Using Machine Learning to Predict the Future Fatigue of Patients with Colorectal Cancer, Endometrial Cancer, Ovarian Cancer, and Multiple Lymphoma Types

Dhirendra Adiprakoso

Academic Year: 2022/2023
Supervisors: dr. J. A. van Til & dr. J. Mikhal
Health Science, S&T Faculty, University of Twente

Abstract

Study objective and motivation

Of the common side effects experienced by cancer patients, fatigue is the most common across cancer types, cancer treatment regimens, and long-term survivors. Previous studies have shown that patients still experience fatigue even after treatment, affecting a patient's health-related quality of life (HRQoL). To mitigate the long-term effects of fatigue, prediction models are a tool that could be used to predict future fatigue. To the best of the author's knowledge, a research gap exists to investigate the use of machine learning (ML) to predict future fatigue within cancer patients, especially on an individual level. Moreover, a research gap exists combining different cancer types when researching the prediction of future fatigue.

Therefore, this study aimed to explore the use of ML algorithms to predict the future fatigue of patients across patients with colorectal cancer, ovarian cancer, endometrial cancer, bladder cancer, and varying types of lymphoma. This consisted of predicting clinically relevant fatigue after 24-36 months, as well as change in fatigue scores after said period. To do this, sociodemographic and clinical factors, as well as HRQoL and symptoms reported within 12 months after diagnosis were used as predictors. Furthermore, data gathered from questionnaires completed within 12 months after diagnosis and within 24-36 months thereafter were used for prediction.

Methods

This study created prediction models predicting the presence of clinically relevant fatigue after 24 to 36 months (classification) and predicting the change in fatigue for a patient after 24 to 36 months (regression). Missing values within predictor variables were imputed. K-Nearest Neighbours (kNN) imputation was done on HRQoL score, functioning scores, and symptom scores, while multivariate imputation using chained equations (MICE) was done on clinical and sociodemographic factors. This separation on imputation method was made to retain the realistic nature of imputed values from missing data.

Extreme gradient boosting (XGBoost), support vector machines (SVM), and artificial neural networks (ANN) were utilised for prediction model development. A regression model was used as a reference for comparison. To avoid overfitting, repeated ten-fold cross-validation was conducted on each prediction model. Model outputs were analysed and compared based on different metrics for predicting the presence of future fatigue (accuracy, balanced accuracy, precision, sensitivity, and specificity) and predicting future change in fatigue (root mean square error, symmetric mean absolute percentage error (SMAPE), and R-squared). Furthermore, external validation was conducted and assessed using the statistical metrics and calibration plots.

Results

When predicting clinically relevant fatigue (classification), all prediction models attained an average area under the receiver operating curve (AUC-ROC) value above 0.85 with low standard deviation. Further, the reference regression model had the highest average AUC-ROC value (0.934) as well as the lowest difference between sensitivity and specificity. This implies a strong preference for the reference regression model to predict the presence of clinically relevant, future fatigue. On the other hand, when predicting the future change in fatigue (regression), the prediction models were not able to perform the prediction task well. The prediction models produced average R-squared values between 0.019 and 0.058 and average SMAPE values between 1.66 and 1.787.

Conclusion

This study shows that while most ML models predict reasonably well, there is no model that performs best on all quality indicators. This study found that the models were able to predict the presence of future fatigue well. However, no model was able to predict the future change in fatigue of patients. Other than this, this study showed the feasibility of combining multiple cancer types into a prediction model. Future research should explore further into future fatigue within patients of multiple cancer types, explore the influence of dichotomised symptom outcomes when developing prediction models within the cancer domain, and compare the use of different tools measuring fatigue in cancer patients for building prediction models using ML.

Introduction

Every year, the cancer survival rate increases as standards for treatment and care are continually improved. In the Netherlands alone, the number of cancer survivors increased from over 800,000 to over 1,000,000 from the 2001-2010 period to the 2011-2020 period [1]. Moreover, although cancer-related mortality in the Netherlands increased from 42,858 to 44,839 between 2010 and 2015, this number plateaued to under 46,000 by 2021 [2]. Currently, in the Netherlands, more than 850,000 people have or survived cancer (20-year prevalence as per 1/1/2022) [3]. Together, these statistics show how much cancer-related treatment has improved over the past decade, especially in terms of handling the increasing prevalence over 20 years. However, even after treatment, cancer patients can still experience long-term side effects, such as fatigue, depression, and problems with infertility [4].

Of the common side effects experienced by cancer patients, fatigue is the most common across cancer types, cancer treatment regimens, and long-term survivors [5]. Regarding the latter, previous studies have shown that patients still experience fatigue even after treatment [4, 6-7], which can have a long-term effect on a patient's health-related quality of life (HRQoL) [4]. Long-term fatigue was found in cancer patients with, among others, non-Hodgkin Lymphoma [4], endometrial cancer, and ovarian cancer [6]. Moreover, significant fatigue was reported across both male and female populations [7]. Notably, Poort et al. highlighted the importance of "developing scalable and effective transdiagnostic interventions to reduce fatigue" [6]. This is especially the case since fatigue is often persistent in the long-term, i.e., 12-24 months after treatment [6]. This implies the prolonged need for supportive care handling fatigue even if a patient had successful curative treatment.

To understand the extent of the required supportive care, prediction models is a tool that can be used for this purpose. Within a clinical context, prediction models allow a clinician to anticipate and be adequately informed of a patient's need for supportive care [8]. The use of prediction models within the cancer domain ranges from predicting cancer survival [9-11] to overall functioning and HRQoL [12-13]. With respect to providing supportive care to handle long-term fatigue, prediction models provide the possibility to predict the future fatigue within a cancer patient. Therefore, clinicians are provided with necessary information for assigning supportive care regimens.

To be able to predict long-term fatigue using prediction models, these models would need to be fitted on a type of measurement. A common method to measure a patient's fatigue is to use patient reported outcome measures (PROM). These are tools or instruments used to measure health-, quality-of-life-, or function-related responses directly reported by patients without the interpretation of clinicians or other stakeholders [14]. An example of this within the cancer domain is the European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire Core-30 (EORTC QLQ-C30), which is a questionnaire that assesses a cancer patient's quality of life, symptoms, and functioning [15]. This PROM is particularly applicable since it has been used and validated in previous studies [16-17].

Within cancer research, machine learning (ML) is a tool that is increasingly utilised for developing prediction models [18]. ML is defined as algorithms that utilise past data to make predictions or decisions [19]. Critically, in developing prediction models, especially when ML is used, attention is paid towards generalisability and reproducibility. This implies, respectively, a focus on how the model's output can be reproduced under modified or new conditions, and whether a model can be effective in different applications [20]. Previous studies have investigated using ML models to predict cancer susceptibility [21-22], recurrence [23-24], and survival [9-11]. Other studies have also used PROMs when developing ML models to predict future HRQoL and symptoms of cancer patients [12-13]. However, within the literature, there is a greater number of studies focusing on predicting cancer susceptibility, recurrence, and survival rather than predicting cancer symptoms, including fatigue. This comes despite cancer symptoms being another aspect that could be investigated with regards to the development of prediction models [14]. Moreover, of the PROMs utilised for prediction, the EORTC QLQ-C30 is not often used within the literature. Hence, this implies a research gap with respect to developing prediction models for future fatigue using ML as well as the EORTC QLQ-C30.

In addition, while there are cancer-related registries available for research use [25-26], previous research mainly focused on one cancer type for predicting future fatigue [e.g., 12-13, 28]. This narrow scope loses the possibility to generalise a prediction model for predicting future fatigue to patients of multiple, different cancer types. By developing a prediction model that can be applied to patients of multiple cancer types, a helpful tool is provided to clinicians whereby an all-round model can be applied into different contexts with ease, simplifying the process for assigning appropriate (supportive) care regimes. Next to this, previous studies predicting fatigue utilised a binary variable, often against a baseline value, implying that the models predicted the presence of fatigue after a defined period. However, there is added value in investigating the change in symptoms of cancer patients, including fatigue, since it can contribute to developing tailored care [29-30]. Therefore, a further research gap is present with respect to developing a prediction model capable of predicting the future change in fatigue of cancer patients.

Following the described research gaps, this study aims to predict the future fatigue of patients across multiple types of cancer, namely colorectal cancer, ovarian cancer, endometrial cancer, and various types of lymphoma, externally validating such a model with bladder cancer patients' data. Specifically, this study will use data from historic longitudinal cohort studies that utilised the EORTC-QLQ-C30 questionnaire to investigate whether such patients experience clinically relevant, future fatigue, using questionnaires answered within 12 months after diagnosis and between 24 and 36 months thereafter. This includes using sociodemographic and clinical factors, as well as HRQoL and symptoms reported within 12 months after diagnosis as predictors. Moreover, using the same prediction parameters, this study aims to predict the future fatigue change of cancer patients within the two timepoints. To understand the extent to which extent ML can be used to achieve these aims, different ML algorithms will be compared to a regression model. Ultimately, this study aims to answer the following questions:

RQ1. Using external validation, can clinically relevant, future fatigue be predicted for patients with different cancer types using clinical factors, sociodemographic factors, and HRQoL and symptoms reported within 12 months after diagnosis, using questionnaires answered within 12 months after diagnosis and between 24 and 36 months thereafter?

RQ2. Using the same prediction parameters as RQ1 as well as external validation, can the same model predict future fatigue change for patients with different cancer types?

RQ3. How do different ML algorithms compare when predicting the future fatigue of cancer patients?

Methodology

Data collection

This study collected data from the Patient-Reported Outcomes Following Initial Treatment and Long-term Evaluation of Survivorship (PROFILES) registry. PROFILES contains longitudinal sociodemographic and EORTC-QLQ-C30 data from various cohort studies of patients diagnosed with different types of cancer [24-25]. Treatment- and tumour-related characteristics from the National Cancer Registry are also linked to the PROFILES registry [26]. Ethical approval was sought after and approved by the University of Twente's Behavioural, Management, and Social Science (BMS) Faculty's Ethics Committee. This study used data from cohort studies of patients with colorectal cancer (PROCORE), ovarian cancer (ROGY), endometrial cancer (ROGY), bladder cancer (BlaZib), and varying types of lymphoma (LYMPHOMA) – i.e., Hodgkin lymphoma (HL), non-Hodgkin lymphoma (NHL), chronic lymphocytic leukaemia (CLL), and multiple myeloma (MM).

Notably, each cohort study collected its data differently in terms of the time intervals between follow-ups. Table 1 provides an overview of the data collection timelines for each cohort. Note that for the ROGY cohort, although patients' data was collected in terms of time since treatment, time since diagnosis was still recorded. To accommodate the difference in time intervals, patients were selected based on the time elapsed between filling out the first questionnaire and filling out the latest follow-up questionnaire. Specifically, patients were selected if the time elapsed within this interval was between 24 and 36 months. Moreover, a patient's baseline was defined as their first observation, given that said patient provided an observation under 12 months after diagnosis. Any patient with a baseline beyond 12 months after diagnosis was not included. This avoided the possibility for a model to assume that an observation from a patient answering immediately after diagnosis had equal weight as an observation from a patient answering beyond 12 months after diagnosis.

Table 1. Overview of data collection time intervals for each cohort study.

Cohort	***Time since diagnosis (months)***							***Time since initial treatment (months)***		
	0	3	6	12	18	24	36	0	12	24
PROCORE [31]	✓			✓		✓				
ROGY [32]								✓	✓	✓
LYMPHOMA [4]	✓*			✓*						

BlaZib [33]	✓		✓	✓		✓

*LYMPHOMA cohort studies do not have a coherent data collection schedule, only those referenced are shown.

Because of the variety of data made available to this study, the cohorts were separated for model development and external validation. The PROCORE, ROGY, and LYMPHOMA sets were used for model development, while BlaZib was used for external validation. Aside from being the datasets made available for this study's research, this division has an added benefit of exploring the generalisability of the developed ML models through external validation, especially across patients with different cancer types. In the following sub-sections, unless specified, the mention of used datasets refer to those used for model development.

Data preparation

Prior to conducting pre-processing, the prescribed cohort datasets were merged into one dataset and each variable was prepared depending on the presence of missing values or differences in categorisation method per cohort. All steps described in this section were done using R version 4.2.1.

Firstly, the following steps were done to harmonise the missing entries within the dataset. Note that these steps did not omit patient data, rather omitting the datapoint directly:

1. Categorical variables were initially given an alphanumeric input to indicate a missing entry. These entries were omitted. Furthermore, categorical variables with unknown values were identified and omitted. Appendix 1 shows which variables had missing values recoded.
2. Datapoints that were unrealistic were omitted. This was done to ensure the dataset maintained a realistic representation of the sample and to avoid overfitting the model [34]. These unrealistic datapoints included BMI below the lowest recorded (i.e., BMI < 6.7) [35], BMI above the highest recorded (i.e., BMI > 105.7) [36], and an unrealistic recorded time since diagnosis to the filling of the questionnaire (i.e., negative time).

