Proceedings of Machine Learning Research 152:1–20, 2021 Conformal and Probabilistic Prediction and Applications

A non–conformity approach towards post-prostatectomy metastasis estimation using a multicenter prostate cancer database

Christos Chatzichristos CCHATZIC@ITS.JNJ.COM
CCHATZIC@ESAT.KULEUVEN.BE
Clinical Innovation Research & Development
Janssen Pharmaceutica, Beerse, Belgium
Dept. of Electrical Engineering, STADIUS Center
KU Leuven, Leuven, Belgium

Jose-Felipe Golib-Dzib JGOLIBDZ@ITS.JNJ.COM
Clinical Innovation Research & Development
Janssen Pharmaceutica, Madrid, Spain

Andries Clinckaert ANDRIES.CLINCKAERT@UZLEUVEN.BE
Department of Urology, University Hospitals Leuven
Leuven, Belgium

Wouter Everaerts WOUTER.EVERAERTS@UZLEUVEN.BE
Department of Urology, University Hospitals Leuven
Leuven, Belgium
Dept. of Development and Regeneration, KULeuven
Leuven, Belgium

Maarten De Vos MAARTEN.DEVOS@KULEUVEN.BE
Dept. of Electrical Engineering, STADIUS Center
KU Leuven, Leuven, Belgium

Martine Lewi MLEWI@ITS.JNJ.COM
Clinical Innovation Research & Development
Janssen Pharmaceutica, Beerse, Belgium

Editor: Lars Carlsson, Zhiyuan Luo, Giovanni Cherubin, and Khuong An Nguyen

Abstract

Prostate cancer is among the most common type of cancer in men worldwide. Despite the use of clinical indicators, as part of simple rule-based strategies, stratifying patients diagnosed with prostate cancer into risk groups to reliably reflect oncological prognosis remains challenging. Machine Learning (ML) offers the possibility to develop estimation models based on routinely evaluated patient or tumor characteristics. In the present study, the estimation of metastasis in prostate patients after primary treatments (radical prostatectomy) with the aid of Support Vector Machines (SVMs) and Conformal Predictors (CP) was evaluated. We show that the use of ML models can complement classical statistical approaches. Moreover, the application of CP, on top of an underlying ML model, renders a probabilistic outcome that combines the simplicity of a clinical indicator with the precision

of a ML approach. The TriNetX Research Network, an electronic health records database with datasets from several United States health care organizations, was used in this study. This approach can be further adapted to support clinical decision making in prostate and other types of cancer.

Keywords: Metastatic prostate cancer, conformal predictors, real world data, machine learning, supervised classification, risk stratification.

1. Introduction

Prostate cancer (PCa) is the second most frequent cancer diagnosis in men worldwide accounting for 25.9% of all new cancers in 2020 and was the fourth most common in the general population (Ferlay et al., 2018; Sung et al., 2021).

PCa is a heterogeneous disease with many differences in clinical evolution, ranging from indolent tumors to lethal castration-resistant PCa (Testa et al., 2019). A wide variety of treatment options is available, and treatment choice is notoriously complex and driven by the assessment of a favorable trade-off between treatment risks (e.g., erectile dysfunction and incontinence) and benefits (cancer control), speed of return to routine activities and the long-term impact on health-related quality of life. Sub-optimal prognosis of possible outcomes following a particular treatment is therefore common and often results in over-treatment of indolent disease or under-treatment of aggressive disease (Loeb et al., 2014; Bratt et al., 2015). This highlights the need of high quality, customized prognostic stratification strategies for clinical counselling and decision-making to tailor treatment to the needs of the individual PCa patient.

To date, patient risk stratification and outcomes assessments hold prominent places in therapeutic decision making of localized PCa. Risk stratification aims to assess the lethality of PCa to help determine choice of treatment, whilst health outcome assessments help to determine the efficacy of the therapies prescribed (Clinckaert et al., 2021).

Several risk stratification systems for patients with PCa have been proposed (Zelic et al., 2020; D'Amico et al., 1998). These stratification systems are easy to obtain provided that the required clinical-pathological variables (more details are given in Sec. 3.2) are available. Given the seminal work in Pound et al. (1999), these variables have gained more attention and are frequently found as part of data collected routinely during PCa patient follow up. For example, recently published online tools (MDCalc, 2021; MSKCC, 2021) are available for patient stratification.

However, data availability remains a primary obstacle for use of these stratification systems at scale across different healthcare settings. Moreover, data need to be standardized to a common terminology due to the variety of clinical procedures and medical coding systems utilized differently across institutions. Additionally, little is known about how to estimate the inherent risk associated with misclassifying PCa patients during risk stratification as there is no probabilistic framework to account for these uncertainties. Finally, the stratification systems are supervised methods, meaning that the set of variables used in the stratification is known in advance. Thus, there is no exploration regarding the potential benefit(s) that incorporating additional variables found in electronical health records (EHRs) may provide.

Patient-level data, collected from EHRs, registries, or claims data, can be de-identified and organized using a standard terminology to enable clinical research, among other appli-

cations, across different healthcare organizations. For instance, TriNetX (TriNetX, 2021), as a global health research network, provides secured access to EHRs from approximately 17,000 patients from 16 healthcare organizations with clinical data linked to tumor registry data.

Conformal Predictors (CP) (Vovk et al., 2005) offers a theoretical framework to generate probabilistic predictions (i.e., a prediction set, allowing for several values) whilst controlling for a defined error rate, a property denoted as validity (see Sec. 2.3 for more details). The prediction sets and the validity property brought by CP are used in several healthcare applications such as early drug development (Lapins et al., 2018; Alvarsson et al., 2021), allergy detection (Forreryd et al., 2018), or Alzheimer's Disease (Pereira et al., 2020). In oncology, CP are applied to early diagnosis of ovarian and breast cancer where validity, prediction sets, and efficiency metrics are key components for interpreting the results (Devetyarov et al., 2012).

Machine Learning (ML) refers to a broad range of algorithms that learn the pattern recognition of the underlying data structure in a dataset to perform intelligent predictions (Uddin et al., 2019). Generally, clinical practice is not predictable, but ML algorithms have a capability to learn from almost any data type and to identify data-driven clinical patterns in large patient populations (MacEachern and Forkert, 2021). Unlike statistical models, ML models make minimal assumptions about the data generation mechanism and can provide accurate prognosis even when the data are originally collected for purposes other than research and in the presence of complex non-linear interactions, as is the case of Real World Data (RWD).

