Title: The effects of data leakage on neuroimaging predictive models

Authors: Matthew Rosenblatt[1], Link Tejavibulya[2], Rongtao Jiang[3], Stephanie Noble[3], Dustin Scheinost[1,2,3,4,5]

[1]Department of Biomedical Engineering, Yale University, New Haven, CT
[2]Interdepartmental Neuroscience Program, Yale University, New Haven, CT
[3]Department of Radiology & Biomedical Imaging, Yale School of Medicine, New Haven, CT
[4]Child Study Center, Yale School of Medicine, New Haven, CT
[5]Department of Statistics & Data Science, Yale University, New Haven, CT

Matthew Rosenblatt, Magnetic Resonance Research Center, 300 Cedar St, P.O. Box 208043, New Haven, CT, USA 06520-8043, USA. matthew.rosenblatt@yale.edu

Abstract

Predictive modeling has now become a central technique in neuroimaging to identify complex brain-behavior relationships and test their generalizability to unseen data. However, data leakage, which unintentionally breaches the separation between data used to train and test the model, undermines the validity of predictive models. Although previous literature suggests that leakage is generally pervasive in machine learning, few studies have empirically evaluated the effects of leakage in neuroimaging data. Here, using over 500 different pipelines spanning four large neuroimaging datasets and three phenotypes, we evaluated six forms of leakage fitting into three broad categories: feature selection, covariate correction, and lack of independence between subjects. As expected, leakage via feature selection and repeated subjects drastically inflated prediction performance. Notably, other forms of leakage had only minor effects (e.g., leaky site correction) or even decreased prediction performance (e.g., leaky covariate regression). In some cases, leakage affected not only prediction performance, but also model coefficients, and thus neurobiological interpretations. Overall, our results illustrate the variable effects of leakage on prediction pipelines and underscore the importance of avoiding data leakage to improve the validity and reproducibility of predictive modeling.

1. Introduction

Understanding individual differences in brain-behavior relationships is a central goal of neuroscience. As part of this goal, machine learning approaches using neuroimaging data, such as functional connectivity, have grown increasingly popular in predicting numerous phenotypes (*1*), including cognitive performance (*2–6*), age (*7–10*), and several clinically-relevant outcomes (*11–13*). Compared to classic statistical inference, prediction offers advantages in replicability and generalizability, as it evaluates models on participants unseen during model training (*14*, *15*). Essentially, the data are split into training and test subsets, such as through k-fold cross-validation or simple train/test splits, so that the model is strictly evaluated on unseen data. Unfortunately, the boundaries between training and test data can be inadvertently violated by

data leakage. Data leakage is when information about the test data is introduced into the model during training (*16*), nullifying the benefits of separating training and test data.

A recent meta-review of machine learning (*17*) highlighted the prevalence of leakage across 17 fields. Three hundred twenty-nine papers were identified as containing leakage. This meta-review described eight types of leakage: not having a separate test set, pre-processing on the training and test sets, feature selection jointly on the training and test sets, duplicate data points, illegitimate features, temporal leakage, non-independence between the training and test sets, and sampling bias. Data leakage often led to inflated model performance and consequently decreased reproducibility (*17*). In another review specific to predictive neuroimaging, ten of the 57 studies may have leaked information by performing dimensionality reduction across the whole dataset prior to train/test splitting (*18*). Since leakage may dramatically change the reported results, it contributes to the ongoing reproducibility crisis in neuroimaging (*17*, *19–21*). Despite the prevalence and concern of leakage, the severity of performance inflation due to leakage in neuroimaging predictive models remains unknown.

In this work, we evaluate the effects of leakage on functional connectome-based predictive models in four large datasets for the prediction of three phenotypes. Specifically, in over 500 pipelines, we test feature leakage, covariate-based leakage, and subject leakage. These leakage types span five of the leakage types described by Kapoor and Narayanan (*17*) (Table S1). We first show the effects of leakage on prediction performance by comparing two performance metrics in various leaky and non-leaky pipelines. Then, we evaluate the effects of leakage on model interpretation by comparing model coefficients. Furthermore, we resample the datasets at four different sample sizes to illustrate how small sample sizes may be most susceptible to leakage. Finally, we extend our analysis to structural connectomes in one open-source dataset. Overall, our results elucidate the consequences, or in some cases lack thereof, of numerous possible forms of leakage in neuroimaging datasets.

2. Methods

2.1 Datasets

Resting-state fMRI data were obtained in each of our four datasets: the Adolescent Brain Cognitive Development (ABCD) Study (*22*), the Healthy Brain Network (HBN) Dataset (*23*), the Human Connectome Project Development (HCPD) Dataset (*24*), and the Philadelphia Neurodevelopmental Cohort (PNC) Dataset (*25*, *26*). Throughout this work, we predicted age, attention, and matrix reasoning in these four datasets. Specific measures for each dataset are described below, but these three broad phenotypes were selected because they are available in all the datasets in this study. In addition, these three phenotypes are appropriate for benchmarking leakage because they span a wide range of effect sizes, with strong prediction performance for age, moderate performance for matrix reasoning, and poor performance for attention.

2.2 Preprocessing

In all cases, data were motion corrected. Additional preprocessing steps were performed using BioImage Suite (*27*). This included regression of covariates of no interest from the functional data, including linear and quadratic drifts, mean cerebrospinal fluid signal, mean white matter signal, and mean global signal. Additional motion control was applied by regressing a 24-parameter motion model, which included six rigid body motion parameters, six temporal derivatives, and the square of these terms, from the data. Subsequently, we applied temporal smoothing with a Gaussian filter (approximate cutoff frequency=0.12 Hz) and gray matter masking, as defined in common space. Then, the Shen 268-node atlas (*28*) was applied to parcellate the denoised data into 268 nodes. Finally, we generated functional connectivity matrices by correlating each pair of node time series data and applying the Fisher transform.

Data were excluded for poor data quality, missing nodes due to lack of full brain coverage, high motion (>0.2mm mean frame-wise motion), or missing behavioral/phenotypic data, which is detailed further for each specific dataset below.

2.3 Adolescent Brain Cognitive Development (ABCD) Data

In this work, we used the first release of the ABCD dataset. This consists of 9-10 year olds imaged across 21 sites in the United States. 4150 resting-state connectomes (up to 10 minutes of resting-state data) remained in this dataset after excluding scans for poor quality or high motion (>0.2mm mean frame-wise displacement [FD]). Among these participants, the mean age was 9.84 (s.d. 0.62) years, with a range of 9-10.92 years, and 48.96% self-reported their sex as female. All but one participant had family information, and among the 4150 participants with family information, there were 3720 participants from unique families.

For the attention measure, we used the Child Behavior Checklist (CBCL) (*29*) Attention Problems Raw Score. One participant was missing the attention score in ABCD. Of the participants with an attention score, the mean was 3.02 (s.d. 3.57), with a range of 0-20.

For the matrix reasoning measure, we used the Wechsler Intelligence Scale for Children (WISC-V) (*30*) Matrix Reasoning Total Raw Score. 71 participants were missing the WISC-V measurement (N=4059). The mean matrix reasoning score in ABCD was 17.78 (s.d. 3.88), with a range of 0-29.

2.4 Healthy Brain Network (HBN) Data

The HBN dataset consists of participants from approximately 5-22 years old. The data were collected from four sites near the New York greater metropolitan area. 1201 resting-state connectomes (10 minute scans) remained after applying the exclusion criteria. 39.80% were female, and the average age was 11.65 (s.d. 3.42) years, with a range of 5.58-21.90. Family information is not available in this dataset.

For the attention measure, we used the CBCL Attention Problems Raw Score (*29*). 51 participants were missing the attention score, but the mean score of the remaining participants was 7.41 (s.d. 4.54), with a range of 0-19.

For the matrix reasoning measure, we also used the WISC-V (*30*) Matrix Reasoning Total Raw Score. 177 participants were excluded for missing this measure. The mean score was 18.36 (s.d. 4.46), with a range of 2-31.

2.5 Human Connectome Project Development (HCPD) Data

The HCPD dataset includes healthy participants ages 8-22, with imaging data acquired at four sites across the United States (Harvard, UCLA, University of Minnesota, Washington University in St. Louis). 605 resting-state connectomes (up to 26 minutes of resting-state data) remained after excluding low-quality or high-motion data. Among these 605 participants, the average age was 14.61 (s.d. 3.90) years, ranging from 8.08-21.92 years. 53.72% of participants self-reported their sex as female, and there were 536 unique families.

