [44], a class of decision trees that employs a statistical hypothesis test based on permutations in order to select optimal variables and their thresholds. This process is different from common

decision tree construction (see **Supplementary Material 3**), which usually selects the variable that maximizes an information measure (e.g., Gini or entropy).

A covariate and a threshold value at a node, split the node's samples into two subsets, and each subset induces a survival curve. To compare the survival curves of the two subsets we use Pan's permutations based hypothesis test [54], as suggested also in [37]. In every node, we test all possible covariates and thresholds, and the one that produces the split with the lowest p-value is selected. Notice that pseudo-objects created from the same subject can end in distinct sub-nodes.

The hypothesis test is based on creating an influence function that maps an object's quadruplet $(L_i, R_i, \delta_i, x_i)$ into a scalar U_i which represents the contribution of sample i to the test statistic. We assume that (l_i, r_i) is the interval in which the true event lies, and denote its contribution to the statistic:

$$U_i = \frac{\hat{S}(l_i) log\, \hat{S}(l_i) - \hat{S}(r_i) log\, \hat{S}(r_i)}{\hat{S}(l_i) - \hat{S}(r_i)} - log\, \hat{S}(L_i)$$

One can show that for failure event at time t ($\delta_i = 1$)

$$U_i = \log\left(\hat{S}(t)\right) + 1$$

and for a right-censored observation at time t ($\delta_i = 0$), assuming $\hat{S}(\infty) = 0$

$$U_i = \log(\hat{S}(t))$$

Now let $U_1, \ldots, U_N$ be the scores of the samples corresponding to the parent node, and suppose n samples reside in the left child and $N - n$ in the right. Write $X = \sum_{left} U_j$. There are $\binom{N}{n}$ ways of choosing n out of the N scores and if k of these have a sum $\leq X$, then assuming all partitions are equi-probable, the probability of obtaining a score of X is $P_{value} = \frac{k}{\binom{N}{n}}$. We estimate it using 1,000 permutations.

The survival function $\hat{S}_l(t)$ for node l is the Kaplan-Meier curve for the samples corresponding to that node. Let C_l be the set on indices of samples in node l, then:

$$\hat{S}_l(t) = \prod_{i \in C_l : t_i \le t} \left(1 - \frac{d_l(t_i)}{Y_l(t_i)}\right)$$

Where $d_l(t_i)$ is the number of failure events that occurred at time t_i in node l and $Y_l(t_i)$ is the total number of objects at risk just before t_i in node l (**Figure 7B**, **Figure 8**).

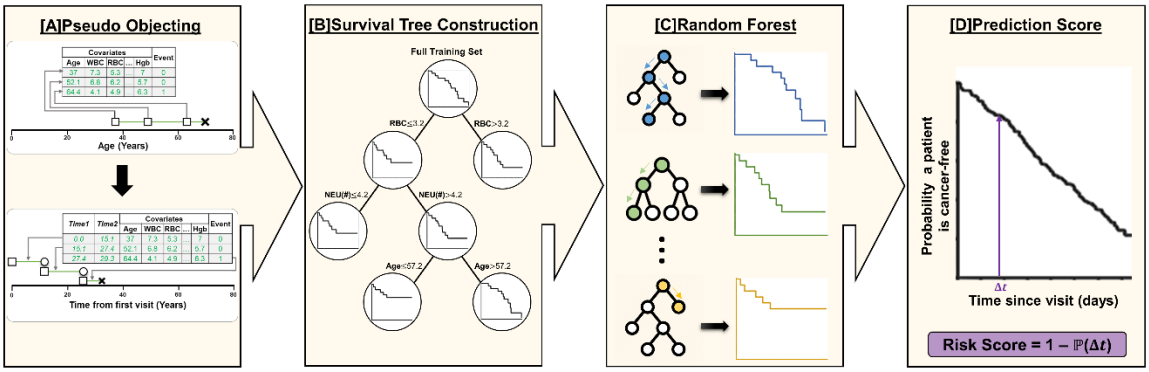


Figure 7: Model construction and evaluation. An illustration of the different parts of our model construction. [A] For each subject we transformed its data into pseudo-objects and change the time axis to time from first visit. [B] An illustration of single survival tree construction [C] Generating 500 survival trees. [D] The trees are combined into a single unified model. Risk score calculation per each sample is based on averaged survival curve.

Algorithm 1: BuildTree (D,K)

Input: Survival data set $D = \{(L_j^i, R_j^i, \delta_j^i, x^i(L_j^i)\}_i^n$, parameter K ;
randomFeatures ← random subset of K features
minP-Value ← ∞
minFeature ← NULL
for *feature in randomFeatures* **do**
  featureUniqueValues ← all the unique values of the feature
  for *val in featureUniqueValues* **do**
    1. D_l,D_r = induced sub-datasets from D based on (val, feature) ;
    2. **P-value** ← LogRankScore(D_l,D_r) ;
    if *P-value* < *minP-Value* **then**
      minP-value ← P-value;
      minFeature ← feature;
      featureVal ← val;
    end
  end
end
if *minPvalue* > 0.05 **then**
  break;
else
  D_l,D_r = induced sub-datasets from D based on (featureVal, minFeature) ;
  BuildTree(D_l,K);
  BuildTree(D_r,K);
end

Algorithm 2: TVsuRF

Input: Survival data set $D = \{(L_j^i, R_j^i, \delta_j^i, x^i(L_j^i)\}_i^n$, number of features per node K, number of trees M ;
minP-Value ← ∞
H ←$\emptyset$;
for $m = 1$ *to M* **do**
  1. h_m ← BuildTree(D,K)
  2. $H \leftarrow H \cup \{h_m\}$
end
return H

Figure 8: Algorithm 1: BuildTree Algorithm. Algorithm 2: TVsuRF Algorithm.

5.3. **Ensemble Model**

We create $M = 500$ survival trees. In each tree, at each internal node, we select at random $K = \sqrt{\#\ Features}$ of the features and split the node according to the feature and threshold giving the least p-value for difference in survival, if that difference is significant (**Figure 8**). The predicted survival curve for a new subject ω is based on the data in all the leaves that ω ended in all the trees. Let $C(l_i^k)$ represent the set of indices of the subjects that are in the i^{th} leaf of the k^{th} tree and let $C_F = \cup \{C(l_i^k) | \ \omega \in l_i^k\}$ be the multiset of all the

subjects in these leaves. If $d_i(t_i)$ is the number of failure events in C_F at time t_i and $Y_i(t_i)$ is the number of objects in C_F in risk at time t_i, then the survival function of ω is (**Figure 7C**):

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

Our model constructs a Kaplan-Meier curve per each subject, producing a continuous risk score (RS) over time.

