

RESEARCH ARTICLE

Predictors of colorectal cancer survival using cox regression and random survival forests models based on gene expression data

Mohanad Mohammed^{1,2*}, Innocent B. Mboya^{1,3}, Henry Mwambi¹, Murtada K. Elbashir⁴, Bernard Omolo^{1,5,6}

1 School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Pietermaritzburg, Scottsville, South Africa, **2** Faculty of Mathematical and Computer Sciences, University of Gezira, Wad Madani, Sudan, **3** Department of Epidemiology and Biostatistics, Kilimanjaro Christian Medical University College (KCMUCo), Moshi, Tanzania, **4** College of Computer and Information Sciences, Jouf University, Sakaka, Saudi Arabia, **5** Division of Mathematics & Computer Science, University of South Carolina-Upstate, Spartanburg, United States of America, **6** School of Public Health, Faculty of Health Sciences, University of Witwatersrand, Johannesburg, South Africa

* mohanadadam32@gmail.com



OPEN ACCESS

Citation: Mohammed M, Mboya IB, Mwambi H, Elbashir MK, Omolo B (2021) Predictors of colorectal cancer survival using cox regression and random survival forests models based on gene expression data. PLoS ONE 16(12): e0261625. <https://doi.org/10.1371/journal.pone.0261625>

Editor: Afrizanfaizal Abdullah, University of Technology Malaysia: Universiti Teknologi Malaysia, MALAYSIA

Received: July 15, 2021

Accepted: December 6, 2021

Published: December 29, 2021

Copyright: This is an open access article, free of all copyright, and may be freely reproduced, distributed, transmitted, modified, built upon, or otherwise used by anyone for any lawful purpose. The work is made available under the [Creative Commons CC0](https://creativecommons.org/licenses/by/4.0/) public domain dedication.

Data Availability Statement: The dataset is publicly available on the Gene Expression Omnibus (GEO) public database (<https://www.ncbi.nlm.nih.gov/geo/>) under the accession number GSE39582.

Funding: This work was funded by GSK Africa Non-Communicable Disease Open Lab through the DELTAS Africa Sub-Saharan African Consortium for Advanced Biostatistics (SSACAB) Grant No. 107754/Z/15/Z- training programme. The views expressed in this publication are those of the

Abstract

Understanding and identifying the markers and clinical information that are associated with colorectal cancer (CRC) patient survival is needed for early detection and diagnosis. In this work, we aimed to build a simple model using Cox proportional hazards (PH) and random survival forest (RSF) and find a robust signature for predicting CRC overall survival. We used stepwise regression to develop Cox PH model to analyse 54 common differentially expressed genes from three mutations. RSF is applied using log-rank and log-rank-score based on 5000 survival trees, and therefore, variables important obtained to find the genes that are most influential for CRC survival. We compared the predictive performance of the Cox PH model and RSF for early CRC detection and diagnosis. The results indicate that *SLC9A8*, *IER5*, *ARSJ*, *ANKRD27*, and *PIPOX* genes were significantly associated with the CRC overall survival. In addition, age, sex, and stages are also affecting the CRC overall survival. The RSF model using log-rank is better than log-rank-score, while log-rank-score needed more trees to stabilize. Overall, the imputation of missing values enhanced the model's predictive performance. In addition, Cox PH predictive performance was better than RSF.

Introduction

Colorectal cancer (CRC) is the second leading cause of mortality in women and third in men [1]. The American cancer society estimate, about 1 in 23 men and 1 in 25 women develop colorectal cancer in their lifetime [2]. Globally, there were about 19.3 million new cancer cases in 2020 alone, while close to 10 million deaths were recorded due to cancer [3]. CRC represents 9.4% of cancer deaths and 10% of newly diagnosed cancer cases [3]. The incidence and mortality in males are 10.6% and 9.3%, respectively, while the incidence and mortality in females are 9.4% and 9.5%, respectively [3]. Early detection of CRC can reduce mortality due improved

author(s) and not necessarily those of GSK. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

chemotherapy regimens and surgical techniques [4–6]. The prognosis and survival of early intervention with CRC patients are linked with tumor staging, where early diagnosis of the tumor is more likely to be curable [7]. The 5-year relative survival rates for patients with localized CRC was 91% in the USA between 2010 and 2016 [8]. However, the 5-year relative survival rates of CRC cases at regional and distant stages are 72% and 14%, respectively [8]. The main characteristics of the CRC are that it has high inter-patient and intra-tumor heterogeneity. Other factors such as environment, lifestyle, and diet can lead to further heterogeneity in the CRC occurrence and progression [9–11]. This heterogeneity leads to variations in response to treatment between individuals. Determining the molecular markers is clinically essential to help detect and precisely predict the prognosis of patients with CRC.

Researchers have developed many methods to determine the prognostic molecular markers to early detect and predict the prognosis of patients with CRC. These methods include univariate and multivariate Cox proportional hazard models, elastic net estimation, and random forests for survival prediction [4, 7, 12–15]. Previous studies such as, Abdul Aziz *et al.* [12] analyzed the CRC death using the Cox proportional hazard model, and they reported a 19 gene signature that could predict the survival of CRC patients with Dukes' B and C stages. In their work, Abdul Aziz *et al.* used SAM, *limma*, and t-test to identify the most significant genes based on microarray gene expression data. Dai *et al.* [4] conducted a survival analysis using univariate and multivariate Cox models based on three microarray datasets from GEO and one dataset from the TCGA database. They used the DEGs from each of the three microarray datasets, and they identified 105 mutual DEGs based on the intersection of the three DEGs lists. They conducted a protein-protein interaction network (PPI) of the DEGs, and they identified hub genes. To investigate the 44 hub genes' prognostic values in CRC, they conducted a survival analysis using the sample splitting and Cox regression models based on the TCGA dataset. Their results showed that two down-regulated and two up-regulated hub genes were significantly associated with the CRC patients' overall survival.

Bian *et al.* [7] analyzed data from four microarray datasets and identified DEGs from each of them. They identified the common genes across the four datasets, and this way, they obtained 53 genes. Then they utilized PPI, which identified ten essential genes according to their degree value, betweenness centrality, and closeness centrality. They used gene expression profiling interactive analysis (GEPIA) to apply survival analysis using the log-rank test based on the expression levels. Their results showed that four low expressed genes out of the ten genes were significantly related to unfavorable prognosis in the patients with CRC. Martinez-Romero *et al.* [14] identified a new set of gene markers associated with CRC to predict tumor progression and evolution towards inferior survival stages based on an integrated gene expression dataset of 1273 CRC samples. They compared the early and late stages of CRC using *limma* to identify the genes (2707 DEGs) that had a significant effect on CRC tumor progression. Then, they applied Kaplan-Meier to rank the genes based on the non-parametric log-rank test. Their results identified 429 essential genes in which overexpression is related to low survival rate and 336 crucial genes in which repression is associated with inferior survival. They validated the top 5 genes using an external cohort study and presented a good separation of the CRC samples into two low and high-risk groups.

A study by Pan *et al.* [13] proposed a predictive model based on RNASeq gene expression data. Their model uses the differentially expressed genes (DEGs) profiles. These profiles were obtained using the univariate and multivariate Cox regression, which was used to compare TNM stages to assess their predictive survival accuracy. Their results showed that 10 DEGs had a significant effect on CRC survival. Yan *et al.* [15] implemented random forests to identify biomarkers associated with survival in CRC based on a set of oligonucleotide microarray data. Their results showed that four genes had the potential to predict CRC survival.

To the best of our knowledge, RSF has not been used with gene expression data in the previous studies to predict CRC survival. The gene expression data is characterized by the problem of the curse of dimensionality and collinearity. To overcome this problem, the CRC survival is predicted based on selecting the differentially expressed genes (DEGs) in colorectal cancer that was based on the three-mutation status (KRAS, BRAF, and TP53) where they serve as a predictive biomarker of response to treatment in CRC. We assume that complex interaction between multiple DEGs contributes to prognostic survival differences between wild-type and mutant patients with CRC.

We developed and compared Cox proportional hazard (Cox PH) model and random survival forests (RSF) in predicting CRC survival and associated biomarkers using a public genome database from Gene Expression Omnibus (GEO). The aim was to assess the CRC survival predictors accounting for missing data based on the gene expression data. We selected 54 common differentially expressed genes from three mutations (KRAS, BRAF, and TP53), using the complete case samples, and performed analysis using Cox PH and RSF models before and after imputation.

Materials and methods

Dataset

The dataset with accession number GSE39582 [16], was downloaded from Gene Expression Omnibus (GEO) public database (<https://www.ncbi.nlm.nih.gov/geo/>) using the BRB-ArrayTools software (<https://brb.nci.nih.gov/BRB-ArrayTools/>). This dataset has 54675 probes taken from 566 samples with colon cancer and 19 non-tumor samples. Usually, the gene expression data includes noisy and or irrelevant genes. Therefore, performing data cleaning and feature (genes) selection are essential steps that should be applied before modeling the data. A pre-processing step was applied to prepare the dataset for modeling. These pre-processing steps are log₂ transformation, quantile normalization, gene filtration, and differentially expressed genes analysis using a two samples t-test. Filtration is a process in data cleaning used to eliminate insufficiently expressed probes and those with excessive missing expression levels across the samples [17–20]. On the other hand, quantile normalization and log₂-transformed steps to eliminate the variation between samples. BRB-ArrayTools is used to implement the filtration and normalization of the dataset. The two-sample t-test, with the 0.001 significance level threshold, was used for gene selection to provide informative genes for building survival models. The overall procedures that we followed in our analysis are summarized in Fig 1.

