algorithms

MDPI

Article

Using Explainable Machine Learning to Explore the Impact of Synoptic Reporting on Prostate Cancer

Femke M. Janssen [1], Katja K. H. Aben [1,2], Berdine L. Heesterman [1], Quirinus J. M. Voorham [3], Paul A. Seegers [3] and Arturo Moncada-Torres [1,*]

[1] The Netherlands Comprehensive Cancer Organization (IKNL), 5612 HZ Eindhoven, The Netherlands; femke-janssen@live.nl (F.M.J.); K.Aben@iknl.nl (K.K.H.A.); b.heesterman@IKNL.NL (B.L.H.)
[2] Radboud Institute for Health Sciences, 6525 EZ Nijmegen, The Netherlands
[3] Nationwide Network and Registry of Histo- and Cytopathology in The Netherlands (PALGA), 1066 CX Amsterdam, The Netherlands; Rinus.Voorham@palga.nl (Q.J.M.V.); Paul.Seegers@palga.nl (P.A.S.)
* Correspondence: a.moncadatorres@iknl.nl

Abstract: Machine learning (ML) models have proven to be an attractive alternative to traditional statistical methods in oncology. However, they are often regarded as *black boxes*, hindering their adoption for answering real-life clinical questions. In this paper, we show a practical application of explainable machine learning (XML). Specifically, we explored the effect that synoptic reporting (SR; i.e., reports where data elements are presented as discrete data items) in Pathology has on the survival of a population of 14,878 Dutch prostate cancer patients. We compared the performance of a Cox Proportional Hazards model (CPH) against that of an eXtreme Gradient Boosting model (XGB) in predicting patient ranked survival. We found that the XGB model (*c*-index = 0.67) performed significantly better than the CPH (*c*-index = 0.58). Moreover, we used Shapley Additive Explanations (SHAP) values to generate a quantitative mathematical representation of how features—including usage of SR—contributed to the models' output. The XGB model in combination with SHAP visualizations revealed interesting interaction effects between SR and the rest of the most important features. These results hint that SR has a moderate positive impact on predicted patient survival. Moreover, adding an explainability layer to predictive ML models can open their *black box*, making them more accessible and easier to understand by the user. This can make XML-based techniques appealing alternatives to the classical methods used in oncological research and in health care in general.

Keywords: Cox Proportional Hazards (CPH); explainable AI; eXtreme Gradient Boosting (XGB); interpretability; oncology; prostatectomy; ranked survival; SHAP

Citation: Janssen, F.M.; Aben, K.K.H.; Heesterman, B.L.; Voorham, Q.J.M.; Seegers, P.A.; Moncada-Torres, A. Using Explainable Machine Learning to Explore the Impact of Synoptic Reporting on Prostate Cancer. *Algorithms* **2022**, *15*, 49. https://doi.org/10.3390/a15020049

Academic Editor: Laurent Risser

Received: 29 October 2021
Accepted: 27 January 2022
Published: 29 January 2022

1. Introduction

Machine learning (ML) methods have been shown to be great complements to classical statistical tools in health care [1–3] and more particularly in oncology, where they have proven to be valuable for improving patient screening, diagnosis, and treatment [4–7]. More interestingly, many ML models have been adapted to handle censored data (i.e., instances that have not yet experienced the event of interest), making them attractive options for studying recurrence and survival in cancer patients [8].

A few of these ML models (such as Survival Trees [9] and Bayesian Methods [10]) have relatively simple inner workings, making them transparent to the user and easy to interpret. Other more complex models (such as Support Vector Machines [11], Neural Networks [12], and Ensemble Methods [13,14]) tend to have higher performance in regression and prediction tasks [15]. However, due to the opacity of their inner workings when producing their results, these are often treated as *black boxes*. This is unfortunate, since it makes it hard to explain how they generated their output [16,17]. This hampers the trust that the users (e.g., oncologists, cancer patients) have on the models' predictions, which is detrimental for their application in real-life scenarios [18].

To mitigate this issue, several techniques have been proposed to open the so-called *black box* of such high-performing, intricate models. These include Permutation Feature Importance [19], Accumulated Local Effects [20], Local Interpretable Model-agnostic Explanations (LIME [16]), Partial Dependence Plots, and SHapley Additive exPlanations (SHAP [21]), among others [22]. Overall, their purpose is to generate an explicit knowledge representation (in terms understandable to humans) of the models' inner workings and of how they generate their predictions [23]. The use of explainable ML (XML) as a novel paradigm has started to grow in health care [24–26] and has been used in a few studies in oncology [27–31], but its potential remains largely unexplored and underused.

Synoptic Reporting

Histopathological reports are of uttermost importance for providing timely and proper information to oncologists, which is crucial in the diagnostic process and delivery of high quality cancer care [32–34]. Traditional pathology reports are in the form of narrative text, which means that they are written without a fixed, structured format [33,35]. Unfortunately, these narrative reports are susceptible to data incompleteness, inferior understandability, and misinterpretation [36], since they lack a fixed structure of scientifically validated data items.

To alleviate these disadvantages, synoptic reporting (SR) has been introduced. SR occurs in the form of a report in which the information elements are presented in a predefined tabular form and stored as discrete data items in a database [37,38]. This structure makes the most important information more accessible to clinicians, which has the potential to reduce medical errors [38,39]. Additionally, SR allows for faster detection of essential data, higher completeness of reported information, improved reporting of clinically relevant data, uniformity, and making information computer readable [40,41].

In the Netherlands, SR was introduced in 2008 by the Nationwide Network and Registry of Histo- and Cytopathology (PALGA), starting with pathological reporting of breast and colorectal cancer. Currently, SR has been implemented in more than 31 different protocols, all of which have been formally approved by the Dutch Society of Pathology. Although the benefits of SR for Dutch patients has been investigated in a few types of cancer (e.g., colorectal [35,42], gallbladder [43]), this has not been done in prostate cancer.

In this paper, we used XML to study the effect of SR on predicted ranked survival of Dutch prostate cancer patients. Additionally, we compared its performance with that of a classical statistical method. The manuscript is organized as follows. Section 2 gives a detailed description of how the data were curated and how the models were developed and validated. Section 3 presents the obtained results, which are further discussed in detail in Section 4. Section 5 closes the paper with our overall conclusions.

2. Materials and Methods

2.1. Data

In this retrospective cohort study, we used data from the Netherlands Cancer Registry (NCR) and the Nationwide Network and Registry of Histo- and Cytopathology in The Netherlands (PALGA) [44] retrieved from the national SR data for radical prostatectomy. These were linked through a trusted third party (ZorgTTP). The data comprised all prostate cancer patients diagnosed between 2011 and 2018 in the Netherlands. Features included demographic, clinical, and pathological data, along with tumor, lymph node, and metastases (TNM)-stage. Figure 1 shows the patient selection process. Originally, the dataset consisted of 55,616 rows. We excluded patients that had distant metastases (defined as a clinical or pathological M1). We only used records of patients who received a radical prostatectomy (i.e., complete surgical removal of the prostate, resection). Lastly, we removed any double records (which occurred when a PALGA excerpt was coupled to more than one tumor or patient) and duplicates (often due to registration artifacts). This resulted in a preliminary dataset of 17,587 patients and 41 columns.

