Supplementary Material: 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

A. Supplementary Information

A.1. Explainable AI Methods

For the attribution maps in Fig. 5 of the main paper, we use Expected Gradients (EG) [4] for the CMNIST dataset and guided Gradient-weighted Class Activation Mapping (GGCAM) [10, 12] for the other two datasets. Due to its up-sampling mechanism, GGCAM is only suitable to a limited extent to compute the attribution of each RGB channel individually. For the results presented in the supplementary section A.2, we additionally apply Integrated Gradients (IG) [13] with a black (i.e. zero) baseline value and the traditional Gradient-weighted Class Activation Mapping (GCAM).

Integrated Gradients. IG calculates a path integral from a baseline value x_0 to the actual value x_j for each of the j input features, in our case pixels or voxels.

$$\text{IG}_j(x, x_0) = (x_j - x_{0j}) \int_{\alpha=0}^{1} \frac{\partial f(x_0 + \alpha(x - x_0))}{\partial x_j} d\alpha \tag{1}$$

However, selecting a baseline value x_0 for IG is often ambiguous, and executing multiple path integrals across different baseline values can be inefficient.

Expected Gradients. To avoid selecting a baseline value as would be necessary for IG, Erion et al. [4] presented a solution based on a probabilistic baseline computed over a sample of observations:

$$EG_j(x) = \int_{x_0} IG_j(x, x_0)\, p_D(x_0)\, dx_0 \tag{2}$$

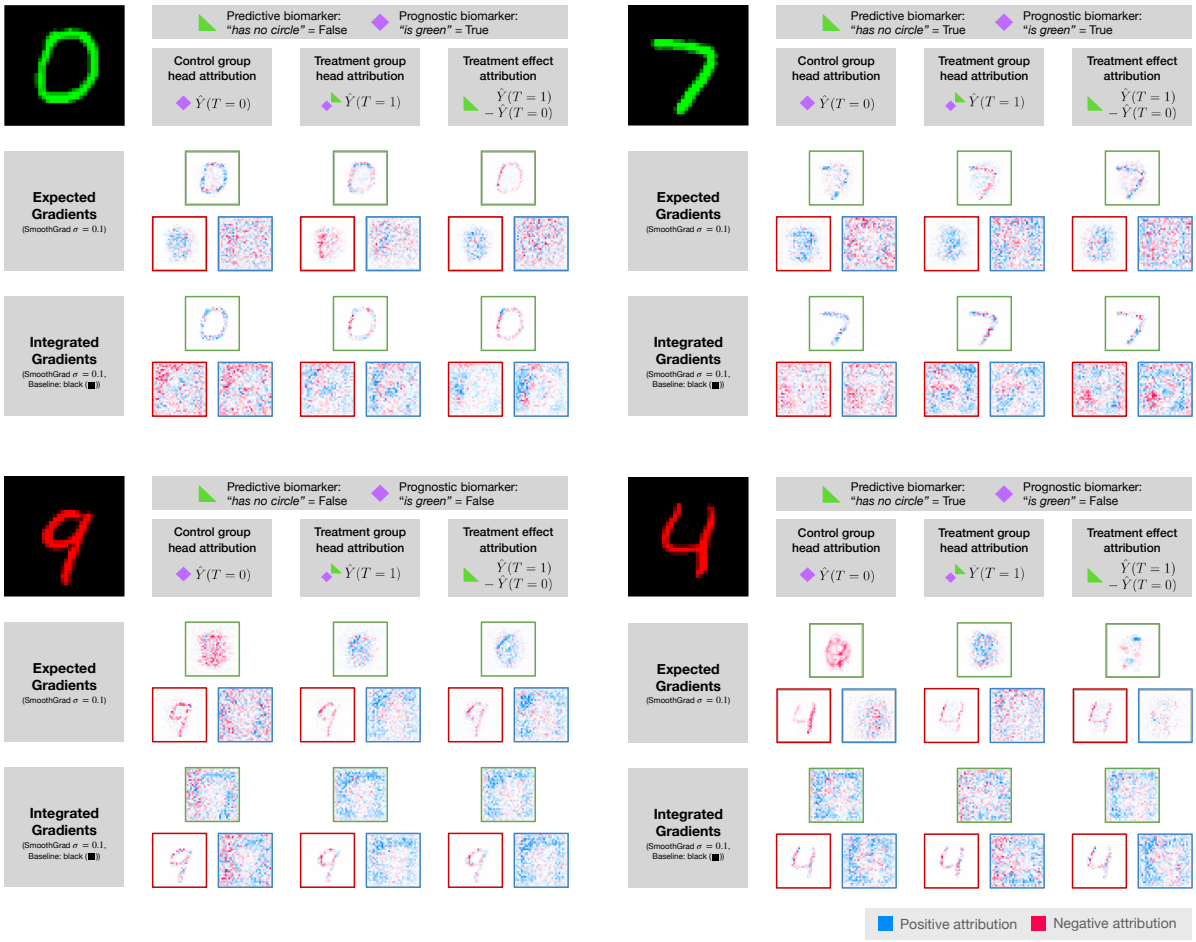
$$= \int_{x_0} \left((x_j - x_{0j}) \int_{\alpha=0}^{1} \frac{\partial f(x_0 + \alpha(x - x_0))}{\partial x_j} d\alpha \right) p_D(x_0)\, dx_0 \tag{3}$$