For the attention measure, we used the Child Behavior Checklist (CBCL) (*29*) Attention Problems Raw Score. 462 participants had this measure available, and the mean was 2.03 (s.d. 2.56), with a range of 0-18.

For the matrix reasoning measure, we used the WISC-V (*30*) Matrix Reasoning Total Raw Score. 424 participants remained in this analysis, with a mean of 21.08 (s.d. 3.96) and a range of 11-31.

2.6 Philadelphia Neurodevelopmental Cohort (PNC) Data

The PNC dataset consists of 8-21 year-olds in the Philadelphia area who received care at the Children's Hospital of Philadelphia. 1126 resting-state scans (six minute scans) passed our exclusion criteria. The average age was 14.80 (s.d. 3.29) years, with a range of 8-21. The percentage of self-reported female participants was 54.62%.

For the attention measure, we used the Structured Interview for Prodromal Symptoms (*31*): Trouble with Focus and Attention Severity Scale (SIP001, accession code: phv00194672.v2.p2). 1104 participants had this measure available, and the mean was 1.03 (s.d. 1.19), with a range of 0-6.

For the matrix reasoning measure, we used the Penn Matrix Reasoning (*32*, *33*) Total Raw Score (PMAT_CR, accession code: phv00194834.v2.p2).1119 participants remained in this analysis, with a mean of 11.99 (s.d. 4.09) and a range of 0-24.

2.7 Baseline Models

For the primary analyses, we trained a ridge regression model using 5-fold cross-validation (*34*). For HBN, HCPD, and PNC, 5 nested folds were used for hyperparameter selection, while only 2 nested folds were used in ABCD to reduce computation time. Within the folds, the top 5% of features most significantly correlated with the phenotypic variable were selected. Furthermore, we performed a grid search over the L2 regularization parameter α ($\alpha=10^{\{-3,-2,-1,0,1,2,3\}}$), with the chosen model being the one with the highest Pearson's correlation value *r* in the nested folds.

For our baseline "gold standard" model (Figures 2-5: labeled "Gold standard"), the data were split accounting for family structure, where applicable (ABCD and HCPD only), such that all members of a single family were included in the same test split. In addition, we performed cross-validated covariate regression, where several covariates were regressed from the functional connectivity data within the cross-validation scheme (*35*, *36*). Covariates were first regressed from the training data, and then those parameters were applied to regress covariates from the test data. The covariates included mean head motion (FD), sex, and age, though age was not regressed from the data for models predicting age. Furthermore, where applicable (ABCD, HBN, and HCPD), site differences were corrected within the cross-validation scheme using ComBat (*37*–*39*). ComBat was performed separately from covariate regression because ComBat is designed for batch effects, not continuous variables (*40*). In addition to the baseline "gold standard" model, we evaluated numerous forms of leakage, as described in the following sections (see also Figure 1).

2.8 Feature leakage

Feature selection is a common step in many connectome-based machine learning pipelines. Often, it consists of determining which features are the most relevant to the phenotype of interest, such as by selecting the edge features most correlated with this phenotype. A possible mistake is selecting features in the entire dataset, then training models and evaluating performance with k-fold cross-validation or a train/test split. Whelan and Garavan (*41*) previously demonstrated inflated performance via non-nested feature selection using random data, though we think it is also useful to demonstrate this with neuroimaging data. To do so, we selected the top 5% of features in the entire dataset, and then evaluated the performance using 5-fold cross-validation.

2.9 Covariate-related leakage

2.9.1 Leaky phenotypic z-scoring

Leakage via the phenotypic variable may occur when standardizing (i.e., z-scoring) the variable across the entire dataset. To avoid leakage, one should calculate the mean and standard deviation of the phenotypic variable for only the training data within each fold, and then apply the mean and standard deviation of the training data to z-score the test data.

2.9.2 Leaky site correction

A common example of leakage is correcting for site effects, such as with ComBat (*37–39*), in the whole dataset prior to splitting data into training and test sets. To avoid leakage, one should apply ComBat to the training data within each fold of cross-validation, and then use those ComBat parameters to correct for the site of the test data, as described in a recent work by Spisak (*40*). Here, we evaluated the effect of leakage via performing ComBat on the entire dataset and compared this to the "gold standard" of applying ComBat within cross-validation folds. As described in several previous works (*37–39*), ComBat performs the correction based on the following equation for feature *v* of the data *X*, with site *i* and scan *j (37–39)*:

$$x_{i,j,v}^{adjusted} = \frac{x_{i,j,v} - \widehat{\alpha_v} - \gamma_{i,v}^{*}}{\delta_{i,v}^{*}} + \widehat{\alpha_v}$$

where α_v is the overall measure for that edge, $\gamma_{i,v}$ is the additive site effects for site *i* and feature *v*, and $\delta_{i,v}$ is the multiplicative site effects for site *i* and feature *v (38)*. However, unless performed in a cross-validated manner, leakage occurs when estimating α_v, $\gamma_{i,v}$, and $\delta_{i,v}$ on the combined training and test data. In practice, these parameters should be estimated only in the training data, and then applied to correct for site in both the training and test data.

2.9.3 Leaky covariate regression

Another common form of leakage is covariate regression. For example, regressing the covariates/confounds out of the whole dataset in a non-cross-validated manner has been shown to negatively affect prediction performance (*36*). The following equations describe regressing covariates out of the whole dataset for feature *X* of size [*N x p*], covariates *C* of size [N x (k+1)], which includes an intercept term, and the OLS solution $\hat{\beta}$ (*35*):

$$\hat{\beta} = (C^T C)^{-1} C^T X$$

$$X_{adjusted} = X - C\hat{\beta}$$

Leakage occurs above by finding the OLS solution $\hat{\beta}$ using both training and test data. Performing covariate regression within the cross-validation folds is the recommended solution to avoid leakage (*35*, *36*):

$$\hat{\beta}_{train} = (C_{train}^T C_{train})^{-1} C_{train}^T X_{train}$$

$$X_{train,adjusted} = X_{train} - C_{train}\hat{\beta}_{train}$$

$$X_{test,adjusted} = X_{test} - C_{test}\hat{\beta}_{train}$$

The above equations avoid leakage by finding the OLS solution $\hat{\beta}$ only in the training data, and subsequently applying it to the test data. Cross-validated and non-cross-validated covariate regression have been compared in several previous studies (*35*, *36*), and we build upon those previous works here by evaluating them in three phenotypes and four datasets through the view of leakage.

2.10 Subject-level leakage

2.10.1 Family leakage

Neuroimaging datasets, such as ABCD and HCPD, are often oversampled for twins or siblings. Given the heritability of brain structure and function (*42–44*), family structure should be specifically accounted for in analysis pipelines, such as in permutation tests (*45*). In the context of this work, having one family member in the training set and another in the test split is a form of leakage. For instance, in a hypothetical case of predicting a highly heritable phenotype from a highly heritable brain network, the model could “memorize” the data from one family member in the training set and strongly predict the phenotype of another family member in the test set. However, splitting the data into cross-validation folds by family instead of on the individual subject level prevents leakage. If a dataset contains 1000 participants from 500 unique families, 5-fold cross-validation should include training data belonging to 400 randomly selected *families*, and test data belonging to the other 100 randomly selected *families.*

2.10.2 Subject leakage

Another possible case of leakage is where various scans from the same participant are treated as separate samples, which we call “Subject leakage.” This could include, for instance, treating longitudinal data as cross-sectional or including various runs (or tasks) of fMRI acquisition for the same participant as separate data points. To evaluate an extreme version of subject leakage, we considered the case when a certain percentage of the subjects in the dataset were repeated. For example, in a hypothetical dataset with 1000 participants, 20% subject leakage would include an additional 200 repeats of random participants in the dataset, for a total dataset size of 1200. Then, in this larger sample, we repeated nested cross-validation and compared the prediction performance results to the original sample. Notably, we did not account for family structure in subject leakage, otherwise the leaked subjects would always be in the same training or test split.

