Predicting personal risk of developing breast and prostate cancer from routine check-up data using survival analysis trees

Dan Coster
Tel-Aviv University https://orcid.org/0000-0001-8514-7965
Eyal Fisher
University of Cambridge
Shani Shenhar-Tsarfaty
Tel Aviv Sourasky Medical Center
Tehillah Menes
Tel-Aviv University
Shlomo Berliner
Tel Aviv Sourasky Medical Center
Ori Rogowski
Tel Aviv Sourasky Medical Center
David Zeltser
Tel-Aviv University
Itzhak Shapira
Tel-Aviv University
Eran Halperin
University of California, Los Angeles
Saharon Rosset
Tel Aviv University https://orcid.org/0000-0002-4458-9545
Malka Gorfine
Tel-Aviv University
Ron Shamir (rshamir@tau.ac.il)
Tel Aviv University

Article

Keywords: prostate cancer, breast cancer, risk prediction, early detection of cancer, machine learning, electronic medical records

Posted Date: February 8th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1298640/v1

Predicting personal risk of developing breast and prostate cancer from routine check-up data using survival analysis trees

Dan Coster[1], Eyal Fisher[4], Shani Shenhar-Tsarfaty[3,8], Tehillah Menes[7,8], Shlomo Berliner[3,8], Ori Rogowski[3,8], David Zeltser [3,8], Itzhak Shapira[3,8], Eran Halperin[5,6], Saharon Rosset[2], Malka Gorfine[2], Ron Shamir*[1]

[1] Blavatnik School of Computer Science, Tel-Aviv University, Tel-Aviv, Israel

[2] Department of Statistics and Operations Research, Tel-Aviv University, Tel-Aviv, Israel

[3] Departments of Internal Medicine "C", "D" and "E", Tel-Aviv Sourasky Medical Center

[4] Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, UK

[5] Department of Computer Science, University of California, Los Angeles, California, USA.

[6] Department of Computational Medicine, University of California, Los Angeles, California, USA

[7] Department of Surgery C & Surgical Oncology, Chaim Sheba Medical Center, Ramat Gan, Israel

[8] Sackler Faculty of Medicine, Tel-Aviv University, Tel-Aviv, Israel

Corresponding Author: Ron Shamir

Email: rshamir@tau.ac.il

Phone Number: +972-54-304-3337

Postal address: School of Computer Science, Tel-Aviv University, 30 Haim Levanon Street, Tel-Aviv, Israel, 69978

Keywords: prostate cancer, breast cancer, risk prediction, early detection of cancer, machine learning, electronic medical records

Word Count: 7,730

ABSTRACT

The challenge of survival prediction is ubiquitous in industry and medicine. Few methods are available for survival prediction of time varying data. Here we propose a novel method for this problem, using a random forest of survival trees for left truncated and right-censored data. We demonstrated the advantage of our method on prediction of 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 and identified those who later developed cancer. In cross-validation, our method predicted future prostate and breast cancers six months before diagnosis with an area under the ROC curve of 0.62 ± 0.05 and 0.6 ± 0.03 respectively, outperforming standard random forest, Cox-regression model and a single survival tree. Our results suggest that computational analysis of data on healthy individuals can improve the detection of those at risk of future cancer development.

INTRODUCTION

In industry and clinical research, survival models are widely employed. In survival analysis, the goal is to develop a model for predicting the failure time of a query sample, based on training examples with their features (covariates) and failure or censoring times. In healthcare, examples correspond to individuals, features are clinical measurements and failure is a particular medical event (death, disease onset, hospitalization, etc.)

Several approaches have been proposed for survival prediction. Gordon and Olsen introduced the survival tree[1], a decision tree where each node contains a survival curve of the corresponding subgroup of individuals. The node splitting criterion usually aims to maximize the difference in survival between the subgroups of the daughter nodes or the within-node homogeneity. Most survival tree methods addressed right-censored data and time-independent covariates. Incorporating time-varying covariates in survival trees was first done by introducing 'pseudo-objects'[2], a concept later used in other studies[3–6]. Additional methods that construct survival trees for time-dependent covariates include the Cox proportional hazard model[10,11] and others[7–9].

Several ensemble methods for survival trees analysis were suggested for time-independent covariates[12,13]. Random survival forests (RSF), introduced by Ishwaran[14], combined Breiman's random forest (RF)[15,16], survival trees and the log-rank test as the splitting criterion. An extension of RSF utilized conditional inference trees, which employ hypothesis testing to select the splitting covariates and also as a stopping criterion[17], and other improvements were also examined[18,19].