Following this, although time since diagnosis was used as the criterion for filtering patients, this data was occasionally missing, particularly within follow-up observations. Using the known follow-up time intervals of each cohort, missing time since diagnosis values were inputted based on their respective follow-up schedules. This step ensured including as many patients as possible for model development since 9.4% of the data contained missing time since diagnosis values. This step applied for the PROCORE, ROGY, and BlaZib cohorts. Meanwhile, because the LYMPHOMA cohort did not have a coherent data collection schedule, the missing time since diagnosis values could not be inputted using a well-defined assumption. Therefore, only patients with known time since diagnosis values were filtered through.

Aside from this, while the PROCORE dataset coded age as a categorical variable, the LYMPHOMA, ROGY, and BlaZib datasets coded age as both a categorical and continuous variable. To harmonise this, all ages were categorised under the same rules (see Appendix 1). Afterwards, the LYMPHOMA, ROGY, and BlaZib cohorts were checked to see whether they contained patients that had categorical age data, but without continuous age data. In such cases, the recorded age category remained. This renewed variable was used as the age variable for pre-processing and model development.

Data pre-processing

Clinical-, sociodemographic-, and EORTC QLQ-C30-related variables were used as predictor variables for this study. Appendix 2 provides an overview of these variables in terms of what they describe and how they were coded within the dataset. Note that the "systemic" variable implied whether a patient underwent either chemotherapy, targeted therapy, or immunotherapy. All pre-processing steps were done using R version 4.2.1.

For this study, two prediction tasks were conducted. Firstly, the prediction models conducted a classification task. This implied that the models sought to predict a label from an outcome variable within a finite set [37]. The prediction outcomes from this task implied how well the model was able to predict the presence of future fatigue within a cancer patient, answering RQ1. To prepare for this, the outcome variable, i.e., the fatigue scores of patients based on the EORTC QLQ-C30, underwent categorisation based on the clinical thresholds defined by Giesinger et al. [37]. This dichotomised the variable into *[0,1]*, of which "*1*" indicated an individual having clinically-relevant fatigue. For fatigue, the cut-off point was set at 39, implying scores above that indicated the presence of clinically-relevant fatigue [38].

The other prediction task conducted by the prediction models was regression. This implied that the models sought to predict a continuous variable [37]. The prediction outcomes from this task implied how well the model was able to predict the extent of future fatigue within a cancer patient, answering RQ2. For this task, the outcome variable was a patient's difference in fatigue scores between baseline and at endpoint. The baseline was defined as a patient's first observation, given that said patient provided an observation under 12 months after diagnosis, while the endpoint was defined as the patient's fatigue score at the last point they filled in the questionnaire, given that said patient completed follow-up between 24 and 36 months. Within the subsequent (sub-)sections of this study, a patient's baseline ($T_{baseline}$) and endpoint ($T_{endpoint}$) will use these definitions.

Multiple imputation

After examining the collected data, missing datapoints were due to patients inadequately answering the questionnaires both at $T_{baseline}$ and at $T_{endpoint}$. If only complete cases were used, this led to an information loss of over 70%. Moreover, a patients' data at baseline could not be copied over to their $T_{endpoint}$ or vice-versa since this risked an unrealistic representation of the patient, especially since the gap between observations is on the long-term. This thusly implied data being missing at random [39]. Hence, to avoid a significant information loss [39-40], imputation was done to handle the missing values present in predictor variables. EORTC QLQ-C30-related predictor variables were imputed using k-Nearest Neighbours (kNN). This implied grouping datapoints based on the proximity of an individual datapoint to a defined group [40]. Following the guidelines set by Aaronson et al., only observations whereby less than half of the EORTC QLQ-C30 questions were answered were treated as missing [15]. Hence, only observations where these occurred were checked and imputed. This process was done using the "VIM" package [42] within R version 4.2.1.

On the other hand, missing values present in clinical- and sociodemographic-related predictor variables were imputed using multivariate imputation by chained equations (MICE). Quoting from Buuren and Groothuis-Oudshoorn [43], this process, "specifies the multivariate imputation model on a variable-by-variable basis by a set of conditional densities, one for each incomplete variable." MICE has been noted for its good imputation performance, leading to smaller standard errors and narrower confidence intervals [44]. Furthermore, MICE is noted for its ability to retain the realistic nature of imputed values over predictor variables that might be related [43]. MICE was conducted using the MICE package available under R-CRAN's repository [45] using R version 4.2.1. The following paragraphs in this sub-section further describe the MICE process for this study.

Prior to conducting the imputation process, imputer variables were defined. This enabled a pre-selection on which variables to use for obtaining prediction values in place of the missing datapoint. Notably, EORTC QLQ-C30-related variables were not used to impute the missing datapoints in other variables. This ensured that the imputed estimates remain close to the patient- and tumour-related characteristics of patients records within the dataset. Aside from this, imputation methods were defined individually for each variable. Binary variables utilised logistic regression, non-ordered multinomial variables utilised polynomial regression, ordered multinomial variables utilised an ordered logit model, and continuous variables utilised predictive mean matching. The latter process obtained a random value from an observation regression-predicted value that is closest to the regression-predicted value for a missing datapoint [46]. Appendix 3 lists the variables imputed along with their respective imputation methods.

Following the imputation process, preliminary analysis was conducted to explore intervariable relationships. This was done through pooling together estimates from each imputed dataset and applying the Pearson correlation test on the pooled dataset. A correlation plot was created to show the correlational relationships between predictor variables and patients' fatigue scores.

Model development and internal validation

Following data pre-processing, the data was split such that the patients' data at $T_{baseline}$ were used as training data while the patients' data at $T_{endpoint}$ were used as testing data. To ensure the robustness of the models' development, repeated cross-validation was utilised. This validation method was selected because of its computational efficiency [47] and advantage over the random split method [48]. This process is further described later in this section. For this study, alongside a reference model that utilises regression, three ML algorithms were used, namely eXtreme gradient-boosted random forest (XG-Boost), support vector machines (SVM), and artificial neural networks (ANN). These models were selected since each have shown their proficiency as tools for developing prediction models in previous studies and literature reviews [9-10, 49-50]. Note that although two different methods for quantifying fatigue were used for analysis, this did not impact the model development procedure. All model development steps were done using R version 4.2.1. This sub-section will first describe the model development process for the reference regression model. Afterwards, the steps taken for developing the XG-Boost, SVM, and ANN models are described.

For the reference regression model, firstly, all predictor variables were fitted. Afterwards, variables were selected by backwards selection using the Bayesian Information Criterion (BIC). This method selected a preferred model by adding a penalty based on the number of predictor variables within the model [see 51]. The process started from fitting all predictor variables into the model and after applying a penalty on a predictor variable, eliminated said variable from the model. The process ceased when no further improvement can be made from eliminating a predictor variable. Neath and Cavanaugh [51] further described possible benefits for using the BIC method for variable selection, such as its consistency characteristic [52] as well as this selection procedure's tendency to choose parsimonious models, i.e., models with fewer variables. Once a preferred model was selected with optimal predictor variables, this model was tuned using ten-times-repeated ten-fold cross-validation. This internally validated the model by using a random subset ten times interchangeably until all observations within the training data had been used for validation [47]. This process was repeated ten times such that, in total, there are 100 instances where the model had been internally validated.

For each ML model, an initial model was created using all predictor variables and with default settings provided by the associated R package ("xgboost" for XG-Boost, "e107c1" for SVM, and "nnet" for ANN). The SVM model used a linear kernel for prediction, while the ANN model utilised a simple feedforward structure with hidden layers defined during hyperparameter tuning [53]. No further variable selection process was done for each ML model because such a process is part of each model's framework [see 24, 28]. Once fitted, each model's hyperparameters were tuned to optimise model performance. Hyperparameter tuning was done using a grid search on the required hyperparameters for each model. This implied that, through a grid containing possible ranges of each hyperparameter, every possible combination was tested to see which was the most optimal [53]. Each hyperparameter was selected based on the set required by the *train()* function under the "caret" package [see 53 for further documentation]. Table 2 lists the hyperparameters tuned for each ML model. Accuracy and root mean square error were used to define the most optimal hyperparameters for each model for the classification and regression tasks respectively. Alongside this step, each model utilised ten-times-repeated ten-fold cross-validation. This prevented any overfitting that can occur with a grid search for hyperparameter tuning [54]. Finally, each model was refitted using the most optimal hyperparameters.

Table 2. List of hyperparameters tuned for each ML algorithm.

Machine Learning Algorithm	***Hyperparameter Tuned***	***Grid range***
eXtreme Gradient Boosted Random Forest (XGBoost)	Maximum tree depth (max_depth)	(3, 5, 7)
	Step size of each boosting step (eta)	(0.01, 0.05, 0.71)
	Minimum loss reduction required to further partition a leaf node (gamma)	(0.1, 1, 10)
	Subsample ratio of the training instance (subsample)	(0.5, 0.6, 0.7)
Support Vector Machine (SVM)	Cost – "C" constant in Lagrange formulation (C)	([0.01, 0.1], [0.2, 1], [2, 10])
Artificial Neural Network (ANN)	Number of hidden units in the network (size)	(5, 6, 7, 8)
	Weight decay (decay)	(0.01, 0.05, 0.1)

Model output, statistical analysis, and external validation

Since this study utilised multiple imputation, this implied that each model had to be applied to each imputed dataset and their respective results pooled [55]. The pooling process is needed since the data will differ per imputed dataset. To do this, the analysis results were averaged after a model was applied to each imputed dataset, and standard deviations were reported to show the extent of variation. Papachristou et al. [see 55, figure 2] provided a description and visualisation in their study on how this process is typically conducted.

To analyse the performance of the models, because both a classification and a regression task was conducted, different metrics were calculated to assess the predictive ability of each model. This provided an answer to RQ3. Firstly, for classification, each model's accuracy, balanced accuracy, precision, sensitivity, and specificity were calculated and compared. While accuracy and precision explain how well a model creates correct predictions, sensitivity and specificity explain a model's discriminative ability [56-57]. On the other hand, balanced accuracy calculates the average accuracy among the different classes, thereby avoiding potential bias due to class imbalance [58].

For each of these metrics, a model can attain a score within the range *[0,1]*, in which a higher score implied a better-performing model. Following these metrics, a receiver operating curve (ROC) was plotted and the area underneath the curve (AUC-ROC) was calculated. This described how well each model retained its predictive performance up to a given threshold, thereby showing their respective discriminative abilities [57, 59]. For this metric, a model can attain a score within the range *[0.5,1]*, in which a higher score implied a better-performing model. Moreover, an AUC-ROC value above "*0.85*" implied a model had excellently performed the classification task.

As for the regression task, each model's root mean squared error (RMSE), symmetric mean absolute percentage error (SMAPE), and R-squared value were calculated and compared. While the RMSE provided an overall picture of the distribution of errors between prediction and observed values [60], the R-squared value showed the "proportion of the variance in the dependent variable that is predictable from independent variables" [61]. For the R-squared value, a model can attain a score within the range *[0,1]*. Hence, this metric described the goodness-of-fit of the predictor variables within each model. Like the R-squared metric, SMAPE informed which model produced a good performance on the regression task [51]. For this metric, a model can attain a score within the range *[0,2]*. While a lower value for RMSE and SMAPE indicated the better-performing model, an R-squared value close to "*1*" provided that indication. Additionally, an R-squared value close to "*0*" and a SMAPE value above "*1.5*" indicated poor performance from a prediction model.

Following statistical analysis, external validation will be conducted based on the steps recommended by Ramspek et al. [62]. This applied for predicting both the presence of future fatigue as well as the future change in fatigue. Firstly, the prediction models predicted the outcomes on the BlaZib dataset. The predicted outcomes are then compared using the same statistical analysis method as previously mentioned. This allows for a fair assessment between internal and external validation. Afterwards, calibration plots were plotted for each model's result to visualise a comparison between the predicted outcomes are and the observed outcomes [63], in this case, with respect to the BlaZib dataset Through these plots, patterns of miscalibration can be identified, which indicated whether the model is sufficiently receptive to completely new data [62].

