

# Data Augmentation for Information Transfer

## Why controlling for confounding effects in radiomic studies is important and how to do it

Michael Götz<sup>1</sup>, Klaus Maier-Hein<sup>1</sup>

<sup>1</sup>Medical Image Computing Group, Deutsches Krebsforschungszentrum (DKFZ)  
m.goetz@dkfz-heidelberg.de

The major goal of radiomics studies is the identification of predictive and reliable markers. It is, therefore, crucial to account for unwanted confounding effects that affect the radiomic features like scanning noise, annotator bias, or the used imaging device and parameter. Usually, these confounding effects are not sufficiently represented in the main cohort of radiomics studies and consequently are investigated in smaller side-studies.

Within our study [1], we looked into two questions: a) are those side-studies necessary and b) how to use the information from those studies on the feature stability in the radiomics modelling process. For this, three different methods for incorporating prior knowledge into a radiomics modelling process were compared: the naïve approach (ignoring feature quality), the most common approach consisting of removing unstable features based on correlation ranking, and a novel approach using data augmentation for information transfer (DAFIT). The predictive power and the ability to estimate the predictive power were assessed by looking at the ROC Area under Curve (AUC) and the difference between the AUC from data with and data without confounding effects present. Synthetic and publicly available real lung imaging patient data were used for the experiments.

The experiments showed the importance of controlling for confounding effects. Differences between the estimated and true performance of a model of up to 20 and 25 percentage points for real and synthetic data, respectively, showed the possible impact of ignoring confounding effects. Removing unstable features improved the performance estimation, while slightly decreasing the model performance, i.e. decreasing the area under curve achieved with the model. We argue that the reduction of features led to an effective reduction of information that is available to build the model. This point is addressed by the proposed approach, which performed superior both in terms of the estimation of the model performance and the actual model performance.

## References

1. Götz M, Maier-Hein KH. Optimal Statistical Incorporation of Independent Feature Stability Information into Radiomics Studies. *Sci Rep.* 2020 01;10(737):2045–2322.