[Artificial Intelligence in the Life Sciences 3 (2023) 100061](https://doi.org/10.1016/j.ailsci.2023.100061)

Contents lists available at [ScienceDirect](http://www.ScienceDirect.com/)

Artificial Intelligence in the Life Sciences

journal homepage: [www.elsevier.com/locate/ailsci](http://www.elsevier.com/locate/ailsci)

Research Article

Machine learning for longitudinal mortality risk prediction in patients with malignant neoplasm in São Paulo, Brazil

GFS Silva[a](#_bookmark0),[∗](#_bookmark3), LS Duarte [b](#_bookmark1), MM Shirassu[b](#_bookmark1), SV Peres[c](#_bookmark2), MA de Moraes [b](#_bookmark1), A Chiavegatto Filho[a](#_bookmark0)

a *School of Public Health, University of São Paulo, São Paulo, Brazil*

b *São Paulo State Health Department, Division of Non-communicable diseases, São Paulo, Brazil*

c *São Paulo State Health Department, Oncocenter Foundation of São Paulo, São Paulo, Brazil*

a r t i c l e i n f o a b s t r a c t

*Keywords:*

Machine learning Artificial intelligence Predictive model Cancer

Artificial intelligence is becoming an important diagnostic and prognostic tool in recent years, as machine learning algorithms have been shown to improve clinical decision-making. These algorithms will have some of their most important applications in developing regions with restricted data collection, but their performance under this condition is still widely unknown. We analyzed longitudinal data from São Paulo, Brazil, to develop machine learning algorithms to predict the risk of death in patients with cancer. We tested different algorithms using nine separate model structures. Considering the area under the ROC curve (AUC-ROC), we obtained values of 0.946 for the general model, 0.945 for the model with the five main cancers, 0.899 for bronchial and lung cancer, 0.947 for breast cancer, 0.866 for stomach cancer, 0.872 for colon cancer, 0.923 for rectum cancer, 0.955 for prostate cancer, and 0.917 for uterine cervix cancer. Our results indicate the potential of building models for predicting mortality risk in cancer patients in developing regions using only routinely-collected data.

# Introduction

Neoplasms are defined by abnormal tissue growth and can be classi- fied as benign or malignant. Benign (noncancerous) neoplasms are char- acterized by slow and organized spread, the presence of well-defined borders, and the absence of an invasive character at both the tissue and organ levels. Malignant (cancerous) neoplasms, on the other hand, are characterized by often rapid and disorganized growth, with poorly de- fined borders and possible invasion of adjacent tissues and organs, im- plying the possibility of metastatic cancer [[1](#_bookmark13),[2](#_bookmark14)].

According to the World Health Organization, about 9.6 million peo- ple died of cancer worldwide in 2018, of which around 70% were in middle- and low-income countries [[3]](#_bookmark15). In Brazil, according to the Na- tional Cancer Institute (INCA) [[4]](#_bookmark16), about 625,000 new cases were ex- pected in 2020, based on estimates from before the SARS-coV-2 pan- demic, and were mainly distributed among cancers of the prostate, fe- male breast, colon and rectum, trachea, bronchus and lung, and stom- ach. As for the total number of deaths, according to the Brazilian Mor- tality Information System data (SIM), 224,829 deaths were caused by malignant neoplasms in 2020, and the most frequent were trachea, bronchus, and lung (28,516 deaths), breast (18,032 deaths), prostate (15,841 deaths), stomach (13,850 deaths), and colon (12,422 deaths)

[[5]](#_bookmark17).

∗ Corresponding author.

*E-mail address:* [gabriel8.silva@usp.br](mailto:gabriel8.silva@usp.br) (G. Silva).

Artificial Intelligence (AI) has become an important tool in the field of medicine. Machine learning algorithms are capable to identify pat- terns and trends from data that may not be readily apparent to the human eye. This allows medical professionals to make more accurate predictions about patient diagnosis and prognosis and make informed decisions about their treatment. Machine learning in healthcare has the potential to greatly improve patient outcomes and make the healthcare system more eﬃcient.

Given the growing scenario of cancer cases in Brazil and around the world [[4](#_bookmark16),[6](#_bookmark18)], it is increasingly important to improve prognostic decisions for cancer patients [[7]](#_bookmark19). The aim of this work is to develop machine learning algorithms to predict the risk of death in cancer patients in order to provide inputs for their clinical management.

# Material and methods

* 1. *Dataset description*

We analyzed data collected from the Hospital Cancer Registry (RHC) of the Oncocenter Foundation of São Paulo (FOSP/SP) [[8]](#_bookmark20), a public reg- istry that monitors patients treated in the state of São Paulo since the year 2000. RHC has information on the cancer diagnosis, treatment, metastases, recurrences, age and sex of the patients, and data on the

<https://doi.org/10.1016/j.ailsci.2023.100061>

Received 21 December 2022; Received in revised form 31 January 2023; Accepted 1 February 2023

Available online 3 February 2023

2667-3185/© 2023 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>)

health facilities/organizations where the consultations were performed. The dataset includes a total 99 variables and 1085,380 patients from 2000 through September 2022.

All variables collected after the cancer diagnosis for each patient were removed. The algorithms were trained with twelve variables: sex, age, days between first physician visit and diagnosis, clinical stage of cancer, category of medical service, previous diagnosis, type of diag- nosis, topography group, health region of residence [[9]](#_bookmark21), morphology, health institution habilitation, and health region of diagnosis. There are three levels to the variable category medical service: 1) private care,

2) public care, 3) private care. The variable previous diagnosis presents binary information, 1 for patients who started longitudinal follow-up with a previous cancer diagnosis and 0 for patients without previous diagnosis. The variable type of diagnosis presents four categories: 1) clinical examination, 2) non-microscopic auxiliary resources, 3) micro- scopic confirmation and 4) no information. The variables cancer topog- raphy and cancer morphology are categorized with ICD-10 and ICD-O, respectively. The variables related to health region have seventeen dis- tinct values, referring to the seventeen health regions in the state of São Paulo, Brazil. The variable health institution habilitation has fifteen categories: 1) High Complexity Oncology Care Unit (UNACON), 2) UNA- CON with Radiotherapy Service, 3) UNACON with Hematology Service,

5) Exclusive UNACON for Pediatric Oncology, 6) High Complexity On- cology Care Center (CACON), 7) CACON with Pediatric Oncology Ser- vice, 8) General Hospital with Oncological Surgery, 9) UNACON with Radiotherapy and Hematology Services, 10) UNACON with Radiother- apy, Hematology and Pediatric Oncology Services, 12) UNACON with Hematology and Pediatric Oncology Services, 13) Volunteer, 14) Inac- tive, 15) Exclusive UNACON for Pediatric Oncology with Radiotherapy Service. The full description of the dataset and its variables can be found in Supplementary Appendix B (Table B1).