Advance warning for high cancer risk in healthy individuals is crucial for providing appropriate care and can improve both prognosis and survival[20–24]. 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. The current strategies for early detection of cancer use specific screening tests that require substantial resources, e.g., serum Prostate-Specific Antigen (PSA) level for Prostate Gland Cancer (PGC), mammography, an X-ray modality, for detecting early signs of Breast Cancer (BC), and clinical breast examination (CBE), a physical examination to recognize abnormalities in the breast[25]. Other approaches to assess cancer risk use models, e.g. Gail's model[26,27], BRCAPRO[28], IBIS[29] and BOADICEA[30] for BC risk, and the Prostate Cancer Prevention Trial Risk Calculator (PCPTRC) for PGC risk[31]. These models use a few clinical and genetic parameters and are not based on routine laboratory measurements, and their performance is relatively limited[32]. Advanced genetic methods, including polygenic risk scores, have been proposed for screening, but they are still not part of the clinical practice[33,34]. In recent years, machine learning has improved screening models in two major ways. It enhanced existing screening tests, including mammography analysis[35–37], Gail's model[38], and the PGC risk score[39]. A new set of cancer risk prediction tools was also developed based on patients' historical Electronic Medical Records (EMR) collected over time as part of routine care. Such models were suggested for lung cancer[40], colorectal cancer[41], and Acute Myeloid Leukemia[42], among others.

Here, we address survival prediction for time-dependent covariates by combining the notions of survival trees, pseudo-objects and ensemble methods. We present a novel method called TVsuRF (Time-Varying SUrvival Random Forest). It is the first to use conditional inference trees in this setting. We demonstrate our method's utility in early detection of cancer, for BC and PGC, the most common cancers among females and males,

respectively. Our model uses EMR data collected from healthy individuals in routine periodic checkups. To the best of our knowledge, this is the first risk model that is based on routine laboratory measurements proposed for these cancer types.

RESULTS

Dataset and Covariates

We analyzed data from routine checkups of individuals at the Tel-Aviv Medical Center Inflammation Survey (TAMCIS), Tel-Aviv Sourasky Medical Center, Israel. Most individuals performed the checkups as an employer-provided benefit while others independently chose to be checked. Participants were men and non-pregnant women with no active current malignant or infectious disease who chose to be tested and signed an informed consent form. In each visit, the individual 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 over several years. We conducted a retrospective analysis of the TAMCIS EMR data collected between November 2001 and February 2017. Our study covered 20,317 adults (age $\geq$ 18). The study was reviewed and approved by the Institutional Review Board (Approval no. 02-049-Tlv). We identified the individuals who later developed cancer using the Israeli National Cancer Registry (INCR) and applied certain inclusion and exclusion criteria (see "Methods", **Figure 1**).

We used only covariates that were available for more than 80% of the individuals. The missing values were imputed by Predictive-Mean-Matching on age[43] using the *mice* package[44]. For BC risk prediction we used 20 covariates (**Table 1**) that include

demographic parameters such as age and body mass index (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**). The study complied with the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement.

The TVsuRF model

We developed a novel method called TVsuRF (time-varying survival random forest) for risk prediction. It uses as training data multivariate time series data and employs a random forest of survival trees for left-truncated and right-censored data. Given query data of a new individual, the method predicts the individual's risk for developing cancer during the next two years.

To handle training individuals with several visits, we used pseudo-objects[2]: For each visit, we created an interval (pseudo-object) that starts at the age of the individual at the visit (left-truncation time). See **Figure 5**. The end of the interval was one of three options: (1) if the individual had a subsequent visit, and was not diagnosed with cancer by that time, the interval ended at the next visit and was right censored. (2) if that was the individual's last visit, and s/he was not diagnosed with cancer by the INCR query time, the interval ended at that time and was censored. (3) otherwise that was the individual's last visit, and s/he was diagnosed with cancer by the INCR query time, so the interval ended at the recorded cancer diagnosis time with failure. In this manner, each interval was treated as a separate left truncated and right censored (LTRC) pseudo-object, with the covariates

recorded at the visit. Finally, the times were shifted so that each individual's first visit was at time 0, and we added the age at each visit as a covariate of the pseudo-object.

This formulation allowed us to utilize a standard Kaplan-Meier (KM) estimator of the survival function for LTRC data[45]. We constructed a survival tree for the pseudo-objects, where each node corresponds to a subset of the pseudo-objects, and compute a survival KM estimator for that subset. We used the framework of conditional inference decision trees[17], which employs a statistical hypothesis test based on permutations, in order to select optimal variables and their thresholds. Each split of a covariate and a threshold induced two subsets of pseudo-objects and we used Pan's permutations based hypothesis test[46] in order to compare their KM estimators and select the optimal split with the lowest p-value.

We created an ensemble model with 500 survival trees. In each tree, at each internal node, we select at random a subset of the covariates and split the node according to the covariate and threshold giving the least p-value for difference in survival, if that difference is significant. The predicted survival curve for a new individual is obtained by identifying the leaf the individual ended in in each tree, and averaging their KM estimators. This curve gives for each time Δt the probability of not having a failure event (the cancer diagnosis) until time $t + \Delta t$ based on the individual's covariates from the latest visit at time t (**Figure 2A-2C**).A full description of the algorithm is presented in **Figure 6** and "Methods"**.**

Evaluation Approach

We used TVsuRF and several other models to predict BC and PGC risk on our cohorts. We denote the covariates of individual i that were measured at time t as $x^i(t)$, where $t =$

0 is the time of the first visit of the individual. 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 individual i is thus $1 - \hat{S}\left(t + \Delta t + \epsilon \mid x^i(t)\right)$ where $\hat{S}$ is the predicted survival curve (**Figure 2D**).