Statistical analysis

We analyzed the gene expression data using the *R* version (R-4.0.4). Summary statistics of the gene expressions are depicted in the supplementary file (see S1 Appendix). These statistics include the minimum, maximum, means, and standard deviations of the expression levels. We used frequency and percentages for the categorical data representing the clinical information, as shown in Table 1. The statistical analysis was conducted in three phases; the first phase is the complete case analysis, followed by imputation of missing values in the outcome based on the covariates and an appropriate imputation model. Then we applied survival analysis on the complete case and imputed datasets. The survival analysis results on these two datasets were compared to evaluate the precision of estimates. Two separate models were fitted before and after imputations; the first is the Cox regression model, while the second is the random survival forests with log-rank and log-rank-score split rules. The missing values were assumed to be missing at random (MAR), where the probability of data being missing does not depend on

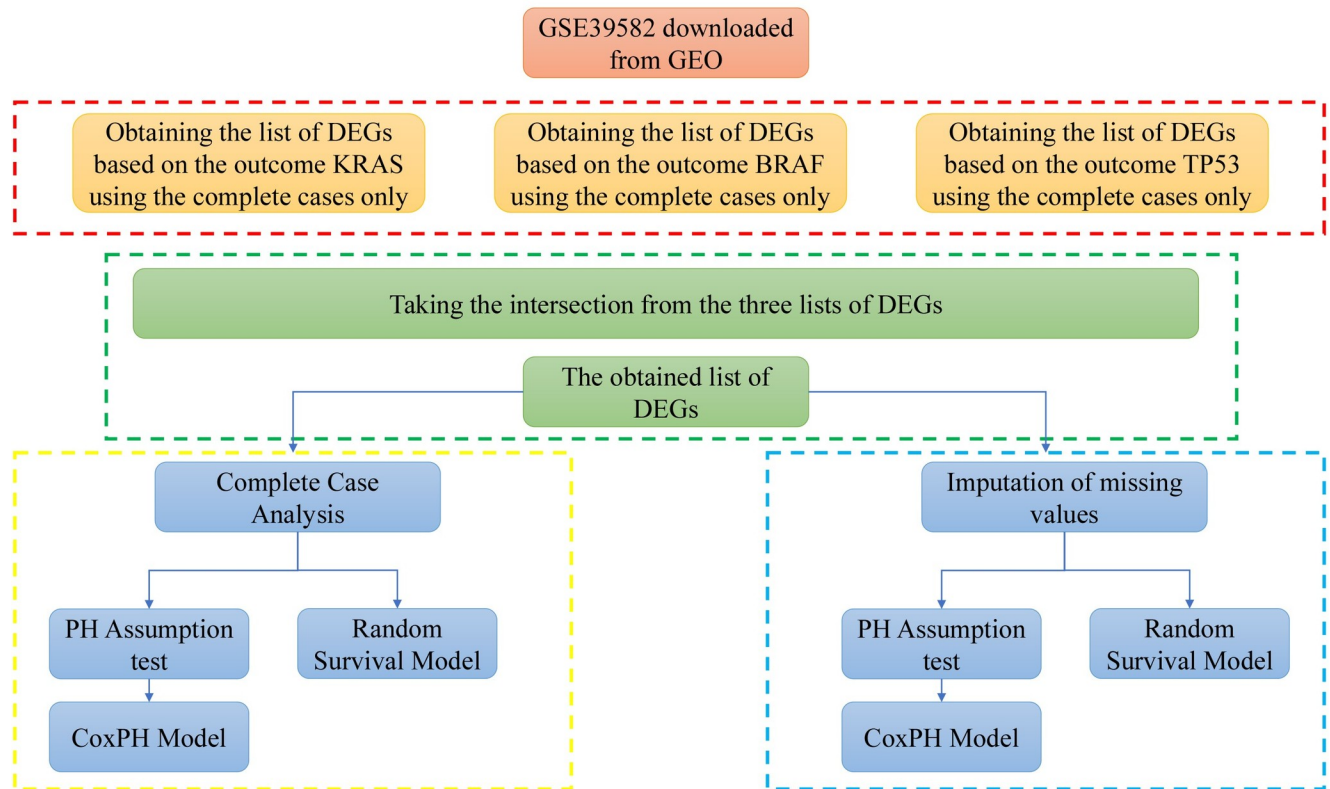


Fig 1. Flow-chart of the procedure followed in the pre-processing and analysis of the dataset.

<https://doi.org/10.1371/journal.pone.0261625.g001>

the unobserved data, conditional on the observed data [21–24]; consequently, the genes and other covariates in the dataset were used to predict missingness.

Complete case analysis. The filtration step resulted in 18865 out of 54675 probes. These 18865 probes were used for further reduction analysis using a t-test. To find the differentially expressed genes (DEGs) that discriminate between the mutant and wild-type mutation, we used the three mutation types, KRAS, BRAF, and TP53. We created three different datasets using the 18865 probes with each of the three mutation types based on these three mutation types. First, we removed the samples with missing values for each of the three datasets according to their clinical outcome. Then, we calculated the correlation matrix for the gene expression data and filtered out one gene from every two genes that show a correlation coefficient greater than 0.6. Subsequently, we extracted three DEGs lists from all three datasets using a two-sample t-test based on 0.001 thresholds. Ultimately, from the three lists of DEGs, there were 54 common genes (see [S1 Appendix](#)). Also, we used the common samples across the three datasets to produce the complete cases in one dataset. The samples with missing or zero values in the event status and time variables were removed. We then converted the five TNM stages into a new categorical variable with two stages (Early and Late), where stages four and five were combined to give the late category. Finally, we used the obtained data for finding the most significant gene markers that may predict survival for CRC patients. [Table 2](#) provides a concise summary of the pre-processed data.

Multiple imputations of the missing values. To compensate for the missing data, we used the R package “mice (Multivariate Imputation by Chained Equations)”, which impute the missing values in the covariates. The mice package takes care of uncertainty related to missing

Table 1. Clinical characteristics of colorectal cancer patients (N = 307).

Variable	Frequency (n)	Percentage (%)
Age at diagnosis in years: Mean (SD)*	66.8 (13.2)	
KRAS Mutation		
Mutant	123	40
WildType	184	60
BRAF Mutation		
Mutant	25	8
WildType	282	92
TP53 Mutation		
Mutant	166	54
WildType	141	46
Tumor Location		
Proximal	124	40
Distal	183	60
Cancer stage		
Early	156	51
Late	151	49
Sex		
Female	137	45
Male	170	55
Molecular subtype		
C1	65	21
C2	49	16
C3	43	14
C4	29	9
C5	29	9
C6	36	12

*SD: Standard deviation

<https://doi.org/10.1371/journal.pone.0261625.t001>

values [23–25]. It assumes that the missing values are missing at random (MAR) see (Fig 2), where the probability of missing data does not depend on the unobserved data, conditional on the observed data [21–24]. The mice package uses the genes and other covariates in the dataset to predict missingness. The missingness pattern in the data is assumed to be non-monotone. In this pattern, some subject values can be observed again after missing values happen [23–25]. For this missing data pattern, it is recommended to use the chained equations (fully conditional specification (FCS)) [26], or the Markov Chain Monte Carlo (MCMC) method to impute missing values [25].

Table 2. Summary of the filtered datasets and the pre-processing steps.

Dataset (GSE39582) *	Number of samples	Complete cases	Common samples	Total number of genes	After filtration	Uncorrelated genes	DEGs (t-test)	Common genes
Clinical outcomes	KRAS	585	545	307	54675	18865	13827	711
	BRAF		512					2388
	TP53		351					629

* Three datasets with the same covariates and different clinical outcome

<https://doi.org/10.1371/journal.pone.0261625.t002>

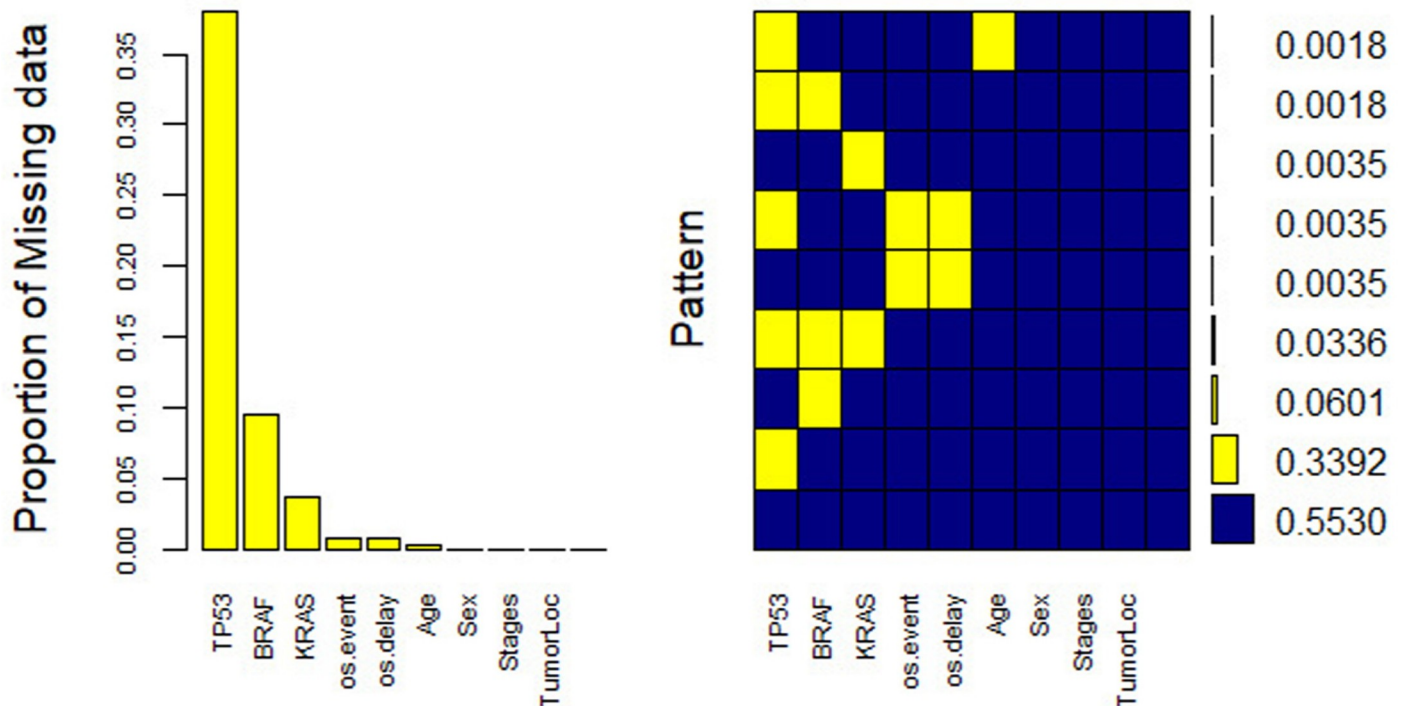


Fig 2. Proportion and patterns of missing values in the clinical characteristics available in the GSE39582 dataset.

<https://doi.org/10.1371/journal.pone.0261625.g002>

We used FCS to handle the missing values in our dataset implemented in the mice package in R using a random forest model. The FCS is considered a powerful and statistically valid method for creating imputations in both categorical and continuous variables [26]. We generated 5 imputed datasets using random forest (rf) imputations after 100 iterations (imputation cycles). We used 1051991 as a random seed to replicate imputation results each time a multiple imputation analysis was performed. In addition, we followed the procedures indicated by the work of Sterne *et al.* [27] for reporting and analysis of missing data. KRAS, BRAF, TP53, and the event status were imputed as binary, while time and age imputed as numeric variables. The rest of the variables did not contain any missing values, and were used as auxiliary variables in the imputation model. Overall, firstly we performed a complete case analysis using Cox PH and random survival forests models. Thereafter, we compared the final models from this analysis to those from the multiply imputed dataset.