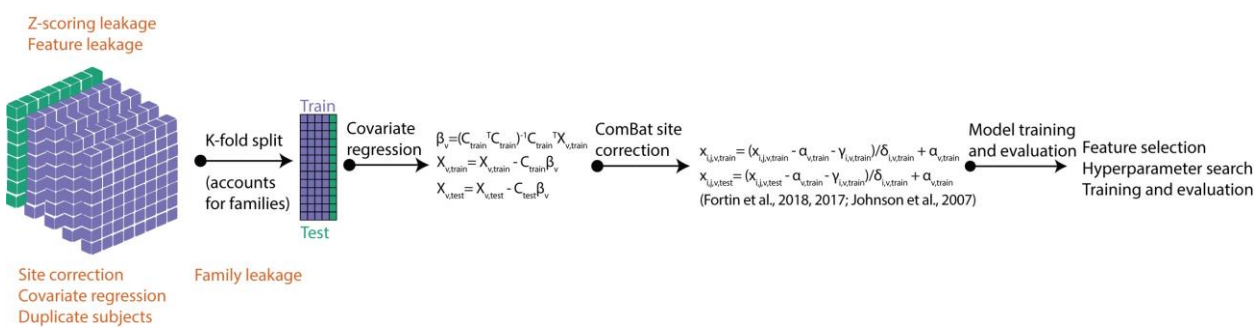

Figure 1. Summary of the prediction pipelines used in this study. The various forms of leakage that may occur are shown in orange. Z-scoring leakage, feature leakage, leaky site correction, leaky covariate regression, and subject leakage may occur prior to splitting the data into training and test sets. Family leakage may occur during the splitting of data.

2.11 Evaluation metrics

Our main evaluation metric is Pearson's correlation *r* between the true and predicted phenotype. This evaluation metric does not necessarily reflect predictive power but is commonly used to establish brain-behavior relationships (*46*). In addition, we reported cross-validation R^2, also called q^2, as defined below (*46*):

$$q^2 = \frac{\Sigma\,(y_{pred} - \bar{y}_{pred})^2}{\Sigma\,(y - \bar{y})^2}$$

2.12 Sample size

To evaluate the interaction between sample size and leakage, we randomly sampled without replacement N=100, 200, 300, and 400 participants from ABCD, HBN, HCPD, and HBN. The ABCD data was only resampled from the four largest sites (total N=1300) to avoid issues with ComBat where only one data point was present from a particular site after resampling. In addition, for the two datasets with family structure (ABCD and HCPD), data were resampled by family, not by individual, to keep approximately the same proportion of related participants in our subsample. We repeated the resampling procedure 10 times for each dataset and sample size, and for each resampled dataset we evaluated the gold standard and leaky prediction performance across 10 repeats of 5-fold nested cross-validation.

2.13 Structural connectome analysis

Our structural connectome analyses included 635 participants from the HCPD dataset. We started with the diffusion tensor data of the 635 participants. Then, we corrected for susceptibility artifacts and applied DSI-studio to reconstruct the diffusion data using generalized q-sampling imaging. Finally, we created structural connectomes using automatic fiber tracking for the Shen 268 atlas (*28*).

2.14 Additional models

We evaluated the effects of leakage in two additional models: support vector regression (SVR) (*34*) and connectome-based predictive modeling (CPM) (*2*). In both cases, we performed 5% feature selection, as described above. For SVR, the radial basis function was used, and we performed a grid search for the L2 regularization parameter ($C=10^{\{-3,-2,-1,0,1,2,3\}}$), where *C* is inversely proportional to regularization strength. For CPM, the positive and negative features were combined into a single number, and a univariate linear regression model was then fitted.

3. Results

We first evaluated the effects of leakage on prediction in HCPD (Sections *3.1-3.4*), and then, we showed effects of leakage in the other three datasets (ABCD, HBN, PNC) (Section *3.5*). Moreover, we compared model coefficients (Section *3.6*), varied the sample size (Section *3.7*), and performed sensitivity analyses (Section *3.8*).

3.1 Performance of non-leaky pipelines

We evaluated five non-leaky pipelines and found that different analysis choices led to different prediction performance (Figure 2). Our "gold standard" model, which included covariate regression and site correction within the cross-validation scheme, exhibited no prediction performance for attention (median r=0.01, q^2=-0.13), strong performance for age (r=0.80, q^2=0.63), and moderate performance for matrix reasoning (r=0.30, q^2=0.08). Performance when excluding site correction or z-scoring the phenotype was nearly identical to the gold standard model ($|\Delta r|$<0.01, Δq^2<0.01). However, not regressing out covariates inflated r but had varying effects on q^2 for all three phenotypes, including attention (Δr=0.05, Δq^2=-0.08), age (Δr=0.06, Δq^2=0.11), and matrix reasoning (Δr=0.05, Δq^2=0.01). Though not the main focus of this paper, these results highlight how prediction performance may change with different analysis choices, especially regarding whether or not covariates are regressed from the data.

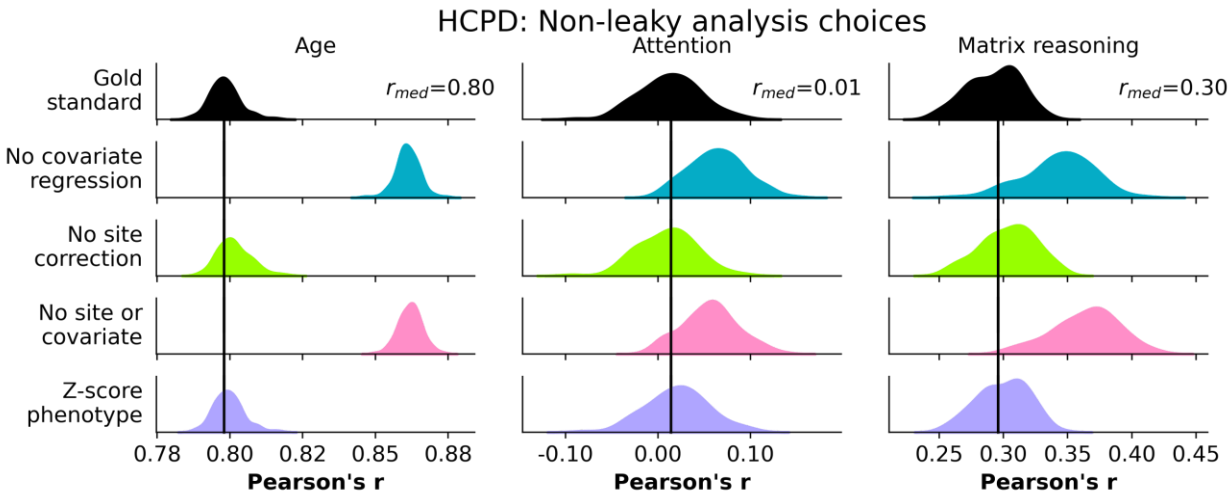

Figure 2. Prediction performance in HCPD of the non-leaky pipelines, including the gold standard, omitting covariate regression, omitting site correction, omitting both covariate regression and site correction, and z-scoring the phenotype within the cross-validation folds. Rows represent different leakage types, and columns show different phenotypes. The black bar represents the median performance of the "gold standard" models across random iterations, and the exact value of the bar is shown as the median r, r_{med}. The histograms show prediction performance across 100 iterations of 5-fold cross-validation. See also Figure S1.

3.2 Feature leakage

Feature leakage inflated prediction performance for each phenotype (Figure 3). The inflation was small for age (Δr=0.03, Δq^2=0.05), larger for matrix reasoning (Δr=0.17, Δq^2=0.13), and largest for attention (Δr=0.47, Δq^2=0.35). Notably, age had a strong baseline performance and was minimally affected by feature leakage, whereas attention had the worst baseline performance and was most affected by feature leakage. Furthermore, the attention prediction went from chance-level (r=0.01, q^2=-0.13) to moderate (r=0.48, q^2=0.22), which highlights the potential for feature leakage to hinder reproducibility efforts.

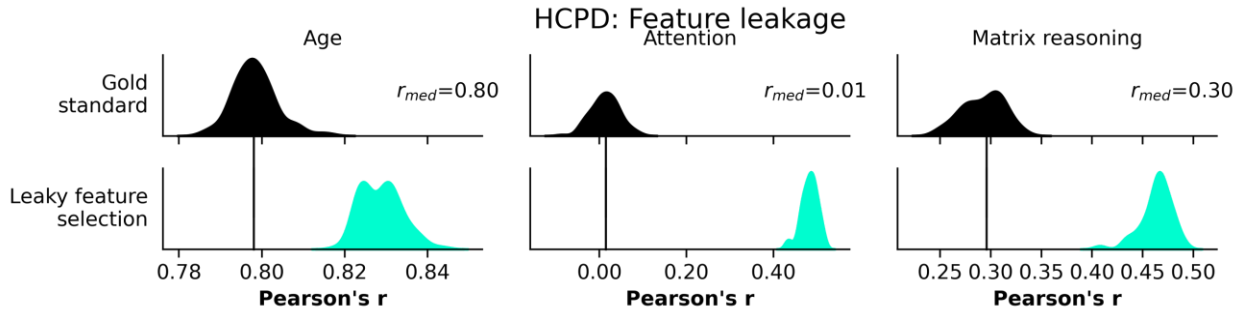

Figure 3. Prediction performance in HCPD of leaky feature selection, compared to the gold standard. Rows represent different leakage types, and columns show different phenotypes. The black bar represents the median performance of the "gold standard" models across random iterations, and the histograms show prediction performance across 100 iterations of 5-fold cross-validation. See also Figure S1.

