Risk Modeling of Time-Varying Covariates Using an Ensemble of Survival Trees: Predicting Future Cancer Events

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] Faculty of Medicine, Tel-Aviv University, Tel-Aviv, Israel
rshamir@tau.ac.il, dancoster@gmail.com

Abstract

The challenge of survival prediction is ubiquitous in medicine, but only a handful of methods are available for survival prediction based on 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 demonstrate 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 analyze electronic medical records of 20,317 healthy individuals who underwent routine check-ups 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, random survival forest, cox-regression model, dynamic deep-hit and a single survival tree. Our work proposes a new framework for survival risk prediction in time-varying data and 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 clinical research, survival models are widely employed. In this work, given a subject's temporal covariate values, we wish to develop a model for predicting the subject's risk of experiencing a failure event over time, or equivalently, to estimate the subject's survival probability function. In healthcare, the covariates could be clinical measurements such as vital signs, laboratory measurements and demographics, and the failure event is a particular medical outcome (death, disease onset, hospitalization, etc.). The estimation is particularly challenging since for many patients the outcomes are censored, i.e., they are unknown due to lack of follow-up information.

In this study, we introduce a method called TVsuRF (Time-Varying Survival Random Forest) for survival prediction using time-varying covariates. Our method relies solely on the proportional-hazard assumption, as opposed to traditional survival analysis methods, which make prior assumptions about the distribution of the data. Our approach combines survival trees, pseudo-objects and ensemble methods. We show that each of these three components is essential for the improvement achieved by our method. Our novel method is the first to use conditional inference trees in this setting. We evaluate TVsuRF by predicting the future risk of breast and prostate cancers in healthy individuals. In contrast to traditional survival analysis models, where the predicted risk is modeled as a function of baseline (time-independent) covariates, our method uses also values of covariates measured along time.

Early detection of cancer is crucial for providing appropriate care and can improve both prognosis and survival (Loomans-Kropp and Umar 2019; Adamson and Welch 2019; Wiens et al. 2019; Crosby et al. 2020). Today, screening tests in the healthy population are used to identify individuals with cancer and without symptoms, but these tests are costly, labor-intensive, and suffer from low accuracy. The current strategies for early detection of cancer use screening tests that require substantial resources. For example, Serum Prostate-Specific Antigen (PSA) level is used for detecting Prostate Gland Cancer (PGC). Mammography, an X-ray modality, is used for detecting early signs of Breast Cancer (BC), and clinical breast examination (CBE), a physical examination is used to recognize abnormalities in the breast (Bancej et al. 2003). Other approaches to assess cancer risk use models, e.g. Gail's model (Gail et al. 1989; Banegas et al. 2017), BRCAPRO (Berry et al. 2002), IBIS (Tyrer, Duffy, and Cuzick 2004) and BOADICEA (Lee et al. 2019) for BC risk, and the Prostate Cancer Prevention Trial Risk Calculator (PCPTRC) for PGC risk (Ankerst et al. 2014). These models use a few clinical and genetic parameters, do not use routine laboratory measurements, and have limited performance (Clendenen et al. 2019). Advanced ge-

netic methods, including polygenic risk scores, have been proposed for screening, but they are still not part of clinical practice (Gourd 2020; Sud, Turnbull, and Houlston 2021).

In recent years, machine learning has improved screening approaches in two major ways. First, it enhanced existing screening tests, including mammography analysis (Kim et al. 2020; McKinney et al. 2020; Akselrod-Ballin et al. 2019), Gail's model (Stark et al. 2019), and the PGC risk score (Strobl et al. 2015) by applying better computational models. Second, a new set of cancer risk prediction tools was developed based on patients' historical Electronic Medical Records (EMR) collected over time as part of routine care. Such models were suggested for lung cancer (Wang et al. 2019), colorectal cancer (Kinar et al. 2016), and Acute Myeloid Leukemia (Abelson et al. 2018), among others, but to the best of our knowledge none of these studies used clinical measures of apparently healthy individuals, and this study presents the first risk model that is based on routine clinical measures proposed for these cancer types.

Related Work

Survival trees for time-varying covariates: (Gordon and Olshen 1985) introduced the survival tree, 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 of the methods which are based on survival trees addressed right-censored data and time-independent covariates (Bou-Hamad, Larocque, and Ben-Ameur 2011b). Incorporating time-varying covariates in survival trees was first done by introducing 'pseudo-objects' (Bacchetti and Segal 1995), a concept adopted later in other studies (Huang, Chen, and Soong 1998; Bou-Hamad, Larocque, and Ben-Ameur 2011a; Wallace 2014; Fu and Simonoff 2017).

An ensemble of survival trees: Several ensemble methods for survival tree analysis were suggested for right-censored data and time-independent covariates (Hothorn et al. 2006; Bellot and van der Schaar 2018). Random survival forests (RSF), introduced by (Ishwaran et al. 2008), combined Breiman's random forest (RF) (Breiman 2001; Ishwaran and Kogalur 2007), survival trees and the log-rank test as the splitting criterion. Another survival trees ensemble model utilized conditional inference trees, which employ hypothesis testing to select the splitting covariates and thresholds and also as a stopping criterion (Hothorn, Hornik, and Zeileis 2006). In subsequent improvements of those methods, (Utkin et al. 2019) optimized the weights of each tree, and (Steingrimsson, Diao, and Strawderman 2019) proposed generalized weighted bootstrap procedures. There is limited work on ensembles of survival trees in the time-varying covariates setting. (Sun, Chiou, and Wang 2020) suggested to use for tree-building a splitting criteria based on time-dependent Receiver Operating Curve (ROC) and (Wongvibulsin, Wu, and Zeger 2020) proposed an adjusted model of RSF.

Time-Varying Survival Prediction Models: Several methods for survival prediction models using time-dependent covariates have been suggested. Joint modeling (Tsiatis and Davidian 2004) attempts to jointly model the longitudinal markers and the event time. Landmarking (Van Houwelingen 2007) sets a landmark time point and uses the value of the time-varying covariate at this point as a time-independent covariate in an analysis of survival from this point onwards, for the subset of subjects at risk at this time point. Cox proportional hazard model (Therneau, Crowson, and Atkinson 2017; Andersen and Gill 1982) can utilize these methods in time-varying setting. Recently, new deep learning techniques for personalized survival prediction were suggested utilizing attention and RNN models with adjusted loss functions for the survival analysis setting (Nagpal, Jeanselme, and Dubrawski 2021; Sun et al. 2021; Jarrett, Yoon, and van der Schaar 2019).

Methods

TVsuRF: We developed a novel method called TVsuRF (time-varying survival random forest) for risk prediction. It is trained on multivariate time series data and employs a random forest of survival trees for left-truncated and right-censored intervals generated by pseudo-objecting of the data. Given query data of a new individual, the method predicts the individual's risk for failure event for different time horizons.