Experimental setup. To evaluate the different methods, the resulting dataset was divided into training set (80%) and testing set (20%). The training set was then divided into 10 subsets to train the methods using 10-fold cross validation approach to avoid overfitting. In the 10-fold cross-validation approach the integrated brier scores (IBS) is calculated on each fold left-out while the model is trained on the other 9 folds. Finally, the trained model is tested on the testing set. The model performance was measured using prediction error curve (pec).

Statistical methods

Cox proportional hazard model (Cox PH). Cox proportional hazard model is the most widely used statistical model for modeling time to event data [28]. The Cox PH evaluates the association of the survival time of patients and one or more predictors/genes variables. The Cox PH model relates the effect of predictors which include genes in our case to the rate or

hazard of occurrence of an event such time to infection, death, recurrence of a condition at a certain point of time, this rate is generally referred as the hazard rate [29, 30]. In order to estimate the association of the gene expression levels and the survival time, consider n cancer samples say from sample $i = 1, 2, \dots, n$ and $\mathbf{g}_i = (g_{i1}, g_{i2}, g_{i3}, \dots, g_{ip})$ is a vector of p genes expression level. The i^{th} patient survival data can be represented by $(T_i, \delta_i, g_{i1}, g_{i2}, g_{i3}, \dots, g_{ip})$, where $i = 1, 2, \dots, n$; T_i and δ_i indicate the survival time and the censor status respectively. The Cox PH model is mathematically represented as follow

$$h_i(t) = h_0(t)e^{\beta' \mathbf{g}_i} \quad (1)$$

where the parameters vector β' is the regression coefficients and \mathbf{g}_i is the covariates (genes) vector. The baseline hazard function $h_0(t)$ is unspecified and non-parametric function of an individual with all expression levels equal to zero [12, 31]. The model has a parametric part specified by the linear predictor and assumed to be proportional to the non-parametric baseline hazard. This means that for two individuals, i and j , the hazard ratio is

$$\frac{h_i(t)}{h_j(t)} = \frac{e^{\beta' \mathbf{g}_i}}{e^{\beta' \mathbf{g}_j}} \quad (2)$$

The hazard ratio is assumed to be independent of time t . The maximum partial likelihood method used to estimate the Cox PH model parameters is given by

$$L(\beta) = \prod_{r \in E} \frac{e^{\beta' \mathbf{g}_r}}{\sum_{j \in R_r} e^{\beta' \mathbf{g}_j}} \quad (3)$$

where E indicates the indices of the events (e.g., deaths) and R_r represents the vector of indices of the individuals at risk at time $t_r - 0$. The results of the Cox PH model are easy to interpret, however, there are key assumptions needed such as linearity and proportional hazards. We used *survival* and *survminer* packages to implement Cox PH model in R.

Moreover, we performed the stepwise regression for developing the Cox PH model at a 5% threshold level to find a simple model that shows the essential genes (markers) and clinical covariates correlated with the CRC. At each time, we removed the genes/ covariates that are not significant at $\alpha = 0.05$ level of significance. Thereafter, we tested for the Cox PH assumption, and the integrative analysis of the CRC data showed five genes (markers) that passed the Cox PH assumption test. Thereafter, we used the five genes and the other clinical information to fit the Cox PH model.

Random survival forests (RSF). Random survival forests are an ensemble of trees and a non-parametric method constructed by bagging of classification trees for right censored data [32, 33]. The RSF are an extension of the random forests method proposed by Breiman [34]. It works on high dimensional data where the number of covariates exceeds the number of the observations. Also it can handle data that consist of complex and non-linear relationships between the dependant and the independent variables and when the covariates violate the proportional hazard assumption [35]. There are several advantageous of using the RSF method, such as, it is not based on any model assumption compared to Cox PH model. It seeks to find a model that best represent the data in the case of limited survival data. In addition, it can handle high dimensional data unlike Cox PH, and it is robust to outliers in the explanatory variables [33]. RSF employs two steps of randomizations to grow the tree. These two steps are the bootstrap sample to select cases randomly and random selection of subset of covariates for splitting the nodes of the tree. These two steps help to decorrelate the tree [20, 33]. The RSF was implemented using the *randomForestSRC* package in R [36].

Random survival forests algorithm. We used the RSF algorithm that was introduced in the work of Ishwaran *et al.* [32] as shown below:

For i in 1: $ntrees$

- Draw bootstrap samples from the original total number of samples. For each bootstrap exclude approximately 37% of the samples as out-of-bag (OOB) samples.
- Build a survival tree for every bootstrap sample by recursively repeating the following steps for each node in a tree
 - Randomly select ν genes at random from the p genes ($\nu = \sqrt{p}$)
 - To split the node, pick the best gene among the ν genes, that maximizes survival differences between daughter nodes. We used log-rank and log-rank-score splitting rules as measures of survival differences.
 - Produce the tree to full size under the constraint that a terminal node should have no less than $d_0 > 0$ unique deaths.
 - Calculate a cumulative hazard function (CHF) for every tree. Average the CHF for all the $ntrees$ trees to find the ensemble CHF.
 - Calculate the OOB prediction error for the ensemble CHF, using OOB samples.

Once the survival tree is built, the ends of the tree are called the terminal nodes. Assume, the terminal node is h and $t_{n,h}$ is the individual's death time at node h , $d_{n,h}$ is the number of deaths, and $M_{n,h}$ is the number of individuals at risk at time $t_{n,h}$. Therefore, the cumulative hazard function (CHF) can be estimated using the Nelson-Aalen estimator [37] as follows

$$\hat{H}_h(t) = \sum_{t_{n,h} \leq t} \frac{d_{n,h}}{M_{n,h}} \tag{4}$$

The CHF was calculated for all the terminal nodes. The CHF for new observation i given a vector of genes as a covariate \mathbf{g}_i , can be calculated for one tree as follows

$$\hat{H}_h(t|\mathbf{g}_i) = \hat{H}_h(t), \quad \text{for } \mathbf{g}_i \in h \tag{5}$$

To compute an ensemble CHF, the average of the $ntrees$ trees is calculated, and the bootstrap ensemble CHF for an observation i is

$$\hat{H}_e(t|\mathbf{g}_i) = \frac{1}{ntrees} \sum_{b=1}^{ntrees} \hat{H}_b(t|\mathbf{g}_i) \tag{6}$$

let,

$$I_{i,b} = \begin{cases} 1 & \text{if } i \text{ is an OOB observation for } ntrees \text{ training sample.} \\ 0 & \text{Otherwise.} \end{cases} \tag{7}$$

then the OOB ensemble CHF for an observation i is given by

$$\hat{H}_e^*(t|\mathbf{g}_i) = \frac{\sum_{b=1}^{ntrees} I_{i,b} \hat{H}_b^*(t|\mathbf{g}_i)}{\sum_{b=1}^{ntrees} I_{i,b}} \tag{8}$$

therefore, $\hat{H}_e^*(t|\mathbf{g}_i)$ is an average over the training samples where i is an OOB observation.

Log-rank split rule. The log-rank split-rule is a measure of a node separation which helps in determining the best split for that node [38]. Let h be a node of a tree and let there are n

individuals with this node. Suppose $(T_1, \sigma_1), (T_2, \sigma_2), \dots, (T_n, \sigma_n)$ are the survival outcomes corresponding to the n individuals. Thus, the best split at node h on covariate x at split point c , is the one that maximize the log-rank statistic between the two daughter nodes [32] given as follow

$$L(x, c) = \frac{\sum_{i=1}^N (d_{i1} - Y_{i1} \frac{d_i}{Y_i})}{\sqrt{\sum_{i=1}^N \frac{Y_{i1}}{Y_i} (1 - \frac{Y_{i1}}{Y_i}) (\frac{Y_i - d_i}{Y_i - 1}) d_i}} \tag{9}$$

The aim is to maximize the log-rank statistic by finding values of x and c that maximize $L(x, c)$. Specifically, we are looking to find a predictor x^* and c^* such that $|L(x^*, c^*)| \geq |L(x, c)|$ for every x and c . This process is repeated at every node until the terminal node is reach.

Log-rank-score split rule. The log-rank-score split rule is a version of the log-rank-score split rule [39]. Consider $r = (r_1, r_2, \dots, r_n)$ as a vector that ranks the survival times $(T, \delta) = ((T_1, \sigma_1), (T_2, \sigma_2), \dots, (T_n, \delta_n))$ [39, 40]. Assume $a = a(T, \delta) = (a_1(r), a_2(r), \dots, a_n(r))$ indicates the ranked score vector. Let the ranked vector r order the genes variables in such a way that $g_1 < g_2 < \dots < g_n$. Therefore, the log rank score for an observation at T_i is given by

$$a_i = a_i(T, \delta) = \delta_i - \sum_{j=1}^{\gamma_i(T)} \frac{\delta_j}{(n - \gamma_j(T) + 1)}, \tag{10}$$

where, $\gamma_j(T) = \sum_{i=1}^n \chi\{T_i \leq T_j\}$ is the number of individuals who died or were censored before or at time T_j .

Performance evaluation

We used integrated brier scores (IBS) measure [41] to assess and compare the accuracy of the predictive performance of all the models in this study. The IBS represent the average squared differences between the observed survival status and the predicted survival probability at time t . However, the value of the IBS is always between 0 and 1, the value of 0 represent the best possible IBS value. We calculated the brier scores (BS) measure using the test sample of size n_{test} as follows

$$BS(t) = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \left\{ [0 - \hat{S}(t|x)]^2 \frac{I(t_i \leq t, \delta_i = 1)}{\hat{G}(t_i|x)} + [1 - \hat{S}(t|x)]^2 \frac{I(t_i > t)}{\hat{G}(t|x)} \right\} \tag{11}$$

where $\hat{G}(t|x) \approx P(C > t|X = x)$ is the Kaplan-Meier estimate for the conditional survival function of the censoring times. Therefore, the IBS is calculated as below

$$IBS = \int_0^{\max(t)} BS(t) dt \tag{12}$$

Results

Cox proportional hazards analysis

The results of the survival problem based on gene expression data were obtained using R . We used the Cox PH model based on the selected covariates that satisfy the Cox PH assumptions. We tested the Cox PH assumptions using the Schoenfeld residual test implemented by the function *cox.zhp*. The Cox PH model assumes the regression parameters are constant over time. Therefore, the hazard ratios for any two individuals are constant over time. However, the covariates that do not satisfy the Cox PH assumptions do not meet the criteria to be entered in our final Cox PH model. As a first step, we fitted the Cox PH model for all the covariates

Table 3. Testing the proportional hazard assumption using scaled Schoenfeld residuals.