3.3 Covariate-related leakage

Covariate-related forms of leakage in this study included z-scoring the phenotype, correcting for site differences, and performing covariate regression outside of the cross-validation folds (Figure 4). Leaky z-scoring had no effect on the performance across all three phenotypes (Δr=0.00, Δq^2=0.00). Similarly, leaky site correction had little effect on performance (Δr=-0.01-0, Δq^2=-0.01-0.01). Unlike the other forms of leakage in this study, leaky covariate regression decreased performance for attention (Δr=-0.06, Δq^2=-0.17), age (Δr=-0.02, Δq^2=-0.03), and matrix reasoning (Δr=-0.09, Δq^2=-0.08). These results illustrate that leakage can not only hamper reproducibility by false inflation of performance, but also by underestimating the true effect sizes.

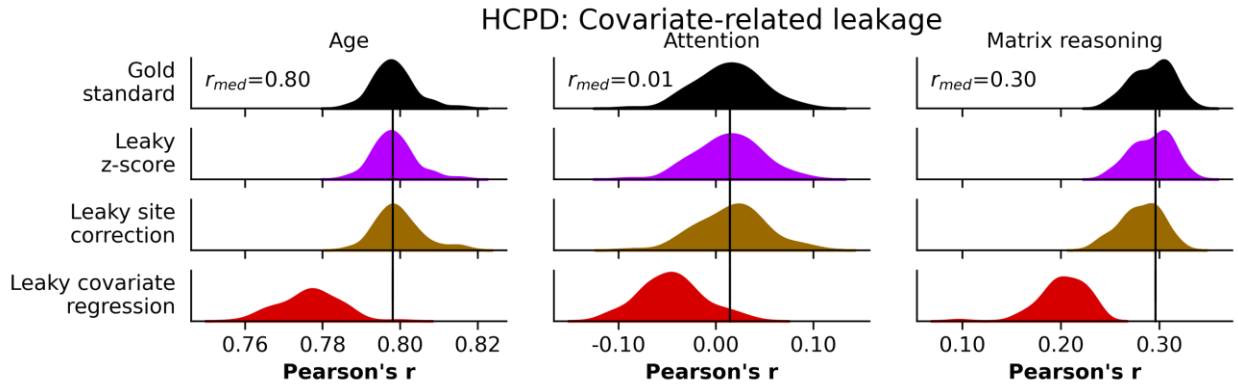

Figure 4. Prediction performance in HCPD of covariate-related forms of leakage, including leaky z-scoring of the phenotype, leaky site correction, and leaky covariate regression. Rows represent different leakage types, and columns show different phenotypes. The black bar represents the median performance of the "gold standard" models across random iterations, and the histograms show prediction performance across 100 iterations of 5-fold cross-validation. See also Figure S1.

3.4 Subject-level leakage

Since families are often oversampled in neuroimaging datasets, leakage via family structure may affect predictive models. Family leakage did not affect prediction performance of age or

matrix reasoning (Δr=0.00, Δq^2=0.00) but did slightly increase prediction performance of attention (Δr=0.02, Δq^2=0.00) (Figure 5). In addition, we evaluated subject leakage at three different levels (5%, 10%, 20%), where a certain percentage of scans were repeated and appended to the dataset (Figure 5). In all cases, subject leakage inflated prediction performance, with 20% subject leakage having the greatest impact on attention (Δr=0.28, Δq^2=0.19), age (Δr=0.04, Δq^2=0.07), and matrix reasoning (Δr=0.14, Δq^2=0.11). Similar to the trend seen in feature leakage, the effects of subject leakage were more dramatic for models with weaker baseline performance. Overall, these results suggest that family leakage may have little effect in certain instances, but potential leakage via repeated measurements (i.e., subject leakage) can largely inflate performance.

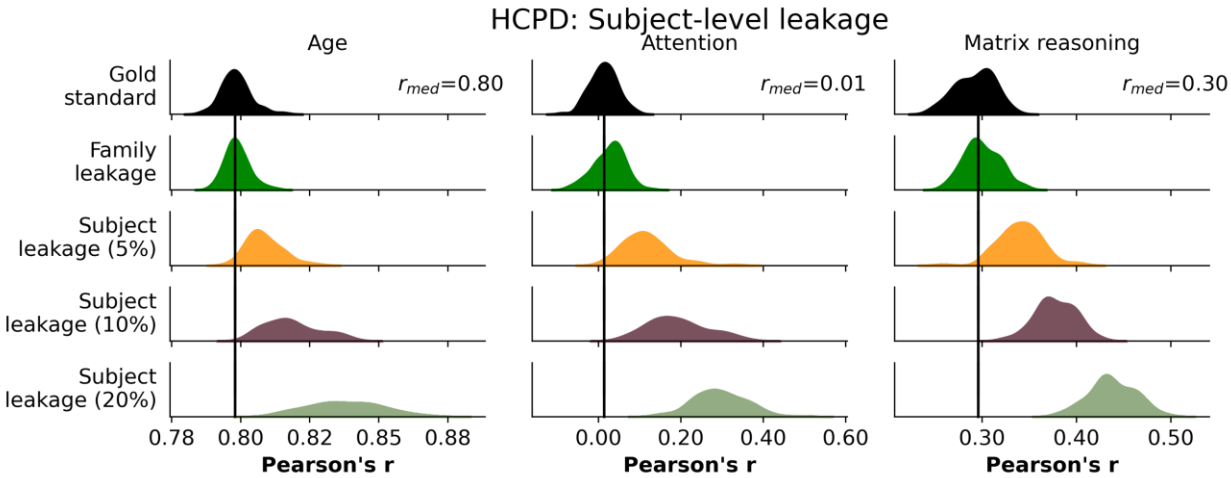

Figure 5. Prediction performance in HCPD of subject-level forms of leakage. These included family leakage and subject leakage at three different levels. Rows represent different leakage types, and columns show different phenotypes. The black bar represents the median performance of the "gold standard" models across random iterations, and the histograms show prediction performance across 100 iterations of 5-fold cross-validation. See also Figure S1.

3.5 Evaluation of performance in additional datasets

Compared to HCPD, we found similar trends in each of the 13 pipelines for ABCD, HBN, and PNC (Figures S2-3). While excluding site correction had little to no effect in HCPD or HBN, there was a small effect in ABCD (Δr=-0.01-0.02, Δq^2=0.00). Furthermore, not performing covariate regression generally inflated performance relative to the baseline for attention (Δr=0.02-0.04, Δq^2=-0.02-0.04), age (Δr=0.00-0.04, Δq^2=0.00-0.05), and matrix reasoning (Δr=0.02-0.05, Δq^2=-0.02-0.01).

Across all datasets and phenotypes, leaky feature selection and subject leakage (20%) led to the greatest performance inflation. Feature leakage had varying effects based on the dataset and phenotype (Δr=0.02-0.52, Δq^2=0.01-0.47). The dataset with the largest sample size (ABCD), was least affected by leaky feature selection, and weaker baseline models were more affected by feature leakage. Subject leakage (20%) also inflated performance across all datasets and phenotypes (Δr=0.04-0.29, Δq^2=0.02-0.24). Corroborating the results in HCPD,

leaky covariate regression was the only form of leakage that consistently deflated performance (Δr=-0.06-0.00, Δq^2=-0.07-0.00). The other forms of leakage had little to no effect (Δr=-0.01-0.00, Δq^2=0.00), as in HCPD.

The change in performance relative to the gold standard for each pipeline across all four datasets and three phenotypes is summarized in Figure 6. Overall, only leaky feature selection and subject leakage inflated prediction performance in this study.

