Enhancing predictive imaging biomarker discovery through treatment effect analysis

Shuhan Xiao[1,2] Lukas Klein[3,4,5] Jens Petersen[1] Philipp Vollmuth[1,6,7],
Paul F. Jaeger[3,5*] Klaus H. Maier-Hein[1,2,5,8*]

[1]German Cancer Research Center (DKFZ) Heidelberg, Division of Medical Image Computing, Germany [2]Faculty of Mathematics and Computer Science, Heidelberg University, Germany [3]DKFZ Heidelberg, Interactive Machine Learning Group, Germany
[4]Institute for Machine Learning, ETH Zürich, Switzerland [5]DKFZ Heidelberg, Helmholtz Imaging, Germany
[6]Division for Computational Radiology Clinical AI (CCIBonn.ai), Clinic for Neuroradiology, University Hospital Bonn, Germany [7]Medical Faculty Bonn, University of Bonn, Germany
[8]Pattern Analysis and Learning Group, Department of Radiation Oncology, Heidelberg University Hospital, Germany
s.xiao@dkfz-heidelberg.de

Abstract

Identifying predictive covariates, which forecast individual treatment effectiveness, is crucial for decision-making across different disciplines such as personalized medicine. These covariates, referred to as biomarkers, are extracted from pre-treatment data, often within randomized controlled trials, and should be distinguished from prognostic biomarkers, which are independent of treatment assignment. Our study focuses on discovering predictive imaging biomarkers, specific image features, by leveraging pre-treatment images to uncover new causal relationships. Unlike labor-intensive approaches relying on handcrafted features prone to bias, we present a novel task of directly learning predictive features from images. We propose an evaluation protocol to assess a model's ability to identify predictive imaging biomarkers and differentiate them from purely prognostic ones by employing statistical testing and a comprehensive analysis of image feature attribution. We explore the suitability of deep learning models originally developed for estimating the conditional average treatment effect (CATE) for this task, which have been assessed primarily for their precision of CATE estimation while overlooking the evaluation of imaging biomarker discovery. Our proof-of-concept analysis demonstrates the feasibility and potential of our approach in discovering and validating predictive imaging biomarkers from synthetic outcomes and real-world image datasets. Our code is available at https://github.com/MIC-DKFZ/predictive_image_biomarker_analysis.

*These authors contributed equally to this work.

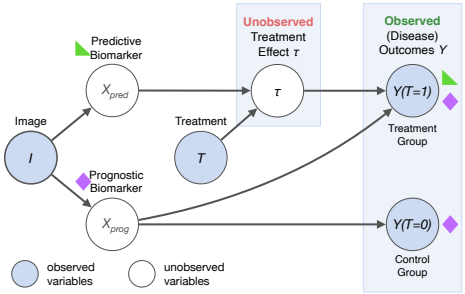

Figure 1. Relationship between biomarkers x_{prog} and x_{pred}, outcomes $Y(T)$ depending on the treatment T and the treatment effect τ. Since both potential outcomes $Y_i(T=0)$ and $Y_i(T=1)$ cannot be observed for the same individual simultaneously it is impossible to infer the individual treatment effect directly.

1. Introduction

Identifying predictive biomarkers is crucial for determining which subgroup of individuals will have a positive treatment effect and ultimately for making informed treatment decisions across different fields such as medical treatments, environmental strategies, and economic policies. Precision medicine, for example, relies on predictive biomarkers to tailor interventions to individual patients and ensure optimized patient outcomes. Generally, a biomarker is a measurable characteristic associated with an individual's outcome such as disease progression or physiologic measures [35]. Although the term originally stems from the biomedical field, we use it more broadly in this paper to refer to features or covariates in general contexts. A biomarker is predictive when it acts as a driver of treatment effect heterogeneity [12]. Predictive biomarkers are

treatment-specific, while prognostic biomarkers are associated with the outcome independent of treatment assignment [6], as illustrated in Fig. 1. The discovery of predictive biomarkers is key to not only explaining the causal mechanisms behind treatment effects and supporting informed treatment decisions but also to driving the development of novel treatments. In particular, there has been a growing interest in leveraging the vast amount of non-invasively acquired information provided by different imaging modalities to discover so-called imaging biomarkers, especially predictive imaging biomarkers [38].

In previous research, the discovery process of predictive imaging biomarkers involves handcrafted radiomics features (*e.g.* shape, intensity, and texture of tumors or lesions [8, 31, 33, 34]) as candidates to determine their predictive performance. This process typically contains several steps including segmentation to define regions of interest, feature extraction, and feature selection.

While machine learning approaches have been employed to facilitate the discovery of imaging biomarkers [9, 29, 33, 34, 36, 39, 40], the training processes rely on handcrafted radiomics-based features and have the risk of introducing human bias, as shown in [25]. Some approaches directly aim at discovering predictive biomarkers and distinguishing them from prognostic ones [5, 7, 43, 55], but are limited to tabular input data. More flexibility and adaptability are offered by deep learning (DL)-based conditional average treatment effect (CATE) estimation methods [3, 13, 22, 45, 46, 54], which have the potential to identify predictive biomarker candidates from a set of tabular covariates as well [7, 11]. CATE estimation differs from a standard supervised learning task and requires different modeling approaches as the ground truth for our quantity of interest – the individual treatment effect – is not available. This is due to the fundamental problem of causal inference [24]: It is impossible to observe both potential outcomes, treated and untreated, from the same individual simultaneously, yet they are necessary to compute the individual treatment effect. For CATE estimation, the presence of strong prognostic biomarkers, which is frequently encountered in practice, can negatively impact the performance of CATE estimators, even though they are not relevant for the treatment effect and, thus, treatment decision-making. For instance, CATE estimators can mistakenly identify prognostic as predictive biomarkers, as studies have shown [11, 23, 43], which may lead to ineffective or even harmful treatment recommendations. It is therefore essential to ensure that these methods can distinguish the two types of biomarkers.

CATE estimation methods have been originally designed for tabular inputs and remain a widely unexplored topic in the context of image inputs. In response to this gap, recent advancements have adapted DL-based CATE estimation methods to estimate treatment effects not only from medical images [16, 17, 27, 37] but also other types of images [26, 28, 50]. Yet, none of these image-based methods directly describe how predictive biomarkers can be identified and interpreted or address how well models manage to do so, which is an important but often overlooked performance metric to consider when evaluating CATE estimation methods, as noted in [12]. To conduct such an evaluation, a benchmarking environment was proposed in [11], albeit only applicable to tabular data.

Adapting the evaluation of predictive biomarker discovery from tabular data to images introduces a significant challenge: Extracting imaging biomarkers is complicated by the high-dimensional and structured nature of image data, which lacks distinct, pre-defined features. Consequently, a critical step in interpreting these biomarkers is determining the specific image features on which a black-box CATE estimation model depends. This step is also vital for drug development and clinical decision-making.