To evaluate the performance of risk predictors in classification, we calculated the area under the receiver operator characteristic curve (AUROC), where the positive class is the set of individuals who were diagnosed with cancer during the next $\Delta t + \epsilon$ days as suggested in[47], but excluding pseudo-objects censored in that period. We also estimated the area under the precision-recall curve (AUPR). We performed 20 iterations of 4-fold cross-validation, where in each iteration the partition of individuals into folds was done at random. For each of the above measures, we calculated the average and standard deviation.

We tested three other models: (1) Cox regression model adapted to time-varying covariates[10,11], (2) single LTRC survival tree as in[6] (denoted LTRCIT), and (3) RF model[15]. Since RF is a classification model, training for prediction was done separately for each time interval Δt, and the class of an individual 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[48]. In addition, we compared our method to a random survival forest (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.

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 that were associated with the covariate (i.e. the covariate was used to split these nodes) in all the trees. 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 modified data. The difference in the AUROC between the original and the modified data was averaged over ten random assignments per each covariate on every fold of the 4-fold cross-validation[14]. We repeated this process 20 times and defined VIMP as the mean difference obtained. Again, higher VIMP indicates more importance.

Breast Cancer Cohort

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 covariates 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 concertation (MCHC). To reduce the effect of age on our model, we created an age-matched cohort ('Matched BC-Free') of3,635 individuals (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.

Predicting Breast Cancer Risk

The performance of each of the methods tested, for different time intervals, is summarized in **Figures 3A and 3B**. We also marked the AUROC of Gail's breast cancer risk estimation for 5 years horizon as reported in[32]. TVsuRF had the highest AUPR on every time interval, and the highest AUROC on all intervals except 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 1**).

Figure 3C summarizes the importance of variables in TVsuRF BC risk prediction model for a time window of 183 days. The variables mean corpuscular volume (MCV), monocytes (MONO), mean platelet volume (MPV), MCHC and age were most important in the model. The importance of immune system-related covariates such as MONO might correlate to the fact BC is an inflammatory and systemic disease.

Notably, when focusing on women who underwent standard BC screening tests (clinical breast examination and mammography), our method gave very high risk scores to several women who tested negative and developed breast cancer later (**Supplementary Text 1, Supplementary Figures 2,3).**

Prostate Gland Cancer Cohort

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 individuals (6,083 visits) as done for BC (**Table 2**). None of the covariates showed significant difference between the PGC and the Matched PGC-Free groups.

Predicting Prostate Gland Cancer Risk

Figures 4A and 4B show the results of five prediction methods, using the same comparison metrics used for BC. Our model had the highest AUROC in prediction window of 0-183 days and close to best 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 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 for 0-183 days, but less for longer time windows (**Supplementary Figure 4**).

Figure 4C 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[50].

DISSCUSSION

In this study 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 approach disposes of the time-independence assumption of the existing random survival forest model. Unlike traditional survival analysis methods, which use prior assumptions concerning the distribution of the data[51], our method relies only on the proportional-hazard assumption.

Our study has several limitations. First, we do not directly address the issue of size imbalance between the negative (here, the majority) and positive classes, as done by methods such as synthetic minority sampling[52]. 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. Second, the limited cohort size made it difficult to evaluate the calibration of our model and extend it for competing risks (e.g. death).

In terms of clinical interpretation of our results, since our dataset did not include the input parameters of existing cancer risk scores (Gail's model for BC, and PCRTRC model for PGC), we could not compare performance to them per individual. Moreover, the small number of visits per individual did not allow us to incorporate into the model time-related covariates[53,54], covariate interactions[55], or to model per-individual random effects across pseudo-objects. Future work should investigate dynamic models that incorporate the full history into the risk prediction[56] and examine different imputation methods, as those might affect classifier performance[57]. In addition, 'out-of-bag' approaches may improve the

evaluation of the prediction[58]. Moreover, the robustness of the approach is yet to be demonstrated in other medical centers and on additional types of cancers. Finally, a prospective clinical trial would provide a more accurate evaluation of the performance and medical utility.

Our method outperformed traditional prediction methods in breast cancer and for short term prediction also in prostate cancer, and demonstrated the potential of using common laboratory tests of healthy individuals to assess cancer risk. They can serve as additional screening tests, complementing current screening methods and forthcoming polygenic risk scores.

METHODS

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 5** shows the number of individuals 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. Individuals who had a different type of cancer prior to diagnosis of BC or PGC were excluded.

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 1**).

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.

A cohort of age-matched individuals was created by using the *matchit* package[49].

