# 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
Faculty of Business Information Systems
Furtwangen University
Furtwangen, Germany
e-mail: zie@hs-furtwangen.de

Norbert Marschner, Dunja Klein
iOMEDICO AG
Medical Department
Freiburg, Germany
e-mail: {norbert.marschner,dunja.klein}@iomedico.com

Benjamin Kasenda
Medical Oncology
University Hospital of Basel
Basel, Switzerland
e-mail: benjamin.kasenda@usb.ch

Nina Haug
iOMEDICO AG
Biostatistics
Freiburg, Germany
e-mail: nina.haug@iomedico.com

***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. Artificial Intelligence (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 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 Explainable AI (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.**

***Keywords-Explainable AI; Oncology; Medical Information Systems.***

## I. INTRODUCTION

In this paper, we investigate the potential of using Artificial Intelligence (AI) and Explainable AI (XAI) technologies to support oncologists with a tool for exploring medical registry data. The paper is an updated version of our earlier technical report [1]. In our work, we envision an analytics system that uses AI for providing oncologists with case-specific information and leverages XAI techniques, specifically Shapley (SHAP) values [2], 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" (RWD); that is, patients are drawn from a large sample of individuals who received treatment for advanced Colorectal Cancer (CRC) during routine clinical care. Our aim is to leverage AI to make the latent knowledge that rests within RWD 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 about likely therapy choices which would have been made by other oncologists for a given (possibly fictious) patient. Here, the AI estimates the probability that an oncologist would prescribe a given therapy to a patient given their characteristics. Using XAI techniques, the underlying reasoning of the algorithm is made transparent. Precisely, the algorithm explains which particular patient features spoke in favor of or against a given therapy choice in a given case. In the use case of "identification of similar patients", the AI model is employed 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 both 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, testing various algorithms
- 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

 ISBN: 978-1-68558-105-3 

- We present and evaluate an AI-based similarity metric for patient records

The remainder of the paper is structured as follows. We discuss related work in Section II. The rationale and motivation from a medical perspective is given in Section III, along with details of the application scenarios and specifics of the analyzed data sets. In Section IV, we present our AI-based approaches to support decision making for oncologists. This is followed by a description of our experimental setup and experimental results in Section V, and the conclusion of the paper in Section VI.

## II. Related Work

Clinical Decision Support Systems (CDSS) constitute an active area of research with applications including diagnostics, prediction of adverse events, and drug control [3], [4]. 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 [5]. Non-knowledge-based (or data-driven) approaches, in contrast, leverage real-world data by techniques from statistics and machine learning (ML). For example, in [6], 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 [7]. Challenges faced include lack of available data or low data quality and the black-box behavior of many ML algorithms, limiting trust placed into them by humans.

In the field of cancer therapy, the CDSS Watson for Oncology (WFO) aimed at providing treatment recommendations to oncologists regarding surgical procedures, radiotherapy, and medication. A description of the underlying technology is given in the supplement of [8]. While a meta-analysis found an overall solid agreement of WFO's recommendations with those by human experts [9], this concordance has been shown to vary by country and tumor entity [10]. For these reasons, the WFO service was discontinued in 2022. While we see analytics as the purpose of our tool rather than recommendation, 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 CRC treatments and analyzing a specific XAI approach.

ML, XAI and SHAP values have been used in medicine and specifically oncology in previous works. For example, Nohara et al. use a SHAP-based approach to analyze models for predicting the risk of colon cancer [11]. Moncada-Torres et al. address breast cancer survival with machine learning for survival analysis and use SHAP to analyze model predictions [12]. Alabi et al. predict the survival of nasopharyngeal cancer with machine learning and analyze the resulting models with the XAI methods SHAP and LIME [13]. Unlike these works, we do not aim at predicting an outcome, but the therapy. Hence, our XAI analysis reflects factors that impact the therapy selection. Our experiments examine the utility of the approach for this use case. To the best of our knowledge this is the first work presenting such an analysis on real-world cancer data (specifically advanced/metastatic CRC).

Using SHAP for the comparison of data points has been proposed in the context of what has been coined as "supervised clustering" by Lundberg et al. [14]. In the medical domain, Cooper et al. use a clustering method based on that idea to identify subgroups in COVID-19 symptoms [15]. Likewise, our similarity metric is based on this idea of supervised clustering. In this paper, we adapt and test the concept for the use case of finding patient records that are similar in a meaningful way.