Only patients with diagnoses from 2014 to 2017 were included, in order to avoid longer clinical effects after the diagnosis. Although the dataset included a small portion of population from other Brazil- ian states, we limited the algorithm development to residents of the state of São Paulo (93% of total patients). We included only patients with malignant neoplasms and excluded cases of non-melanoma of the skin as they had a low mortality rate. We analyzed adult patients re- gardless of sex. The final sample was composed of a total of 29,194 patients.

* 1. *Outcome definition*

The original dataset contains four categories regarding the last avail- able information about the patient: 1) alive without cancer, 2) alive with cancer, 3) death from cancer, and 4) death without further infor- mation. Our outcome of interest was patients with a confirmed cancer death between 12 and 24 months after the date of diagnosis. For the negative outcome, we included patients 1) alive without cancer or 2) alive with cancer between 12 and 36 months after the date of diag- nosis. Patients were removed if categorized as 4) death without other information.

* 1. *Model design*

Considering the distinct cancer types, we tested different models to assess whether changing the strategy increased model performance. We first developed a general model for all cancer types. We then developed a model for the top five causes of cancer mortality (bronchus and lung, breast, stomach, colon, and rectum). We also trained specific models for the five most frequent causes and added two other models based on its growing importance for health vigilance: prostate cancer and cervix uteri cancer (in both cases the sex variable was not used as a predic- tor). We evaluated the models independently, without sharing any in- formation during algorithm training. A summary of the model design is provided in Supplementary Appendix A (Figure A1).

* 1. *Machine learning techniques*

For quantitative variables, we performed normalization using the z- score (separately in training and test). For all qualitative variables, we separated each category using one-hot encoding. The variable type of di- agnosis presented categorized missing (value 9) for twenty-four patients. We considered this a new category for the one hot encoding procedure. We also removed 75 patients due to the lack of any information in two variables: cancer stage (two patients) and difference in days between first medical appointment dates and diagnosis (73 patients).

We tested the predictive performance of six different machine learn- ing algorithms: catboost [[10]](#_bookmark22), xgboost [[11]](#_bookmark23), lightgbm [[12]](#_bookmark24), gradient boosting classifier, random forest, logistic regression. For catboost, xg- boost and lightgbm, we used their own Python packages. For the other algorithms, we used the scikit-learn library [[13]](#_bookmark25).

We used 10-fold cross-validation to select the hyperparameters in the training set with Hyperopt [[14]](#_bookmark27), which applies a Bayesian strategy for optimization, and RandomSearch. In the case of high class imbalance (minority class representing under 25% of total outcomes), we applied the Synthetic Minority Oversampling Technique (SMOTE). Also in the training set, we applied the BORUTA [[15]](#_bookmark28) method for variable selection. We then selected the best performing models from the training set (70% of the data) to evaluate their performance in the test set (30%).

The complete structure of the datasets is presented in [Fig. 1](#_bookmark4).

The predictive performance of the models was evaluated on the test set using metrics such as area under the ROC curve (AUC-ROC), area un- der the precision-recall curve (AUC-PR), precision, recall, positive pre- dicted value, negative predicted value, and F1-score. We also evaluated the performance of the algorithms in the 20% highest risk patients (20% k-tops), with metrics such as true positive, false positive, precision and recall. Finally, the interpretation and evaluation of the contribution of each variable to the outcome was obtained by calculating the Shapley values [[16](#_bookmark30),[17](#_bookmark31),[18](#_bookmark32)] for the test set. We followed the guidelines of the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) [[19]](#_bookmark24).

# Results and discussion

* 1. *Descriptive data analysis*

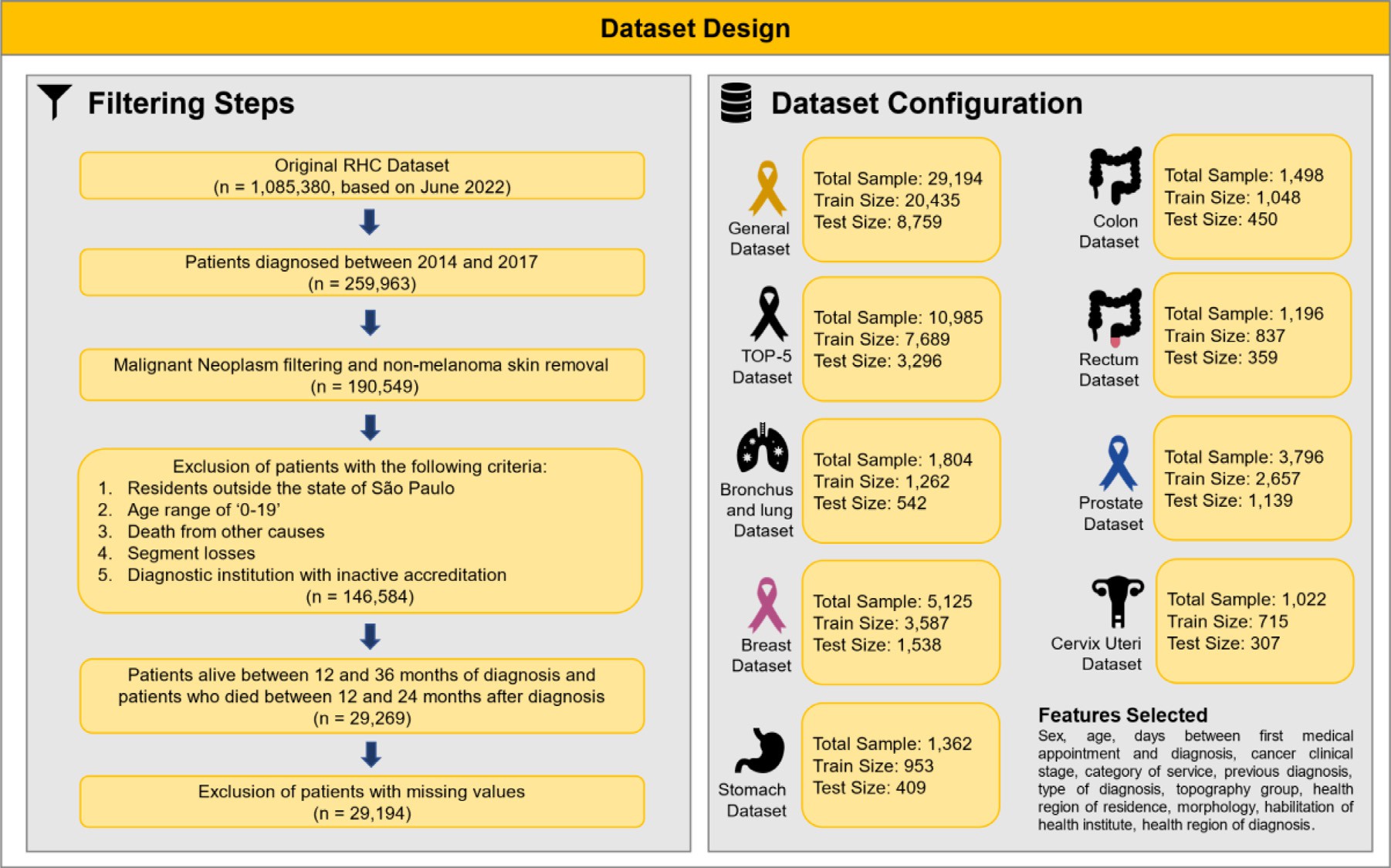
After data preprocessing, a total 29,194 patients were included in the study, most of whom were female (51.0%). A total of 27.7% of pa- tients were between 60 and 69 years old and 23.5% seventy or more. Regarding the clinical classification of the cancer, there was a relative balance between stages I (20.9%), II (20.3%), III (20.1%), IV (26.0%). The other categories accounted for about 12.6% of the total number of patients.