Results

Patient characteristics

Figure 1 depicts how patients were filtered for analysis within this study. After pre-processing, 511 patients' data was used for model development, representing 26.5% of the raw data. Respectively 334 patients came from the PROCORE cohort (65.4% of cohort data), 162 patients from the ROGY cohort (31.7% of cohort data), and 15 patients from the LYMPHOMA cohort (2.9% of cohort data). Table 3 shows the descriptive statistics of each cohort after pre-processing and imputation, indicating the patient-, tumour-, and treatment-characteristics of patients at $T_{baseline}$ and at $T_{endpoint}$. In describing these characteristics, denoted percentages are with respect to the respective cohorts' population. Note that for the BlaZib cohort, only data at $T_{endpoint}$ is reported since this was the data used for external validation.

Across the cohorts used for model development (PROCORE, ROGY, and LYMPHOMA), between $T_{baseline}$ and $T_{endpoint}$, there were more patients who were married, were actively drinking alcohol, had two or more comorbidities, and underwent systemic therapy compared to radiotherapy. At $T_{baseline}$, the PROCORE cohort had a greater proportion of patients with stage 3 cancer (122; 36.5%) while the ROGY cohort had a greater proportion of patients with stage 1 cancer (114; 70.4%). Moreover, the spread of patients across age groups within each cohort were similar at both $T_{baseline}$ and $T_{endpoint}$. Regarding the proportion of patients with clinically relevant fatigue at $T_{baseline}$, the ROGY cohort had the greatest proportion of such patients (76; 46.9%) while the PROCORE cohort had the smallest proportion of such patients (46; 13.8%). Meanwhile, at $T_{endpoint}$, the ROGY cohort had the greatest proportion of patients with clinically relevant fatigue (46; 27.2%) and the ROGY cohort had the smallest proportion of patients with clinically relevant fatigue (39; 11.7%). Finally, while patients in the ROGY cohort experienced the greatest decrease in fatigue scores (-13.13) and patients in the PROCORE cohort experienced the smallest decrease in fatigue scores (-0.25), the difference between patients were similar across all the cohorts.

As for the BlaZib cohort, there were a greater number of patients who completed vocational school (625; 39.7%), were divorced (1,438; 91.4%), and were actively drinking alcohol (1,147; 72.9%). Like the ROGY cohort, patients within the BlaZib cohort were in stage 1 cancer at $T_{endpoint}$ (930; 59.1%). There were a greater proportion of older patients within the BlaZib cohort. A relatively small proportion of patients within the BlaZib cohort had clinically relevant fatigue at $T_{endpoint}$ (185; 11.8%). Finally, patients in the BlaZib cohort experienced a small decrease in fatigue scores (-1.53), but the difference between patients were in line with the other cohorts.

Figure 1. Patient selection from data collection and final data composition for analysis per cohort.

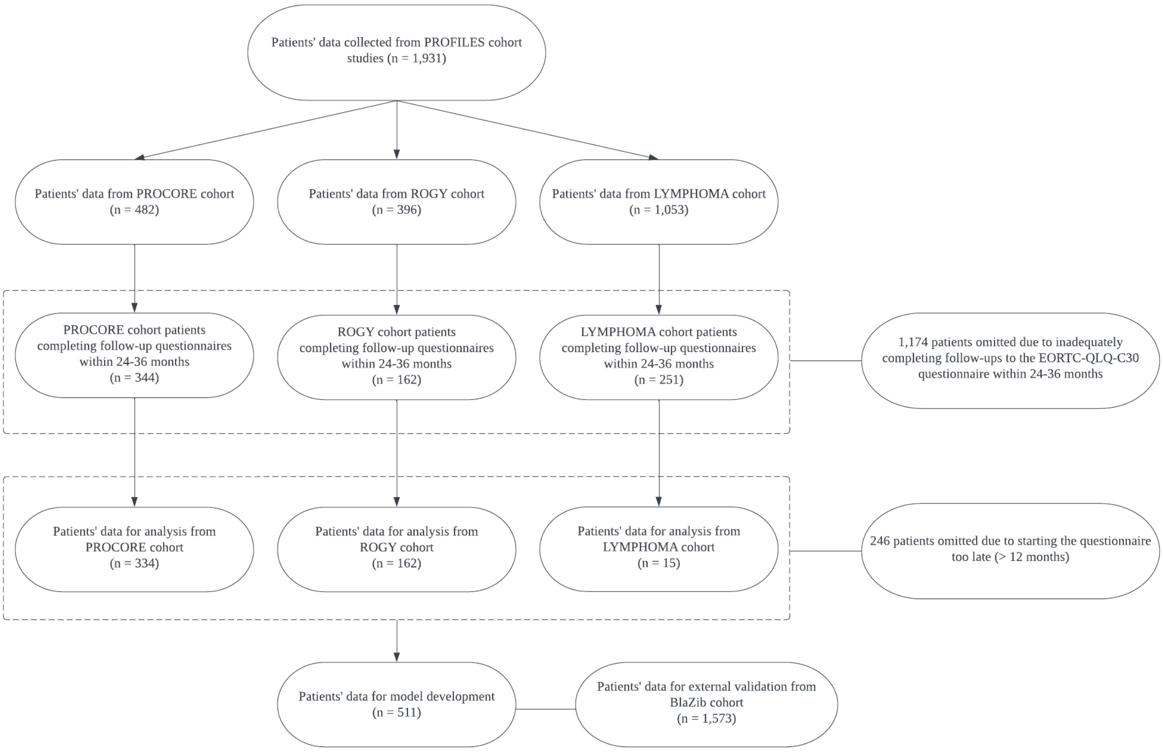

Patients' data collected from PROFILES cohort studies (n = 1,931)
Patients' data from PROCORE cohort (n = 482)
Patients' data from ROGY cohort (n = 396)
Patients' data from LYMPHOMA cohort (n = 1,053)
PROCORE cohort patients completing follow-up questionnaires within 24-36 months (n = 344)
ROGY cohort patients completing follow-up questionnaires within 24-36 months (n = 162)
LYMPHOMA cohort patients completing follow-up questionnaires within 24-36 months (n = 251)
1,174 patients omitted due to inadequately completing follow-ups to the EORTC-QLQ-C30 questionnaire within 24-36 months
Patients' data for analysis from PROCORE cohort (n = 334)
Patients' data for analysis from ROGY cohort (n = 162)
Patients' data for analysis from LYMPHOMA cohort (n = 15)
246 patients omitted due to starting the questionnaire too late (> 12 months)
Patients' data for model development (n = 511)
Patients' data for external validation from BlaZib cohort (n = 1,573)

Table 3. Patient-, tumour-, and treatment-related characteristics and patient-reported outcomes at $T_{baseline}$ and $T_{endpoint}$.

Variable	***PROCORE Cohort (N = 334, N%)***		***ROGY Cohort (N = 162, N%)***		***LYMPHOMA Cohort (N = 15, N%)***		***BlaZib Cohort (N =1,573, N%)^1***
	$T_{baseline}$	**$T_{endpoint}$**	**$T_{baseline}$**	**$T_{endpoint}$**	**$T_{baseline}$**	**$T_{endpoint}$**	
Education level							
Lower	28 (8.4)	28 (8.4)	20 (12.3)	22 (13.6)	4 (26.7)	3 (20.0)	189 (12.0)
Middle	75 (22.5)	74 (22.2)	120 (74.1)	51 (31.5)	2 (13.3)	2 (13.3)	383 (24.3)
Vocational	136 (40.1)	137 (41.0)	22 (13.6)	66 (40.7)	4 (26.7)	4 (26.7)	625 (39.7)
Higher	95 (28.4)	97 (29.0)	0 (0.0)	23 (14.2)	5 (33.3)	6 (40.0)	376 (23.9)
Marital status							
Married	280 (83.8)	276 (82.6)	134 (82.7)	133 (82.1)	11 (73.3)	10 (66.7)	135 (8.6)
Divorced	54 (16.2)	58 (17.4)	28 (17.3)	29 (17.9)	4 (26.7)	5 (33.3)	1,438 (91.4)
Smoking history							
No, never	104 (31.1)	107 (32.0)	84 (51.9)	80 (49.4)	9 (60.0)	5 (33.3)	263 (16.7)
No, but used to	195 (58.4)	202 (60.5)	60 (37.0)	68 (42.0)	6 (40.0)	9 (60.0)	1,135 (72.2)
Yes	35 (10.5)	25 (7.5)	18 (11.1)	14 (8.6)	0 (0.00)	1 (6.7)	175 (11.1)
Alcohol use							
No, never	57 (17.1)	76 (22.8)	66 (40.7)	60 (37.0)	5 (33.3)	2 (13.3)	266 (16.9)
No, but used to	22 (6.6)	23 (6.9)	22 (13.6)	10 (6.2)	5 (33.3)	7 (46.7)	160 (10.2)
Yes	255 (74.1)	235 (70.4)	74 (45.7)	92 (56.8)	5 (33.3)	6 (40.0)	1,147 (72.9)
Comorbidities							
None	87 (26.0)	108 (32.3)	43 (26.5)	36 (22.2)	1 (6.7)	2 (13.3)	336 (21.4)
One	115 (34.4)	97 (29.0)	73 (45.1)	65 (40.1)	6 (40.0)	5 (33.3)	343 (21.8)
Two or more	132 (39.5)	129 (38.6)	46 (28.4)	61 (37.7)	8 (53.3)	8 (53.3)	894 (56.8)
Quality of life (mean, SD)	76.82 (18.54)	79.82 (17.46)	70.01 (18.59)	75.46 (18.32)	72.78 (9.16)	75.00 (17.25)	80.91 (12.58)
Physical functioning (mean, SD)	90.69 (13.65)	85.96 (17.27)	81.91 (16.68)	76.69 (21.36)	71.56 (19.59)	80.89 (11.78)	83.82 (15.32)

Role functioning (mean, SD)	87.13 (22.97)	87.08 (23.43)	71.19 (25.54)	76.13 (29.07)	68.89 (31.41)	78.89 (22.24)	85.97 (19.17)
Emotional functioning (mean, SD)	78.75 (19.86)	89.15 (15.78)	78.22 (21.01)	83.56 (17.76)	77.04 (26.96)	86.11 (16.27)	92.46 (13.59)
Cognitive functioning (mean, SD)	90.07 (15.29)	86.78 (18.78)	81.58 (22.49)	83.74 (19.39)	80.00 (26.87)	83.33 (25.97)	90.53 (14.36)
Social functioning (mean, SD)	90.77 (16.50)	91.22 (17.46)	76.34 (25.11)	85.60 (21.22)	77.78 (27.94)	88.89 (19.59)	93.94 (13.73)
Fatigue (mean, SD)	17.63 (22.48)	17.45 (20.21)	40.26 (24.01)	26.89 (25.13)	33.33 (20.14)	24.44 (15.83)	20.00 (17.72)
Nausea/vomiting (mean, SD)	2.79 (8.79)	2.30 (8.35)	7.82 (17.04)	5.25 (12.93)	4.44 (9.89)	2.22 (5.86)	0.91 (4.82)
Pain (mean, SD)	9.58 (18.62)	12.38 (21.49)	25.10 (26.35)	23.97 (28.11)	21.11 (19.38)	10.00 (12.28)	7.41 (15.69)
Dyspnoea (mean, SD)	8.68 (19.36)	11.28 (21.05)	16.87 (24.99)	12.35 (23.17)	17.78 (24.77)	11.11 (24.12)	13.99 (20.88)
Sleeping disturbance (mean, SD)	20.76 (25.34)	17.76 (26.03)	32.10 (32.58)	24.90 (29.56)	31.11 (38.76)	13.33 (27.60)	11.26 (20.71)
Appetite loss (mean, SD)	7.49 (18.64)	3.89 (13.96)	13.17 (23.01)	8.64 (18.78)	20.00 (27.60)	13.33 (24.56)	2.67 (11.71)
Constipation (mean, SD)	10.38 (20.47)	8.28 (17.93)	20.58 (28.57)	13.37 (24.49)	20.00 (27.60)	4.44 (11.73)	5.40 (14.69)
Diarrhoea (mean, SD)	14.87 (24.68)	9.18 (19.73)	9.88 (21.31)	6.58 (16.96)	4.44 (11.73)	0 (0.00)	3.33 (12.39)
Financial difficulties (mean, SD)	2.30 (9.56)	4.09 (15.06)	4.53 (15.09)	5.14 (16.85)	11.11 (27.22)	13.33 (30.34)	1.91 (9.69)
Tumour type							
Non-Hodgkin Lymphoma					4 (26.7)	4 (26.7)	
Hodgkin Lymphoma					2 (13.3)	2 (13.3)	
Chronic Lymphocytic Leukaemia					3 (20.0)	3 (20.0)	
Multiple Myeloma					6 (40.0)	6 (40.0)	
Ovarian Cancer			99 (61.1)	99 (61.1)			
Endometrial Cancer			63 (38.9)	63 (38.9)			
Colon Cancer	245 (73.4)	245 (73.4)					