Probeset ID (Symbol)	χ^2 (df)	p-value
204014_at (DUSP4)	10.219 (1)	0.0014
212947_at (SLC9A8)	1.345 (1)	0.2462
218611_at (IER5)	2.045 (1)	0.1527
219973_at (ARSJ)	3.601 (1)	0.0577
221522_at (ANKRD27)	1.583 (1)	0.2083
221605_s_at (PIPOX)	1.651 (1)	0.1988
227134_at (SYTL1)	4.699 (1)	0.0302
Age at diagnosis (years)	2.589 (1)	0.1076
Molecular subtype	15.824 (5)	0.0074
Disease stages	1.173 (1)	0.2787
Sex	0.378 (1)	0.5388
Tumor location	0.951 (1)	0.3294

*Chi-square statistic

<https://doi.org/10.1371/journal.pone.0261625.t003>

(genes and clinical variables) in our dataset and then obtained the Cox PH assumption using the Schoenfeld residuals [Table 3](#). The genes and variables in violation of the Cox PH assumption ($p < 0.05$) were DUSP4, SYTL1, and molecular subtype.

From the Cox PH model in [Table 3](#), three variables violated the Cox PH assumption, and therefore, these genes and molecular subtype were not included in the final Cox PH model. We fitted the Cox PH model on the genes and variables that did not violate the Cox PH assumptions before and after imputation. The results from this analysis are shown in [Table 4](#). Results before imputation of missing values indicated that *218611_at (IER5)* (HR = 9.51, 95% CI 1.30, 69.58), *221522_at (ANKRD27)* (HR = 34.89, 95%CI 1.91, 635.90), and late disease

Table 4. Multivariable Cox PH results for predictors of colorectal cancer survival among adults aged 24 years and above.

Probeset ID (Symbol) / Variables	Before imputation (N = 307)			After imputation (N = 566)		
	HR* (SE)	95%CI	P-value	HR* (SE)	95%CI	P-value
212947_at (SLC9A8)	0.09 (0.84)	(0.02, 0.49)	0.005**	0.30 (0.66)	(0.08, 1.07)	0.066
218611_at (IER5)	9.51 (1.02)	(1.30, 69.58)	0.027*	6.48 (0.79)	(1.37, 30.53)	0.019*
219973_at (ARSJ)	0.23 (0.48)	(0.09, 0.58)	0.002**	0.44 (0.36)	(0.22, 0.89)	0.024*
221522_at (ANKRD27)	34.89 (1.48)	(1.91, 635.90)	0.016*	2.49 (1.06)	(0.31, 19.95)	0.393
221605_s_at (PIPOX)	0.43 (0.34)	(0.22, 0.85)	0.014*	0.49 (0.27)	(0.28, 0.83)	0.009**
Age diagnosis (years)	1.03 (0.01)	(1.01, 1.05)	0.001***	1.03 (0.01)	(1.01, 1.04)	<0.000***
Sex						
Female	1.00			1.00		
Male	1.23 (0.20)	(0.84, 1.81)	0.281	1.40 (0.15)	(1.05, 1.88)	0.024
Stages						
Early	1.00			1.00		
Late	1.97 (0.20)	(1.33, 2.93)	0.001***	1.96 (0.15)	(1.47, 2.63)	<0.000***
Tumor location						
Proximal	1.00			1.00		
Distal	1.06 (0.21)	(0.71, 1.58)	0.783	0.86 (0.16)	(0.63, 1.18)	0.356

HR: Hazard ratio, SE: Standard error, adjusted for 212947_at, 218611_at, 219973_at, 221522_at, 221605_s_at, age at first diagnosis, sex, disease stage, and tumor location.

<https://doi.org/10.1371/journal.pone.0261625.t004>

stage (HR = 1.97, 95%CI 1.33, 2.93) were associated with higher hazards of death. However, we note that two confidence intervals for *IER5* and *ANKRD27* are quite wide; therefore, they should be interpreted with caution. For every year increase, the hazards of death increased by 1.03 (95%CI 1.01, 1.05). Significantly lower hazards were observed in *212947_at (SLC9A8)* (HR = 0.09, 95%CI 0.02, 0.49), *219973_at (ARSJ)* (HR = 0.23, 95%CI 0.09, 0.58), and *221605_s_at (PIPOX)* (HR = 0.43, 95%CI 0.22, 0.85) differentially expressed genes.

After imputation of missing values, the Cox PH model showed that sex was a significant predictor of males having higher death hazards (HR = 1.40, 95%CI 1.05, 1.88) than females. Also, the disease stage covariate was a significant predictor where those with late disease stage had higher death hazards (HR = 1.96, 95%CI 1.47, 2.63) than early cases. Moreover, the results illustrated that *219973_at (ARSJ)* (HR = 0.44, 95%CI 0.22, 0.89), *221605_s_at (PIPOX)* (HR = 0.49, 95%CI 0.28, 0.83) were related with lower hazards of death. For every year increase, the hazards of death increased by 1.03 (95%CI 1.01, 1.04). Significantly higher hazards were detected with gene *218611_at (IER5)* (HR = 6.48, 95%CI 1.37, 30.53) gene.

Random survival forests analysis

We fitted two random survival forests models, including survival trees built using log-rank and the log-rank-score split rules on the datasets before and after imputation. These two models were built using the 54 genes and the other clinical information as covariates. The characteristics of the two fitted models are summarized in Table 5 below.

Permutation importance measure used to identify the most important genes/ clinical variables associated with the survival of the colon patients [42–44]. We fitted a random survival forest model before imputation and after imputation with 5000 survival trees built using log-rank and log-rank-score and their results presented in Figs 3 and 4.

Table 5 and Fig 3 show that the log-rank split-rule is more stable than the log-rank-score split-rule. Moreover, we fitted the model with 1000, 2000, and 3000 survival trees and noticed that the log-rank-score split-rule needs more survival trees to stabilize. In addition, the error rate for the forest built with survival trees based on the log-rank and log-rank-score split-rules are 41.26 and 49.05, respectively. These error rates of the RSF before imputation are much

Table 5. Random survival forests results before and after imputation using log-rank and log-rank-score split rules.

	Before imputation (N = 246)*		After imputation (N = 453)*	
	Log-rank	Log-rank-score	Log-rank	Log-rank-score
Number of deaths	88	88	157	157
Number of trees	5000	5000	5000	5000
Forest terminal node size	15	15	15	15
Average no. of terminal nodes	13.58	11.92	25.34	22.14
No. of variables tried at each split	8	8	8	8
Total no. of variables	62	62	62	62
Resampling used to grow trees	swor	swor	swor	swor
Resample size used to grow trees	155	155	286	286
Analysis	RSF	RSF	RSF	RSF
Family	surv	surv	surv	surv
Splitting rule	log-rank	log-rank-score	log-rank	log-rank-score
Number of random split points	10	10	10	10
Error rate	41.26%	49.05%	33.22%	43.01%

* Analysis performed using the 80% training set

<https://doi.org/10.1371/journal.pone.0261625.t005>

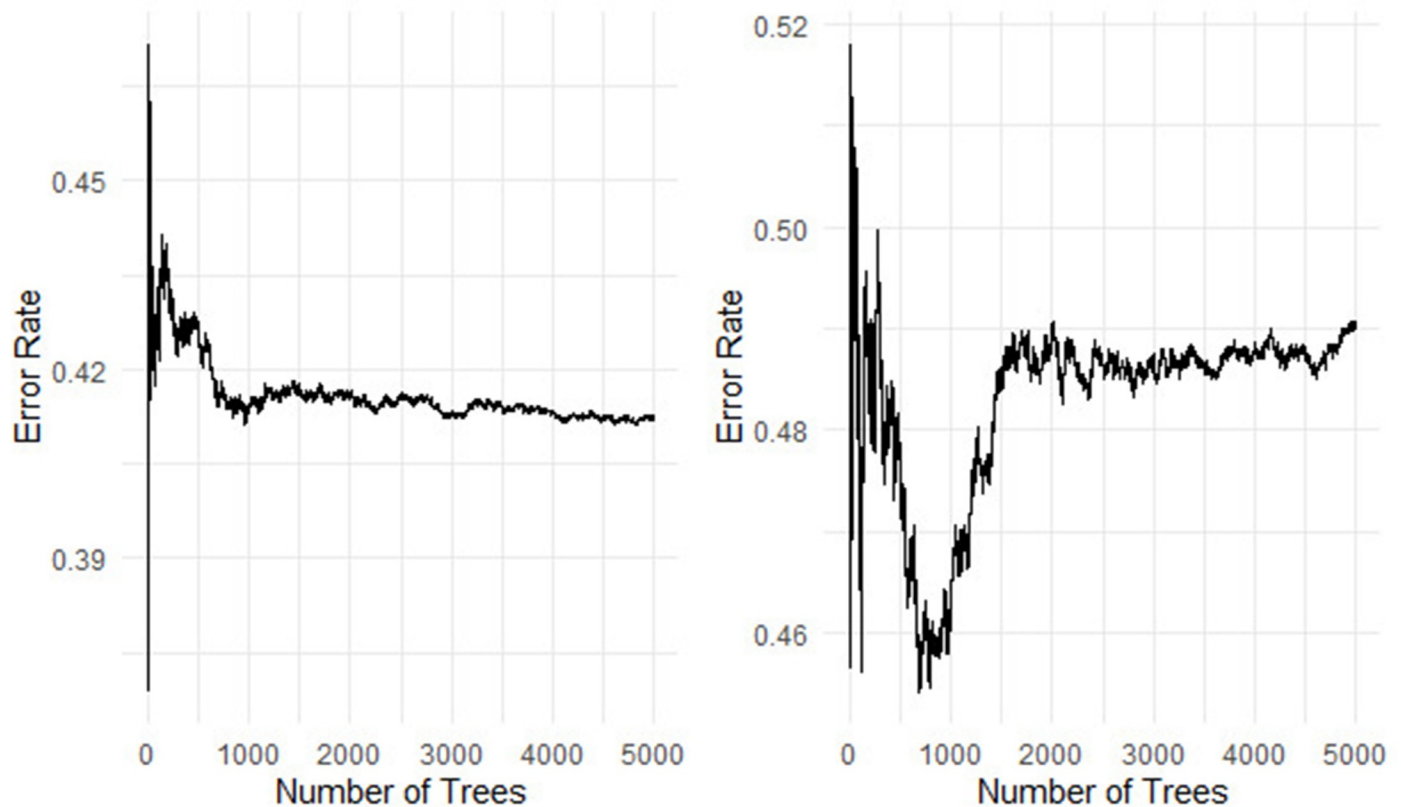


Fig 3. The prediction error rate for the random survival forests of 5000 trees before imputation and the log-rank and log-rank-score in the left and right panel used 80% training dataset.