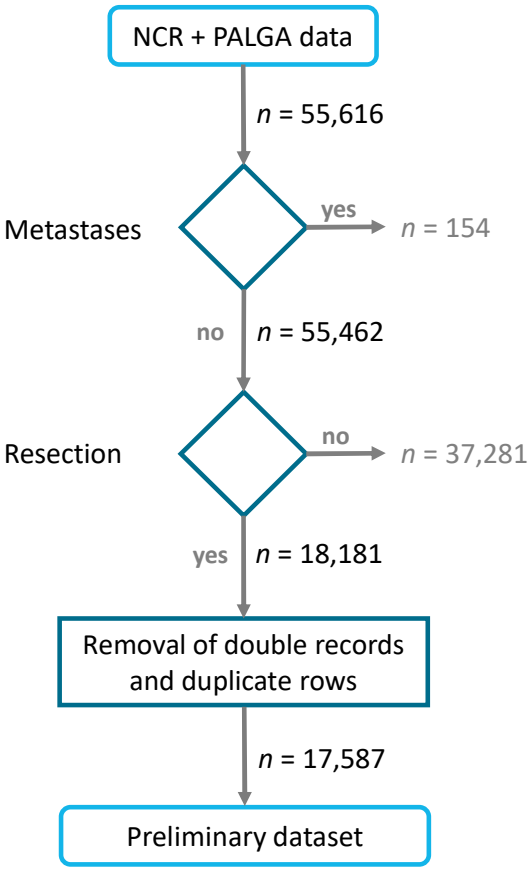

Figure 1. Flowchart depicting inclusion/exclusion criteria. The preliminary dataset consisted of 17,587 patients and 41 columns.

2.2. Pre-Processing

We used `Python v3.7.1` for the analyses of the whole study.

2.2.1. Feature Engineering

Firstly, we engineered a new feature: the European Association of Urology risk group (`EAU`), which classifies patients into risk groups for biochemical recurrence of localized and locally advanced prostate cancer [45]. It provides a prognostic profile of a patient's tumor by assigning it to one of four categories based on prostate-specific antigen (PSA), Gleason score, and clinical T and clinical N values (Table 1). For the latter, the 7th edition of the Union for International Cancer Control (UICC) classification of malignant tumors was used for patients diagnosed between 2011 and 2017 [46], and the 8th edition was used for patients diagnosed between 2017 and 2018 [47]. However, this did not have an influence on the `EAU` feature, since the definition of cT and cN did not change between versions [48].

Table 1. Risk groups for biochemical recurrence of localized and locally advanced prostate cancer as defined by the European Association of Urology.

Localized			Locally Advanced
Low Risk	**Intermediate Risk**	**High Risk**	**High Risk**
PSA $<$ 10 ng/mL and GS $<$ 7 and cT1-2a	PSA 10–20 ng/mL or GS $=$ 7 or cT2b	PSA $>$ 20 ng/mL or GS $>$ 7 or cT2c	Any PSA Any GS cT3-4 or cN+

PSA: prostate-specific antigen; GS: Gleason score; cT: clinical T; cN: clinical N.

2.2.2. Feature Selection

Afterwards, we chose a subset of features that were clinically relevant predictors of prostate cancer survival based on the recommendation of clinical experts. These were `age`, `EAU` (as defined in Section 2.2), `incidence year` (i.e., year in which the patient was diagnosed), `academic hospital` (i.e., whether a patient was treated at an academic hospital or not), and `SR`.

2.2.3. Dealing with Missing Values

Not all the models used in this study (Section 2.3) are capable of dealing with missing values. Therefore, in order to make a comparison that was as fair as possible, we chose to use a complete case analysis approach (i.e., removing patients that had any missing values).

A summary of the data can be found in Table 2. For continuous variables, we report their mean and standard deviation, whereas for categorical variables we report their absolute and relative numbers (as a percentage). In all cases, we also show completeness of each variable (before deletion). In the end, the final dataset consisted of five features and a total of 14,878 patients.

Table 2. General overview of the dataset used in this study.

Variable	Mean	SD	N	%	Original Completeness (%)
Input					
`age` (in years)	65.11	5.97	-	-	100.0
`EAU`					84.6
localized—low risk	-	-	506	3.40	
localized—intermediate risk	-	-	10,930	73.46	
localized—high risk	-	-	3094	20.80	
Locally advanced—high risk	-	-	348	2.34	
`incidence year`					100.0
2011	-	-	1599	10.75	
2012	-	-	1784	11.99	
2013	-	-	1957	13.15	
2014	-	-	1792	12.05	
2015	-	-	1829	12.29	
2016	-	-	1982	13.32	
2017	-	-	1930	12.97	
2018	-	-	2005	13.48	
`academic hospital`					100.0
Yes	-	-	3485	23.42	
No	-	-	11,393	76.58	
`SR`					100.0
Yes	-	-	7568	50.87	
No	-	-	7310	49.13	

Completeness refers to the percentage of patients with no missing values (Section 2.2). SD: standard deviation; `EAU`: European Association of Urology risk group; SR: synoptic reporting.

2.3. Models

We defined ranked patient survival as the outcome to be predicted, which is calculated based on follow-up time and event occurrence. We used two different models for this task.

First, we performed a multiple CPH analysis (using `scikit-survival v0.15.0` [49]), since this is one of the most common and well-known methods used in oncology. It is a semi-parametric approach that evaluates the effects of different covariates (i.e., features) on the hazard ratio (HR) of the occurrence of death. In a CPH regression, a patient's hazard is modelled as a combination of the population time-variant baseline hazard and of his/her time-invariant predictors (multiplied by their corresponding coefficients). Moreover, it

makes no assumptions on the underlying hazard function. We also performed a classical CPH multiple regression (using `lifelines v0.25.11` [50]), which yielded HRs for each feature in the model.

Additionally, we used eXtreme Gradient Boosting (XGB, using `xgboost v1.3.3`) as a representative ML technique for survival analysis [14]. We chose this model based on the results of our previous work, where we compared the performance of several ML algorithms for predicting ranked survival in a similar dataset and found that XGB performed the best [28]. These results are also in line with those of a few other studies in literature that have successfully used XGB for predictive modelling in oncology [29,51,52]. XGB poses its learning task as a numerical optimization process. It performs the gradient descent procedure by calculating the loss function and minimizes it by adding decision (regression) trees as weak learners (i.e., classifiers that perform slightly better than chance), as shown in Equation (1):

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i{}^{(t-1)} + f_t(\mathrm{x}_i)) + \Omega(f_t) \quad (1)$$

Here, l is a (differentiable) convex loss function measuring the difference between the target y_i and the prediction $\hat{y}_i$ at the iteration $t-1$. The tree $f_t(\mathrm{x}_i)$, is built using data x_i that most improves the model [14]. These trees are parameterized and added one at a time until the loss reaches an acceptable value or until there is no further improvement. The final output is given by the weighted sum of the individual trees' predictions. Moreover, XGB actively tries to reduce overfitting by using subsets of the data for generating each new tree, by constraining the trees characteristics (such as number of trees or tree depth), and by weighting the updates (i.e., applying a learning rate), as given by the term $\Omega(f_t)$ [53].