In this paper, we propose a methodology to rank men at risk of developing metastatic PCa using a retrospective analysis of patient data with the potential to be used at scale across different healthcare institutions. To mitigate the risk of data unavailability, we use a comprehensive PCa dataset from TriNetX (see Sec. 3 for details). To complement existing stratification systems, we incorporate CP as a probabilistic framework to assist clinical interpretation by means of calibrated probabilities, prediction regions, and a way to estimate the risk of patient misclassification given a confidence level. We provide groundwork on the ML framework to computationally extract features from EHRs along the longitudinal follow up of each patient to generate the models, which can be extended to incorporate additional clinical variables that are currently absent in the existing methods that use a supervised approach.

2. Methods

In this section, we describe our problem definition, followed by an introduction to the ML method and the probabilistic framework used to address it.

2.1. Problem Definition

The problem consists of estimating, retrospectively, the occurrence of a metastasis diagnosis in PCa patients initially diagnosed as non-metastatic. This problem can modeled as a binary classification task, since there are only two classes (either metastatic or non-metastatic). The true class of the patient is determined based on retrospective data reported and collected in EHRs. If the patient's initial diagnosis of non-metastatic PCa is followed by a

subsequent diagnosis of metastatic PCa, the true class is defined as 'metastatic'. If there is no indication of diagnosis of metastasis in the patient's medical history and only the initial diagnosis of non-metastatic PCa is found, then the true class is defined as 'non-metastatic'.

In our use case, we are facing what is known as class imbalance. It is a more difficult type of classification task due to the imbalance on the diagnoses, as there are more cases of patients in the non-metastatic class. Therefore, ML classifiers will tend to have poor performance on the minority class (metastatic class) which is harder to estimate since there are less examples to learn from. We aim to estimate metastatic cases, so we will use performance metrics that are specific to the minority class or those that take into account class imbalance. For instance, the F1-score with average macro.

Let us assume that a set of $n \in \mathbb{N}$ patients is in scope for our study (with $n > 1$). Let $\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_{n-1}, y_{n-1})\}$ be a set of samples $(\boldsymbol{x}_i, y_i)$ for $i \in \{1, 2, \ldots, n-1\}$ used to build a predictive model (namely, a training set), where $\boldsymbol{x}_i$ is a vector of features constituting a clinical profile for patient i and y_i is the label of patient i. We define $y_i = 1$, if a diagnosis of metastatic disease is recorded after the earliest occurrence of a localized prostate cancer diagnosis (ICD-O, International Classification of Diseases for Oncology, code C61.9). Otherwise, $y_i = -1$.

Metastatic disease is defined as ICD-9 (International Coding System) codes: 196, 197, and 198, and their descendants, or ICD-10 codes: C77, C78, and C79, and their descendants, or having M1 status from the TNM coding system. Patients who do not fall into this metastatic disease definition are considered as non-metastatic. Our goal is to predict the label of a patient n, alien to the training of the predictive model.

2.2. Support Vector Machines (SVMs)

Support Vector Machines (SVMs) are widely used as classifiers in ML problems, for both binary or multi-class problems. The basic principle of the algorithm is the identification of the optimal margin between two classes in a certain feature space. SVMs find a decision hyperplane that separates two classes with a maximal margin. Any hyperplane can be written as a function of a set of data points $\mathbf{x}$, each one corresponding to a feature vector in the feature space:

$$f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b, \tag{1}$$

where $\mathbf{w}$ is the normal vector to the hyperplane and b a scalar offset.

The distance from this decision hyperplane to the closest data point determines the margin of the classifier. The goal of the learning phase is to find the optimal hyperplane such that the distance of the margins to the decision boundary is maximized. Hence, in a binary problem we can define two classes $+1, -1$ and two hyperplanes passing by the closest point of each class to the decision hyperplane. The function of these hyperplanes can be $h_1 : \mathbf{w}^T\mathbf{x} + b = 1$ and $h_2 : \mathbf{w}^T\mathbf{x} + b = -1$, any point lying above the first boundary will lie in the class c_1 with label 1 and any data point lying below the second one in the class c_2 with label -1, respectively. The distance of the two hyperplanes will be $\frac{2}{\|\mathbf{w}\|_2}$ (with $\|.\|_2$ denoting the Frobenius norm of the vector). The distance of a random data point x_i to the decision hyperplane, h, is given as:

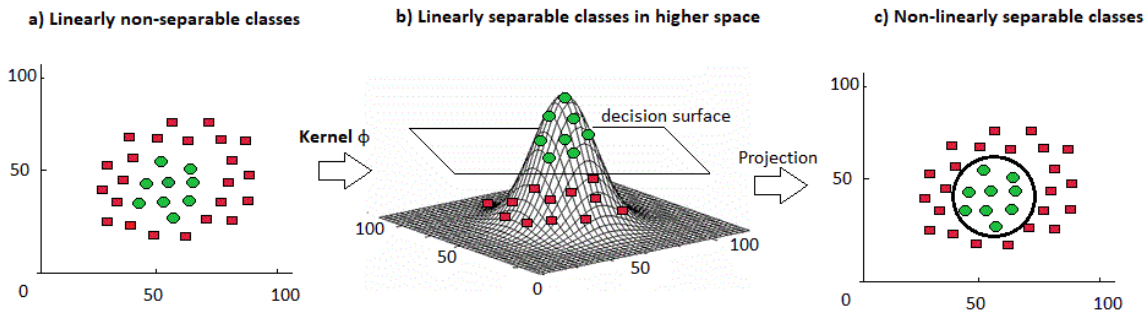

Figure 1: Finding the decision hyperplane in non-linearly separable data with the use of the 2D Gaussian kernel function. The data points of the two classes are depicted with red squares and green circles, respectively.

$$d_i^h = \frac{|f(x_i)|}{\|\mathbf{w}\|_2} \tag{2}$$

and the distance to the margin, m of one of the two classes $(+1, -1)$ is given as:

$$d_i^m = \frac{|f(x_i)| - 1}{\|\mathbf{w}\|_2}. \tag{3}$$

Hence, the maximization of the margin, is equivalent to the minimization of $\|\mathbf{w}\|_2$. SVM algorithm also identifies the support vectors, which lie either on or within the margin and are the data points that are solely responsible for the classification solution. For multi-class problems one decision function for each class and the one amongst those with the maximum value are selected.

The initial algorithm of SVM assumes that the data are linearly separable. In real case scenarios, this assumption is not always valid, due to non-linearities or noise in the underlying data generation process. non-linear versions of the SVM algorithm are also available. The most common approach for classifying non-linear data with the use of SVMs is through the adoption of kernels and the mapping of the data into a higher dimensional space, where they can be linearly classified (Fig. 1). The mapping of the original feature space into some higher-dimensional feature space, where the training set is separable, must be done in a way that any coherence information between the data points will be preserved. Kernel functions facilitate the computation of the inner product of any given two points in a suitable feature space, providing a notion of similarity, with low computational cost, even in high-dimensional spaces. There are different types of kernel functions that can be used such as linear, Gaussian, radial basis, polynomial, Laplacian, Bessel, sigmoid, and hyperbolic tangent kernels. The selection of the suitable kernel function is problem- and data-dependent and usually selected based on trial and error. The RBF, which is considered the most generalized form of kernelization and is one of the most widely used kernels due to its similarity to the Gaussian distribution, is defined for two points x_i and x_j, as:

$$K(x_i, x_j) = e^{-\gamma \|x_i - x_j\|_2^2}, \tag{4}$$

where γ is a parameter that sets the "spread" of the kernel.

