Technical report: Identification of factors guiding treatment decision in oncology by rapid data insights using AI and xAI - a pilot study on real-world data

Holger Ziekow, Norbert Marschner, Dunja Klein, Benjamin Kasenda, Nina Haug

Abstract

Real-world data on the treatment histories of patients in everyday care contain a large amount of latent knowledge which to date is almost only made available via publication with a considerable time lag and only in relation to specific issues. AI models can capture the knowledge contained in this kind of data and transfer it to new scenarios. We aim to develop an AI-based information tool that enables dynamic data exploration and analysis of real-world datasets on medical treatments. The purpose of the tool is to support oncologists in their decision-making process through a system that is trained with prospectively documented real-world data on historical treatment decisions for a large population of patients. It will facilitate research on treatment routines for specific patient populations by providing information on likely therapy choices. Leveraging xAI techniques, the reasoning of the analytics system is made transparent to the user. In this paper, we describe and test a system that follows this concept. Specifically, we address the two use cases (a) "therapy selection" and (b) "identification of similar patients". We test respective AI and xAI mechanisms with real-world data. Our analysis provides insights into the potential of the approach of using AI/xAI as supporting analytics system for oncologists as well as on the data requirements.

Introduction

Artificial intelligence (AI) is a branch of computer science which is concerned with the automation of intelligent behavior as displayed by humans and animals. Machine Learning (ML) is a subbranch of AI which deals with the automatic learning of patterns from data, and the usage of these learned patterns for prediction and classification tasks. Explainable AI (xAI) refers to methods that allow humans to understand the decision process of AI models. In this paper, we investigate the potential of using these technologies to support oncologists with a tool for exploring registry data. That is, we envision an analytics system that uses AI for providing oncologists with case-specific information and leverages xAI techniques to make its reasoning transparent. The core idea is to learn from historical records of applied treatments and transfer the learned patterns into new settings. These historical records constitute "real-world data"; that is, patients are drawn from a large sample of individuals who received treatment for advanced colorectal cancer during routine clinical care. This stands in contrast with randomized clinical trials, where mostly patients meeting specific and often highly restrictive criteria can participate. Our aim is to leverage AI to make the latent knowledge that rests within real-world data accessible to oncologists. Specifically, we address the two use cases of (a) "therapy selection" and (b) "identification of similar patients".

The use case of "therapy selection" is about informing oncologists of the estimated distribution function of therapies conditional on a set of medical covariates. Here the AI estimates, for a given set of patient characteristics and each possible therapy choice the probability that an oncologist would prescribe this therapy to a patient having these characteristics. Using xAI techniques, the underlying reasoning of the algorithm is made transparent. Precisely, the algorithm explains which particular patient features spoke for or against a given therapy choice in a given case. In the use case of “identification of similar patients”, the AI model is used to define a meaningful similarity metric between patients. This metric is based on clinical characteristics available to the treating oncologist.

Within this paper, we present and evaluate solutions for the implementation of the two described use cases. The evaluation includes quantitative tests as well as qualitative analyses by oncology domain experts. Our main contributions are:

- We present a concept for using AI as supporting analytics system for oncologists
- We analyze the applicability of an AI-based analytics system for estimating probability distributions of therapies
- We analyze the dependence of the analytics system’s performance on the amount of available training data
- We analyze how xAI can render the reasoning of an AI-based data analytics system transparent to the treating physician
- We present and evaluate an AI-based similarity metric for patient records

The remainder of the paper is structured as follows. The rationale and motivation from a medical perspective is given in **Medical Background**. We describe details of the application scenarios in section **Aims and Scope** as well as specifics of the analyzed data sets in **Patient Sample**. In section **Technical Approach, Concepts** we present our AI-based approaches to support decision making for oncologists. This is followed by a description of our experiments to test the AI-based solution in section **Experiments**. We review related work in section **Related Work** and conclude the paper in section **Summary and Discussion**.

Medical Background

New treatment options for cancer patients have emerged over the last decades providing oncologists and patients with an increased number of treatment options. However, with the growing number of options, complexity in treatment decision making grows and increasingly challenges the medical expertise (1). What is the best treatment for this patient? Currently, treatment recommendations and guidelines are mainly based on evidence from randomized clinical trials (RCTs) comparing new drugs to standard treatments or placebo. Although RCTs are the best way to compare drugs or treatment strategies, most patients recruited into such RCTs are not representative for most patients who are intended to receive these treatments during routine clinical care. This is because current RCTs have strict in- and exclusion criteria that typically exclude many elderly patients and those with relevant co-morbidities. Consequently, such RCTs can have a high degree of internal validity, but only a low level of external validity (2). In other words, there can be high uncertainty as to whether tested treatments are feasible and whether the observed treatment effect is transferable to patient populations that would not have met the inclusion criteria to enter the respective RCT. In addition, to predict a patient’s prognosis based on data from RCTs is also questionable, because populations of patients being included in RCTs often have a different risk profile (2). To close this evidence gap, insights into data collected during routine clinical care (real-world data) are necessary and should also be considered when making treatment decisions. However, it is important that such real-world data are of high quality to exclude biased inference (e.g., selection or reporting bias). To investigate the potentials of machine learning in supporting treatment decision making, we have set up this project using a high-quality cohort of patients being enrolled in the prospective and multicenter Tumor Registry Colorectal Cancer (TKK) registry (3).

Aims and Scope

Treatment decision making relies on many factors such as patient characteristics (e.g., age, co-morbidities, and clinical performance status), tumor characteristics, available evidence, patient preferences, and the physician's expertise. In the TKK database, we have information on the first two aspects: patient characteristics and tumor characteristics. Both are very important factors when it comes to treatment decision making, given that the available evidence is the same and that physician's expertise is usually similar among trained oncologists. Based on this, we have three aims:

- To investigate whether AI can satisfactorily model the probability distribution of treatments a clinician would administer to an individual based on given patient and tumor characteristics
- To investigate whether xAI methods can render the reasoning of the AI model interpretable
- To investigate whether AI techniques can be used to define a meaningful similarity metric for patients.

Patient Sample

For all experiments outlined below, we used a dataset of patients with advanced/metastatic colorectal cancer from the TKK. This is a prospective, multicenter, longitudinal, nation-wide cohort study in Germany which started in 2006. Since then, 269 medical oncologists have recruited more than 4000 patients with advanced/metastatic disease. This study was reviewed by an ethics committee and is registered at ClinicalTrial.gov (NCT00910819). Eligible patients are 18 or older with histologically confirmed colorectal cancer. Patients also received at least one systemic chemo- or targeted therapy (e.g., antibodies) for advanced/metastatic disease. Written informed consent was obtained from all patients. All patients are treated according to physician's choice and are followed for a minimum of 3 years (or until death, loss to follow-up or withdrawal of consent). At the time of enrolment, data on patient and tumor characteristics are documented. From 2008 to 2013, the KRAS-mutation status was collected without further information on the tested/mutated exon(s). Since 2014, data on the extended RAS-testing routine have been documented (KRAS-exons 2, 3 and 4 and NRAS-exons 2, 3 and 4), further referred to as (K)RAS and (N)RAS mutation testing, respectively.

For all experiments described herein, we used a sample of 3,563 patients who were prospectively enrolled with start of the first systemic treatment for their advanced/metastatic colorectal cancer. Further details of the TKK have been previously reported (3).

Technical Approach, Concepts

In this section we describe a concept for using ML and xAI to support decision making in therapy selection. We describe the basic architecture of a corresponding software system and the concept for its application. Figure 1 provides an overview of the key components and the workflow for their use. We discuss each component and the specific instantiation of our test implementation blow.

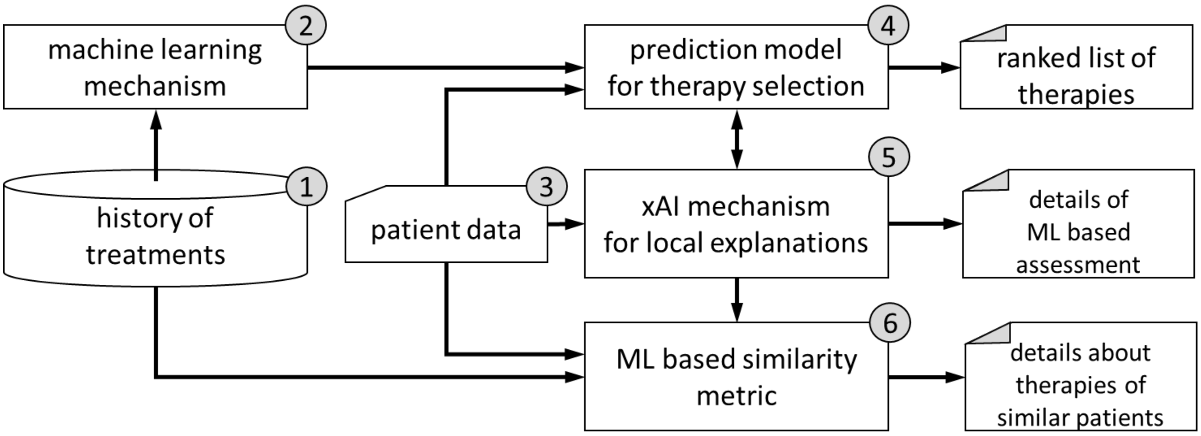

machine learning mechanism
2
prediction model for therapy selection
4
ranked list of therapies
5
history of treatments
1
patient data
3
xAI mechanism for local explanations
details of ML based assessment
6
ML based similarity metric
details about therapies of similar patients

Figure 1: Architecture and key components of the system concept

Component 1: history of treatments

Treatment data include patient characteristics such as demographics (e.g., age, sex, education), medical history, comorbidities and disease characteristics (e.g., timing of tumor onset, metastases, mutational status, pattern of disease). In our study this refers to each patient included in the TKK registry, alongside with their chosen therapy (see section **Patient Sample**). The patient and disease characteristics are available to a clinician when discussing treatment options with a patient and constitute the basis for the supervised learning of therapy decisions (component 2). In our study, chosen therapies are specified by the therapy backbone (e.g., FOLFOX/CAP+IRI) and the used antibody (e.g., anti-EGFR, anti-VEGF), if applicable. This leads us to n=15 distinct therapy schemes in the data. When we discard the information about the exact substances in the backbone and only consider its principle (monotherapy, doublet chemotherapy, or triplet chemotherapy), we obtain n=12 distinct therapies. When also discarding the substance details of the chosen antibody to only consider if an antibody was given or not, we obtain n=8 therapy classes. Our main analysis is carried out on the variant with 8 distinct therapies, but we also report the results of the other two variants when evaluating the algorithm’s performance.

Component 2: machine learning mechanism

The ML mechanism is used to learn a prediction model for therapy selection. That is, it learns to predict which therapy is chosen for a patient based on given information about patient and disease characteristics at the beginning of treatment. It thereby aims at mimicking the decision made by oncologists. In principle, any supervised learning mechanism could be used for this task. In our test implementation we used the XGBoost algorithm (4). We chose this algorithm because it generally shows good performance on tabular data (5) and is well supported in the ML ecosystem.

Component 3: patient data

Patient data in component 3 comprise the available information about a specific patient for whom the system should support the therapy selection. This information contains a subset of features that are available in the history of treatments (component 1). Ideally this subset has complete data for the features that are used by the prediction model (component 4). Depending on the prediction method, missing information may need to be replaced by suitable values. In our implementation we use XGBoost, which has inbuilt mechanisms for dealing with missing values.