Rectal Cancer	89 (26.6)	89 (26.6)					
Bladder Cancer							1,573 (100.0)
Cancer stage[2]							
1	112 (33.5)	112 (33.5)	114 (70.4)	114 (70.4)	0* (0.0)	0* (0.0)	930 (59.1)
2	94 (28.1)	94 (28.1)	10 (6.2)	9 (5.6)	1* (6.7)	1* (6.7)	407 (25.9)
3	122 (36.5)	122 (36.5)	31 (19.1)	32 (19.8)	1* (6.7)	1* (6.7)	137 (8.7)
4	6 (1.8)	6 (1.8)	7 (4.3)	7 (4.3)	0* (0.0)	0* (0.0)	99 (6.3)
Age category at questionnaire							
15 – 45	6 (1.8)	4 (1.2)	2 (1.2)	0 (0.0)	1 (6.7)	0 (0.0)	11 (0.7)
46 – 50	6 (1.8)	0 (0.0)	9 (5.6)	4 (2.5)	0 (0.0)	1 (6.7)	13 (0.8)
51 – 55	29 (8.7)	16 (4.8)	17 (10.5)	12 (7.4)	3 (20.0)	1 (6.7)	37 (2.4)
56 – 60	43 (12.9)	28 (8.4)	25 (15.4)	29 (17.9)	1 (6.7)	2 (13.3)	71 (4.5)
61 – 65	64 (19.2)	66 (19.8)	42 (25.9)	26 (16.0)	0 (0.0)	1 (6.7)	135 (8.6)
66 – 70	59 (17.7)	108 (32.3)	26 (16.0)	38 (23.5)	4 (26.7)	2 (13.3)	237 (15.1)
71 – 75	86 (25.7)	63 (18.9)	22 (13.6)	29 (17.9)	2 (13.3)	2 (13.3)	355 (22.6)
76 – 80	29 (8.7)	32 (9.6)	15 (9.3)	17 (10.5)	3 (20.0)	3 (20.0)	291 (18.5)
81 - 85	12 (3.4)	17 (5.1)	4 (2.5)	7 (4.3)	1 (6.7)	3 (20.0)	236 (15.0)
85 <	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	187 (11.9)
Sex							
Male	209 (62.6)	209 (62.6)	0 (0.0)	0 (0.0)	7 (46.7)	7 (46.7)	***
Female	125 (37.4)	125 (37.4)	148 (100.0)	148 (100.0)	8 (53.3)	8 (53.3)	***
BMI (mean, SD)	26.57 (3.82)	23.29 (1.92)	28.71 (6.99)	28.83 (5.84)	27.84 (4.25)	27.49 (5.22)	26.40 (5.62)
Time since diagnosis (mean, SD)	0.07 (0.05)	2.06 (0.06)	0.24 (0.13)	2.33 (0.20)	0.79 (0.21)	3.22 (0.42)	3.33 (32.24)

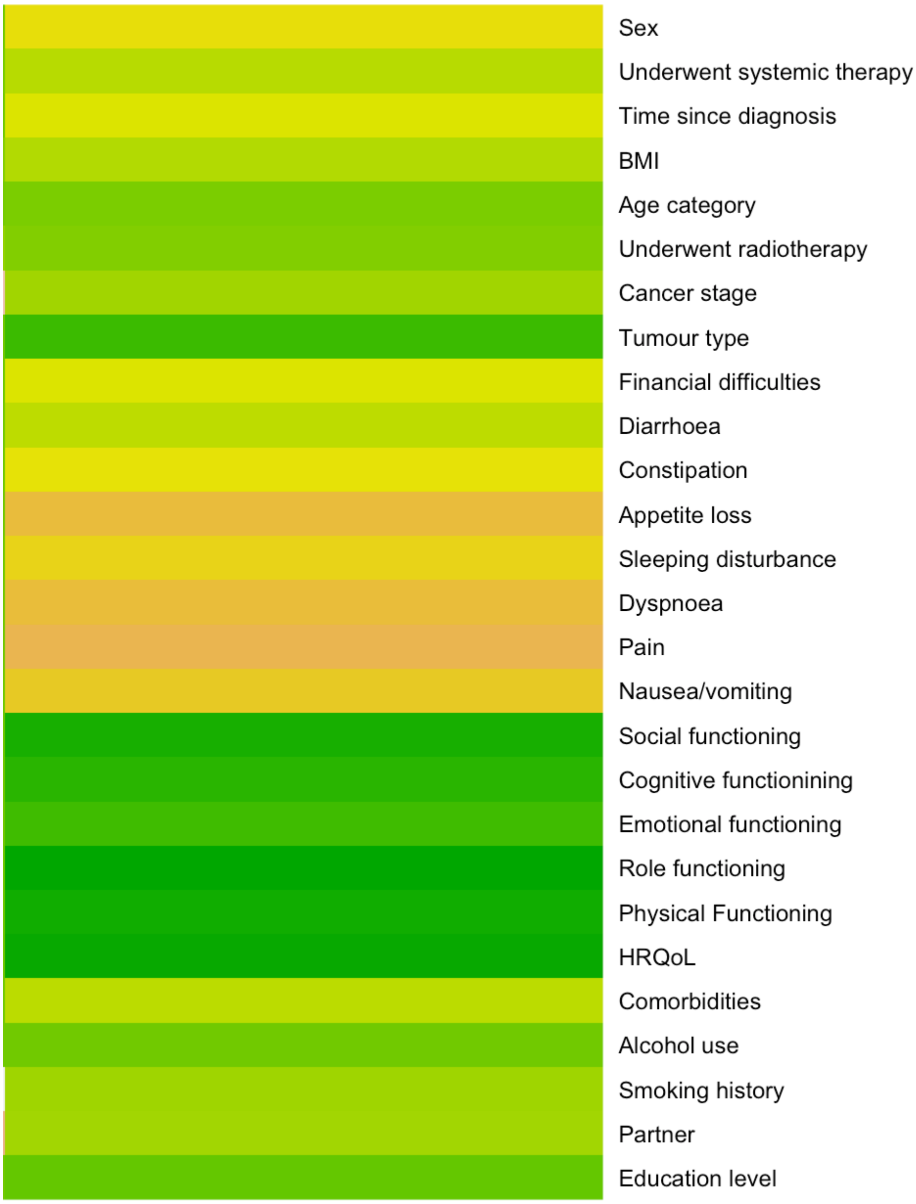
Underwent systemic treatment	101 (30.2)	101 (30.2)	50 (30.9)	50 (30.9)	13 (86.7)	13 (86.7)	110 (7.0)
Underwent radiotherapy	49 (14.7)	49 (14.7)	27 (16.7)	27 (16.7)	2 (13.3)	2 (13.3)	206 (13.1)
Has clinically-relevant fatigue	46 (13.8)	39 (11.7)	76 (46.9)	44 (27.2)	4 (26.7)	2 (13.3)	185 (11.8)
Fatigue difference (mean, SD)	0 (0.00)**	-0.25 (21.34)	0 (0.00)**	-13.13 (25.51)	0 (0.00)**	-5.93 (22.56)	-1.53 (20.31)

*Data was averaged between imputed datasets. [1]Patient characteristics from the BlaZib are only recorded for $T_{endpoint}$. [2]Reported cancer stage combines the TNM and Ann Arbor cancer stages.*Cancer stages for patients with indolent Non-Hodgkin Lymphoma cannot be determined, thus imputation was not done. **Fatigue difference at $T_{baseline}$ is measured null since no fatigue change has occurred. ***Sex for BlaZib cohort is unreported due to oversight from the author.*

Figure 2 shows a correlation plot between the predictor variables used in this study and the fatigue scores of patients using Pearson's correlation test. Correlation values are colour-coded with more positive relationships highlighted in green while negative relationships highlighted in brown.

From the figure, it is evident that most of the predictor variables have a statistically significant relationship with fatigue. Most obviously, patients with more fatigue had lower levels of role functioning (-0.67; p = <0.01), HRQoL (-0.62; p = <0.01), physical functioning (-0.58; p = <0.01), and social functioning (-0.55; p = <0.01). On the other hand, patients with more fatigue had higher levels of pain (0.58; p = <0.01), appetite loss (0.50; p = <0.01), and dyspnoea (0.50; p = <0.01). Finally, no statistically significant correlation was found between fatigue scores and smoking history (0.02; p = 0.60), cancer stage (0.03; p = 0.49), and marital status (0.03; p = 0.42).

Figure 2. correlation plot of predictor variables against fatigue.



Predicting presence of clinically relevant future fatigue

For predicting the presence of fatigue after 24-36 months, all four prediction models attained average AUC-ROC scores above 0.85 with low standard deviation. Figure 3 shows sample ROC curves for each prediction model. Table 3 shows the statistical output of each prediction model, averaged between each imputed dataset. Note that no standard deviation was reported with the reference regression model since it was negligible.

The reference regression model had the highest average value for accuracy (0.908), balanced accuracy (0.818), sensitivity (0.682), and the area under the ROC curve (0.934). This indicates that with respect to these metrics, the ML models were not able to perform better than the reference regression model. Next to this, the XGBoost model has the highest average precision (0.812; SD = 0.054) and specificity (0.978; SD = 0.009), performing better than the reference regression model on these metrics. Next to this, the SVM model also performed better than the reference regression model with respect to precision

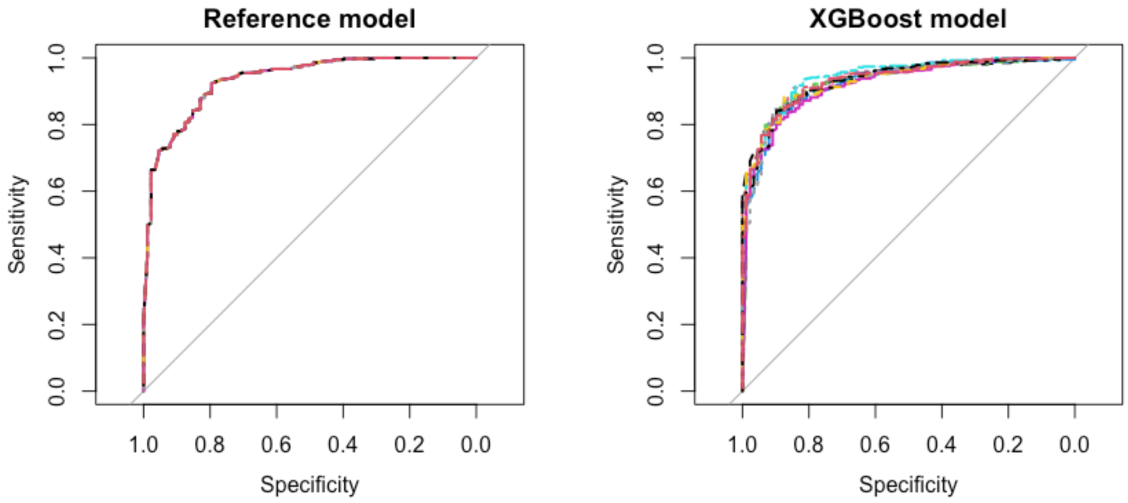
(0.775; SD = 0.036) and specificity (0.963; SD = 0.01). Notably, despite a very high average specificity, the XGBoost model has the lowest average sensitivity value (0.434; SD = 0.099). Finally, the ANN model has the lowest average value for accuracy (0.878; SD = 0.012), precision (0.651; SD = 0.048), specificity (0.927; SD = 0.017), and AUC-ROC (0.902; SD = 0.01).

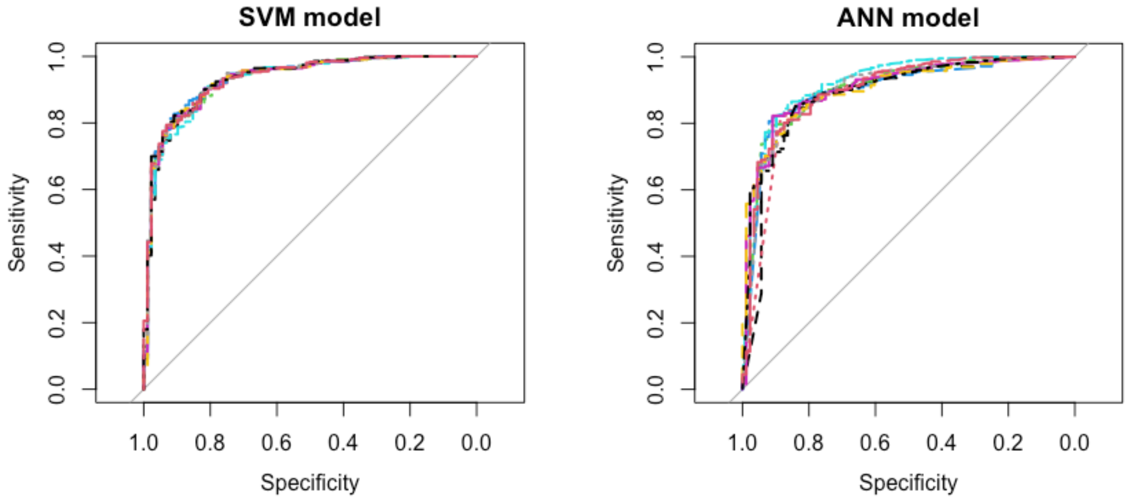
Table 3. Statistical output of each prediction model when predicting presence of future fatigue (RQ1).