$$= \mathop{\mathbb{E}}_{x_0 \sim D,\ \alpha \sim U(0,1)} \left[\frac{\partial f(x_0 + \alpha(x - x_0))}{\partial x_j}\, d\alpha \right], \tag{4}$$

*These authors contributed equally to this work.

with x_0 as the baseline, x_j the input feature number j and D as the underlying data distribution. In practice, EG is computed via a mini-batch procedure by drawing samples for x_0 and α, computing the expression inside the expectation, and averaging over the mini-batch.

(Guided)-GradCAM. GGCAM is the combination of Guided Backpropagation (GB) [12] and GradCAM [10]. GradCAM computes the attribution via backpropagation into a selected hidden layer, usually the last convolutional layer, and up-samples it to the input size. GradCAM leverages the idea that convolutional neural networks transform spatial to semantic information by attributing to the semantic information, which is then up-sampled back into the input space. GB, on the other hand, backpropagates directly from a target output into the original image but via a guiding function, overriding non-negative gradients from Rectified Linear Unit (ReLU) activation functions. GGCAM takes the element-wise product between GB and the non-negative GradCAM attributions, leveraging both the semantic information from GradCAM and the more fine-grained spatial information in the input space from GB.

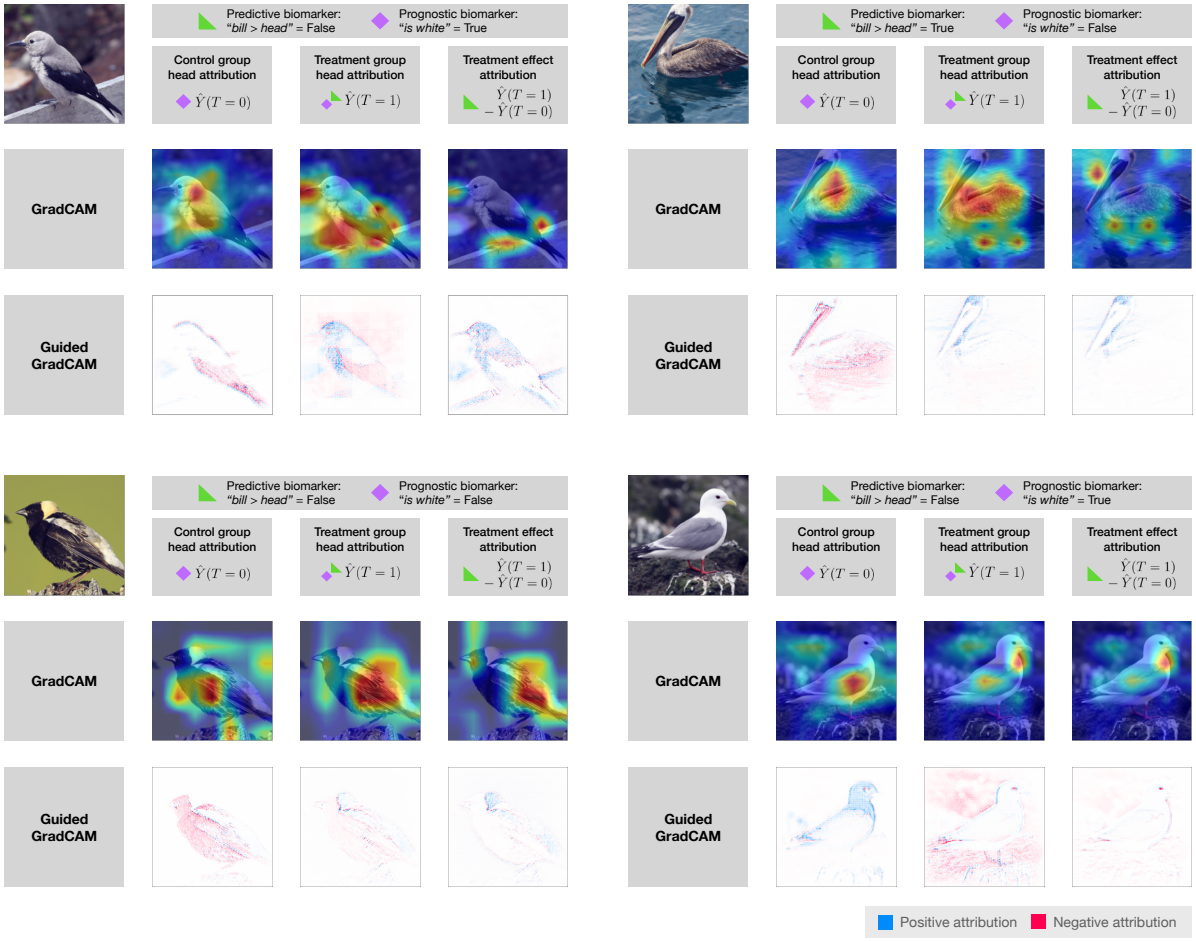


Supplementary Figure 1. EG and IG attribution maps of our model for the control head, treatment group head, and CATE output for the four CMNIST dataset samples, showcasing the presence or absence of the predictive and prognostic biomarker. We present the attribution map for each RGB color channel (left: red, top: green, right: blue).

A.2. Additional feature attributions results for interpreting predictive imaging biomarkers