In this paper, we define a novel task in response to the challenges above: discovering predictive imaging biomarkers directly from image data in a data-driven way, without requiring handcrafted features or a separate feature extraction step. We introduce a new evaluation protocol tailored to this task and demonstrate as a proof-of-concept how a DL-based CATE estimation model can be applied in practice (Fig. 2). Our evaluation protocol includes two components: (1) statistical testing to investigate the estimated predictive biomarker strength, and (2) explainable artificial intelligence (XAI) methods [18, 44, 47–49] to enable the verification and interpretation of the discovered predictive imaging biomarker candidates.

We also propose and conduct experiments to validate our evaluation protocol on real image data using pre-defined imaging biomarkers with varying strengths of predictive and prognostic effects on synthetic outcomes. This setup, for benchmarking and model development, enables assessing a model's ability to identify and interpret predictive imaging biomarkers. Experiments on natural and medical images highlight the potential of an image-based CATE estimator to address our task, showcasing the model's capability to identify predictive imaging biomarkers with greater predictive strength compared to a baseline that does not distinguish between prognostic and predictive effects.

2. Methods

2.1. Treatment heterogeneity and predictive biomarkers

We describe how treatment effects, which cannot be observed directly, can be estimated from data by introducing the concept of potential outcomes. Here, we consider pre-treatment images and data collected through randomized controlled trials (RCTs), the typical experimental setting for

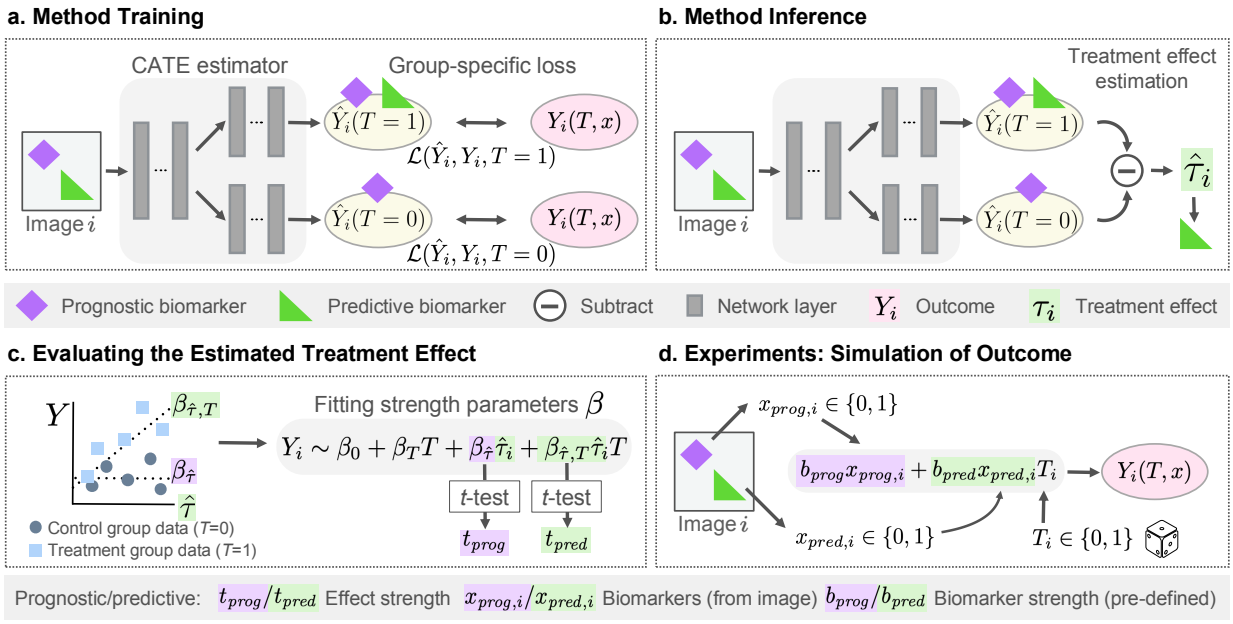

Figure 2. Overview of the identification of predictive biomarkers from pre-treatment images. The (**a**) training and (**b**) inference step employs a two-headed architecture to estimate treatment effects $\hat{\tau}$ from images. In the evaluation step (**c**) the predictive strength of the estimated $\hat{\tau}$, the predictive biomarker candidate, is assessed using regression. In our simulation experiments (**d**), the outcome data Y_i used in our experiments are simulated with image features from ground truth annotations and randomly assigned treatments T.

discovering biomarkers. The relation between outcomes, defined by a problem-specific measure of interest, and treatment effect has been described by the Neyman-Rubin causal model [42], where the individual treatment effect (ITE) for an individual i is defined as the difference between potential outcomes $Y_i(T)$, ITE $:= Y_i(T=1) - Y_i(T=0)$. Here, we assume a binary treatment variable $T \in \{0,1\}$ for whether a treatment is applied or not. In RCTs, T is randomly assigned and indicates whether an individual belongs to the control group ($T=0$) or treatment group ($T=1$). Since it is not possible to observe the counterfactual outcomes and thus measure the ITE due to the fundamental problem of causal inference, in practice, the conditional average treatment effect (CATE) τ

$$\tau(x) := \mathbb{E}\left[Y(T=1) - Y(T=0) | X = x\right] \qquad (1)$$

is estimated instead. The CATE depends on observable pre-treatment covariates $x \in X$, which can for example be extracted from images I. While such covariates that measure image features are often called imaging biomarkers in biomedical applications, we use "imaging biomarkers" as a more general term. Only heterogeneous treatment effects, *i.e.* effects that vary among individuals and covariates x, are relevant for making treatment decisions or subgroup selection. Therefore, we are interested in identifying covariates that directly contribute towards the heterogeneous treatment effect and interact with the treatment, also known as predictive biomarkers. Under the common assumption that prognostic effects f_{prog} and predictive effects f_{pred} are additive [12, 23, 32, 43] as in

$$\mathbb{E}\left[Y(x)\right] = f_{prog}(x) + f_{pred}(x)T, \qquad (2)$$

the CATE defined in Eq. (1) yields $f_{pred}(x)$, which only depends on predictive biomarkers x_{pred}. In this case, treatment effect estimation automatically separates prognostic and predictive effects and thus identifies predictive biomarkers x_{pred}. Generally, a biomarker can be both prognostic and predictive at the same time if it contributes to both $f_{prog}(x)$ and $f_{pred}(x)$. Figure 1 depicts the relationship between biomarkers x_{pred} or x_{prog} and outcomes Y.

2.2. Image-based treatment effect estimator

To enable the discovery of predictive imaging biomarkers, we leverage neural network-based CATE estimators adapted for image inputs. For our experiments, we modify a TARNet model [45], originally designed for tabular inputs, similar to the adaptation described in [17]. The network has shared convolutional layers as encoders for learning the similarities between the control and treatment groups arising from prognostic effects [14], and two treatment-specific heads for predicting the outcomes $Y(T)$. During the training (Fig. 2a), we apply the loss to the corresponding head, depending on which RCT group the input data belongs to. In each training step, the total loss is the sum of the loss

of the control group head output and the treatment group head output, so that the weights of both heads are updated. During inference (see Fig. 2b), the CATE is estimated by subtracting the model's control group output from the treatment group output: $\hat{\tau} = \hat{Y}_i(T=1) - \hat{Y}_i(T=0)$.