## III. Medical Background

New treatment options for cancer patients have emerged over the last decades, providing oncologists and patients with an increased number of treatment possibilities. However, with the growing number of options, treatment decision making becomes increasingly complex, thereby challenging medical expertise [16]. What is the best treatment for a 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, due to their strict in- and exclusion criteria, patients recruited into them are often not representative for those who are intended to receive these treatments in routine clinical care. Consequently, such RCTs can have a high degree of internal validity, but only a low level of external validity [17]. To close this evidence gap, insights drawn from data collected during routine clinical care should also be considered when making treatment decisions. However, it is important that such Real-World Data (RWD) are of high quality to exclude biased inference (e.g., selection or reporting bias). To investigate the potentials of ML in supporting treatment decision making, we set up this project using a high-quality cohort of patients being enrolled in the prospective and multicenter Tumor Registry Colorectal Cancer (TKK) [18].

### A. Aims and Scope

Treatment decision making relies on many factors, such as patient characteristics (e.g., age and co-morbidities), tumor characteristics, available evidence, patient preferences, and the physician's expertise. The TKK database provides 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 predict — based on given patient and tumor characteristics — what treatment a clinician would have given to a patient.
- 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.

 ISBN: 978-1-68558-105-3 

### B. Patient Sample

For all experiments outlined below, we used a dataset of patients with advanced/metastatic CRC 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 4,000 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 CRC. 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 were 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 prospectively enrolled patients with start of the first systemic treatment for their advanced/metastatic CRC. Further details of the TKK have been reported previously [18].

## IV. Technical Approach and Concepts

In this section, we provide an overview of our concept for using ML and XAI to support decision making in therapy selection. Figure 1 shows key components and the workflow for their use. We discuss each component and the specific instantiation of our test implementation blow. Further details can be found in [1].

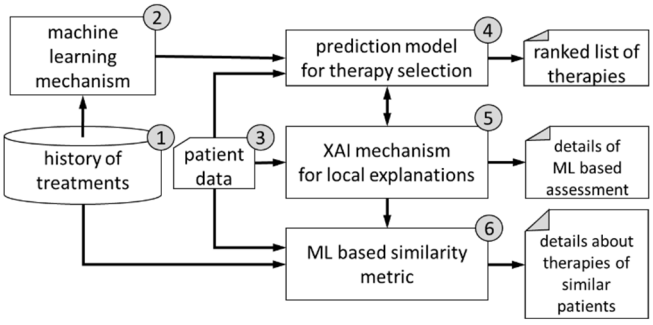


Figure 1. Architecture and key components of the system concept

*Component 1* includes patient characteristics such as demographics, medical history, comorbidities, and tumor characteristics. In our study, we used patient data from the TKK registry, alongside with the chosen therapy (see Section III B). 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 results in up to $n = 15$ distinct therapy schemes, $n = 12$ when reducing the principle (monotherapy, doublet chemotherapy, or triplet chemotherapy), and $n = 8$ when only indicating if an antibody was given and discarding the antibody details. Our main analysis is carried out with 8 distinct therapies, but we also report results of the other two variants when evaluating the algorithm's performance.

*Component 2* refers to an ML algorithm that learns to predict which therapy is chosen for a patient based on 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 main implementation we used XGBoost [19], which generally shows good performance on tabular data [20]. We compare its classification performance to other well-established methods and a baseline in our experiments.

*Component 3* refers to data of a specific patient for whom the system should support therapy selection. This information contains the subset of features that are available in the history of treatments (component 1), that is relevant for the specific prediction model.

*Component 4* is the prediction model resulting from training an ML algorithm on historical treatment decisions. The model is used to estimate, for new instances, the probability distribution of therapies given a set of patient characteristics.

*Component 5* provides local explanations for the prediction of the ML model for a given patient. Such mechanisms give insights on how important a given feature was for the decision for a given instance (e.g., how a certain mutation impacted therapy prediction for that patient). This contrasts with global feature importance that assesses the general importance of a feature (e.g., average importance across many cases). In our implementation, we use the SHAP library to compute Shapley values [2], reflecting the fair contribution of each feature to the outcome prediction [21]. Here we take two inputs: (a) the prediction model and (b) the patient data for which we explain the prediction. The output is the Shapley value (case-specific importance) of each feature in the given patient record. Oncologists can use this information to reason about the model's decision for and against different therapies.

*Component 6* identifies patients that are similar to a reference patient (see aim #3). 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. Here, we build upon the idea of "supervised clustering" as presented in [14]. The concept is to define similarity via local feature importance instead of raw feature values. In our case, the importance stems from what the prediction model has learned about case-specific therapy selection. With this concept, similarity focuses on features that the model finds relevant in a 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, we concatenate vectors of Shapley values for each feature and each target class. The intuition is that similar cases have similar importance for the same features in therapy prediction. Precisely, we represent each patient by a vector $v = (v_1, v_2, \dots, v_m)$ 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 in terms of the Manhattan distance of their vector representations. In our main analysis, we have $n = 8$ and $p = 116$.