Regarding the category of health services, 72.7% of patients were diagnosed in public institutions, 26.6% in private services and 0.7% in individual private services. Of the total patients, 43.6% died between 12 and 24 months after the date of the malignant cancer diagnosis and 56.4% remained alive. The characteristics of the patients were similar in the training and testing data, as well as in the general dataset ([Table 1](#_bookmark5)). Additional summaries of cancer morphology and topography are pro- vided in Tables B2 and B3 in Supplementary Appendix B.

* 1. *Model data structure*

We developed a total of 42 machine learning algorithms considering nine root structures: 1) general model, 2) top 5 cause of death model,

3) bronchus and lung cancer model, 4) breast cancer model, 5) stom- ach cancer model, 6) colon cancer model, 7) rectum cancer model, 8) prostate cancer model, and 9) cervix uteri cancer model. [Table 2](#_bookmark6) pro- vides a descriptive summary of each model considering the total number of patients, total number of deaths and nondeaths, total mortality, and



**Fig. 1.** Dataset filtering and configurations for predictive models’ development.

**Table 1**

Descriptive summary of full, train and test datasets.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Full Dataset | Death | Non-death | Train | Test |
| **Sex** |  |  |  |  |  |
| Male | 14,313 (49.0%) | 7027 (55.2%) | 7286 (44.2%) | 9972 (48.8%) | 4341 (49.6%) |
| Female | 14,881 (51.0%) | 5697 (44.8%) | 9184 (55.8%) | 10,463 (51.2%) | 4418 (50.4%) |
| **Age** |  |  |  |  |  |
| 20–29 | 929 (3.2%) | 263 (2.1%) | 666 (4.0%) | 652 (3.2%) | 277 (3.2%) |
| 30–39 | 2176 (7.5%) | 636 (5.0%) | 1540 (9.4%) | 1534 (7.5%) | 642 (7.3%) |
| 40–49 | 3862 (13.2%) | 1453 (11.4%) | 2409 (14.6%) | 2714 (13.3%) | 1148 (13.1%) |
| 50–59 | 7270 (24.9%) | 3234 (25.4%) | 4036 (24.5%) | 5077 (24.8%) | 2193 (25.0%) |
| 60–69 | 8093 (27.7%) | 3690 (29.0%) | 4403 (26.7%) | 5658 (27.7%) | 2435 (27.8%) |
| 70+ | 6864 (23.5%) | 3448 (27.1%) | 3416 (20.7%) | 4800 (23.5%) | 2064 (23.6%) |
| **Clinical Stage** |  |  |  |  |  |
| I | 6107 (20.9%) | 570 (4.5%) | 5537 (33.6%) | 4292 (21.0%) | 1815 (20.7%) |
| II | 5917 (20.3%) | 1371 (10.8%) | 4546 (27.6%) | 4119 (20.2%) | 1798 (20.5%) |
| III | 5876 (20.1%) | 2843 (22.3%) | 3033 (18.4%) | 4088 (20.0%) | 1788 (20.4%) |
| IV | 7602 (26.0%) | 6090 (47.9%) | 1544 (9.4%) | 5361 (26.2%) | 2241 (25.6%) |
| X | 712 (2.4%) | 414 (3.3%) | 1512 (9.2%) | 482 (2.4%) | 230 (2.6%) |
| Y | 2980 (10.2%) | 1436 (11.3%) | 298 (1.8%) | 2093 (10.2%) | 887 (10.1%) |
| **Service Category** |  |  |  |  |  |
| Public | 21,224 (72.7%) | 11,865 (93.2%) | 9359 (56.8%) | 14,840 (72.6%) | 6384 (72.9%) |
| Private | 7755 (26.6%) | 803 (6.3%) | 6952 (42.2%) | 5448 (26.7%) | 2307 (26.3%) |
| Particular | 215 (0.7%) | 56 (0.4%) | 159 (1.0%) | 147 (0.7%) | 68 (0.8%) |
| **utcome** |  |  |  |  |  |
| Death | 12,724 (43.6%) | – | – | 8906 (43.6%) | 3818 (43.6%) |
| Non-death | 16,470 (56.4%) | – | – | 11,529 (56.4%) | 4941 (56.4%) |

the number of patients in the training and test groups. There were no- table imbalances according to the different models, with 29,194 patients in the general model and 1022 in the cervix uteri cancer model, which may have affected the predictive performance of the algorithms.

* 1. *Algorithms performance*

The catboost algorithm presented the best performance on all models except stomach, where Gradient Boosting performed better in terms of AUC-ROC. [Fig. 2](#_bookmark7) presents the performance of the best prediction algo- rithms for each of the models considering the AUC-ROC test set criterion

(a) and AUC-PR (b). We found good discrimination performance for the

nine models. All models presented AUC-ROC values of at least 0.871 and six of them reached values above 0.900. The general model presented the best overall performance, with AUC-ROC of 0.946 and AUC-PR of 0.932. The model for the five main causes of death (top-5) presented an AUC-ROC of 0.945 and an AUC-PR of 0.937.