In contrast to the two-headed model, we expect a single-head model to learn to predict the average outcome across groups from both predictive and prognostic biomarkers and not differentiate between the treatment group or control group. The predicted outcome of such a network is the composition of both predictive and prognostic effects. It is used as a baseline to validate whether the CATE estimator could successfully discover a predictive biomarker. Implementation details are described in the Supplementary section **A.4**.

2.3. Proposed evaluation protocol

2.3.1 Statistical evaluation of the predictive strength

To verify whether the model has identified a predictive effect – that is, whether the estimated CATE $\hat{\tau}$ is indeed predictive and can be considered a predictive biomarker candidate – we test the interaction between biomarker candidate and treatment, as seen in Fig. 2c. Such an evaluation is also performed in clinical practice [6, 41]. We assume a linear relationship between biomarkers and outcome (Eq. (2)) and perform a linear regression of the outcomes Y using

$$\beta_0 + \beta_T\, T + \beta_{\hat{\tau}}\, \hat{\tau} + \beta_{\hat{\tau},T}\, \hat{\tau}\, T \sim Y, \tag{3}$$

which includes an interaction term $\beta_{\hat{\tau},T}\, \hat{\tau}$ and coefficients β_i. We test the null hypothesis that the biomarker-treatment interaction coefficient is $\beta_{\hat{\tau},T} = 0$ using the Student's t-test with the t-value $t_{\beta_{\hat{\tau},T}}$ test statistic, which is proportional to the estimated $\hat{\beta}_{\hat{\tau},T}$. This test is additionally repeated with the other fit coefficients β_i. The t-value ratio $t_{\beta_{\hat{\tau},T}}/t_{\beta_{\hat{\tau}}} =: t_{pred}/t_{prog}$ can be used as an indicator for the predictive strength of the estimated CATE $\hat{\tau}$ compared to its prognostic strength. To estimate the experimental lower (indicating a prognostic biomarker) and upper (indicating a predictive biomarker) bound for the relative predictive strength, we conduct the same evaluation, replacing $\hat{\tau}$ in Eq. (3) with either the purely prognostic or a purely predictive ground truth biomarker $x_{prog,pred}$.

2.3.2 Interpretation using feature attribution methods

We also investigate which input image features the trained model is sensitive to when predicting the CATE $\hat{\tau}$ and whether they correspond to predictive imaging biomarkers. Since a direct quantitative assessment is not straightforward for general image features, unlike for tabular data, we rely on visual explanations through attribution maps [47] instead. To this end, we employ the XAI methods expected gradients (EG) [18] and guided gradient-weighted class activation mapping (GGCAM) [44, 48] to generate attribution maps from the trained model and input images. The attribution maps enable us to visually analyze how much individual pixels contribute to either the prognostic effect via the attribution map of the control group head prediction $\hat{Y}(T=0)$ or the predictive effect via the attribution map of the estimated CATE $\hat{Y}(T=1) - \hat{Y}(T=0)$.

2.4. Simulation of imaging biomarkers and outcomes for validation

To study the CATE estimator's ability to identify predictive imaging biomarkers in the presence of prognostic ones, we conduct experiments on data with varying predictive and prognostic biomarker strengths. Since ground truth counterfactual treatment outcomes are unavailable in real data, we generate synthetic data to experimentally verify the model and simulate the ground truth treatment outcomes (Fig. 2d). Our proposed approach simulates outcomes based on imaging biomarkers by assigning image features to biomarker values $x_{prog,pred}$ instead of simulating outcomes directly from tabular biomarkers. This entails selecting features from available image information such as attributes, class labels, or radiomics features, as shown in Fig. 3. In our examples, the biomarkers are either purely prognostic or predictive and may be binary or continuous depending on the dataset. The outcomes Y are then generated according to a simple linear function:

$$Y(T, x) = b_{prog} x_{prog} + b_{pred}\, x_{pred} T, \tag{4}$$

assuming no offset b_0 and constant treatment effect b_T for simplicity, similar to a case considered in [30]. An important aspect of using simulated outcomes is that we can control the size of prognostic or predictive effects by adjusting the parameters $b_{prog,pred}$. The biomarker parameter strength ratio b_{pred}/b_{prog} can be interpreted as a measure of the signal-to-noise ratio of the predictive effect in the input data. Here, in an RCT setting, the treatment variable $T \in \{0, 1\}$ is assigned with probabilities $p(T) = 0.5$.

2.5. Experimental Setup

2.5.1 Datasets and imaging biomarker features

We evaluate our CATE estimator on four diverse publicly available datasets also shown in Fig. 3: colored digits (MNIST [4, 15]) with semi-synthetic image features, images of birds (CUB-200-2011 [53]) as an example of a natural image dataset, as well as skin lesion images (ISIC 2018 [10, 51]) and 3D lung computed tomography (CT) scans of non-small cell lung cancer (NSCLC) tumors (NSCLC-Radiomics [2]) as real-world medical datasets.

Colored MNIST (CMNIST). We adapt the MNIST dataset and introduce color as an image feature. The color of the digits is determined based on the random variable x_i sampled from a binomial distribution (with $p = 0.5$). We

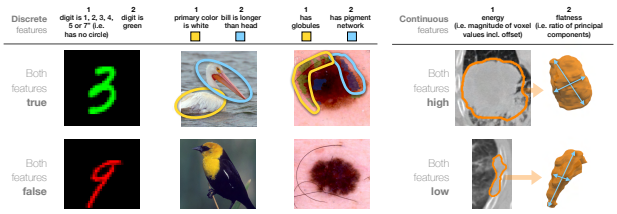

Figure 3. Image features from the four datasets, where either feature 1 or 2 is designated as predictive or prognostic biomarkers. ISIC 2018 skin lesion features are shown with ground truth masks. Globules (light green mask) manifest as darker dots, pigment networks have dark grid-like patterns of streaks with lighter "holes" (dark blue mask). The NSCLC-Radiomics images display tumor segmentation outlines of a 2D slice (left) or corresponding 3D volumes (right). Examples on the bottom row depict images where both biomarkers are either absent or have a low value.

define binary features as imaging biomarkers $x_{pred,prog} \in \{0, 1\}$: (a) the color (green or not green) as prognostic feature and whether digits lack or contain a circle or loop (*i.e.* $\{1, 2, 3, 4, 5, 7\}$ vs. $\{0, 6, 8, 9\}$) as the predictive feature or (b) vice versa. For intuition, a treatment might involve applying an image filter to alter the digit's appearance, while the outcome might be a digit classifier's confidence score.