Model Development

Consider a dataset of N individuals, where for each of them data from one or more visits were recorded. Individual 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 individual i. The last time point individual 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 individual 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 $\mathcal{D} = \{(\mathcal{X}^i, \tau^i, \delta^i)\}_{i=1}^N$ (**Figure 5A**). $\mathcal{X}^i(t)$ denotes the data of individual i that were measured until time t, i.e., $\mathcal{X}^i(t) = \{x^i(t_j^i): 0 \le t_j^i \le t\}$. We assume time homogeneity so that w.l.o.g. we can shift times per individual to set $\forall i: t_1^i = 0$, i.e., all first visits were at time 0 (**Figure 5B**). We also assume that the age of the individual 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 individual's covariates at the latest visit before that time. That is, let $t_*^i = \max\{t_j^i < t | j\}$. We wish to estimate the survival function:

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

In order to model the time-dependent covariates, we use pseudo-objects[2]. We split the data of each individual into disjoint intervals $[t_j^i, t_{j+1}^i)$ and we assume that the covariates $x^i(t_j)$ are constant in the interval (**Figure 5C**). 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 individual i then we view time t_{j+1}^i as censoring time. We denote the pseudo-object of the j^{th} interval of individual i as $[L_j^i, R_j^i)$ where:

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

Hence, the transformation is:

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

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

The standard Kaplan-Meier (KM) estimator of the survival function can now be generalized for LTRC data [59], 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 individuals 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 is calculated in a step ahead prediction manner - we calculate the probability of an individual in time t to experience failure in the next time window Δt given its covariates at time t, namely $\mathbb{P}\left(\tau^i < t + \Delta t, \delta_i = 1 \middle| \tau^i > t, x_i(t)\right)$.

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 2A**). 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[17], which 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, 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 [46], as suggested also in [6]. 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 individual 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 that represents the contribution of sample i to the test statistic.

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 $\leq X$ is $P_{value} = \frac{k}{\binom{N}{n}}$. We estimate it using 1000 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:

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

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 2B**, **Figure 6**).

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. The predicted survival curve for a new individual ω is computed 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 individuals 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 individuals 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 2C**):

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

This gives the risk score of individual ω over time.

DATA AVAILABILITY

Data cannot be shared due to regulations governing privacy protection.

CODE AVAILABILITY

Code for model development and implementation as well as the deployed model are available upon request.

ACKNOWLEDGEMENTS

None.

AUTHOR CONTRIBUTIONS

D.C., R.S., E.F., E.H., S.R., M.G., contributed to the conception of the work and to developing the model.

S.S.T, S.B, O.R, D.Z, I.S collected the data.

D.C. and R.S. developed the model, analyzed the data and assessed the performance

T.M manually reviewed medical records.

D.C., R.S., S.S.T , E.F., E.H. M.G., contributed to the study design.

S.S.T, S.B., T.M assisted in evaluation of the clinical aspects of the study (data interpretation).

D.C and R.S. wrote the manuscript.

All authors contributed to the review of the manuscript.

All authors read and approved the final version of the manuscript.

COMPETING INTERESTS

None.

FUNDING

Supported in part by Israel Science Foundation (ISF) grant No. 1339/18 (RS); ISF grant No. 3165/19, within the Israel Precision Medicine Partnership program (RS); grant 2016694 from the US - Israel Binational Science Foundation (BSF), and the US National Science Foundation (NSF) (RS); ELROV grant (S.S.T). D.C. was supported, in part, by fellowships from the Edmond J. Safra Center for Bioinformatics at Tel Aviv University and from Google.

REFRENCES

1. Gordon, L. & Olshen, R. A. Tree-structured survival analysis. *Cancer Treat. Rep.* **69**, 1065–1068 (1985).

2. Bacchetti, P. & Segal, M. R. Survival trees with time-dependent covariates: Application to estimating changes in the incubation period of AIDS. *Lifetime Data Anal.* **1**, 35–47 (1995).

3. Huang, X., Chen, S. & Soong, S. Piecewise exponential survival trees with time-

dependent covariates. *Biometrics* **54**, 1420 (1998).

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

5. Wallace, M. L. Time-dependent tree-structured survival analysis with unbiased variable selection through permutation tests. *Stat. Med.* **33**, 4790–4804 (2014).

6. Fu, W. & Simonoff, J. S. Survival trees for left-truncated and right-censored data, with application to time-varying covariate data. *Biostatistics* **18**, 352–369 (2017).

7. Bou-Hamad, I., Larocque, D. & Ben-Ameur, H. A review of survival trees. *Stat. Surv.* **5**, 44–71 (2011).

8. Wongvibulsin, S., Wu, K. C. & Zeger, S. L. Clinical risk prediction with random forests for survival, longitudinal, and multivariate (RF-SLAM) data analysis. *BMC Med. Res. Methodol.* **20**, 1 (2019).

9. Sun, Y., Chiou, S. H. & Wang, M. C. ROC-guided survival trees and ensembles. *Biometrics* (2019) doi:10.1111/biom.13213.

10. Therneau, T., Crowson, C. & Atkinson, E. Using time dependent covariates and time dependent coefficients in the cox model. *Surviv. Vignettes* 1–8 (2017).

11. Andersen, P. K. & Gill, R. D. Cox's regression model for counting processes: a large sample study. *Ann. Stat.* **10**, 1100–1120 (1982).

