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
iOMEDICO AG
Biostatistics
Freiburg, Germany
nina.haug@iomedico.com

Resume

2013 – 2017	**PhD (Applied Mathematics)** Queen Mary University of London London, UK
2017 – 2020	**Postdoctoral Researcher** Center for Medical Data Science Medical University of Vienna, Vienna, Austria
2017 – 2020	**Resident Scientist** Complexity Science Hub Vienna, Vienna, Austria
2020 – 2021	**Data Scientist** Wiener Linien, Vienna, Austria
since 2021	**Senior Data Scientist — Statistics** iOMEDICO AG, Freiburg, Germany

Background

- Knowledge in oncology expands rapidly, with an estimated **175,000** new research papers published in 2020 alone
- **Medical guidelines,** summarize results from many research papers, are the main source of information regarding therapy standards for physicians
- Until present, medical knowledge mainly comes from **randomized controlled trials (RCTs)** with **strict in- and exclusion criteria**
- Patients included into RCTs are often **not representative** for patients encountered during routine clinical care
- Results from RCTs are only made available with a considerable **time lag**, in the form of medical guidelines
- RCT results suffer from **publication bias**

Background

- At iOMEDICO, we collect data about treatments in routine clinical care — **Real World Data (RWD)**
- They also capture information on treament decisions and outcomes for patients who would not be included into an RCT (e.g., very old patients)
- RWD thus contain a large amount of **latent knowledge**

Can AI and XAI be used to make the latent knowledge contained in RWD accessible to the treating physician?

Research questions

Can AI predict what treatment a clinician would give to a colorectal cancer patient?

Can XAI methods render the reasoning of the AI model interpretable?

Can AI techniques be used to define a meaningful distance metric for patients?

How does data availability impact performance of AI-based therapy prediction?

Our data

Tabular data on $n = 3{,}563$ patients treated for advanced colorectal cancer between 2006 and 2018, including

- patient demographics (e.g. age, sex)
- concomittant diseases (e.g. diabetes mellitus, hypertension)
- disease characteristics (e.g. tumor size, metastasis locations)
- prior therapies (e.g. surgeries, curative chemotherapy)
- palliative first-line therapy (can be grouped by principle)

Our data

Tabular data on $n = 3{,}563$ patients treated for advanced colorectal cancer between 2006 and 2018, including

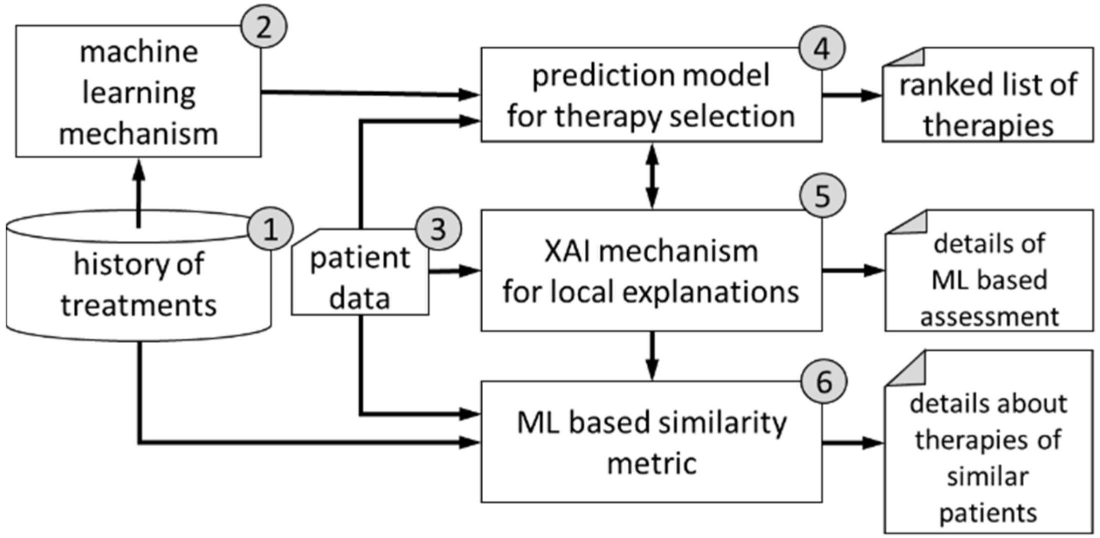
- patient demographics (e.g. age, sex)
- concomittant diseases (e.g. diabetes mellitus, hypertension)
- disease characteristics (e.g. tumor size, metastasis locations)
- prior therapies (e.g. surgeries, curative chemotherapy)