To describe the method, we start with some notation. Consider data of N individuals, where for each of them data from one or more records is available. Individual i had M^i records at times $t_1^i < \ldots < t_{M^i}^i$. 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 measured in every record). Note that covariates can be either time-dependent or time-independent. Hence, $X^i = (x^i(t_1^i), \ldots, x^i(t_{M^i}^i))$ 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 $\mathcal{D} = (X^i, \tau^i, \delta^i)_{i=1}^N$ (Figure 1[A](i)). Denote by $X^i(t)$ the data of individual i that was measured until time t, i.e., $X^i(t) = \{x^i(t_j^i) : 0 \leq t_j^i \leq t\}$. 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 records were at time 0 (Figure 1[A](ii)). We also assume that the age of the individual at each record is one of the covariates.

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

$$S\left(t|x^i(t_*^i)\right) = P\left(\tau^i > t|x^i(t_*^i)\right)$$

In order to model the time-dependent covariates where there are multiple records from different time per subject, we transform the data into pseudo-objects (Bacchetti and Segal 1995). 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 1[A](iii)). In that manner, we consider t_j as the left-truncation time. If $[t_j^i, t_{j+1}^i)$ is

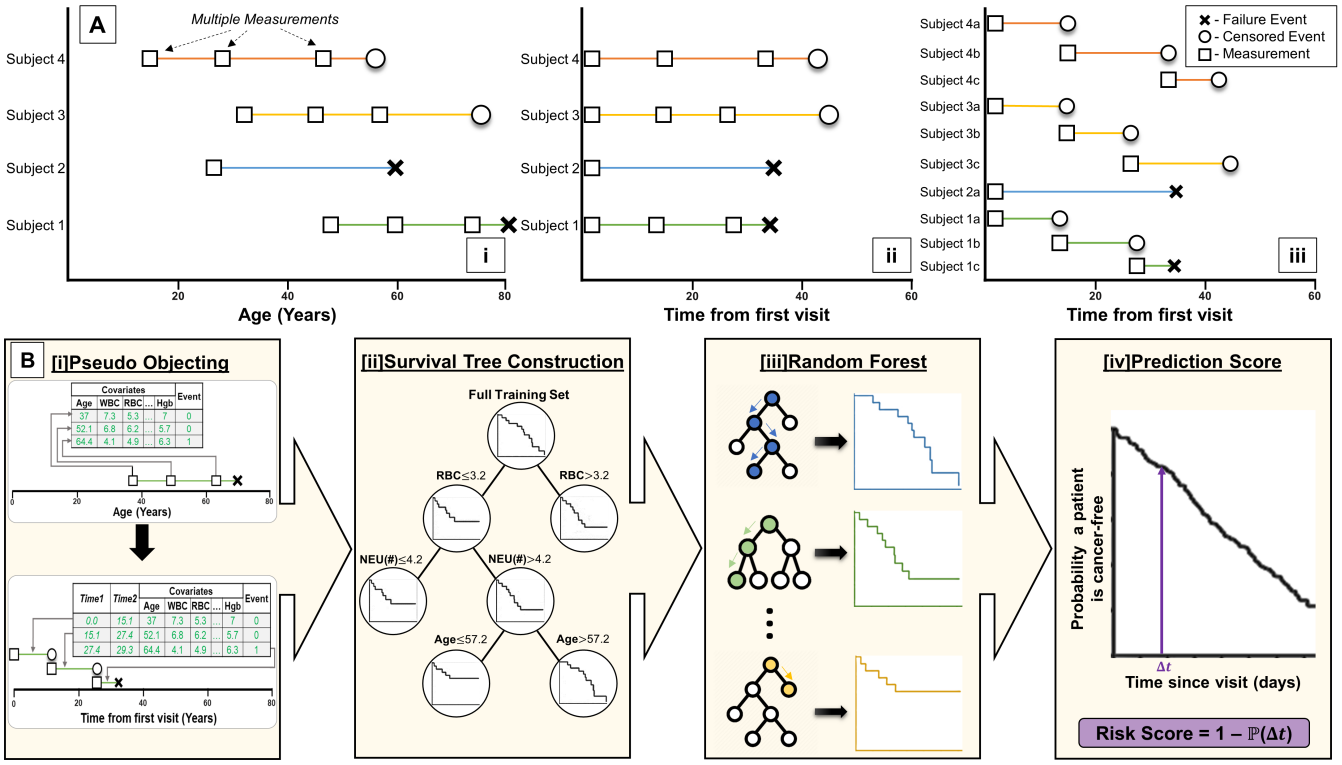

Figure 1: [A] Transforming longitudinal data of multiple visits. (i)Longitudinal measurements and survival analysis setting. Squares indicate the times of the measurements, Crosses indicate failure events, and circles indicate censoring events. (ii) The data after shifting all first visit times to 0. (iii)The same data after transforming into pseudo-objects. [B] Model construction and evaluation. (i)For each individual we transformed its data into pseudo-objects and changed the time axis to time from first visit. (ii)A single survival tree. (iii) Generating 500 survival trees. (iv)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.

not the last interval of individual i then we view time t^i_{j+1} as censoring time. We denote the pseudo-object of the j^{th} interval of individual i as $[L^i_j, R^i_j)$ where $L^i_j = t^i_j$ and

$$R^i_j = \begin{cases} t^i_{j+1} & \text{if } 1 \le j < M^i \\ \tau^i & \text{otherwise} \end{cases}$$

$$\delta^i_j = \begin{cases} 0 & \text{if } 1 \le j < M^i \\ \delta^i & \text{otherwise} \end{cases},$$

Hence, the transformation is: $\left(X^i, \tau^i, \delta^i\right) \rightarrow \left\{\left(t^i_1, t^i_2, \delta^i_1, x^i(t^i_1)\right), \left(t^i_2, t^i_3, \delta^i_2, x^i(t^i_2)\right), \ldots\right\} = \left\{\left(L^i_1, R^i_1, \delta^i_1, x^i(t^i_1)\right), \left(L^i_2, R^i_2, \delta^i_2, x^i(t^i_2)\right), \ldots\right\}$

The standard Kaplan-Meier (KM) estimator of the survival function can now be generalized for left-truncated and right-censored (LTRC) data (Klein and Moeschberger 2003), as follows. Assume that there were D failure events and they occurred at distinct times $t_1 < \ldots < t_D$. We denote by Y_j the number of pseudo-objects at risk at time t_j, $Y_j = \sum_{i=1}^{N} \sum_{k=1}^{M^i} (L^k_i \le t_j \le R^k_i)$ i.e., the number of individuals who entered the study before time t_j and did not experience a failure or censoring event until $t_j \cdot d_j$ 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:

$$\hat{S}(t) = \begin{cases} 1 & \text{if } t < t_1 \\ \prod_{t_j \le t} \left[1 - \frac{d_j}{Y_j}\right] & \text{Otherwise} \end{cases}$$

The probability of an individual whose record was obtained at time t^i_j to be free of a failure event in the next time window of size Δt, is $P\left(\tau^i > t^i_j + \Delta t | x_i(t^i_j)\right)$.