## V. EXPERIMENTS

Our experiments are designed to test feasibility of using AI to aid therapy selection for patients with advanced/metastatic CRC. Specifically, we address four questions: (1) What is the quality of AI-based therapy selection, (2) do local explanations with Shapley values render the AI algorithms' selection interpretable, (3) is the AI-based similarity metric meaningful and (4) how does data availability impact the performance of the AI-based therapy selection? We describe the corresponding experiments below. We use accuracy and $f_1$-score as metrics for the predictive performance of the algorithm. All analyses were done in python using the libraries xgboost (v1.5.0) [19], scikit-learn (v1.0.2), scikit-optimize (v0.9.0) and shap (v0.40.0).

### A. 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 unknown. 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.

#### 1) Experimental Setup

In the experiments, we used records of 3,586 individual CRC patients. After removing implausible records, we were left with 3,563 patients. We extracted 67 variables containing information about patient's health status at the beginning of their first palliative therapy. We used one-hot encoding for non-binary categorical variables if they had less than 10 possible values and label encoding otherwise. Ordinal variables were encoded numerically. This gave us $p = 116$ features as predictors for our ML model. As label for model training, we used the chosen therapies, as described in Section IV, *component 1*. This resulted 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. We fitted a classifier using XGBoost with balanced class weighting. Predictive performance was measured in terms of macro-averaged $f_1$-score (the harmonic mean of precision and recall). We optimized hyperparameters of the classifier with respect to this metric by Bayesian hyperparameter search using the class BayesSearchCV from scikit-optimize. As a benchmark, we trained several alternative ML algorithms, where we optimized hyperparameters using the same method. The tested algorithms were a Random Forest (RF), decision tree, logistic regression with $L^2$ regularization, linear Support Vector Classifier (SVC) and a dummy classifier always predicting the most frequent class. For logistic regression and SVC, we binned continuous variables by using the bin edges [0, 50, 60, 70, 80, 100] for age, [0, 10, 50, 100, 200, 500, 2000] for the number of weeks since primary diagnosis, [0, 18.5, 25, 30, 100] for Body Mass Index (BMI) and [0, 5, 30] for the Charlson Comorbidity Index. The bins were chosen to allow sufficient numbers of examples in each category or to reflect a common categorization (in the case of BMI) [22]. We used one-hot encoding for all categorical variables thereafter, combining missing or unknown values into separate categories. This led us to $p = 158$ variables.

#### 2) Evaluation

Figures 2 and 3 show the confusion matrix and ROC curves for the classifier's predictive performance on the test set, respectively. Macro-averaged $f_1$-scores for the three different levels of granularity of the therapy classes are reported in Table I. There, we also report performances of the benchmark methods.

According to Figure 2, predictive performance increases with the number of examples for a given class. For the most frequent class (doublet therapy with antibody), we observe fair agreement between the model's prediction and the actual given treatment and thus therapy choice of the treating physician. Rare therapies, on the other hand, are rarely predicted, therefore yielding poor evaluation results. This is expected, since therapy selection is influenced by personal opinions and preferences of the corresponding physician and patient. This means that different physicians would often select different therapy strategies for the same individual. Moreover, the training set includes a low — and possibly insufficient — number of examples for rare therapies. However, it is encouraging that, according to Figure 3, AUC values of the ROC curves are high for some of the less frequent therapy classes. RF performs similar to XGBoost, and these two methods clearly outperform the other algorithms. This suggests the presence of relevant interactions between variables which cannot be captured by linear models. One can expect improved results with bigger training sets. We investigate this effect in our fourth experiment.

### B. Experiment 2: Insights with global and local feature importance measures

Here, we aim at testing the benefits of using local feature importance to support the therapy decision of oncologists. This is qualitative by nature and aims at providing insights into the use of XAI in the targeted use case. We computed and visualized Shapley values for therapy predictions. This comprises global Shapley values (i.e., local Shapley values averaged over the entire test set) and local Shapley values for individual predictions. The visualizations are then analyzed by domain experts regarding validity from a medical perspective. We provide a sample result and the corresponding medical analysis below. To protect privacy of patients in the displayed figures, Gaussian noise was added to the features Age at start of 1st-line, Date of inclusion, Weeks since primary diagnosis, and BMI. Note that the noise was

added to the entire data set after training the model, but before computing the Shapley values.