2.3. Conformal Predictors

The intuitive notion of CP (Gammerman and Vovk, 2007; Vovk et al., 2005) is to quantify how 'unusual' a completion $z_n = (\boldsymbol{x}_n, y_n)$ is when compared to z_i from the training set ($i \neq n$). If the completion $(\boldsymbol{x}_n, 1)$ 'conforms' with those of the training set (all other patients in the study) while the alternative completion $(\boldsymbol{x}_n, 0)$ does not, then we can assign $\hat{y}_n = 1$ as the label of patient n. But in other situations, a patient can be assigned to both classes (metastatic and non-metastatic) or remains unassigned. For our use case, assignments to a single final status are preferred, but other cases could be informative as well.

To construct a CP, we need to define a non–conformity measure to capture the extent to which given patient data conforms to each class (e.g., metastatic, non-metastatic) of patients. The definition is dependent on the underlying algorithm and several alternatives can be available and play an important role in the outcome of CPs (Balasubramanian et al., 2009; Toccaceli et al., 2017). The non–conformity measure we use is:

$$\alpha_i^c = -cd(x_i), \tag{5}$$

where i refers to a given patient profile, c is the class (metastatic or non-metastatic), and d is the (signed) distance proportional to that of the profile x_i to the SVM hyperplane.

Then, the non-conformity measure, together with the underlying algorithm, are used to compute the p-values for each class. We use label-conditional (Mondrian) CP (Vovk et al., 2005), because our classification problem has class imbalance. The p-value is implemented as:

$$p(y) = \frac{A_y + 1}{B_y + 1} \tag{6}$$

where A_y is the count of calibration scores that are associated to class y and are equal or greater than the calibration score obtained in the test set for (x_i, y), B_y is the count of labels in the calibration set that are class y.

Finally, for a given a significance level, $\epsilon \in [0, 1]$, for each instance in the test set, we compute a prediction region which is a set of labels for which the true label is not one of its elements $\epsilon\%$ of the times or less. If the prediction region contains a single element then the prediction is a singleton, if there are more than one then the prediction is uncertain, and in those cases where the prediction region has no elements, we call the prediction empty. A more generic and formal definition can be found for instance in (Schafer and Vovk, 2008).

Validity is an important advantage for the applicability of CPs. As other confidence predictors, CPs can be used to generate a prediction region based on an underlying machine learning algorithm like SVM or random forest (RF) (Vovk et al., 2005; Schafer and Vovk, 2008). Validity means the frequency of errors that the underlying algorithm commits does not exceed a pre-specified chosen confidence level (Vovk et al., 2016). Validity relies on similar assumptions like those imposed by SVMs or on one that is even less restrictive, namely exchangeability (Vovk et al., 2003; Fedorova et al., 2012). In addition to validity,

the performance of a CP can be measured by means of efficiency defined as the proportion of single-label predictions at a given significance level.

Advances in CPs theory allow for targeted approaches to address challenges arising from large datasets or highly imbalanced ones in classification problems (Norinder et al., 2015; Toccaceli et al., 2017) with the flexibility to incorporate diverse underlying ML models (Balasubramanian et al., 2009; Lapins et al., 2018; Forreryd et al., 2018; Nouretdinov et al., 2010), including artificial neural networks (Papadopoulos et al., 2007; Toccaceli and Gammerman, 2017).

3. Data Release

The licensed dataset from TriNetX is harmonized to a standardized terminology[1]. Yet, since different coding systems are used across the healthcare institutions a mapping of the respective codes of interest was a prerequisite, also provided by TriNetX.

3.1. Cohort Definition

To define the cohort of patients in scope, we considered the following conjunct (all have to be fulfilled) inclusion and exclusion criteria:

- Inclusion criteria:
  - Any record linked to tumor registry data. For instance, any occurrence of codes for tumor annotation or tumor properties.
  - Gleason score registered at the time of diagnosis.
  - Radical Prostatectomy (RP) following a prostate cancer diagnosis.
  - PSA measures taken after RP and a minimum of 3 PSA values in the first year after RP and at least one PSA measurement afterwards.
- Exclusion criteria:
  - Metastasis at time of diagnosis.
  - Received radiation, chemotherapy, or hormonal therapy prior to RP (i.e., neoadjuvant therapy).

1. TriNetX is compliant with the Health Insurance Portability and Accountability Act (HIPAA), the US federal law which protects the privacy and security of healthcare data. TriNetX is certified to the ISO 27001:2013 standard and maintains an Information Security Management System (ISMS) to ensure the protection of the healthcare data it has access to and to meet the requirements of the HIPAA Security Rule. Any data displayed on the TriNetX Platform in aggregate form, or any patient level data provided in a data set generated by the TriNetX Platform, only contains de-identified data as per the de-identification standard defined in Section §164.514(a) of the HIPAA Privacy Rule. The process by which the data is de-identified is attested to through a formal determination by a qualified expert as defined in Section §164.514(b)(1) of the HIPAA Privacy Rule.

3.2. Feature Engineering

Different types of features have been proposed in the literature for the prediction of metastasis and biochemical recurrence (BCR) in patients with PCa. The features that we have identified and used in this analysis are of a different nature (endpoints, clinical characteristics, specific measurements) and include the following:

- **Gleason score (GS).** The Gleason scoring system is the most common PCa grading system used to determine the aggressiveness of PCa (net, 2020). The pathologist quantifies the differentiation of the cancer from the arrangement of the cells and glands obtained from the biopsy or RP. The Gleason Grades range from 3 to 5 (most differentiated tissue); the GS, which is a sum of a primary and secondary Gleason grade subsequently range from 6-10. The Gleason sum score is assessed differently from biopsy and RP (Egevad et al., 2002). In specimens taken from the biopsy, two different grades are summed up, a primary grade is given to describe the cells that make up the largest area of the tumor ("the most") and a secondary grade is referred to the Gleason grade which is the second worst in all specimens ("the remaining worst"). In specimens taken from RP, also two different grades are summed up, a primary grade is given to describe the cells that make up the largest area of the tumor ("the most") and a secondary grade, is referred to the Gleason grade which is the second most common in the entire specimen and is given to describe the cells of the next largest area ("the second most").