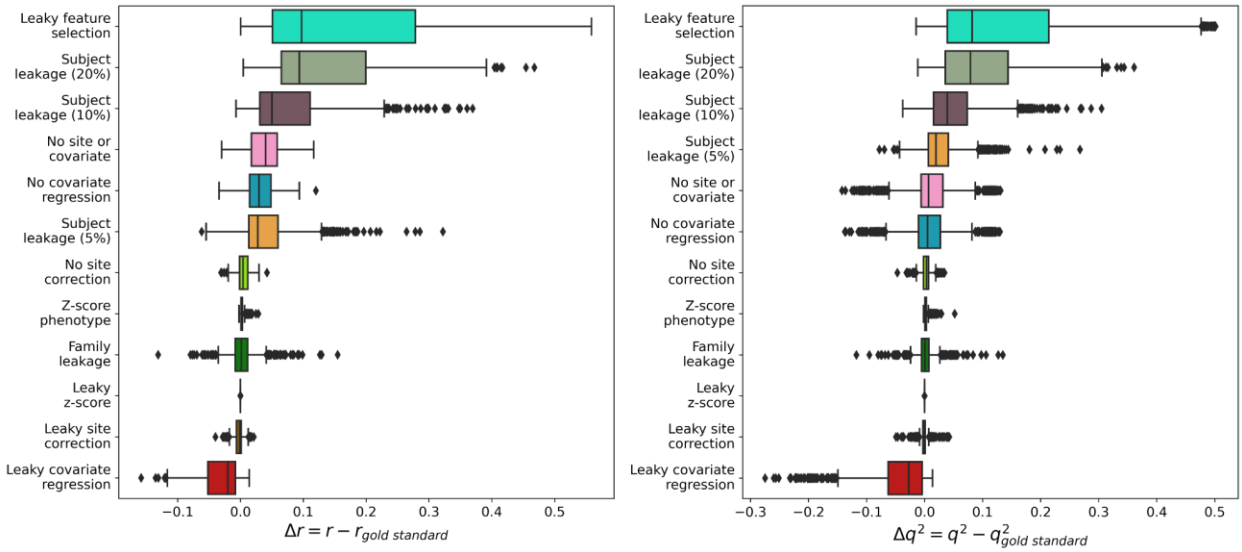

Figure 6. Evaluation of performance differences between all pipelines and the gold standard pipeline across all datasets and phenotypes for Pearson's r (left) and q^2 (right). For each of r and q^2, the plots are ranked from the most inflated performance (top) to the most deflated performance (bottom). See also Figure S2, Figure S3.

3.6 Comparing coefficients in leaky and non-leaky pipelines

Determining whether the performance of leaky and non-leaky pipelines is similar only tells part of the story, as two models could have similar prediction performance but learn entirely different brain-behavior relationships. As such, establishing how model coefficients may change for various forms of leakage is an equally important facet of understanding the possible effects of leakage. We first averaged the coefficients over the five folds of cross-validation and calculated the correlation between those coefficients and the coefficients of the gold standard model (Figure 7). Excluding covariate regression (median r_{coef}=0.31-0.86), site correction (median r_{coef}=0.86-0.99), or both (median r_{coef}=0.32-0.82) resulted in moderate changes in the coefficients. Amongst the forms of leakage, leaky feature selection unsurprisingly was the most dissimilar to the gold standard coefficients (median r_{coef}=0.39-0.76). Other forms of leakage that notably affected the coefficients included family leakage (median r_{coef}=0.79-0.94) and 20% subject leakage (median r_{coef}=0.68-0.93). Otherwise, the coefficients between the leaky and gold standard pipelines were very similar. We also compared the coefficients for each pair of 13 analysis pipelines (Figure S4). Interestingly, although the coefficients when excluding covariate regression or performing leaky feature selection were relatively dissimilar to the gold standard coefficients (median r_{coef}=0.31-0.86), these coefficients were relatively similar to each other

(median r_{coef}=0.68-0.92). This result could possibly be explained by covariates contributing to brain-behavior associations in the whole dataset.

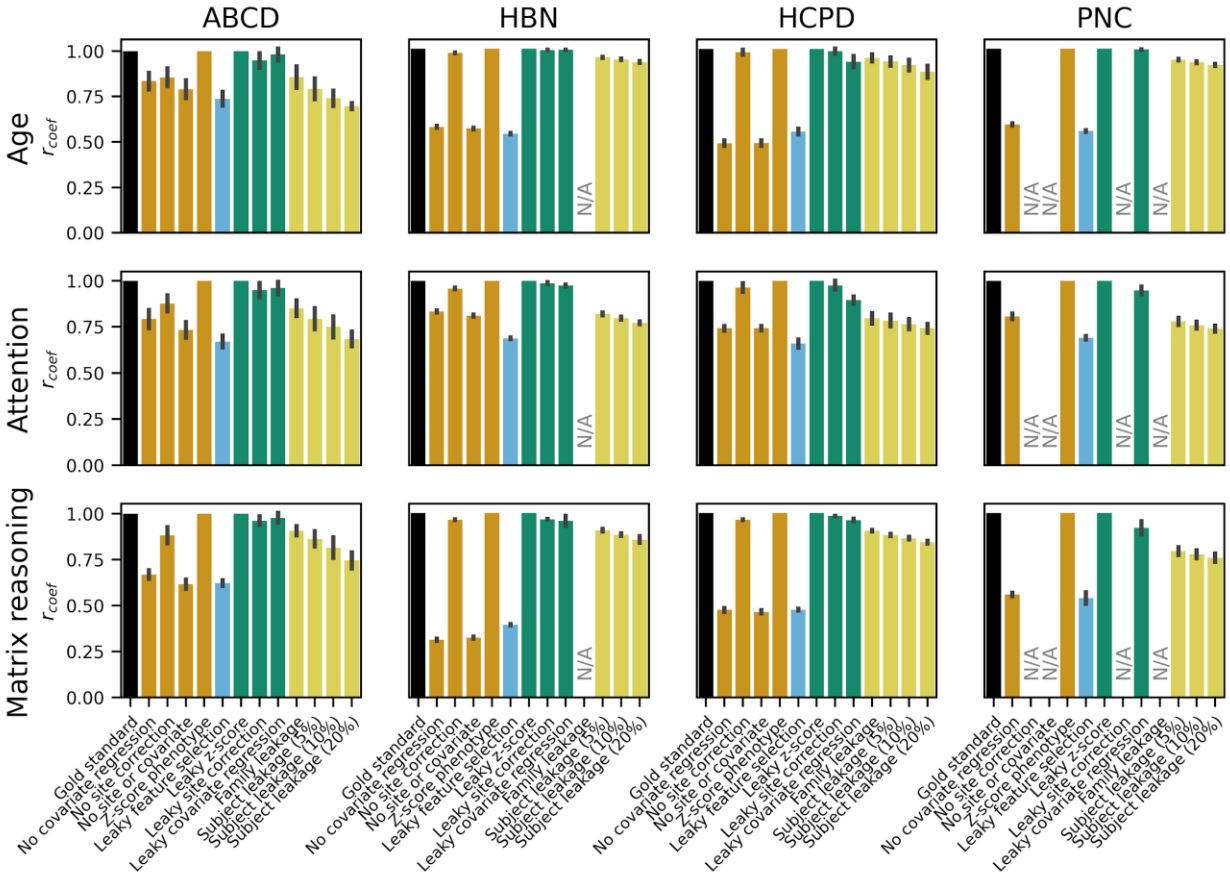

Figure 7. Similarity of coefficients between the gold standard and various forms of leakage. The bars are colored by leakage family: black (gold standard), orange (non-leaky analysis choices), blue (feature leakage), green (covariate-related leakage), yellow (subject-level leakage). The bars show the average correlations across 100 iterations, and the error bars represent the standard deviation across the 100 iterations. Missing values (i.e., no site information in PNC) are marked with "N/A." See also Figures S4-5.

In addition to correlating the coefficients at the edge level, we also considered the similarity of feature selection across 10 canonical networks (*47*) (Figure S5). We calculated the number of edges selected as features in each of the 55 subnetworks, which are defined as a connection between a specific pair of the 10 canonical networks. We then adjusted for subnetwork size and compared the rank correlation across different leakage types. Similar to the previous analysis, not performing covariate regression changed the distribution of features across subnetworks (median $r_{spearman,network}$=0.28-0.90). Amongst the forms of leakage, leaky feature selection showed the greatest network differences compared to the gold standard model (median $r_{spearman,network}$=0.25-0.85), while the other forms of leakage showed smaller differences (median $r_{spearman,network}$=0.75-1.00).