We performed a nested cross-validation procedure. The inner loop consisted of a fivefold cross validation used for tuning the hyperparameters of the XGB model using a randomized search of 2000 different parameter settings (shown in Table 3). The outer loop consisted of a tenfold cross validation used to assess the model performance. In all cases, the target was to maximize Harrell's concordance index (Section 2.3.1).

Table 3. Hyperparameters of the XGB model.

Hyperparameter	Hyperparameter Space	Chosen Value
Max. depth	[1, 2, 3, 4, 5, 10, 15, 25, 50, 100, 250, 500]	2
Max. number of trees	[25, 50, 75, 100, 250, 500, 750, 1000, 1500]	75
Learning rate	Logarithmic space ranging from 10^{-2} to 10^{0}	0.054
Subsamples	[0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8]	0.5

2.3.1. Model Evaluation

The models' output is a prediction of the sequence of events (i.e., which patients have a higher risk to die) on an arbitrary scale. The models do so based on a combination of a patient's follow-up time and his corresponding censorship indicator. Hence, we evaluated the models' accuracy using Harrell's concordance index (c-index) [54].

The c-index is a measure of how discriminant a model is in a ranking prediction task. It can be interpreted as the probability that a certain patient with a shorter time-to-event (as compared to another patient) is assigned a higher predicted probability of having the event (in our case, death). Mathematically, it is the proportion of concordant patient pairs divided by the total number of patient pairs. A concordant pair is defined as the case when a patient with a shorter follow-up time than another patient indeed received a higher probability of an event. In essence, the c-index could be interpreted in a similar way as the area under the receiver operating curve [54,55], with a value of $c = 1$ indicating a capacity to perfectly predict patient ordering. In other words, a model with a high c-index value would be able to predict between two patients which one will have a shorter survival time

with high reliability. We used tenfold cross validation to calculate a mean *c*-index for the CPH and XGB models. Then, we compared these values using a paired Student *t*-test [56].

2.4. Explainability

Lastly, in order to understand how our models yielded their predictions, we decided to use one of the many explainability techniques that are available [22,57]. Namely, we chose SHapley Additive exPlanations (SHAP) as proposed and implemented by Lundberg and Lee (`v0.39.0` [21]), since in our previous work [31] (where we compared it against the LIME approach [16]) we found that it was capable of providing detailed enough explanations of a model's inner workings at a local *and* at a global level, the latter being crucial to our model at hand. Moreover, this technique has numerous advantages. First, it offers a unified approach that presents the mathematical properties of (local) accuracy (i.e., an approximate model built to explain the original model should match the output of the original model for a given feature), missingness (i.e., when a feature is missing, there should be no impact on its attribution), and consistency (i.e., if a feature's contribution stays the same or increases regardless of other features, then its attribution should not decrease), which are not found simultaneously in other techniques [21,58]. These are desirable properties that guarantee sound behaviour of the obtained explanations. More importantly, SHAP values are model agnostic, meaning that they can be applied to models of different nature, which was imperative to our comparison. Their execution on tree-based models is extremely efficient [59], which was particularly advantageous for the XGB model. Lastly, their implementation is open-source and is being actively supported and developed by the original authors, as well as by the community.

The solid theoretical background behind SHAP values is derived from coalitional game theory, where the original purpose was to quantify the fair distribution of the payout of players in a game scenario [60]. In our case, SHAP values quantify the contribution of features to a model's output in a prediction task as given by Equation (2).

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z'_j \tag{2}$$

Here, g is the explanation model and z' is the coalition vector (i.e., a vector where a value of 1 means that the corresponding feature is present, while a value of 0 means that it is absent). The value of g is composed of the model's base rate ϕ_0 and the sum of the individual contributions of all features ϕ_j (i.e., the features' Shapley values).

Moreover, SHAP values are capable of quantifying feature contribution not only at an individual level, but also among all pairs of them. This allowed us to study local interaction effects out-of-the-box, providing new insights into the relations between the model's input [29]. For a more in-depth explanation about SHAP values, we refer the reader to the original papers [21,59].

3. Results

Table 4 shows the results of the CPH multiple regression. Based on the corresponding *z*- and *p*-values, `age` and `EAU` were the most important predictors of ranked survival for the data at hand. Additionally, `SR` had a significant positive effect on overall predicted survival. The features `academic hospital` and `incidence year` had no significant impact on the model's output.

Figure 2 shows a bar plot comparing the performance of the CPH and XGB models' ranked survival predictions using the *c*-index. The latter is presented as a mean with its corresponding 95% confidence intervals from the tenfold cross-validation. XGB performed significantly better in predicting ranked survival predictions as compared to CPH, with *c*-index values of 0.67 and 0.58, respectively ($p < 0.0001$).

Next, we calculated SHAP values of the CPH and XGB models, which are shown in Figure 3. In short, these plots show the impact that each feature had on the model output.

Features are ordered from top to bottom in decreasing importance, given by the mean of their absolute Shapley values. For each of them, each dot corresponds to a patient. Their location on the x-axis is determined by their SHAP value. In our case, positive SHAP values correspond to a higher chance of death, while negative values correspond to the opposite (i.e., a higher chance of survival). In other words, a patient with a higher SHAP value has a higher mortality risk relative to a patient with a lower SHAP value. The color indicates the value of the feature it represents and is depicted from low (blue) to high (pink).

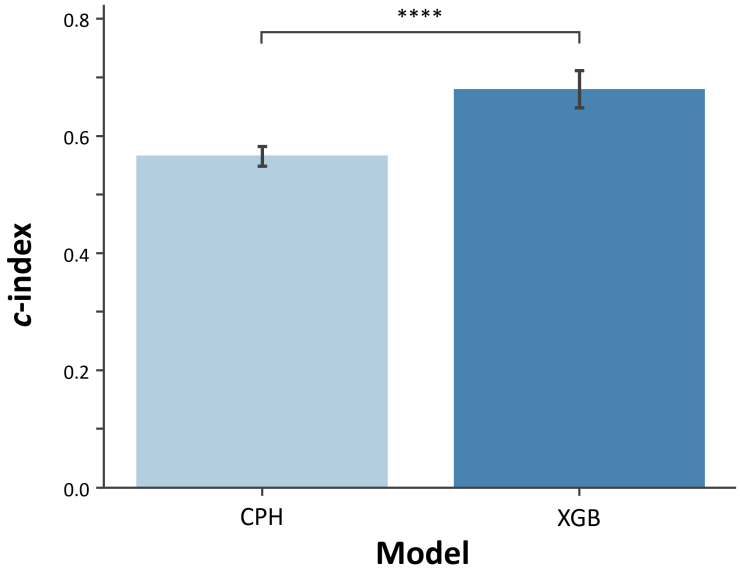

Figure 2. Mean c-index of the CPH and XGB models using tenfold cross-validation. Error bars represent the 95% confidence intervals across folds. **** $p < 0.0001$.