12. Hothorn, T., Bühlmann, P., Dudoit, S., Molinaro, A. & Van Der Laan, M. J. Survival ensembles. *Biostatistics* **7**, 355–373 (2006).

13. Bellot, A. Boosted trees for risk prognosis. *Proc. Mach. Learn. Res.* **85**, 1–15 (2018).

14. Ishwaran, H., Kogalur, U. B., Blackstone, E. H. & Lauer, M. S. Random survival forests. *Ann. Appl. Stat.* **2**, 841–860 (2008).

15. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).

16. Ishwaran, H. & Kogalur, U. B. random survival forests for R. *New Funct. Multivar. Anal.* **7**, 25–31 (2007).

17. Hothorn, T., Hornik, K. & Zeileis, A. Unbiased recursive partitioning: A conditional inference framework. *J. Comput. Graph. Stat.* **15**, 651–674 (2006).

18. Steingrimsson, J. A., Diao, L. & Strawderman, R. L. Censoring unbiased regression trees and ensembles. *J. Am. Stat. Assoc.* **114**, 370–383 (2019).

19. Utkin, L. V. *et al.* A weighted random survival forest. *Knowledge-Based Syst.* **177**, 136–144 (2019).

20. Early detection: A long road ahead. *Nature Reviews Cancer* vol. 18 401 (2018).

21. Loomans-Kropp, H. A. & Umar, A. Cancer prevention and screening: the next step in the era of precision medicine. *npj Precis. Oncol.* **3**, 1–8 (2019).

22. Adamson, A. S. & Welch, H. G. Machine learning and the cancer-diagnosis problem — No gold standard. *New England Journal of Medicine* vol. 381 2285–2287 (2019).

23. Wiens, J. *et al.* Do no harm: a roadmap for responsible machine learning for health care. *Nat. Med.* **25**, 1337–1340 (2019).

24. Crosby, D. *et al.* A roadmap for the early detection and diagnosis of cancer. *Lancet Oncol.* **21**, 1397–1399 (2020).

25. Bancej, C. *et al.* Contribution of clinical breast examination to mammography screening in the early detection of breast cancer. *J. Med. Screen.* **10**, 16–21 (2003).

26. Gail, M. H. *et al.* Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *J. Natl. Cancer Inst.* **81**, 1879–1886 (1989).

27. Banegas, M. P. *et al.* Projecting individualized absolute invasive breast cancer risk in US hispanic women. *J. Natl. Cancer Inst.* **109**, (2017).

28. Berry, D. A. *et al.* BRCAPRO validation, sensitivity of genetic testing of BRCA1/BRCA2, and prevalence of other breast cancer susceptibility genes. *J. Clin. Oncol.* **20**, 2701–2712 (2002).

29. Tyrer, J., Duffy, S. W. & Cuzick, J. A breast cancer prediction model incorporating familial and personal risk factors. *Stat. Med.* **23**, 1111–1130 (2004).

30. Lee, A. *et al.* BOADICEA: a comprehensive breast cancer risk prediction model incorporating genetic and nongenetic risk factors. *Genet. Med.* **21**, 1708–1718 (2019).

31. Ankerst, D. P. *et al.* Prostate cancer prevention trial risk calculator 2.0 for the prediction of low- vs high-grade prostate cancer. *Urology* **83**, 1362–1368 (2014).

32. Clendenen, T. V. *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.* **21**, 42 (2019).

33. Gourd, E. New advances in prostate cancer screening and monitoring. *Lancet. Oncol.* **21**, 887 (2020).

34. Sud, A., Turnbull, C. & Houlston, R. Will polygenic risk scores for cancer ever be clinically useful? *npj Precis. Oncol. 2021 51* **5**, 1–5 (2021).

35. Kim, H. E. *et al.* Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. *Lancet Digit. Heal.* **2**,

e138–e148 (2020).

36. McKinney, S. M. *et al.* International evaluation of an AI system for breast cancer screening. *Nature* **577**, 89–94 (2020).

37. Akselrod-Ballin, A. *et al.* Predicting breast cancer by applying deep learning to linked health records and mammograms. *Radiology* **292**, 331–342 (2019).

38. Stark, G. F., Hart, G. R., Nartowt, B. J. & Deng, J. Predicting breast cancer risk using personal health data and machine learning models. *PLoS One* **14**, e0226765 (2019).

39. Strobl, A. N. *et al.* Improving patient prostate cancer risk assessment: Moving from static, globally-applied to dynamic, practice-specific risk calculators. *J. Biomed. Inform.* **56**, 87–93 (2015).

40. Wang, X. *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.* **21**, e13260–e13260 (2019).

41. Kinar, Y. *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.* **23**, 879–890 (2016).

42. Abelson, S. *et al.* Prediction of acute myeloid leukaemia risk in healthy individuals. *Nature* **559**, 400–404 (2018).