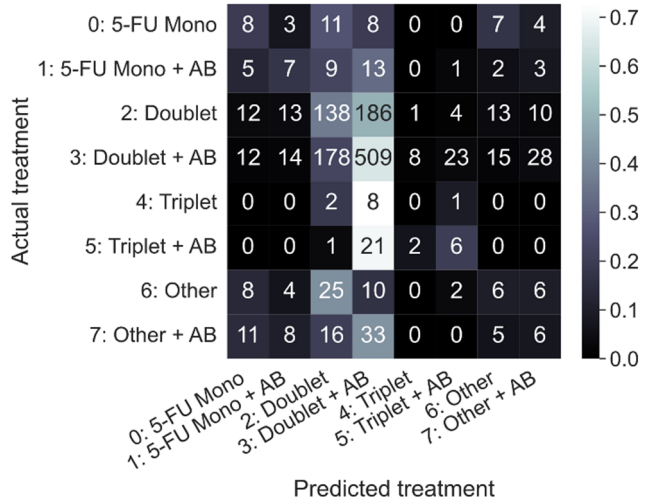


Figure 2. Confusion matrix for 8 distinct therapy classes. The cell in row *j* and column *k* is colored according to the fraction of patients who were predicted to receive therapy *k* among those patients who actually received therapy *j*.

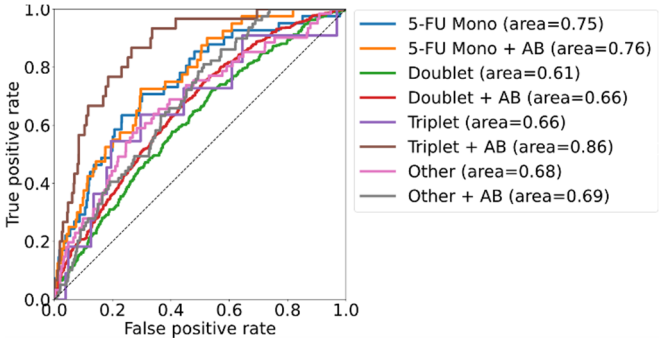


Figure 3. ROC curves for the classifier's predictive performance (case of 8 distinct therapy classes).

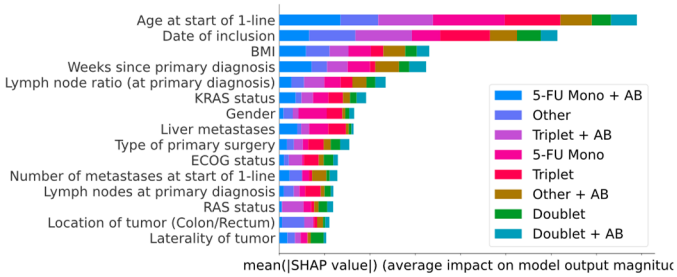


Figure 4. The 15 most important features for therapy prediction, with importance measured in terms of their global Shapley value. Color coding shows the contribution of the different therapy classes.

Figure 4 shows the 15 most important features used for therapy prediction. Here, we 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

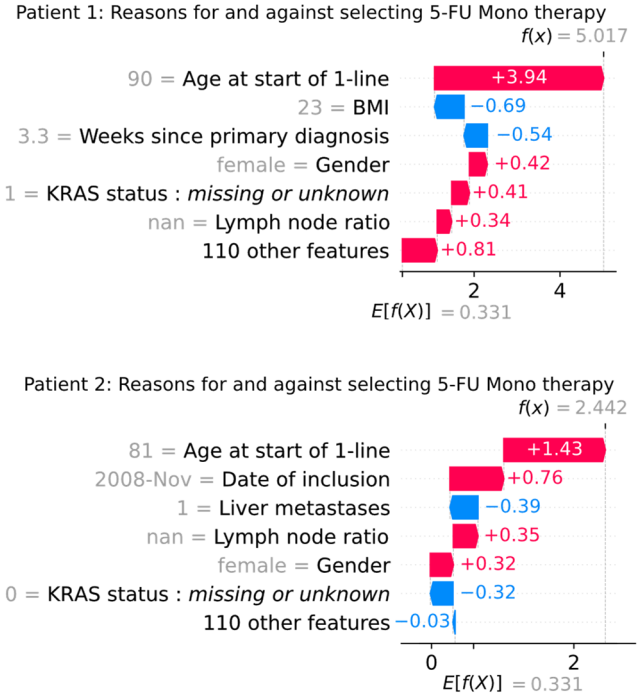


Figure 5. Shapley values for and against 5-FU monotherapy for two representative patients where the algorithm correctly predicted 5-FU monotherapy.

set. 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.