- **PSA Doubling Time (PSADT).** The doubling time of PSA for prostate carcinoma cells is faster than the growth of healthy tissue. Furthermore, the tumor cells contribute significantly more to the increase of PSA (Nowroozi et al., 2009). PSADT is the number of months it would take for PSA to double and has been used as an indicator of the presence or absence of malignancy (or metastasis after the removal of the initial tumor) (Pound et al., 1999). Two main methods are used for the estimation of the PSADT for every patient, with their difference being the way that the slope, b_i of PSA is computed. Best-Line Fit (BLF) method opts for fitting a least-squares regression line to the log-scale PSA measurements. The least squares regression line of a set of data points is computed by minimizing the sum of the offsets of the data points from the plotted line. As an alternative, the slope can be estimated by computing the difference between the First and Last log-scale Observations (FLO) of PSA in relation to the time interval between the two observations. FLO has been reported to be more robust in the presence of outliers (Svatek et al., 2006). However, if the outlier happens to be the first or last point of the PSA measurements then the PSADT metric becomes practically useless. To alleviate such cases, we propose a slightly updated version of the FLO method which computes the PSA slope between the mean of the first observations and the last two observations. Therefore, in this study, PSADT is:

$$\text{PSADT}_i = log(2)/b_i \tag{7}$$

where b_i is:

$$b_i = \frac{log(mean(\text{PSA}_{i,k_i}, \text{PSA}_{i,k_i-1})) - log(mean(\text{PSA}_{i,1}, \text{PSA}_{i,2}))}{mean(t_{i,k_i}, t_{i,k_i-1}) - mean(t_{i,1}, t_{i,2})}$$

with k_i being the last PSA value in time of patient i. The maximum value of PSADT is set to 100 months as in Pound et al. (1999). In cases that have negative values of PSA or $b_i = 0$ (and hence the division cannot be defined), PSADT is set to zero.

- **Time to BCR.** Time to BCR is defined as the time from surgery to the first measurement of PSA above 0.2 ng/ml [2]. To avoid the influence of outliers or wrong entries (which were common in our dataset) to the metric, we considered time of BCR as the time to the first of at least 2 consecutive PSA measurements above 0.2 ng/ml. The maximum BCR time is set to 5 years as in Pound et al. (1999).

- **Post-prostatectomy PSA.** A persistent PSA after RP (or following other forms of treatment) can potentially mean cancer has progressed and metastasized.

- **Age at the time of surgery.** We chose the earliest occurrence of a surgical procedure recorded as any of the following CPT codes: 55810, 55815, 55840, 55842, 55845, or an ICD-9 code of 60.5.

3.3. Sampling, Partitioning, and Folding

We split the data into (proper) training, calibration, and test sets using 60%, 20%, and 20% of the data release to enable an Inductive CP setup. The splits are stratified with respect to the label, to keep the class imbalance 1 to 3 in favor of the non-metastatic class. Thus, the minority class is the metastatic one. Variables are centered and scaled. Moreover, we generated 100 different samples to allow variance estimation on the performance metrics of the Inductive CP. In term, each training set was further divided into 3 folds for the ML setup. Thus, each fold corresponds to 20% of the data release. Folds are also stratified with respect to the label.

4. Results

Deriving a PCa dataset with pre-selected clinical variables from several healthcare organizations is facilitated using TriNetX research platform. We generate a data release from the TriNetX licensed dataset consisting of a single table of 204 PCa patient profiles (rows), and 7 attributes (columns) including a profile identifier (id001, id002, etc.), 5 clinical features (age, GS, time to BCR, PSADT, and PSA after surgery), and one label variable (metastatic status, +1 or -1). More details are found in Sec. 3.

In Fig. 2, the cohort diagram highlights the impact of subsequent selection criteria on data availability for analysis (the N value). Among patients diagnosed with PCa having tumor registry annotation, estimated in more than 17K cases (at the top of the diagram), only 6186 were initially diagnosed as non-metastatic PCa patients (and had a GS recorded at diagnosis), representing an impact (decrease in data availability) of 70%. Next, the impact of requiring patients having RP after diagnosis accounts for an additional drop of more than 5K cases (or 29%). The final selection criterion is about PSA measurements. The availability of PSA measurements as a lone criterion has a minor impact (less than

2. For subjects that exhibit abnormally high PSA following RP procedure, time to BCR was computed as the time from the first PSA value below to the first measurement of PSA above 0.2 ng/ml.

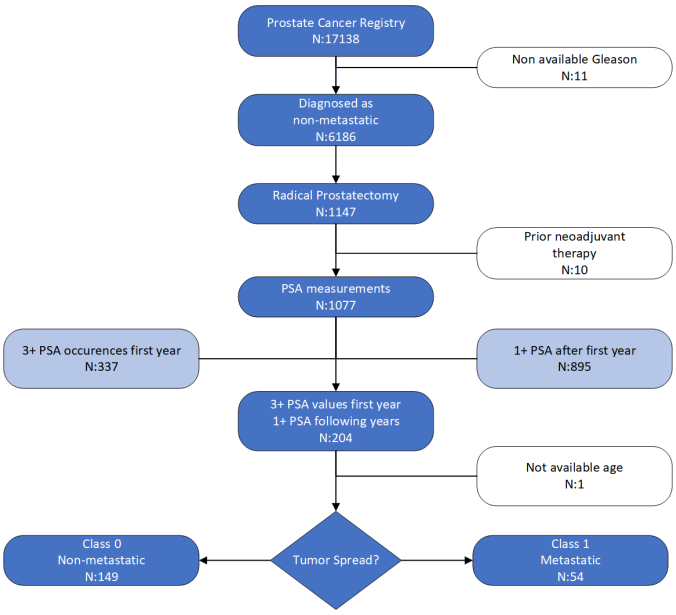

Figure 2: Cohort diagram. Dark blue boxes contain information about the patients that continue in the flow diagram of the cohort definition, light blue boxes offer additional information about the patients' cohort and the white boxes include information about patients that are excluded from the final cohort in this analysis. The rhombus shape represents a question.

1%) in regard to data availability. However, an additional criterion on PSA measurements accounts for a decrease in data availability from 1,077 cases to only 204 (i.e., an additional 5% decrease). Some additional cases are dismissed along the cohort diagram (shown in white boxes) due to missing data. Fulfillment of all required inclusion and exclusion criteria, provide a final selection of 204 cases, from which 150 are class non-metastatic and 54 are class metastatic.

4.1. Inductive Mondrian Conformal Predictors

To generate a scalable patient risk stratification methodology across multiple healthcare organizations, we applied Inductive Mondrian CPs on data release. The inductive setup allows for computational scalability and the Mondrian variant is used to account for class imbalance present in data release. The performance is evaluated on a series of 100 different rounds of training, calibrating, and testing, which uses 60%, 20%, and 20% of the data release, respectively. More details can be found at Sec. 3.3. The results are shown in Table 1.