3.7 Effect of sample size

All previously presented results investigated the four datasets at their full sample sizes (ABCD: N=4079-4150, HBN: N=1024-1201, HCPD: N=424-605, PNC: N=1104-1126). Yet, although these can lead to less reproducible results (*21*), smaller sample sizes are common in neuroimaging studies. As such, consideration of how leakage may affect reported prediction performance at various sample sizes is crucial. For leaky feature selection, leaky site correction, leaky covariate regression, family leakage, and 20% subject leakage, we computed $\Delta r = r_{leaky} - r_{gold}$, where r_{leaky} is the performance of the leaky pipeline and r_{gold} is the performance of the gold standard non-leaky pipeline for a single seed of 5-fold cross-validation. For each combination of leakage type, sample size, and dataset, Δr was evaluated for 10 different resamples for 10 iterations of 5-fold cross-validation each (over 20,000 evaluations of 5-fold cross-validation in total) (Figure 8). In general, the variability of Δr was much greater for the smallest sample size (N=100) compared to the largest sample size (N=400). For instance, for matrix reasoning prediction in ABCD, Δr for family leakage ranged from -0.28 to 0.28 for N=100 and from -0.13 to 0.13 for N=400. Another example is site correction in ABCD matrix reasoning prediction, where Δr ranged from -0.14 to 0.05 for N=100 and from -0.06 to 0.04 for N=400. While not every dataset and phenotype prediction had large variability in performance for leaky pipelines at small sample sizes (e.g., HBN age prediction), the overall trends suggest that leakage may be more unpredictable and thus dangerous in small samples compared to large samples.

However, when taking the median performance across multiple k-fold splits for a given subsample, the effects of most leakage types, except feature and subject leakage, decreased (Figure S6). In general, the best practice is to perform at least 100 iterations of k-fold splits, but due to the many analyses and pipelines in this study, we only performed 10 iterations. For instance, for the ABCD prediction of matrix reasoning, taking the median across 10 iterations resulted in a smaller range of Δr values for all forms of leakage (N=400), including feature leakage ($\Delta r_{multiple}$=0.12-0.39 for multiple iterations vs. Δr_{single}=0.11-0.46 for a single iteration), leaky site correction ($\Delta r_{multiple}$=-0.03-0.01 vs. Δr_{single}=-0.06-0.04), leaky covariate regression ($\Delta r_{multiple}$=-0.06-0.13 vs. Δr_{single}=-0.11-0.15), family leakage ($\Delta r_{multiple}$=-0.02-0.02 vs. Δr_{single}=-0.13-0.13), and 20% subject leakage ($\Delta r_{multiple}$=0.11-0.27 vs. Δr_{single}=0.04-0.38). Overall, performing multiple iterations of k-fold cross-validation reduced, but did not eliminate, the effects of leakage. Leakage still led to large changes in performance in certain cases, particularly at small sample sizes.

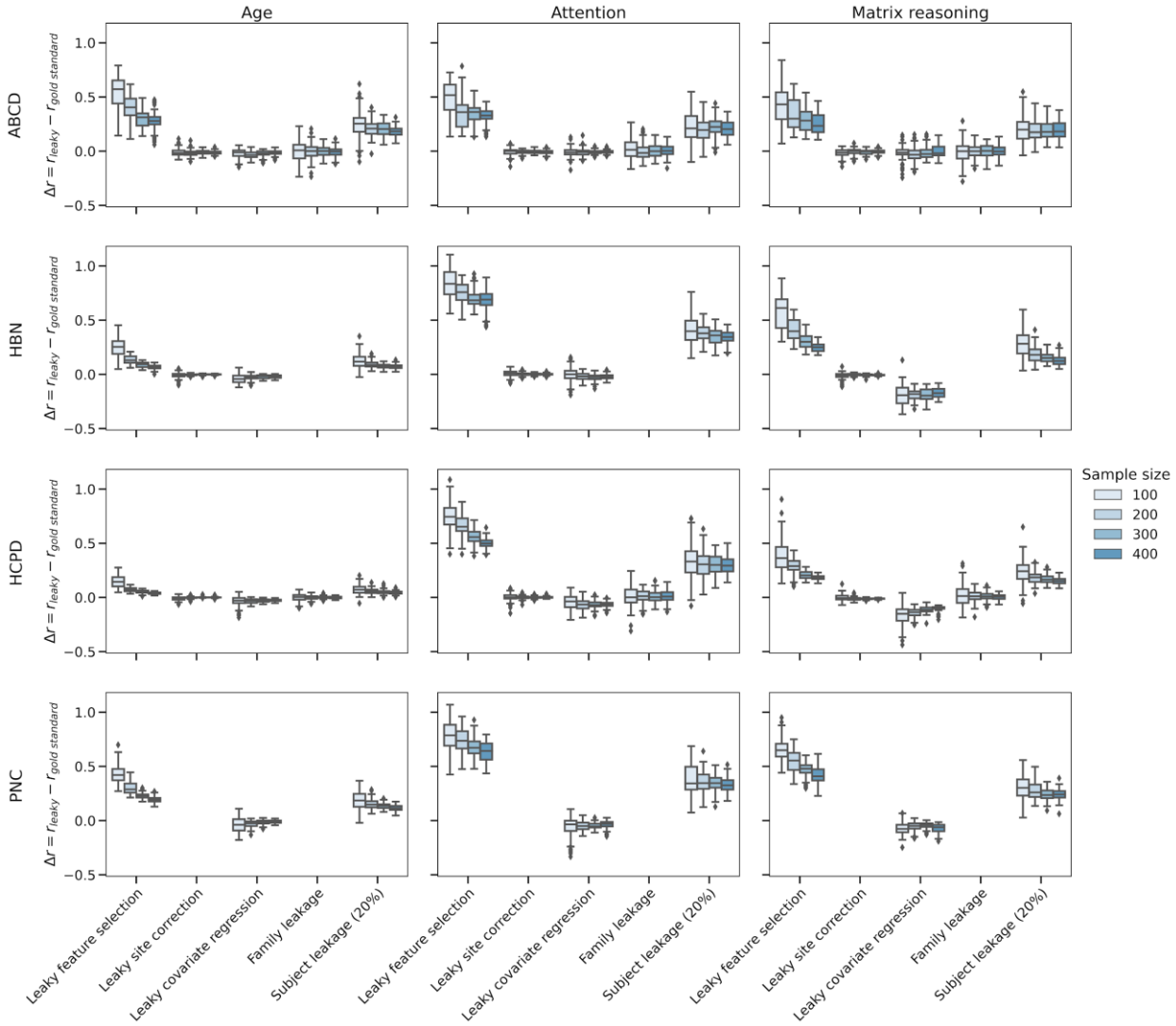

Figure 8. Difference between leaky and gold standard performance for various types of leakage and two sample sizes (N=100, 200, 300, 400). Rows represent the dataset, and columns show the phenotype. For each leakage type (x-axis), there are four results (N=100, 200, 300, 400). For each sample size, we repeated 10 random seeds of resampling for 10 iterations of 5-fold cross-validation. The boxplots show the range of performance across all seeds and iterations. See also Figure S6.

3.8 Sensitivity analyses

Two main sensitivity analyses were performed to support the robustness of our findings. First, we analyzed the effects of leakage in two other models (SVR, CPM). Second, we performed similar analyses using structural connectomes to demonstrate the effects of leakage beyond functional connectivity.

3.8.1 Other models

We repeated the analysis for SVR (Figures S7, S9) and CPM (Figures S8, S10) and found similar trends in the effects of leakage. CPM generally had a slightly lower “gold standard”

performance (age: median r=0.08-0.61, q^2=0.00-0.37; attention: r=-0.04-0.04, q^2=-0.25-0.00; matrix reasoning: r=0.07-0.27, q^2=0.00-0.05) than ridge regression (age: r=0.13-0.80, q^2=0.01-0.63; attention: r=-0.01-0.06, q^2=-0.21 to -0.01; matrix reasoning: r=0.10-0.30, q^2=0.01-0.08) and SVR (age: r=0.12-0.80, q^2=-0.04-0.64; attention: r=0.00-0.05, q^2=-0.15 to -0.11; matrix reasoning: r=0.12-0.34, q^2=-0.02-0.10). Notably, CPM was less affected by leaky feature selection (Δr=-0.04-0.39, Δq^2=0.00-0.38) compared to ridge regression (Δr=0.02-0.52, Δq^2=0.01-0.47) and SVR (Δr=0.03-0.48, Δq^2=0.01-0.38). In addition, subject leakage had the largest effect on SVR (Δr=0.07-0.43, Δq^2=0.12-0.31), followed by ridge regression (Δr=0.04-0.29, Δq^2=0.02-0.24) and CPM (Δr=0.01-0.17, Δq^2=0.00-0.16). Regardless of differences in the size of the effects across models, the trends were generally similar.