67 variables = predictors

- palliative first-line therapy (can be grouped by principle)

target

Architecture and key components


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

Experiments

1) We test how well AI models can predict therapy selection for advanced colorectal cancer patients

2) We use Shapley values to render the predictions from the algorithm explainable and discuss several examples

3) We define a similarity metric between patients based on Shapley values and test the performance of the metric in a KNN-classifier compared to a baseline metric

4) We evaluate the dependency of the amount of training data on the quality of AI-based therapy predictions

Experiment 1 — Therapy prediction

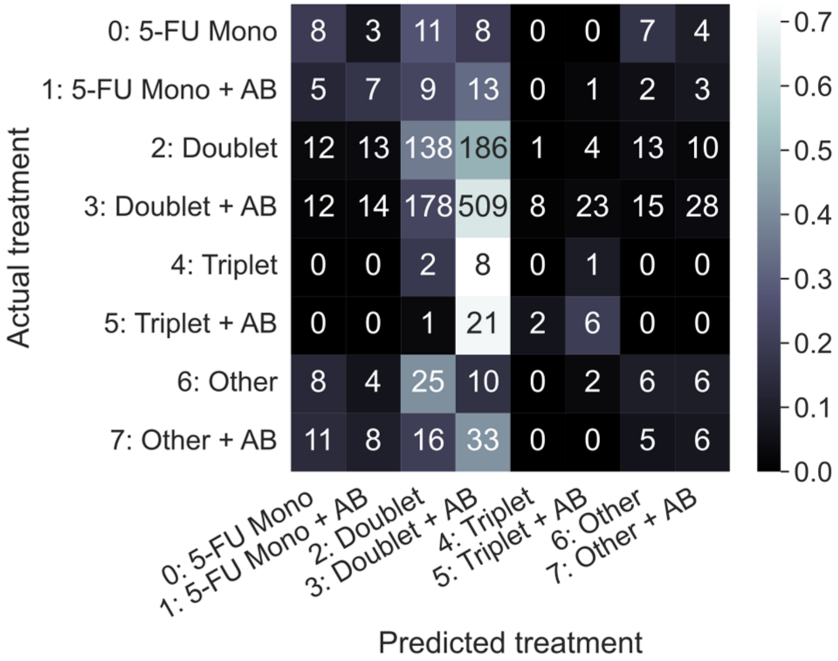
- We selected a stratified random set of 60% of our data for training
- We trained an XGBoost classifier with balanced class weighting to predict therapy decisions based on patient and disease characteristics. Hyperparameters were selected using Bayesian optimization
- As benchmark methods we used a random forest, multinomial logistic regression, a linear support vector classifier, a decision tree and a dummy classifier
- We evaluated the quality of predictions on the 40% of data held out for testing, using macro-averaged f_1 score, and plotted confusion matrices and ROC curves

Experiment 1 — Therapy prediction

Classifier	macro-averaged f_1
XGBoost	0.21
Random Forest	**0.23**
Logistic Regression	0.17
Linear Support Vector Classifier	0.17
Decision Tree	0.19
Dummy Classifier	0.09

Table: Performance of different algorithms in therapy prediction — case of 8 different therapy classes.