Statistical Metric	*Regression (Reference)*	*XGBoost*	*Support Vector Machine*	*Artificial Neural Network*
Accuracy	0.908	0.884 (± 0.01)	0.9 (± 0.006)	0.878 (± 0.012)
Balanced Accuracy	0.818	0.706 (± 0.045)	0.78 (± 0.033)	0.787 (± 0.018)
Precision	0.759	0.812 (± 0.054)	0.775 (± 0.036)	0.651 (± 0.048)
Sensitivity	0.682	0.434 (± 0.099)	0.597 (± 0.075)	0.647 (± 0.042)
Specificity	0.955	0.978 (± 0.009)	0.963 (± 0.01)	0.927 (± 0.017)
AUC-ROC	0.934	0.933 (± 0.007)	0.932 (± 0.002)	0.902 (± 0.01)

Each metric was calculated per imputed dataset and averaged. Standard deviations are shown in brackets.

Figure 3. Receiver operating curves for each prediction model.





Each imputed dataset was plotted and overlayed onto the overall chart.

Predicting change in fatigue

With regards to predicting the extent to which fatigue changes for a patient after 24 to 36 months, all models produced a better performance in terms of the RMSE and R-squared compared to the reference regression model. Furthermore, the SVM model performed the best since it has the lowest average SMAPE (1.66; SD = 0.013) and highest average R-squared (0.058). However, all the prediction models obtained average R-squared values close to "*0*" as well as average SMAPE values above "*1.5*", indicating poor performance from the prediction models. Table 4 shows the statistical output of each prediction model when predicting the difference in fatigue scores of each patient between T_{baseline} and T_{endpoint}, averaged between each imputed dataset.

Table 4. Statistical output of each prediction model when predicting future change in fatigue (RQ2).

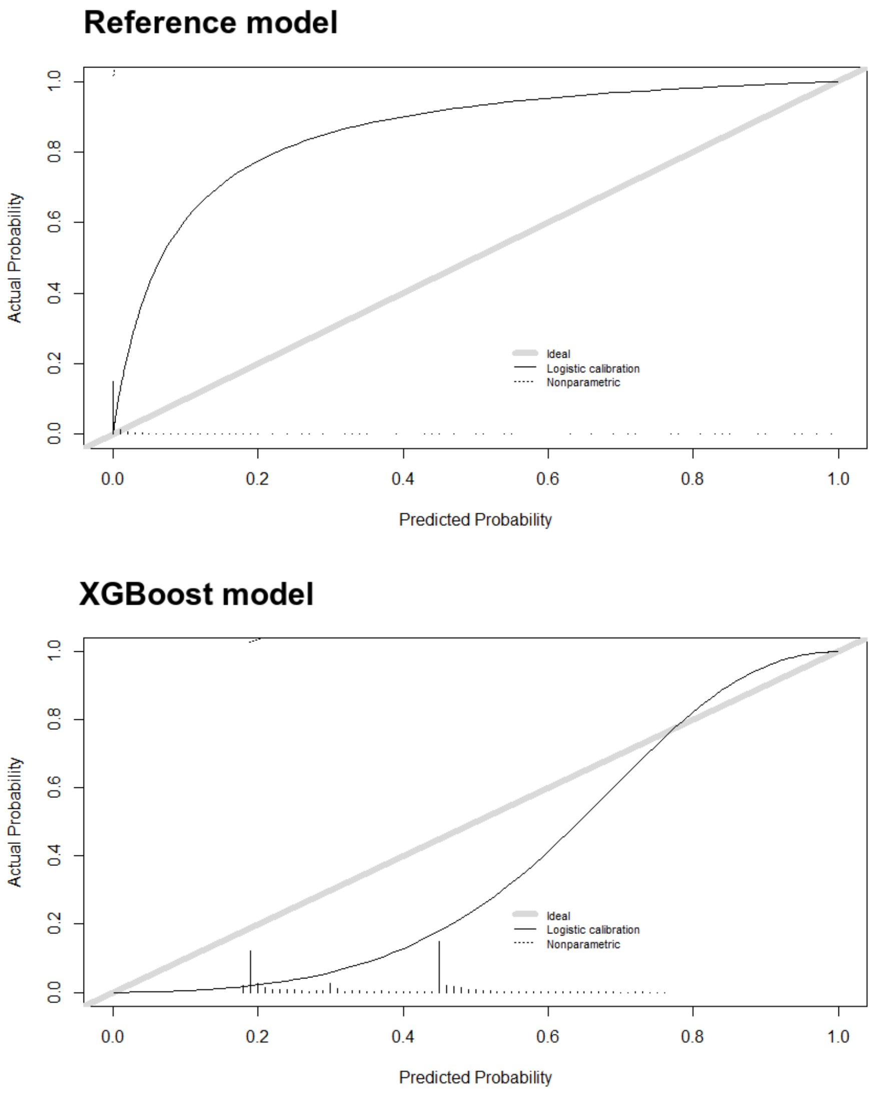
Statistical Metric	***Regression (Reference)***	***XGBoost***	***Support Vector Machine***	***Artificial Neural Network***
Root Mean Square Error	26.045 (± 0.024)	24.887 (± 0.313)	25.347 (± 0.186)	24.046 (± 0.058)
Symmetric Mean Absolute Percentage Error	1.669 (± 0.001)	1.725 (± 0.023)	1.66 (± 0.013)	1.787 (± 0.082)
R-squared	0.019	0.054 (± 0.003)	0.058	0.029 (± 0.013)

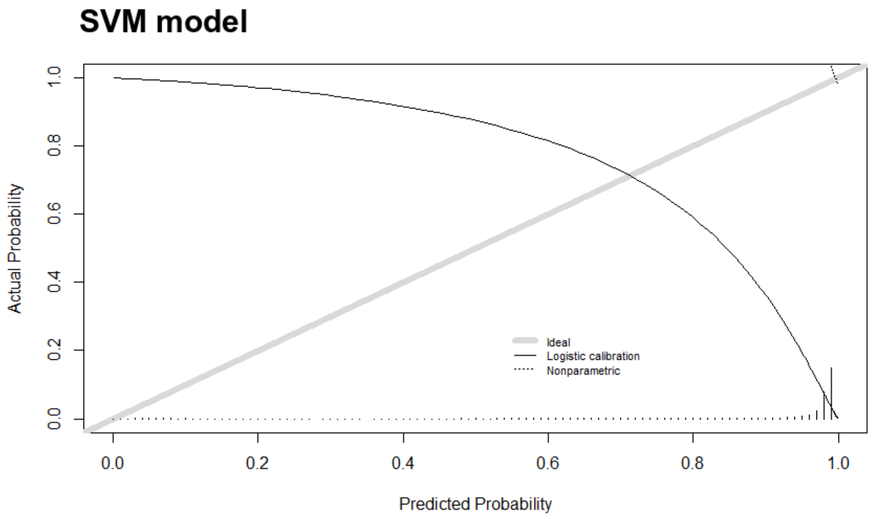
Each metric was calculated per imputed dataset and averaged. Standard deviations are shown in brackets.

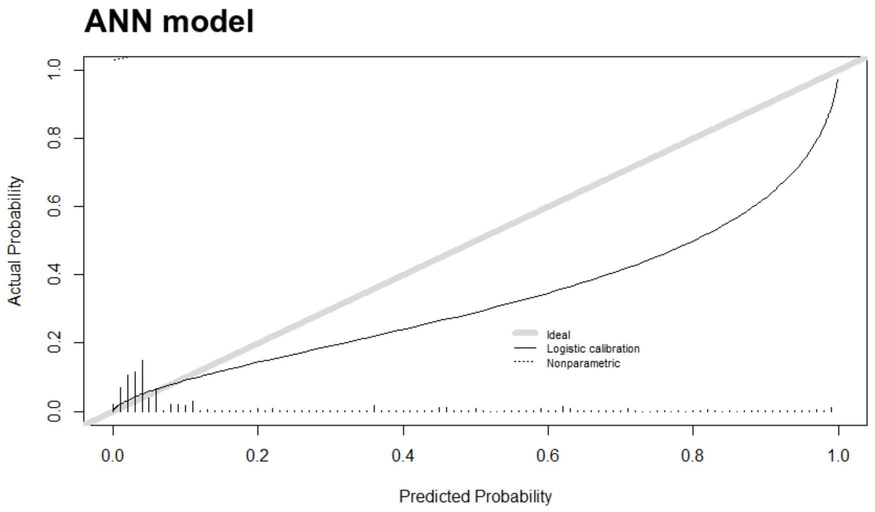
External validation

Figure 4 depicts the calibration plots from external validation using the BlaZib cohort on all the prediction models. Note that these calibration plots were done with respect to predicting the presence of clinically relevant, future fatigue since this prediction task produced favourable statistical outputs. Based on the calibration plots, the XGBoost and ANN models performed better than the reference regression model. This better performance is seen through the distance between the calibration line and the "ideal" calibration line. Next to this, while the reference regression model had a greater tendency to underpredict, i.e., produce predictions smaller than the actual probabilities, the XGBoost and ANN models had a greater tendency to overpredict, i.e., produce predictions greater than the actual probabilities.

Figure 4. Calibration plots of prediction models predicting presence clinically relevant, future fatigue after external validation.







Discussion

The aim of this study was to investigate the future fatigue of such patients given that questionnaires are started within 12 months after diagnosis and completed in between 24 and 36 months. In doing so, two prediction tasks were conducted, namely predicting the presence of clinically relevant fatigue using dichotomised outcomes and predicting the change in fatigue from T_{baseline} to T_{endpoint}. This section interprets the results per prediction task and elaborates on a general comparison between prediction models as well as with respect to clinical application.

Predicting presence of clinically relevant future fatigue

In the case of predicting the presence of clinically relevant future fatigue at 24-36 months after diagnosis, all the prediction models performed very well since they each had an AUC-ROC value of above 0.85. Moreover, because of the stable prediction output across the imputed datasets, one could argue that this shows the strength in using any of the ML algorithms [40]. Despite so, the strong performance of the reference regression model suggests that predicting the presence of future fatigue does not require an overtly complex model. Within this study, after BIC selection, 13 predictor variables were used in the final reference regression model. Amongst others, this included education level, alcohol use, HRQoL at $T_{baseline}$, cancer stage, and BMI.

With respect to the ML models when predicting the presence of clinically relevant future fatigue, the ANN model is better with respect to discriminative ability. This is because the ANN model had the smallest difference between average sensitivity and specificity values of the ML models (0.28 – ANN, 0.366 – SVM, 0.544 – XGBoost). Thus, the ANN model performed better in properly predicting patients with clinically relevant fatigue as well as those without than the other ML algorithms [56]. Within a clinical application, the difference between sensitivity and specificity values provides further insight to the AUC-ROC value since a more reliable assessment can be carried out to predict whether a patient will have future fatigue [64]. This echoes findings from previous studies that noted the strength of ANNs in terms of their discriminative ability when used as prediction models [64-65]. Therefore, within the clinical context, the ANN model might be preferred for predicting the presence of fatigue in the future within cancer patients.

Predicting change in fatigue between $T_{baseline}$ and $T_{endpoint}$

In the case of predicting the change in future fatigue of between $T_{baseline}$ and $T_{endpoint}$, all the prediction models performed very poorly. This is shown through the high average SMAPE and low average R-squared values, both of which indicate poor performance in terms of making predictions comparable with actual outcomes [61]. Following the attained results, this would imply that the prediction models are better at predicting the presence of fatigue within a cancer patient in the future but is unable to predict how much more (or less) fatigued said patient that would be. To the best of the author’s knowledge, not many studies have explored whether this difference in performance holds within the context of predicting other symptoms or HRQoL within cancer patients. Future research would, therefore, be needed to investigate and verify whether this is the case.