Bird species dataset (CUB-200-2011). The dataset includes images of 200 bird species, 5,794 for testing and 5,994 for training, which we further split into training and validation data with an $80\%/20\%$ split. From the binary attributes of the birds, we select two visually distinct biomarkers $x_{pred,prog} \in \{0, 1\}$ with high annotator certainty: (a) "*has primary color: white*" as prognostic and "*has bill length: longer than head*" as the predictive feature or (b) vice versa. To illustrate, the imaging biomarkers here might relate to the bird's observed behavior as an outcome, and habitat modification might serve as the treatment.

Skin lesion dataset (ISIC 2018). The ISIC 2018 dataset contains skin lesion images with a designated training dataset of 2,594 images, which is split into a training and validation set of sizes 2,075 and 519 respectively. Final evaluations are performed on the designated validation set with 100 images. We identify dermoscopic attributes, *i.e.* visual skin lesion patterns, using ground truth segmentation masks and assign their presence to biomarkers. In feature set (a) the presence of globules is prognostic and the presence of a pigment network is predictive, or in (b) vice versa. Both features have been evaluated as imaging biomarkers for diagnosing melanoma [19, 20] making them realistic examples of biomarkers. Unlike the features of the previous datasets, these features are based on the presence of patterns rather than localized features or color values.

Lung cancer CT dataset (NSCLC-Radiomics). This dataset comprises 415 3D CT volumes of pre-treatment scans from NSCLC patients and ground truth segmentation masks of the lung tumors. We crop the volumes to the largest connected tumor volume bounding box, use 332 samples for 5-fold cross-validation, and reserve 83 for testing. We define two continuous, uncorrelated radiomics features described in [56] as biomarkers, which have both been evaluated for their prognostic or predictive value before [1, 8]: (a) the shaped-based feature "flatness" describing the ratio between the smallest and largest principal tumor components as a prognostic feature and the first-order statistics feature "energy" characterizing the sum of squares of tumor intensity values as a predictive feature or (b) vice versa. The flatness feature is inverse to the actual flatness of the tumor. Values close to 0 indicate flat shapes, whereas values close to 1 indicate sphere-like shapes. Energy depends strongly on both volume and minimum pixel intensity as the minimum intensity value is added as an offset. The radiomics features were extracted from the ground truth tumor segmentation volumes with PyRadiomics [52].

We split all datasets randomly into two equally sized subsets, a control ($T = 0$) and a treatment group dataset ($T = 1$), and generate group-specific outcomes $Y(T, x)$ according to Eq. (4). For each CMNIST feature, we choose the biomarker strength parameters $b_{pred,prog} \in \{0.0, 0.1, \ldots, 1.0\}$, resulting in training 121 models. For the remaining datasets, we choose the biomarker strength parameters $b_{pred,prog} \in \{0.0, 0.2, 0.4, 0.6, 0.8, 1.0\}$, resulting in 36 different trained models.

3. Results

3.1. Predictive strength of the estimated CATE

We present the results of our quantitative experimental validation protocol in Fig. 4, where the estimated relative predictive strength $|t_{pred}/t_{prog}|$ reflects its dependency on the relative size of the true predictive effect b_{pred}/b_{prog} in the outcome simulation described in Fig. 2. Across the four datasets, the CATE estimation model shows higher relative predictive strength t_{pred}/t_{prog} with higher relative predictive biomarker signal strength b_{pred}/b_{prog}, often surpassing the baseline models, especially for low b_{pred}/b_{prog}. While the results are similar for models (a) and (b), the difference is more pronounced for the other datasets, indicating a greater influence of the type of biomarkers.

Our model performs best on CMNIST among all four datasets, with a significantly larger gap from the baseline. For example, it reaches a factor of 10^2 for b_{pred}/b_{prog} in the range of 0 to 1, and has results much closer to the upper bound than the lower bound.

While the relative predictive strength for CUB-200-2011 is lower than on CMNIST, it remains above the lower and near the upper bound. For b_{pred}/b_{prog} between 0 and 1, the median t_{pred}/t_{prog} differs from the baseline by factors of 10 and 5 for sets (a) and (b), respectively. The dependency

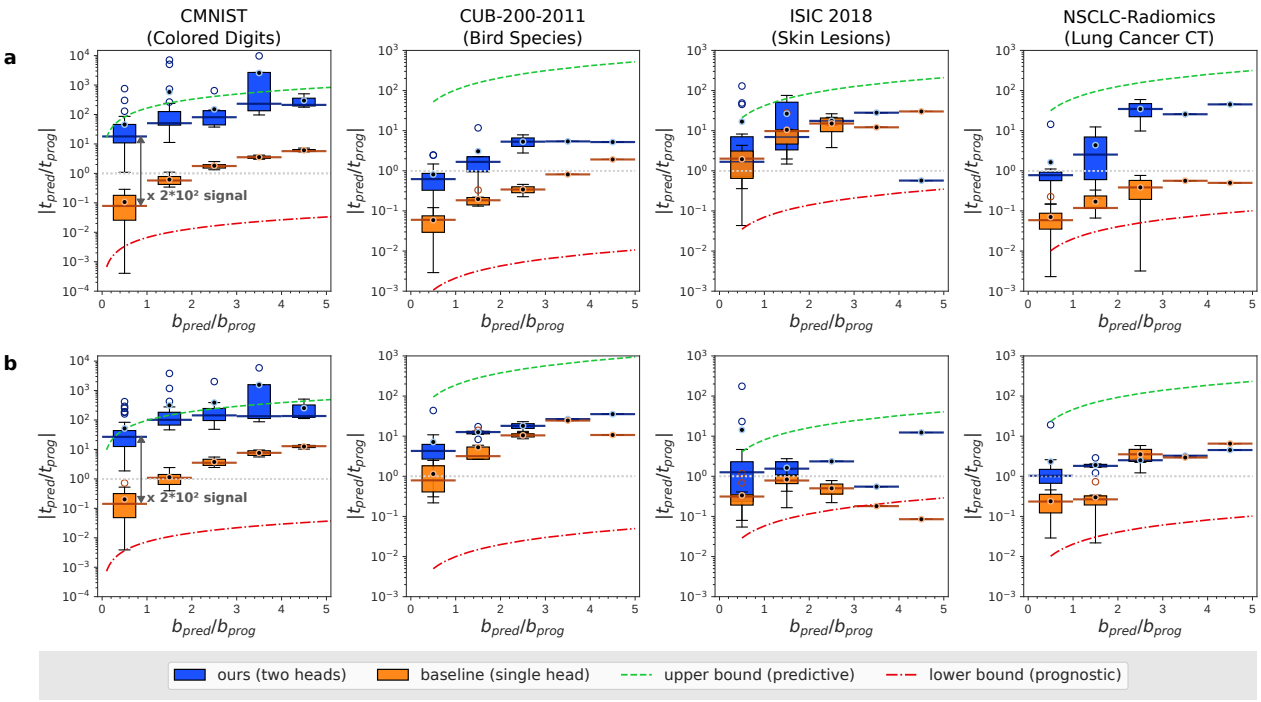

Figure 4. Model performance based on the relative predictive strength t_{pred}/t_{prog} of the CATE, shown on a logarithmic scale. We compare our two-headed CATE estimator with a one-headed baseline model across different simulation parameters b_{pred}/b_{prog} (*i.e.* relative size of the predictive effect in the simulated outcomes). Boxplots summarize data averaged over b_{pred}/b_{prog}-bin widths, indicated by the horizontal error bars over the median line. Rows (**a**) and (**b**) correspond to different sets of prognostic and predictive features used for generating the data (see Sec. 2.5.1 and Fig. 3). The variance of the boxplots is affected by the differing number of samples each bin contains.

on the biomarker choice is evident from the smaller gap between our model and the baseline in set (b) versus (a).