Experiment 1 — Therapy prediction



	0: 5-FU Mono	1: 5-FU Mono + AB	2: Doublet	3: Doublet + AB	4: Triplet	5: Triplet + AB	6: Other	7: Other + AB
0: 5-FU Mono	8	3	11	8	0	0	7	4
1: 5-FU Mono + AB	5	7	9	13	0	1	2	3
2: Doublet	12	13	138	186	1	4	13	10
3: Doublet + AB	12	14	178	509	8	23	15	28
4: Triplet	0	0	2	8	0	1	0	0
5: Triplet + AB	0	0	1	21	2	6	0	0
6: Other	8	4	25	10	0	2	6	6
7: Other + AB	11	8	16	33	0	0	5	6

Figure: Confusion matrix for therapy prediction by XGBoost classifier (8 therapy classes)

Experiment 1 — Therapy prediction

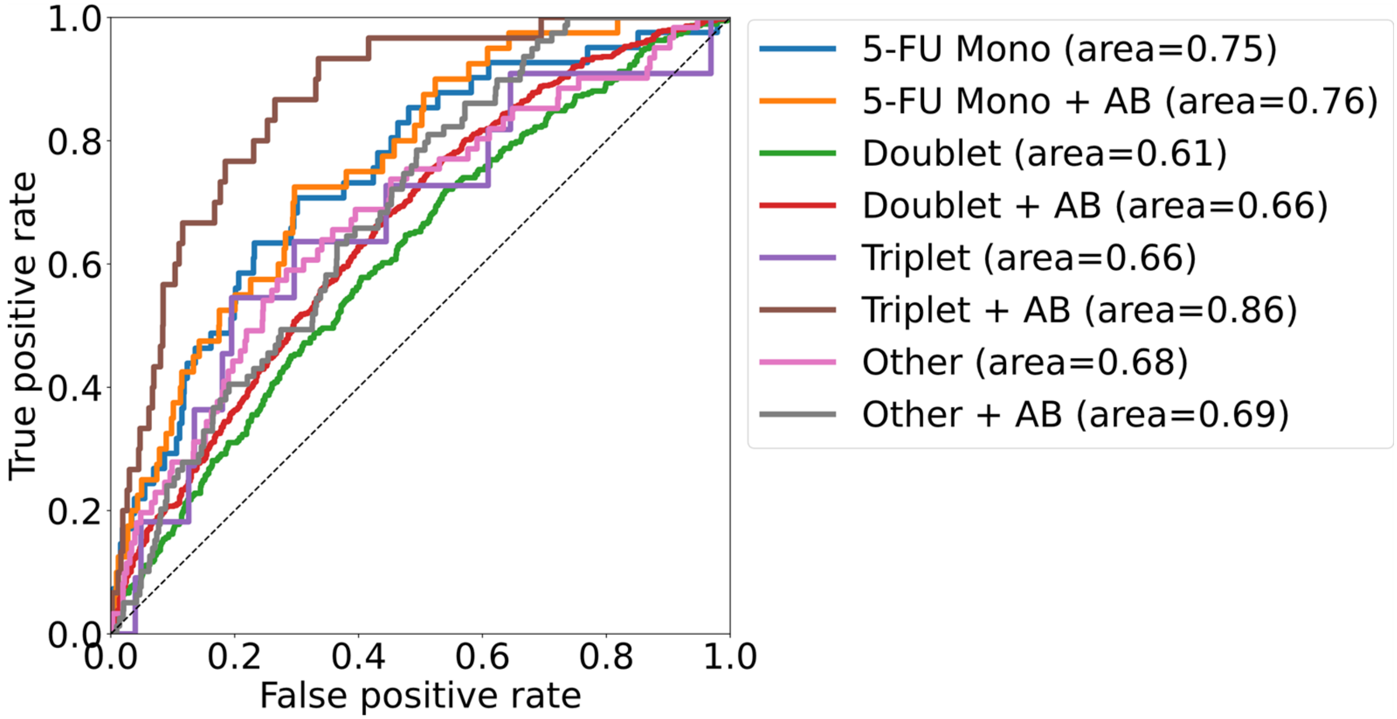


Figure: ROC curves for therapy prediction by XGBoost classifier.