Further attribution maps for samples from the CMNIST, CUB-200-2011, ISIC 2018, and NSCLC-Radiomics dataset and the control head $\hat{Y}(T=0)$, CATE $\hat{Y}(T=1) - \hat{Y}(T=0)$ and additionally the treatment group head output $\hat{Y}(T=1)$ are presented in Supplementary Fig. 1, 2, 3 and 5, respectively. Each figure showcases the results of four examples with varying presence or values of predictive and prognostic biomarkers.

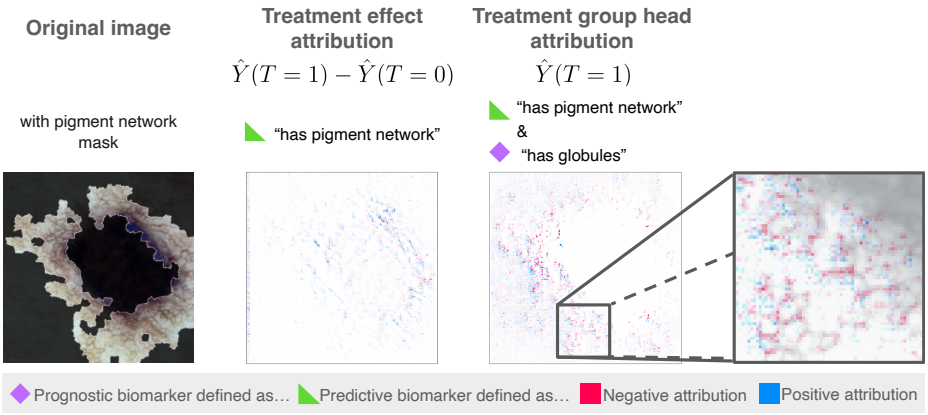
CMNIST. We extend our analysis of the main paper and show two additional CMNIST images and IG attribution maps in Supplementary Fig. 1. The blue color channel only shows noisy attribution for both EG and IG attribution maps, suggesting minimal impact on the model's predictions. While attribution maps for the digit color channel show consistent patterns across EG and IG methods, they differ for the other remaining color channel. For the control group head output $\hat{Y}(T=0)$, a positive attribution is observed from the green color channel for both the digit 0 and 7 in the top row, whereas a negative attribution is observed in both the red and also green color channel of the digit 9 and 4 in the bottom row. This indicates that the model correctly identifies the prognostic biomarker "digit is green" from the relevant color channel. The attribution maps of the predicted CATE output $\hat{Y}(T=1) - \hat{Y}(T=0)$ exhibited notable negative attribution in the color channel of the digit color for both the digit 0 and 9 in the left column, indicating the absence of the predictive biomarker "has no circle". An overall positive attribution was observed in the green color channel of digit 7, suggesting the presence of the predictive biomarker. However, a more negative attribution with some noisy positive attribution is seen for the red channel of the digit four, indicating some ambiguity in identifying the presence of the predictive biomarker.



Supplementary Figure 2. GCAM and GGradCAM attribution maps of our model for the control head, treatment group head, and CATE output for four CUB-200-2011 dataset samples, showcasing the presence or absence of the predictive and prognostic biomarker.