The ISIC 2018 results show smaller absolute t_{pred}/t_{prog} values, yet the relative predictive strength mean values remain above 1, except for two outliers at high b_{pred}/b_{prog}, based on a single sample. In set (a), the absolute t_{pred}/t_{prog} values are higher and much closer to the upper bound, but exhibit greater boxplot overlaps with the baseline for low b_{pred}/b_{prog} compared to set (b), where "has globules" is predictive. In set (b), the medians differ by a factor of 4 for relative b_{pred}/b_{prog} in the range of 0 to 1. The large baseline values suggest the baseline model also strongly relies on the predictive biomarker "has pigment networks".

On NSCLC-Radiomics, our model demonstrates larger t_{pred}/t_{prog} gaps between model and baseline, particularly for smaller b_{pred}/b_{prog}, with gaps decreasing slightly as b_{pred}/b_{prog} increases for set (b). The performance differs between biomarker sets (a) and (b), with medians of our models and baseline differing by a factor of 13 and 4 respectively for b_{pred}/b_{prog} in the range of 0 to 1.

3.2. Interpreting predictive imaging biomarkers

In Fig. 5, we illustrate our XAI-based evaluation scheme to assess whether the image features identified by our CATE estimation model as predictive or prognostic correspond to the ground truth biomarkers. By applying attribution methods [18,44,48,49] to our model and an input image, we generate an attribution map, indicating positive (blue) and negative (red) contributions to the prediction. We show attribution maps of the predicted CATE, $\hat{Y}(T=1) - \hat{Y}(T=0)$, which is expected to be sensitive only to the predictive biomarker (Fig. 2b), and the control group head, $\hat{Y}(T=0)$, which should be sensitive to the prognostic biomarker.

For CMNIST, the attribution maps of the predicted CATE $\hat{Y}(T=1) - \hat{Y}(T=0)$ show mostly negative attribution in the green channel of the first example, which corresponds to the absence of the predictive biomarker "has no circle" in the input image. Similarly, the treatment effect attribution maps for the second example (red digit four) show weaker negative attribution from the digit in the red channel with some noisy positive attribution in the background. More positive attribution is observed in the green channel, indicating that the model correctly infers that the predictive biomarker "has no circle" is present. The control group head output $\hat{Y}(T=0)$ correctly identifies the prognostic biomarker, *i.e.* "digit is green", in the respective color channel, which is evident from the mainly positive attribution in the green color channel in the first example and negative at-

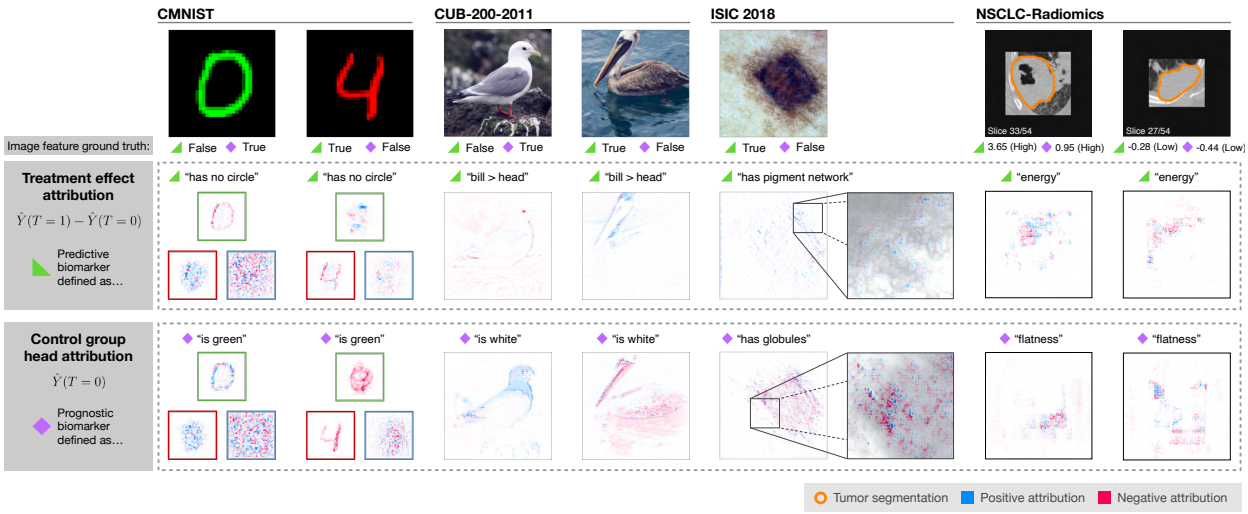

Figure 5. Attribution maps for the control group prediction head (last row) and the predicted CATE output (middle row) for different example images from each dataset (top row). For the CMNIST dataset, the attribution is shown for each RGB color channel (red: left, green: top, blue: right), as the color information is important for the biomarker prediction. An additional zoomed-in patch of the ISIC 2018 attribution map is overlaid with a grayscale version of the original image. For the NSCLC-Radiomics dataset, sagittal slices of the 3D patches are shown with orange outlines of segmented tumors. Here, results are based on models trained with $b_{pred}, b_{prog} = 1.0$.

tribution in the red color channel for the second example. For both outputs' attribution maps, only noisy attribution is present for the blue channel, suggesting that the model does not use this channel for prediction.

In the first CUB-200-2011 example, where the predictive biomarker "bill longer than head" is absent, the attribution map for $\hat{Y}(T=1) - \hat{Y}(T=0)$ is mostly negative and focusing on the eye and outlines of the throat and breast. The attribution is not as localized as in the second example, where the predictive biomarker is present and the overall attribution is positive. Here, features of the head are primarily used for the predictions, while the main body and wings are ignored, reinforcing the importance of the bill and head region for determining the predictive biomarker. The $\hat{Y}(T=0)$ attribution map shows overall positive attribution, especially in the white head and breast region from the first, primarily white bird. For the second bird, the attribution map is overall negative, particularly in the dark wing, main body, and pouch region. These patterns indicate that the model correctly identifies the presence or absence of the prognostic biomarker in the corresponding example.