Experiment 2 — Insights with feature importance measures

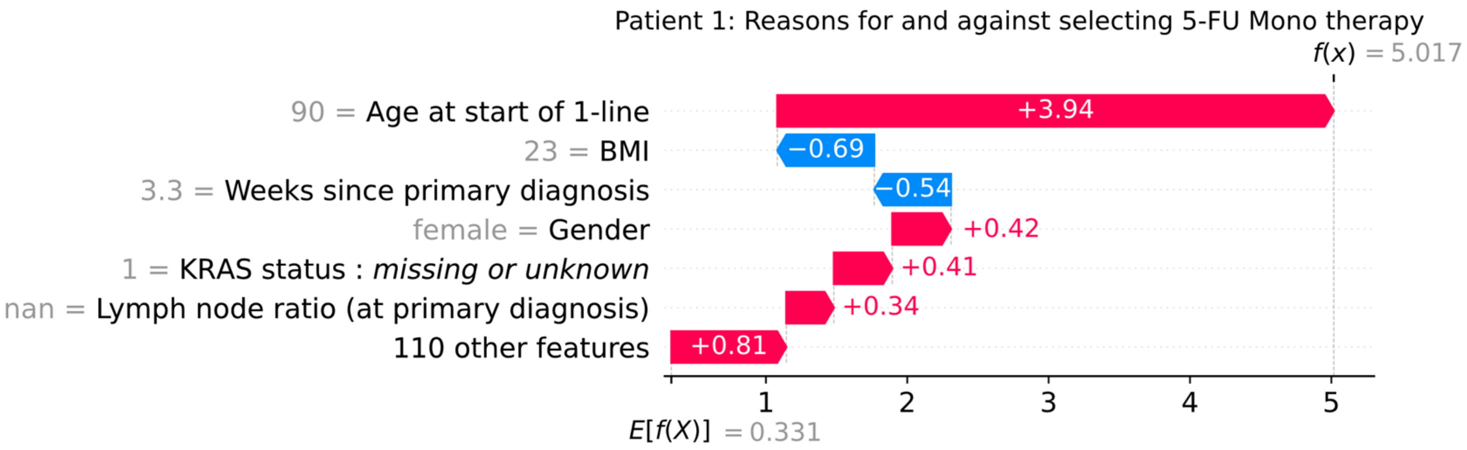


Figure: Shapley values for and against 5-FU mono therapy for a patient where the algorithm correctly predicted 5-FU mono therapy.