3.8.2 Structural connectome data

Gold standard predictions of matrix reasoning, attention, and age exhibited low to moderate performance in the HCPD structural connectome data (Figure 9, S11) (matrix reasoning: median r=0.34, q^2=0.12; attention: r=0.11, q^2=-0.07; age: r=0.73, q^2=0.53). The forms of leakage that most inflated the performance were feature leakage (Δr=0.07-0.57, Δq^2=0.12-0.52) and subject leakage (Δr=0.05-0.27, Δq^2=0.06-0.20). Compared to its effect on functional connectivity data, leaky covariate regression in this particular instance showed milder reductions in performance (Δr=-0.04-0.00, Δq^2=-0.04-0.00). Despite minor differences, these results in structural connectivity data follow similar trends as the functional connectivity data.

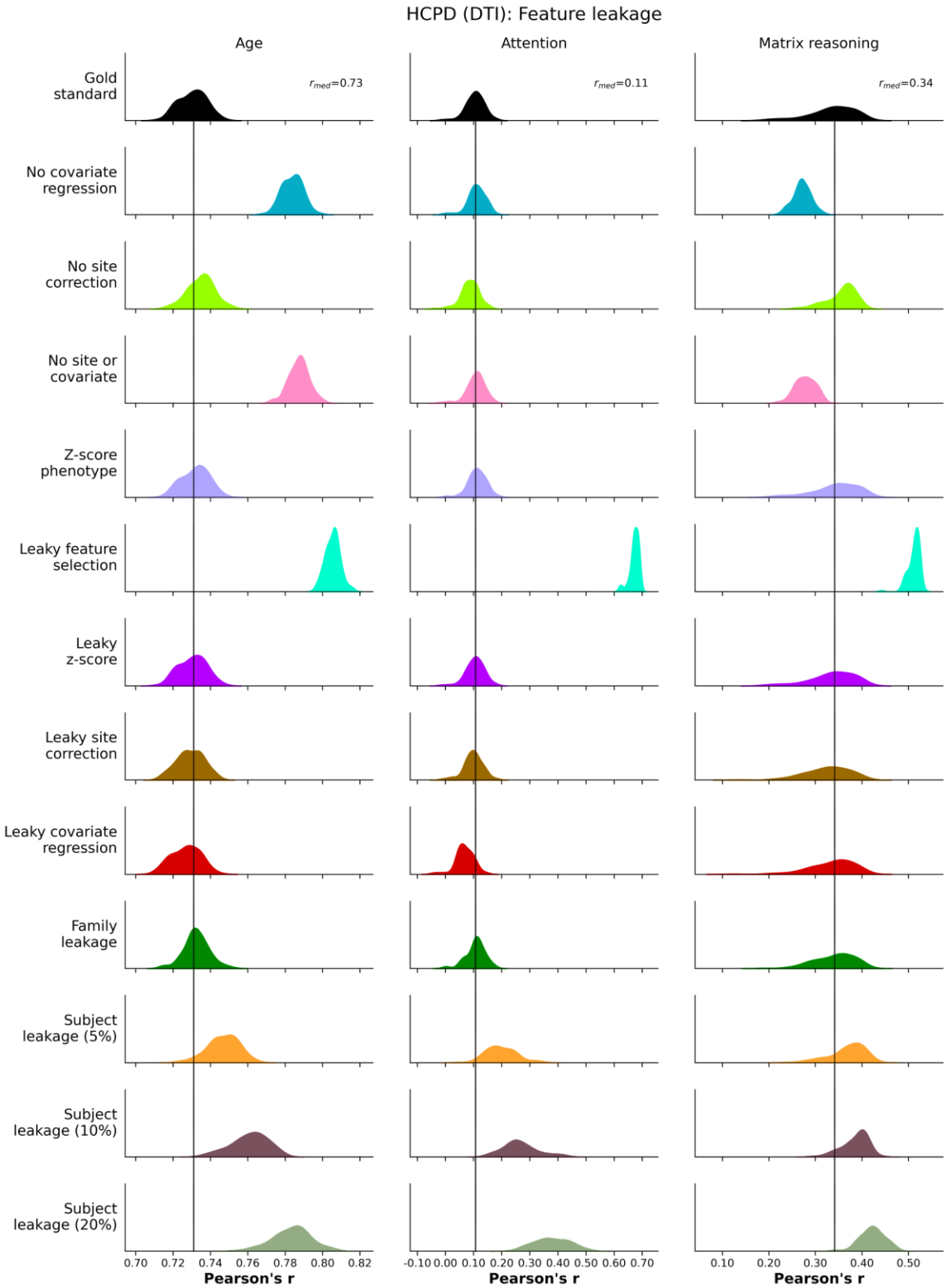

Figure 9. Evaluation of leakage types for matrix reasoning, attention, and age prediction in structural connectomes using *r*. The rows show different leakage types, and the columns show different phenotypes. The black bar represents the median performance of the "gold standard" models across random iterations, and the histograms show prediction performance across 100 iterations of 5-fold cross-validation. See also Figure S11.

4. Discussion

In this work, we demonstrated the effects of six possible forms of leakage on connectome-based predictive models in the ABCD, HBN, HCPD, and PNC datasets. In some cases, leakage led to severe inflation of prediction (e.g., leaky feature selection). In others, there was little to no difference (e.g., leaky site correction). The overall effects of leaky pipelines showed similar trends across the different phenotypes, models, and connectomes investigated in this work. Furthermore, subsampling to smaller sizes--typical in the neuroimaging literature--led to an increased effect of leakage. These results highlight the importance of the strict splitting of testing and training samples to accurately estimate the performance of a predictive model.

Feature leakage is widely accepted as a bad practice, and as expected, it severely inflated prediction performance. Although feature leakage is likely rare in the literature, it can dramatically increase model performance and thus hinder reproducibility. For instance, a recent work (*48*) demonstrated that a high-profile article “predicting” suicidal ideation in youth had no predictive power after removing feature selection leakage. The original paper, which has now been retracted, received 243 citations since its publication in 2017 according to Google Scholar. As such, it is essential to re-emphasize the importance of avoiding feature leakage. Though avoiding feature leakage may appear obvious, it can occur in more subtle ways. For example, one may investigate which networks significantly differ between two groups in the whole dataset and then create a predictive model using those networks. Notably, the effects of feature leakage were smaller in ABCD due to its large sample size. In other words, when using thousands of samples, the selected features are likely robust across different folds of training data. This result is consistent with recent findings regarding association studies (*21*). In general, the occurrence of feature leakage can be reduced by sharing code on public repositories, which would allow the community to quickly investigate suspicious results.

Similarly, subject leakage led to inflated effects. It is more likely in datasets with multiple runs of fMRI, time points, or tasks. For example, a high-profile preprint used deep learning to predict pneumonia from chest X-rays, and the authors did not account for patients having multiple scans, causing leakage across the training and test sets. Fortunately, this leakage was identified by a community member and quickly corrected in a subsequent version of the preprint (*49*), which illustrates the importance of writing detailed methods and sharing code. Extra care should be taken for prediction when using multiple scans per individual, such as longitudinal data.

We typically associate leakage with inflated prediction performance. However, leaky covariate regression deflated prediction performance. Our results corroborate previous works, suggesting that covariate regression must be performed within the cross-validation loop to avoid false deflation of effect sizes (*35*, *36*, *50*). Interestingly, performing covariate regression itself may lead to leakage (*51*) and is another consideration when deciding whether and how to implement covariate regression.

Aside from feature and subject leakage, no other leakage significantly increased prediction performance when using large sample sizes. Notably, this result included leaky site correction and family leakage. Large-open source datasets, like ABCD and the broader HCP lifespan datasets, are increasingly multi-site with complex nested dependency between participants (i.e., twins) (*52*). These factors facilitate larger sample sizes for better statistical power and more representative samples, which can minimize model bias (*1*, *53–55*). However, accounting for these factors can rapidly increase the complexity of a prediction pipeline. Thus, these results are reassuring for the broader field. Overall, they likely mean that any results with these forms of leakage remain valid. Although, there is no guarantee that any of these forms of leakage will not inflate performance. Thus, avoiding data leakage is still necessary to ensure valid results.

Among the phenotypes predicted, age, which had the best prediction performance, was the least affected by leakage. This result suggests that leakage may moreso affect phenotypes with a weaker brain-behavior association. When there is a strong brain-behavior relationship, a model may capture the relevant pattern regardless of leakage. However, when there is a weak brain-behavior relationship, the model may predominantly rely on patterns arising from leakage.

Crucially, leakage exhibited more variable effects in smaller datasets. As such, accounting for leakage is even more crucial with small samples. Taking the median performance of the models over multiple iterations of k-fold cross-validation (i.e., different random seeds) mitigated the inflation. This example underscores the benefits of performing many (>=100) iterations of k-fold cross-validation (*46*, *56*). Though k-fold cross-validation is the most common form of evaluation in neuroimaging, train/test splits are not uncommon. Given that train/test splits are often only performed for one random seed, leakage with small sample sizes may be a greater issue when using train/test splits.