Component 4: prediction model for estimating probability distributions of therapies

The prediction model is the output of the ML mechanism. Trained on historical treatment decisions, it aims at modeling the decision making of oncologists. The model is used to make predictions on new instances, i.e., to estimate the probability distribution of therapies given a set of patient characteristics.

Component 5: xAI mechanism for local explanations

This component provides local explanations for the prediction that the ML model made for a given patient. Local explanations offer humans insights into the decision making of an AI-based system. Specifically, they provide information on how important a given feature was for the decision for a given instance. In our application, that is the impact of different pieces of information about the patient and their disease (e.g., a certain mutation or age of the patient) on the therapy prediction for that patient. Note that this contrasts with global feature importance that assesses the general importance of a feature (e.g., average importance across many cases). With local feature importance, we can assess the impact of features on the prediction for a particular case. This may differ greatly from the average importance of a feature and give precise insight into the decision for an individual patient. In our implementation, we use the SHAP library to compute Shapley values for local explanations (6). These values reflect the fair contribution of each feature to the prediction outcome (7). Here we take two inputs: (a) the prediction model[1] and (b) the patient data for which we explain the prediction. The output is the Shapley value of each feature in the analyzed patient. This information can provide oncologists with insights into what factors were considered by the model and how they impacted the prediction. That is, oncologists get information if a certain feature impacted the model output in favor or against a given therapy. The magnitude of the value furthermore shows the strength of that impact. Oncologists can use this information to reason about the model's decision for and against different therapies.

Component 6: AI-based similarity metric

The proposed system includes a component to identify patients that are similar to a reference patient (see aim #2). This enables oncologists to inform themselves about past treatment routines applied to similar patients. The key challenge is to find a suitable definition of similarity between patients. Here we build upon the idea of "supervised clustering" as presented in (17). The concept is to use local feature importance - obtained through supervised prediction models – instead of raw feature values. The similarity metric then uses a distance defined via these importance values. Thereby, the metric applies a case specific weighting. This weighting stems from what the prediction model has learned about case-specific therapy selection. With this concept, the similarity focuses on the features that the model finds relevant in the given case. This is in contrast to a similarity metric that factors in the features for all patients in the same way, regardless of the specifics of their case.

For our implementation of this general concept we use the vectors of Shapley values for each feature and for each class of the therapy prediction (i.e., one vector per therapy class). We then concatenate all vectors and use their distance to define their similarity metric. The intuition is that similar cases have similar importance for the same features in therapy prediction. Precisely, we represent each patient by a vector **v**= $(v_1, v_2, \ldots, v_m)$ with $m = n \cdot p$. Here p is the number of patient features and n is the number of target classes (therapies). The entry $v_{(k-1)p+j}$ is the Shapley value of feature j in the prediction of therapy class k , where $1 \leq k \leq n$. Similarity between two patients is then defined

[1] Note that the SHAP library typically uses a set of data instances (e.g. training data) as background information to compute SHAP values. We do not mention this input in the general concept, because alternative implementations without such data are possible.

in terms of the Manhattan distance of their vector representations. In our main analysis, we have $n = 8$ and $p = 96$.

Experiments

Our experiments are designed to test the feasibility of using AI to aid the therapy selection for patients with advanced/metastatic colorectal cancer. Specifically, we address four questions: (1) What is the quality of AI-based therapy selection, (2) do the local explanations with Shapley values render the AI algorithms' selection interpretable, (3) is the AI-based similarity metric meaningful and (4) how does the availability of data impact the performance of the AI-based therapy selection? We describe the corresponding experiments below. We use accuracy and the macro-averaged $f1$-score as metrics for the predictive performance of the algorithm.

Experiment 1: Quality of Predictions

Assessing the quality of the predictions made by the AI algorithm yields a conceptual challenge since in general, the best therapy for a given patient is not known. We therefore here resort to comparing the AI's predictions with the therapy decisions made by humans. However, even human experts may disagree regarding the optimal therapy for a specific case. This limits the quality that we can expect to observe but provides us with an indication about the quality of AI-based predictions.

Experimental Setup

In the experiments we used the records of 3,586 individual cancer patients. After removing implausible records, we were left with 3,563 patients. We extracted 71 variables containing information about the patient's health status at the beginning of their first palliative therapy. We used one-hot encoding for non-binary categorical variables if they have less than 10 possible values and label encoding otherwise. Ordinal variables were encoded numerically. This gave us $p = 96$ features as predictors for our ML model. As label for model training, we used the chosen therapies, defined as the combination of the principle of the chosen chemotherapy backbone and whether an antibody was given or not. This results in 8 different first-line

therapies within the data set. From this data set we selected a stratified random sample of 60% of the records as training set and held out the rest for testing. With this data we trained a classifier using the XGBoost library (version 1.5.0) (4) and balanced class weighting. Predictive performance is measured in terms of accuracy (the sum of the diagonal entries of the confusion matrix divided by the total sum of its entries), and the f_1 score (the harmonic mean of precision and recall), averaged over all target classes. We used both macro averaging (the unweighted arithmetic mean over all target classes) and weighted averaging with weights proportional to the number of true positives of the given class.

Evaluation of prediction performance

Figures 2 and 3 show the confusion matrix and ROC curves for the classifier's predictive performance on the test set, respectively (horizontal: predicted treatments; vertical: actual treatments applied). To highlight the algorithm's ability to distinguish between the therapy classes 'DOUBLET Anti-VEGR' and 'DOUBLET Anti-EGFR', we also show the confusion matrix for n=12 distinct therapy classes (i.e. including detailed antibody information). Accuracy, weighted-average and macro-averaged $f1$-scores are reported in Table 1.

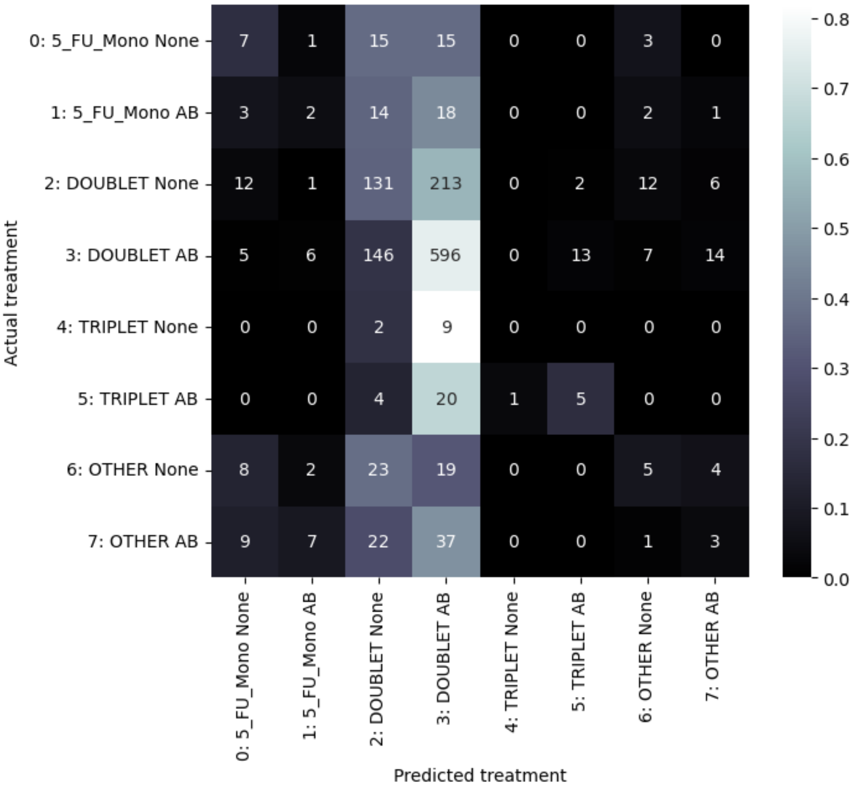


Figure 2: Confusion matrix for 8 distinct therapy classes

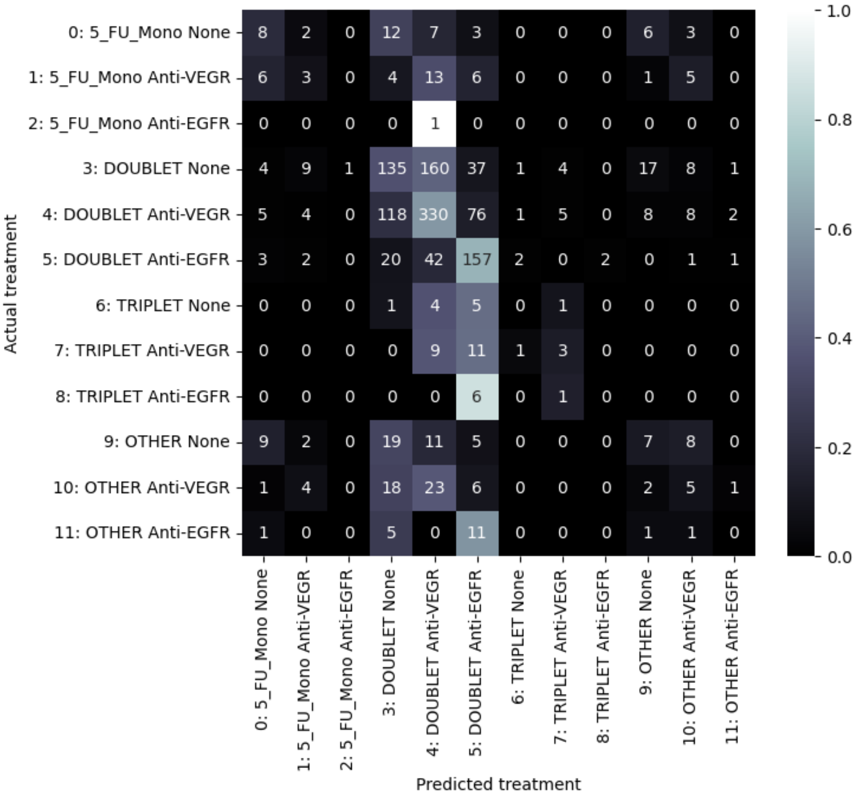


Figure 2: Confusion matrix for 8 (upper panel) and 12 (lower panel) distinct therapy classes. The cell in row j and column k is colored according to the fraction of patients who were predicted to receive therapy k among those patients who actually received therapy j .

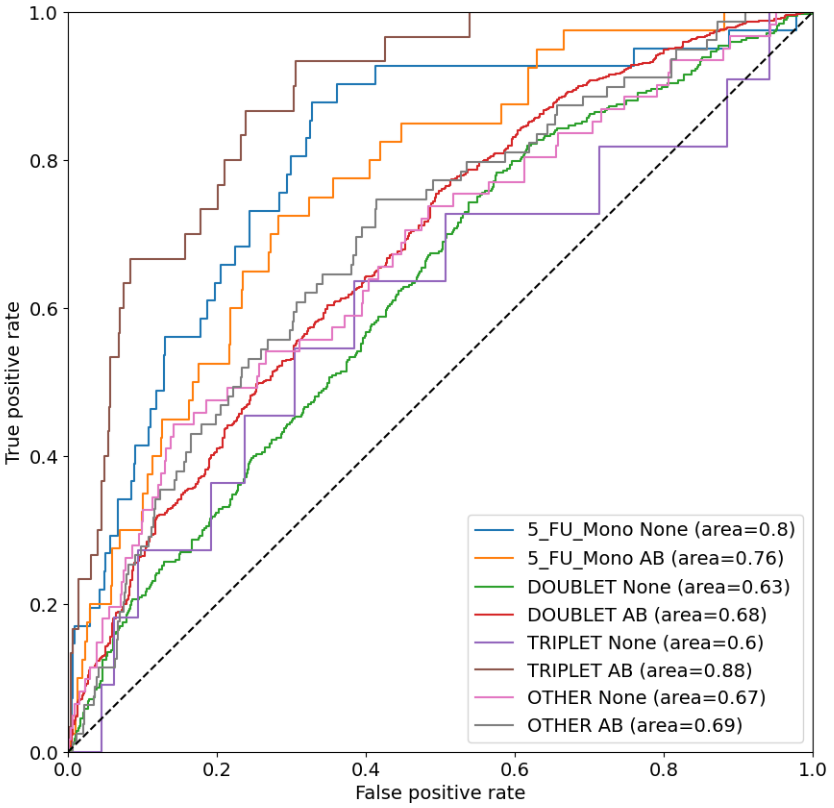


Figure 3: ROC curves for the classifier's predictive performance in the case of 8 distinct therapy classes.