Survival tree construction: We now describe the construction of the survival tree for the data (Figure 1[B](iii)). Given data $\mathcal{D} = \left\{\left(X^i, \tau^i, \delta^i\right)\right\}_{i=1}^{N}$ as described above, we transform $\mathcal{D}$ to pseudo-objects, and wish to build a binary decision tree such that the two subgroups corresponding to the children of each node will have survival functions that are as different as possible. We use the framework of conditional inference trees (Hothorn, Hornik, and Zeileis 2006), which employs a statistical hypothesis test in order to split nodes. This process is different from common decision tree construction, where usually an information measure (e.g., Gini or entropy) is used.

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 (Pan 1998), as suggested also in (Fu and Simonoff 2017). In every node, we test all possible covariates and thresholds, and the one that produces the split with the lowest p-value is selected. 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. U_i is based on the maximum likelihood estimator under the proportional hazard assumption, and one can use the sum of U_j-s for the samples in a node to test the null hypothesis that there is no difference between the survival functions of the populations of the two nodes. This test is equivalent to the log-rank test. Notice that pseudo-objects created from the same individual can end in distinct subnodes. 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 $\leq X$ is $P_{value} = k/\binom{N}{n}$. We estimate k using 1000 permutations. The survival function $\hat{S}_l(t)$ for node l is the KM curve for the samples corresponding to that node. Let C_l be the set on indices of samples in node l, then:

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

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.

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 no difference in survival, if that difference is significant. The predicted survival curve for a new sample 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 samples that are in the i^{th} leaf of the k^{th} tree and let $C_F = \cup \left\{C(l_i^k) | \omega \in l_i^k\right\}$ be the multiset of all the samples 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 samples in C_F in risk at time t_i, then the survival function of ω is (Figure 1[B](iv)):

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

This gives the risk score of individual ω over time. The algorithm is summarized in Algorithm 1 and Algorithm 2.

Variable Importance: We assessed the importance of each covariate in our model in two ways. First, we computed the fraction of internal nodes in all trees for which the covariate was used to split the node. We call this fraction Vprop; higher Vprop indicates more importance. Second, for each covariate, we replaced its values in the data 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 between the AUROC on the original and the modified data was computed for ten random assignments on every fold of the 4-fold cross-validation as done in (Ishwaran et al. 2008). We repeated this process 20 times and defined VIMP as the mean difference obtained. Again, higher VIMP indicates more importance.

Computational Complexity: To compute the complexity of TVsuRF, assume covariate values are integers and that logrank p-value can be computed in $\mathcal{O}(1)$ using a lookup table. Let N be the number of observations (pseudo objects), k the total number of covariates, K the number of covariates considered for splitting per tree, and M the number of trees. For constructing a survival tree, per each node, we will need to sort all the covariates that are considered. We can perform this operation only once for the full data, in time $\mathcal{O}(kN\log N)$. Using the sorted data, for the construction of a survival tree, per each node we need to examine the

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

$randomFeatures \leftarrow$ random subset of K features
$minP-Value \leftarrow \infty$
$minFeature \leftarrow$ NULL
for $feature$ in $randomFeatures$ **do**
  $featureUniqueValues \leftarrow$ all the unique values of the feature
  for val in $featureUniqueValues$ **do**
    D_l,D_r = induced sub-datasets from D based on (val, $feature$)
    $P-value \leftarrow$ LogRankScore(D_l,D_r)
    if $P-value < minP-Value$ **then**
      $minP-value \leftarrow P-value$
      $minFeature \leftarrow feature$
      $featureVal \leftarrow val$
    end if
  end for
end for
if $minP-value > 0.05$ **then**
  break
else
  BuildTree(D_l,K)
  BuildTree(D_r,K)
end if

Algorithm 2: TVsuRF (D,K,M)

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

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

best split of the K covariates, which will take $\mathcal{O}(KN)$. As a tree can have at most $N+1$ nodes, the overall construction of a tree will take $\mathcal{O}(KN^2)$. As we construct M trees the overall complexity is $\mathcal{O}(MKN^2)$ + $\mathcal{O}(kN\log N)$ which is $\mathcal{O}(MkN^2)$.

Data

Cohort: We analyzed data from routine checkups of individuals at the Tel-Aviv Sourasky Medical Center. The individuals enrolled in a large screening program (the medical center's executive health program), and their checkups were typically scheduled three months in advance, without regard to any special medical condition and were performed mostly due to an employer-provided benefit. Enrollees were asked to arrive only if they felt healthy, and postpone if they felt sick. Participants were men and non-pregnant women with no active current malignant or infectious disease and signed an informed consent form. All subjects underwent the same medical evaluation in each visit, which includes a comprehensive medical review of their medical history, a complete physical examination, blood and urine tests, vital signs measurements, a respiratory function test, etc.. Some individuals had multiple visits over several years. We conducted a retrospective analysis of the 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). We identified the individuals who later developed cancer using the National Cancer Registry (NCR) and applied certain inclusion and exclusion criteria (Appendix A). There are no links between these data and the hospitalization data in the EMR system at the medical center. An evaluation of the clinical utility of our model was carried out on a subset of our cohort that underwent standard BC screening tests (CBE and mammography) and PGC screening tests (PSA).

Covariates Selection: 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 (Little 1988) using the *mice* package (Van Buuren and Groothuis-Oudshoorn 2011). For BC risk prediction we used 20 covariates (Appendix B) 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 (Appendix C).

Results

Evaluation Approach: We tested several models that can 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 record 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 $\varepsilon = 31$ days to Δt. The risk for individual i is thus $1-\hat{S}(t+\Delta t+\varepsilon|x^i(t))$ where $\hat{S}$ is the predicted survival function (Figure 1[B](iv)). 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 $\Delta t+\varepsilon$ days as suggested in (Blanche, Kattan, and Gerds 2019), 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 but keeping all pseudo-objects of a subject in the same fold. This is aimed to simulate a real-world situation, where after each visit to the screening center a prediction can be made. For each of the above measures, we calculated the average and standard deviation.

We tested TVsuRF and five other models: (1) Cox regression model adapted to time-varying covariates (Therneau, Crowson, and Atkinson 2017; Andersen and Gill 1982). (2) A single LTRC survival tree as in (Fu and Simonoff 2017) (denoted LTRCIT), (3) RF model (Breiman 2001). 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 $\Delta t+\varepsilon$ 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 (Wright and Ziegler 2015). (4) 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. (5) The dynamic deep-hit model, which incorporates longitudinal data to predict a survival curve (Lee, Yoon, and Van Der Schaar 2019). The hyperparameters of the model such as the coefficients, the activation functions, and the number of hidden layers and nodes of each subnetwork were determined using a grid search over predefined ranges of possible values (See Appendix D).

Predicting Breast Cancer Risk: Our cohort contained data on 6,424 women with a total of 11,831 visits. Out of those, 77 were diagnosed with BC and had one or more visits less than 730 days before the diagnosis date (90 visits in total; this group was denoted as the BC group). Further details are in Appendix B.

The performance of each of the methods tested, for different time intervals, is summarized in Figures 2A and 2B. We also marked the AUROC of Gail's breast cancer risk estimation for 5 years horizon as reported in (Clendenen et al. 2019). 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 (Appendix E). See Appendix F re the increasing values of AUPR.