CUB-200-2011. The extended results for the CUB-200-2011 dataset are shown in Supplementary Fig. 2. The GCAM attribution maps of the control group head output $\hat{Y}(T=0)$ indicate that the model focuses the most on the main body of the bird. The GGCAM attribution maps reveal overall negative attributions from the wing and tail (top left and top right) or from the belly region (bottom left) and an overall positive attribution from the tail and head/neck region (bottom right). While this suggests that the model incorrectly identifies the prognostic biomarker "has primary color: white" of the top left bird as black due to the black wings, it correctly identifies it in the other three cases. For the predicted CATE output

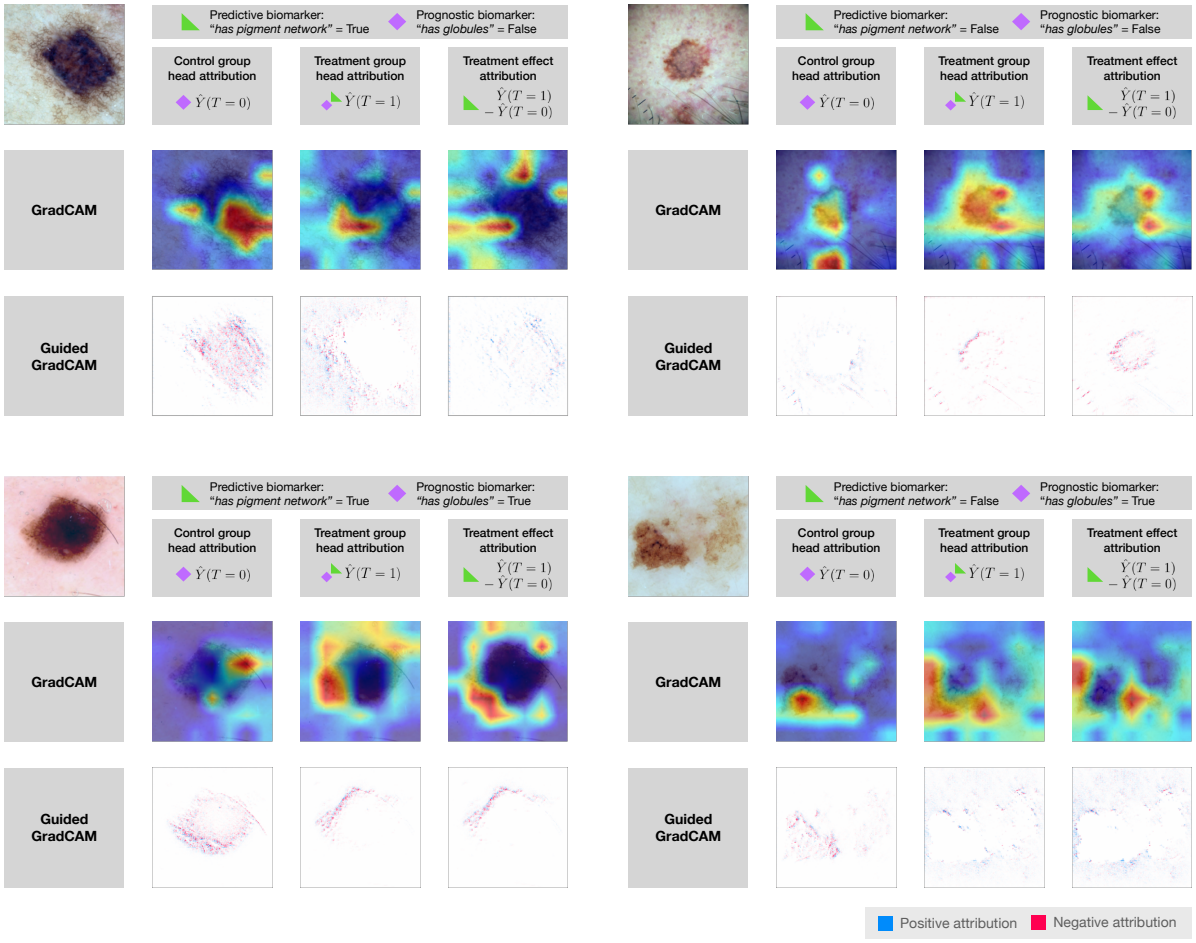
$\hat{Y}(T=1)-\hat{Y}(T=0)$, the GCAM attribution maps highlight the beak and neck areas, which corresponds to the areas where the predictive biomarker "has bill length: longer than head", but also parts of the bottom areas of the birds. The GGCAM attribution maps show overall positive attributions in all examples except the bottom right, indicating that it is more difficult for the model to correctly distinguish the relative bill lengths, except for the top and bottom right examples.



Supplementary Figure 3. Attribution maps for an example image containing a pigment network (shown as a mask) but no globules for an image of the ISIC 2018 dataset.