We observe that the AI-based prediction has a fair agreement with the actual given treatment and thus treatment choice of the treating physician. This is encouraging, given that also human experts may disagree on the best therapy and hence, we cannot expect perfect agreement in these experiments. However, rare therapies are never predicted and yield poor results in the evaluation. This is expected, as the training set includes a low – and insufficient – number of examples for rare therapies. One can expect improved results with bigger training sets. We investigate this effect in our fourth experiment.

Number n of distinct therapies	Accuracy	Weighted average f_1-score	Macro-average f_1-score
8	0.53	0.50	0.21
12	0.45	0.44	0.18
15	0.27	0.26	0.16

Table 1: Accuracy, weighted-average and macro-averaged $f1$-scores of the prediction for the three different levels of aggregation of the therapy classes.

Experiment 2: Insights with global and local feature importance measures

In these experiments we aim at testing the benefits of using local feature importance to support the therapy decision of oncologists. This experiment is qualitative by nature and aims at providing insights into the use of xAI in the targeted use case. For the experiment we computed and visualized the Shapley values for therapy predictions. This comprises the global Shapley values (i.e. local Shapley values averaged over the entire test set) as well as local Shapley values for individual predictions. The visualizations are then analyzed by domain experts regarding their validity from a medical perspective. We provide a sample result and the corresponding medical analysis below. To protect the privacy of patients in the displayed figures, Gaussian noise was added to the features *Age at start of 1-line*, *Date of inclusion*, *Disease free interval* and *BMI*[2].

Qualitative assessment of xAI results

We can obtain a global measure of feature importance by averaging the magnitude of the Shapley value of each feature and therapy over all patients in the test set. The results are then stacked to obtain the total feature importance. Figure 4 shows the 30 most important features used for therapy prediction.

[2] The noise was added to the entire data set after training the model but before the computation of SHAP values.

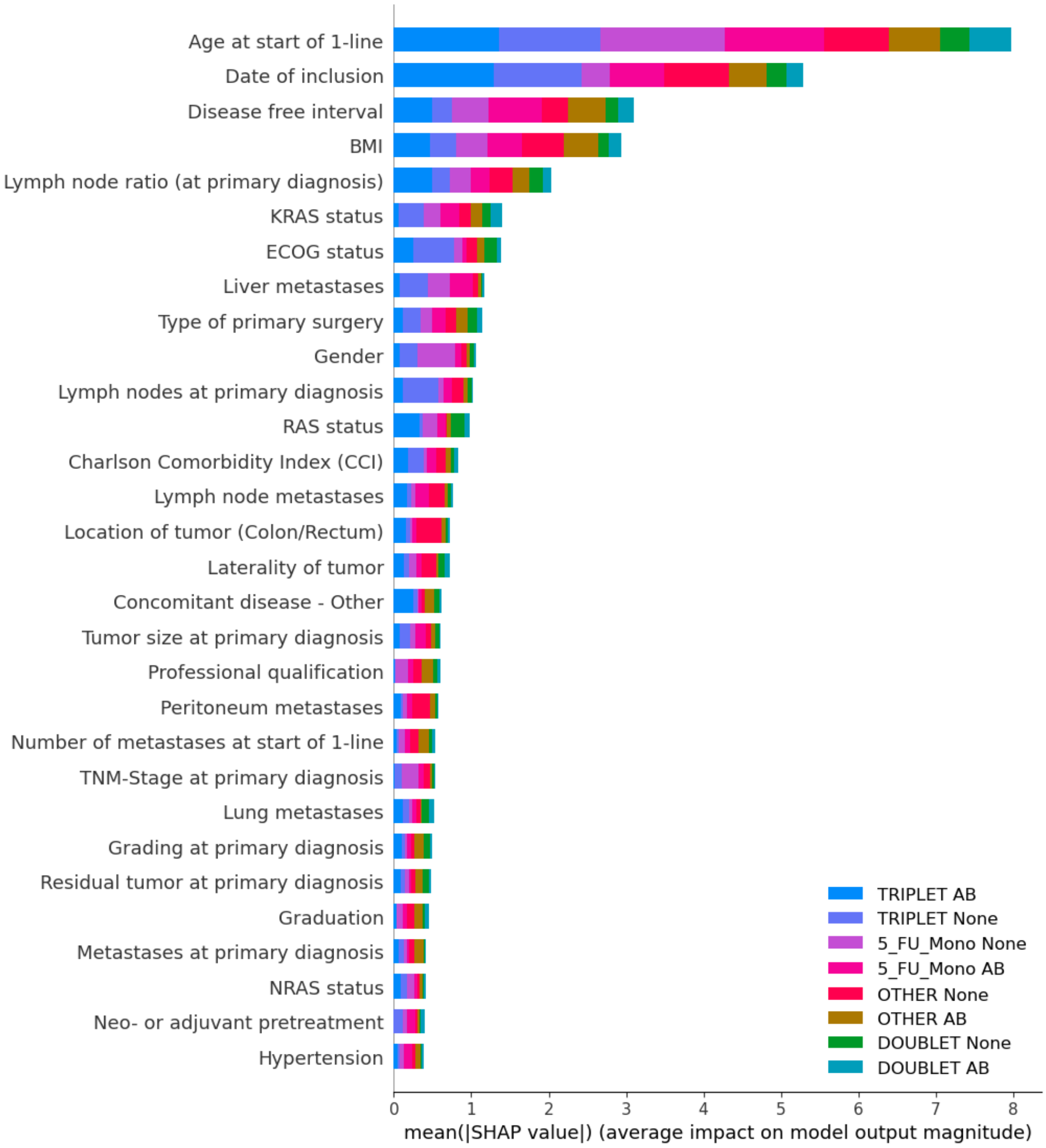
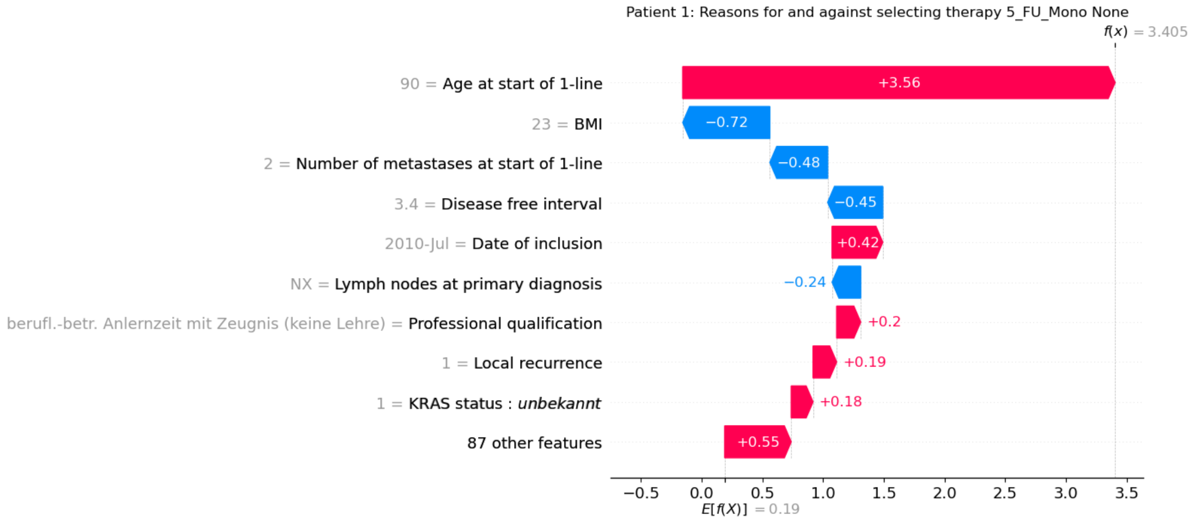
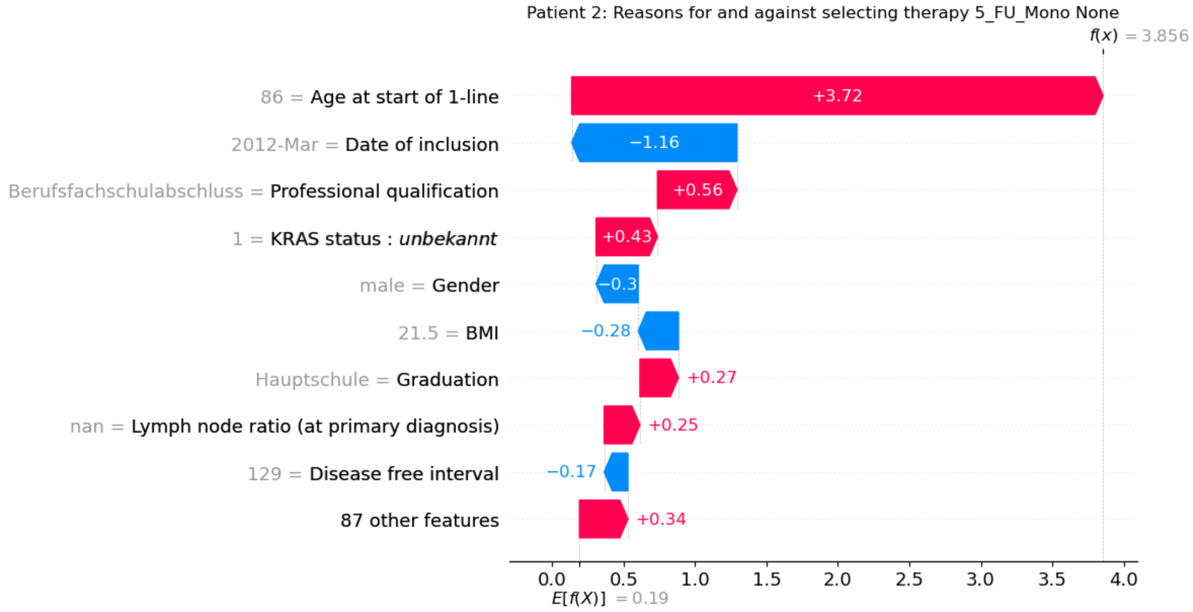