Significance	Correctly Classified Metastatic	Correctly Classified Non-metastatic	Misclassified Metastatic	Misclassified Non-metastatic	Empty Predictions	Uncertain Predictions
1	0	0	0	0	0	100
5	4.7±4.4	0	0	2.6±3.7	0	93±7.4
10	9.7±4.6	13±14	2.2±2.8	6.4±5.8	0	69± 18
15	11±4.7	13±14	2.2±2.8	8.5±6.3	0	65± 18
20	14±4.6	28±16	4.5± 4	13±7.2	0.11±0.84	41±19
25	15±4.4	37±16	6.2±4.1	16±7.8	0.47±2.1	26±18
30	17±4.2	36±15	6±3.8	20±7.8	1.5±4.8	20±17

Table 1: Conformal predictors results for data release using SVM variants for different significance levels on 100 test sets. Figures represent averages and standard deviations over the best models, in percentages. For each sampling the best performing SVM (based on F1-weighted, ties were allowed) is used as the underlying algorithm.

4.2. Performance of ML models under the current data release

For training models, we use SVMs. Several variants are built to assess the relative performance between different kernels (linear or RBF) with or without correction for class imbalance, inspired by King and Zeng (2001). Thus, we explore 4 SVM variants:

- **SVM-linear.** Linear kernel with class weight 1 on metastatic class.
- **SVM-linear-balanced**. Linear kernel with class weight 3 on metastatic class.
- **SVM-RBF**. RBF kernel with class weight 1 on metastatic class.
- **SVM-RBF-balanced**. RBF kernel with class weight 3 on metastatic class.

Parameter tuning of models built on the training set is achieved using a brute-force parameter grid search. Model parameters (C for linear SVMs, and C and γ for non-linear SVMs) are tuned via grid search from the sets:

$$C \in \{0.25, 0.50, 0.75, 1.00, 5, 10, 15, \ldots, 100\} \quad \text{and}$$
$$\gamma \in \{0.01, 0.02, \ldots, 0.1, 0.2, 0.3\} \cup \{\text{“auto”}, \text{“scale”}\},$$

To select the best configuration of the tuned parameters, including kernel and class weight, a 3-fold cross-validation is used on the training set (60% of instances in data release). Results for the 4 SVM variants based on kernel selection and choice of class weight, are illustrated in Fig. 3. To assess the robustness of SVM predictions in the data release, we use a collection of 100 different samplings. The best performing models (based on f1-weighted), for each sampling, are carried over for subsequent phases of calibration and testing.

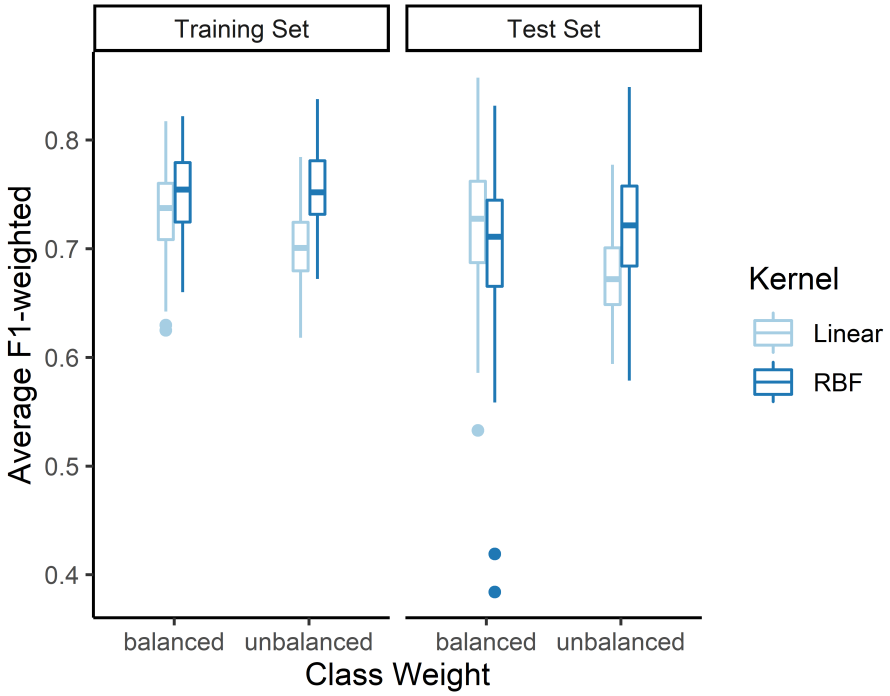

Figure 3: Performance of SVMs in training and test sets. Boxplots of the F1-score (weighted average) metric using 100 samplings. The larger the F1-score, the better the model performance is.

Using the same method as in Pound et al. (1999), where patient profiles composed of Gleason score, PSADT, and Time to BCR were used, we classify patient profiles from our study cohort into six risk groups or clusters. For each risk group a probability to develop metastasis is predefined as described in Fig. 4. To compare the performance of the suggested ML algorithms with the risk stratification every patient with a metastasis

probability higher than 0.5 (classification threshold set in our ML models) were assigned in the metastatic class, whether patients with probability lower than 0.5 were considered non-metastatic. Moreover, to contrast this methodology (denoted as thresholding in Table 2) against that of SVMs, we use the patient profiles from the test set to assess the performance of the classification which yields inferior results compared to those of SVMs.

Method	**mean** $\pm$ std	**minimum**	**maximum**
SVM-linear	0.676 $\pm$ 0.047	0.777	0.594
SVM-linear-balanced	0.72 $\pm$ 0.055	0.857	0.533
SVM-RBF	0.722 $\pm$ 0.053	0.849	0.579
SVM-RBF -balanced	0.703 $\pm$ 0.078	0.83	0.384
Thresholding	0.69 $\pm$ 0.135	0.83	0.57

Table 2: Performance of the underlying algorithms evaluated in the test set. The values used for thresholding were those reported in Pound et al. (1999) for the six different patient risk groups.

Implementation of data processing and machine learning was performed using `Python` 3.8 and `RStudio Server` v. 1.3.1076.1 (RStudio Team, 2020), for visualization. Open-source libraries were employed such as `NumPy` (Harris et al., 2020), `pandas` (McKinney, 2010; pandas development team, 2020), and `scikit-learn` (Pedregosa et al., 2011).

4.3. CPs for risk patient stratification

To illustrate an application of CPs and ML on risk stratification for PCa patients, we use an example with two scenarios: one for 'high' confidence to avoid misclassification of metastatic cases, and the second for 'moderate' confidence to minimize the number of uncertain predictions.

Let us assume that G^ϵ is the set of predictor regions $\Gamma_i^\epsilon = \{y \in Y : p_i^y > \epsilon\}$, a given test set (having labels Y) of the results presented in Table 1, and p_i^y is the p-value obtained from the Inductive Mondrian CP as prepared in this study.