Appendix G 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), mean

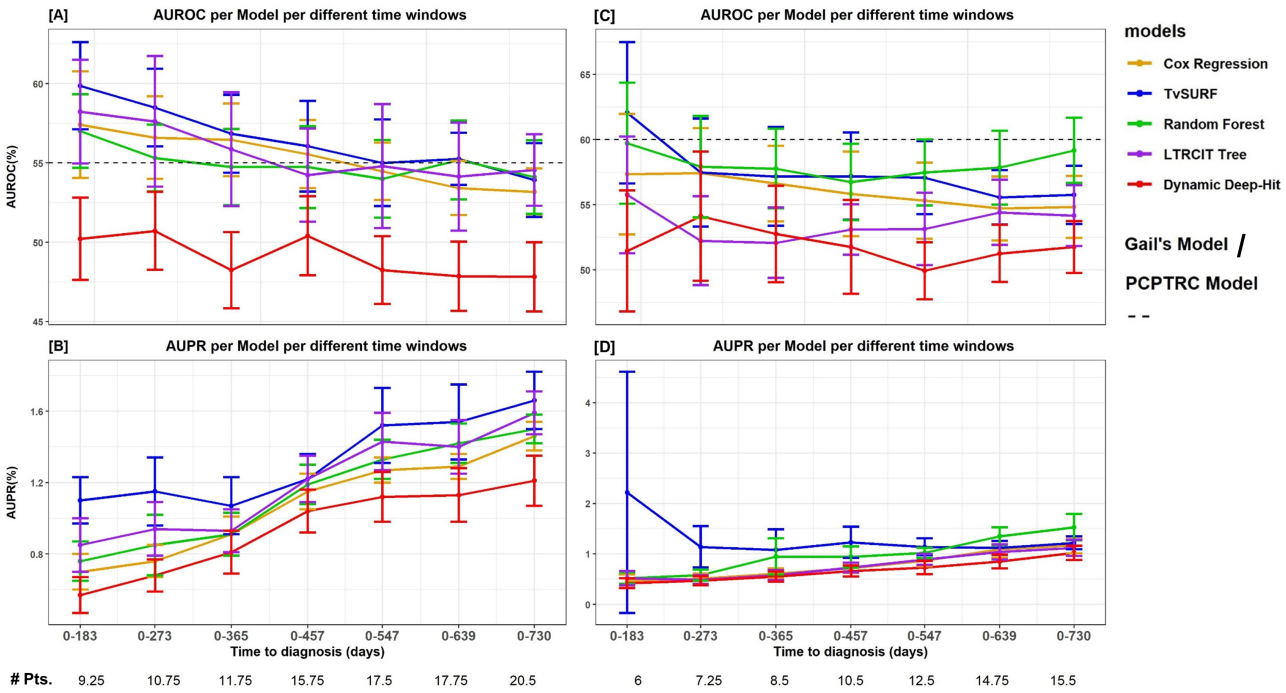

Figure 2: [A] AUROC (mean±SD) of five BC prediction models for different time intervals. The dashed line represents the (time-independent) AUROC for Gail's Risk model (Clendenen et al. 2019).[B] AUPR (mean±SD). #Pts. is the average number of BC individuals that were available across the cross-validation folds for each time interval. [C,D] PGC risk prediction, the dashed line represents the (time-independent) AUROC for PCPTRC model(Ankerst et al. 2014)

corpuscular hemoglobin concentration (MCHC) and age were most important in the model. The importance of immune system-related covariates such as MONO might reflect the fact BC is an inflammatory and systemic disease.

For the clinical impact of our model, we evaluated its performance on a subgroup who underwent standard BC screening tests (CBE and mammography).Note that the results are not directly comparable, as mammography and CBE identify current malignancy and TVsuRF computes future disease risk. Our model achieved 18% sensitivity and 94% specificity for BC subjects with normal CBE (N=1,975). Our model was 60% sensitive and 68% specific for BC subjects with normal mammography results (N=504). The method also gave very high-risk scores to several women who tested negative and later developed 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 (Appendix H).

Predicting Prostate Gland Cancer Risk: This cohort consisted of 11,416 males who made a total of 24,567 visits. 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 group. Further details are in Appendix C.

Figures 2C and 2D show the results of five prediction methods, using the same comparison metrics used for BC. Our model had the highest AUROC in the prediction window 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. See Appendix F re the increasing values of AUPR. When testing variants of RSF, TVsuRF had better performance for 0-183 days, but less for longer time windows (Appendix E).

Appendix G 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 - LDL, high-density lipoprotein (HDL), cholesterol and triglycerides - had high importance risk according to at least one criterion, in agreement with previous reports(Tewari et al. 2014). We evaluated the clinical impact of TVsuRF on a subgroup of 1,918 subjects with normal PSA levels, and TVsuRF showed 40% sensitivity and 84% specificity.

Discussion

In this study we introduced a new method for survival prediction based on time-varying covariates utilizing an ensemble of survival trees, and applied it for predicting a future emergence of BC and PGC. Our approach disposes of the time-independence assumption of the established RSF model (Ishwaran et al. 2008). Unlike traditional survival analysis methods, which use prior assumptions concerning the distribution of the data (LeBlanc and Crowley 1993), our

method relies only on the proportional-hazard assumption. The same data was collected to all the subjects in our cohort and since it did not depend on the subject's medical condition, we could avoid ubiquitous hidden confounders that stem from the physician's choice of data to be collected for each patient, and identify real cancer risk among apparently healthy subjects. Our results could enable population cancer screening at a low cost, before ad-hoc expensive and labor-intensive tests are performed.

Our study has several methodological 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 ((Afrin et al. 2018). That could affect the splitting criteria and produce nodes with a small number of samples or nodes without failure events. Second, the limited cohort size made it difficult to evaluate the calibration of our model and extend it for competing risks (e.g. death). Moreover, the small number of records per individual did not allow us to use time-related features (Kinar et al. 2016; Karnes et al. 2018), feature interactions (Hayashi et al. 2017), 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 (Lee, Yoon, and Van Der Schaar 2019). Additionally, we used the Predictive-Mean-Matching imputation method (Little 1988), which was shown to be effective in analyzing EMR data (Beaulieu-Jones et al. 2018). Further work should examine the effect of additional imputation methods on model performance (Che et al. 2018). Furthermore, to handle settings with a large feature space (both in terms of number of features and in terms of possible values per feature), future works should investigate the effect of selecting a subset of possible thresholds evaluated for splitting a feature randomly, as proposed by (Ishwaran et al. 2008). Additionally, since our data did not include all 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 study was limited to a single medical center and two cancer types. We are currently expanding the study to additional cohorts from other screening centers to overcome these limitations. Finally, a prospective clinical trial would provide a more accurate evaluation of the performance and clinical utility.