43. Little, R. J. A. Missing-data adjustments in large surveys. *J. Bus. Econ. Stat.* **6**, 287–296 (1988).

44. van Buuren, S. & Groothuis-Oudshoorn, K. mice: Multivariate imputation by chained equations in R. *J. Stat. Softw.* **45**, 1–67 (2011).

45. Klein, J. P. & Moeschberger, M. L. Censoring and Truncation. in 63–90 (2003). doi:10.1007/0-387-21645-6_3.

46. Pan, W. Rank invariant tests with left truncated and interval censored data. *J. Stat. Comput. Simul.* **61**, 163–174 (1998).

47. Blanche, P., Kattan, M. W. & Gerds, T. A. The c-index is not proper for the evaluation of t-year predicted risks. *Biostatistics* **20**, 347–357 (2019).

48. Wright, M. N. & Ziegler, A. Ranger: A fast implementation of random forests for high dimensional data in C++ and R. *J. Stat. Softw.* **77**, (2017).

49. Ho, D. E., Imai, K., King, G. & Stuart, E. A. MatchIt: Nonparametric preprocessing for parametric causal inference. *J. Stat. Softw.* **42**, 1–28 (2011).

50. Tewari, R. *et al.* Significant association of metabolic indices, lipid profile, and androgen levels with prostate cancer. *Asian Pacific J. Cancer Prev.* **15**, 9841–9846 (2014).

51. LeBlanc, M. & Crowley, J. Survival trees by goodness of split. *J. Am. Stat. Assoc.* **88**, 457–467 (1993).

52. Afrin, K., Illangovan, G., Srivatsa, S. S. & Bukkapatnam, S. T. S. Balanced random survival forests for extremely unbalanced, right censored data. *arXiv* **preprint**, 1803.09177 (2018).

53. Kinar, Y. *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.* **23**, 879–890 (2016).

54. Karnes, R. J. *et al.* Prostate-specific antigen trends predict the probability of prostate cancer in a very large U.S. Veterans affairs cohort. *Front. Oncol.* **8**, 296 (2018).

55. Hayashi, T. *et al.* Serum monocyte fraction of white blood cells is increased in patients with high Gleason score prostate cancer. *Oncotarget* **8**, 35255–35261 (2017).

56. 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.* **67**, 122–133 (2020).

57. Che, Z., Purushotham, S., Cho, K., Sontag, D. & Liu, Y. Recurrent neural networks for multivariate time series with missing values. *Sci. Rep.* **8**, 6085 (2018).

58. Hothorn, T., Lausen, B., Benner, A. & Radespiel-Tröger, M. Bagging survival trees. *Stat. Med.* **23**, 77–91 (2004).

59. Klein, J. P. & Moeschberger, M. L. *Survival Analysis*. (Springer New York, 2003). doi:10.1007/b97377.

FIGURES

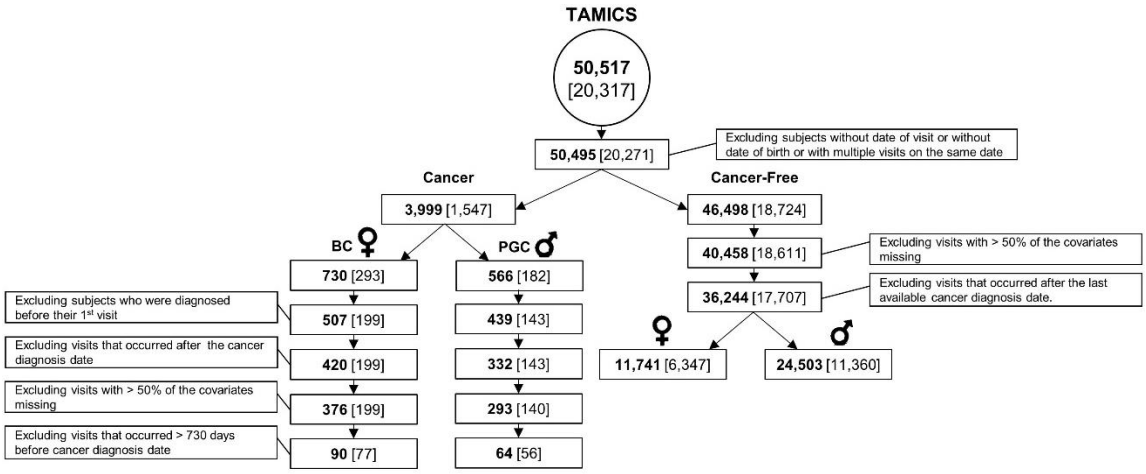


Figure 1: Study design. The bold number is the number of TAMICS visits; the number of individuals appears in parentheses.