In Figure 5, we show representative Shapley values for two patients where the algorithm correctly predicted therapy 5-FU monotherapy (that is, intravenous 5-FU without an antibody) out of 8 possible choices. For privacy reasons, data shown have been overlaid with noise. These examples represent 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. Both patients in Figure 5 are over 80 years when starting therapy. 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. For patient 1, BMI, which was within the normal range, was a factor rather speaking against choice of 5-FU monotherapy, although the effect was not very strong. BMI is also a surrogate for morbidity; in the context of CRC, low BMI can be associated with frailty and is a sign of malnutrition and disease activity. Thus the "normal" BMI may have been considered by the model as a factor allowing more intense treatment than 5-FU monotherapy.

Figures 6 and 7 show Shapley values for two different representative patients for which the algorithm predicted 5-FU monotherapy when actually doublet chemotherapy was applied. Such cases provide insights into potential causes for

TABLE I. PREDICTIVE PERFORMANCE OF THE COMPARED CLASSIFIERS, AS MEASURED BY THE MACRO-AVERAGED $f_1$-SCORES FOR THE THREE DIFFERENT LEVELS OF AGGREGATION OF THERAPY CLASSES.

| **Number of distinct therapies** | **8** | **12** | **15** |
|---|---|---|---|
| **XGBoost** | 0.21 | 0.19 | 0.15 |
| **Random Forest** | 0.23 | 0.20 | 0.17 |
| **Logistic Regression** | 0.17 | 0.16 | 0.13 |
| **Linear Support Vector Classifier** | 0.17 | 0.16 | 0.15 |
| **Decision Tree** | 0.19 | 0.14 | 0.14 |
| **Dummy Classifier** | 0.09 | 0.05 | 0.02 |

divergence between the AI-based prediction and the treatment decision. The figures show the most important Shapley values for the prediction of 5-FU monotherapy and doublet chemotherapy, respectively. For patient 1, similar to the true positives in Figure 5, increased age (85 years) was speaking for 5-FU monotherapy. We can see in Figure 7 that this factor is reversed, speaking against the treatment with doublet chemotherapy. In patient 2, missing grading status and time since primary diagnosis were factors favoring doublet chemotherapy. In addition, presence of anemia — which is also considered a surrogate for morbidity — was speaking against doublet chemotherapy. As outlined above, age is an important factor for treatment-decision making, but chronological age does not necessarily mirror the frailty status of a patient (fit elderly patients). Although we have information on clinical performance status for most patients in our dataset, other important factors also driving treatment decisions are not captured (e.g., patient preference). For instance, some patients may opt for a more intense treatment despite higher risk for side effects. Such factors outside our database might have driven the treatment decision. Interestingly, one would have assumed co-morbidities and clinical performance status to have more weight in the decision, but their effect are rather modest in either direction.

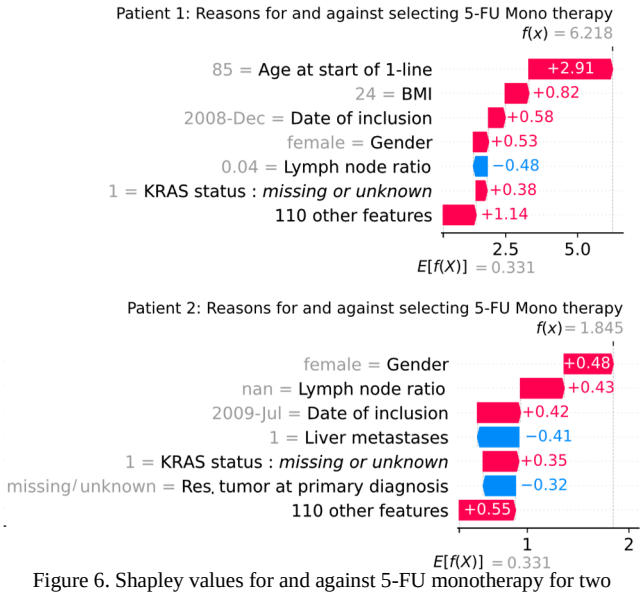


Figure 6. Shapley values for and against 5-FU monotherapy for two different representative patients for which the algorithm predicted 5-FU monotherapy when actually doublet chemotherapy was applied.

## C. *Experiment 3: Benefits of AI-based similarity metric*

Here, we analyze the benefits of using the proposed AI-based similarity metric. The goal is to show that this metric helps to identify patients that are similar in a meaningful way. A direct way to evaluate this would be to ask domain experts to assess the results. Here 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 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.