5.4. **Variable Importance**

We assessed the importance of each covariate in our model in two ways. In the first, we counted the fraction of internal nodes in all the trees that were associated with the covariate (i.e. the covariate was used to split these nodes). We call this fraction Vprop; higher Vprop indicates more importance. In the second approach, for each object, we replaced the values of the covariate by random values sampled independently from its original distribution, while keeping the other covariates in their true values, and recomputed the performance with the new data. The difference in the area under the receiver-operator characteristic curve (AUROC) between the original and the modified data was computed and averaged over ten random assignments per each covariate on every fold of the 4-fold cross-validation [41]. We repeated this process 20 times and defined VIMP as the mean difference obtained. Again, higher VIMP indicates more importance.

5.5. **Comparison to BC Screening Tests**

For a subset of the TAMICS females, we had data concerning BC screening. Mammography was available for 6,526 women and Clinical Breast Exam (CBE) was available for 17,958. We excluded women with mutated BRCA genes, those who refused to conduct a CBE, lacked ID, had more than one record per visit, or were diagnosed with another type of cancer (see **Supplementary Figure 4** for study design).

The result of the mammography was provided in free text written by the physician and transformed by us into binary labels (normal/abnormal) by natural language processing of the physician's notes (see **Supplementary Material 5** for details). The CBE result was available as free text written by a physician and four binary values that represent an

abnormal finding in the left/right breast or axilla. We considered the CBE result abnormal if one of the binary values was positive. In case that no values were reported, a breast cancer surgeon reviewed the physician's text and decided if there was a positive finding.

We compared the recommendations of these screening tests to our predictions, in order to evaluate the added value of our approach. We binned the risk scores into deciles and the average risk score was calculated for each subject.

5.6. **Evaluation Approach**

We used TVsuRF and several other models to predict BC and PGC risk on our cohorts. If a subject's covariates were measured at time t, we aimed to predict cancer at time $t + \Delta t$, for values of Δt ranging between 183 and 730 days. Since there might be a delay between the cancer diagnosis time and the time it was reported to the cancer registry, we added ϵ =31 days to Δt. The risk for patient i is thus:

$$RS^i(t, \Delta t) = 1 - \hat{S}\left(t + \Delta t + \epsilon \mid x^i(t)\right)$$

To evaluate the performance of this score for classification, we calculated AUROC, where the positive class is the set of individuals that were diagnosed with cancer during the next $\Delta t + \epsilon$ days as suggested in [55] (but excluding patients censored in $[t, t + \Delta t + \epsilon]$). We also estimated the area under the precision-recall (AUPR) curve.

We performed 20 iterations of 4-fold cross-validation, where in each iteration the partition of patients into folds was done at random. For each of the above measures, we calculated the average and standard deviation.

We compared our method to three others: (1) Cox regression model adapted to time-varying covariates [38,39], (2) single LTRC survival tree as in [37] (denoted LTRCIT), and (3) RF model [42]. Since RF is a classification model, training for prediction was done separately for each time interval Δt, and the class of a subject was positive if the diagnosis of cancer occurred during the next $\Delta t + \epsilon$ days, and negative otherwise. We used 500 trees, and the Gini index as a splitting rule, with the rest of the parameters at the default values in the *ranger* package [56] (**Figure 7D**).

In addition, we compared our method to a RSF model that predicts a survival curve per sample. Since RSF was originally designed for handling time-independent covariates, we adapted it to our setting.

6. **Results**

6.1. **Breast Cancer**

Dataset: Our cohort contained data on 6,424 women with a total of 11,831 visits to TAMICS. Out of those, 77 were diagnosed with breast cancer and had one or more visits less than 730 days before the diagnosis date (90 visits in total). These constituted the positive (BC) group. The covariates that were included in the model were CBC (18 parameters), age, and BMI. The statistics of these values are summarized in **Table 1.**

Women in the positive group were significantly older on average than in the BC-free group and had significantly lower levels of mean corpuscular hemoglobin concentration (MCHC). To reduce the effect of age on our model, we created an age-matched cohort ('Matched BC-Free') using the approach of [57] (3,635 subjects, 5,884 visits). When comparing the BC and the Matched BC-free group (**Table 1**) none of the parameters was significantly different between the groups.

Prediction accuracy: The performance of each of the methods tested, for different time ranges, is summarized in **Figures 9A and 9B**. We also marked the AUROC of Gail's breast cancer risk estimation for 5 years horizon as reported in [58]. TVsuRF had the highest AUPR on every time interval, and the highest AUROC on all intervals except one (though differences were not statistically significant) for 730 days, where Gail's score was best. We also tested two versions of RSF and our model was better for time windows until 273 days in terms of AUPR and AUROC (**Supplementary Figure 6**).