Comparing prediction models

Regarding the external validation process, the prediction models within this study performed somewhat well in receiving new, unseen data. During the process itself, statistical outputs obtained from the prediction models were in line with those obtained from internal validation. This suggests that the model can perform on a consistent level with new data [62]. However, the models tended to either underpredict or overpredict when fitted with data from the BlaZib cohort. This deviance implied that the models showed signs of underfitting [62]. This could be explained by how the development data consisted of fewer patients than the external validation data, implying that the model tested on data with relationships that could have been missed within the development data.

With respect to the prediction models within this study, interestingly, the ANN model did not exceedingly outperform the XGBoost and SVM models. While the ANN model showed very good discriminative ability when predicting the presence of clinically relevant, future fatigue, the model performed poorly in comparison to the XGBoost and SVM models when predicting the future change in fatigue. This contradicts the findings of various studies that supported ANN as being the preferred algorithm for prediction models on cancer patients [21, 65, 68]. Moreover, notably, ANN has been remarked as the gold standard algorithm within cancer research [24]. However, this level of performance might instead be explained by ANN’s strength as an algorithm that is able to explain complex relationships [69]. Therefore, the ANN model is more capable to simplify the relationship between the predictor variables and clinically relevant fatigue.

Within a clinical context, prediction models can be applied in the cancer domain to decide on the type of treatment and supportive care needed for a patient. Such an approach is thusly able to meet a patient's needs as well as reduce healthcare costs [70]. In the case of fatigue, this implies using a prediction model to predict the extent to which a patient will experience fatigue and whether it is clinically relevant enough to require supportive care. Moreover, this implies deciding on care that reduces the potential for persistent levels of fatigue as a side effect [71]. It is, therefore, beneficial for a clinician to have a prediction model that not only can accurately predict whether a patient is going to experience clinically relevant fatigue, but also sufficiently predict patients who will not experience such levels of fatigue. Hence, this study noted and elaborated on the influence of the difference between the sensitivity and specificity values of each prediction model towards the attained AUC-ROC values.

Aside from this, there is also a benefit to investigate the extent to which a patient's fatigue level changes over time so that care can be better adapted to a patient [72]. Due to this, it is, therefore, unfortunate that the prediction models could not satisfactorily conduct the regression task with this objective in mind. However, the difference in performance level between the classification and regression tasks might indicate the complexity behind predicting future fatigue. This point has been raised in previous studies [73-74] and could be further due to the complex nature of fatigue since it is dependent on the patient [73]. This complex nature is also applicable to other long-term symptoms, such as depression and anxiety [4]. Next to this, depending on the task at hand, ML applications can perform differently even when given the same set of parameters (i.e., dataset, predictor variables, etc.) [75]. This might be due to the complexity of the task required [76], or due to different hyperparameter requirements during optimisation [77]. Hence, this could lead to different outcomes for different prediction tasks. Therefore, future research is needed to explore the extent of the difference in prediction outcomes for ML algorithms conducting different prediction tasks. This includes predicting other long-term symptoms using similar approaches as outlined in this study. Moreover, future research is also needed to develop better models that can predict the extent of future fatigue experienced by patients.

Limitations

Like all pieces of research, this study has its limitations. First is a limitation related to the data collection process for this study. Although covering multiple cancer types is useful for the generalisability and ubiquity of ML applications in the cancer domain [20], within larger datasets that incorporate equal or more types of cancers, if the data is skewed towards a specific cancer type, this will influence the development of prediction models in favour of said cancer type [78]. Within this study, this did imply that certain predictor variables could not be used within the prediction models. For example, treatment types inapplicable to haematological cancer types could not be used as predictor variables, despite their common usage in other cancer types. This included having whether a patient underwent surgery and chemotherapy as a predictor variable. On top of this, the low number of patients included from the LYMPHOMA cohort further raised doubts regarding the possibility of producing prediction models for multiple cancer types. However, in this case, it was decided to still include the LYMPHOMA cohort to ensure generalisability remained. Therefore, future research is needed to see whether issues relating to including data from patients of multiple cancer types persists when predicting clinically relevant, future fatigue.

Next to this, the strong performance of the reference regression model in predicting the presence of future fatigue could also be a result of the predictor variable selection method. Due to the punitive nature of the BIC method, a highly favourable prediction model from this method could have led to information loss through omitting predictor variable. Appendix 4 shows the results of applying the reference regression model for predicting future fatigue but using LASSO [79] and Ridge [80] for predictor variable selection. These methods were selected since they have shown to be capable to conduct variable selection while minimising prediction error [79]. Following these results, the reference regression model still performed better than the ML models when predicting future fatigue. However, the LASSO selection method produced more favourable results than BIC selection, since it has a higher AUC-ROC value (0.941; SD = ± 0.001) and smaller difference between average sensitivity and specificity (0.113). Interestingly, the LASSO method produced a final model with more predictor variables than the BIC method. Therefore, the LASSO method could have been used within this study since it produced more favourable outputs and included more predictor variables, implying a reduced risk of information loss.

Following this, the extent of skewness within a dataset could also influence the subsequent analysis. This study mitigated this issue by conducting ten-times-repeated ten-fold cross-validation for internal validation. This method was chosen due to its computational efficiency as well as reliability for application across different prediction tasks [47, 48]. However, other internal validation methods could be used for this step. For example, Steyerberg et al. recommended the use of bootstrapping for internal validation of a logistic regression model [48]. Moreover, a randomised grid search, whereby every possible hyperparameter combination within a defined parameter space is used for tuning [53], could have been used during hyperparameter tuning. This search method has been noted to cover the parameter space well during model development [53]. Within this study, a streamlined method was preferred for developing the prediction models prior to conducting the prediction tasks. Hence, because the bootstrapping method was shown to be preferable only for classification tasks, it was not selected. In future research, this method could be used as part of investigating the future fatigue of cancer patients. However, the influence bootstrapping has on predicting the difference in fatigue with respect to computational efficiency should be considered.

Secondly, although the prediction models within this study classified future fatigue quite well, in clinical application, the entire process can still be left up to a clinician's and patient's interpretation. This is because, despite using a literature-based clinically-relevant threshold for dichotomising fatigue scores, patients and clinicians can still have a different interpretation on the level of fatigue experienced [38]. To add insight on this aspect, this study conducted additional analysis to predict the changes in of a patient's future fatigue. Yet aside from achieving poor results, the interpretation of the scores attained during clinical application could still differ depending on the individual [38, 81]. Overall, this difference in interpretation influences the type of supportive care needed. Therefore, within the domain of predicting future symptoms within cancer patients, this aspect should be considered.

Another aspect that could support the clinical interpretation of the models is through the variable selection step during model development. While the predictor variables within this study's data were collected with support from previous literature [26], this study utilised ML methods to select predictor variables for the prediction models. Although the merits of this approach have been discussed in previous literature [24, 28, 51], a literature-based or combined approach could have been conducted. In this case, predictor variables are selected based on aspects previously found to be influential for predicting fatigue. This would make the prediction models more relatable and understandable for clinicians during application since the predictors are selected based on clinical opinion. Therefore, future research could investigate either using a literature-based approach for variable selection or a mixture of ML- and literature-based variable selection when developing prediction models for future fatigue.

Third is a limitation regarding the method used for multiple imputation. This study utilised kNN and MICE for imputing EORTC QLQ-C30-related variables, and clinical- and sociodemographic-related variables respectively. Aside from the possibility of utilising other imputation methods, e.g., missForest [82], since the follow-up schedules of each cohort study were known, missing data could have been imputed on this basis. Notably, imputation using kNN uses this as one of the criteria for clustering [42]. However, this was only done on the time since diagnosis variable. This was not extended to the other cohorts because of the incoherency of the follow-up schedules within the LYMPHOMA cohort. Yet this could be mitigated by imputing the cohorts separately and only combining them during analysis. Therefore, should future research decide to use the same patient cohorts, not only different imputation methods can be considered, but also imputing the cohorts separately and investigating the effect this has on the analysis.

Conclusion

A research gap existed in the literature regarding the use of ML models to predict long-term cancer symptoms on an individual level in patients with different types of cancer. This study showed that while most ML models predict reasonably well, there is no model that performs best on all quality indicators. Moreover, no model was able to predict the future change in fatigue of patients. For use in the clinical setting, the reference regression model is preferred, because of its ability to distinguish between patients with and without fatigue 24-36 months after treatment. Other than this, this study showed that combining multiple cancer types into a single model was feasible, which is beneficial for use in a clinical setting. Future research should explore the influence of internal validation and multiple imputation methods when developing prediction models for this purpose.

References

[1] NCR Cancer Survival Dashboard [Internet]. NKR CIJFERS. 2020 [cited 2023Feb20]. Available from: https://applicatie.nkr-cijfers.nl/?fs%7Cepidemiologie_id=527&fs%7Ctumor_id=1&fs%7Coverlevingssoort_id=532&fs%7Cperiode_van_diagnose_id=601%2C600%2C599%2C598%2C597%2C596&fs%7Cjaren_na_diagnose_id=688%2C689%2C690%2C691%2C692%2C693%2C694%2C695%2C696%2C697%2C698%2C699&cs%7Ctype=line&cs%7CxAxis=jaren_na_diagnose_id&cs%7Cseries=periode_van_diagnose_id&ts%7CrowDimensions=periode_van_diagnose_id&ts%7CcolumnDimensions=jaren_na_diagnose_id&lang%7Clanguage=en.

[2] NKR Visuals [Internet]. nkr-cijfers.iknl.nl. 2021 [cited 2023 Feb 20]. Available from: https://nkr-cijfers.iknl.nl/#/viewer/62d06e64-74a2-4ee3-82af-53d772cd13a3.

[3] NKR Visuals [Internet]. nkr-cijfers.iknl.nl. 2021 [cited 2023 Mar 20]. Available from: https://nkr-cijfers.iknl.nl/#/viewer/753dc5d7-aab9-496a-bb5c-462e65a4e752.

[4] Oerlemans S, Mols F, Issa DE, Pruijt JH, Peters WG, Lybeert M, Zijlstra W, Coebergh JW, van de Poll-Franse LV. A high level of fatigue among long-term survivors of non-Hodgkin's lymphoma: results from the longitudinal population-based PROFILES registry in the south of the Netherlands. haematologica. 2013 Mar;98(3):479.

[5] de Rooij BH, Oerlemans S, van Deun K, Mols F, de Ligt KM, Husson O, Ezendam NP, Hoedjes M, van de Poll-Franse LV, Schoormans D. Symptom clusters in 1330 survivors of 7 cancer types from the PROFILES registry: A network analysis. Cancer. 2021 Dec 15;127(24):4665-74.

[6] Poort H, de Rooij BH, Uno H, Weng S, Ezendam NP, van de Poll-Franse L, Wright AA. Patterns and predictors of cancer-related fatigue in ovarian and endometrial cancers: 1-year longitudinal study. Cancer. 2020 Aug 1;126(15):3526-33.

[7] Oertelt-Prigione S, de Rooij BH, Mols F, Oerlemans S, Husson O, Schoormans D, Haanen JB, van de Poll-Franse LV. Sex-differences in symptoms and functioning in> 5000 cancer survivors: results from the PROFILES registry. European Journal of Cancer. 2021 Oct 1;156:24-34.

[8] Vickers AJ, Cronin AM, Kattan MW, Gonen M, Scardino PT, Milowsky MI, Dalbagni G, Bochner BH, International Bladder Cancer Nomogram Consortium. Clinical benefits of a multivariate prediction model for bladder cancer: a decision analytic approach. Cancer. 2009 Dec 1;115(23):5460-9.

[9] Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Computational and structural biotechnology journal. 2015 Jan 1;13:8-17.

[10] Cruz JA, Wishart DS. Applications of machine learning in cancer prediction and prognosis. Cancer informatics. 2006 Jan;2:117693510600200030.

[11] Li J, Zhou Z, Dong J, Fu Y, Li Y, Luan Z, Peng X. Predicting breast cancer 5-year survival using machine learning: A systematic review. PloS one. 2021 Apr 16;16(4):e0250370.

[12] Shi HY, Tsai JT, Chen YM, Culbertson R, Chang HT, Hou MF. Predicting two-year quality of life after breast cancer surgery using artificial neural network and linear regression models. Breast cancer research and treatment. 2012 Aug;135:221-9.

[13] Huber M, Kurz C, Leidl R. Predicting patient-reported outcomes following hip and knee replacement surgery using supervised machine learning. BMC medical informatics and decision making. 2019 Dec;19(1):1-3.