<https://doi.org/10.1371/journal.pone.0261625.g003>

higher than the error rates for RSF built after imputation, as shown in [Table 5](#). This result indicates that the imputation can improve the performance of RSF.

The genes/ covariates associated with CRC ranked using RSF according to their importance before and after imputation based on the log-rank, and log-rank-score split-rules are presented in [Figs 5](#) and [6](#). Using RSF allows all 54 genes and other covariates regardless of their satisfying the Cox PH assumption. However, this is a very important characteristic of the RSF, as explained in the model building stage. The selection of the genes/ covariates in the model does not need to satisfy the too restrictive Cox PH assumption. RSF is purely non-parametric; hence there is no requirement of the Cox PH assumption being satisfied a prior.

We implemented RSF with 5000 survival trees built using two split-rules before and after imputation. The RSF identified the most important genes/ covariates that explain the survival of CRC patients by calculating the measure of the permutation importance as a variable's importance [[32](#), [43](#)]. For the RSF before imputation see ([Fig 5](#)), the top 20 genes/ covariates that are most important and strongly associated with the CRC obtained using the log-rank split-rule are age, *SLC28A3*, stages, *TNFSF9*, *EGLN3*, molecular subtype, *CTSV*, *ANKRD27*, *POLR3B*, *CTSA*, *SYTL1*, *MYRF*, *RPS27L*, *L3MBTL1*, *PIPOX*, *ADPRM*, *SLC6A4*, *LDLRAD3*, *MSRA*, and *SCAND1*. While the top 20 genes/ covariates that were identified by RSF using log-rank-score are *POLR3B*, *L3MBTL1*, *CTSV*, *EGLN3*, *SYTL1*, age, molecular subtype, *LDLRAD3*, *MAP7D2*, *SLC28A3*, *ANKRD27*, stages, *SLC6A4*, *CTSA*, *CABLES2*, *TNFSF9*, *GIF*, *SCAND1*, *PTP4A3*, and *MSRA*.

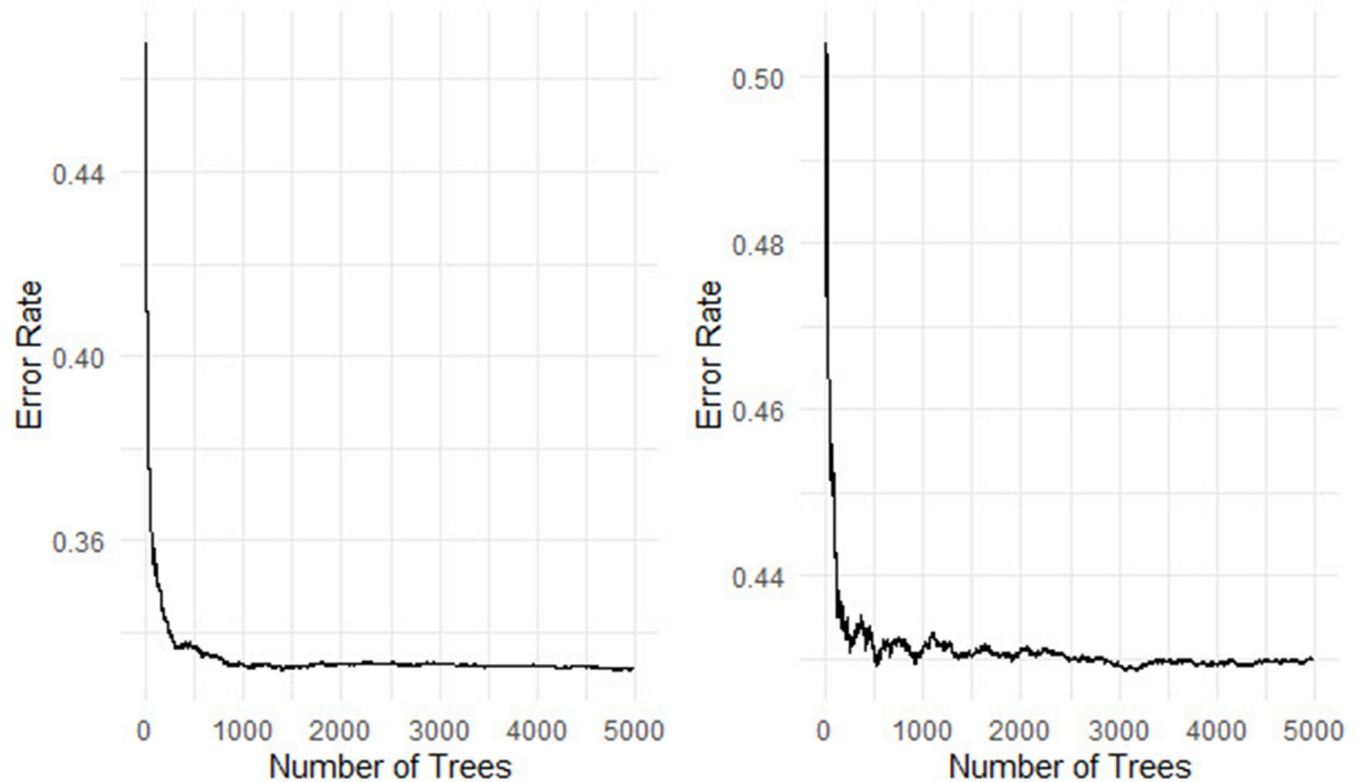


Fig 4. The prediction error rate for random survival forests of 5000 trees after imputation and the log-rank and log-rank-score in the left and right panel, respectively, using 80% training dataset.

<https://doi.org/10.1371/journal.pone.0261625.g004>

However, for the RSF after imputation (Fig 6), the top 20 genes/ covariates strongly related to CRC identified using RSF with log-rank split-rule are age, stages, molecular subtype, *PIPOX*, *ADPRM*, *CLCN2*, *RPS27L*, *IER5*, *POLR3B*, *SLC6A4*, *KRAS*, *SGMS2*, *DUSP4*, *SLC28A3*, *SLC9A8*, *ACOT8*, *SYTL1*, *CABLES2*, *SCAND1*, and *MAP7D2*. Although the RSF with log-rank-score obtains a top 20 genes/ covariates strongly relevant to CRC, these genes/ covariates are molecular subtypes, *POLR3B*, *CLCN2*, *IER5*, *SLC9A8*, *MAP7D2*, *CABLES2*, *SYTL1*, stages, *KRAS*, *SLC6A4*, *LDLRAD3*, *CTSA*, *SCAND1*, *PIPOX*, *ARSJ*, *PHACTR3*, *SLC28A3*, *SGMS2*, and *CTSV*.

The RSF with log-rank split-rule after imputation performed better in terms of the error rate. Age and disease stage were the most important covariates that affecting CRC. However, the *PIPOX*, *IER5*, and *SLC9A8* were among the most important genes strongly associated with CRC. These results agree with the results achieved from fitting the Cox PH model presented in Table 4. As far as significant effects are concerned, the most striking result to emerge was that the RSF model did pick other genes and covariates as substantial, e.g., molecular subtype and *DUSP4* which could not be included in the Cox PH model because of not satisfying the Cox PH assumption.

Predictive performance

We assessed the predictive performance of the models using the integrated brier scores measure in *R* using the *pec* package [45, 46]. The model with lower prediction error rates is therefore considered useful [43, 47]. Figs 7 and 8 show the prediction error curve of the RSF (log-

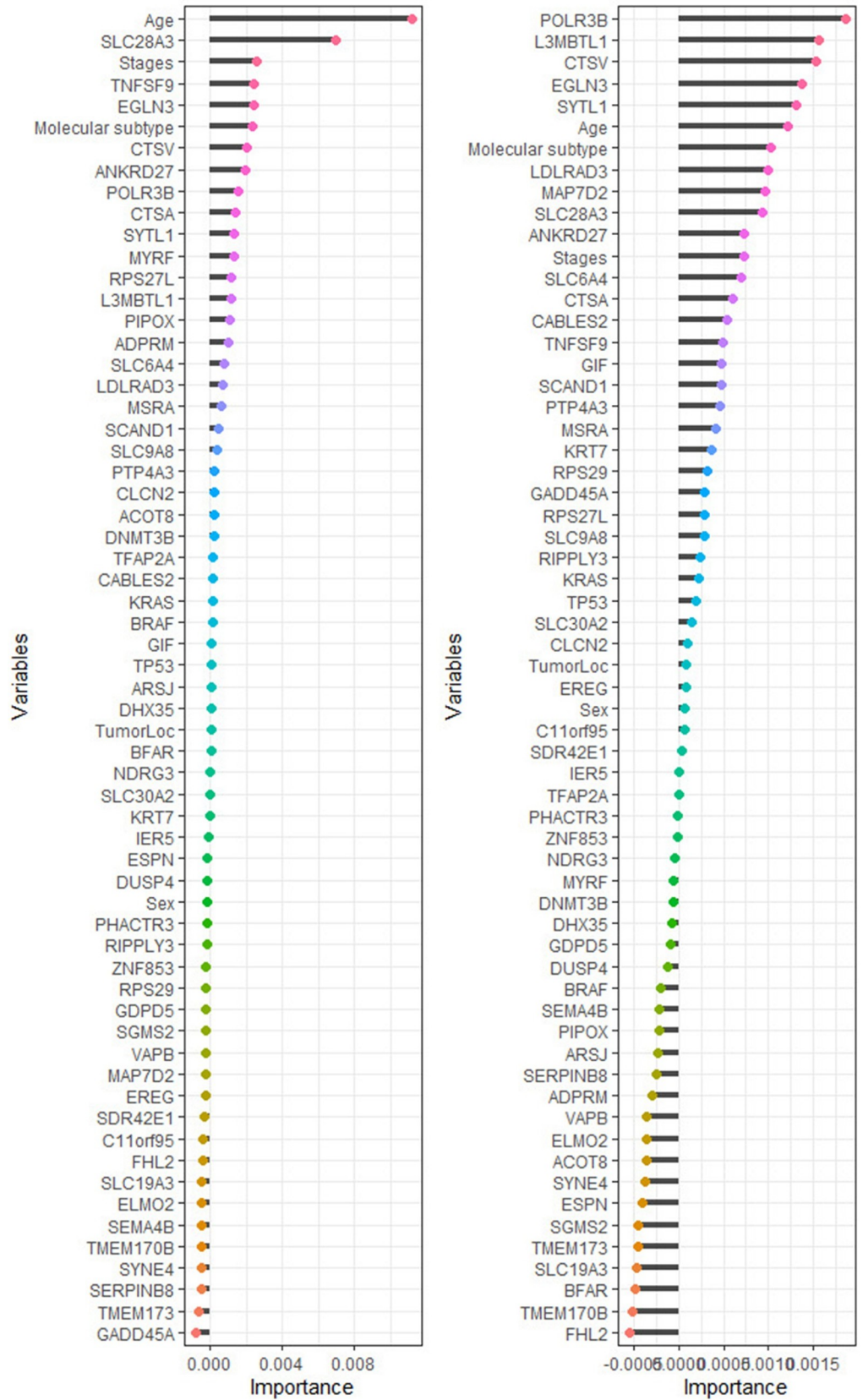


Fig 5. The rank of most predictive genes and clinical variables for colorectal cancer patients' survival before the imputation is based on how they influence the survival outcome. The variables importance is built using log-rank and log-rank-score split-rules in the left and right panel, respectively.