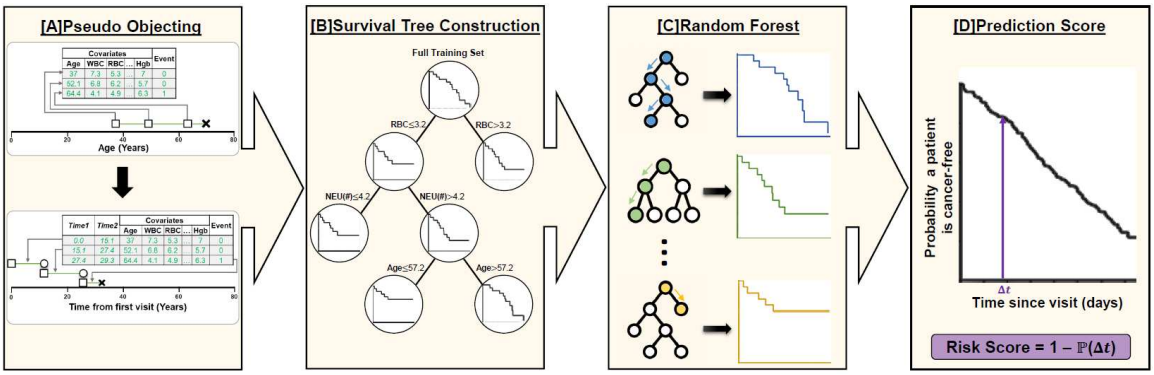


Figure 2: Model construction and evaluation. [A] For each individual we transformed its data into pseudo-objects and changed the time axis to time from first visit. [B] A single survival tree. [C] Generating 500 survival trees. [D] The trees are combined into a single unified model. Risk score calculation for a new sample is the averaged survival curve of the leaves it corresponds to in all trees.

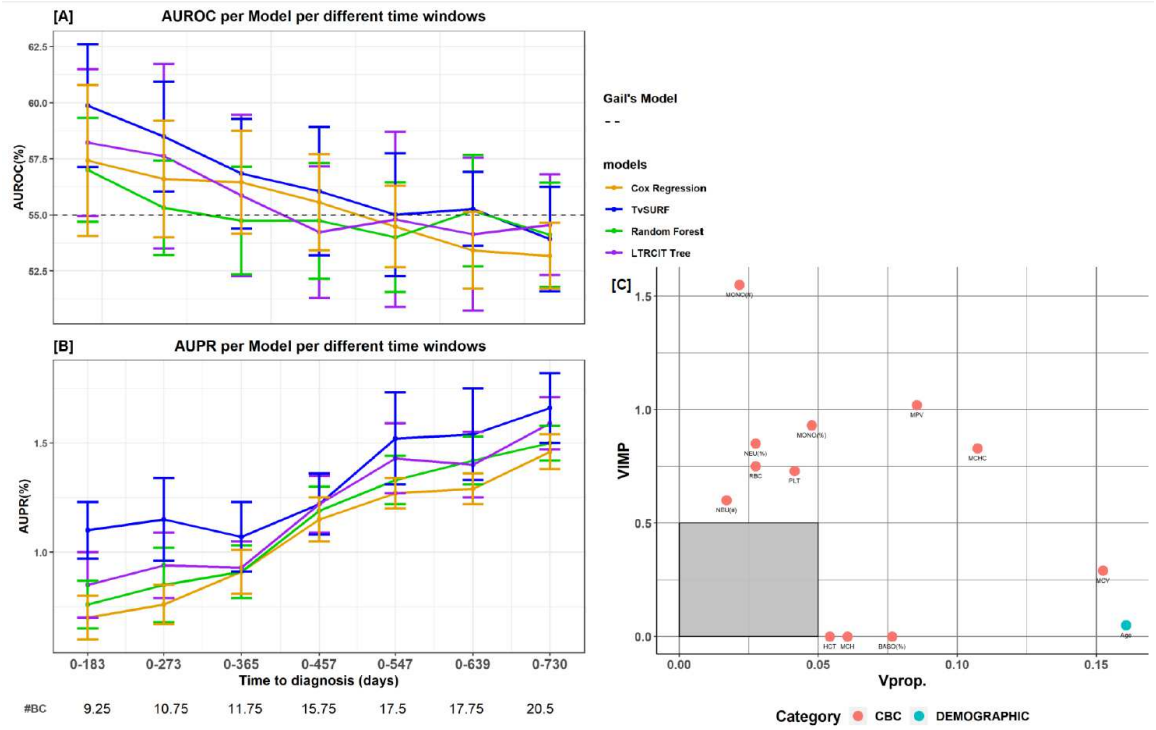


Figure 3: **BC risk prediction and variable importance** [A] AUROC (mean±SD) of five prediction models for different time intervals. The grey dashed line represents the (time-independent) AUROC reported for Gail's Risk factor model [32]. [B] AUPR (mean±SD). The numbers below the x-axis labels are the average number of BC individuals that were available across the cross-validation folds for each time interval. [C] Variable importance for model prediction in the 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. Covariates of low importance (Vprop <0.05 and VIMP<0.5) are not shown.