ISIC 2018. As mentioned in the Sec. 3.2 and shown in Fig. 5 of the main paper, the model likely uses the lighter gaps or "holes" between the dark vein-like grid structure to detect pigment networks. We expect the treatment group head $\hat{Y}(T=1)$ to be sensitive to both predictive and prognostic biomarkers. In Supplementary Fig. 3 we observe in the treatment group head's attribution map that there is positive attribution again (shown in blue) in the light interspaces but also negative attribution (in red) in the dark veins. This observation also supports the hypothesis that the model associates the dark veins with the globules but detects the pigment network correctly through the light interspaces as seen for the treatment effect attribution map, as the lighter interspaces are uniquely present in pigment networks. The Supplementary Fig. 4 shows the results for three additional examples from the ISIC 2018 dataset. The GCAM attribution maps show a lack of clear localization, except for the bottom left example, indicating that the network does not identify a clear localization of the biomarkers. When comparing the GGCAM attribution maps, even though the attribution maps show that the network identifies some structures, only the location of the predictive biomarker "has pigment network" is likely correctly identified by the network in the bottom left example when comparing to the ground truth segmentation.

NSCLC-Radiomics. The additional attribution map results for the NSCLC-Radiomics dataset are depicted in Supplementary Fig. 5. Both the GCAM and GGCAM attribution maps for the control group output $\hat{Y}(T=0)$ show the most attribution from areas surrounding the tumors. This is especially evident in areas where the tumor shape is not spherical or round, such as in the bottom right color of the upper left tumor example or the thinner section in the middle of the lower left tumor example. GCAM attribution maps for the CATE output $\hat{Y}(T=1)-\hat{Y}(T=0)$ show that the model tends to focus on the darker parts of the image slices. The corresponding GGCAM attribution maps reveal more negative attributions from the darker parts of the images, but strong positive attributions from the neighboring lighter parts in all four examples. This observation aligns with the fact that the minimum pixel intensity value contributes strongly to the predictive biomarker feature "energy". To provide deeper insights into how the model identifies 3D features, we also show the 3D attribution maps for one example in Supplementary Fig. 6. Here, the 3D attribution map for the control group output $\hat{Y}(T=0)$ highlights an area above the tumor, likely erroneously. The CATE output $\hat{Y}(T=1)-\hat{Y}(T=0)$ shows a high attribution on the upper right side of the tumor where it appears flat. This area likely corresponds to the region where the principal components lie and which contributes to the predictive biomarker "flatness".



Supplementary Figure 4. GCAM and GGradCAM attribution maps of our model for the control head, treatment group head, and CATE output for four ISIC 2018 dataset samples, showcasing the presence or absence of the predictive and prognostic biomarker.

A.3. CATE estimation performance and biomarker identification performance

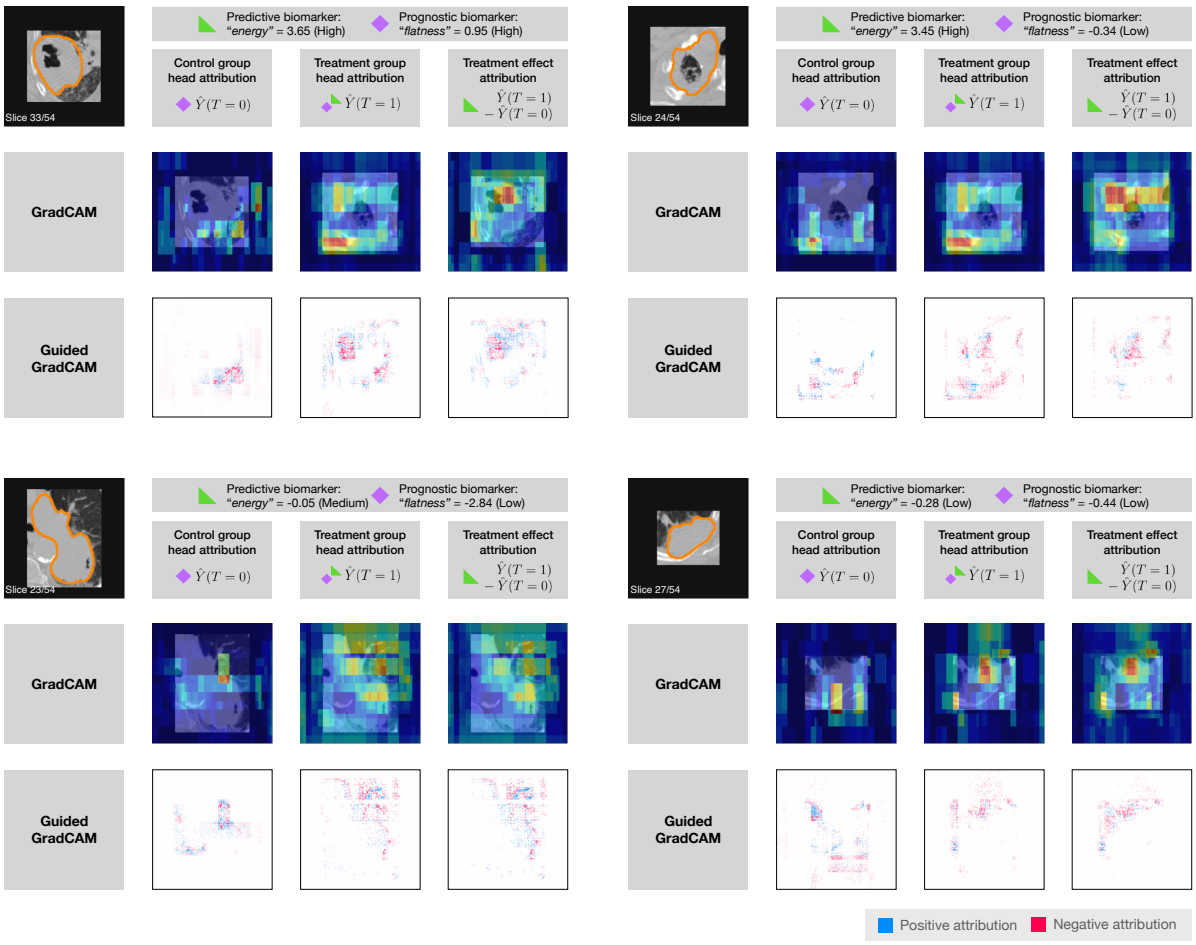
In CATE estimation, a model's performance is usually evaluated using the Precision of Estimating Heterogeneous Effects (PEHE) [6] metric, which is defined as