The ISIC 2018 image shows a pigment network surrounding a darker center. Several patterns become apparent in the attribution map overlaid with the original image. However, the allocation of positive or negative attribution provides only limited insight, possibly due to the biomarker features' complexity. In the $\hat{Y}(T=1) - \hat{Y}(T=0)$ attribution map, positive attributions are given to the periphery surrounding the dark center where the pigment network is located. Notably, the model relies on the less pigmented gaps between the dark grid-like structures to detect the pigment network, suggesting that the gaps contain sufficient information for their detection. The $\hat{Y}(T=0)$ attribution map reveals that the model uses the dark lesion center for control group predictions, with red and blue spots indicating the model's search for the small globule dots.

In the first NSCLC-Radiomics example, the highest absolute values in the $\hat{Y}(T=1) - \hat{Y}(T=0)$ attribution maps are observed within the tumor area. While the attention maps show negative attributions in the darker tumor regions, positive attributions can be seen in the surrounding areas, indicating the presence of a strong predictive biomarker. This observation is consistent with the ground truth, where the energy value is comparably high, whereas mostly negative attributions are observed for the second example with a lower energy value. However, the attributions are mainly given to the areas outside the outline of the tumor, potentially due to the network's difficulty in correctly identifying the tumor boundary. The $\hat{Y}(T=0)$ attribution maps show strong attributions mainly outside the tumor outline, which relates to the prognostic biomarker flatness. Additionally, artifacts around the border suggest that the patch shapes contribute partially to the prediction. Further results and a more detailed qualitative XAI analysis can be found in the Supplementary section **A.2**.

4. Discussion

The results suggest that the estimated CATE used in our quantitative evaluation approach is a reliable measure both for the predictive effect and the predictive biomarker it-

self under the assumption of a linear biomarker-outcome relation. The experiments also highlight how an image-based CATE estimator can be employed to identify predictive biomarkers from our simulated data while not being affected by prognostic biomarkers across various types of biomarkers and input images. This was validated by comparing the relative predictive strength t_{pred}/t_{prog} to our experimental baseline as well as our experimental upper and lower bound in our proposed experiments. Even in scenarios where predictive effects are smaller than prognostic effects for $b_{pred}/b_{prog} < 1$, which is often observed in real-world data, the model demonstrated the ability to identify predictive imaging biomarkers.

However, for the specific image-based CATE estimator we used, weaker performance is observed for CUB-200-2011, ISIC 2018, and NSCLC-Radiomics, particularly when b_{pred}/b_{prog} is high and where t_{pred}/t_{prog} is close to the baseline. This may be due to the model's lower accuracy in predicting outcomes Y when facing more abstract features, along with the imbalance and distribution of image features found in the datasets, issues that could be addressed by CATE estimators designed for this purpose. In practical applications, where a single model is trained on data with unknown predictive effects, a quantitative evaluation would entail performing regression and t-tests on the parameters, as described in Sec. 2.3, to assess the model's ability to identify information relevant for treatment effects (*i.e*. predictive biomarkers).

Our qualitative experimental results empirically demonstrate how an image-based CATE estimator's ability to identify predictive biomarkers can be assessed by comparing whether the treatment effect attribution maps to the selected ground truth predictive imaging biomarkers features. This is effective for both localized features based on color and shape (CMNIST, CUB-200-2011, NSCLC-Radiomics), as well as first-order statistics (NSCLC-Radiomics) or patterns (ISIC 2018). In applications, our XAI analysis is essential for identifying and interpreting predictive and prognostic imaging biomarkers. Unlike tabular data, images lack discrete candidates for feature importance scores. Distinguishing between predictive and prognostic imaging biomarkers using attribution maps becomes challenging when located in the same image areas. The heatmap focuses on the same pixels (as with energy and flatness in the NSCLC-Radiomics example), making it difficult to discern whether an image feature that is both predictive and prognostic is present, or if two independent imaging biomarkers with distinct meanings are spatially overlapping. In such cases, other XAI methods like counterfactual explanations [21] could quantify the effect of different properties of the same feature. Despite potential ambiguities for more abstract biomarkers, our evaluation can offer valuable insights into the features used by the model for its predictions.

While we acknowledge the limitations of using only semi-synthetic data, due to the current unavailability of public RCT image datasets with verified predictive imaging biomarkers, we also emphasize its advantages. Semi-synthetic data enables us to demonstrate the performance of CATE estimation models in a reproducible way, as discussed in [12] and [11]. Our approach to predictive imaging biomarker discovery and evaluation does not rely on handcrafted features such as radiomics. Instead, we use radiomics features as biomarkers to simulate outcomes in our experiments, serving merely as a baseline for conducting performance comparisons.

5. Conclusion

In this paper, we introduce the task of identifying predictive imaging biomarkers and show how a candidate identified by a model can be evaluated through (1) a statistical evaluation comparing the predictive strength relative to prognostic interactions, and (2) attribution maps to support the interpretation of the identified candidate. We outline an approach using an image-based CATE estimator to solve this task, enabling the discovery of new predictive imaging biomarkers without relying on potentially biased handcrafted features or image feature extractors. This also facilitates the detection of even abstract concepts from high-dimensional data, as demonstrated by our experiments. Our proposed experiments and analysis for assessing a model's qualitative and quantitative performance offer valuable insights for developing image-based CATE estimation methods tailored to specific challenges, such as adapting various network architectures for vision tasks and using CATE estimators previously applied to only tabular data. Our evaluation provides a foundation for future research addressing different imaging modalities and problem settings. This may include addressing non-linear biomarker-outcome relations, *e.g*. survival or time-to-event data, and mitigating confounding effects in observational data. Overall, we believe that applying image-based CATE estimators to discover unknown predictive biomarkers from imaging data can significantly enhance image-based treatment decision-making for personalized medicine and applications beyond.

Acknowledgements. We acknowledge funding from the German Research Foundation (DFG) as part of the Priority Programme 2177 Radiomics: Next Generation of Biomedical Imaging (project identifier: 428223917) and Collaborative Research Center 1389 (UNITE Glioblastoma – project identifier: 404521405). PV is funded through an Else Kröner Clinician Scientist Endowed Professorship by the Else Kröner Fresenius Foundation (reference number: 2022 EKCS.17). This work was partly funded by Helmholtz Imaging (HI), a platform of the Helmholtz Incubator on Information and Data Science. We thank David Zimmerer for the feedback on the manuscript.