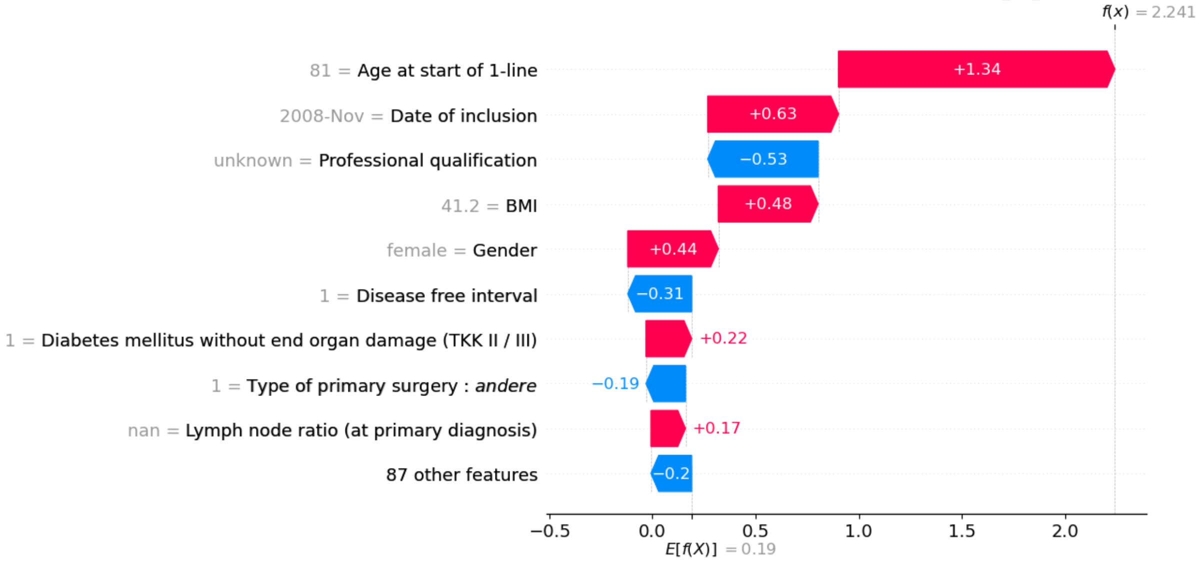
Figure 4: The 30 most important features used for therapy prediction. Here the importance of a feature is defined as the class specific global Shapley value, summed over all therapy classes. The length of the horizontal bars represents the importance of the given feature and the color coding shows the contributions of the different therapy classes.

In Figure 5 we show the Shapley values for five different patients for which the algorithm correctly predicted therapy 5-FU monotherapy (5-FU Mono None, that is, intravenous 5-FU without an antibody) out of 8 different possible therapies. As mentioned above, patient data is overlaid with noise for privacy protection reasons. We chose these examples because they resemble interesting cases where a less common therapy was chosen. Such cases are well suited to check if the special reasons for using such a therapy are well reflected in the Shapley values. All five patients in Figure 5 are over 80 years when starting therapy (range 81 to 90 years). Age is known to be a very important factor in clinical decision-making because it strongly correlates with frailty and increased risk of treatment-related side effects. Interestingly, in patient number 3, it is also the additional diagnosis of diabetes without end organ damage that contributed to prediction of treatment with 5-FU monotherapy. In almost all patients (except patient 3), the body mass index (BMI) was a factor rather speaking against choice of 5-FU monotherapy, although the effect was not very strong. These four patients had a BMI within the range considered to be normal weight. BMI is also a surrogate for morbidity; in the context of colorectal cancer, a low BMI can be associated with frailty and being a sign of malnutrition and disease activity. Thus the “normal” BMI may have been considered by the model as a factor that would have allowed more intense treatment than 5-FU monotherapy.

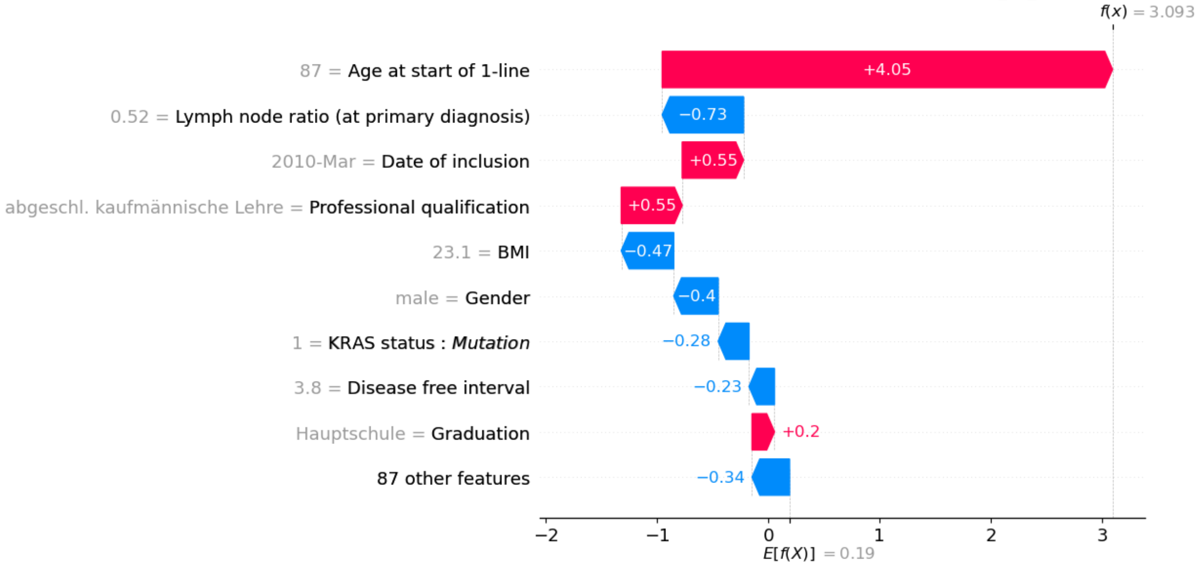

Patient 1: Reasons for and against selecting therapy 5_FU_Mono None
f(x) = 3.405
90 = Age at start of 1-line
23 = BMI
2 = Number of metastases at start of 1-line
3.4 = Disease free interval
2010-Jul = Date of inclusion
NX = Lymph nodes at primary diagnosis
berufl.-betr. Anlernzeit mit Zeugnis (keine Lehre) = Professional qualification
1 = Local recurrence
1 = KRAS status : unbekannt
87 other features
+3.56
−0.72
−0.48
−0.45
+0.42
−0.24
+0.2
+0.19
+0.18
+0.55
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
E[f(X)] = 0.19


Patient 2: Reasons for and against selecting therapy 5_FU_Mono None
f(x) = 3.856
86 = Age at start of 1-line
2012-Mar = Date of inclusion
Berufsfachschulabschluss = Professional qualification
1 = KRAS status : unbekannt
male = Gender
21.5 = BMI
Hauptschule = Graduation
nan = Lymph node ratio (at primary diagnosis)
129 = Disease free interval
87 other features
+3.72
−1.16
+0.56
+0.43
−0.3
−0.28
+0.27
+0.25
−0.17
+0.34
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
E[f(X)] = 0.19

Patient 3: Reasons for and against selecting therapy 5_FU_Mono None



Patient 4: Reasons for and against selecting therapy 5_FU_Mono None



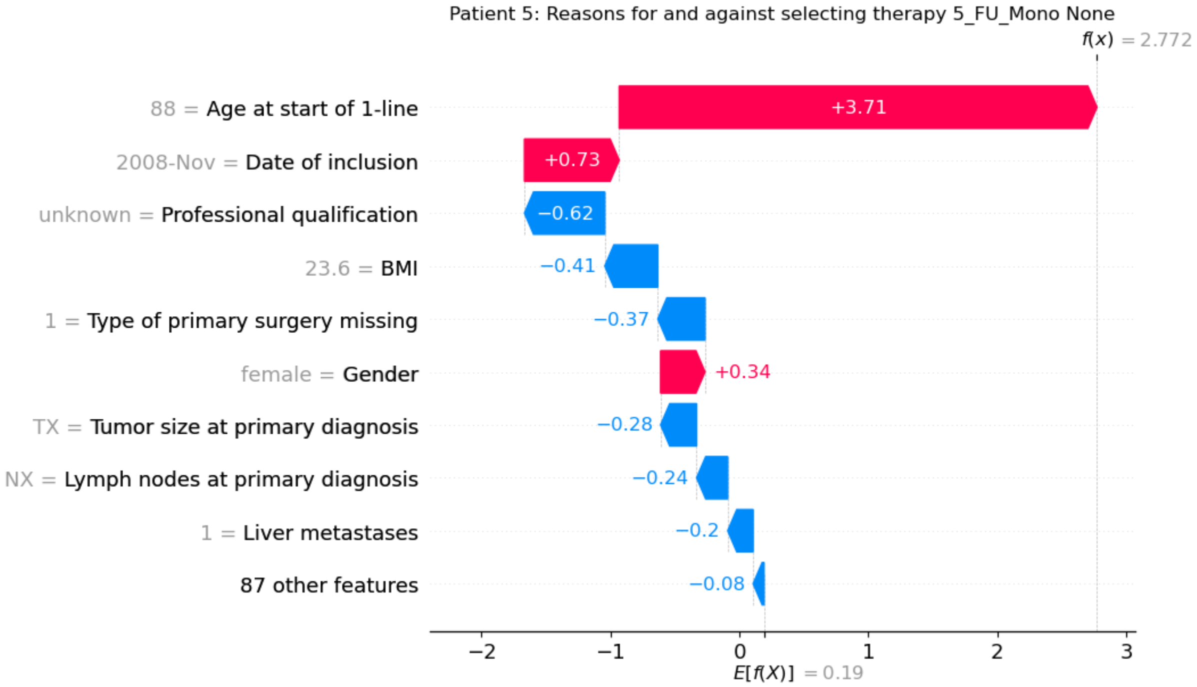
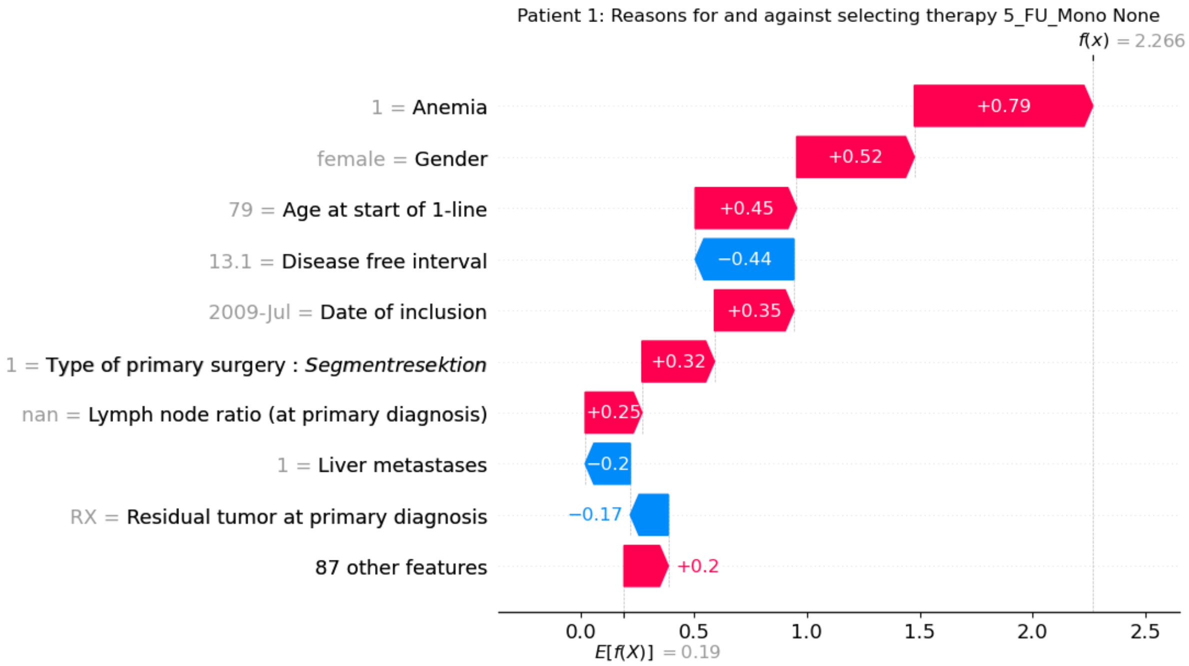
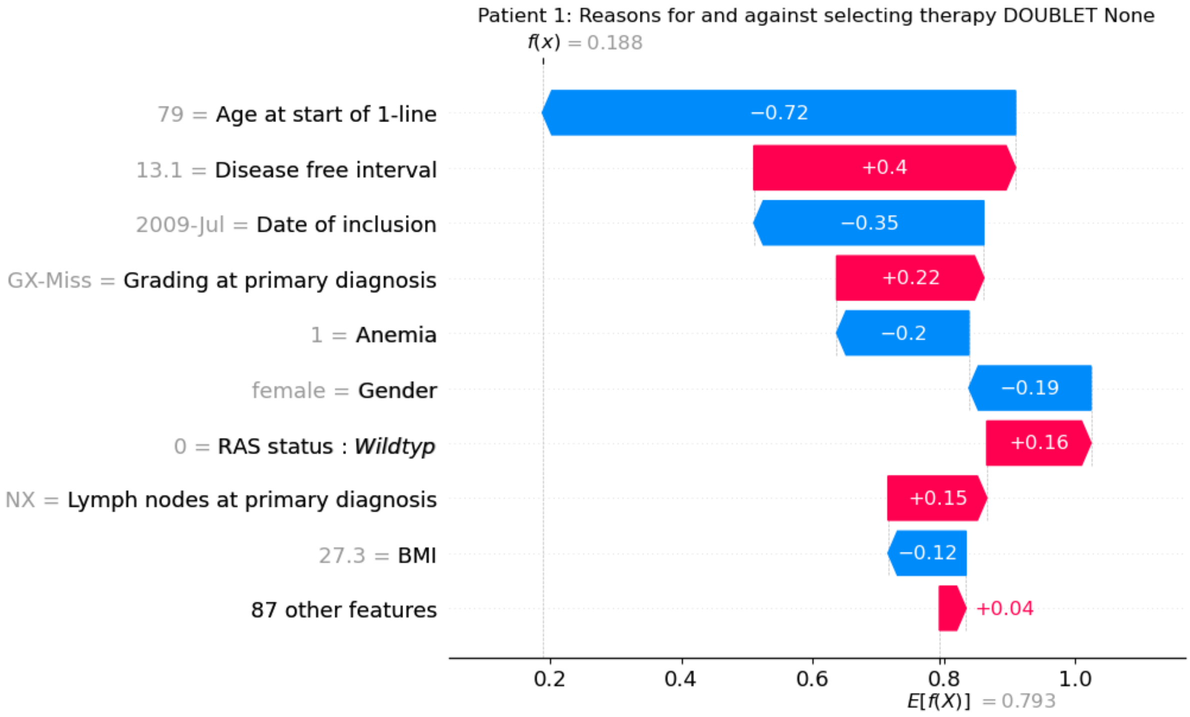