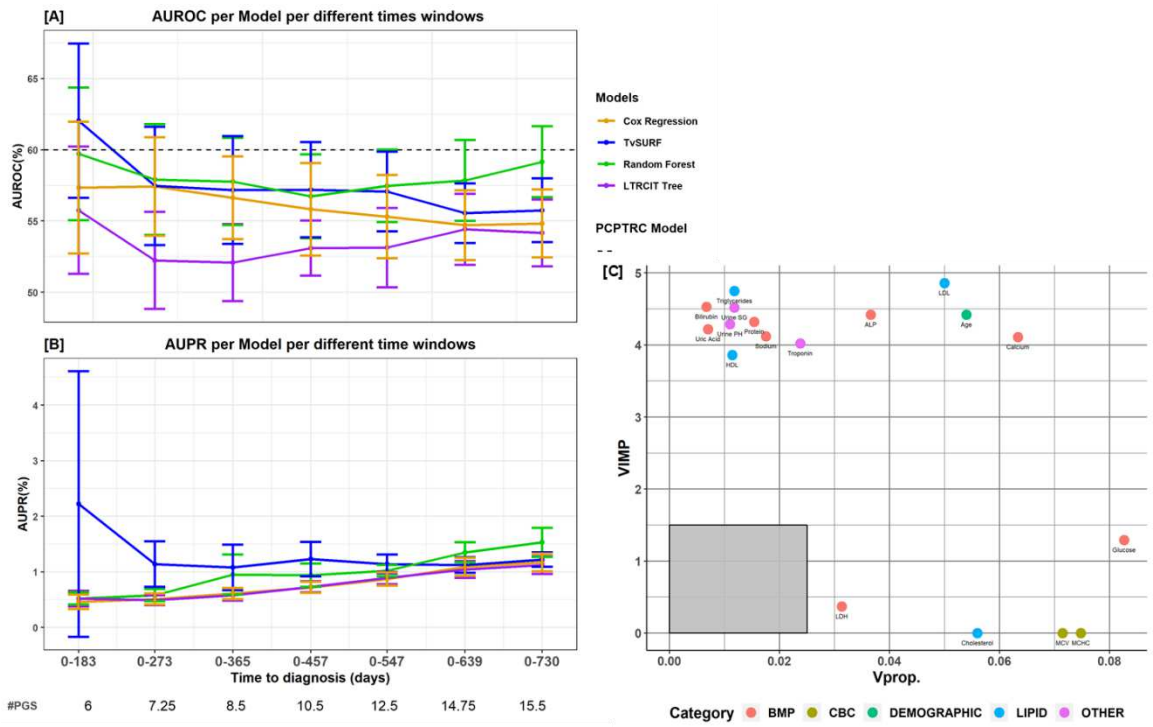
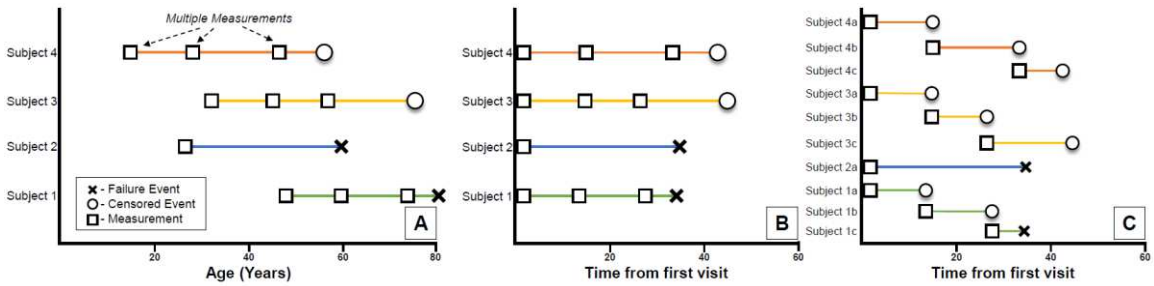


Figure 4: PGC risk prediction and variable importance. [A] Performance (AUROC mean±SD) of five prediction models for different time windows. The grey dashed line represents the (time-independent) AUROC previously reported for the PCPTRC model. [B] AUPR. The numbers below the x-axis labels are the average number of individuals with PGC that were available across the cross-validation folds for each time interval. [C] Variable importance for model prediction in the 183-day window. Points indicate the different variables. Axis definitions are as in Figure 3. The color of a point represents the variable's category. Covariates of low importance (Vprop <0.025 and VIMP<1.5) are not shown.



<u>Figure 5</u>: Transforming longitudinal data of multiple visits.

(A) Longitudinal measurements and survival analysis setting. Squares indicate the times of the 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.

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 ← ∞
$H \leftarrow \emptyset$;
for $m = 1$ *to M* **do**
  1. h_m ← BuildTree(D,K)
  2. $H \leftarrow H \cup \{h_m\}$
end
return H

Figure 6: BuildTree and TVsuRF Algorithms.

TABLES

	BC			BC-Free			Matched BC-Free			BC vs. BC-Free P-value		BC vs. Matched BC-Free P-value	
Parameter	**Visits**	**Individuals**	**Mean±STD**	**Visits**	**Individuals**	**Mean±STD**	**Visits**	**Individuals**	**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 Student's 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**	**Individuals**	**Mean+STD**	**Visits**	**Individuals**	**Mean+STD**	**Visits**	**Individuals**	**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 urine test.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFile.pdf