$$\text{PEHE} = \sqrt{\frac{1}{n}\sum_i (\tau_i - \hat{\tau}_i)^2}, \tag{5}$$

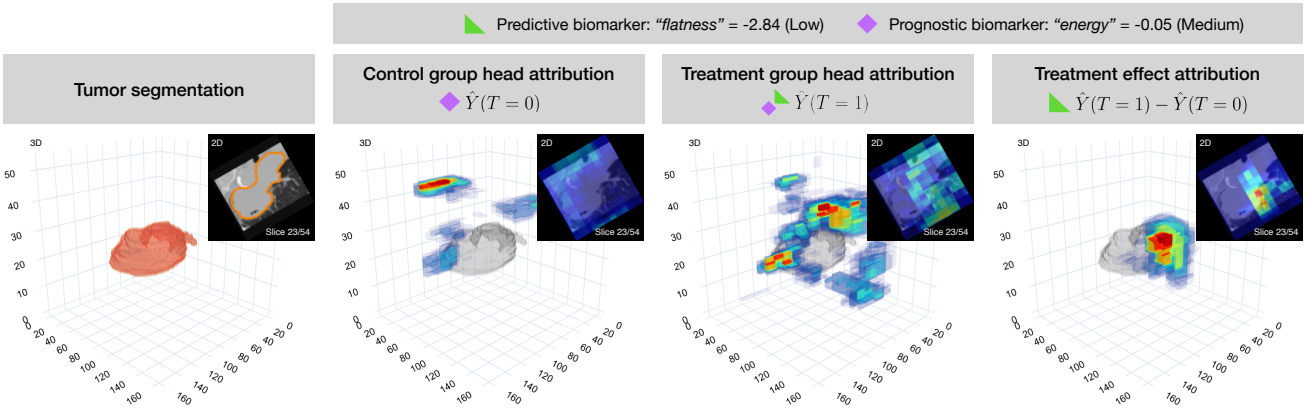
where n denotes the number of test samples, τ_i the true CATE and $\hat{\tau}_i$ the estimated CATE for a test sample i. In Supplementary Fig. 7 and Supplementary Table 1, where the PEHE metric is reported alongside the root mean square error (RMSE) of the prediction of factual outcomes, we observe a better CMNIST compared to the other three datasets. This observation also corresponds to the performance for the relative predictive strength as observed in Fig. 4 of the main paper.

There are variations within the same dataset between models trained with biomarkers on feature sets (a) and (b), which is likely since it is slightly easier for the models to identify one type of prognostic and predictive biomarker combinations than the others. A lower RMSE but a high PEHE indicates that the model can only predict the factual outcomes well but not the counterfactual outcomes, which slight effects of overfitting could cause. Due to the different sampling space of the simulation parameters b_{pred} and b_{prog}, only a limited comparison can be made for CMNIST with the other two datasets, however. As the scale of the CATE automatically changes with parameters b_{pred} and b_{prog} [2], also the PEHE changes, which therefore also depends on the absolute value of b_{pred} and b_{prog}. This phenomenon further the comparability across different ratios b_{pred}/b_{prog}.