<https://doi.org/10.1371/journal.pone.0261625.g005>

rank and log-rank score) and Cox PH models before and after imputation. These prediction curves show that Cox PH outperformed RSF with log-rank and log-rank score split rules. The Cox PH model before and after imputation had similar prediction errors, while RSF models under the two split-rules (log-rank and log-rank-score, respectively) after imputation had lower prediction error rates compared to before imputation as can be seen (Fig 8). Their predictive performance exhibited that the log-rank split-rule is better than the log-rank-score split-rule. Moreover, we noticed that the Cox PH model showed good predictive performance compared to the two RSF under the two split-rules before and after imputation models. Thus it is safer to say that if all covariates satisfy the Cox PH assumption, the Cox PH model can be used [44].

Although the Cox PH model before and after imputation had better performance in terms of the prediction error rate, we can still not use it in the event of a violation of the proportionality of hazards assumption. Thus, in the presence of the non-proportional hazards genes/covariates, using RSF is an appealing option in the analysis of survival data, especially for high dimensional genomics data. Genomics data are usually presented in a matrix, with the columns indicating the samples and the rows showing a genomic feature such as genes [48].

Table 6 shows a comparison of the model performance using the integrated brier scores. We can notice that the prediction error estimates are lower for RSF, especially in the case of using the log-rank as a split rule. In addition, RSF models perform substantially better than Kaplan-Meier and Cox PH models.

Discussion

Cancer incidence and mortality are rapidly growing worldwide, exerting big physical, emotional, and financial problems on individual, families, communities, and health systems levels. Cancer is the first or second leading cause of death in 112 countries and is considered the third or fourth in 23 countries [3]. According to estimates from the World Health Organization (WHO), cancer is the leading cause of death around the world and accounting for nearly 10 million deaths in 2020. Moreover, WHO reported that CRC is the third common new cases, and it is also the second leading cause of death worldwide since 2020 [49]. The study aimed to determine the association between the genes and clinical covariates with CRC survival in the presence of missing values data. We also compared the predictive performance of the Cox PH and RSF models. The study provides essential information for CRC early detection and diagnosis.

The traditional regression-based methods to analyse survival data usually suffer from many problems such as restrictive assumptions including the proportionality, multicollinearity, curse of dimensionality, and lack of ability to rank the predictive performance. However, RSF models are frequently becoming a successful alternative for the analysis of the time to event data. In particular, the RSF is viewed as an appropriate analysing model for survival data, especially when the proportional hazards assumption is violated [39, 50]. When it comes to CRC survival analysis the gene expression and clinical information are utilized as covariates. The gene expression data contains many genes and most of these genes do not discriminate between normal cells and tumors. Therefore, we select the genes in which the change or difference in read counts between two conditions of experiment is statistically significant and such genes are known as the differentially expressed genes. In this study, the differentially expressed genes were obtained using three mutations based on the complete cases. The preliminary

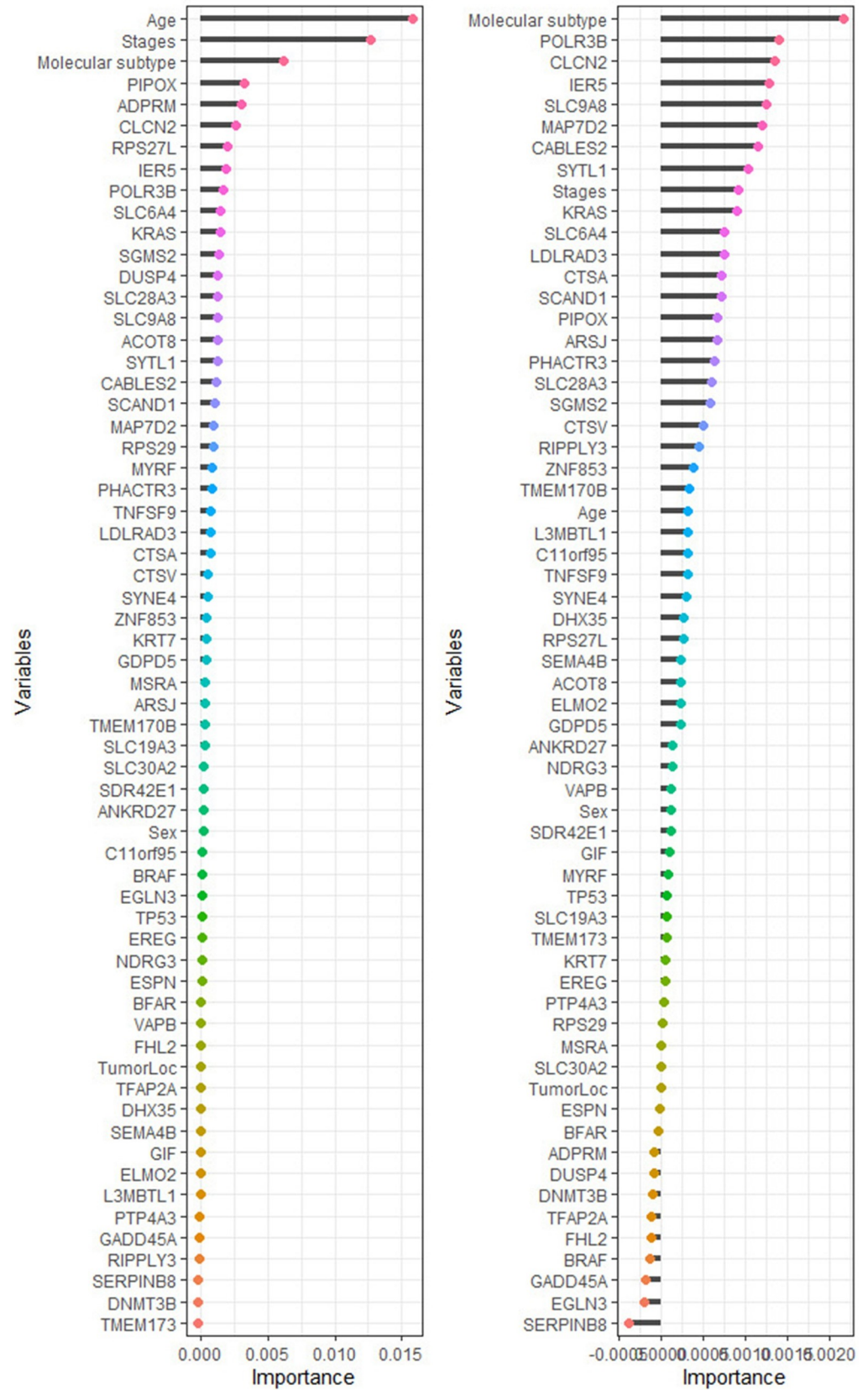


Fig 6. The rank of most predictive genes and clinical variables for colorectal cancer patients' survival after the imputation is based on how they influence the survival outcome. The variables importance is built using log-rank and log-rank-score split-rules in the left and right panel, respectively.

<https://doi.org/10.1371/journal.pone.0261625.g006>

analysis showed that 54 potentially differentially expressed genes could be correlated with CRC survival and important for understanding the initiation and progression of CRC. The differentially expressed genes together with the clinical data were used to compare the predictive performance of the Cox PH model and RSF model before and after imputation on the CRC gene expression data.

We used stepwise regression for developing the Cox PH model at a 5% threshold level to get a simple model capturing the association between the top genes and CRC patient survival. Only five genes did not violate the Cox PH assumption in the final Cox PH model. The results show that the error rates of the RSF before imputation are much higher than the error rates for RSF built after imputation. Thus, the imputation can improve the performance of RSF. Although the Cox PH model had a better performance than RSF, the results from the current study demonstrate that the random survival forests models are more flexible than the models based on the Cox PH assumption as a prerequisite for variable inclusion in the model.

After imputation, the Cox PH model indicated *SLC9A8* and *ANKRD27* genes were no longer significant predictors of CRC survival. This because it is expected that the number of observations to increase, hence, statistical power to detect an effect. The variables that were not statistically significant before imputation may now be seen as statistically significant and vice versa. Therefore, this might affect the statistical power of some variables after imputation. Overall, the most prominent finding to emerge from the analysis based on Cox PH is that for one year increase in age, the hazards of death increase by 1.03, also the males are the most exposed to the hazards of death compared to females. Thus, this study supports evidence from previous observations [51–55].

The results of the RSF using both split-rules before and after imputation identified other genes/ covariates such as molecular subtype, *SLC6A4*, *KRAS*, *SGMS2*, *DUSP4*, and *SLC28A3*. These genes/ covariates show up as important in explaining CRC survival rates. However, these genes/ covariates did not appear very strongly associated with CRC survival in the Cox PH model. Thus, one interesting finding to note is that RSF models give additional information about variable importance.

Furthermore, the results from the two RSF models before and after imputation show that age, stages, molecular subtype, *SLC9A8*, *IER5*, *ARSJ*, *ANKRD27*, and *PIPOX* greatly affected

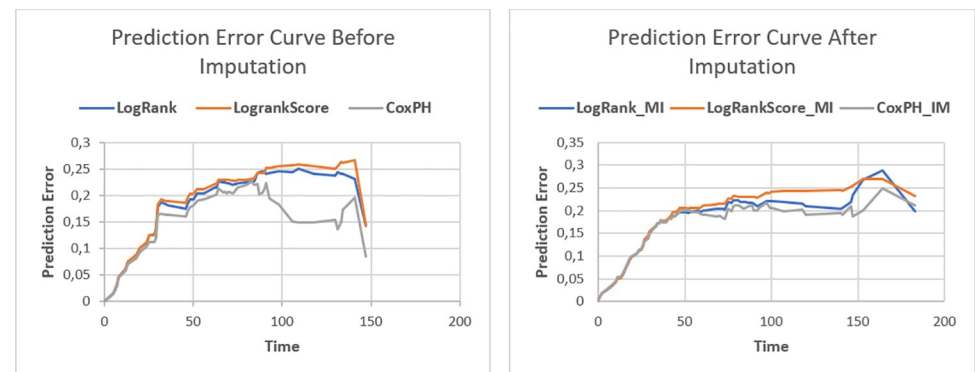


Fig 7. RSF with (log-rank and log-rank score) and Cox PH prediction error curve using 20% test set. The complete case and imputed dataset plots are in the left and right panel, respectively.