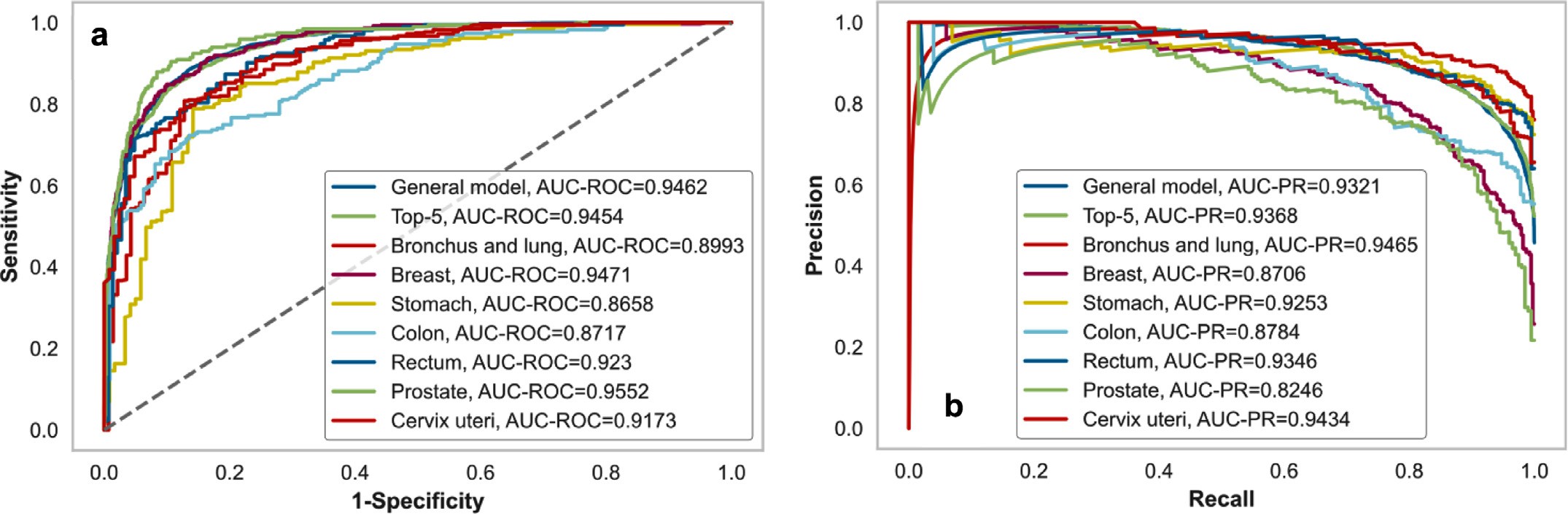
In the general model, we tested combinations of hyperparame- ter optimization (Hyperopt and RandomSearch) and variable selection (BORUTA). However, the best predictive performance was achieved for the raw model, with AUC-ROC of 0.946, recall 0.855, specificity 0.889, precision 0.857, F1-score 0.856, and AUC-PR 0.932 ([Table 3](#_bookmark8)). Although changes in algorithm settings improved at least one scoring metric (i.e., the general model with RandomSearch had a recall of 0.858), the raw

**Table 2**

Description of the eight models developed for prediction of cancer mortality.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ID | Model | Variables | Total cases | Non-death | Death | Mortality rate | Train Size (70%) | Test Size (30%) |
| 1 | General | sex, age, medsvtodiag, cancerstage, servicecat, prevdiag, diagbase, topogroup, rras, morpho, habilit, rrasofserv | 29,194 | 16,470 | 12,724 | 43.6% | 20,435 | 8.759 |
| 2 | Top-5 cause of  death | sex, age, medsvtodiag, cancerstage, servicecat, prevdiag,  diagbase, topogroup, rras, morpho, habilit, rrasofserv | 10,985 | 5937 | 5048 | 46.0% | 7689 | 3.296 |
| 3 | Bronchus and  lung | sex, age, medsvtodiag, cancerstage, servicecat, prevdiag,  diagbase, rras, morpho, habilit, rrasofserv | 1804 | 468 | 1336 | 74.1% | 1262 | 542 |
| 4 | Breast | sex, age, medsvtodiag, cancerstage, servicecat, prevdiag,  diagbase, rras, morpho, habilit, rrasofserv | 5125 | 3845 | 1280 | 25.0% | 3587 | 1.538 |
| 5 | Stomach | sex, age, medsvtodiag, cancerstage, servicecat, prevdiag,  diagbase, rras, morpho, habilit, rrasofserv | 1362 | 401 | 961 | 70.6% | 953 | 409 |
| 6 | Colon | sex, age, medsvtodiag, cancerstage, servicecat, prevdiag,  diagbase, rras, morpho, habilit, rrasofserv | 1498 | 740 | 758 | 50.6% | 1048 | 450 |
| 7 | Rectum | sex, age, medsvtodiag, cancerstage, servicecat, prevdiag,  diagbase, rras, morpho, habilit, rrasofserv | 1196 | 483 | 713 | 59.6% | 837 | 359 |
| 8 | Prostate | age, medsvtodiag, cancerstage, servicecat, prevdiag, diagbase,  rras, morpho, habilit, rrasofserv | 3796 | 3232 | 664 | 17.5% | 2657 | 1.139 |
| 9 | Cervix uteri | age, medsvtodiag, cancerstage, servicecat, prevdiag, diagbase,  rras, morpho, habilit, rrasofserv | 1022 | 416 | 609 | 59.6% | 715 | 307 |

**medsvtodiag**: difference in days between first medical appointment dates and diagnosis, **cancerstage**: cancer clinical stage, **servicecat**: category of service, **prevdiag**: previous diagnosis, **diagbase**: type of diagnosis, **topogroup**: cancer topography, **rras**: regional net of healthcare (residence), **morpho**: cancer morphology, **habilit**: qualification of the health establishment, **rrasofserv**: regional net of healthcare (service).