Our example uses two complementary steps. First, we rank patient profiles according to a risk stratification system, like in Pound et al. (1999), where patient profiles consist of GS, PSADT, and Time to BCR. We classify patient profiles from our study cohort into six risk groups. For each risk group, a probability to develop metastasis is predefined as described in Fig. 4. Based on this predefined probability, the (ascending) order of the cluster numbers is: 1,2,3,5,6,4. Second, we use a CP classifier to generate an alternative ranking of the same patient profiles based on their probability of being classified as metastatic at the given significance level. The final ranking is obtained adding up the two rankings previously described (without ML and with ML support) for those cases where the prediction region has a single element. Otherwise for empty predictions or uncertain predictions, only the output of the first ranking is used.

For example, in a scenario where 'high' confidence is desired (say, Table 1, line where $\epsilon = 0.05$), we expect that out of the 41 cases of the training set, around 97% will be

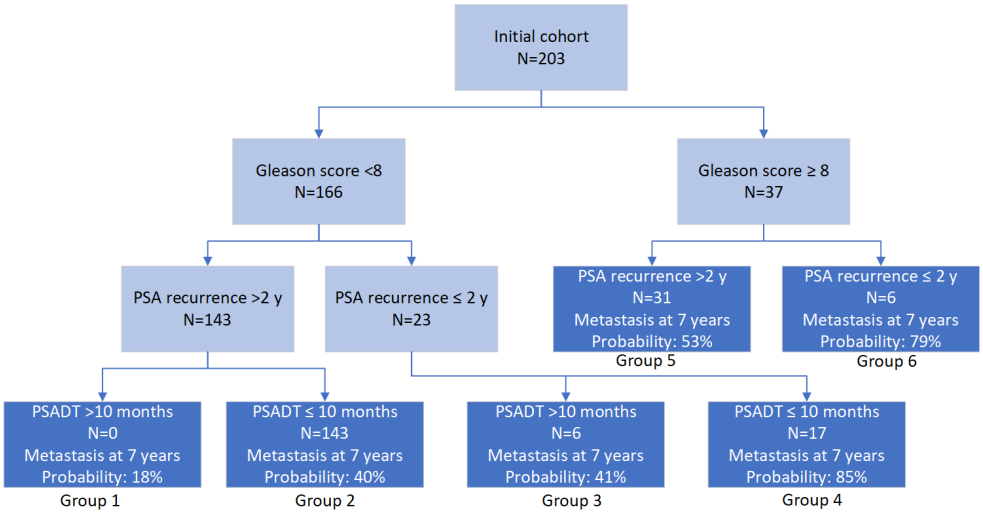

Figure 4: Example of patient classification based on risk groups as defined in Pound et al. (1999). Data availability (indicated by N) is shown in each box and it is derived using the data release (see Sec. 3).

uncertain. Thus, only around 3% may be used to complement the existing ranking at this level of confidence. If we were to allow a more 'moderate' confidence, for instance $\epsilon = 0.20$, the number of expected empty or uncertain predictions drops to around 41%, allowing more contribution from the CP ranking knowing the maximum level of errors that may be introduced.

This combination of rankings may incorporate more granularity because p-value ranking further stratifies patients even when assigned to the same risk group. CP framework allows us to tune the trade-off between gaining increased granularity for risk stratification and the introduction of potentially incorrect estimations for the risk of developing metastasis. The more efficient the CP, the more precise the granularity. This is demonstrated by an increased confidence in risk-stratification rank due to less potential misclassifications.

5. Discussion and Conclusions

In this paper, we present different SVM models which can be combined with CPs in order to estimate metastasis for patients who underwent RP following a diagnosis of non-metastatic PCa. To the best of our knowledge, this is the first report on application of CP as a more precise model in the retrospective estimation of metastasis in patients with PCa.

Under the current data release, no single combination of parameter, or SVM variant, in the ML models outperforms all others. As can be seen in Fig. 3 (left panel) the RBF SVMs (both balanced and unbalanced) outperform the linear kernels in the training set. But as we can see in the test set (right panel of 3) and in Table 2 the performance of the RBF is not as good in the test set. This implies that the dataset is very heterogeneous

and the tuning in the training set do not provide (always) optimal parameters for the test. This is the reason also that the linear kernel (which has less parameters) generalizes better in the test set. Hence, with the current dataset an optimal algorithm can not be chosen. Still some indications for modeling choices that could be beneficial for future work can be derived. For instance, accounting for class imbalance had a positive effect in performance of the underlying algorithm, especially in the linear models, as showed in Fig. 3. Furthermore, we can see that the SVM models slightly outperform the simple thresholding approach used in Pound et al. (1999), which suffers from high standard deviation pointing out also the heterogeneity of the dataset.

Despite the contribution from several healthcare organizations through a global hospital network, data completeness remains a critical requirement to improve the performance of the proposed model. As shown in Fig. 2, in our application use case, the final selection of 204 patients represents a small subset of the initial cohort. The major data drop comes from missing features related to tumor annotation. Future work should focus on expanding the study population, either based on the same data set or applying the analysis to other datasets to compare results and performance. Furthermore, methodologies for handling bias (Cave et al., 2019; Berger et al., 2016), outlier detection (Estiri et al., 2019; Hauskrecht et al., 2016), and data missingness (Beaulieu-Jones et al., 2018; Cave et al., 2019) shall be explored in an attempt to avoid the decrease of population due to incomplete data.

Additional data curation processes designed to increase data quality can help to alleviate the impact of data unavailability. There is a trade-off between data inclusion and data heterogeneity. In the currently used dataset, we can choose a broader definition for prostatectomy as opposed to only RP. This impacts the range and variability of the PSA values, which in turn may require a more robust definition of PSADT or basal PSA after surgery, to account for more scenarios in the longitudinal recollection of PSA measurements per patient. Even when only codes of RP are selected, errors in the codification may not be excluded and can be associated with outliers in PSA values. For instance, an abnormally high PSA following a procedure coded as RP, may be indicative of a partial rather than a RP.

The inclusion of additional data modalities, such as medical imaging and genomic markers, offers a promising roadmap for improvements in future data releases. Datasets enriched with augmented Real Word Data (RWD) (Cave et al., 2019; Sarker, 2021) are a promising line of research. In addition to lab measurements and diagnosis codes, inclusion of additional type of data, such as genomic markers (Cooperberg et al., 2015; Spratt et al., 2018), imaging data (Yu et al., 2016), or electronic Patient-Reported Outcomes (ePROs) (Sellers et al., 2016) can help us to advance our understanding of disease progression in prostate cancer (Rawla, 2019).