The RMSE and PEHE results for NSCLC-Radiomics are worse and have a slightly larger variance than for the CUB-200-



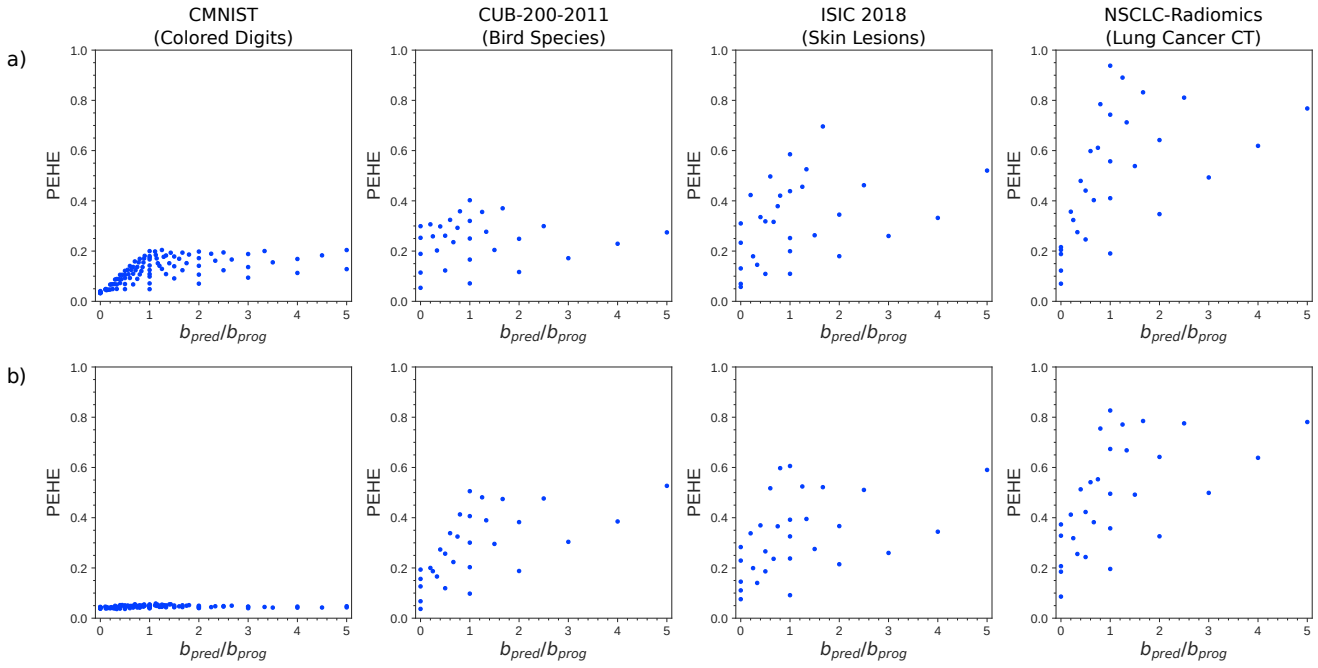
Supplementary Figure 5. GCAM and GGCAM attribution maps of our model for the control head, treatment group head, and CATE output for one sagittal slice of each of the four NSCLC-Radiomics dataset samples, showcasing the predictive and prognostic biomarker with varying strengths. The tumor segmentation outlines are shown in orange.



Supplementary Figure 6. 3D GCAM attribution maps of our model for the control head, treatment group head, and CATE output, illustrated for a 3D patch from the NSCLC-Radiomics dataset. Additionally, a 3D render of the segmented tumor and a 2D sagittal slice are shown.

2011 and ISIC 2018 datasets, which also explains the models' worse performance in identifying predictive biomarkers as shown in Fig. 4 of the main paper.

However, it is generally not possible to directly conclude a model's performance concerning identifying the correct imaging biomarkers just from the PEHE metrics alone [2, 3]. In our case, the exact value of the CATE is not directly important for the evaluation, as our main task of interest is identifying predictive imaging biomarkers. Therefore, the PEHE is only suitable as a secondary evaluation metric alongside the evaluation methods mentioned in the Sec. 2.3 of the main paper.



Supplementary Figure 7. Performance of our treatment effect estimation models trained with biomarkers from feature set (**a**) or **(b)** with respect to the precision of estimating heterogeneous effects (PEHE) for different simulation parameters b_{pred}/b_{prog} (i.e. relative size of the predictive effect in the simulated outcomes). The lower the PEHE the better the performance.