By combining survival trees, pseudo-objects, and ensemble methods, TVsuRF achieves better performance. Each of these parts improves the final outcome. The advantage of survival trees was illustrated by comparing them to RF; the advantage of using pseudo-objects in TVsuRF was demonstrated by comparing it with standard RSF models; and the advantage of ensemble methods was shown by comparing to a single survival tree. TVsuRF also performed better than dynamic deep-hit models, which incorporate longitudinal data to predict a survival curve. This may be due to the relatively small size of the dataset and to its low dimension. Furthermore, TVsuRF outperformed traditional prediction methods in BC and for short-term prediction also in PGC, and demonstrated the potential of using common laboratory tests of apparently healthy individuals to assess cancer risk. Such predictions can serve as additional screening tests, complementing current screening methods.

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

Code Availability. The code for TVsuRF is in Github: https://github.com/Shamir-Lab/TVsuRF.

References

Abelson, S.; Collord, G.; Ng, S. W.; Weissbrod, O.; Mendelson Cohen, N.; Niemeyer, E.; Barda, N.; Zuzarte, P. C.; Heisler, L.; Sundaravadanam, Y.; et al. 2018. Prediction of acute myeloid leukaemia risk in healthy individuals. *Nature*, 559(7714): 400–404.

Adamson, A. S.; and Welch, H. G. 2019. Machine learning and the cancer-diagnosis problem-no gold standard. *The New England journal of medicine*, 381(24): 2285–2287.

Afrin, K.; Illangovan, G.; Srivatsa, S. S.; and Bukkapatnam, S. T. 2018. Balanced random survival forests for extremely unbalanced, right censored data. *arXiv preprint arXiv:1803.09177*.

Akselrod-Ballin, A.; Chorev, M.; Shoshan, Y.; Spiro, A.; Hazan, A.; Melamed, R.; Barkan, E.; Herzel, E.; Naor, S.; Karavani, E.; et al. 2019. Predicting breast cancer by applying deep learning to linked health records and mammograms. *Radiology*, 292(2): 331–342.

Andersen, P. K.; and Gill, R. D. 1982. Cox's regression model for counting processes: a large sample study. *The annals of statistics*, 1100–1120.

Ankerst, D. P.; Hoefler, J.; Bock, S.; Goodman, P. J.; Vickers, A.; Hernandez, J.; Sokoll, L. J.; Sanda, M. G.; Wei, J. T.; Leach, R. J.; et al. 2014. Prostate Cancer Prevention Trial risk calculator 2.0 for the prediction of low-vs high-grade prostate cancer. *Urology*, 83(6): 1362–1368.

Bacchetti, P.; and Segal, M. R. 1995. Survival trees with time-dependent covariates: application to estimating changes in the incubation period of AIDS. *Lifetime data analysis*, 1(1): 35–47.

Bancej, C.; Decker, K.; Chiarelli, A.; Harrison, M.; Turner, D.; and Brisson, J. 2003. Contribution of clinical breast examination to mammography screening in the early detection of breast cancer. *Journal of Medical Screening*, 10(1): 16–21.

Banegas, M. P.; John, E. M.; Slattery, M. L.; Gomez, S. L.; Yu, M.; LaCroix, A. Z.; Pee, D.; Chlebowski, R. T.; Hines, L. M.; Thompson, C. A.; et al. 2017. Projecting individualized absolute invasive breast cancer risk in US Hispanic women. *Journal of the National Cancer Institute*, 109(2): djw215.

Beaulieu-Jones, B. K.; Lavage, D. R.; Snyder, J. W.; Moore, J. H.; Pendergrass, S. A.; and Bauer, C. R. 2018. Characterizing and managing missing structured data in electronic health records: data analysis. *JMIR medical informatics*, 6(1): e8960.

Bellot, A.; and van der Schaar, M. 2018. Boosted trees for risk prognosis. In *Proceedings of the 3rd Machine Learning for Healthcare Conference*, volume 85, 2–16. PMLR.

Berry, D. A.; Iversen Jr, E. S.; Gudbjartsson, D. F.; Hiller, E. H.; Garber, J. E.; Peshkin, B. N.; Lerman, C.; Watson, P.; Lynch, H. T.; Hilsenbeck, S. G.; et al. 2002. BRCAPRO validation, sensitivity of genetic testing of BRCA1/BRCA2, and prevalence of other breast cancer susceptibility genes. *Journal of Clinical Oncology*, 20(11): 2701–2712.

Blanche, P.; Kattan, M. W.; and Gerds, T. A. 2019. The c-index is not proper for the evaluation of-year predicted risks. *Biostatistics*, 20(2): 347–357.

Bou-Hamad, I.; Larocque, D.; and Ben-Ameur, H. 2011a. Discrete-time survival trees and forests with time-varying covariates: application to bankruptcy data. *Statistical Modelling*, 11(5): 429–446.

Bou-Hamad, I.; Larocque, D.; and Ben-Ameur, H. 2011b. A review of survival trees. *Statistics surveys*, 5: 44–71.

Breiman, L. 2001. Random forests. *Machine learning*, 45(1): 5–32.

Che, Z.; Purushotham, S.; Cho, K.; Sontag, D.; and Liu, Y. 2018. Recurrent neural networks for multivariate time series with missing values. *Scientific reports*, 8(1): 1–12.

Clendenen, T. V.; Ge, W.; Koenig, K. L.; Afanasyeva, Y.; Agnoli, C.; Brinton, L. A.; Darvishian, F.; Dorgan, J. F.; Eliassen, A. H.; Falk, R. T.; et al. 2019. Breast cancer risk prediction in women aged 35–50 years: impact of including sex hormone concentrations in the Gail model. *Breast Cancer Research*, 21(1): 1–12.

Crosby, D.; Lyons, N.; Greenwood, E.; Harrison, S.; Hiom, S.; Moffat, J.; Quallo, T.; Samuel, E.; and Walker, I. 2020. A roadmap for the early detection and diagnosis of cancer. *The Lancet Oncology*, 21(11): 1397–1399.

Fu, W.; and Simonoff, J. S. 2017. Survival trees for left-truncated and right-censored data, with application to time-varying covariate data. *Biostatistics*, 18(2): 352–369.

Gail, M. H.; Brinton, L. A.; Byar, D. P.; Corle, D. K.; Green, S. B.; Schairer, C.; and Mulvihill, J. J. 1989. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *JNCI: Journal of the National Cancer Institute*, 81(24): 1879–1886.

Gordon, L.; and Olshen, R. A. 1985. Tree-structured survival analysis. *Cancer treatment reports*, 69(10): 1065–1069.

Gourd, E. 2020. New advances in prostate cancer screening and monitoring. *The Lancet Oncology*, 21(7): 887.

Hayashi, T.; Fujita, K.; Tanigawa, G.; Kawashima, A.; Nagahara, A.; Ujike, T.; Uemura, M.; Takao, T.; Yamaguchi, S.; and Nonomura, N. 2017. Serum monocyte fraction of white blood cells is increased in patients with high Gleason score prostate cancer. *Oncotarget*, 8(21): 35255.

Hothorn, T.; Bühlmann, P.; Dudoit, S.; Molinaro, A.; and Van Der Laan, M. J. 2006. Survival ensembles. *Biostatistics*, 7(3): 355–373.