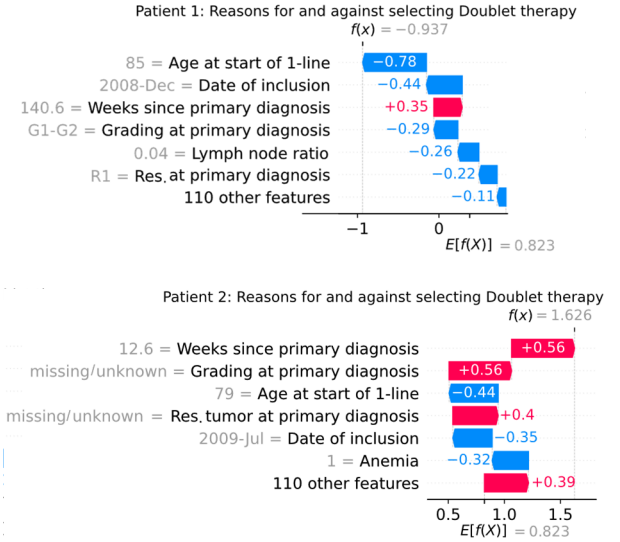


Figure 7. Shapley values for and against doublet chemotherapy for two representative patients where the algorithm predicted 5-FU monotherapy when actually doublet chemotherapy was applied.

### 1) *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 $k = 5$ and inverse distance weighting to the training data. Implementation is done with the KNeighborsClassifier from scikit-learn. For our baseline metric, we represent each patient by the vector $\mathbf{w} = (w_1, w_2, \ldots, w_p)$ of their features, with the same preprocessing as for logistic regression and SVC in Experiment 1. 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 performance of the two classifiers using the same metrics as in Experiment 1.

 ISBN: 978-1-68558-105-3 

*2) 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 II). Although improvements are small, 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, 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.

TABLE II. COMPARISON OF THE PERFORMANCE OF KNN CLASSIFIERS BASED ON THE SHAP-BASED METRIC AND THE BASELINE METRIC

| **Score type** | **Classifier** | **Number of distinct therapies** | | |
|---|---|---|---|---|
| | | 8 | 12 | 15 |
| $f_1$ (macro average) | KNN (Shapley) | 0.18 | 0.16 | 0.15 |
| | KNN (Baseline) | 0.16 | 0.13 | 0.12 |
| $f_1$ (weighted average) | KNN (Shapley) | 0.49 | 0.43 | 0.23 |
| | KNN (Baseline) | 0.49 | 0.38 | 0.22 |
| Accuracy | KNN (Shapley) | 0.54 | 0.46 | 0.25 |
| | KNN (Baseline) | 0.55 | 0.42 | 0.24 |

## D. Experiment 4: Impact of data availability

The amount of training data impacts the performance of ML models, but strength of this impact varies from case to case. Here, we analyze this effect for the case of CRC therapy prediction.

*1) Experimental Setup*

For our experiments we have in total 3,563 patient records available (see Experiment 1 – Experimental setup). To investigate the effect of 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 training set size). We then computed performance measures ($f_1$-score for one-versus-all) on the initially held out test sets for each model and therapy class. The process was repeated 10 times with different random seeds.