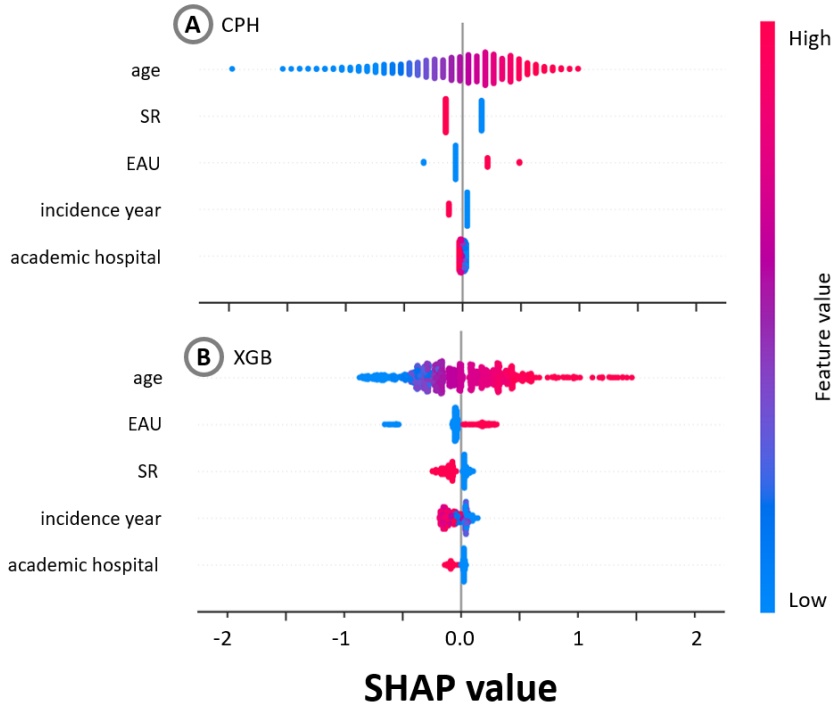

Figure 3. SHAP swarm plots of (**A**) CPH model and (**B**) XGB model. In each panel, features are arranged from top to bottom in decreasing order of importance. For each of them, each dot corresponds to a patient. Their location on the x-axis is determined by their SHAP value. Higher SHAP values correspond to a higher chance of death. The color indicates the value of the feature it represents and is depicted from low (blue) to high (pink).

Afterwards, we generated dependence plots for all five features, as shown in Figure 4. In these, we also compared the CPH and XGB models (in light and dark blue, respectively). Once more, each dot corresponds to one patient. The feature's values are represented along the *x*-axis, while the SHAP values are shown along the *y*-axis (their interpretation remains unchanged: positive SHAP values correspond to a higher chance of death and viceversa).

Since we were particularly interested in studying the effect of SR, we also studied its interaction effect with the two other most important features: `age` and `EAU`, which is shown in Figure 5. In this case, color represents the value of the interacting feature, going from low (in blue) to highest (in pink).

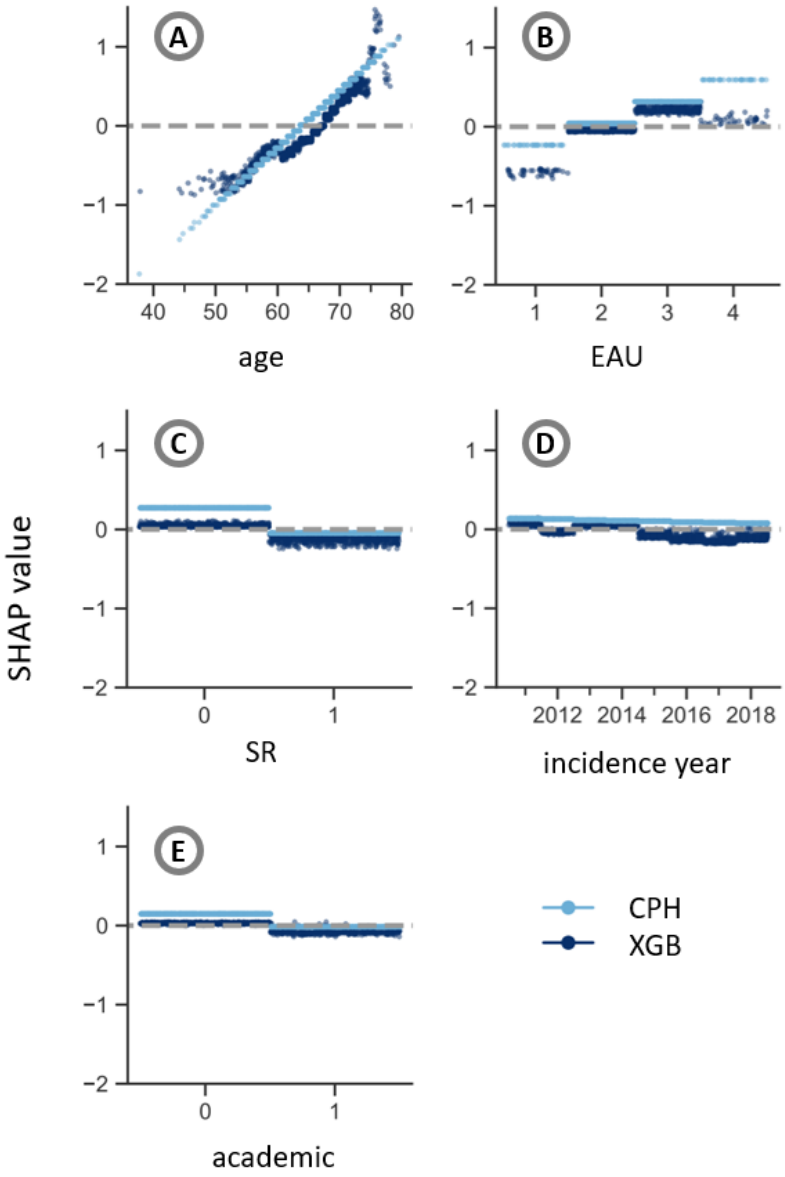

Figure 4. Dependence plot for all features: (**A**) `age`; (**B**) `EAU` (encoded as 1—low risk/localised, 2—intermediate risk/localised, 3—high risk/localised, and 4—high risk/locally advanced); (**C**) `SR` (encoded as 0—False, 1—True); (**D**) `incidence year`; (**E**) `academic hospital` (encoded as 0—False, 1—True). Each dot corresponds to one patient. In the case of categorical features (i.e., all features except `age`), artificial jitter was added along the *x*-axis for the sake of easier representation. The *y*-axis scale is the same for all features in order to give a proper idea of their contributions to the model output.