References

[1] Hugo JWL Aerts, Emmanuel Rios Velazquez, Ralph TH Leijenaar, Chintan Parmar, Patrick Grossmann, Sara Carvalho, Johan Bussink, René Monshouwer, Benjamin Haibe-Kains, Derek Rietveld, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature communications*, 5(1):4006, 2014. 5

[2] H. J. W. L. Aerts, L. Wee, E. Rios Velazquez, R. T. H. Leijenaar, C. Parmar, P. Grossmann, S. Carvalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld, F. Hoebers, M. M. Rietbergen, C. R. Leemans, A. Dekker, J. Quackenbush, R. J. Gillies, and P. Lambin. Data From NSCLC-Radiomics (version 4). https://doi.org/10.7937/K9/TCIA.2015.PF0M9REI, 2014. Data set. 4

[3] Ahmed M Alaa, Michael Weisz, and Mihaela Van Der Schaar. Deep counterfactual networks with propensity-dropout. *arXiv preprint arXiv:1706.05966*, 2017. 2

[4] Martín Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *CoRR*, abs/1907.02893, 2019. 4

[5] Asma Bahamyirou, Mireille E Schnitzer, Edward H Kennedy, Lucie Blais, and Yi Yang. Doubly robust adaptive lasso for effect modifier discovery. *The International Journal of Biostatistics*, 18(2):307–327, 2022. 2

[6] Karla V Ballman. Biomarker: predictive or prognostic? *Journal of clinical oncology: official journal of the American Society of Clinical Oncology*, 33(33):3968–3971, 2015. 2, 4

[7] Philippe Boileau, Nina Ting Qi, Mark J van der Laan, Sandrine Dudoit, and Ning Leng. A flexible approach for predictive biomarker discovery. *Biostatistics*, 24(4):1085–1105, 2023. 2

[8] Chandra Bortolotto, Andrea Lancia, Chiara Stelitano, Marianna Montesano, Elisa Merizzoli, Francesco Agustoni, Giulia Stella, Lorenzo Preda, and Andrea Riccardo Filippi. Radiomics features as predictive and prognostic biomarkers in nsclc. *Expert Review of Anticancer Therapy*, 21(3):257–266, 2021. 2, 5

[9] Ahmad Chaddad, Michael Jonathan Kucharczyk, Paul Daniel, Siham Sabri, Bertrand J Jean-Claude, Tamim Niazi, and Bassam Abdulkarim. Radiomics in glioblastoma: current status and challenges facing clinical implementation. *Frontiers in oncology*, 9:374, 2019. 2

[10] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M. Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, Harald Kittler, and Allan Halpern. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019. 4

[11] Jonathan Crabbé, Alicia Curth, Ioana Bica, and Mihaela van der Schaar. Benchmarking heterogeneous treatment effect models through the lens of interpretability. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 2, 8

[12] Alicia Curth, David Svensson, Jim Weatherall, and Mihaela van der Schaar. Really doing great at estimating CATE? a critical look at ML benchmarking practices in treatment effect estimation. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. 1, 2, 3, 8

[13] Alicia Curth and Mihaela van der Schaar. Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms. In *International Conference on Artificial Intelligence and Statistics*, pages 1810–1818. PMLR, 2021. 2

[14] Alicia Curth and Mihaela van der Schaar. On inductive biases for heterogeneous treatment effect estimation. *Advances in Neural Information Processing Systems*, 34:15883–15894, 2021. 3

[15] Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012. 4

[16] Joshua Durso-Finley, Jean-Pierre Falet, Raghav Mehta, Douglas L. Arnold, Nick Pawlowski, and Tal Arbel. Improving Image-Based Precision Medicine with Uncertainty-Aware Causal Models. In Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood, and Russell Taylor, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 472–481, Cham, 2023. Springer Nature Switzerland. 2

[17] Joshua Durso-Finley, Jean-Pierre Falet, Brennan Nichyporuk, Arnold Douglas, and Tal Arbel. Personalized prediction of future lesion activity and treatment effect in multiple sclerosis from baseline mri. In *International Conference on Medical Imaging with Deep Learning*, pages 387–406. PMLR, 2022. 2, 3

[18] Gabriel Erion, Joseph D Janizek, Pascal Sturmfels, Scott M Lundberg, and Su-In Lee. Improving performance of deep learning models with axiomatic attribution priors and expected gradients. *Nature machine intelligence*, 3(7):620–631, 2021. 2, 4, 6

[19] Daniel S Gareau, James Browning, Joel Correa Da Rosa, Mayte Suarez-Farinas, Samantha Lish, Amanda M Zong, Benjamin Firester, Charles Vrattos, Yael Renert-Yuval, Mauricio Gamboa, et al. Deep learning-level melanoma detection by interpretable machine learning and imaging biomarker cues. *Journal of Biomedical Optics*, 25(11):112906–112906, 2020. 5

[20] Daniel S Gareau, Joel Correa da Rosa, Sarah Yagerman, John A Carucci, Nicholas Gulati, Ferran Hueto, Jennifer L DeFazio, Mayte Suárez-Fariñas, Ashfaq Marghoob, and James G Krueger. Digital imaging biomarkers feed machine learning for melanoma screening. *Experimental dermatology*, 26(7):615–618, 2017. 5

[21] Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual visual explanations. In *International Conference on Machine Learning*, pages 2376–2384. PMLR, 2019. 8

[22] Negar Hassanpour and Russell Greiner. Learning disentangled representations for counterfactual regression. In *International Conference on Learning Representations*, 2019. 2

[23] Erik Hermansson and David Svensson. On discovering treatment-effect modifiers using virtual twins and causal for-

est ml in the presence of prognostic biomarkers. In *International Conference on Computational Science and Its Applications*, pages 624–640. Springer, 2021. 2, 3

[24] Paul W Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960, 1986. 2

[25] Ahmed Hosny, Hugo J Aerts, and Raymond H Mak. Handcrafted versus deep learning radiomics for prediction of cancer therapy response. *The Lancet Digital Health*, 1(3):e106–e107, 2019. 2

[26] Connor T Jerzak, Fredrik Johansson, and Adel Daoud. Image-based treatment effect heterogeneity. *arXiv preprint arXiv:2206.06417*, 2022. 2

[27] Xiaotong Jiang, Xin Zhou, and Michael R Kosorok. Deep doubly robust outcome weighted learning. *Machine Learning*, 113(2):815–842, 2024. 2

[28] Ziyang Jiang, Zhuoran Hou, Yiling Liu, Yiman Ren, Keyu Li, and David Carlson. Estimating causal effects using a multi-task deep ensemble. *arXiv preprint arXiv:2301.11351*, 2023. 2

[29] Philipp Kickingereder, Michael Götz, John Muschelli, Antje Wick, Ulf Neuberger, Russell T Shinohara, Martin Sill, Martha Nowosielski, Heinz-Peter Schlemmer, Alexander Radbruch, et al. Large-scale radiomic profiling of recurrent glioblastoma identifies an imaging predictor for stratifying anti-angiogenic treatment responseradiomic profiling of bev efficacy in glioblastoma. *Clinical Cancer Research*, 22(23):5765–5771, 2016. 2

[30] Julia Krzykalla, Axel Benner, and Annette Kopp-Schneider. Exploratory identification of predictive biomarkers in randomized trials with normal endpoints. *Statistics in Medicine*, 39(7):923–939, 2020. 4