Dataset	Feature Set	PEHE	RMSE
CMNIST	(a)	0.121	0.094
	(b)	0.045	0.115
CUB-200-2011	(a)	0.227	0.304
	(b)	0.277	0.261
ISIC 2018	(a)	0.304	0.352
	(b)	0.308	0.362
NSCLC-Radiomics	(a)	0.475	0.561
	(b)	0.469	0.633

Supplementary Table 1. Performance with respect to the mean PEHE and RMSE for the prediction of factual outcomes per dataset for our treatment effect estimation models trained with biomarkers from feature set (a) or (b).

A.4. Implementation details

In our experiments, the two-headed CATE estimation models are all based on the ResNet [5] architecture tailored to each dataset. For the CMNIST experiments, we utilize a MiniResNet (ResNet-14) with 14 layers, 0.20 M parameters, and only three building blocks. In the CUB-200-2011 and ISIC 2018 experiments, we employ a two-headed ResNet-18 with 11.18 M parameters, and for the NSCLC-Radiomics a two-headed 3D ResNet with 33.30 M parameters. In all architectures, the

treatment-specific heads consist of either the last fully connected layer or the last four fully connected layers for NSCLC-Radiomics experiments. Its preceding convolutional layers learn shared presentations of control and treatment group data. We use the classic (one-headed) version of the corresponding ResNet architectures as our baseline models. The models for CMNIST are trained for 400 epochs with a mini-batch size of 1000. For CUB-200-2011 and ISIC 2018, the models are trained with a mini-batch size of 64 and for 1000 or 2000 epochs respectively. The NSCLC-Radiomics models are trained with a batch size of 8 and 2000 epochs. For data all datasets, we use the mean squared error loss function, a learning rate of $\mathrm{lr} = 0.001$, and the SGD optimizer.

For preprocessing, we apply zero padding of size 2 to each edge of the CMNIST images. The CUB-200-2011 images are resized so their smaller edge has the size 256. We augment the data by performing random crop and horizontal flips so that all final images have the size 224×224. We resize the ISIC 2018 images to 224 for the shorter edge, crop them to between 40% and 100% of their previous size, and resize them again to size 224×224. We augment them with random horizontal and vertical flips, randomly applied rotations by 90 degrees and color jitters. During the inference of both CUB-200-2011 and ISIC 2018 images, center crops are used. All 2D images are normalized by subtracting the mean and dividing by the standard deviation of the respective channel from the training dataset. For the NSCLC-Radiomics dataset, we added padding of value -1024 (HU) so that all 3D patches are of the size $162 \times 162 \times 54$. All radiomics features are normalized by subtracting the mean and dividing by the standard deviation of each feature. 3D image augmentations are implemented using the MONAI deep-learning framework [1] and include random flipping, random rotation by 90 degrees along the xy-axis, and random zooming with probability 0.5 by a factor in the range $[0.9, 1.1]$. Resampling to the median spacing of the dataset [0.9765625, 0.9765625, 3.0] mm is based on Isensee et. al [7] and uses a third-order spline in-plane and nearest-neighbor interpolation out-of-plane.

For the statistical evaluations, linear regression using ordinary least squares and t-tests for the fit coefficients as described in Sec. 2.3 of the main paper are performed using the statsmodels python module [9]. To create attribution maps, we use expected gradients (EG) [4] for CMNIST and guided gradient-weighted class activation mapping (GGCAM) [10, 12] for the other three datasets. Using EG allows us to determine the attribution of each color channel in contrast to CAM methods, which is vital for discovering the color-related CMNIST biomarkers. Both methods are implemented using Captum [8] and enhanced by SmoothGrad [11] to make the attribution maps less noisy and more robust.

References

[1] M Jorge Cardoso, Wenqi Li, Richard Brown, Nic Ma, Eric Kerfoot, Yiheng Wang, Benjamin Murrey, Andriy Myronenko, Can Zhao, Dong Yang, et al. Monai: An open-source framework for deep learning in healthcare. *arXiv preprint arXiv:2211.02701*, 2022. 8

[2] 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. 5, 7

[3] 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. 7

[4] 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. 1, 8

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 7

[6] Jennifer L Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011. 5

[7] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021. 8

[8] Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. Captum: A unified and generic model interpretability library for pytorch, 2020. 8

[9] Skipper Seabold and Josef Perktold. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010. 8

[10] 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. 1, 2, 8

[11] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda B. Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *CoRR*, abs/1706.03825, 2017. 8

[12] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. In *ICLR (workshop track)*, 2015. 1, 2, 8

[13] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic Attribution for Deep Networks. *arXiv:1703.01365 [cs]*, June 2017. arXiv: 1703.01365. 1