**Fig. 2.** Predictive performance of best algorithm for each model regarding AUC-ROC (a) and AUC-PR (b).

**Table 3**

Predictive performance of best algorithm for each model.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ID | Model | Best Algorithm | Hypermeter Tunning | Feature Selection | Resample | Accuracy | AUC-ROC | Recall | Specificity | Prec. | F1 | AUC-PR |
| 1 | General | CatBoost Classifier | None | None | None | 0.8743 | 0.9462 | 0.8549 | 0.8893 | 0.8565 | 0.8557 | 0.9321 |
| 2 | Top-5 cause  of death | CatBoost  Classifier | None | None | None | 0.8686 | 0.9454 | 0.8581 | 0.8776 | 0.8564 | 0.8572 | 0.9368 |
| 3 | Bronchus  and lung | CatBoost  Classifier | None | None | SMOTE | 0.8561 | 0.8993 | 0.9152 | 0.6879 | 0.8929 | 0.9039 | 0.9465 |
| 4 | Breast | CatBoost  Classifier | None | None | None | 0.8973 | 0.9471 | 0.7214 | 0.9558 | 0.8445 | 0.7781 | 0.8706 |
| 5 | Stomach | Gradient  Boosting | RandomSearch | None | None | 0.8093 | 0.8658 | 0.9343 | 0.5083 | 0.8207 | 0.8738 | 0.9253 |
| 6 | Colon | CatBoost  Classifier | None | None | None | 0.7578 | 0.8717 | 0.7763 | 0.7387 | 0.7532 | 0.7646 | 0.8784 |
| 7 | Rectum | CatBoost  Classifier | Hyperopt | None | None | 0.8412 | 0.9230 | 0.9159 | 0.7310 | 0.8340 | 0.8731 | 0.9346 |
| 8 | Prostate | CatBoost  Classifier | None | None | None | 0.9210 | 0.9552 | 0.7487 | 0.9574 | 0.7884 | 0.7680 | 0.8246 |
| 9 | Cervix uteri | CatBoost  Classifier | Hyperopt | None | None | 0.8306 | 0.9173 | 0.9235 | 0.6935 | 0.8164 | 0.8667 | 0.9434 |

**Table 4**

Predictive performance of best algorithm for each model based on 20% individuals with the highest risk of death.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ID | Model | Total Patients | Real Positive | Positive Prediction | True Positive | False Positive | Precision | Recall |
| 1 | General | 1749 | 1703 | 1749 | 1703 | 46 | 0.9737 | 1.0000 |
| 2 | Top-5 cause of  death | 659 | 645 | 659 | 645 | 14 | 0.9788 | 1.0000 |
| 3 | Bronchus and  lung | 109 | 107 | 109 | 107 | 2 | 0.9817 | 1.0000 |
| 4 | Breast | 308 | 263 | 308 | 263 | 45 | 0.8539 | 1.0000 |
| 5 | Stomach | 82 | 78 | 82 | 78 | 4 | 0.9512 | 1.0000 |
| 6 | Colon | 90 | 88 | 90 | 88 | 2 | 0.9778 | 1.0000 |
| 7 | Rectum | 72 | 70 | 72 | 70 | 2 | 0.9722 | 1.0000 |
| 8 | Prostate | 228 | 166 | 189 | 149 | 40 | 0.7884 | 0.8976 |
| 9 | Cervix uteri | 62 | 60 | 62 | 60 | 2 | 0.9677 | 1.0000 |

model was the one that presented the best AUC-ROC. When BORUTA was used, there was a significant reduction in the number of predictors (497 to 93) without a large loss in predictive performance (AUC-ROC of 0.946 for the raw model versus 0.945 for the model with BORUTA and without hyperparameter optimization). A similar pattern to the general model was observed for the top 5 causes of death model. Complete re- sults for all training strategies can be found in Supplementary Appendix B (Table B4). The hyperparameters of each model are also available in Supplementary Appendix B (Tables B5, B6, B7, B8, B9, B10, B11, B12, B13).

We also evaluated the performance of the algorithms in the top 20% (20% k-tops) of patients with the highest mortality risk ([Table 4](#_bookmark9)). The general model had 1749 patients in the group, of whom 1703 died, giv- ing the algorithm a precision of 97,37% and a recall of 100% in this high-risk group. For the top 5 causes of death model, 659 individuals were in the 20% highest risk patients, of which 645 died, resulting in a precision of 97.88% and a recall of 100%.

* 1. *Model interpretation*

In order to interpret the decision-making process of the algorithms, we calculated the Shapley values. In the general model ([Fig. 3](#_bookmark10)), the can- cer stage during the first diagnosis was the most important predictor. Stage I patients were more likely to be classified negatively (non-death), whereas stage IV patients were more significant for the positive out- come (death). The variable on the category of service provided was also important for the outcome. Category 2 (public service) increased mor- tality prediction, whereas category 1 (private service) showed a greater propensity for patient survival. The other main predictive variables refer to the topography of cancer, regional net of healthcare service (rrasof- serv) and regional net of healthcare service (rras). The plots of Shapley values for the other models can be found in Supplementary Appendix A (Figures A2, A4, A6, A8, A10, A12, A14, A16).

We also randomly selected three patients (high risk, medium risk, and low risk) to highlight the individual interpretation of results ([Fig. 4](#_bookmark12)). The first (a) was a true positive (risk of 0.972) with the expected Shap- ley value was 2.92. The variables that contributed for the prediction

of positive outcome were public care service (servicecat\_2 = 1), cancer

stage different from I (cancerstage\_*I*= 0), cancer topography ICD C-34

(rrasofserv\_RRAS01 = 1) and the qualification of the healthcare institu- (malignant neoplasm of bronchus and lung), regional net of service 01

tion (code 12, UNACON with Hematology and Pediatric Oncology Ser- vices). A second patient (b) classified as a false negative was selected. The total risk score was 0.4782, which led the algorithm to classify the patient incorrectly as alive during the period. We observed that there was balance in the aggregate of the contribution of the predictors, high- lighting the importance of cancer stage IV to increase Shapley value and the non-public health service to decrease it. For patient c, a true negative classified as low risk, the most important characteristic to a low expected Shapley value were cancer stage I and non-public health

service. Visualizations of the individual Shapley values for the other models are available in Supplement A (figures A3, A5, A7, A9, A11, A13, A15, A17).

* 1. *General vs specific models*

To understand the best strategy regarding the types of models (gen- eral or specific for each cancer), we performed a comparison between the performance of the general algorithm for all cancers versus the spe- cific algorithms for bronchus and lung, breast, stomach, colon, rectum, prostate, bronchus and lung, colon, and uterine cervix ([Table 5](#_bookmark11)). Based on the AUC-ROC, the general model performed better in bronchus and long, stomach and colon. For breast, rectum, prostate, and cervix uteri the model performed better for the specific case. For bronchus and lung cancer, the area under the curve increased from 0.899 (model specific) to 0.927 (general model) and the precision from 0.893 to 0.907. For stomach cancer, there was an increase in AUC-ROC (0.866 to 0.926), precision (0.821 to 0.924). This scenario was repeated for colon cancer (AUC-ROC 0.753 to 0.848).