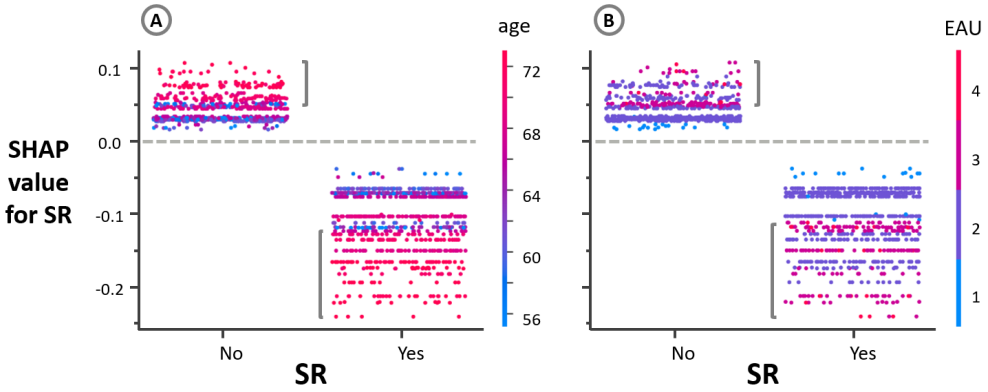

Figure 5. Interaction effects between `SR` and the two most relevant features: (**A**) `age` and (**B**) `EAU` (encoded as 1—low risk/localised, 2—intermediate risk/localised, 3—high risk/localised, and 4—high risk/locally advanced). In both cases, artificial jitter was added along the *x*-axis to better show the point density.

Table 4. Results of the CPH multiple regression

Feature	HR	95% CI		*z*-Value	*p*-Value
`age`	1.07	1.06	1.09	10.72	<0.005
`EAU`					
Localized—low risk	1.00	-	-	-	-
Localized—intermediate risk	2.39	1.28	4.46	2.73	0.01
Localized—high risk	3.27	1.73	6.18	3.65	<0.005
Locally advanced—high risk	2.82	1.47	5.31	3.16	0.01
`SR`					
No	1.00	-	-	-	-
Yes	0.78	0.63	0.97	−2.27	0.02
`academic hospital`					
No	1.00	-	-	-	-
Yes	0.90	0.76	1.06	−1.23	0.22
`incidence year`	0.99	0.94	1.05	−0.25	0.80

HR: hazard ratio; CI: confidence interval.

4. Discussion

In this paper, we showed a practical application of XML in oncology. More specifically, we looked at the effect that SR had on the survival of a Dutch prostate cancer subpopulation of 14,878 patients. For this, we used two different approaches: classical statistical survival analysis using a CPH model and state-of-the-art ML in the shape of an XGB model. More importantly, we included an explainability layer in the form of SHAP values to generate an explicit knowledge representation of how the models generated their predictions.

The comparison between the models (Figure 2) showed that the XGB approach performed significantly better ($p < 0.0001$) with a *c*-index of 0.67 (which can practically be considered reasonable performance [61]), in contrast to the CPH approach, with a *c*-index of 0.58. This is in line with findings in the literature that performed similar comparisons between CPH and ML models [28,62–65]. These works found that ML-based techniques can perform at least as well as classical statistical methods (if not better) than classical statistical methods in predicting ranked occurrence of patient events. This highlights the potential of a more extensive use of ML-based methods when studying patient survival or recurrence.

Afterwards, we opened the models' *black boxes* by computing their SHAP values. This allowed us to generate valuable insights on many different aspects. First of all, we were able to generate an explicit representation (Figure 3) of what impact the different features had on the models' output (and thus on their performance). On one hand, we found that

for the CPH model, the most important feature was `age`, followed by `SR` and `EAU`. On the other hand, for the XGB model the most important feature was also `age`, but was followed by `EAU` and *then* by `SR`. We believe that the latter ranking makes more sense, given that the `EAU` feature has more information about the condition of a patient (since it is calculated from his TNM staging, Gleason score, and PSA levels, Section 2.2). In both cases, `incidence year` and `academic hospital` were the least important features. We hypothesize that this could be because there were no important changes in the care pathway of this particular patient subpopulation dependent on time (`incidence year`) or on the type of hospital where the patients were treated (`academic hospital`). These results are also consistent with the HRs calculated earlier (Table 4), where those of `age`, `EAU`, and `SR` were significant, while those of `incidence year` and `academic hospital` were not. Therefore, we will focus on the top three (significant) features (`age`, `EAU`, and `SR`) for the rest of this discussion.

Dependence plots (Figure 4) can provide a few insights on the feature distributions. For example, we can see that the number of patients that are 50 years old or younger is lower than in other age groups. We can also observe that there are more patients with an `EAU` of 2 or 3 (i.e., intermediate and high risk, respectively; both localised) compared to the rest. A more interesting practical application was to provide additional information on how the models deal with their inputs. Particularly, we can clearly see how the CPH approach is only capable of modelling linear relations between the features and its output. It could be argued that this model could be extended [66–68] to account for non-linearities in the data through, for example, covariate transformations, step-wise regression, or specialized functions. However, these alternatives are often highly dependent on the (subjective) expertise of the researcher, which could also reduce the model generalizability, making it prone to overfitting. In contrast, the XGB approach is capable of capturing and modelling said non-linearities out-of-the-box, without any additional effort from the researcher in a more generalizable, data-driven approach. For example, in the case of `age` (Figure 4A), we can identify a plateau for patients between ∼40 and ∼55 years on their SHAP values. Then, there is a relatively constant rise until ∼75 years, with a sharp increase after that. Moreover, dependence plots also allow to compare the model's inner workings with real life and clinical intuition (based on the user's medical expertise) to better understand the phenomenon at hand. For example, `EAU` (Figure 4B) shows that the XGB model gives patients with a low risk a much better chance of survival compared to the rest, which is expected. Lastly, we can see that the XGB model found that when a patient received `SR` (Figure 4C), his SHAP values shifted from positive to negative, further suggesting that SR has a positive impact on predicted patient survival. Since the latter was of particular interest, we also computed the interaction effects `SR` had with `age` and `EAU` (Figure 5). Figure 5A reveals that the negative/positive effect of not receiving/receiving SR is more impactful for older patients. Figure 5B shows a similar trend: not receiving/receiving SR is more detrimental/beneficial for patients that are in a higher EAU risk group (either localised or locally advanced).

The results presented here are the first step in exploring the effect that SR has on the survival of prostate cancer patients. They should be expanded and confirmed with further research. For example, it could be interesting to incorporate additional prognostic factors (such as patient performance or comorbidities) and look at cause-specific survival (which were not available in the dataset at hand). It could also be valuable to investigate what items were missing in the narrative reports in comparison with those with SR and analyze what impact they had on the patient care pathway. Given the volume of reports, using natural language processing (NLP) could be a suitable approach [69].