*2) Evaluation*

Figure 8 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

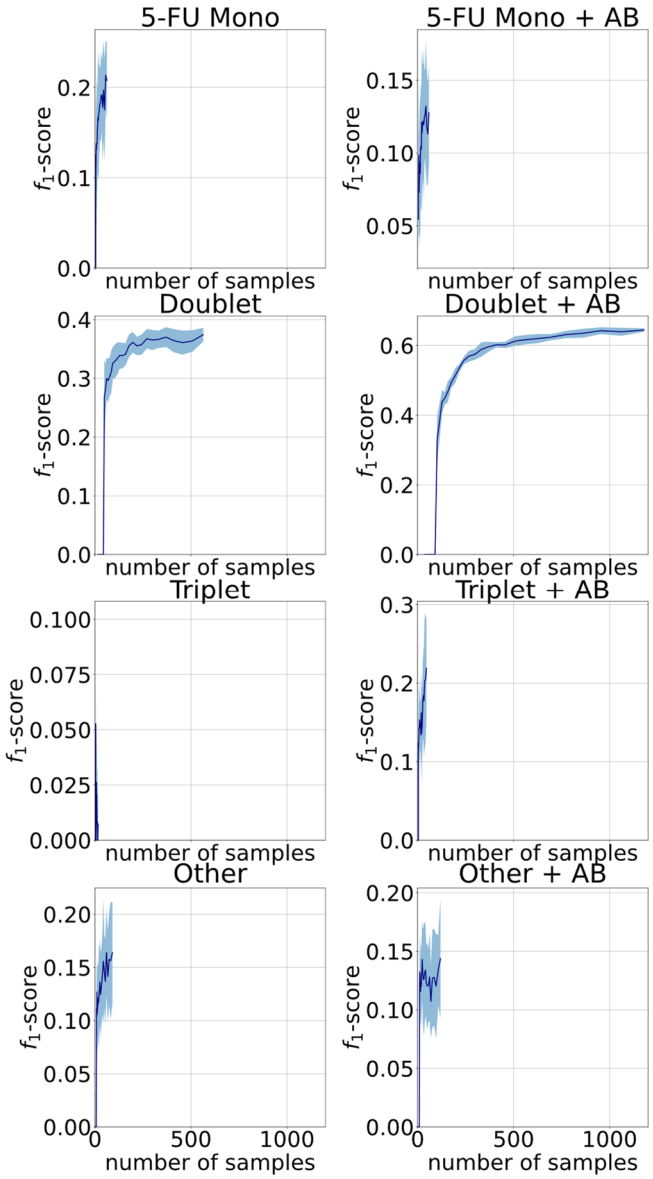


Figure 8. Impact of the number of training samples of a given therapy class on the model's performance, measured in terms of $f_1$-score for one-vs-all. Dark blue is the mean value and light blue is the standard deviation.

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 saturation. Note that for several therapies, the number of training examples is rather small, resulting in poor prediction performance of the model for those therapies. However, the shape of the curve suggests that improvements with larger training sets may be possible. We draw two main conclusions from this analysis. One is that performance results for therapy prediction (observed in Experiment 1) would likely improve with additional training data. The other conclusion is that about 500 training instances per therapy may be sufficient for the chosen ML setup.

## VI. CONCLUSION

In this paper, we analyzed the potential of using AI to build an information tool that enables dynamic data exploration and analysis of RWD. 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 RWD and thus make this broad knowledge accessible to individual oncologists. For these use cases, we proposed a system setup using supervised learning and XAI techniques.

We have shown applicability of the concept and obtained insights on the required amount of training data, but additional work should be done to assess viability of our approach. While we have demonstrated superiority of the AI-based approach against baseline methods, our experiments show a certain degree of disagreement between predicted and chosen therapies. However, disagreement is 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 which should be addressed in future work. One challenge is posed by the fact that the 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, best practice about therapy decision may change over time and change the probability of selecting a therapy for a given patient. This concept drift makes the algorithm prone to the cold-start problem of AI-based recommender systems. Solutions to this problem could involve non-uniform weighting of observations based on treatment date or incorporation of expert knowledge.

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.

Furthermore, feature selection 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. Incorporation of feature reduction techniques may therefore lead to better results. We also note that, while Shapley values are a measure for the impact of a given feature on a model's prediction for a given subject, they do not imply causation.

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 expansion to further use cases in the future.

## REFERENCES


[1] H. Ziekow, N. Marschner, D. Klein, B. Kasenda, and N. Haug, "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." 2022. Accessed: Oct. 04, 2023. [Online]. Available: https://opus.hs-furtwangen.de/frontdoor/index/index/docId/8454

[2] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, in NIPS'17. Red Hook, NY, USA: Curran Associates Inc., Dec. 2017, pp. 4768–4777.

[3] R. T. Sutton, D. Pincock, D. C. Baumgart, D. C. Sadowski, R. N. Fedorak, and K. I. Kroeker, "An overview of clinical decision support systems: benefits, risks, and strategies for success," *npj Digit. Med.*, vol. 3, no. 1, p. 17, Dec. 2020, doi: 10.1038/s41746-020-0221-y.

[4] T. N. T. Tran, A. Felfernig, C. Trattner, and A. Holzinger, "Recommender systems in the healthcare domain: state-of-the-art and research issues," *J Intell Inf Syst*, vol. 57, no. 1, pp. 171–201, Aug. 2021, doi: 10.1007/s10844-020-00633-6.

[5] C. D. Mahoney, C. M. Berard-Collins, R. Coleman, J. F. Amaral, and C. M. Cotter, "Effects of an integrated clinical information system on medication safety in a multi-hospital setting," *American Journal of Health-System Pharmacy*, vol. 64, no. 18, pp. 1969–1977, Sep. 2007, doi: 10.2146/ajhp060617.