Hothorn, T.; Hornik, K.; and Zeileis, A. 2006. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics*, 15(3): 651–674.

Huang, X.; Chen, S.; and Soong, S.-j. 1998. Piecewise exponential survival trees with time-dependent covariates. *Biometrics*, 1420–1433.

Ishwaran, H.; and Kogalur, U. B. 2007. Random survival forests for R. *R news*, 7(2): 25–31.

Ishwaran, H.; Kogalur, U. B.; Blackstone, E. H.; and Lauer, M. S. 2008. Random survival forests. *The annals of applied statistics*, 2(3): 841–860.

Jarrett, D.; Yoon, J.; and van der Schaar, M. 2019. Dynamic prediction in clinical survival analysis using temporal convolutional networks. *IEEE journal of biomedical and health informatics*, 24(2): 424–436.

Karnes, R. J.; MacKintosh, F. R.; Morrell, C. H.; Rawson, L.; Sprenkle, P. C.; Kattan, M. W.; Colicchia, M.; and Neville, T. B. 2018. Prostate-specific antigen trends predict the probability of prostate cancer in a very large US Veterans affairs cohort. *Frontiers in Oncology*, 8: 296.

Kim, H.-E.; Kim, H. H.; Han, B.-K.; Kim, K. H.; Han, K.; Nam, H.; Lee, E. H.; and Kim, E.-K. 2020. Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. *The Lancet Digital Health*, 2(3): e138–e148.

Kinar, Y.; Kalkstein, N.; Akiva, P.; Levin, B.; Half, E. E.; Goldshtein, I.; Chodick, G.; and Shalev, V. 2016. 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. *Journal of the American Medical Informatics Association*, 23(5): 879–890.

Klein, J. P.; and Moeschberger, M. L. 2003. *Survival analysis: techniques for censored and truncated data*, volume 2. Springer.

LeBlanc, M.; and Crowley, J. 1993. Survival trees by goodness of split. *Journal of the American Statistical Association*, 88(422): 457–467.

Lee, A.; Mavaddat, N.; Wilcox, A. N.; Cunningham, A. P.; Carver, T.; Hartley, S.; Babb de Villiers, C.; Izquierdo, A.; Simard, J.; Schmidt, M. K.; et al. 2019. BOADICEA: a comprehensive breast cancer risk prediction model incorporating genetic and nongenetic risk factors. *Genetics in Medicine*, 21(8): 1708–1718.

Lee, C.; Yoon, J.; and Van Der Schaar, M. 2019. Dynamic-deephit: A deep learning approach for dynamic survival analysis with competing risks based on longitudinal data. *IEEE Transactions on Biomedical Engineering*, 67(1): 122–133.

Little, R. J. 1988. Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics*, 6(3): 287–296.

Loomans-Kropp, H. A.; and Umar, A. 2019. Cancer prevention and screening: the next step in the era of precision medicine. *NPJ precision oncology*, 3(1): 1–8.

McKinney, S. M.; Sieniek, M.; Godbole, V.; Godwin, J.; Antropova, N.; Ashrafian, H.; Back, T.; Chesus, M.; Corrado, G. S.; Darzi, A.; et al. 2020. International evaluation of an AI system for breast cancer screening. *Nature*, 577(7788): 89–94.

Nagpal, C.; Jeanselme, V.; and Dubrawski, A. 2021. Deep Parametric Time-to-Event Regression with Time-Varying Covariates. In *Survival Prediction-Algorithms, Challenges and Applications*, 184–193. PMLR.

Pan, W. 1998. Rank invariant tests with left truncated and interval censored data. *Journal of Statistical Computation and Simulation*, 61(1-2): 163–174.

Stark, G. F.; Hart, G. R.; Nartowt, B. J.; and Deng, J. 2019. Predicting breast cancer risk using personal health data and machine learning models. *Plos one*, 14(12): e0226765.

Steingrimsson, J. A.; Diao, L.; and Strawderman, R. L. 2019. Censoring unbiased regression trees and ensembles. *Journal of the American Statistical Association*, 114(525): 370–383.

Strobl, A. N.; Vickers, A. J.; Van Calster, B.; Steyerberg, E.; Leach, R. J.; Thompson, I. M.; and Ankerst, D. P. 2015. Improving patient prostate cancer risk assessment: Moving from static, globally-applied to dynamic, practice-specific risk calculators. *Journal of biomedical informatics*, 56: 87–93.

Sud, A.; Turnbull, C.; and Houlston, R. 2021. Will polygenic risk scores for cancer ever be clinically useful? *NPJ precision oncology*, 5(1): 1–5.

Sun, Y.; Chiou, S. H.; and Wang, M.-C. 2020. ROC-guided survival trees and ensembles. *Biometrics*, 76(4): 1177–1189.

Sun, Z.; Dong, W.; Shi, J.; He, K.; and Huang, Z. 2021. Attention-Based Deep Recurrent Model for Survival Prediction. *ACM Transactions on Computing for Healthcare*, 2(4): 1–18.

Tewari, R.; Chhabra, M.; Natu, S. M.; Goel, A.; Dalela, D.; Goel, M. M.; and Rajender, S. 2014. Significant association of metabolic indices, lipid profile, and androgen levels with prostate cancer. *Asian Pacific Journal of Cancer Prevention*, 15(22): 9841–9846.

Therneau, T.; Crowson, C.; and Atkinson, E. 2017. Using time dependent covariates and time dependent coefficients in the cox model. *Survival Vignettes*, 2: 3.

Tsiatis, A. A.; and Davidian, M. 2004. Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, 809–834.

Tyrer, J.; Duffy, S. W.; and Cuzick, J. 2004. A breast cancer prediction model incorporating familial and personal risk factors. *Statistics in medicine*, 23(7): 1111–1130.

Utkin, L. V.; Konstantinov, A. V.; Chukanov, V. S.; Kots, M. V.; Ryabinin, M. A.; and Meldo, A. A. 2019. A weighted random survival forest. *Knowledge-Based Systems*, 177: 136–144.

Van Buuren, S.; and Groothuis-Oudshoorn, K. 2011. mice: Multivariate imputation by chained equations in R. *Journal of statistical software*, 45: 1–67.

Van Houwelingen, H. C. 2007. Dynamic prediction by landmarking in event history analysis. *Scandinavian Journal of Statistics*, 34(1): 70–85.

Wallace, M. 2014. Time-dependent tree-structured survival analysis with unbiased variable selection through permutation tests. *Statistics in medicine*, 33(27): 4790–4804.

Wang, X.; Zhang, Y.; Hao, S.; Zheng, L.; Liao, J.; Ye, C.; Xia, M.; Wang, O.; Liu, M.; Weng, C. H.; et al. 2019. Prediction of the 1-year risk of incident lung cancer: prospective study using electronic health records from the state of Maine. *Journal of medical Internet research*, 21(5): e13260.