Lastly, it is worth mentioning that (X)ML-based models still present a few limitations. For instance, they are well-known to be data hungry, requiring large amounts of data to achieve acceptable performance. Very often, they require optimization and tuning of different parameters, which can be a cumbersome process. Lastly, depending on their complexity, these type of models can be very resource consuming and computationally demanding.

5. Conclusions

In this paper, we explored the impact that SR has on predicted ranked survival in a population of 14,878 Dutch prostate cancer patients. We used classical statistical methods (in the form of a CPH model) as well as more novel XML techniques (in the form of an XGB model in combination with SHAP values). Our results show that XGB approach performed significantly better than CPH in our patient cohort. The explainability layer of the analysis pipeline (in the form of SHAP values) revealed that this difference in performance was due to the XGB's capability of capturing and modelling non-linearities as well as interaction effects present in the data. SHAP values hint that SR has a moderate positive impact on predicted patient survival. Combining the XGB model with SHAP visualizations revealed interesting interaction effects between `SR` and other important features of interest, such as `age` and `EAU`.

These findings show how XML-based techniques are capable of competitive performance, while at the same time opening their *black box* by generating an explicit knowledge representation of how models derive at their predictions. While increasing the trust that end users (e.g., clinicians, patients) have on (complex) ML models remains a huge challenge [17,70,71], we believe that XML is a step in the right direction, potentially making them even more attractive tools in oncology and in health care in general.

Author Contributions: Conceptualization, K.K.H.A., Q.J.M.V., P.A.S. and A.M.-T.; methodology, F.M.J. and A.M.-T.; software, F.M.J.; validation, all authors; formal analysis, F.M.J., K.K.H.A., B.L.H. and A.M.-T.; investigation, F.M.J. and A.M.-T.; resources, A.M.-T.; data curation, F.M.J., Q.J.M.V., P.A.S.; writing—original draft preparation, F.M.J. and A.M.-T.; writing—review and editing, all authors; visualization, F.M.J. and A.M.-T.; supervision, A.M.-T.; project administration, A.M.-T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: This study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Supervisory Committee of the Netherlands Cancer Registry (NCR). It used data from the NCR and the Nationwide Network and Registry of Histo- and Cytopathology in the Netherlands (PALGA), granted under requests K19.328 and lzv2016-137-FU, respectively, on October 2020.

Informed Consent Statement: Patient consent was waived by the NCR Supervisory Committee since this study had a national, non-interventional retrospective design and all data were analyzed anonymously.

Data Availability Statement: The data used in this study are not publicly available due to privacy restrictions. However, they are available through the standard data usage request process from the NCR upon reasonable request.

Acknowledgments: We would like to thank the registration team of the Netherlands Comprehensive Cancer Organization (IKNL) for the collection of the data for the NCR.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CPH	Cox Proportional Hazards
`EAU`	European Association of Urology risk group
GS	Gleason score
HR	Hazard ratio
LIME	Local Interpretable Model-agnostic Explanations
NCR	Netherlands Cancer Registry
NLP	Natural Language Processing
PALGA	Nationwide Network and Registry of Histo- and Cytopathology in the Netherlands

PSA	Prostate-specific antigen
SD	Standard deviation
SHAP	SHapley Additive exPlanations
SR	Synoptic Reporting
TNM	Tumor, nodes, metastases
UICC	Union for International Cancer Control
XGB	eXtreme Gradient Boosting

References

1. Ravì, D.; Wong, C.; Deligianni, F.; Berthelot, M.; Andreu-Perez, J.; Lo, B.; Yang, G.Z. Deep learning for health informatics. *IEEE J. Biomed. Health Inform.* **2016**, *21*, 4–21. [CrossRef] [PubMed]
2. Panch, T.; Szolovits, P.; Atun, R. Artificial intelligence, machine learning and health systems. *J. Glob. Health* **2018**, *8*, 020303. [CrossRef] [PubMed]
3. Ghassemi, M.; Naumann, T.; Schulam, P.; Beam, A.L.; Chen, I.Y.; Ranganath, R. A review of challenges and opportunities in machine learning for health. *AMIA Summits Transl. Sci. Proc.* **2020**, *2020*, 191. [PubMed]
4. Kourou, K.; Exarchos, T.P.; Exarchos, K.P.; Karamouzis, M.V.; Fotiadis, D.I. Machine learning applications in cancer prognosis and prediction. *Comput. Struct. Biotechnol. J.* **2015**, *13*, 8–17. [CrossRef]
5. Cuocolo, R.; Caruso, M.; Perillo, T.; Ugga, L.; Petretta, M. Machine Learning in oncology: A clinical appraisal. *Cancer Lett.* **2020**, *481*, 55–62. [CrossRef]
6. Ngiam, K.Y.; Khor, W. Big data and machine learning algorithms for health-care delivery. *Lancet Oncol.* **2019**, *20*, e262–e273. [CrossRef]
7. Munir, K.; Elahi, H.; Ayub, A.; Frezza, F.; Rizzi, A. Cancer Diagnosis Using Deep Learning: A Bibliographic Review. *Cancers* **2019**, *11*, 1235. [CrossRef]
8. Wang, P.; Li, Y.; Reddy, C.K. Machine learning for survival analysis: A survey. *ACM Comput. Surv.* **2019**, *51*, 1–36. [CrossRef]
9. Bou-Hamad, I.; Larocque, D.; Ben-Ameur, H. A review of survival trees. *Stat. Surv.* **2011**, *5*, 44–71. [CrossRef]
10. Raftery, A.E.; Madigan, D.; Volinsky, C.T. Accounting for model uncertainty in survival analysis improves predictive performance. *Bayesian Stat.* **1996**, *5*, 323–349.
11. Pölsterl, S.; Navab, N.; Katouzian, A. Fast training of support vector machines for survival analysis. In Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Porto, Portugal, 7–11 September 2015; pp. 243–259.
12. Katzman, J.L.; Shaham, U.; Cloninger, A.; Bates, J.; Jiang, T.; Kluger, Y. DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Med. Res. Methodol.* **2018**, *18*, 24. [CrossRef] [PubMed]
13. Ishwaran, H.; Kogalur, U.B.; Blackstone, E.H.; Lauer, M.S. Random survival forests. *Ann. Appl. Stat.* **2008**, *2*, 841–860. [CrossRef]
14. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794. [CrossRef]
15. Duval, A. *Explainable Artificial Intelligence (XAI)*; Mathematics Institute, The University of Warwick: Coventry, UK, 2019.
16. Ribeiro, M.T.; Singh, S.; Guestrin, C. Why should I trust you? Explaining the Predictions of Any Classifier. *arXiv* **2016**, arXiv:1602.04938.
17. Doshi-Velez, F.; Kim, B. Towards a rigorous science of interpretable machine learning. *arXiv* **2017**, arXiv:1702.08608.
18. Holzinger, A.; Biemann, C.; Pattichis, C.S.; Kell, D.B. What do we need to build explainable AI systems for the medical domain? *arXiv* **2017**, arXiv:1712.09923.
19. Altmann, A.; Toloşi, L.; Sander, O.; Lengauer, T. Permutation importance: A corrected feature importance measure. *Bioinformatics* **2010**, *26*, 1340–1347. [CrossRef]
20. Apley, D.W.; Zhu, J. Visualizing the effects of predictor variables in black box supervised learning models. *J. R. Stat. Soc. Ser. B* **2020**, *82*, 1059–1086. [CrossRef]
21. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **2017**, *2017*, 4765–4774.
22. Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; Pedreschi, D. A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv.* **2018**, *51*, 93:1–93:42. [CrossRef]
23. Adadi, A.; Berrada, M. Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access* **2018**, *6*, 52138–52160. [CrossRef]
24. Ahmad, M.A.; Eckert, C.; Teredesai, A. Interpretable machine learning in healthcare. In Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, Washington, DC, USA, 29 August–1 September 2018; pp. 559–560.
25. Pawar, U.; O'Shea, D.; Rea, S.; O'Reilly, R. Explainable ai in healthcare. In Proceedings of the 2020 International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA), Dublin, Ireland, 15–19 June 2020; pp. 1–2.
26. Tonekaboni, S.; Joshi, S.; McCradden, M.D.; Goldenberg, A. What clinicians want: Contextualizing explainable machine learning for clinical end use. In Proceedings of the Machine Learning for Healthcare Conference, PMLR, Ann Arbor, MI, USA, 9–10 August 2019; pp. 359–380.