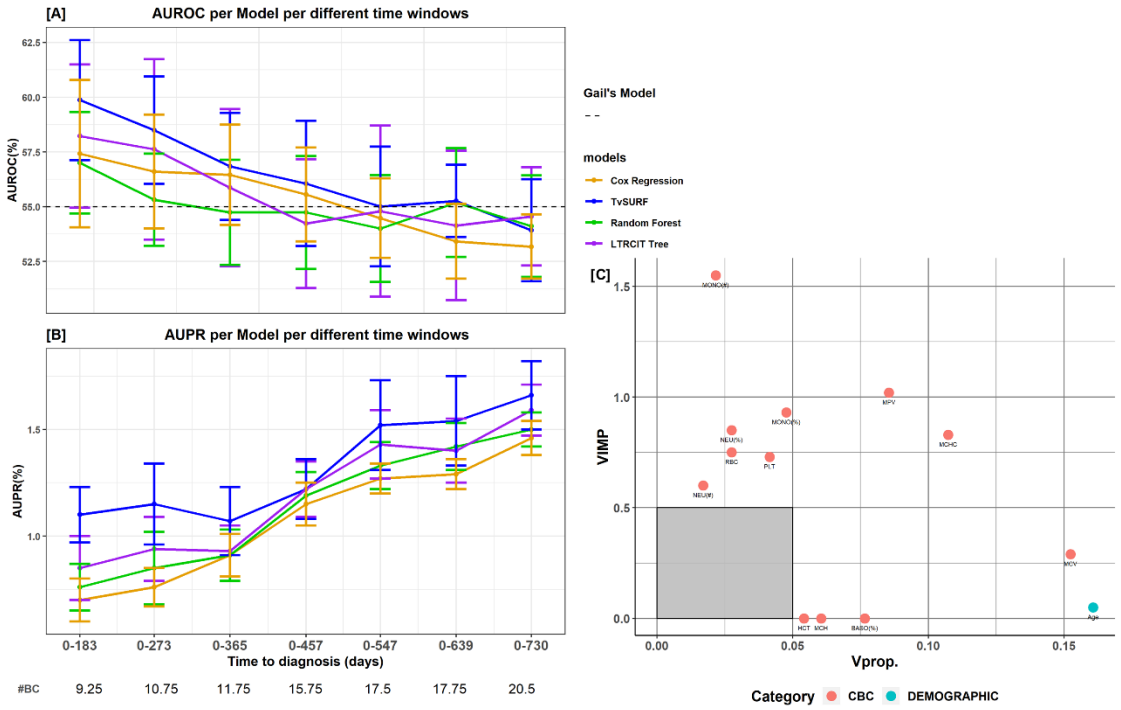


Figure 9: BC risk prediction and variable importance. [A] Performance (AUROC mean±SD) of five prediction models for different time intervals. The grey dashed line represents the (time-independent) AUROC reported for Gail's Risk factor model [13]. [B] AUPR. The numbers below the x-axis labels are the average number of BC patients that were available across the cross-validation folds for each time interval. [C] Variable importance for model prediction in a 183-day window. Points indicate the different variables. The y-axis presents VIMP, the decrease in AUROC following random assignment of values to the variable. The x-axis plots Vprop, the variable's inclusion frequency in the trees of the model. For both measures, higher values indicate more importance. The color of a point represents the category of the parameter. Features of low importance (Vprop <0.05 and VIMP<0.5) are not shown.

Variable importance: Figure 9C summarizes the importance of variables in TVsuRF BC risk prediction model for a time window of 183 days. The most important variables in the TVsuRF model were mean corpuscular volume (MCV), monocytes (MONO), mean platelet volume (MPV), mean corpuscular hemoglobin concentration (MCHC), and age.

The importance of immune system-related covariates such as MONO might correlate to the fact BC is an inflammatory and systemic disease.

Comparison to mammography and CBE: For every woman who underwent mammography or CBE in her checkup visits, we compared the results of the 730-day predictor, computed using data only from her latest visit. CBE had 29.1% sensitivity and 93.7% specificity, while TVsuRF had 12.5% sensitivity for the same specificity. Mammography sensitivity and specificity were 58.3% and 66.1%, and TVsuRF had 41.7% sensitivity for similar specificity (Note that the results are not directly comparable, as mammography and CBE identify current malignancy and TVsuRF computes future disease risk). The results in **Supplementary Figure 7** show the three predictions for women that were subsequently diagnosed with BC. Remarkably, the three women with the highest risk score estimated by our model were not detected by CBE, and one of them tested negative in mammography as well. In contrast, some of the women had lower risk scores but were detected by other screening tests.

6.2. **Prostate Gland Cancer**

Dataset: This cohort consisted of 11,416 males who made a total of 24,567 visits to TAMICS. Out of them 56 were subsequently diagnosed with PGC and had 64 visits less than 730 days before the PGC diagnosis. We call this group the PGC subset. The covariates included in the model were CBC (20 parameters), basic metabolic panel data (BMP, 16 parameters), lipids (4 parameters), vital signs (5 parameters), urine tests (2 parameters), troponin, age and BMI. The characteristics of the covariates are summarized in **Table 2**. Since PGC individuals were significantly older than the PGC-free individuals, to reduce the effect of age on our model, we created an age-matched cohort ('Matched PGC-Free') of 3,320 subjects (6,083 visits) using the approach of [57] (**Table 2**). None of the covariates showed significant difference between the PGC and the Matched PGC-Free groups.

Prediction accuracy: Figures 10A and 10B show the results of five prediction methods, using the same comparison metrics as in the BC section. Our model had the highest AUROC in prediction window of 0-183 days and similar performance for intermediate size

time windows. For windows of 547 days and longer, RF had the highest AUROC. In terms of AUPR, our model performed best in until 547 days and the advantage was significant in the windows of up to 273 days. When testing variants of RSF, TVsuRF had better performance on the prediction windows of 0-183 days, but less for longer time windows (**Supplementary Figure 8**).

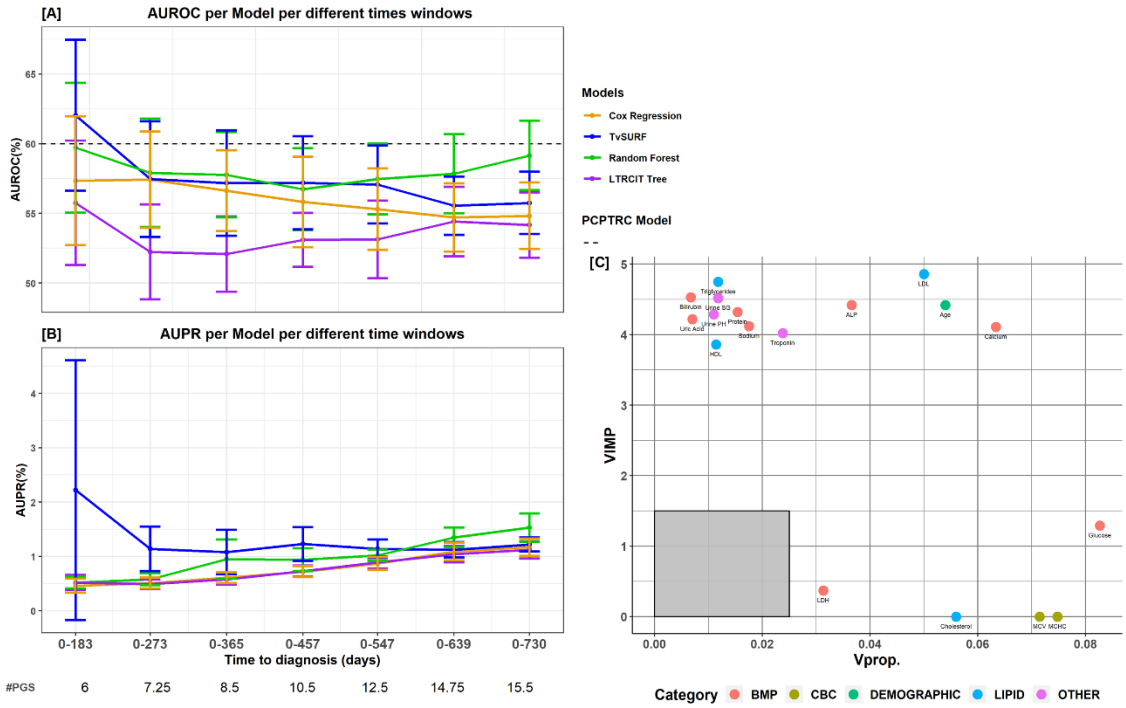


Figure 10: PGC risk prediction and variable importance. [A] Performance (AUROC mean±SD) of five prediction models for different time windows. The grey dashed line represents the (time-independent) AUROC previously reported for the PCPTRC model. [B] AUPR. The numbers below the x-axis labels are the average number of individuals with PGC that were available across the cross-validation folds for each time interval. [C] Variable importance for model prediction in a 183 day window. Points indicate the different variables. Axis definitions are as in **Figure 9**. The color of a point represents the variable's category. Features of low importance (Vprop <0.025 and VIMP<1.5) are not shown.