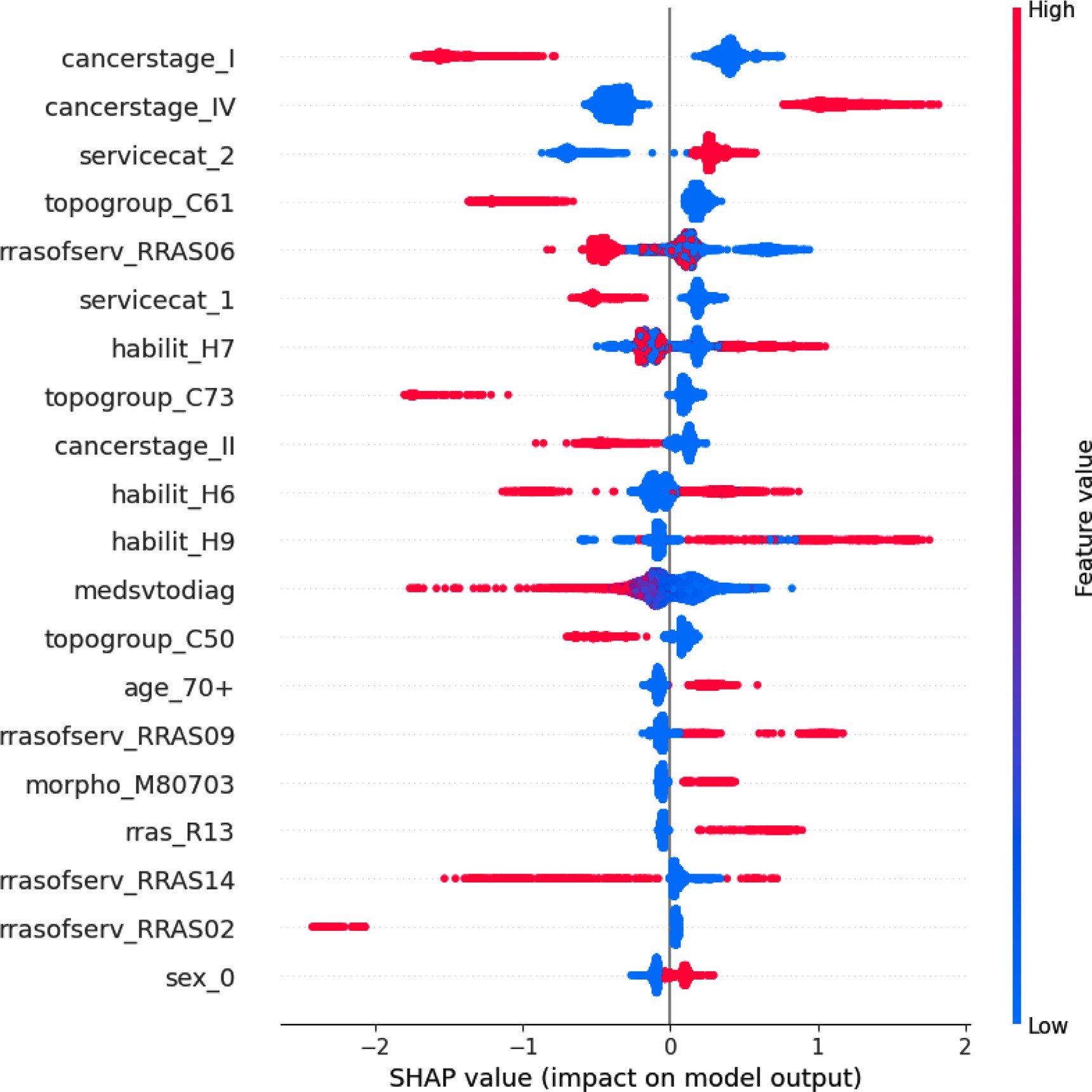
* 1. *Discussion*

We found that all models achieved an AUC-ROC higher than 0.86 to predict cancer mortality using only routinely-collected data. Our results also indicated that a general algorithm, that included all cancer mortal- ity, performed in most cases better than cancer-specific algorithms.

Information about mortality risk after cancer diagnosis can be an important input to support clinical decisions. These algorithms can be integrated into mobile devices, electronic medical records, or on- line resources, to help doctors in making more informed decisions about treatment options and to allocate healthcare resources more effectively.

We obtained a high predictive performance without the use of omics or image data, which is a promising result in the field of oncology espe- cially in low-income regions. We were able to develop an approach to compare the use of a general model with a model for the main causes of death, and with models specific to each type of cancer. Most of the studies developed in the literature are specific for a given type of cancer, due to the selected data sets, using either from image data or through structured data [[20](#_bookmark26),[21](#_bookmark29)]. The use of data from a cancer registry allowed for achieving consistent results in all proposed models, while at the same time providing a real-world dataset, with recent cases and dif- ferent types of cancer.

The study has a few limitations. First, the algorithms were developed with data from São Paulo, Brazil, so there should be caution in transfer- ring its conclusions to other contexts. Second, considering the filtered sample, we had 1391 follow-up losses that could have disproportionally altered the results. Third, we excluded patients younger than 20 years due to the presence of different biological mechanisms that lead to can- cer deaths for this group, so the results refer only to adult patients.



**Fig. 3.** Top twenty predictors of risk of death from cancer 12 to 24 months after diagnosis. General Model, with Catboost Classifier. **cancerstage\_I**: cancer stage I, **can- cerstage\_IV**: cancer stage IV, **servicecat\_2**: public care service, **topogroup\_C61**: cancer topography ICD C-61 (malignant neoplasm of prostate), **rrasofserv\_RRAS06**: regional net of healthcare (service) 06, **servicecat\_1**: private care service, **habilit\_H7**: qualification H7 CACON with Pediatric Oncology Service, **topogroup\_C73**: can- cer topography ICD C-73 (malignant neoplasm of thyroid gland), **cancerstage\_II**: cancer stage II, **habilit\_H6**: qualification H6 CACON, **habilit\_H9**: qualification H9