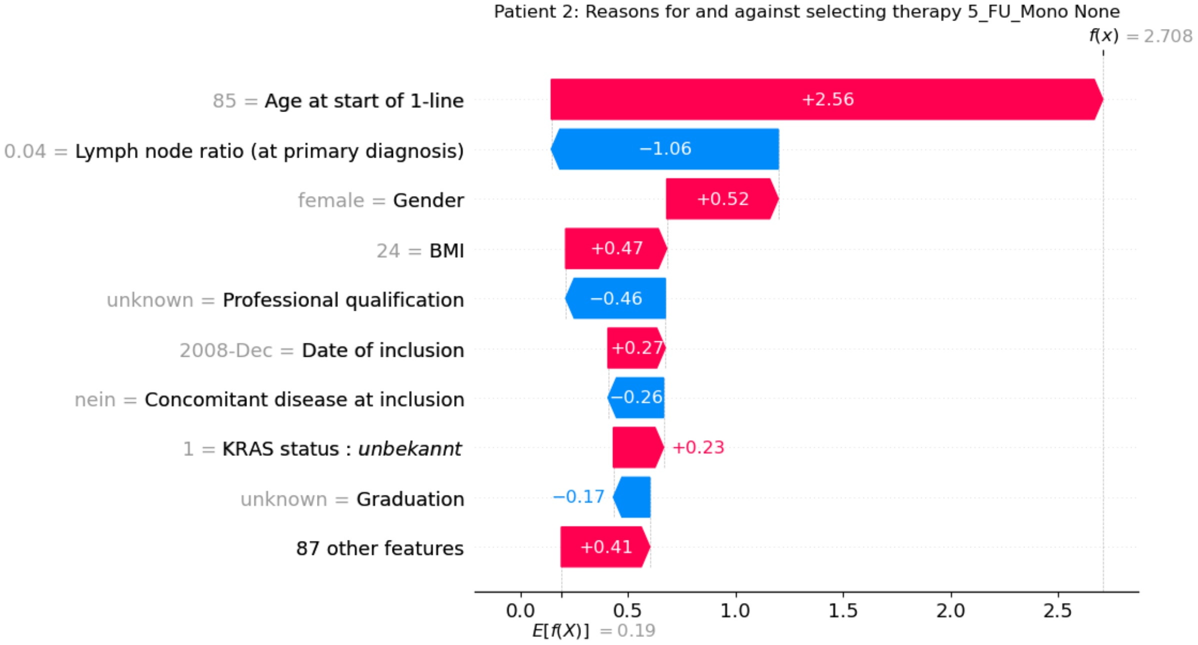
Figure 5: Shapley values for five patients for which the algorithm correctly predicted 5-FU monotherapy. For privacy protection reasons, patient data displayed in the figures has been overlaid with noise.

In Figure 6 we show the Shapley values for five different patients for which the algorithm predicted 5-FU monotherapy, but the patient actually received doublet chemotherapy without an antibody (false positives). We chose these examples to be transparent about the existing uncertainty when predicting less common therapies. Furthermore, such cases provide insights into potential causes for divergence between the AI-based prediction and the treatment decision. The figures show the ten most important Shapley values for the prediction of this therapy class. Upper panel, reasons for and against treatment with 5-FU monotherapy; lower panel, reasons for and against treatment with doublet chemotherapy.

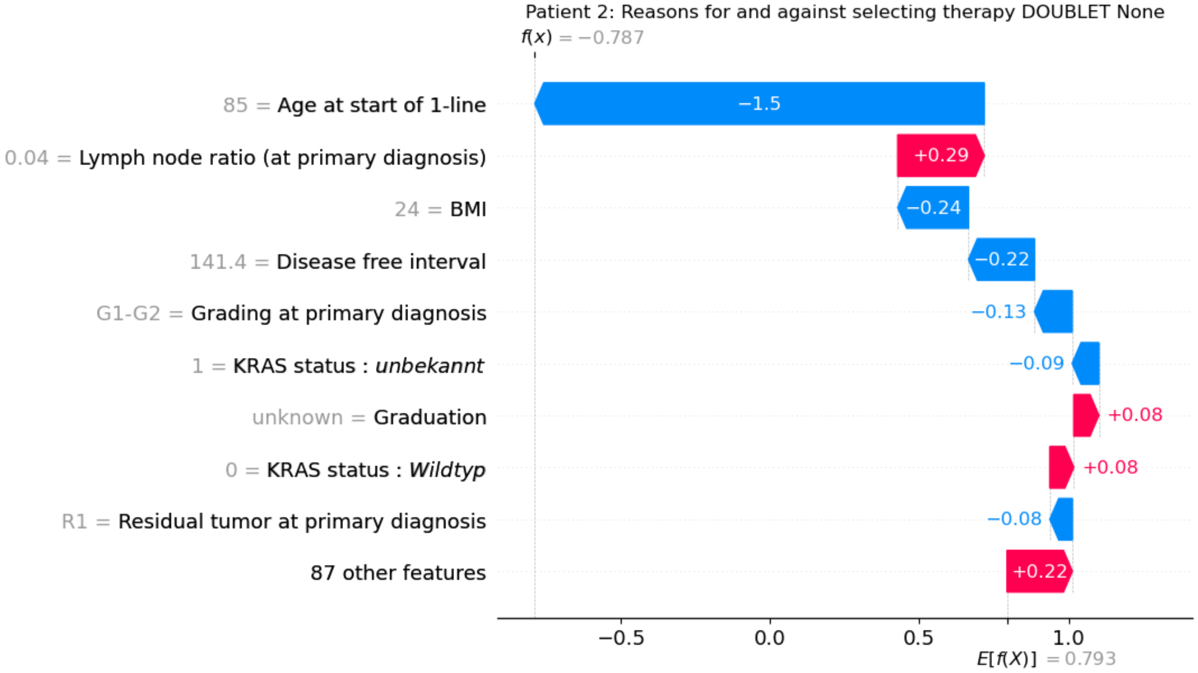
In patient 1, similar to the true positives in Figure 5, increased age (79 years) was the factor speaking for 5-FU monotherapy and also the presence of anemia (patient 1, upper panel). We can see in the lower panel, that these factors are inversed, speaking against the treatment with doublet chemotherapy, whereas missing grading status and disease-free interval were factors favoring doublet chemotherapy. As outlined above, age is an important factor for treatment-decision making, but there are situations in which the perceived chronological age does not necessarily mirror the frailty status of the patient (fit elderly patients). Although we have information on the clinical performance status for most patients in our dataset, many important factors also driving treatment-decision making are not captured (e.g. the actual patient preference). For instance, some patients may opt for a more intense treatment, although risk for side effects is higher. Such factors outside our database might have driven the treatment decision. Interestingly, one would have assumed that co-morbidities and clinical performance status would have more weight in the recommended treatment, but the effect of these are rather modest in either direction. A similar pattern as for patient 1 can be seen in patients 2, 3, 4 and 5.