Variable importance: Figure 10C summarizes the importance of the variables used by TVsuRF in PGC risk prediction, for the 183-day window. The covariates alkaline phosphatase (ALP), low-density lipoprotein (LDL), age, calcium, and glucose had the largest impact on the model. Most of the lipids that were measured - LDL, high-density lipoprotein (HDL), cholesterol and triglycerides - had high importance risk according to at least one criterion, in agreement with previous reports [59].

7. **Discussion**

In this article, we introduced a method for survival prediction based on time-varying covariates utilizing an ensemble of survival trees, and applied it for predicting future emergence of breast and prostate cancer. Our method outperformed traditional prediction methods in breast cancer and for short-term prediction also in prostate cancer. While traditional survival analysis methods use prior assumptions concerning the distribution of the data [60], our method relies only on the proportional-hazard assumption.

Our work has several limitations. First, we do not directly address the issue of size imbalance between the negative (here, the majority) and positive classes. That could affect the splitting criteria and produce nodes with a small number of samples or nodes without failure events, especially in datasets with high-dimensional feature space. Methods such as synthetic minority sampling might address this point [61]. Second, since our dataset did not record the existing clinical models for cancer risk (Gail's model for BC, and PCRTRC model for PGC), we could not compare performance to them on individual patients in our cohort. Incorporating them as additional features in our models may improve prediction. Third, the small number of visits per patient did not allow us to incorporate into the model time-related features, as suggested, e.g., in [30,62] engineered features that capture interactions [63], or to model per-patient random effects across pseudo-intervals. Other model extensions such as competing risks (e.g. death) and accounting for cardiovascular background were not possible for lack of data. Moreover, the limited cohort size made it difficult to evaluate the calibration of our model.

Future work should examine different imputation methods, as those might affect the performance of classifiers when modeling EMR data [64], and investigate sequential models that incorporate the full history in predicting the personalized survival curve [65]. In addition, 'out-of-bag' approaches may improve the evaluation of the prediction, as previously suggested [66]. Moreover, the robustness of the approach is yet to be demonstrated on EMR data from other medical centers. Predictions for additional types of cancers should also be tested, given sufficient data. Finally, a prospective clinical study would provide a more accurate evaluation of the performance.

In conclusion, our models demonstrate the potential of using common laboratory tests of healthy individuals to assess cancer risk. They can serve as additional screening tests and complement the existing BC screening methods.

8. References

[1] Loomans-Kropp HA, Umar A. Cancer prevention and screening: the next step in the era of precision medicine. *Npj Precis Oncol* 2019;3:1–8.

[2] Adamson AS, Welch HG. Machine learning and the cancer-diagnosis problem — No gold standard. *N Engl J Med* 2019;381:2285–2287.

[3] Wiens J, Saria S, Sendak M, et al. Do no harm: a roadmap for responsible machine learning for health care. *Nat Med* 2019;25:1337–1340.

[4] Crosby D, Lyons N, Greenwood E, et al. A roadmap for the early detection and diagnosis of cancer. *Lancet Oncol* 2020;21:1397–1399.

[5] Early detection: A long road ahead. *Nat Rev Cancer* 2018;18:401.

[6] Gourd E. New advances in prostate cancer screening and monitoring. *Lancet Oncol* 2020;21:887.

[7] Gordon L, Olshen RA. Tree-structured survival analysis. *Cancer Treat Rep* 1985;69:1065–1068.

[8] Hothorn T, Bühlmann P, Dudoit S, et al. Survival ensembles. *Biostatistics* 2006;7:355–373.

[9] Stratton MR, Campbell PJ, Futreal PA. The cancer genome. *Nature* 2009;458:719–724.

[10] Prostate Cancer — Cancer Stat Facts n.d. https://seer.cancer.gov/statfacts/html/prost.html (accessed January 18, 2021).

[11] מערכת נתוני הרישום הלאומי לסרטן, משרד הבריאות n.d. https://www.health.gov.il/UnitsOffice/HD/ICDC/ICR/Pages/tableau_loby.aspx (accessed January 18, 2021).

[12] Perez-Cornago A, Key TJ, Allen NE, et al. Prospective investigation of risk factors for prostate cancer in the UK Biobank cohort study. *Br J Cancer* 2017;117:1562–1571.

[13] Ilic D, Neuberger MM, Djulbegovic M, et al. Screening for prostate cancer. *Cochrane Database Syst Rev* 2013;2013.

[14] Ankerst DP, Hoefler J, Bock S, et al. Prostate cancer prevention trial risk calculator 2.0 for the prediction of low- vs high-grade prostate cancer. *Urology* 2014;83:1362–1368.

[15] Prostate Cancer Test | 4Kscore | OPKO Health n.d. https://4kscore.com/ (accessed January 18, 2021).

[16] Carlsson S V., Roobol MJ. Improving the evaluation and diagnosis of clinically significant prostate cancer in 2017. *Curr Opin Urol* 2017;27:198–204.

[17] Female Breast Cancer — Cancer Stat Facts n.d. https://seer.cancer.gov/statfacts/html/breast.html (accessed January 18, 2021).

[18] Al-Ajmi K, Lophatananon A, Ollier W, et al. Risk of breast cancer in the UK biobank female cohort and its relationship to anthropometric and reproductive factors. *PLoS One* 2018;13:e0201097.

[19] Bancej C, Decker K, Chiarelli A, et al. Contribution of clinical breast examination to mammography screening in the early detection of breast cancer. *J Med Screen* 2003;10:16–21.

[20] Gail MH, Brinton LA, Byar DP, et al. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *J Natl Cancer Inst* 1989;81:1879–1886.

[21] Banegas MP, John EM, Slattery ML, et al. Projecting individualized absolute invasive breast cancer risk in US hispanic women. *J Natl Cancer Inst* 2017;109.