27. Okagbue, H.I.; Adamu, P.I.; Oguntunde, P.E.; Obasi, E.C.M.; Odetunmibi, O.A. Machine learning prediction of breast cancer survival using age, sex, length of stay, mode of diagnosis and location of cancer. *Health Technol.* **2021**, *11*, 887–893. [CrossRef]
28. Moncada-Torres, A.; van Maaren, M.C.; Hendriks, M.P.; Siesling, S.; Geleijnse, G. Explainable machine learning can outperform Cox regression predictions and provide insights in breast cancer survival. *Sci. Rep.* **2021**, *11*, 6968. [CrossRef]
29. Li, R.; Shinde, A.; Liu, A.; Glaser, S.; Lyou, Y.; Yuh, B.; Wong, J.; Amini, A. Machine Learning-Based Interpretation and Visualization of Nonlinear Interactions in Prostate Cancer Survival. *JCO Clin. Cancer Inform.* **2020**, *4*, 637–646. [CrossRef] [PubMed]
30. Giraud, P.; Giraud, P.; Nicolas, E.; Boisselier, P.; Alfonsi, M.; Rives, M.; Bardet, E.; Calugaru, V.; Noel, G.; Chajon, E.; et al. Interpretable Machine Learning Model for Locoregional Relapse Prediction in Oropharyngeal Cancers. *Cancers* **2021**, *13*, 57. 10.3390/cancers13010057. [CrossRef] [PubMed]
31. Jansen, T.; Geleijnse, G.; van Maaren, M.; Hendriks, M.P.; Ten Teije, A.; Moncada-Torres, A. Machine Learning Explainability in Breast Cancer Survival. In *Studies in Health Technology and Informatics. Digital Personalized Health and Medicine*; IOS Press: Amsterdam, The Netherlands, 2020 ; Volume 270, pp. 307–311.
32. Valenstein, P.N. Formatting pathology reports: Applying four design principles to improve communication and patient safety. *Arch. Pathol. Lab. Med.* **2008**, *132*, 84–94. [CrossRef] [PubMed]
33. Aumann, K.; Niermann, K.; Asberger, J.; Wellner, U.; Bronsert, P.; Erbes, T.; Hauschke, D.; Stickeler, E.; Gitsch, G.; Kayser, G.; et al. Structured reporting ensures complete content and quick detection of essential data in pathology reports of oncological breast resection specimens. *Breast Cancer Res. Treat.* **2016**, *156*, 495–500. [CrossRef] [PubMed]
34. Nakhleh, R.E. Quality in surgical pathology communication and reporting. *Arch. Pathol. Lab. Med.* **2011**, *135*, 1394–1397. [CrossRef]
35. Sluijter, C.E.; van Workum, F.; Wiggers, T.; van de Water, C.; Visser, O.; van Slooten, H.J.; Overbeek, L.I.H.; Nagtegaal, I.D. Improvement of Care in Patients With Colorectal Cancer: Influence of the Introduction of Standardized Structured Reporting for Pathology. *JCO Clin. Cancer Inform.* **2019**, *3*, 1–12. [CrossRef] [PubMed]
36. Powsner, S.M.; Costa, J.; Homer, R.J. Clinicians are from Mars and pathologists are from Venus. *Arch. Pathol. Lab. Med.* **2000**, *124*, 1040–1046. [CrossRef] [PubMed]
37. Leslie, K.O.; Rosai, J. Standardization of the surgical pathology report: Formats, templates, and synoptic reports. *Semin. Diagn. Pathol.* **1994**, *11*, 253–257. [PubMed]
38. Williams, C.L.; Bjugn, R.; Hassell, L. Current status of discrete data capture in synoptic surgical pathology and cancer reporting. *Pathol. Lab. Med. Int.* **2015**. [CrossRef]
39. Ellis, D.W. Surgical pathology reporting at the crossroads: Beyond synoptic reporting. *Pathology* **2011**, *43*, 404–409. [CrossRef] [PubMed]
40. Ellis, D.W.; Srigley, J. Does standardised structured reporting contribute to quality in diagnostic pathology? The importance of evidence-based datasets. *Virchows Arch. Int. J. Pathol.* **2016**, *468*, 51–59. [CrossRef] [PubMed]
41. Qu, Z.; Ninan, S.; Almosa, A.; Chang, K.G.; Kuruvilla, S.; Nguyen, N. Synoptic reporting in tumor pathology: Advantages of a web-based system. *Am. J. Clin. Pathol.* **2007**, *127*, 898–903. [CrossRef]
42. Baranov, N.S.; Nagtegaal, I.D.; van Grieken, N.C.T.; Verhoeven, R.H.A.; Voorham, Q.J.M.; Rosman, C.; van der Post, R.S. Synoptic reporting increases quality of upper gastrointestinal cancer pathology reports. *Virchows Arch.* **2019**, *475*, 255–259. [CrossRef]
43. Bitter, T.; Savornin-Lohman, E.; Reuver, P.; Versteeg, V.; Vink, E.; Verheij, J.; Nagtegaal, I.; Post, R. Quality Assessment of Gallbladder Cancer Pathology Reports: A Dutch Nationwide Study. *Cancers* **2021**, *13*, 2977. [CrossRef]
44. Casparie, M.; Tiebosch, A.; Burger, G.; Blauwgeers, H.; Van de Pol, A.; Van Krieken, J.; Meijer, G. Pathology databanking and biobanking in the Netherlands, a central role for PALGA, the nationwide histopathology and cytopathology data network and archive. *Anal. Cell. Pathol.* **2007**, *29*, 19–24. [CrossRef]
45. Professionals, S.O. EAU Guidelines: Prostate Cancer. Available online: https://uroweb.org/wp-content/uploads/EAU-EANM-ESUR-ESTRO-SIOG-Guidelines-on-Prostate-Cancer-2019-1.pdf (accessed on 20 October 2021).
46. Sobin, L.H.; Gospodarowicz, M.K.; Wittekind, C. *TNM Classification of Malignant Tumours*, 7th ed.; Wiley-Blackwell: Chicester, UK, 2009; p. 332.
47. Brierley, J.D.; Gospodarowicz, M.K.; Wittekind, C. *TNM Classification of Malignant Tumours*, 8th ed.; Wiley: Hoboken, NJ, USA, 2016.
48. Bertero, L.; Massa, F.; Metovic, J.; Zanetti, R.; Castellano, I.; Ricardi, U.; Papotti, M.; Cassoni, P. Eighth Edition of the UICC Classification of Malignant Tumours: An overview of the changes in the pathological TNM classification criteria-What has changed and why? *Virchows Arch. Int. J. Pathol.* **2018**, *472*, 519–531. [CrossRef]
49. Pölsterl, S. scikit-survival: A Library for Time-to-Event Analysis Built on Top of scikit-learn. *J. Mach. Learn. Res.* **2020**, *21*, 1–6.
50. Davidson-Pilon, C.; et al. lifelines - Survival analysis in Python. *Zenodo* **2019**, *4*, 1317. [CrossRef]
51. Koyasu, S.; Nishio, M.; Isoda, H.; Nakamoto, Y.; Togashi, K. Usefulness of gradient tree boosting for predicting histological subtype and EGFR mutation status of non-small cell lung cancer on 18 F FDG-PET/CT. *Ann. Nucl. Med.* **2020**, *34*, 49–57. [CrossRef] [PubMed]
52. Li, Y.; Chen, T.; Chen, T.; Li, X.; Zeng, C.; Liu, Z.; Xie, G. An Interpretable Machine Learning Survival Model for Predicting Long-term Kidney Outcomes in IgA Nephropathy. In Proceedings of the AMIA Annual Symposium, Online, 14–18 November 2020; Volume 2020, p. 737.