[14] Weldring T, Smith SM. Patient-Reported Outcomes (PROs) and Patient-Reported Outcome Measures (PROMs). Health services insights. 2013 Jan;6: HSI-S11093.

[15] Aaronson NK, Ahmedzai S, Bergman B, et al. The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. J Natl Cancer Inst 1993;85(5):365–76.

[16] Singer S, Wollbrück D, Wulke C, Dietz A, Klemm E, Oeken J, Meister EF, Gudziol H, Bindewald J, Schwarz R. Validation of the EORTC QLQ-C30 and EORTC QLQ-H&N35 in patients with laryngeal cancer after surgery. Head & Neck: Journal for the Sciences and Specialties of the Head and Neck. 2009 Jan;31(1):64-76.

[17] Arraras JI, Arias F, Tejedor M, Pruja E, Marcos M, Martínez E, Valerdi J. The EORTC QLQ-C30 (version 3.0) quality of life questionnaire: validation study for Spain with head and neck cancer patients. Psycho-Oncology: Journal of the Psychological, Social and Behavioral Dimensions of Cancer. 2002 May;11(3):249-56.

[18] Valdes G, Simone II CB, Chen J, Lin A, Yom SS, Pattison AJ, Carpenter CM, Solberg TD. Clinical decision support of radiotherapy treatment planning: A data-driven machine learning strategy for patient-specific dosimetric decision making. Radiotherapy and Oncology. 2017 Dec 1;125(3):392-7.

[19] Koza JR, Bennett FH, Andre D, Keane MA. Automated design of both the topology and sizing of analogue electrical circuits using genetic programming. Artificial intelligence in design'96. 1996:151-70.

[20] Azad TD, Ehresman J, Ahmed AK, Staartjes VE, Lubelski D, Stienen MN, Veeravagu A, Ratliff JK. Fostering reproducibility and generalizability in machine learning for clinical prediction modeling in spine surgery. The Spine Journal. 2021 Oct 1;21(10):1610-6.

[21] Futschik ME, Sullivan M, Reeve A, Kasabov N. Prediction of clinical behaviour and treatment for cancers. Applied bioinformatics. 2003 Jan 1;2:S53-8.

[22] Ayer T, Alagoz O, Chhatwal J, Shavlik JW, Kahn Jr CE, Burnside ES. Breast cancer risk estimation with artificial neural networks revisited: discrimination and calibration. Cancer. 2010 Jul 15;116(14):3310-21.

[23] De Laurentiis M, De Placido S, Bianco AR, Clark GM, Ravdin PM. A prognostic model that makes quantitative estimates of the probability of relapse for breast cancer patients. Clinical Cancer Research. 1999 Dec;5(12):4133-9.

[24] Exarchos KP, Goletsis Y, Fotiadis DI. Multiparametric decision support system for the prediction of oral cancer reoccurrence. IEEE Transactions on Information Technology in Biomedicine. 2011 Aug 18;16(6):1127-34.

[25] van de Poll-Franse LV, Horevoorts N, Schoormans D, Beijer S, Ezendam NP, Husson O, Oerlemans S, Schagen SB, Hageman GJ, Van Deun K, van den Hurk C. Measuring Clinical, Biological, and Behavioral Variables to Elucidate Trajectories of Patient-Reported Outcomes: The PROFILES Registry. JNCI: Journal of the National Cancer Institute. 2022 Jun;114(6):800-7.

[26] van de Poll-Franse LV, Horevoorts N, van Eenbergen M, Denollet J, Roukema JA, Aaronson NK, Vingerhoets A, Coebergh JW, de Vries J, Essink-Bot ML, Mols F, Profiles Registry Group. The Patient Reported Outcomes Following Initial treatment and Long-term Evaluation of Survivorship registry: scope, rationale, and design of infrastructure for the study of physical and psychosocial outcomes in cancer survivorship cohorts. Eur J Cancer. 2011 Sep;47(14):2188- 94. Epub 2011 May 27.

[27] Warren JL, Klabunde CN, Schrag D, Bach PB, Riley GF. Overview of the SEER-Medicare data: content, research applications, and generalizability to the United States elderly population. Medical care. 2002 Aug 1:IV3-18.

[28] Pfob A, Mehrara BJ, Nelson JA, Wilkins EG, Pusic AL, Sidey-Gibbons C. Machine learning to predict individual patient-reported outcomes at 2-year follow-up for women undergoing cancer-related mastectomy and breast reconstruction (INSPiRED-001). The Breast. 2021 Dec 1;60:111-22.

[29] Williams AM, Khan CP, Heckler CE, Barton DL, Ontko M, Geer J, Kleckner AS, Dakhil S, Mitchell J, Mustian KM, Peppone LJ. Fatigue, anxiety, and quality of life in breast cancer patients compared to non-cancer controls: a nationwide longitudinal analysis. Breast cancer research and treatment. 2021 May;187:275-85.

[30] Liu Y, Pettersson E, Schandl A, Markar S, Johar A, Lagergren P. Higher dispositional optimism predicts better health-related quality of life after oesophageal cancer surgery: a nationwide population-based longitudinal study. Annals of Surgical Oncology. 2021 Nov;28:7196-205.

[31] Bonhof CS, van de Graaf DL, Wasowicz DK, Vreugdenhil G, Mols F. Symptoms of pre-treatment anxiety are associated with the development of chronic peripheral neuropathy among colorectal cancer patients. Supportive Care in Cancer. 2022 Jun;30(6):5421-9.

[32] Zandbergen N, de Rooij BH, Vos MC, Pijnenborg JM, Boll D, Kruitwagen RF, van de Poll-Franse LV, Ezendam NP. Changes in health-related quality of life among gynaecologic cancer survivors during the two years after initial treatment: a longitudinal analysis. Acta Oncologica. 2019 May 4;58(5):790-800.

[33] Ripping TM, Kiemeney LA, Van Hoogstraten LM, Witjes JA, Aben KK. Insight into bladder cancer care: study protocol of a large nationwide prospective cohort study (BlaZIB). BMC cancer. 2020 Dec;20:1-7.

[34] Jabbar H, Khan RZ. Methods to avoid over-fitting and under-fitting in supervised machine learning (comparative study). Computer Science, Communication, and Instrumentation Devices. 2015 Dec;70:163-72.

[35] Suszko M, Sobocki J, Imieliński C. Mortality in extremely low BMI anorexia nervosa patients-implications of gastrointestinal and endocrine system dysfunction. Psychiatria Polska. 2022 Feb 27;56(1):89-100.

[36] Eman Ahmed Abd El Aty [Internet]. Wikipedia. Wikimedia Foundation; 2023 [cited 2023Mar27]. Available from: https://en.wikipedia.org/wiki/Eman_Ahmed_Abd_El_Aty.

[37] Nasteski V. An overview of the supervised machine learning methods. Horizons. b. 2017 Dec;4:51-62.

[38] Giesinger JM, Loth FL, Aaronson NK, Arraras JI, Caocci G, Efficace F, Groenvold M, van Leeuwen M, Petersen MA, Ramage J, Tomaszewski KA. Thresholds for clinical importance were established to improve interpretation of the EORTC QLQ-C30 in clinical practice and research. Journal of clinical epidemiology. 2020 Feb 1;118:1-8.

[39] King MT, Bell ML, Costa D, Butow P, Oh B. The Quality-of-Life Questionnaire Core 30 (QLQ-C30) and Functional Assessment of Cancer-General (FACT-G) differ in responsiveness, relative efficiency, and therefore required sample size. Journal of clinical epidemiology. 2014 Jan 1;67(1):100-7.

[40] Afshar HL, Jabbari N, Khalkhali HR, Esnaashari O. Prediction of breast cancer survival by machine learning methods: An application of multiple imputation. Iranian Journal of Public Health. 2021 Mar;50(3):598.

[41] What is the K-nearest neighbours' algorithm? [Internet]. [cited 2023 May 16]. Available from: https://www.ibm.com/topics/knn.

[42] Kowarik A, Templ M. Imputation with the R Package VIM. Journal of statistical software. 2016 Oct 20;74:1-6.

[43] Van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. Journal of statistical software. 2011 Dec 12;45:1-67.

[44] Bouhlila DS, Sellaouti F. Multiple imputation using chained equations for missing data in TIMSS: a case study. Large-scale Assessments in Education. 2013 Dec;1:1-33.

[45] van Buuren S, Groothuis-Oudshoorn K, Robitzsch A, Vink G, Doove L, Jolani S. Package 'mice'. Computer software. 2015 Nov 9.

[46] Schenker N, Taylor JM. Partially parametric techniques for multiple imputation. Computational statistics & data analysis. 1996 Aug 10;22(4):425-46.
[47] Wong TT. Performance evaluation of classification algorithms by k-fold and leave-one-out cross-validation. Pattern Recognition. 2015 Sep 1;48(9):2839-46.
[48] Steyerberg EW, Harrell Jr FE, Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. Journal of clinical epidemiology. 2001 Aug 1;54(8):774-81.
[49] Karri R, Chen YP, Drummond KJ. Using machine learning to predict health-related quality of life outcomes in patients with low grade glioma, meningioma, and acoustic neuroma. Plos one. 2022 May 4;17(5):e0267931.
[50] Lou SJ, Hou MF, Chang HT, Lee HH, Chiu CC, Yeh SC, Shi HY. Breast Cancer Surgery 10-Year Survival Prediction by Machine Learning: A Large Prospective Cohort Study. Biology. 2021 Dec 29;11(1):47.
[51] Neath AA, Cavanaugh JE. The Bayesian information criterion: background, derivation, and applications. Wiley Interdisciplinary Reviews: Computational Statistics. 2012 Mar;4(2):199-203.
[52] Claeskens G, Hjort NL. Model selection and model averaging. Cambridge Books. 2008.
[53] Kuhn M. The Caret package [Internet]. 2019 [cited 2023Apr18]. Available from: https://topepo.github.io/caret/train-models-by-tag.html.
[54] Shekar BH, Dagnew G. Grid search-based hyperparameter tuning and classification of microarray cancer data. In2019 second international conference on advanced computational and communication paradigms (ICACCP) 2019 Feb 25 (pp. 1-8). IEEE.
[55] Papachristou N, Puschmann D, Barnaghi P, Cooper B, Hu X, Maguire R, Apostolidis K, P. Conley Y, Hammer M, Katsaragakis S, M. Kober K. Learning from data to predict future symptoms of oncology patients. PloS one. 2018 Dec 31;13(12):e0208808.
[56] Trevethan R. Sensitivity, Specificity, and Predictive values: Foundations, Pliabilities, and Pitfalls in research and Practice. Frontiers in public health. 2017 Nov 20;5:307.
[57] Adlung L, Cohen Y, Mor U, Elinav E. Machine learning in clinical decision making. Med. 2021 Jun 11;2(6):642-65.
[58] Brodersen KH, Ong CS, Stephan KE, Buhmann JM. The balanced accuracy and its posterior distribution. In2010 20th international conference on pattern recognition 2010 Aug 23 (pp. 3121-3124). IEEE.
[59] Huber M, Kurz C, Leidl R. Predicting patient-reported outcomes following hip and knee replacement surgery using supervised machine learning. BMC medical informatics and decision making. 2019 Dec;19(1):1-3.
[60] Chai T, Draxler RR. Root mean square error (RMSE) or mean absolute error (MAE). Geoscientific model development discussions. 2014 Feb;7(1):1525-34.
[61] Chicco D, Warrens MJ, Jurman G. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Computer Science. 2021 Jul 5;7:e623.
[62] Ramspek CL, Jager KJ, Dekker FW, Zoccali C, van Diepen M. External validation of prognostic models: what, why, how, when and where? Clinical Kidney Journal. 2021 Jan;14(1):49-58.
[63] Royston P, Altman DG. External validation of a Cox prognostic model: principles and methods. BMC medical research methodology. 2013 Dec;13:1-5.
[64] Rossi LA, Melstrom LG, Fong Y, Sun V. Predicting post-discharge cancer surgery complications via telemonitoring of patient-reported outcomes and patient-generated health data. Journal of surgical oncology. 2021 Apr;123(5):1345-52.

[65] Sidey-Gibbons CJ, Sun C, Schneider A, Lu SC, Lu K, Wright A, Meyer L. Predicting 180-day mortality for women with ovarian cancer using machine learning and patient-reported outcome data. Scientific reports. 2022 Dec 8;12(1):21269.

[66] Ahmad LG, Eshlaghy AT, Poorebrahimi A, Ebrahimi M, Razavi AR. Using three machine learning techniques for predicting breast cancer recurrence. J Health Med Inform. 2013 Jan;4(124):3.