[22] Berry DA, Iversen ES, Gudbjartsson DF, et al. BRCAPRO validation, sensitivity of genetic testing of BRCA1/BRCA2, and prevalence of other breast cancer susceptibility genes. *J Clin Oncol* 2002;20:2701–2712.

[23] Tyrer J, Duffy SW, Cuzick J. A breast cancer prediction model incorporating familial and personal risk factors. *Stat Med* 2004;23:1111–1130.

[24] Lee A, Mavaddat N, Wilcox AN, et al. BOADICEA: a comprehensive breast cancer risk prediction model incorporating genetic and nongenetic risk factors. *Genet Med* 2019;21:1708–1718.

[25] Kim HE, Kim HH, Han BK, et al. Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. *Lancet Digit Heal* 2020;2:e138–e148.

[26] McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for

breast cancer screening. *Nature* 2020;577:89–94.

[27] Akselrod-Ballin A, Chorev M, Shoshan Y, et al. Predicting breast cancer by applying deep learning to linked health records and mammograms. *Radiology* 2019;292:331–342.

[28] Stark GF, Hart GR, Nartowt BJ, et al. Predicting breast cancer risk using personal health data and machine learning models. *PLoS One* 2019;14:e0226765.

[29] Wang X, Zhang Y, Hao S, et al. Prediction of the 1-year risk of incident lung cancer: Prospective study using electronic health records from the state of Maine. *J Med Internet Res* 2019;21:e13260–e13260.

[30] Kinar Y, Kalkstein N, Akiva P, et al. Development and validation of a predictive model for detection of colorectal cancer in primary care by analysis of complete blood counts: A binational retrospective study. *J Am Med Informatics Assoc* 2016;23:879–890.

[31] Abelson S, Collord G, Ng SWK, et al. Prediction of acute myeloid leukaemia risk in healthy individuals. *Nature* 2018;559:400–404.

[32] Klein JP, Moeschberger ML. Survival Analysis. New York, NY: Springer New York; 2003.

[33] Bacchetti P, Segal MR. Survival trees with time-dependent covariates: Application to estimating changes in the incubation period of AIDS. *Lifetime Data Anal* 1995;1:35–47.

[34] Huang X, Chen S, Soong S. Piecewise exponential survival trees with time-dependent covariates. *Biometrics* 1998;54:1420.

[35] Bou-Hamad I, Larocque D, Ben-Ameur H. Discrete-time survival trees and forests with time-varying covariates: Application to bankruptcy data. *Stat Modelling* 2011;11:429–446.

[36] Wallace ML. Time-dependent tree-structured survival analysis with unbiased variable selection through permutation tests. *Stat Med* 2014;33:4790–4804.

[37] Fu W, Simonoff JS. Survival trees for left-truncated and right-censored data, with application to time-varying covariate data. *Biostatistics* 2017;18:352–369.

[38] Therneau T, Crowson C, Atkinson E. Using time dependent covariates and time dependent coefficients in the cox model. *Surviv Vignettes* 2017:1–8.

[39] Andersen PK, Gill RD. Cox's regression model for counting processes: a large sample

study. *Ann Stat* 1982;10:1100–1120.

[40] Bellot A. Boosted trees for risk prognosis. *Proc Mach Learn Res* 2018;85:1–15.

[41] Ishwaran H, Kogalur UB, Blackstone EH, et al. Random survival forests. *Ann Appl Stat* 2008;2:841–860.

[42] Breiman L. Random forests. *Mach Learn* 2001;45:5–32.

[43] Ishwaran H, Kogalur UB. random survival forests for R. *New Funct Multivar Anal* 2007;7:25–31.

[44] Hothorn T, Hornik K, Zeileis A. Unbiased recursive partitioning: A conditional inference framework. *J Comput Graph Stat* 2006;15:651–674.

[45] Utkin L V., Konstantinov A V., Chukanov VS, et al. A weighted random survival forest. *Knowledge-Based Syst* 2019;177:136–144.

[46] Steingrimsson JA, Diao L, Strawderman RL. Censoring unbiased regression trees and ensembles. *J Am Stat Assoc* 2019;114:370–383.

[47] Bou-Hamad I, Larocque D, Ben-Ameur H. A review of survival trees. *Stat Surv* 2011;5:44–71.

[48] Sun Y, Chiou SH, Wang MC. ROC-guided survival trees and ensembles. *Biometrics* 2019.

[49] Wongvibulsin S, Wu KC, Zeger SL. Clinical risk prediction with random forests for survival, longitudinal, and multivariate (RF-SLAM) data analysis. *BMC Med Res Methodol* 2019;20:1.

[50] Mantel N. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother Rep* 1966;50:163–170.

[51] 3 Likelihood and Censored (or Truncated) Survival Data Review of Parametric Likelihood Inference. n.d.

[52] Little RJA. Missing-data adjustments in large surveys. *J Bus Econ Stat* 1988;6:287–296.

[53] van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. *J Stat Softw* 2011;45:1–67.

[54] Pan W. Rank invariant tests with left truncated and interval censored data. *J Stat Comput Simul* 1998;61:163–174.

[55] Blanche P, Kattan MW, Gerds TA. The c-index is not proper for the evaluation of t-year predicted risks. *Biostatistics* 2019;20:347–357.

[56] Wright MN, Ziegler A. Ranger: A fast implementation of random forests for high dimensional data in C++ and R. *J Stat Softw* 2017;77.

[57] Ho DE, Imai K, King G, et al. MatchIt: Nonparametric preprocessing for parametric causal inference. *J Stat Softw* 2011;42:1–28.

[58] Clendenen T V., Ge W, Koenig KL, et al. Breast cancer risk prediction in women aged 35-50 years: impact of including sex hormone concentrations in the Gail model. *Breast Cancer Res* 2019;21:42.

[59] Tewari R, Chhabra M, Natu SM, et al. Significant association of metabolic indices, lipid profile, and androgen levels with prostate cancer. *Asian Pacific J Cancer Prev* 2014;15:9841–9846.

[60] LeBlanc M, Crowley J. Survival trees by goodness of split. *J Am Stat Assoc* 1993;88:457–467.

[61] Afrin K, Illangovan G, Srivatsa SS, et al. Balanced random survival forests for extremely unbalanced, right censored data. *ArXiv* 2018;preprint:1803.09177.

[62] Karnes RJ, MacKintosh FR, Morrell CH, et al. Prostate-specific antigen trends predict the probability of prostate cancer in a very large U.S. Veterans affairs cohort. *Front Oncol* 2018;8:296.