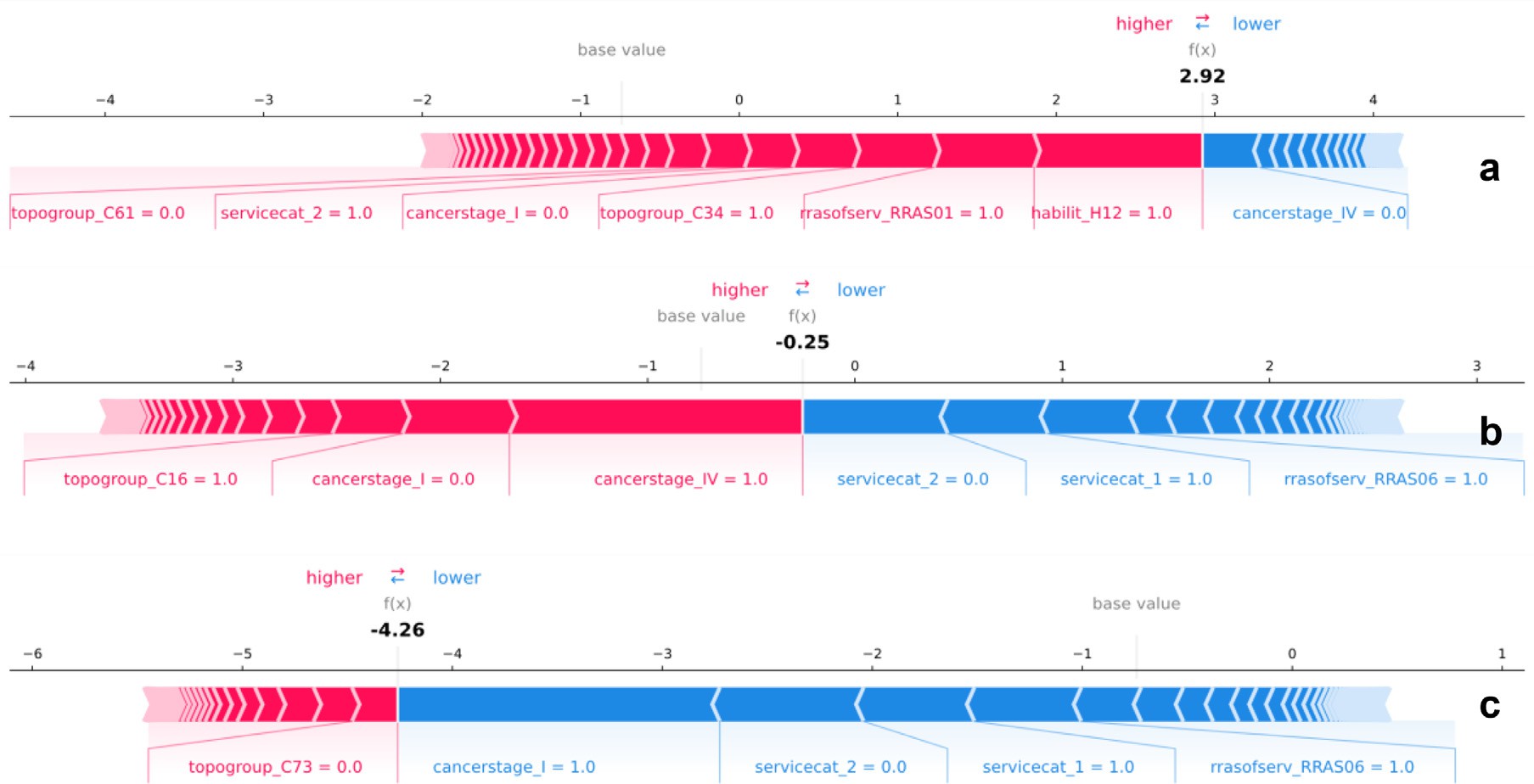
cancer topography ICD C-50 (malignant neoplasm of breast), **age\_70+**: age group of 70 years or more), **rrasofserv\_RRAS09**: regional net of healthcare (service) 09, 9 - UNACON with Radiotherapy and Hematology Services, **medsvtodiag**: difference in days between first medical appointment dates and diagnosis, **topogroup\_C50**:

**morpho\_M80703**: cancer morphology 80,703 (squamous cell carcinoma, NOS), **rras\_R13**: regional net of healthcare (residence) 13, **rrasofserv\_RRAS14**: regional net of healthcare (service) 14, **rrasofserv\_RRAS02**: regional net of healthcare (service) 02.

**Table 5**

Comparison of the predictive performance between the specific algorithms for each type of cancer and the general algorithm.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Cancer Type | Model | Test size | Real Positive | True Positive | False Positive | True Negative | False Negative | Precision | Recall | AUC-ROC |
| Bronchus and Lung | General | 486 | 367 | 340 | 35 | 84 | 27 | 0.9067 | 0.9264 | 0.9265 |
| Specific | 542 | 401 | 367 | 44 | 97 | 34 | 0.8929 | 0.9152 | 0.8993 |
| Breast | General | 1192 | 206 | 146 | 29 | 957 | 60 | 0.8343 | 0.7087 | 0.9460 |
|  | Specific | 1538 | 384 | 277 | 51 | 1103 | 107 | 0.8445 | 0.7214 | 0.9471 |
| Stomach | General | 383 | 277 | 258 | 26 | 80 | 19 | 0.9085 | 0.9314 | 0.9255 |
|  | Specific | 409 | 289 | 270 | 59 | 61 | 19 | 0.8207 | 0.9343 | 0.8658 |
| Colon | General | 468 | 241 | 200 | 36 | 191 | 41 | 0.8475 | 0.8299 | 0.9241 |
|  | Colon | 450 | 228 | 177 | 58 | 164 | 51 | 0.7532 | 0.7763 | 0.8717 |
| Rectum | General | 319 | 197 | 177 | 26 | 96 | 20 | 0.8719 | 0.8985 | 0.9163 |
|  | Colon | 359 | 214 | 196 | 39 | 106 | 18 | 0.8340 | 0.9159 | 0.9230 |
| Prostate | General | 1192 | 206 | 146 | 29 | 957 | 60 | 0.8343 | 0.7087 | 0.9460 |
|  | Specific | 1139 | 199 | 149 | 40 | 900 | 50 | 0.7884 | 0.7487 | 0.9552 |
| Cervix Uteri | General | 326 | 215 | 177 | 21 | 90 | 38 | 0.8939 | 0.8233 | 0.8850 |
|  | Specific | 307 | 183 | 170 | 42 | 82 | 13 | 0.8164 | 0.9235 | 0.9173 |



**Fig. 4.** Main predictors of risk of death from cancer between 12 and 24 months after diagnosis for three randomly selected individuals: a) high risk of death (true positive with 0.972 score), b) medium risk of death (false negative with 0.478 score) and c) low risk of death (true negative with 0.017 score), general model with Catboost Classifier. **Patient a) topogroup\_C61:** cancer topography ICD C-61 (malignant neoplasm of prostate), **servicecat\_2:** public care service, **cancer\_stage1:** cancer stage I, **topogroup\_C34:** cancer topography ICD C-34 (malignant neoplasm of bronchus and lung), **rrasofserv\_RRAS01:** regional net of healthcare (service) 01, **habilit\_H12:** qualification H12 UNACON with Hematology and Pediatric Oncology Services, **cancerstage\_IV:** cancer stage I. **Patient b) servicecat\_1:** private care service, **morpho\_80,703:** cancer morphology 80,703 (squamous cell carcinoma, NOS), **cancerstage\_I:** cancer stage I, **cancerstage\_IV:** cancer stage IV, **servicecat\_2:** public private care service, **rrasofserv\_RRAS06:** regional net of healthcare (service) 06. **Patient c) topogroup\_C73:** cancer topography ICD C-73 (malignant neoplasm of thyroid gland), **cancerstage\_I:** cancer stage I, **servicecat\_2:** public care service, **servicecat\_1:** private care service, **rrasofserv\_RRAS06:** regional net of healthcare (service) 06. Zero value are interpreted as the absence of the characteristic and one as the presence.