The efficiency of the proposed CP needs to be improved in order to best contribute to our use case in the clinical setting. However, the CP framework offers a way to measure the impact of those improvements to meet a certain goal. For instance, if we aim to have 99% confidence in the correct classification of metastatic PCa patients (top row in Table 1), it is not possible using the current data release because all predictions are uncertain. We may continue working on improving data curation or improved methods of risk stratification until more cases are correctly classified and there is less uncertainty and more confidence

in the ranking and diagnoses. This level of precision may be possible, and may be tracked along the way, due to the CP framework.

This is in line with previous applications in clinical use cases where the application of label-conditional CP is beneficial (Devetyarov et al., 2012). Indeed, the focus of this study is foundational in terms of the underlying algorithm and the feature engineering and other approaches in ML such as artificial neural networks that can be used to automate the feature extraction process (Poulakis et al., 2004; Papadopoulos et al., 2007).

We give an example of the use of CP to gain granularity in the patient risk stratification (see Sec. 4.3). With the advent of more efficient CP classifiers, misclassification will be reduced leading to greater prognostic precision and diagnostic confidence. Improved methods of risk stratification and rank may then support personalized medicine, meaning that patients will not only be stratified into risk groups, but the improved granularity of the method will lead to ranking at an individual level.

In conclusion, our results signal a favorable application of the proposed framework in a retrospective estimation of metastasis in PCa patients. Our results also indicate the possibility to design a patient-specific risk stratification strategy based on such estimation. Further improvements on the highlighted modeling configuration, together with data enhancement strategies, are needed in order to deploy it as a decision support tool in clinical practice. In this analysis, we define the foundational modeling choices for future analyses where validity and efficacy of CPs or other calibration methodologies can be explored to assist healthcare providers in efficient and precise medical management of patients with PCa. The CP methodology has potential for use in ranking patients with other types of pathologies documented in the EHR systems.

Acknowledgments

The authors would like to thank the TriNetX Support team and Dr. Wouter Botermans, Janssen Pharmaceutica, for fruitful discussions around the definition of the cohort used in the analysis.

References

J. Alvarsson, Arvidsson McShane S., Norinder U., and O. Spjuth. Predicting with confidence: Using conformal prediction in drug discovery. *J. Pharmaceutical Sciences*, 110(1): 42–49, Jan. 2021.

V.N. Balasubramanian, R. Gouripeddi, S. Panchanathan, J. Vermillion, A. Bhaskaran, and R.M. Siegel. Support vector machine based conformal predictors for risk of complications following a coronary drug eluting stent procedure. In *2009 Computers in Cardiology Conference (CinC)*, Park City, Utah, USA., Sep. 2009.

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

M. Berger, M. Curtis, G. Smith, J. Harnett, and A. Abernethy. Opportunities and challenges in leveraging electronic health record data in oncology. *Future oncology*, 12(10), May 2016.

O. Bratt, Y. Folkvaljon, H. M. Eriksson, O. Akre, S. Carlsson, L. Drevin, F. I. Lissbrant, D. Makarov, S. Loeb, and P. Stattin. Undertreatment of men in their seventies with high-risk nonmetastatic prostate cancer. *Eur Urol.*, 68(1):53–58, Jul. 2015.

A. Cave, X. Kurz, and P. Arlett. Real-world data for regulatory decision making: Challenges and possible solutions for europe. *Clinical Pharmacology & Therapeutics*, 106(1):36–39, Apr. 2019.

A. Clinckaert, G. Devos, E. Roussel, and S. Joniau. Risk stratification tools in prostate cancer, where do we stand? *Transl Androl Urol.*, 10(1):12–18, Jan. 2021.

M. R. Cooperberg, E. Davicioni, A. Crisan, R. B. Jenkins, M. Ghadessi, and R. J. Karnes. Combined value of validated clinical and genomic risk stratification tools for predicting prostate cancer mortality in a high-risk prostatectomy cohort. *European urology*, 67(2): 326–333, Feb. 2015.

A. V. D'Amico, R. Whittington, S.B. Malkowicz, D. Schultz, K. Blank, G.A. Broderick, J.E. Tomaszewski, A. Renshaw, I. Kaplan, C.J. Beard, and A. Wein. Biochemical outcome after radical prostatectomy, external beam radiation therapy, or interstitial radiation therapy for clinically localized prostate cancer. *JAMA*, 280(11):969–974, Sep. 1998.

D. Devetyarov, I. Nouretdinov, B. Burford, S. Camuzeaux, A. Gentry-Maharaj, A. Tiss, Celia J. Smith, Z. Luo, A. Chervonenkis, R. Hallett, V. Vovk, M. Waterfield, R. Cramer, J. Timms, J. Sinclair, U. Menon, I. Jacobs, and A. Gammerman. Conformal predictors in early diagnostics of ovarian and breast cancers. *Prog. Artificial Intell.*, 1:245–257, 2012.

L. Egevad, T. Granfors, L. Karlberg, A. Bergh, and P. Stattin. Prognostic value of the gleason score in prostate cancer. *BJU International*, 89(6):538–542, 2002.

H. Estiri, J.G. Klann, and S.N. Murphy. A clustering approach for detecting implausible observation values in electronic health records data. *BMC. Med. Inform. Decis.*, 19, Jul. 2019.

V. Fedorova, A. Gammerman, I. Nouretdinov, and V. Vovk. Plug-in martingales for testing exchangeability on-line. In *Int. Conf. Machine Learning*, pages 1639–1646, Edinburg, Scottland, Jun. 2012.

J. Ferlay, M. Colombet, I. Soerjomataram, T. Dyba, G. Randi, M. Bettio, A. Gavin, O. Visser, and F. Bray. Cancer incidence and mortality patterns in europe: Estimates for 40 countries and 25 major cancers in 2018. *Eur. J. Cancer*, 103:356–387, Nov. 2018.

A. Forreryd, U. Norinder, T. Lindberg, and M. Lindstedt. Predicting skin sensitizers with confidence - using conformal prediction to determine applicability domain of GARD. *Toxicol In Vitro*, 48:179–187, Apr. 2018.

A. Gammerman and V. Vovk. Hedging predictions in machine learning. *Comput. J.*, 50(2): 151–163, 2007.

C. Harris, J. Millman, S. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, and et. al Smith, N.J. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020.

M. Hauskrecht, I. Batal, C. Hong, Q. Nguyen, G. F. Cooper, S. Visweswaran, and G. Clermont. Outlier-based detection of unusual patient-management actions: An icu study. *J. Biomed. Inf.*, 64:211–221, Dec. 2016.

G. King and L. Zeng. Logistic regression in rare events data. *Political Analysis*, 9:137–163, Feb. 2001.

M. Lapins, S. Arvidsson, S. Lampa, A. Berg, W. Schaal, Alvarsson J., and O. Spjuth. A confdence predictor for logD using conformal regression and a support-vector machine. *J. Cheminform.*, 10(17), Apr. 2018.