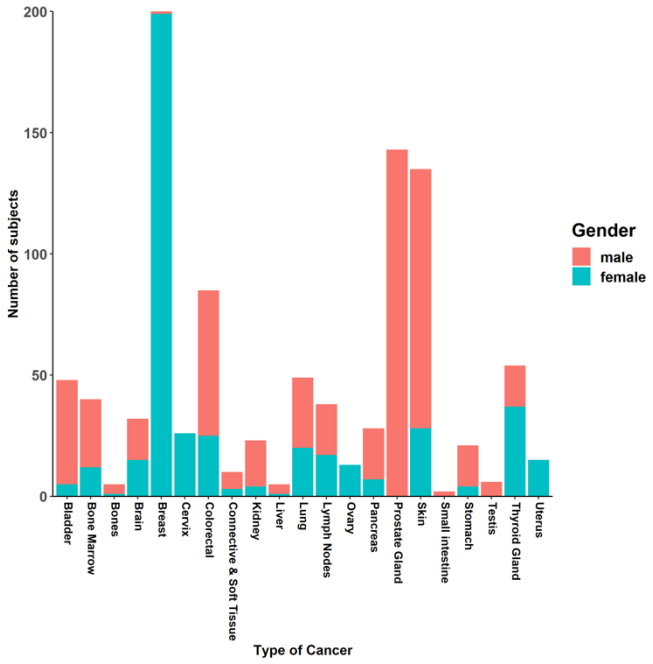
[63] Hayashi T, Fujita K, Tanigawa G, et al. Serum monocyte fraction of white blood cells is increased in patients with high Gleason score prostate cancer. *Oncotarget* 2017;8:35255–35261.

[64] Che Z, Purushotham S, Cho K, et al. Recurrent neural networks for multivariate time series with missing values. *Sci Rep* 2018;8:6085.

[65] Lee C, Yoon J, Van Der Schaar M. Dynamic-DeepHit: a deep learning approach for dynamic survival analysis with competing risks based on longitudinal data. *IEEE Trans Biomed Eng* 2020;67:122–133.

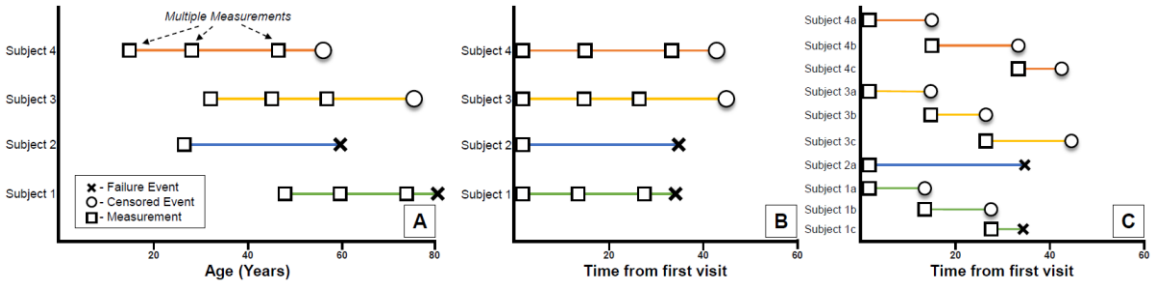
[66] Hothorn T, Lausen B, Benner A, et al. Bagging survival trees. *Stat Med* 2004;23:77–91.

9. Supplementary Material



<u>Supplementary Figure 1:</u> Number of patients per cancer type.

Bar plot of the number of individuals who were surveyed in TAMICS and later diagnosed with cancer, categorized by gender and type of cancer.



10.

<u>Supplementary Figure 2</u>: Presenting longitudinal data of multiple visits. [A] Longitudinal measurements and survival analysis setting. Squares indicate the times of the longitudinal measurements, Crosses indicate failure events, and circles indicate

censoring events. [B] The data after shifting all first visit times to 0. [C] The same data after transforming into pseudo-objects.

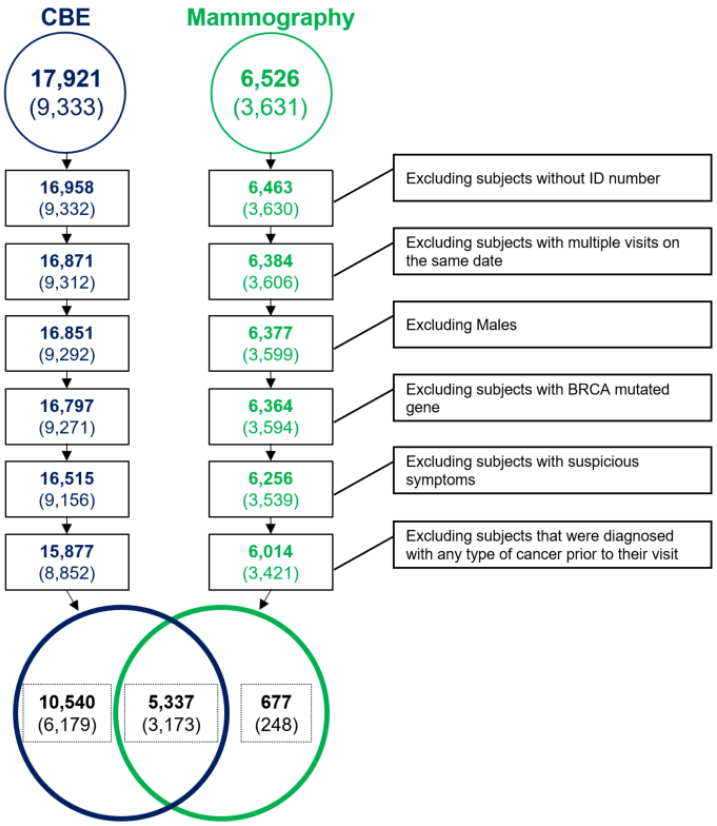
Supplementary Material 3: Decision tree basics

Binary trees:

A rooted tree T is a connected acyclic graph with a designated node r called the *root*. Other nodes of degree 1 are called *leaves*. In such a tree there is a single simple path from r to every node and the number of edges in the path is the *depth* of the node. If there exists a simple path from r to v that passes through u then u is called an *ancestor* of v. If also (u, v) is an edge then u is the *father* of v and v is *child* of u. If every non-leaf has two children then T is called a *binary tree*.

Decision trees:

A binary rooted tree can be used as a decision tree for classification as follows: Each internal (non-leaf) node is associated with a certain covariate and a threshold value. Samples with the covariate value above the threshold are assigned to the right child, and the rest are assigned to the left. This way, a sample starts at the root and descends left or right depending on the corresponding covariate values until it is associated with a leaf. If leaves are assigned with a class label (e.g. case/control), the tree assigns a class for the sample. Similarly, a set of samples can be partitioned into disjoint subsets corresponding to the leaves. Note that in our case the samples are the LTRC pseudo-intervals.