[67] Polce EM, Kunze KN, Fu MC, Garrigues GE, Forsythe B, Nicholson GP, Cole BJ, Verma NN. Development of supervised machine learning algorithms for prediction of satisfaction at 2 years following total shoulder arthroplasty. Journal of Shoulder and Elbow Surgery. 2021 Jun 1;30(6):e290-9.

[68] Chen YC, Ke WC, Chiu HW. Risk classification of cancer survival using ANN with gene expression data from multiple laboratories. Computers in biology and medicine. 2014 May 1;48:1-7.

[69] Ganggayah MD, Taib NA, Har YC, Lio P, Dhillon SK. Predicting factors for survival of breast cancer patients using machine learning techniques. BMC medical informatics and decision making. 2019 Dec;19:1-7.

[70] Mayer DK, Alfano CM. Personalized risk-stratified cancer follow-up care: its potential for healthier survivors, happier clinicians, and lower costs. JNCI: Journal of the National Cancer Institute. 2019 May 1;111(5):442-8.

[71] Street Jr RL, Voigt B. Patient participation in deciding breast cancer treatment and subsequent quality of life. Medical Decision Making. 1997 Jul;17(3):298-306.

[72] Gift AG, Stommel M, Jablonski A, Given W. A cluster of symptoms over time in patients with lung cancer. Nursing research. 2003 Nov 1;52(6):393-400.

[73] Strober LB, Arnett PA. An examination of four models predicting fatigue in multiple sclerosis. Archives of Clinical Neuropsychology. 2005 Jul 1;20(5):631-46.

[74] Du L, Du J, Yang M, Xu Q, Huang J, Tan W, Xu T, Wang L, Nie W, Zhao L. Development, and external validation of a machine learning-based prediction model for the cancer-related fatigue diagnostic screening in adult cancer patients: a cross-sectional study in China. Supportive Care in Cancer. 2023 Feb;31(2):106.

[75] Singh A, Thakur N, Sharma A. A review of supervised machine learning algorithms. In2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom) 2016 Mar 16 (pp. 1310-1315). Ieee.

[76] Yao Y, Xiao Z, Wang B, Viswanath B, Zheng H, Zhao BY. Complexity vs. performance: empirical analysis of machine learning as a service. InProceedings of the 2017 Internet Measurement Conference 2017 Nov 1 (pp. 384-397).

[77] Tran N, Schneider JG, Weber I, Qin AK. Hyper-parameter optimization in classification: To-do or not-to-do. Pattern Recognition. 2020 Jul 1;103:107245.

[78] Verma D, Bach K, Mork PJ. Application of machine learning methods on patient reported outcome measurements for predicting outcomes: a literature review. Informatics 2021 Aug 25 (Vol. 8, No. 3, p. 56). MDPI.

[79] Ranstam J, Cook JA. LASSO regression. Journal of British Surgery. 2018 Sep;105(10):1348-.

[80] McDonald GC. Ridge regression. Wiley Interdisciplinary Reviews: Computational Statistics. 2009 Jul;1(1):93-100.

[81] Giesinger JM, Kuijpers W, Young T, Tomaszewski KA, Friend E, Zabernigg A, Holzner B, Aaronson NK. Thresholds for clinical importance for four key domains of the EORTC QLQ-C30: physical functioning, emotional functioning, fatigue, and pain. Health and Quality of Life Outcomes. 2016 Dec;14(1):1-8.

[82] Stekhoven DJ, Bühlmann P. MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics. 2012 Jan 1;28(1):112-8.

Appendix 1

Variables with missing values but coded differently:

Variables Names	***Coded as…***
"education"	99
"smoking"	99
"alcoholuse"	99
"stage_annarbour"	" " or "."
"age_questionnaire_cat"	1
"age_diagnosis_cat"	1

Appendix 2

Variables used during analysis, their descriptions, and their types.

Variable Names	***Description***	***Variable Type***
"education"	Patient's highest education level	Categorical (>2 factors, ordered)
"partner"	Patient's marital history	Categorical (2 factors)
"smoking"	Patient's smoking history	Categorical (>2 factors)
"alcoholuse"	Patient's history of alcohol use	Categorical (>2 factors)
"comorbidities"	# of patient's comorbidities	Categorical (>2 factors, ordered)
"tumortype"	Patient's tumour type	Categorical (>2 factors)
"stage"	Patient's cancer stage	Categorical (>2 factors, ordered)
"radiotherapy"	Whether a patient underwent radiotherapy	Categorical (2 factors)
"Age_questionnaire_cat"	Patient's age category at the time of questionnaire	Categorical (>2 factors, ordered)
"bmi"	Patient's BMI at the time of questionnaire	Continuous
"time_diagnosis"	Time since diagnosis to the filling of the questionnaire	Continuous
"systemic"	Whether a patient underwent systemic treatment	Categorical (2 factors)
"sex"	Patient's sex	Categorical (2 factors)
"ql"	Patient's quality of life score based on EORTC-QLQ-C30	Continuous
"pf"	Patient's physical functioning score based on EORTC-QLQ-C30	Continuous
"rf"	Patient's role functioning score based on EORTC-QLQ-C30	Continuous
"ef"	Patient's emotional functioning score based on EORTC-QLQ-C30	Continuous

“cf”	Patient’s cognitive functioning score based on EORTC-QLQ-C30	Continuous
“sf”	Patient’s social functioning score based on EORTC-QLQ-C30	Continuous
“nv”	Patient’s nausea/vomit score based on EORTC-QLQ-C30	Continuous
“pa”	Patient’s pain score based on EORTC-QLQ-C30	Continuous
“dy”	Patient’s dyspnoea score based on EORTC-QLQ-C30	Continuous
“sl”	Patient’s sleep disturbance score based on EORTC-QLQ-C30	Continuous
“ap”	Patient’s appetite loss score based on EORTC-QLQ-C30	Continuous
“co”	Patient’s constipation score based on EORTC-QLQ-C30	Continuous
“di”	Patient’s diarrhoea score based on EORTC-QLQ-C30	Continuous
“fi”	Patient’s financial impact score based on EORTC-QLQ-C30	Continuous

Appendix 3

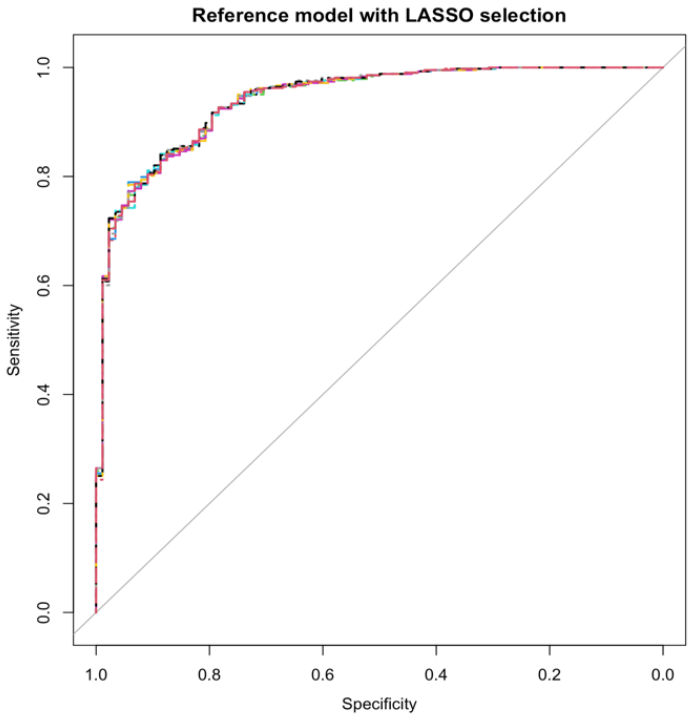
Methods to impute variables with missing values.

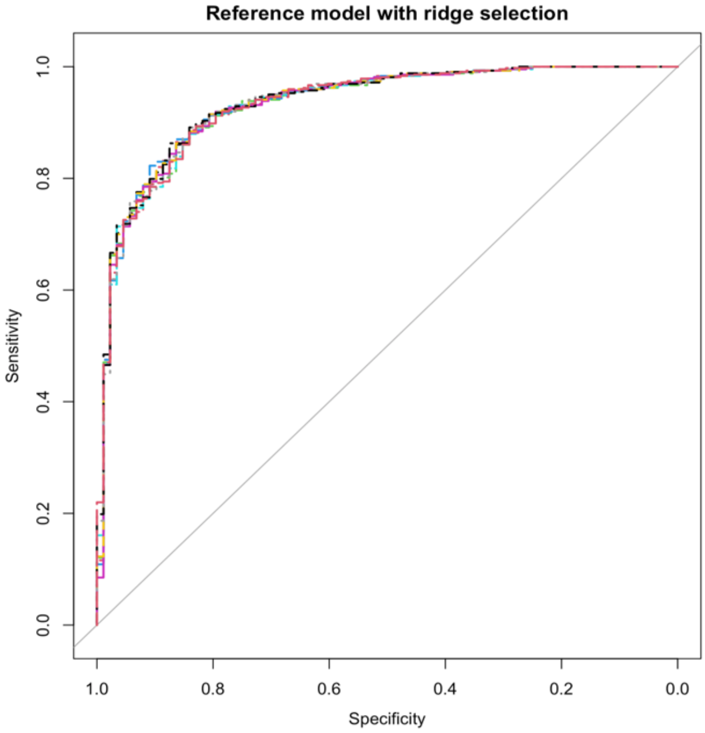
Variable Names	***Imputation Method***
“education”	Ordered polynomial regression
“partner”	Logistic regression
“smoking”	Polynomial regression
“alcoholuse”	Polynomial regression
“comorbidities”	Ordered polynomial regression
“stage”	Ordered polynomial regression
“radiotherapy”	Logistic regression
“bmi”	Predictive mean matching

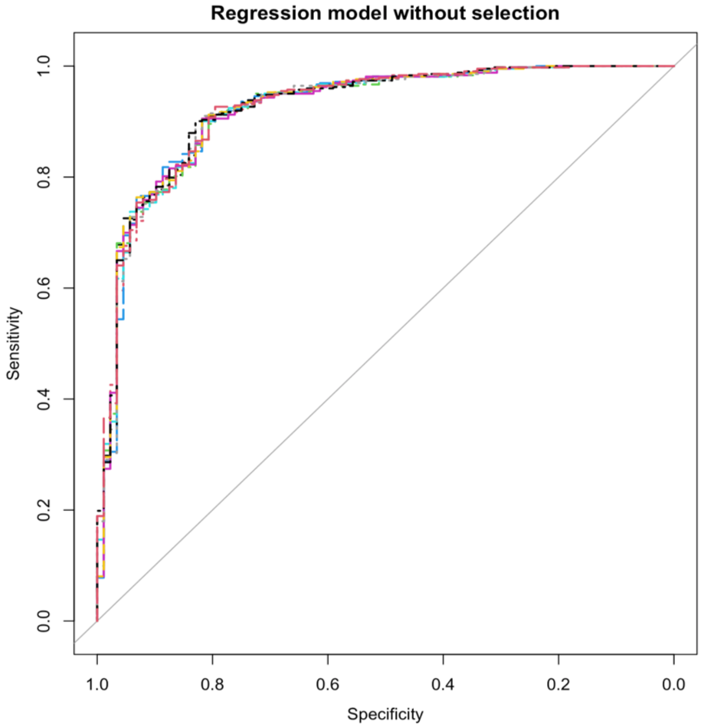
Appendix 4

The table below shows the statistical outputs of the reference regression model using the BIC, LASSO, and Ridge methods for predictor selection. Statistical outputs from a regression model without predictor variable selection is also included. The accompanying figure depicts the ROC curve of the models using the LASSO method, Ridge method, and with no selection method.

Statistical Metric	***Reference (BIC Selection)***	***Full model (Without selection)***	***LASSO Selection***	***Ridge Selection***
Accuracy	0.908	0.903 (± 0.003)	0.888 (± 0.001)	0.897 (± 0.003)
Balanced Accuracy	0.818	0.827 (± 0.01)	0.852	0.845 (± 0.006)
Precision	0.759	0.724 (± 0.021)	0.642 (± 0.002)	0.678 (± 0.014)
Sensitivity	0.682	0.711 (± 0.028)	0.795	0.766 (± 0.016)
Specificity	0.955	0.943 (± 0.008)	0.908 (± 0.001)	0.924 (± 0.006)
AUC-ROC	0.934	0.922 (± 0.001)	0.941 (± 0.001)	0.933 (± 0.001)




Reference model with ridge selection
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
Specificity
1.0
0.8
0.6
0.4
0.2
0.0


Regression model without selection
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
Specificity
1.0
0.8
0.6
0.4
0.2
0.0