[6] F. Gräßer *et al.*, "Therapy Decision Support Based on Recommender System Methods," *Journal of Healthcare Engineering*, vol. 2017, pp. 1–11, 2017, doi: 10.1155/2017/8659460.

[7] F. Gräßer, F. Tesch, J. Schmitt, S. Abraham, H. Malberg, and S. Zaunseder, "A pharmaceutical therapy recommender system enabling shared decision-making," *User Model User-Adap Inter*, pp. 1019–1062, Aug. 2021, doi: 10.1007/s11257-021-09298-4.

[8] S. P. Somashekhar, "Watson for Oncology and breast cancer treatment recommendations: agreement with an expert multidisciplinary tumor board," *Annals of Oncology*, vol. 29, no. 2, pp. 418–423, 2018, doi: https://doi.org/10.1093/annonc/mdx781.

[9] Z. Jie, Z. Zhiying, and L. Li, "A meta-analysis of Watson for Oncology in clinical application," *Sci Rep*, vol. 11, no. 1, Art. no. 5792, Mar. 2021, doi: 10.1038/s41598-021-84973-5.

[10] N. Zhou *et al.*, "Concordance Study Between IBM Watson for Oncology and Clinical Practice for Patients with Cancer in China," *The Oncologist*, vol. 24, no. 6, pp. 812–819, Jun. 2019, doi: 10.1634/theoncologist.2018-0255.

[11] Y. Nohara, K. Matsumoto, H. Soejima, and N. Nakashima, "Explanation of Machine Learning Models Using Improved Shapley Additive Explanation," in *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, in BCB '19. New York, NY, USA: Association for Computing Machinery, Sep. 2019, p. 546. doi: 10.1145/3307339.3343255.

[12] A. Moncada-Torres, M. C. Van Maaren, M. P. Hendriks, S. Siesling, and G. Geleijnse, "Explainable machine learning can outperform Cox regression predictions and provide insights in breast cancer survival," *Sci Rep*, vol. 11, no. 1, Art. no. 6968, Mar. 2021, doi: 10.1038/s41598-021-86327-7.

[13] R. O. Alabi, M. Elmusrati, I. Leivo, A. Almangush, and A. A. Mäkitie, "Machine learning explainability in nasopharyngeal

cancer survival using LIME and SHAP," *Sci Rep*, vol. 13, no. 1, Art. no. 8984, Jun. 2023, doi: 10.1038/s41598-023-35795-0.

[14] S. M. Lundberg, G. G. Erion, and S.-I. Lee, "Consistent Individualized Feature Attribution for Tree Ensembles," Feb. 2018, Accessed: Oct. 04, 2023. [Online]. Available: http://arxiv.org/abs/1802.03888

[15] A. Cooper, O. Doyle, and A. Bourke, "Supervised Clustering for Subgroup Discovery: An Application to COVID-19 Symptomatology," in *Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, M. Kamp, I. Koprinska, A. Bibal, T. Bouadi, B. Frénay, L. Galárraga, J. Oramas, L. Adilova, G. Graça, et al., Eds., in Communications in Computer and Information Science. Cham: Springer International Publishing, 2021, pp. 408–422. doi: 10.1007/978-3-030-93733-1_29.

[16] P. Bossaerts and C. Murawski, "Computational Complexity and Human Decision-Making," *Trends in Cognitive Sciences*, vol. 21, no. 12, pp. 917–929, Dec. 2017, doi: 10.1016/j.tics.2017.09.005.

[17] S. Khozin, G. M. Blumenthal, and R. Pazdur, "Real-world Data for Clinical Evidence Generation in Oncology," *J Natl Cancer Inst*, vol. 109, no. 11, pp. 1–5, Nov. 2017, doi: 10.1093/jnci/djx187.

[18] N. Marschner *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*, vol. 7, pp. 295–303, 2015, doi: 10.2147/CLEP.S73857.

[19] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco California USA: ACM, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[20] R. Shwartz-Ziv and A. Armon, "Tabular data: Deep learning is not all you need," *Information Fusion*, vol. 81, pp. 84–90, May 2022, doi: 10.1016/j.inffus.2021.11.011.

[21] C. Molnar, *Interpretable Machine Learning*. Accessed: Oct. 04, 2023. [Online]. Available: https://christophm.github.io/interpretable-ml-book/

[22] World Health Organisation, "The SuRF report 2: surveillance of chronic disease risk factors." Accessed: Oct. 04, 2023. [Online]. Available: https://iris.who.int/bitstream/handle/10665/43190/9241593024_eng.pdf