<https://doi.org/10.1371/journal.pone.0261625.g007>

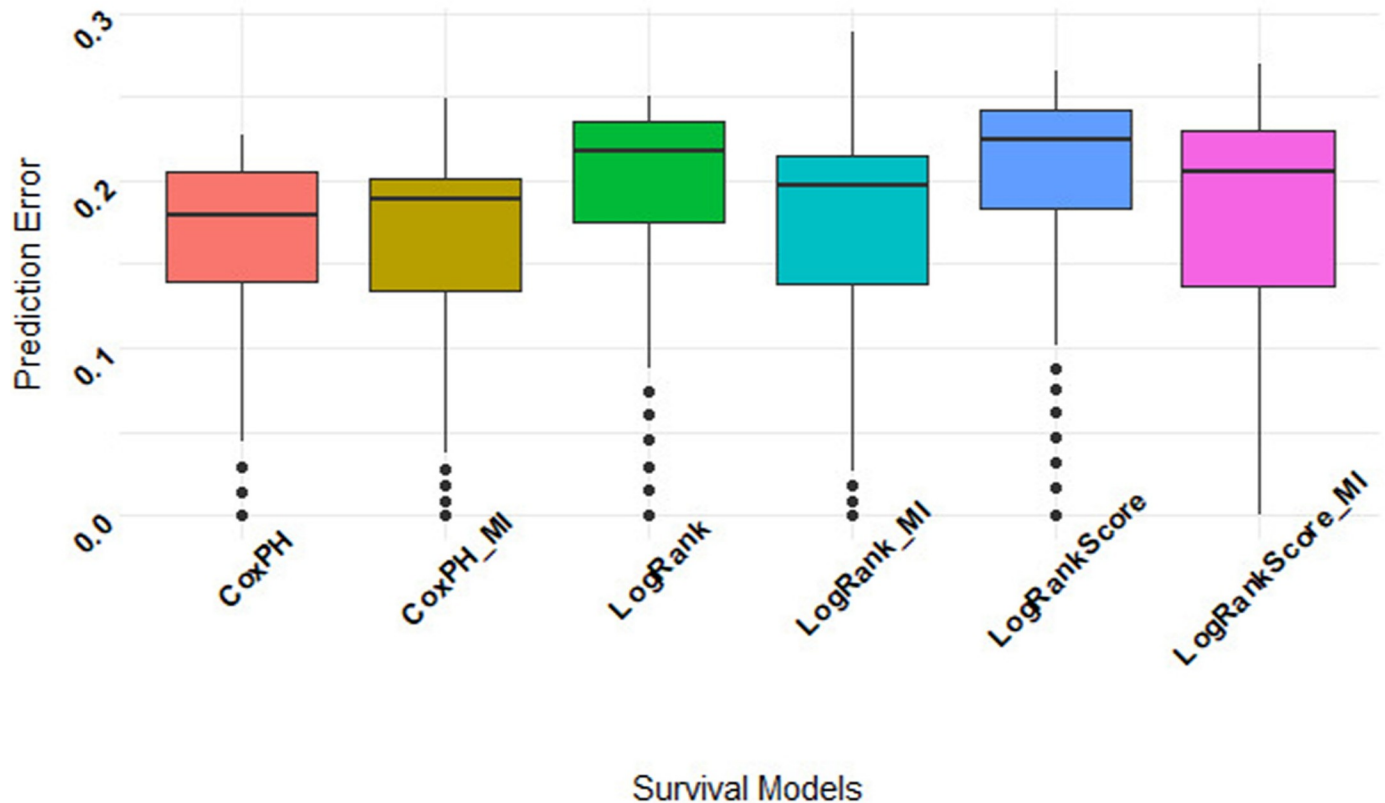


Fig 8. RSF with (log-rank and log-rank score) and Cox PH boxplot prediction error using 20% testing set together with the complete case dataset and the imputed data.

<https://doi.org/10.1371/journal.pone.0261625.g008>

the CRC mortality rates. These are ranked in the top 20 variables important in the two RSF models and agree with the Cox PH model results. Contrary to expectations, the RSF model did not pick sex as an important variable, while it is significant in the Cox PH model.

The Cox PH model had a better predictive performance in the presence of only those covariates that satisfy the Cox PH assumption compared to the RSF models. This result provides further support for the hypothesis that the Cox PH model works best under this assumption. In contrast, the out-of-bag error rate for the RSF with (log-rank and log-rank-score) before imputation is higher than that after imputation. This result implies that the imputation of missing values is a critical step and enormously improves the model's performance.

The most striking result to emerge from the analysis of the RSF is that log-rank has a better performance compared to the log-rank-score split-rule [44]. However, with more survival trees the log-rank-score seems to be stabilize compared to a smaller number of survival trees.

Table 6. Comparison of the models using the integrated brier scores.

Methods	Before Imputation	After Imputation
Kaplan Meier	0.199	0.201
RSF (Log-rank)	0.192	0.198
RSF (Log-rank score)	0.198	0.202
Cox PH	0.228	0.212

<https://doi.org/10.1371/journal.pone.0261625.t006>

We presented the development and validation of a robust five-gene signature (*SLC9A8*, *IER5*, *ARSJ*, *ANKRD27*, and *PIPOX*), which predicted overall survival (OS) for CRC patients. This gene signature was captured using Cox PH and RSF models based on two different scenarios. However, our study results successfully confirmed genes (markers) associated with CRC directly and identified new markers to enrich the field's literature further. Furthermore, the results support previous studies such as Mohammed *et al.* [56], where age, sex, and stages were also shown to be related to CRC survival.

Conclusion

Colorectal cancer (CRC) is a major cause of morbidity and mortality worldwide annually, making CRC the fourth common cause of death from cancer. However, the incidence of CRC has been steadily growing around the world, especially in developing countries. Therefore, the recent advances in technologies such as microarrays allowed for early detection screening using the individual's gene expression profiles.

The present study was designed to identify the genes prognosis of CRC. We developed a robust gene marker associated with the CRC overall survival based on gene expression data generated from microarray, using Cox PH and RSF models before and after missing data imputation. The most prominent finding to emerge from this study is that the Cox PH model identified five genes (*SLC9A8*, *IER5*, *ARSJ*, *ANKRD27*, and *PIPOX*) related to CRC overall survival in addition to age, sex (after imputation), and clinical stages. The RSF model further confirmed these results and had five additional gene markers predicting CRC survival. In addition, imputation improved the model's performance, and the current findings support the relevance of the missing data imputation. In summary, we recommend using a random survival forests model for survival data, especially in the high dimensional data where many genes might violate the Cox PH assumption.

Supporting information

S1 Appendix. Summary statistics of the 54 genes selected for survival analysis.
(PDF)

Acknowledgments

The authors wish to acknowledge and thank Dr. Justine B Naseje for her helping in understanding and implementing the random survival forest methods.

Author Contributions

Conceptualization: Mohanad Mohammed, Henry Mwambi, Bernard Omolo.

Data curation: Mohanad Mohammed, Innocent B. Mboya.

Formal analysis: Mohanad Mohammed, Innocent B. Mboya.

Methodology: Mohanad Mohammed, Innocent B. Mboya, Henry Mwambi, Murtada K. Elbashir, Bernard Omolo.

Software: Mohanad Mohammed.

Supervision: Henry Mwambi, Murtada K. Elbashir, Bernard Omolo.

Writing – original draft: Mohanad Mohammed.

Writing – review & editing: Mohanad Mohammed, Innocent B. Mboya, Henry Mwambi, Murtada K. Elbashir, Bernard Omolo.

References

1. Favoriti P, Carbone G, Greco M, Pirozzi F, Pirozzi RE, Corcione F. Worldwide burden of colorectal cancer: a review. *Updates Surg.* 2016; 68(1):7–11. Epub 2016/04/14. <https://doi.org/10.1007/s13304-016-0359-y> PMID: 27067591.
2. Society AC. Colorectal Cancer Facts & Figures 2020–2022. Published online. 2020: 48.
3. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians.* 2021; 71(3):209–49. <https://doi.org/10.3322/caac.21660> PMID: 33538338
4. Dai GP, Wang LP, Wen YQ, Ren XQ, Zuo SG. Identification of key genes for predicting colorectal cancer prognosis by integrated bioinformatics analysis. *Oncology letters.* 2020; 19(1):388–98. <https://doi.org/10.3892/ol.2019.11068> PMID: 31897151
5. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians.* 2018; 68(6):394–424. <https://doi.org/10.3322/caac.21492> PMID: 30207593
6. Stintzing S. Management of colorectal cancer. *F1000Prime reports.* 2014; 6:108. <https://doi.org/10.12703/P6-108> PMID: 25580262
7. Bian Q, Chen J, Qiu W, Peng C, Song M, Sun X, et al. Four targeted genes for predicting the prognosis of colorectal cancer: A bioinformatics analysis case. *Oncology letters.* 2019; 18(5):5043–54. <https://doi.org/10.3892/ol.2019.10866> PMID: 31612015
8. Society AC. Colorectal Cancer Early Detection, Diagnosis, and Staging. Published online. 2020: 40.
9. Molinari C, Marisi G, Passardi A, Matteucci L, De Maio G, Ulivi P. Heterogeneity in Colorectal Cancer: A Challenge for Personalized Medicine? *International journal of molecular sciences.* 2018; 19(12):3733. <https://doi.org/10.3390/ijms19123733> PMID: 30477151
10. Bramsen JB, Rasmussen MH, Ongen H, Mattesen TB, Orntoft MW, Arnadottir SS, et al. Molecular-Subtype-Specific Biomarkers Improve Prediction of Prognosis in Colorectal Cancer. *Cell reports.* 2017; 19(6):1268–80. <https://doi.org/10.1016/j.celrep.2017.04.045> PMID: 28494874
11. Ogino S, Nowak JA, Hamada T, Phipps AI, Peters U, Milner DA Jr., et al. Integrative analysis of exogenous, endogenous, tumour and immune factors for precision medicine. *Gut.* 2018; 67(6):1168–80. <https://doi.org/10.1136/gutjnl-2017-315537> PMID: 29437869
12. Abdul Aziz NA, Mokhtar NM, Harun R, Mollah MM, Mohamed Rose I, Sagap I, et al. A 19-Gene expression signature as a predictor of survival in colorectal cancer. *BMC medical genomics.* 2016; 9(1):1–13. <https://doi.org/10.1186/s12920-016-0218-1> PMID: 27609023
13. Pan F, Chen T, Sun X, Li K, Jiang X, Försti A, et al. Prognosis prediction of colorectal cancer using gene expression profiles. *Frontiers in oncology.* 2019; 9:252. <https://doi.org/10.3389/fonc.2019.00252> PMID: 31024853
14. Martinez-Romero J, Bueno-Fortes S, Martin-Merino M, Ramirez de Molina A, De Las Rivas J. Survival marker genes of colorectal cancer derived from consistent transcriptomic profiling. *BMC Genomics.* 2018; 19(8):45–60. <https://doi.org/10.1186/s12864-018-5193-9> PMID: 30537927
15. Yan Z, Li J, Xiong Y, Xu W, Zheng G. Identification of candidate colon cancer biomarkers by applying a random forest approach on microarray data. *Oncology reports.* 2012; 28(3):1036–42. <https://doi.org/10.3892/or.2012.1891> PMID: 22752057
16. Marisa L, de Reynies A, Duval A, Selves J, Gaub MP, Vescovo L, et al. Gene Expression Classification of Colon Cancer into Molecular Subtypes: Characterization, Validation, and Prognostic Value. *Plos Medicine.* 2013; 10(5):e1001453. <https://doi.org/10.1371/journal.pmed.1001453> PMID: 23700391
17. Simon R, Lam A, Li M-C, Ngan M, Meneses S, Zhao Y. Analysis of gene expression data using BRB-array tools. *Cancer informatics.* 2007; 3:11–7. PMID: 19455231
18. Chaba L, Odhiambo J, Omolo B. Evaluation of methods for gene selection in melanoma cell lines. *International Journal of Statistics in Medical Research.* 2017; 6(1):1–9. <http://dx.doi.org/10.6000/1929-6029>.
19. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics.* 2003; 19(2):185–93. <https://doi.org/10.1093/bioinformatics/19.2.185> PMID: 12538238