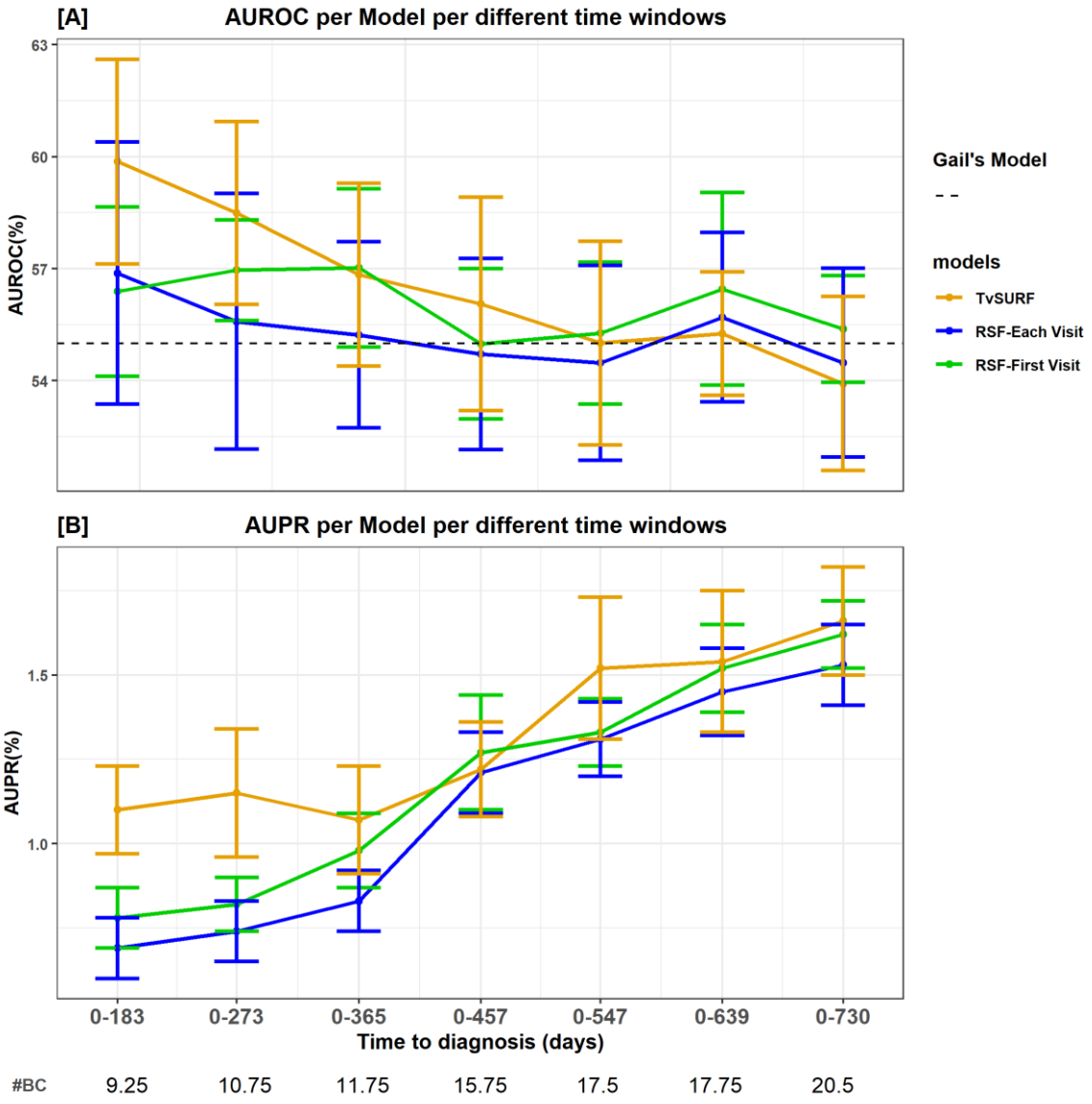


Supplementary Figure 4: **The CBE and mammography cohorts.** Effect of exclusion criteria on the members of the TAMICS cohort who conducted a mammography screening test for BC and CBE.

Supplementary Material 5: Cohort of subjects with CBE and Mammography tests.

We removed all the visits that occurred less than 31 days after the previous one. We excluded all subjects with two or more types of cancer unless the only other type was skin cancer. In case of more than one BC diagnosis we considered only the first one. We used natural language processing to classify each subject who was recommended to conduct any BC-related follow-up test as positive (abnormal mammography). The extraction of the recommendation from the physician's notes was done using a pattern detection script. All phrases after an action verb, such as ‘is required’; ‘recommend’; were extracted and a dictionary of words that indicate BC follow-up test (ultrasound, biopsy, trucut etc.) was created. We manually reviewed the mammography results and added more

action verbs and recommendations in several iterations. Finally, we randomly sampled and manually reviewed 100 cases to confirm the efficacy of our pattern recognition script.