S. Loeb, M.A. Bjurlin, J. Nicholson, T.L. Tammela, D.F. Penson, H.B. Carter, P. Carroll, and R. Etzioni. Overdiagnosis and overtreatment of prostate cancer. *Eur Urol.*, 65(6): 1046–1055, Jun. 2014.

S. J. MacEachern and N. D. Forkert. Machine learning for precision medicine. *Genome*, 64 (4):416–425, Feb. 2021.

W. McKinney. Data Structures for Statistical Computing in Python. In *Proc. Python in Science Conference*, pages 56 – 61, Austin, Texas, USA, Jun. 2010.

MDCalc. PSA Doubling Time (PSADT) calculator. Calculates rate of PSA doubling in prostate cancer (correlates with survival), 2021. URL https://www.mdcalc.com/psa-doubling-time-psadt-calculator.

Nomograms MSKCC. Prostate cancer Nomograms: PSA doubling time, 2021. URL http://nomograms.mskcc.org/Prostate/PsaDoublingTime.aspx.

Cancer net. Prostate cancer: Stages and grades, 2020. URL https://www.cancer.net/cancer-types/prostate-cancer/stages-and-grades.

U. Norinder, L. Carlsson, S. Boyer, and M. Eklund. Introducing conformal prediction in predictive modeling for regulatory purposes. a transparent and flexible alternative to applicability domain determination. *Regulatory Toxicology and Pharmacology*, 71(2): 279–284, Mar. 2015.

I. Nouretdinov, G. Li, A. Gammerman, and Z. Luo. Application of conformal predictors to tea classification based on electronic nose. In *Artificial Intelligence Applications and Innovations*, Berlin, Heidelberg, 2010.

M.R. Nowroozi, S. Zeighami, M. Ayati, H. Jamshidian, A.R. Ranjbaran, A. Moradi, and F. Afsar. Prostate-specific antigen doubling time as a predictor of gleason grade in prostate cancer. *Urol J.*, 6(1):27–30, Nov. 2009.

The pandas development team. pandas-dev/pandas: Pandas, Feb. 2020. URL https://doi.org/10.5281/zenodo.3509134.

H. Papadopoulos, V. Vovk, and A. Gammerman. Conformal prediction with neural networks. In *IEEE Tools with Artificial Intelligence(ICTAI 2007)*, volume 2, Patras, Greece, 2007.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

T. Pereira, S. Cardoso, M. Guerreiro, A. Mendonça, and S. Madeira. Targeting the uncertainty of predictions at patient-level using an ensemble of classifiers coupled with calibration methods, Venn-ABERS, and conformal predictors: A case study in AD. *J. Biomedical Inf.*, 101, 2020.

V. Poulakis, U. Witzsch, R. De Vries, V. Emmerlich, M. Meves, H.M. Altmannsberger, and E. Becht. Preoperative neural network using combined magnetic resonance imaging variables, prostate-specific antigen, and gleason score for predicting prostate cancer biochemical recurrence after radical prostatectomy. *Urology*, 64(6):1165–1170, Dec. 2004.

C.R. Pound, A. W. Partin, M. Eisenberger, D. Chan, J. Pearson, and P. C. Walsh. Natural history of progression after PSA elevation following radical prostatectomy. *JAMA*, 281 (17):1591–1597, May 1999.

Prashanth Rawla. Epidemiology of prostate cancer. *World J. Onc.*, 10(2):63–89, Apr. 2019.

RStudio Team. *RStudio: Integrated Development Environment for R.* RStudio, PBC, Boston, MA, 2020. URL http://www.rstudio.com/.

I. H. Sarker. Machine learning: Algorithms, real-world applications and research directions. *Sn. COMPUT. SCI.*, 2:28–32, Mar. 2021.

G. Schafer and V. Vovk. A tutorial on conformal prediction. *J. Mach. Learning*, 9:371–421, Aug. 2008.

L. Sellers, A. Savas, R. Davda, K. Ricketts, and H. Payne. Patient-reported outcome measures in metastatic prostate cancer. *Trends in Urology & Men's Health*, 7(1):28–32, Jan. 2016.

D. E. Spratt, J. Zhang, M. Santiago-Jiménez, R. T. Dess, and J. et. al. Davis. Development and validation of a novel integrated clinical-genomic risk group classification for localized prostate cancer. *Journal of Clinical Oncology*, 36(6):581–590, Jan. 2018.

H. Sung, J.s Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer J. for Clinicians*, 71(3):209–249, Feb. 2021.

R. S. Svatek, M. Shulman, P. K. Choudhary, and E. Benaim. Critical analysis of prostate-specific antigen doubling time calculation methodology. *Cancer*, 106(5):1047–1053, May 2006.

U. Testa, G. Castelli, and E. Pelosi. Cellular and molecular mechanisms underlying prostate cancer development: Therapeutic implications. *Medicines*, 6(3), Jul. 2019.

P. Toccaceli and A. Gammerman. Combination of conformal predictors for classification. In *Proc. Workshop on Conformal and Probabilistic Prediction and Applications*, volume 60, pages 39–61, Stockholm, Sweden, 13–16 Jun 2017.

P. Toccaceli, I Nouretdinov, and A. Gammerman. Conformal prediction of biological activity of chemical compounds. *Annals of Mathematics and Artificial Intelligence*, 81(1):105–123, Oct 2017.

TriNetX. Real-world-data for the life sciences and healthcare, 2021. URL https://trinetx.com/.

S. Uddin, A. Khan, E. Hossain, and M. A. Moni. Comparing different supervised machine learning algorithms for disease prediction. *BMC Medical Informatics and Decision Making*, 19(1), Dec. 2019.

V. Vovk, I. Nouretdinov, and A. Gammerman. Testing exchangeability on-line. In *Proceedings of the 20th International Conference on Machine Learning*, ICML, pages 768–775, 2003.

V. Vovk, A. Gammerman, and Schafer G. *Algorithmic Learning in a Random World*. Springer, 2005.

V. Vovk, I. Fedorova, V.and Nouretdinov, and A. Gammerman. Criteria of efficiency for conformal prediction. In *Conformal and Probabilistic Prediction with Applications (COPA*, Cham, Apr. 2016.

KH. Yu, C. Zhang, G. Berry, B. A. Russ, C. Re, D. L. Rubin, and M. Snyder. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nat. Commun.*, 7(1):28–32, Aug. 2016.

R. Zelic, H. Garmo, D. Zugna, P. Stattin, L. Richiardi, O. Akre, and A. Pettersson. Predicting prostate cancer death with different pretreatment risk stratification tools: A head-to-head comparison in a nationwide cohort study. *Eur Urol.*, 77(2):180–188, Feb. 2020.