[31] Virendra Kumar, Yuhua Gu, Satrajit Basu, Anders Berglund, Steven A Eschrich, Matthew B Schabath, Kenneth Forster, Hugo JWL Aerts, Andre Dekker, David Fenstermacher, et al. Radiomics: the process and the challenges. *Magnetic resonance imaging*, 30(9):1234–1248, 2012. 2

[32] Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165, 2019. 3

[33] Philippe Lambin, Ralph TH Leijenaar, Timo M Deist, Jurgen Peerlings, Evelyn EC De Jong, Janita Van Timmeren, Sebastian Sanduleanu, Ruben THM Larue, Aniek JG Even, Arthur Jochems, et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nature reviews Clinical oncology*, 14(12):749–762, 2017. 2

[34] EJ Limkin, Roger Sun, Laurent Dercle, EI Zacharaki, Charlotte Robert, Sylvain Reuzé, Antoine Schernberg, Nikos Paragios, Eric Deutsch, and Charles Ferté. Promises and challenges for the implementation of computational medical imaging (radiomics) in oncology. *Annals of Oncology*, 28(6):1191–1206, 2017. 2

[35] Kathleen N Lohr. Outcome measurement: concepts and questions. *Inquiry*, pages 37–50, 1988. 1

[36] Bin Lou, Semihcan Doken, Tingliang Zhuang, Danielle Wingerter, Mishka Gidwani, Nilesh Mistry, Lance Ladic, Ali Kamen, and Mohamed E Abazeed. An image-based deep learning framework for individualising radiotherapy dose: a retrospective analysis of outcome prediction. *The Lancet Digital Health*, 1(3):e136–e147, 2019. 2

[37] Wenao Ma, Cheng Chen, Jill Abrigo, Calvin Hoi-Kwan Mak, Yuqi Gong, Nga Yan Chan, Chu Han, Zaiyi Liu, and Qi Dou. Treatment outcome prediction for intracerebral hemorrhage via generative prognostic model with imaging and tabular data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 715–725. Springer, 2023. 2

[38] James P. B. O'Connor, Eric O. Aboagye, Judith E. Adams, Hugo J. W. L. Aerts, Sally F. Barrington, Ambros J. Beer, Ronald Boellaard, Sarah E. Bohndiek, Michael Brady, Gina Brown, David L. Buckley, Thomas L. Chenevert, Laurence P. Clarke, Sandra Collette, Gary J. Cook, Nandita M. deSouza, John C. Dickson, Caroline Dive, Jeffrey L. Evelhoch, Corinne Faivre-Finn, Ferdia A. Gallagher, Fiona J. Gilbert, Robert J. Gillies, Vicky Goh, John R. Griffiths, Ashley M. Groves, Steve Halligan, Adrian L. Harris, David J. Hawkes, Otto S. Hoekstra, Erich P. Huang, Brian F. Hutton, Edward F. Jackson, Gordon C. Jayson, Andrew Jones, Dow-Mu Koh, Denis Lacombe, Philippe Lambin, Nathalie Lassau, Martin O. Leach, Ting-Yim Lee, Edward L. Leen, Jason S. Lewis, Yan Liu, Mark F. Lythgoe, Prakash Manoharan, Ross J. Maxwell, Kenneth A. Miles, Bruno Morgan, Steve Morris, Tony Ng, Anwar R. Padhani, Geoff J. M. Parker, Mike Partridge, Arvind P. Pathak, Andrew C. Peet, Shonit Punwani, Andrew R. Reynolds, Simon P. Robinson, Lalitha K. Shankar, Ricky A. Sharma, Dmitry Soloviev, Sigrid Stroobants, Daniel C. Sullivan, Stuart A. Taylor, Paul S. Tofts, Gillian M. Tozer, Marcel van Herk, Simon Walker-Samuel, James Wason, Kaye J. Williams, Paul Workman, Thomas E. Yankeelov, Kevin M. Brindle, Lisa M. McShane, Alan Jackson, and John C. Waterton. Imaging biomarker roadmap for cancer studies. *Nature Reviews Clinical Oncology*, 14(3):169–186, Mar. 2017. 2

[39] Ji Eun Park and Ho Sung Kim. Radiomics as a quantitative imaging biomarker: practical considerations and the current standpoint in neuro-oncologic studies. *Nuclear medicine and molecular imaging*, 52(2):99–108, 2018. 2

[40] Chintan Parmar, Patrick Grossmann, Johan Bussink, Philippe Lambin, and Hugo JWL Aerts. Machine learning methods for quantitative radiomic biomarkers. *Scientific reports*, 5(1):1–11, 2015. 2

[41] Mei-Yin C Polley, Boris Freidlin, Edward L Korn, Barbara A Conley, Jeffrey S Abrams, and Lisa M McShane. Statistical and practical considerations for clinical evaluation of predictive biomarkers. *Journal of the National Cancer Institute*, 105(22):1677–1683, 2013. 4

[42] Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005. 3

[43] Konstantinos Sechidis, Konstantinos Papangelou, Paul D Metcalfe, David Svensson, James Weatherall, and Gavin Brown. Distinguishing prognostic and predictive biomarkers: an information theoretic approach. *Bioinformatics*, 34(19):3365–3376, 2018. 2, 3

[44] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017. 2, 4, 6

[45] Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pages 3076–3085. PMLR, 2017. 2, 3

[46] Claudia Shi, David Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. *Advances in neural information processing systems*, 32, 2019. 2

[47] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Workshop at International Conference on Learning Representations*, 2014. 2, 4

[48] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. In *ICLR (workshop track)*, 2015. 2, 4, 6

[49] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic Attribution for Deep Networks. *arXiv:1703.01365 [cs]*, June 2017. arXiv: 1703.01365. 2, 6

[50] Koh Takeuchi, Ryo Nishida, Hisashi Kashima, and Masaki Onishi. Grab the reins of crowds: Estimating the effects of crowd movement guidance using causal inference. *arXiv preprint arXiv:2102.03980*, 2021. 2

[51] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018. 4

[52] Joost JM Van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina GH Beets-Tan, Jean-Christophe Fillion-Robin, Steve Pieper, and Hugo JWL Aerts. Computational radiomics system to decode the radiographic phenotype. *Cancer research*, 77(21):e104–e107, 2017. 5

[53] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 4

[54] Weijia Zhang, Lin Liu, and Jiuyong Li. Treatment effect estimation with disentangled latent factors. In *AAAI Conference on Artificial Intelligence*, 2020. 2

[55] Wencan Zhu, Céline Lévy-Leduc, and Nils Ternès. Identification of prognostic and predictive biomarkers in high-dimensional data with pplasso. *BMC bioinformatics*, 24(1):25, 2023. 2

[56] Alex Zwanenburg, Martin Vallières, Mahmoud A Abdalah, Hugo JWL Aerts, Vincent Andrearczyk, Aditya Apte, Saeed Ashrafinia, Spyridon Bakas, Roelof J Beukinga, Ronald Boellaard, et al. The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology*, 295(2):328–338, 2020. 5