Experiment 2 — Insights with feature importance measures

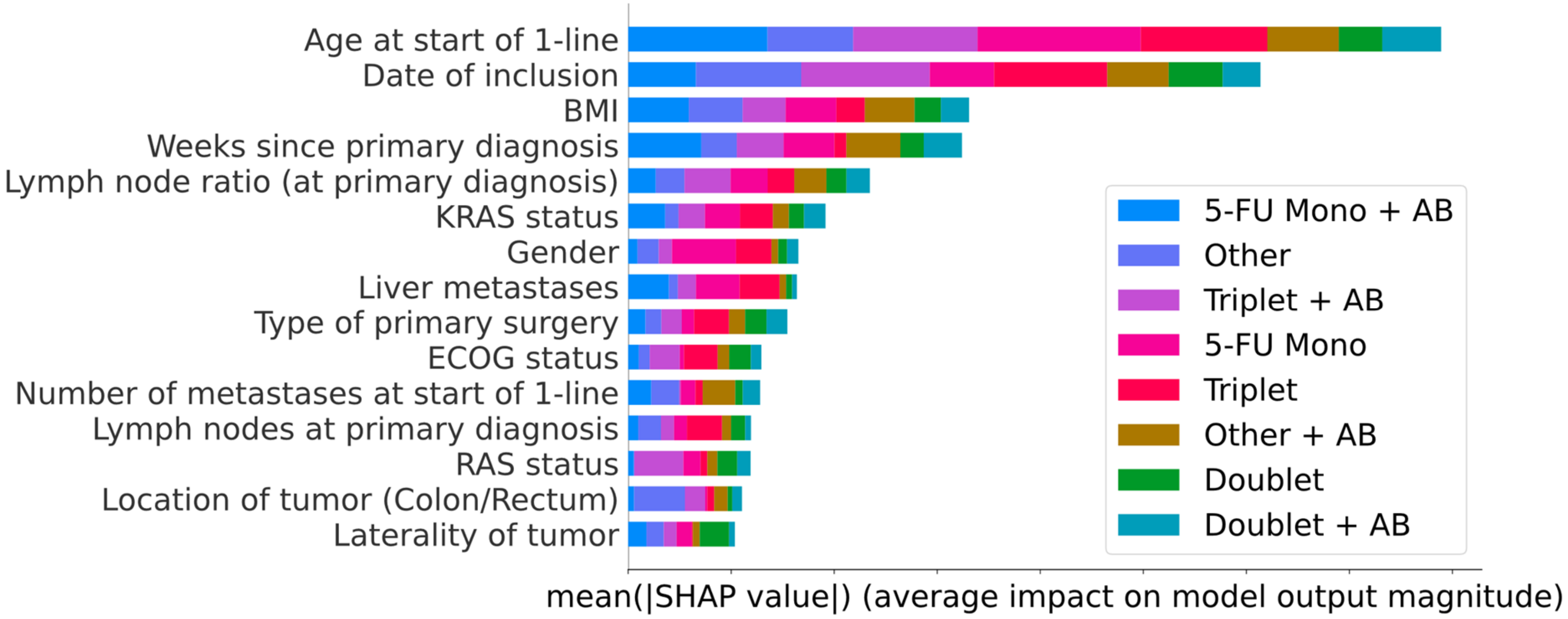


Figure: The 15 most important features for therapy prediction, with importance measured in terms of their global Shapley value.

Experiment 3 — Benefits of AI-based similarity metric

- We represent each patient by a vector $\boldsymbol{v} = (v_1, v_2, \dots, v_m)$, with $m = p \cdot n$. Here, p is the number of patient features and n is the number of target classes
- For each therapy class k and feature j, the entry $v_{(k-1)\cdot p+j}$ is the Shapley value of feature j for the one-vs-all prediction of therapy class k
- The distance between two patients with vector representations $v^{(1)}$ and $v^{(2)}$ is then defined as $d = \left\|\boldsymbol{v}^{(1)} - \boldsymbol{v}^{(2)}\right\|_1$ (the Manhattan distance)
- For a benchmark metric, we represent each patient by a vector $\boldsymbol{w} = \left(w_1, w_2, \dots, w_p\right)$, for a feature j, the entry w_j is the value of feature j (with categorical variables one-hot encoded) and use Manhattan distance
- We test the performance in therapy prediction of two KNN-classifiers based on the Shapley-based metric and the benchmark metric, respectively

Experiment 3 — Benefits of AI-based similarity metric

Score type	Classifier	Score value
f_1 (macro average)	KNN (Shapley)	0.18
	KNN (Baseline)	0.16
f_1 (weighted average)	KNN (Shapley)	0.49
	KNN (Baseline)	0.49
Accuracy	KNN (Shapley)	0.54
	KNN (Baseline)	0.55

Experiment 4 — Impact of data availability

- From the 3,563 patients, we set aside a fixed random subset of 40% for testing
- We iteratively took 90% stratified subsets of the remaining 60% of the data and on each subset, we fitted an XGBoost classification model
- For each model and therapy class, we tested performance of the trained classifier on the test set (f_1 score for one-vs-all)
- The process was repeated 10 times with different random seeds