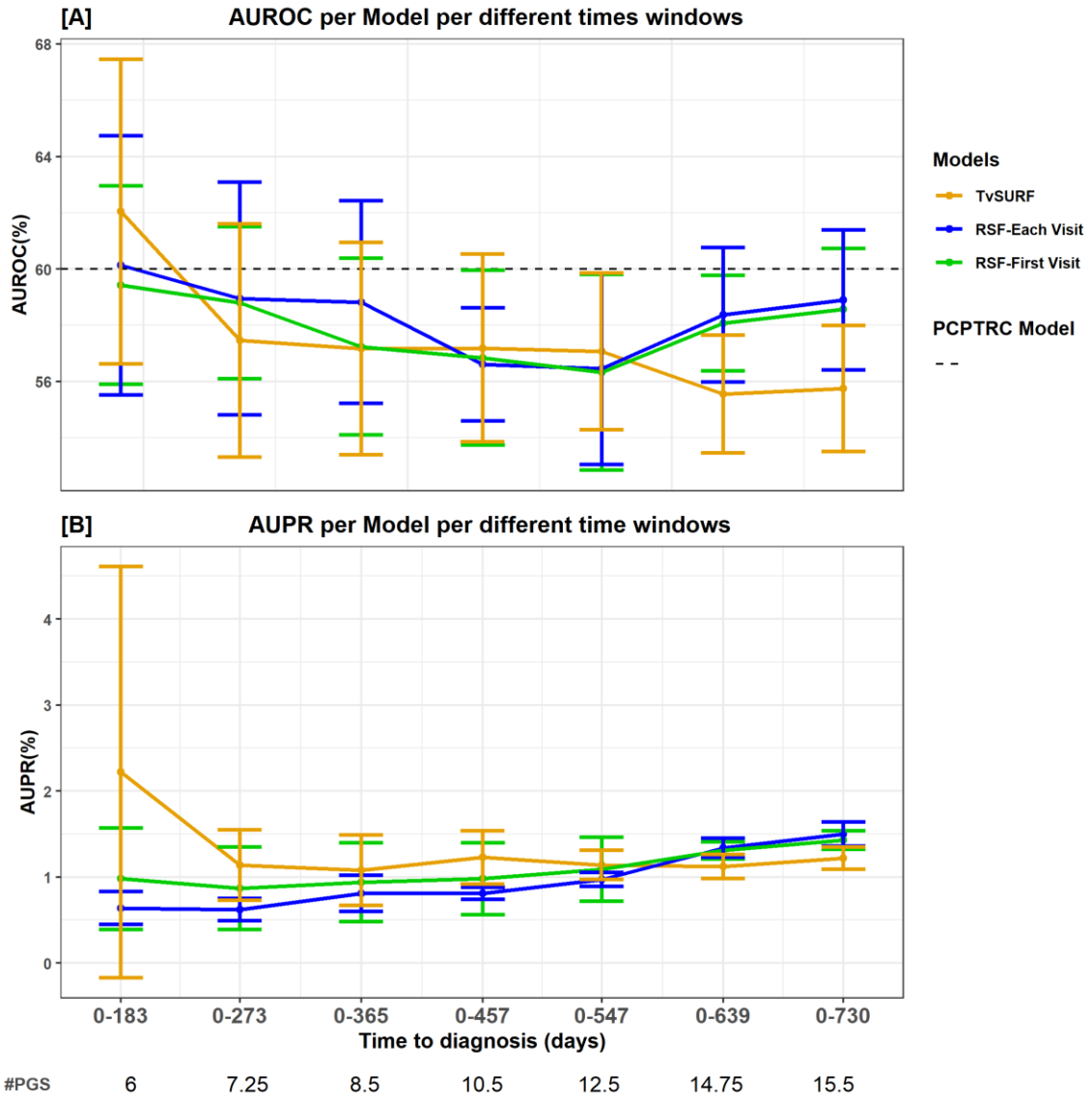


Supplementary Figure 6: BC risk prediction – comparison of TVsuRF to random survival forest. Two versions of RSF were applied: Each Visit: All pseudo-intervals were used. First Visit: Every visit creates an interval starting at the visit time and ending at the time of failure or censoring of the subject. In the two versions, all pseudo-intervals were linearly shifted to start at time t=0 since the RSF models are time-independent. The

numbers below the x-axis labels are the average number of BC patients that were available across the cross-validation folds for each time interval.

CBE																		
Mammography																		
Risk Score	10	9.6	7.53	7.07	6.93	6.67	6	6	5.73	5.67	5.6	5.4	5.2	4.67	4.6	4.13	4.07	3.13
Time-to-Diag.	461	389	54	407	9	434	14	14	653	3	249	563	5	254	42	38	603	33
Patient ID	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***11***	***12***	***13***	***14***	***15***	***16***	***17***	***18***

<u>Supplementary Figure 7:</u> TVsuRF risk score and BC screening tests results for women who subsequently were diagnosed with BC. Green: a normal result; Red: an abnormal test; Grey: test unavailable. 1^{s} line: CBE result; 2nd line: mammography result; 3rd line: the risk score calculated by the TVsuRF model. Patients were ordered from high (dark blue) to low (light blue) risk score. 4th line: time from visit to cancer diagnosis.



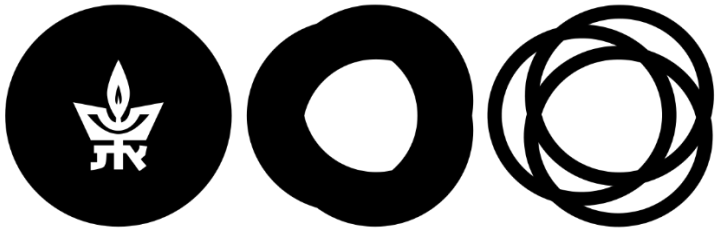
<u>Supplementary Figure 8:</u> PGC risk prediction – comparison of TVsuRF to random survival forest. The same two RSF variants in SFig. 6 were used. The grey dashed line represents the (time-independent) AUROC previously reported for the PCPTRC model.

תקציר

מטרות: לנבא סיכון לסרטן שד וסרטן ערמונית בקרב אוכלוסיה בריאה ע"י ניתוח בדיקות מעבדה רוטיניות, מדדים חיוניים וגיל.

שיטות: ניתחנו רשומות רפואיות אלקטרונית של 20,317 אנשים בריאים אשר עברו בדיקות תקופתיות שגרתיות. בכל בדיקה כזו נאספו יותר מ-600 פרמטרים. בעזרת רשם הסרטן הישראלי, זיהינו את האנשים אשר חלו בסרטן לאחר הבדיקה. פיתחנו מודל לניבוי הסיכון, המתבסס על מידע רב-מימדי ותלוי-זמן. המודל משתמש בעצי הישרדות עבור נתונים שמצונזרים מימין וקטומים משמאל והאלגוריתם מבוסס על שיטת Random Forest.

תוצאות: במבחני cross-validation השיטה שפיתחנו השיגה תוצאות של AUROC 0.62 ± 0.05 בניבוי הסיכון לסרטן ערמונית ו- 0.6 ± 0.03 בניבוי סרטן שד, לאירוע שיתרחש תוך שישה חודשים ממועד הבדיקה. הביצועים הללו היו טובים יותר מביצועים של עץ הישרדות בודד, רגרסיית Cox ואלגוריתם Random Forest. להערכתנו, השיטה שלנו עשויה להיות משלימה לבדיקות הסקר הקיימות (ממוגרפיה, בדיקת מישוש שד) ולסייע בזיהוי של פציינטים אשר לא זוהו ע"י הבדיקות הללו.



אוניברסיטת תל אביב

הפקולטה למדעים מדוייקים ע"ש ריימונד ובברלי סאקלר

בית הספר למדעי המחשב ע"ש בלווטניק

זיהוי מוקדם של סיכון לסרטן שד וערמונית על בסיס נתוני בדיקות סקר בעזרת עצי הישרדות מותאמים לנתונים מצונזרים מימין וקטומים משמאל

חיבור זה הוגש כעבודת גמר לתואר 'מוסמך אוניברסיטה'

בבית הספר למדעי המחשב על ידי

דן קוסטר

בהנחיית

פרופ' רון שמיר

שבט תשפ"א