# Conclusion

In conclusion, the nine final models developed for predicting risk of death in cancer patients presented high predictive performance. The al- gorithms can be an important tool to help prioritize treatment decisions and patient allocation in cancer treatments, especially in low-income re- gions. Future work should explore the proposed methodological struc- ture and evaluate its predictive performance in new settings with differ- ent routinely collected data.

# Ethical statement

This work was evaluated and approved by the Research Ethics Com- mittee of the Faculty of Public Health of the University of São Paulo (CAAE: 65,375,722.9.0000.5421)

# Data and code availability

The main results of this research were published in this article and in supplementary appendix A and B. The RHC/FOSP data are publicly [available in https://www.fosp.saude.sp.gov.br/fosp/diretoria-adjunta- de-informacao-e-epidemiologia/rhc-registro-hospitalar-de-cancer/. The](https://www.fosp.saude.sp.gov.br/fosp/diretoria-adjunta-de-informacao-e-epidemiologia/rhc-registro-hospitalar-de-cancer/) code developed for predictive modeling can be obtained upon request.

# Funding

This work was supported by the São Paulo State Health Department, Special Health Fund for Mass Immunization and Disease Control (FES- IMA/SES/SP).

# Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

# Data availability

The RHC/FOSP is publicly available on FOSP website. The code de- veloped for predictive modeling can be obtained upon request.

# Acknowledgement

We would like to acknowledge the São Paulo State Health Depart- ment, Special Health Fund for Mass Immunization and Disease Control (FESIMA/SES/SP) for supporting and funding this research.

# Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:[10.1016/j.ailsci.2023.100061](https://doi.org/10.1016/j.ailsci.2023.100061).

# References

1. [Patel A. Benign vs Malignant Tumors. JAMA Oncol 2020;6(9) 1488–1488](http://refhub.elsevier.com/S2667-3185(23)00005-3/sbref0001).
2. [Thuler LCS, Sant’Ana DR, Rezende MCR. Abc do câncer: abordagens básicas para o controle do câncer. ABC do câncer: abordagens para o controle do câncer; 2011. pages 127–127.](http://refhub.elsevier.com/S2667-3185(23)00005-3/sbref0002)
3. World Health Organization (2022). Fact sheets: cancer.
4. INCA (2019). Estimativa 2020: incidência de câncer no Brasil.
5. Ministério da Saúde, Brasil. (2022). Sistema de informação sobre mortalidade (SIM).
6. [Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statis- tics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA 2018;68(6):394–424 [1].](http://refhub.elsevier.com/S2667-3185(23)00005-3/sbref0006)
7. [Iqbal MJ, Javed Z, Sadia H, Qureshi IA, Irshad A, Ahmed R, Malik K, Raza S, Abbas A, Pezzani R, et al. Clinical applications of artificial intelligence and machine learning in cancer diagnosis: looking into the future. Cancer Cell Int 2021;21(1):1–11.](http://refhub.elsevier.com/S2667-3185(23)00005-3/sbref0007)
8. Secretaria de Estado da Saúde. (2022). Fundação Oncocentro de São Paulo. Registro hospitalar de câncer: banco de dados. São Paulo, Brasil.
9. [Lavras C. Atenção primária à saúde e a organização de redes regionais de atenção à saúde no brasil, 20. Saúde e Sociedade; 2011. p. 867–74.](http://refhub.elsevier.com/S2667-3185(23)00005-3/sbref0009)
10. [Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. In: Advances in neural information processing systems; 2018. p. 31.](http://refhub.elsevier.com/S2667-3185(23)00005-3/sbref0010)
11. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Min- ing. New York, NY, USA: ACM; 2016. p. 785–94. doi:[10.1145/2939672.2939785](https://doi.org/10.1145/2939672.2939785).
12. [Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Liu T-Y. Lightgbm: A highly eﬃcient gradient boosting decision tree. Adv Neur Inform Process Syst 2017;30:3146–54.](http://refhub.elsevier.com/S2667-3185(23)00005-3/sbref0012)
13. [Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Duchesnay E. Scikit-learn: Machine learning in Python. J Mach Learn Res 2011;12:2825–30.](http://refhub.elsevier.com/S2667-3185(23)00005-3/sbref0013)
14. [Bergstra J, Yamins D, Cox DD. Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. In: To appear in Proc. of the 30th International Conference on Machine Learning (ICML 2013); 2013.](http://refhub.elsevier.com/S2667-3185(23)00005-3/sbref0014)
15. Kursa MB, Rudnicki WR. Feature Selection with the Boruta Package. J Stat Softw 2010;36(11):1–13. doi:[10.18637/jss.v036.i11](https://doi.org/10.18637/jss.v036.i11).
16. [Lundberg S, Lee S-I. A unified approach to interpreting model predictions. NIPS; 2017.](http://refhub.elsevier.com/S2667-3185(23)00005-3/sbref0016)
17. Lundberg SM, Nair B, Vavilala MS, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nat Biomed Eng 2018;2:749–60. doi:[10.1038/s41551-018-0304-0](https://doi.org/10.1038/s41551-018-0304-0).
18. Lundberg SM, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2020;2:56–67. doi:[10.1038/s42256-019-0138-9](https://doi.org/10.1038/s42256-019-0138-9).
19. [Moons KGM, et al. Transparent reporting of a multivariable prediction model for in- dividual prognosis or diagnosis (TRIPOD): Explanation and elaboration. Ann Intern Med 2015;162(1):W1–73.](http://refhub.elsevier.com/S2667-3185(23)00005-3/sbref0019)
20. Sharma A, Rani R. A Systematic Review of Applications of Machine Learning in Cancer Prediction and Diagnosis. Arch Computat Methods Eng 2021;28:4875–96. doi:[10.1007/s11831-021-09556-z](https://doi.org/10.1007/s11831-021-09556-z).
21. Kumar Y, Gupta S, Singla R, et al. A Systematic Review of Artificial Intelli- gence Techniques in Cancer Prediction and Diagnosis. Arch Computat Methods Eng 2022;29:2043–70. doi:[10.1007/s11831-021-09648-w](https://doi.org/10.1007/s11831-021-09648-w).