Patient 1: Reasons for and against selecting therapy 5_FU_Mono None
f(x) = 2.266
1 = Anemia
+0.79
female = Gender
+0.52
79 = Age at start of 1-line
+0.45
13.1 = Disease free interval
−0.44
2009-Jul = Date of inclusion
+0.35
1 = Type of primary surgery : Segmentresektion
+0.32
nan = Lymph node ratio (at primary diagnosis)
+0.25
1 = Liver metastases
−0.2
RX = Residual tumor at primary diagnosis
−0.17
87 other features
+0.2
0.0
0.5
1.0
1.5
2.0
2.5
E[f(X)] = 0.19

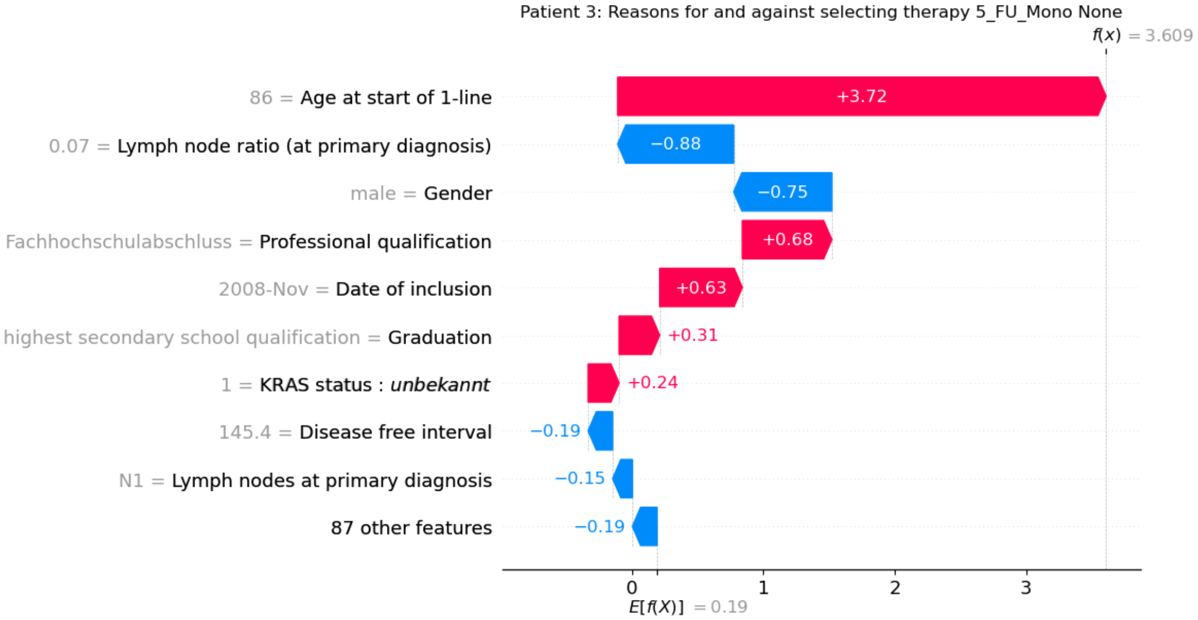

Patient 1: Reasons for and against selecting therapy DOUBLET None
f(x) = 0.188
79 = Age at start of 1-line
−0.72
13.1 = Disease free interval
+0.4
2009-Jul = Date of inclusion
−0.35
GX-Miss = Grading at primary diagnosis
+0.22
1 = Anemia
−0.2
female = Gender
−0.19
0 = RAS status : Wildtyp
+0.16
NX = Lymph nodes at primary diagnosis
+0.15
27.3 = BMI
−0.12
87 other features
+0.04
0.2
0.4
0.6
0.8
1.0
E[f(X)] = 0.793

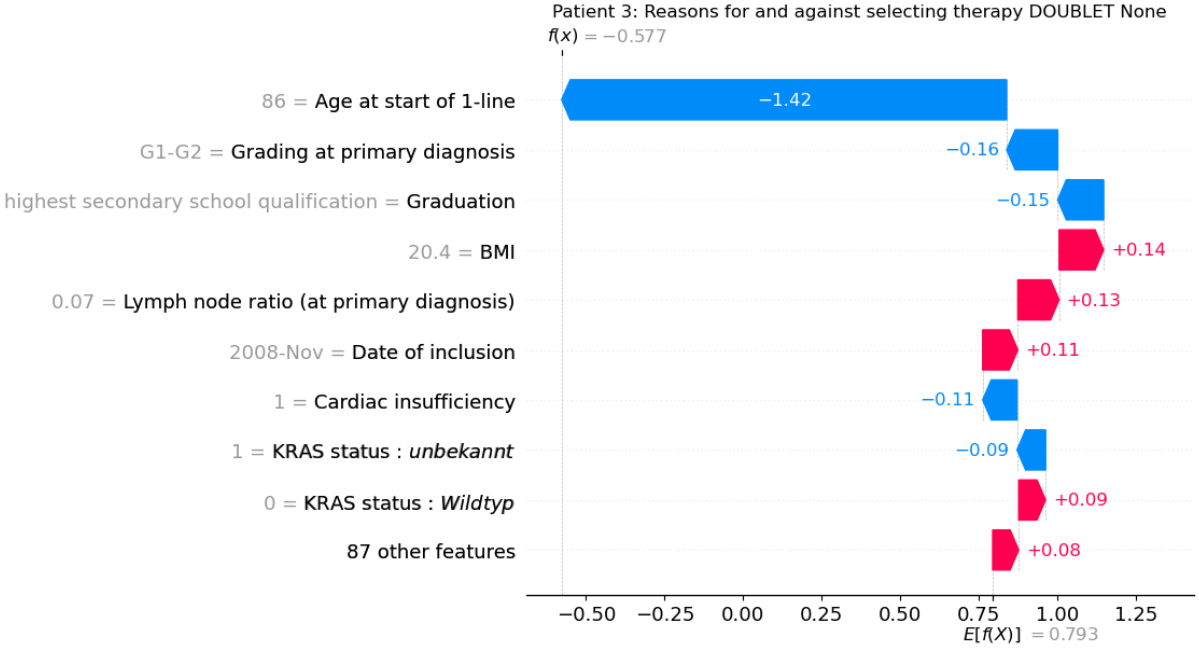


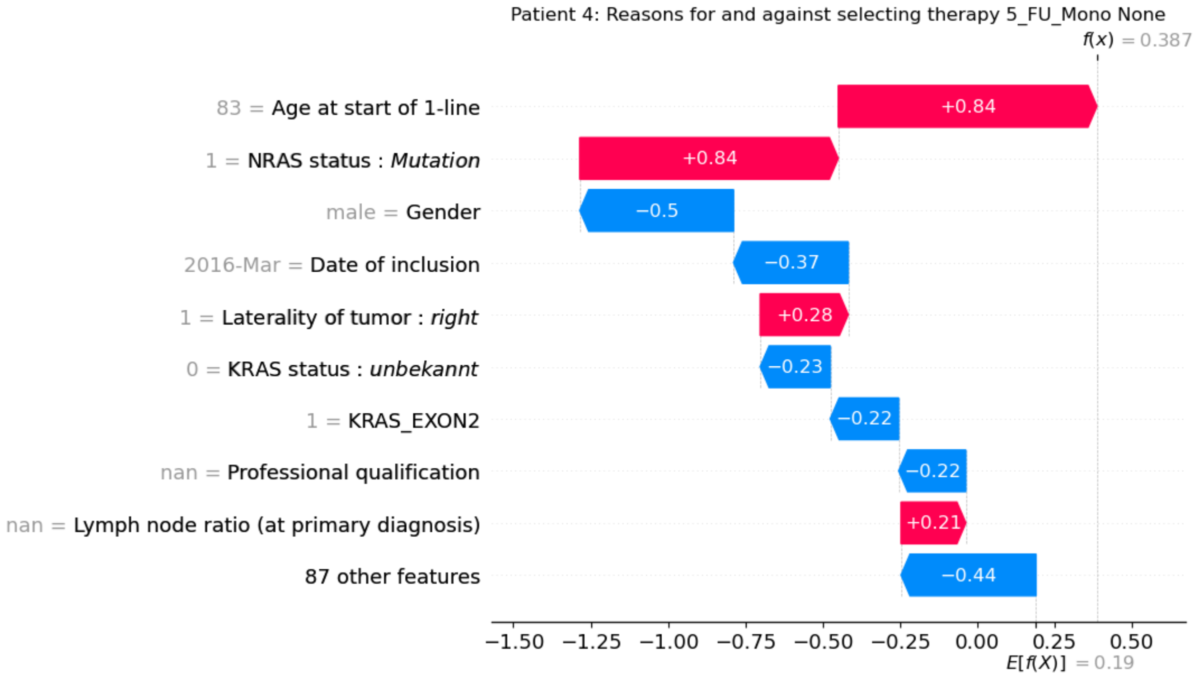
Patient 2: Reasons for and against selecting therapy 5_FU_Mono None

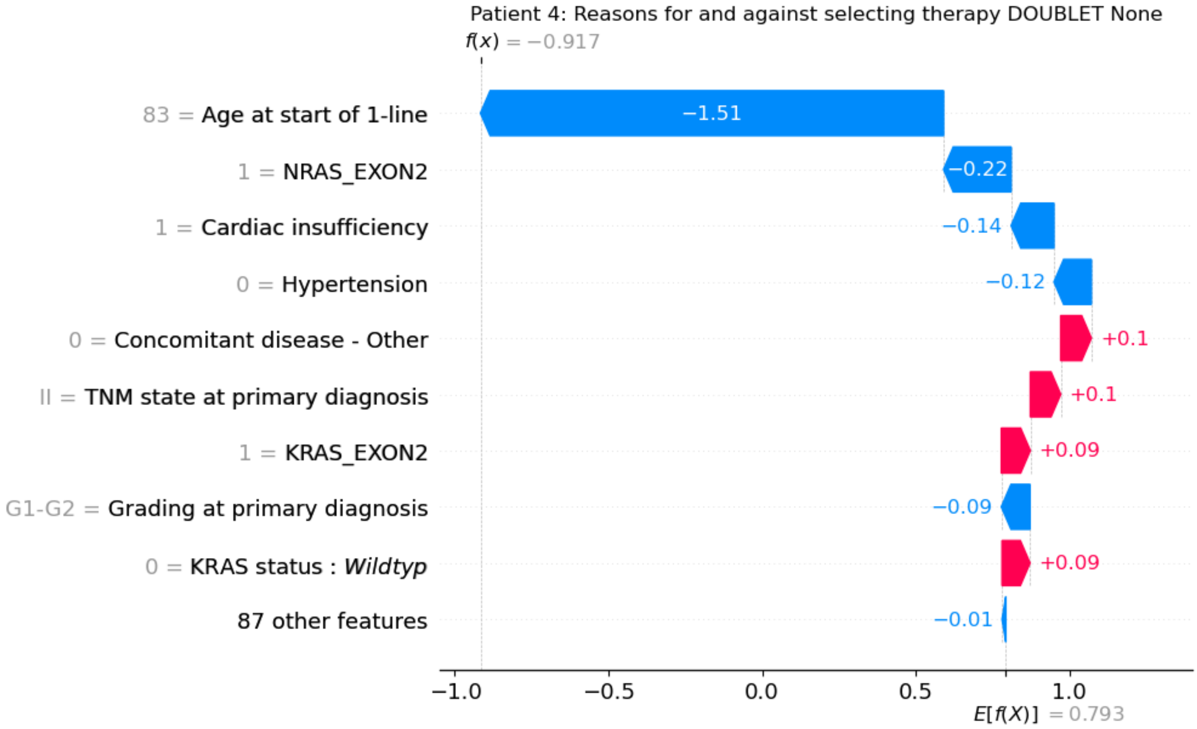


Patient 2: Reasons for and against selecting therapy DOUBLET None


Patient 3: Reasons for and against selecting therapy 5_FU_Mono None


Patient 3: Reasons for and against selecting therapy DOUBLET None


Patient 4: Reasons for and against selecting therapy 5_FU_Mono None
f(x) = 0.387
83 = Age at start of 1-line
1 = NRAS status : *Mutation*
male = Gender
2016-Mar = Date of inclusion
1 = Laterality of tumor : *right*
0 = KRAS status : *unbekannt*
1 = KRAS_EXON2
nan = Professional qualification
nan = Lymph node ratio (at primary diagnosis)
87 other features
+0.84
+0.84
−0.5
−0.37
+0.28
−0.23
−0.22
−0.22
+0.21
−0.44
−1.50
−1.25
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
E[f(X)] = 0.19