53. Sheridan, R.P.; Wang, W.M.; Liaw, A.; Ma, J.; Gifford, E.M. Extreme Gradient Boosting as a Method for Quantitative Structure-Activity Relationships. *J. Chem. Inf. Model.* **2016**, *56*, 2353–2360. [CrossRef]
54. Harrell, F.E.; Califf, R.M.; Pryor, D.B.; Lee, K.L.; Rosati, R.A. Evaluating the yield of medical tests. *JAMA* **1982**, *247*, 2543–2546. [CrossRef] [PubMed]
55. Harrell, F.E., Jr.; Lee, K.L.; Mark, D.B. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat. Med.* **1996**, *15*, 361–387. [CrossRef]
56. Haibe-Kains, B.; Desmedt, C.; Sotiriou, C.; Bontempi, G. A comparative study of survival models for breast cancer prognostication based on microarray data: Does a single gene beat them all? *Bioinformatics* **2008**, *24*, 2200–2208. [CrossRef]
57. Linardatos, P.; Papastefanopoulos, V.; Kotsiantis, S. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy* **2021**, *23*, 18. [CrossRef]
58. Lundberg, S.M.; Nair, B.; Vavilala, M.S.; Horibe, M.; Eisses, M.J.; Adams, T.; Liston, D.E.; Low, D.K.W.; Newman, S.F.; Kim, J.; et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat. Biomed. Eng.* **2018**, *2*, 749–760. [CrossRef]
59. Lundberg, S.M.; Erion, G.G.; Lee, S.I. Consistent individualized feature attribution for tree ensembles. *arXiv* **2018**, arXiv:1802.03888.
60. Shapley, L.S. A value for n-person games. *Contrib. Theory Games* **1953**, *2*, 307–317.
61. Hosmer, D.W., Jr.; Lemeshow, S.; Sturdivant, R.X. *Applied Logistic Regression*; John Wiley & Sons: Hoboken, NJ, USA, 2013; Volume 398.
62. Nicolò, C.; Périer, C.; Prague, M.; Bellera, C.; MacGrogan, G.; Saut, O.; Benzekry, S. Machine Learning and Mechanistic Modeling for Prediction of Metastatic Relapse in Early-Stage Breast Cancer. *JCO Clin. Cancer Inform.* **2020**, *4*, 259–274. [CrossRef]
63. Kim, D.W.; Lee, S.; Kwon, S.; Nam, W.; Cha, I.H.; Kim, H.J. Deep learning-based survival prediction of oral cancer patients. *Sci. Rep.* **2019**, *9*, 6994. [CrossRef] [PubMed]
64. Du, M.; Haag, D.G.; Lynch, J.W.; Mittinty, M.N. Comparison of the Tree-Based Machine Learning Algorithms to Cox Regression in Predicting the Survival of Oral and Pharyngeal Cancers: Analyses Based on SEER Database. *Cancers* **2020**, *12*, 2802. [CrossRef]
65. Huang, Y.; Chen, H.; Zeng, Y.; Liu, Z.; Ma, H.; Liu, J. Development and Validation of a Machine Learning Prognostic Model for Hepatocellular Carcinoma Recurrence After Surgical Resection. *Front. Oncol.* **2021**, *10*, 3327. [CrossRef] [PubMed]
66. Perera, M.; Tsokos, C. A Statistical Model with Non-Linear Effects and Non-Proportional Hazards for Breast Cancer Survival Analysis. *Adv. Breast Cancer Res.* **2018**, *07*, 65–89. [CrossRef]
67. Nagpal, C.; Sangave, R.; Chahar, A.; Shah, P.; Dubrawski, A.; Raj, B. Nonlinear Semi-Parametric Models for Survival Analysis. *arXiv* **2019**, arXiv:1905.05865.
68. Roshani, D.; Ghaderi, E. Comparing Smoothing Techniques for Fitting the Nonlinear Effect of Covariate in Cox Models. *Acta Inform. Med.* **2016**, *24*, 38–41. [CrossRef] [PubMed]
69. Abedian, S.; Sholle, E.T.; Adekkanattu, P.M.; Cusick, M.M.; Weiner, S.E.; Shoag, J.E.; Hu, J.C.; Campion, T.R., Jr. Automated Extraction of Tumor Staging and Diagnosis Information From Surgical Pathology Reports. *JCO Clin. Cancer Inform.* **2021**, *5*, 1054–1061. [CrossRef]
70. Miller, T. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.* **2019**, *267*, 1–38. [CrossRef]
71. Payrovnaziri, S.N.; Chen, Z.; Rengifo-Moreno, P.; Miller, T.; Bian, J.; Chen, J.H.; Liu, X.; He, Z. Explainable artificial intelligence models using real-world electronic health record data: A systematic scoping review. *J. Am. Med. Inform. Assoc.* **2020**, *27*, 1173–1185. [CrossRef]