20. Mohammed M, Mwambi H, Omolo B, Elbashir MK. Using stacking ensemble for microarray-based cancer classification. 2018 International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE): IEEE; 2018. p. 1–8.
21. Pedersen AB, Mikkelsen EM, Cronin-Fenton D, Kristensen NR, Pham TM, Pedersen L, et al. Missing data and multiple imputation in clinical epidemiological research. *Clinical epidemiology*. 2017; 9:157–66. <https://doi.org/10.2147/CLEP.S129785> PMID: 28352203
22. Hayati Rezvan P, Lee KJ, Simpson JA. The rise of multiple imputation: a review of the reporting and implementation of the method in medical research. *BMC medical research methodology*. 2015; 15(1):1–14. <https://doi.org/10.1186/s12874-015-0022-1> PMID: 25880850
23. Mboya IB, Mahande MJ, Obure J, Mwambi HG. Predictors of perinatal death in the presence of missing data: A birth registry-based study in northern Tanzania. *PLoS One*. 2020; 15(4):e0231636. <https://doi.org/10.1371/journal.pone.0231636> PMID: 32298332
24. Mboya IB, Mahande MJ, Obure J, Mwambi HG. Predictors of singleton preterm birth using multinomial regression models accounting for missing data: A birth registry-based cohort study in northern Tanzania. *Plos one*. 2021; 16(4):e0249411. <https://doi.org/10.1371/journal.pone.0249411> PMID: 33793638
25. Jakobsen JC, Gluud C, Wetterslev J, Winkel P. When and how should multiple imputation be used for handling missing data in randomised clinical trials—a practical guide with flowcharts. *BMC medical research methodology*. 2017; 17(1):1–10. <https://doi.org/10.1186/s12874-016-0277-1> PMID: 28056835
26. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? *International journal of methods in psychiatric research*. 2011; 20(1):40–9. <https://doi.org/10.1002/mpr.329> PMID: 21499542
27. Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*. 2009; 338:b2393. <https://doi.org/10.1136/bmj.b2393> PMID: 19564179
28. Bradburn MJ, Clark TG, Love SB, Altman DG. Survival analysis part II: multivariate data analysis—an introduction to concepts and methods. *British journal of cancer*. 2003; 89(3):431–6. <https://doi.org/10.1038/sj.bjc.6601119> PMID: 12888808
29. Ajagbe OB, Kabir Z, O'Connor T. Survival analysis of adult tuberculosis disease. *PLoS One*. 2014; 9(11):e112838. <https://doi.org/10.1371/journal.pone.0112838> PMID: 25409024
30. Kleinbaum DG, Klein M. *Survival analysis*: Springer; 2010.
31. Myte R. Covariate selection for colorectal cancer survival data: A Comparison case study between random survival forests and the cox proportional-hazards model: Umeå University; 2013.
32. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random Survival Forests. *The annals of applied statistics*. 2008; 2(3):841–60. <https://doi.org/10.1214/08-Aoas169>
33. Wang H, Li G. A Selective Review on Random Survival Forests for High Dimensional Data. *Quant Biosci*. 2017; 36(2):85–96. <https://doi.org/10.22283/qbs.2017.36.2.85> PMID: 30740388
34. Breiman L. Random forests. *Machine Learning*. 2001; 45(1):5–32. <https://doi.org/10.1023/A:1010933404324>
35. Jiang S. Prediction Based on Random Survival Forest. *American Journal of Biomedical Science & Research*. 2019; 6(2):109–11. <https://doi.org/10.34297/ajbsr.2019.06.001005>
36. Ishwaran H, Kogalur UB, Kogalur MUB. Package 'randomForestSRC'. 2021; 6:1–125.
37. Nelson W. Theory and applications of hazard plotting for censored failure data. *Technometrics*. 1972; 14(4):945–66. <https://doi.org/10.2307/1267144>
38. Ciampi A, Chang C-H, Hogg S, McKinney S. Recursive partition: A versatile method for exploratory-data analysis in biostatistics. *Biostatistics*: Springer; 1987. p. 23–50.
39. Nasejje JB, Mwambi H, Dheda K, Lesosky M. A comparison of the conditional inference survival forest model to random survival forests based on a simulation study as well as on two applications with time-to-event data. *BMC medical research methodology*. 2017; 17(1):1–17. <https://doi.org/10.1186/s12874-016-0277-1> PMID: 28056835
40. Hothorn T, Lausen B. On the exact distribution of maximally selected rank statistics. *Computational Statistics & Data Analysis*. 2003; 43(2):121–37. [https://doi.org/10.1016/S0167-9473\(02\)00225-6](https://doi.org/10.1016/S0167-9473(02)00225-6)
41. Graf E, Schmoor C, Sauerbrei W, Schumacher M. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in medicine*. 1999; 18(17–18):2529–45. [https://doi.org/10.1002/\(sici\)1097-0258\(19990915/30\)18:17/18<2529::aid-sim274>3.0.co;2-5](https://doi.org/10.1002/(sici)1097-0258(19990915/30)18:17/18<2529::aid-sim274>3.0.co;2-5) PMID: 10474158
42. Ehrlinger J. ggRandomForests: Exploring random forest survival. arXiv preprint arXiv:161208974. 2016.
43. Taylor JM. Random Survival Forests. *Journal of Thoracic Oncology*. 2011; 6(12):1974–5. <https://doi.org/10.1097/JTO.0b013e318233d835> PMID: 22088987

44. Nasejje JB, Mwambi H. Application of random survival forests in understanding the determinants of under-five child mortality in Uganda in the presence of covariates that satisfy the proportional and non-proportional hazards assumption. *BMC research notes*. 2017; 10(1):1–18. <https://doi.org/10.1186/s13104-016-2345-3> PMID: 28057050
45. Gerds TA. Package 'pec'. R package version. 2020.
46. Team RC. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, 2021.
47. Chen G, Kim S, Taylor JM, Wang Z, Lee O, Ramnath N, et al. Development and validation of a quantitative real-time polymerase chain reaction classifier for lung cancer prognosis. *Journal of Thoracic Oncology*. 2011; 6(9):1481–7. <https://doi.org/10.1097/JTO.0b013e31822918bd> PMID: 21792073
48. Zang CZ, Wang T, Deng K, Li B, Hu SE, Qin Q, et al. High-dimensional genomic data bias correction and data integration using MANCIE. *Nature Communications*. 2016; 7(1):1–8. <https://doi.org/10.1038/ncomms11305> PMID: 27072482
49. WHO. Cancer: WHO; 2021 [updated 03/03/2021; cited 2021 25/05/2021]. Available from: <https://www.who.int/news-room/fact-sheets/detail/cancer>.
50. Gerds TA, Kattan MW, Schumacher M, Yu C. Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring. *Statistics in Medicine*. 2013; 32(13):2173–84. <https://doi.org/10.1002/sim.5681> PMID: 23172755
51. van Eeghen EE, Bakker SD, van Bochove A, Loffeld RJ. Impact of age and comorbidity on survival in colorectal cancer. *Journal of gastrointestinal oncology*. 2015; 6(6):605–12. <https://doi.org/10.3978/j.issn.2078-6891.2015.070> PMID: 26697191
52. Jiang Z, Wang X, Tan X, Fan Z. Effect of Age on Survival Outcome in Operated and Non-Operated Patients with Colon Cancer: A Population-Based Study. *PLoS One*. 2016; 11(1):e0147383. <https://doi.org/10.1371/journal.pone.0147383> PMID: 26789841
53. Chandrasinghe PC, Ediriweera DS, Nazar T, Kumarage S, Hewavisenthi J, Deen KI. Overall Survival of Elderly Patients Having Surgery for Colorectal Cancer Is Comparable to Younger Patients: Results from a South Asian Population. *Gastroenterology Research and Practice*. 2017; 2017:1–7. <https://doi.org/10.1155/2017/9670512> PMID: 28811822
54. White A, Ironmonger L, Steele RJC, Ormiston-Smith N, Crawford C, Seims A. A review of sex-related differences in colorectal cancer incidence, screening uptake, routes to diagnosis, cancer stage and survival in the UK. *BMC Cancer*. 2018; 18(1):1–11. <https://doi.org/10.1186/s12885-017-3892-2> PMID: 29291726
55. Abancens M, Bustos V, Harvey H, McBryan J, Harvey BJ. Sexual Dimorphism in Colon Cancer. *Frontiers in Oncology*. 2020; 10:1–27. <https://doi.org/10.3389/fonc.2020.00001> PMID: 32076595
56. Mohammed M, Mwambi H, Omolo B. Colorectal Cancer Classification and Survival Analysis Based on an Integrated RNA and DNA Molecular Signature. *Current Bioinformatics*. 2021; 16(4):583–600. <https://doi.org/10.2174/1574893615999200711170445>