Experiment 4 — Impact of data availability

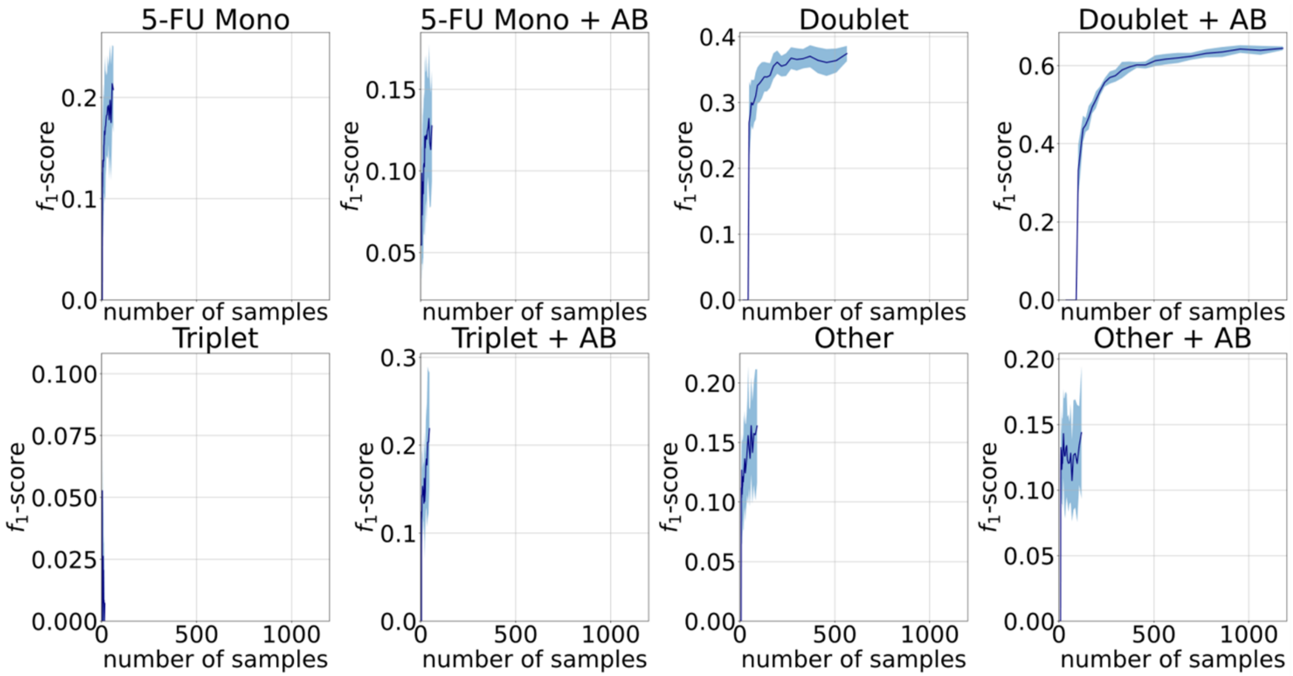


Figure 7: Impact of the number of training samples of a given therapy class on the model’s performance in labeling these samples.

Summary and conclusion

- Tree-based methods performed better than other statistical and ML algorithms in predicting therapy decisions
- Classification performance was rather poor, but this may be expected since different experts may prescribe different treatments to the same patient
- We demonstrated how Shapley values can be used to render therapy predictions interpretable
- We assessed the impact of the number of training examples on prediction quality

Limitations

- Our approach learns therapy decisions from past records. However, the therapy landscape in oncology changes rapidly, leading to concept drift
- Therapy outcomes of patients, such as overall survival, were not considered
- Feature selection was not done in the present work. Due to the high number of features, the Shapley-based distance metric may suffer from the curse of dimensionality
- Shapley values are a measure for the impact of a feature on a prediction. However, they do not imply causality