Patient 4: Reasons for and against selecting therapy DOUBLET None
f(x) = −0.917
83 = Age at start of 1-line
1 = NRAS_EXON2
1 = Cardiac insufficiency
0 = Hypertension
0 = Concomitant disease - Other
II = TNM state at primary diagnosis
1 = KRAS_EXON2
G1-G2 = Grading at primary diagnosis
0 = KRAS status : *Wildtyp*
87 other features
−1.51
−0.22
−0.14
−0.12
+0.1
+0.1
+0.09
−0.09
+0.09
−0.01
−1.0
−0.5
0.0
0.5
1.0
E[f(X)] = 0.793

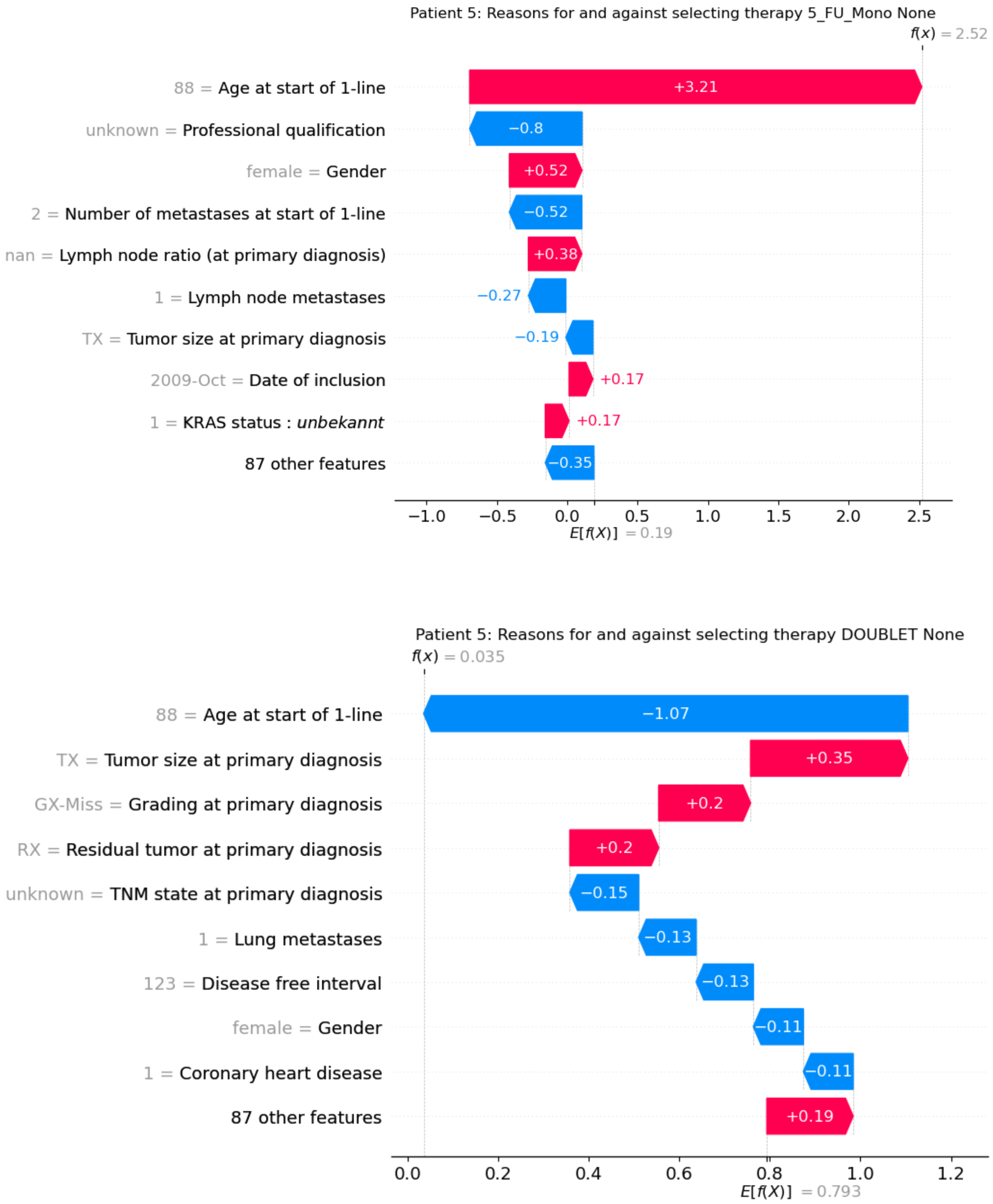


Figure 6: Shapley values for five different patients for which the algorithm predicted 5-FU monotherapy when actually doublet chemotherapy was applied, i.e., false positive prediction. Shapley values for and against 5-FU monotherapy (upper panel) and for and against doublet chemotherapy (lower panel) are depicted. For privacy protection reasons, patient data displayed in the figure has been overlaid with noise.

Experiment 3: Benefits of AI-based similarity metric

In this experiment we analyze the benefits of using the proposed AI-based similarity metric. The goal is to show that the metric improves the identification of patients that are similar in a meaningful way. A direct way to evaluate this aspect would be to enquire domain experts about their assessment of the results. For our experiments we take an indirect approach. That is, we use our similarity metric as input for a K-Nearest-Neighbor (KNN) classifier for therapy prediction and compare the classification results against a baseline metric. We argue that the KNN-classifier yields better prediction results if the underlying metric is more meaningful from a medical perspective.

Experimental Setup

For the distance between two patients, we use the metric based on Shapley values as described above. We use the same partitioning of the data into training and test set as in experiment 1 and fit a KNN-classifier with number of neighbors $k = 30$ and inverse distance weighting to the training data. The implementation is done with the KNeighborsClassifier of the scikit-learn library (version 1.0.2). For our baseline metric, we represent each patient by the vector **w**= $(w_1, w_2, \dots, w_p)$ of their features, where each feature is centered and normalized by its empirical standard deviation. Mean imputation is used for missing values[3]. As for the Shapley value-based metric, similarity of two patients is then defined in terms of the Manhattan distance of their vector representations and a KNN-classifier with the same parameters is fitted to the training data. We compare the performance of the two classifiers using the same metrics as in experiment 1.

Evaluation

The experiments show improvements of the prediction quality when using KNN with the Shapley value-based distance metric, compared to a naïve baseline-distance metric (Table 2). Although the improvements are small, the results indicate that the Shapley-based distance metric may provide a meaningful similarity measure. Note that this experiment evaluates the desired effect only indirectly and classification is not the aim of the addressed use case. For many instances, a less elaborate metric may find less similar patients but lead to the same therapy prediction. In such cases we would observe no benefits. However, our approach aims at identifying patients that are similar in a meaningful way, so that they can serve as reference cases. Here, a better similarity is beneficial even if the recorded therapies are the same. Since the tests with a KNN classifier can only reveal benefits for certain cases, we find the observed improvement encouraging. An analysis with human experts who directly assess the usefulness of the similarity metric may further clarify the benefits of the approach.

Score type	Macro-averaged score		Weighted average		Accuracy	
Number n of distinct therapies	KNN (Shapley)	KNN (Baseline)	KNN (Shapley)	KNN (Baseline)	KNN (Shapley)	KNN (Baseline)
8	0.17	0.13	0.49	0.48	0.58	0.58
12	0.15	0.13	0.44	0.42	0.50	0.48
15	0.14	0.11	0.26	0.24	0.30	0.28

Table 2: Prediction quality improvement using KNN.

[3] Note that 12 out of 96 features were encoded with the LabelEncoder from the scikit-learn package (www.scikit-learn.org), which impacts the distance measure to some extent.

Experiment 4: Impact of data availability

The available amount of training data impacts the performance of machine learning models. However, the strength of this impact varies from case to case. In this experiment we analyze this effect for the case of colorectal cancer therapy predictions.

Experimental Setup

For our experiments we have in total 3,563 patient records available (see section Experiment 1 – Experimental setup). To investigate the effect of the training size, we trained multiple models on differently sized subsets of the data. Specifically, we set aside a 40% stratified sample for testing and used the rest as source for training data. From the training data, we iteratively took 90% stratified subsets to train models (thereby iteratively reducing the training set size). We then computed performance measures (f1-socre for one-versus-all) on the initially held out test sets for each model. The process was repeated 10 times with different random seeds.

Evaluation

Figure 7 shows the results of the experiment about the impact of the training data size. From visual inspection of these plots, it appears that the learning curves for the prediction of two therapies reach a saturation point at about 500 training instances. For all other therapies we have less than 500 instances in the training set and do not observe a point of saturation. Note that for several therapies the number of training examples is rather small. This results in a poor prediction performance of the model for those therapies. However, the shape of the curve suggests that improvements with larger training sets are possible.

We draw two main conclusions from this analysis. One is that the performance results for therapy prediction (observed in section Experiment 1) would likely improve with additional training data. The other conclusion is that about 600 training instances per therapy may be sufficient for the chosen machine learning setup.

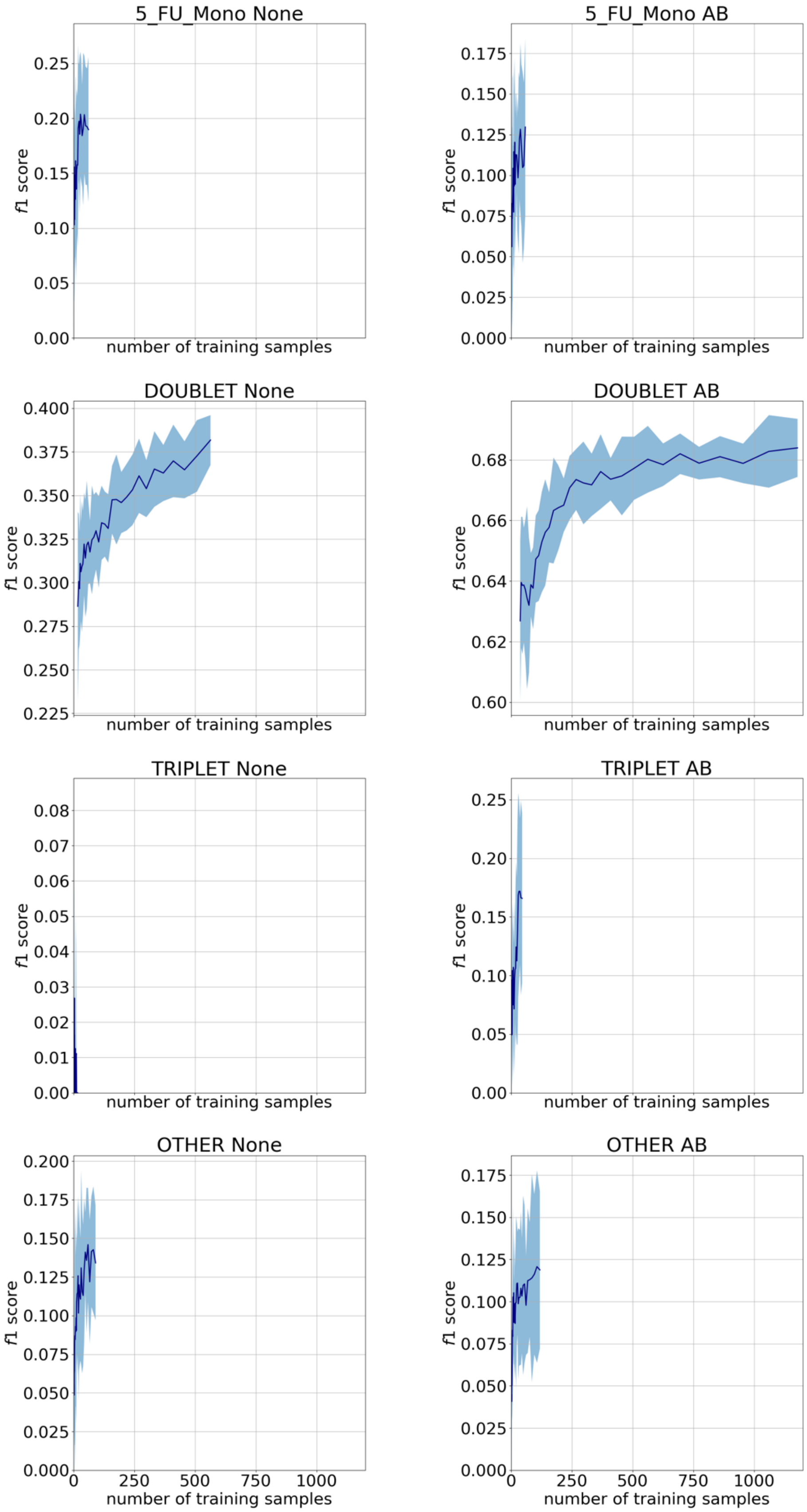


Figure 7: Impact of the number of training samples of a given therapy class on the model’s performance in correctly labeling these samples, measured in terms of f_1 score. Dark blue is the mean value and light blue the standard deviation.