Wiens, J.; Saria, S.; Sendak, M.; Ghassemi, M.; Liu, V. X.; Doshi-Velez, F.; Jung, K.; Heller, K.; Kale, D.; Saeed, M.; et al. 2019. Do no harm: a roadmap for responsible machine learning for health care. *Nature medicine*, 25(9): 1337–1340.

Wongvibulsin, S.; Wu, K. C.; and Zeger, S. L. 2020. Clinical risk prediction with random forests for survival, longitudinal, and multivariate (RF-SLAM) data analysis. *BMC medical research methodology*, 20(1): 1–14.

Wright, M. N.; and Ziegler, A. 2015. ranger: A fast implementation of random forests for high dimensional data in C++ and R. *arXiv preprint arXiv:1508.04409*.

Appendices

A Study Details

Cancer Registry: Participants who later developed cancer were identified (using their national IDs) in the National Cancer Registry (NCR), which records all cancer cases in the country. NCR contains for each case the cancer type (ICD9 code) and diagnosis date, and we used all cancer diagnoses until January 1^{st}, 2016 which we considered as the end time of our study. Figure 4 shows the number of individuals in the cohort with each cancer type. We focused on the two most common cancer types per sex: BC for females and PGC for males and used relevant ICD-10 codes (BC:C50.1-50.6, C50.8-50.9 and PGC: C61.9). Individuals who had a different type of cancer prior to diagnosis of BC or PGC were excluded.

Inclusion criteria: All individuals who had birth, sex, 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 NCR were identified ($n_p = 1,547$, $n_v = 3,999$), along with their cancer type (Figure 3).

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 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

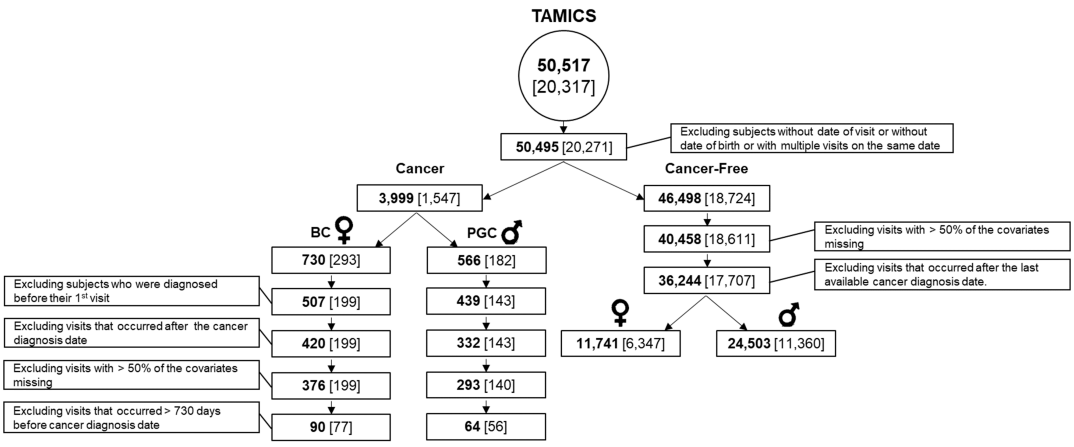

Figure 3: Study design. The bold number is the number of visits; the number of individuals appears in parentheses.

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 NCR ($n_p = 934$, $n_v = 4,214$). 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[50] with default parameters.

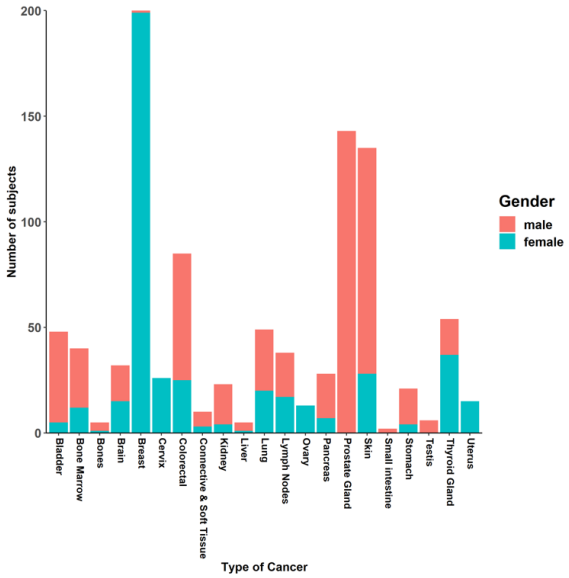

Figure 4: The number of patients per cancer type. A bar plot of the number of individuals who were surveyed in the Center and later diagnosed with cancer, categorized by sex and type of cancer.

B Breast Cancer Cohort

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 concentration (MCHC). To reduce the effect of age on our model, we created an age-matched cohort ('Matched BC-Free') of 3,635 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.

C Prostate Gland Cancer Cohort

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 3. 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. None of the covariates showed significant difference between the PGC and the Matched PGC-Free groups.

D Hyperparameter Optimization

For the hyperparameter search in the Dynamic-DeepHit model, we used a grid search (Table 2. The hyperparameter combinations were evaluated on each fold in the cross-validation and the average performance was computed. ReLU function was used as an activation function with 100 nodes and 2 layers in the RNN and tanh was used in the fully connected component with 100 nodes and 2 layers. In bold are the parameters that were chosen for the BC model and in italics are the parameters that were chosen for the PGC model. We used Adam Optimizer.

	BC			BC-Free			Matched BC-Free			BC vs. BC-Free	BC vs. Matched BC-Free
Parameter	**V**	**N**	**Mean±STD**	**V**	**N**	**Mean±STD**	**V**	**N**	**Mean±STD**	**MW**	**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
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
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
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
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
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
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
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	0.05	0.16
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
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
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
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
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	0.74
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
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
PLT (K/μL)	90	77	262.7±53.0	11,740	6,347	263.2±61.6	5,884	3,635	261.4±61.3	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
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	0.38
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
Age (Years)	90	77	53.46±7.97	11,741	6,347	47.16±10.56	5,884	3,635	53.2±7.66	< 0.001	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. 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. V- number of visits, N-number of individuals

Parameter	**Set**
RNN architecture	***LSTM***,GRU
Dropout	0.4, ***0.6***, 0.8
Learning rate	10^{-3}, $\mathbf{10^{-4}}$,10^{-5},10^{-6}
Mini-batch size	***32***,64,128
α	0.1, ***1***, 3 ,5
β	***0.1***, 1 ,3 ,5
α	0.1, ***1***, 3 ,5

Table 2: Hyperparameter values tested in Dynamic-DeepHit model.

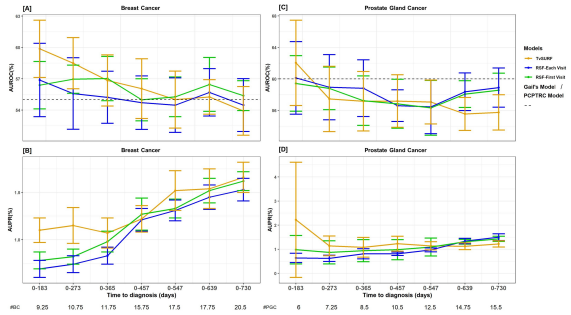

Figure 5: A comparison of TVsuRF to random survival forest. X-axis labels are the average number of patients with cases that were available across the cross-validation folds for each time interval. [A,B] BC risk prediction. [A] AUROC. The dashed line represents the (time-independent) AUROC reported for the Gail's model (Clendenen et al. 2019). [B] AUPR. [C,D] PGC risk prediction. The dashed line represents the PCPTRC model(Ankerst et al. 2014)

E Results of RSF Variants

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. Figure 5 summarizes the results.