Related Work

Clinical decision support systems (CDSS) constitute an active area of research with applications including diagnostics, prediction of adverse events and drug control (8,9). Existing approaches are often categorized as either knowledge-based or non-knowledge-based. Knowledge-based CDSS build on expert knowledge fed into the system in the form of if-then rules. For example, such systems have been shown to successfully decrease the risk of medication errors in a hospital setting (10). Non-knowledge-based (or data-driven) approaches on the other hand leverage real-world data by techniques from statistics and machine learning. The ideas of existing works on the topic are often borrowed from eCommerce, where data-based recommender systems are already being applied on a large scale. For example, in (11) a system based on collaborative filtering for treatment recommendations to psoriasis patients was presented. However, the application of non-knowledge-based CDSS remains relatively scarce until today (12). Challenges faced include lack of available data or low data quality and the black-box behavior of many ML algorithms, limiting the trust placed into them by humans. In the field of cancer therapy, the CDSS Watson for Oncology (WFO) aims at providing treatment recommendations to oncologists regarding surgical procedures, radiotherapy and medication. For a given cancer patient, the system outputs a list of potential therapies for that patient based on characteristics such as comorbidities, tumor characteristics and laboratory findings. Those therapies are categorized as either recommended, for consideration or not recommended. A description of the underlying technology is given in the supplement of (13). Briefly, the system uses natural language processing to retrieve knowledge from a large corpus of medical journal articles relevant to the treatment decision for a particular case. Moreover, an ML algorithm is trained with data from manually selected cases thought to be representative of the full population of patients with a particular tumor entity. WFO can therefore be seen as a hybrid CDSS consisting of both a knowledge-based and a non-knowledge-based component. While a meta-analysis found an overall solid agreement of WFO's recommendations with those by human experts (14), this concordance has been shown to vary by country and tumor entity (15). For the field of hematology, the KAIT system is currently under development. Like WFO, it will aim at providing medical doctors with therapy recommendations and case-specific insights and features both data-driven and knowledge-based components (16). While we understand our tool as an analytics tool rather than a recommender system, the underlying methodology could also be employed in the framework of a CDSS. Therefore, our work adds to the body of knowledge about data driven decision support by providing tests on real world data about colorectal cancer treatments and analyzing a specific xAI approach.

Summary and Discussion

In this paper we analyzed the potential of using AI to build an information tool that enables dynamic data exploration and analysis of real-world datasets. Specifically, we analyzed the two use cases "therapy selection" and "identification of similar patients". Both objectives aim to provide a second view built on the large amount of real-world data and thus make this broad knowledge accessible to individual oncologists. For these use cases we proposed a system setup that uses supervised learning and xAI techniques.

We have shown the applicability of the concept and obtained insights on the required amount of training data, but additional work is required to further assess the viability of our approach. While we have demonstrated superiority of the AI-based approach against simple baseline methods, our experiments show a certain degree of disagreement between the predicted and the chosen therapy. However, disagreement is certainly also expected if different human experts are asked to give a second opinion. Quantifying the level of human disagreement and comparing this to the AI-based results is subject to future work.

Our approach has limitations. One fundamental problem consists in the fact that our current AI algorithm learns therapy selection from prospectively recorded past records. However, the therapy landscape in oncology develops quickly causing concept drift; that is, historic decisions learned by the algorithm may have better alternatives by now. Also, the best practice about therapy decision may change over time and change the probability of selecting a therapy for a given patient. The discussed concept drift makes the algorithm prone to the cold-start problem of AI-based recommender systems: New, effective therapies will not be considered by the system if they have not been administered to enough patients yet. One possible solution to this problem could be to weight samples in such a way that the algorithm gives higher attention to more recent records. This may be combined with concept drift-detection techniques to adjust the training set accordingly. Another solution could comprise human expert knowledge. This could be realized e.g., through implementing a hybrid system where the AI-based recommender is complemented by a rule-based component. Furthermore, human knowledge may be applied to relabel older data records with the current knowledge about state-of-the-art therapies. Note that the experiments in this paper do not account for the concept drift problem. That is, the algorithm is trained and tested on random subsets of the data, not considering temporal order. Also, the algorithm is provided with the starting date of the treatment. This allows the mechanism to learn about treatment regimens that were present in certain time frames and use that knowledge in the prediction. Testing the impact of concept drift on the prediction performance therefore remains an open issue.

It is important to stress that therapy outcomes of patients such as overall survival, progression-free survival, and quality of life, which are also documented in the TKK database, were not considered. This means that the information tool may reproduce and even reinforce suboptimal, yet common practice in treatment routine. Future work should address this issue by including available outcome data when selecting therapies for patients.

Furthermore, feature selection and hyperparameter tuning remains subject to future work. Due to the high number of features and therapy classes, the Shapley value-based similarity measure may be subject to the curse of dimensionality, which means that in high dimensions, the distribution of pairwise distances between points tends to concentrate. Feature reduction techniques applied upstream may therefore lead to better results. Hyperparameter tuning may yield further improvements, as the reported results were obtained using fixed and mainly default parameters of the XGBoost classifier. In future work, a systematic optimization of hyperparameters such as maximum tree depth and learning rate should be performed. Together with optimized feature selection this may lead to improvements of the model performance.

We believe that the investigated concepts have great potential to support information processes in cancer care using dynamic data exploration and analysis of real-world datasets. Our findings show promising results that call for further analysis and development of the outlined ideas. Beyond expanding on these ideas and addressing the discussed limitations, we plan to address further use cases in future work.

References

1. Bossaerts P, Murawski C. Computational Complexity and Human Decision-Making. Trends Cogn Sci [Internet]. 2017 Dec;21(12):917–29. Available from: http://dx.doi.org/10.1016/j.tics.2017.09.005

2. Khozin S, Blumenthal GM, Pazdur R. Real-world Data for Clinical Evidence Generation in Oncology. J Natl Cancer Inst [Internet]. 2017 Nov 1;109(11). Available from: http://dx.doi.org/10.1093/jnci/djx187

3. Marschner N, Arnold D, Engel E, Hutzschenreuter U, Rauh J, Freier W, et al. Oxaliplatin-based first-line chemotherapy is associated with improved overall survival compared to first-line treatment with irinotecan-based chemotherapy in patients with metastatic colorectal cancer - Results from a prospective cohort study. Clin Epidemiol [Internet]. 2015 Apr 20;7:295–303. Available from: http://dx.doi.org/10.2147/CLEP.S73857

4. 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 [Internet]. New York, NY, USA: Association for Computing Machinery; 2016 [cited 2022 Mar 29]. p. 785–94. (KDD '16). Available from: https://doi.org/10.1145/2939672.2939785

5. Shwartz-Ziv R, Armon A. Tabular data: Deep learning is not all you need. Inf Fusion [Internet]. 2022 May 1;81:84–90. Available from: https://www.sciencedirect.com/science/article/pii/S1566253521002360

6. Lundberg SM, Lee S-I. A Unified Approach to Interpreting Model Predictions. In: 31st Conference on Neural Information Processing Systems (NIPS 2017), [Internet]. Available from: https://papers.nips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

7. Molnar C. Interpretable Machine Learning [Internet]. 2022 [cited 2022 Mar 29]. Available from: https://christophm.github.io/interpretable-ml-book/index.html#summary

8. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med [Internet]. 2020 Feb 6;3:17. Available from: http://dx.doi.org/10.1038/s41746-020-0221-y

9. Tran TNT, Felfernig A, Trattner C, Holzinger A. Recommender systems in the healthcare domain: state-of-the-art and research issues. J Intell Inf Syst [Internet]. 2021 Aug 1;57(1):171–201. Available from: https://doi.org/10.1007/s10844-020-00633-6

10. Mahoney CD, Berard-Collins CM, Coleman R, Amaral JF, Cotter CM. Effects of an integrated clinical information system on medication safety in a multi-hospital setting. Am J Health Syst Pharm [Internet]. 2007 Sep 15;64(18):1969–77. Available from: http://dx.doi.org/10.2146/ajhp060617

11. Gräßer F, Beckert S, Küster D, Schmitt J, Abraham S, Malberg H, et al. Therapy Decision Support Based on Recommender System Methods. J Healthc Eng [Internet]. 2017 Mar 28;2017:8659460. Available from: http://dx.doi.org/10.1155/2017/8659460

12. Gräßer F, Tesch F, Schmitt J, Abraham S, Malberg H, Zaunseder S. A pharmaceutical therapy recommender system enabling shared decision-making. User Model User-adapt Interact [Internet]. 2021 Aug 5; Available from: https://doi.org/10.1007/s11257-021-09298-4

13. Somashekhar SP, Sepúlveda M-J, Puglielli S, Norden AD, Shortliffe EH, Rohit Kumar C, et al. Watson for Oncology and breast cancer treatment recommendations: agreement with an expert

multidisciplinary tumor board. Ann Oncol [Internet]. 2018 Feb 1;29(2):418–23. Available from: http://dx.doi.org/10.1093/annonc/mdx781

14. Jie Z, Zhiying Z, Li L. A meta-analysis of Watson for Oncology in clinical application. Sci Rep [Internet]. 2021 Mar 11;11(1):5792. Available from: http://dx.doi.org/10.1038/s41598-021-84973-5

15. Zhou N, Zhang C-T, Lv H-Y, Hao C-X, Li T-J, Zhu J-J, et al. Concordance Study Between IBM Watson for Oncology and Clinical Practice for Patients with Cancer in China. Oncologist [Internet]. 2019 Jun;24(6):812–9. Available from: http://dx.doi.org/10.1634/theoncologist.2018-0255

16. Alexander Oeser, Anne Sophie Kubasch, Tim Meschke, Nora Grieb, Lukas Schmierer, Uwe Platzbecker, Thomas Neumuth. KAIT - Knowledge-Driven and Artificial Intelligence-Based Platform for Therapy Decision Support in Hematology. Available from: https://www.uniklinikum-leipzig.de/einrichtungen/medizinische-klinik-1/Freigegebene%20Dokumente/Dokumente%20H%c3%a4ma%20Forschung/KAIT_White_Paper.pdf

17. Lundberg, Scott M., Gabriel G. Erion, and Su-In Lee. "Consistent individualized feature attribution for tree ensembles." arXiv preprint arXiv:1802.03888 (2018).