F AUPR Trends

The AUPR scores of all algorithms tend to grow with the prediction horizon (Figure 2B,D). This can be explained by the rise in the baseline AUPR. Recall that the expected AUPR when using a baseline random classifier is the proportion of the positive class in the data. As we evaluate the models over growing time horizons, the proportion of the

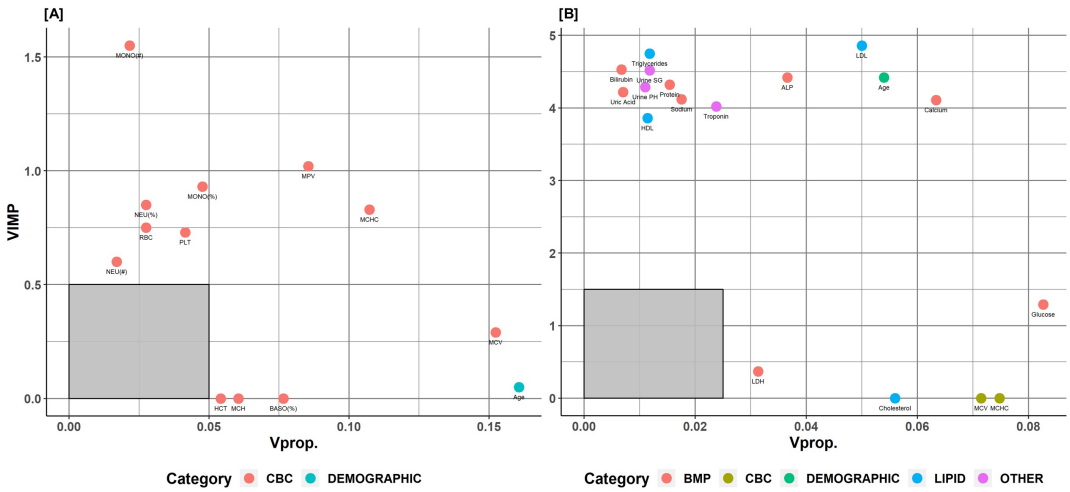

Figure 6: Variable importance for model prediction in the 183-day window. Points indicate the different variables. The y-axis presents VIMP, the decrease in AUROC following random assignment of values to the variable. The x-axis plots Vprop, the variable's inclusion frequency in the trees of the model. For both measures higher values indicate more importance. The color of a point represents the category of the parameter. [A] Variable importance of BC prediction model. features of low importance (Vprop <0.05 and VIMP<0.5) are not shown [B] Variable importance of PGC prediction model. features of low importance (Vprop<0.025 and VIMP<1.5) are not shown.

positive class increases with increasing time-horizons, since additional cancer cases are detected, and therefore the baseline AUPR grows.

G Variable Importance

Variable importance for model prediction in the 183-day window is presented in Figure 6.

H Comparison to Subjects With CBE and Mammography Tests

Some of the participants underwent BC screening tests during the visits. 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. 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 (see Figure 3 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. We 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. 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. If no values were reported, a BC surgeon reviewed the physician's text and determined if there was a positive finding. We compared the recommendations that were done by these screening tests to the results of the 730-day predictor, computed using data only from the latest visit, 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. CBE had 29.1% sensitivity and 93.7% specificity and 5.1% positive predictive value (PPV), while TVsuRF had 12.5% sensitivity and 2.3% PPV for the same specificity. Mammography sensitivity, specificity and PPV were 58.3%, 66.1% and 2.7%, respectively, and TVsuRF had 41.7% sensitivity and 1.3% PPV 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 Figure 8 show the three predictions for women who 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.

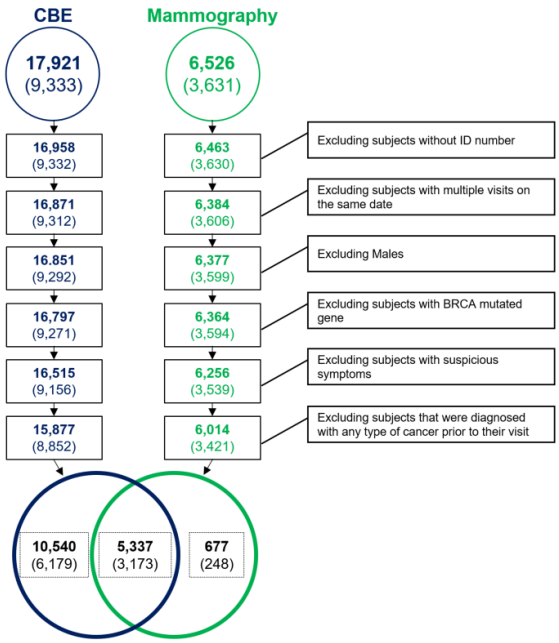

Figure 7: The CBE and mammography cohorts. Effect of exclusion criteria on the members of the study cohort who conducted a mammography screening test for BC and CBE.

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*

Figure 8: 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, 2^{nd} line: mammography result; 3^{rd} line: the risk score calculated by the TVsuRF model. Patients were ordered from high (dark blue) to low (light blue) risk score. 4^{th} line: time from visit to cancer diagnosis.

	PGC			PGC-Free			Matched PGC-Free			PGC vs. PGC-Free	PGC vs. Matched PGC-Free
Parameter	**V**	**N**	**Mean±STD**	**V**	**N**	**Mean±STD**	**V**	**N**	**Mean±STD**	**MW**	**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
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
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
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
Lym (%)	64	56	27.52±6.89	24,382	11,344	29.79±6.74	6,080	3,320	28.58±6.78	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	0.748	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
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
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
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
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
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
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
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
Neu (%)	64	56	61.34±8.05	24,382	11,344	58.82±7.52	6,080	3,320	59.75±7.52	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
RDW (%)	64	56	13.34±0.86	24,389	11,344	13.01±0.79	6,083	3,320	13.2±0.84	0.138	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
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
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
SBP (mmHg)	59	53	131.44±15.59	23,326	10,896	125.1±14.32	5,671	3,155	131.08±15.48	0.099	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
Temp (C^{o})	59	53	36.34±0.33	22,104	10,947	36.35±0.34	5,397	3,184	36.33±0.33	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
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
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
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
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	0.002	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	0.511	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	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
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
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
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
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
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	0.049	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
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
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
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
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	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
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
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
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
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
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
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	1

Table 3: Characteristics of the PGC, PGC-free and Matched PGC-free groups.Values are mean±SD. MW: p-value of the Mann–Whitney test. All p-values were 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.V-number of visits, N-number of individuals