Along with its effects on performance, we found that leakage also affected model interpretation, and therefore neurobiological interpretation. Coefficients for feature leakage were unsurprisingly dissimilar to the gold standard because the leaky feature selection relies on one feature subset, whereas the gold standard pipeline selects a different subset of features for each fold of cross-validation. Otherwise, the most notable differences in coefficients arose from omitting covariate regression. This result highlights that, in addition to avoiding leakage, researchers should consider how various analysis choices may affect results (*57*, *58*).

Among many other benefits, code sharing could decrease leakage by allowing external reviewers to investigate published pipelines for leakage. Furthermore, model checklists (*17*) may contribute to more accurate and reproducible results by identifying and preventing leakage prior to publication.

While this study investigated several datasets, modalities, phenotypes, and models, we cannot exhaustively test every permutation. Other factors may contribute to the severity of leakage. Complex models like neural networks are likely more susceptible to leakage due to their ability to memorize data (*59*, *60*). Other possible forms of leakage, like leakage via hyperparameter selection (*61*), were not considered in this study. Furthermore, in this work, we considered

omission of covariate regression and site correction to be a “non-leaky analysis choice.” However, it could arguably be considered leakage via illegitimate features, meaning that the model predicts using features to which it should not have access (*17*).

Concerns about reproducibility in machine learning (*17*) can be partially attributed to leakage. As expected, feature and subject leakage inflated prediction performance. Positively, many forms of leakage did not exhibit inflated results. Additionally, larger samples and running multiple train and test splits mitigated inflated results. Since the effects of leakage are wide-varying and not known beforehand, the best practice remains to be vigilant and avoid data leakage altogether.

References

1. L. Tejavibulya, M. Rolison, S. Gao, Q. Liang, H. Peterson, J. Dadashkarimi, M. C. Farruggia, C. A. Hahn, S. Noble, S. D. Lichenstein, A. Pollatou, A. J. Dufford, D. Scheinost, Predicting the future of neuroimaging predictive models in mental health. *Mol. Psychiatry*. **27**, 3129–3137 (2022).

2. X. Shen, E. S. Finn, D. Scheinost, M. D. Rosenberg, M. M. Chun, X. Papademetris, R. T. Constable, Using connectome-based predictive modeling to predict individual behavior from brain connectivity. *Nat. Protoc.* **12**, 506–518 (2017).

3. J. Sui, R. Jiang, J. Bustillo, V. Calhoun, Neuroimaging-based Individualized Prediction of Cognition and Behavior for Mental Disorders and Health: Methods and Promises. *Biol. Psychiatry*. **88**, 818–828 (2020).

4. D. S. Barron, S. Gao, J. Dadashkarimi, A. S. Greene, M. N. Spann, S. Noble, E. M. R. Lake, J. H. Krystal, R. T. Constable, D. Scheinost, Transdiagnostic, Connectome-Based Prediction of Memory Constructs Across Psychiatric Disorders. *Cereb. Cortex*. **31**, 2523–2533 (2020).

5. H. Song, M. D. Rosenberg, Predicting attention across time and contexts with functional brain connectivity. *Current Opinion in Behavioral Sciences*. **40**, 33–44 (2021).

6. R. Boyle, M. Connaughton, E. McGlinchey, S. P. Knight, C. De Looze, D. Carey, Y. Stern, I. H. Robertson, R. A. Kenny, R. Whelan, Connectome-based predictive modelling of cognitive reserve using task-based functional connectivity. *Eur. J. Neurosci.* **57**, 490–510 (2023).

7. A. Tetereva, N. Pat, The (Limited?) Utility of Brain Age as a Biomarker for Capturing Cognitive Decline. *bioRxiv* (2023), p. 2022.12.31.522374.

8. M. J. Lund, D. Alnæs, A.-M. G. de Lange, O. A. Andreassen, L. T. Westlye, T. Kaufmann, Brain age prediction using fMRI network coupling in youths and associations with psychiatric symptoms. *Neuroimage Clin*. **33**, 102921 (2022).

9. O. Kardan, S. Kaplan, M. D. Wheelock, E. Feczko, T. K. M. Day, Ó. Miranda-Domínguez, D. Meyer, A. T. Eggebrecht, L. A. Moore, S. Sung, T. A. Chamberlain, E. Earl, K. Snider, A. Graham, M. G. Berman, K. Uğurbil, E. Yacoub, J. T. Elison, C. D. Smyser, D. A. Fair, M. D. Rosenberg, Resting-state functional connectivity identifies individuals and predicts age in 8-

to-26-month-olds. *Dev. Cogn. Neurosci.* **56**, 101123 (2022).

10. J. Zhai, K. Li, Predicting Brain Age Based on Spatial and Temporal Features of Human Brain Functional Networks. *Front. Hum. Neurosci.* **13**, 62 (2019).

11. M. Gao, N. M. L. Wong, C. Lin, C.-M. Huang, H.-L. Liu, C.-H. Toh, C. Wu, Y.-F. Tsai, S.-H. Lee, T. M. C. Lee, Multimodal brain connectome-based prediction of suicide risk in people with late-life depression. *Nature Mental Health*. **1**, 100–113 (2023).

12. X. Wu, Q. Yang, C. Xu, H. Huo, C. A. Seger, Z. Peng, Q. Chen, Connectome-based predictive modeling of compulsion in obsessive–compulsive disorder. *Cereb. Cortex*. **33**, 1412–1425 (2022).

13. K. A. Garrison, R. Sinha, M. N. Potenza, S. Gao, Q. Liang, C. Lacadie, D. Scheinost, Transdiagnostic Connectome-Based Prediction of Craving. *Am. J. Psychiatry*. **180**, 445–453 (2023).

14. T. Spisak, U. Bingel, T. D. Wager, Multivariate BWAS can be replicable with moderate sample sizes. *Nature*. **615** (2023), pp. E4–E7.

15. D. Bzdok, G. Varoquaux, E. W. Steyerberg, Prediction, Not Association, Paves the Road to Precision Medicine. *JAMA Psychiatry*. **78**, 127–128 (2021).

16. S. Kaufman, S. Rosset, C. Perlich, O. Stitelman, Leakage in data mining: Formulation, detection, and avoidance. *ACM Trans. Knowl. Discov. Data*. **6**, 1–21 (2012).

17. S. Kapoor, A. Narayanan, Leakage and the Reproducibility Crisis in ML-based Science. *arXiv [cs.LG]* (2022), (available at http://arxiv.org/abs/2207.07048).

18. R. A. Poldrack, G. Huckins, G. Varoquaux, Establishment of Best Practices for Evidence for Prediction: A Review. *JAMA Psychiatry*. **77**, 534–540 (2020).

19. R. A. Poldrack, C. I. Baker, J. Durnez, K. J. Gorgolewski, P. M. Matthews, M. R. Munafò, T. E. Nichols, J.-B. Poline, E. Vul, T. Yarkoni, Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nat. Rev. Neurosci.* **18**, 115–126 (2017).

20. R. Botvinik-Nezer, T. D. Wager, Reproducibility in Neuroimaging Analysis: Challenges and Solutions. *Biol Psychiatry Cogn Neurosci Neuroimaging* (2022), doi:10.1016/j.bpsc.2022.12.006.

21. S. Marek, B. Tervo-Clemmens, F. J. Calabro, D. F. Montez, B. P. Kay, A. S. Hatoum, M. R. Donohue, W. Foran, R. L. Miller, T. J. Hendrickson, S. M. Malone, S. Kandala, E. Feczko, O. Miranda-Dominguez, A. M. Graham, E. A. Earl, A. J. Perrone, M. Cordova, O. Doyle, L. A. Moore, G. M. Conan, J. Uriarte, K. Snider, B. J. Lynch, J. C. Wilgenbusch, T. Pengo, A. Tam, J. Chen, D. J. Newbold, A. Zheng, N. A. Seider, A. N. Van, A. Metoki, R. J. Chauvin, T. O. Laumann, D. J. Greene, S. E. Petersen, H. Garavan, W. K. Thompson, T. E. Nichols, B. T. T. Yeo, D. M. Barch, B. Luna, D. A. Fair, N. U. F. Dosenbach, Reproducible brain-wide association studies require thousands of individuals. *Nature*. **605**, E11 (2022).

22. B. J. Casey, T. Cannonier, M. I. Conley, A. O. Cohen, D. M. Barch, M. M. Heitzeg, M. E. Soules, T. Teslovich, D. V. Dellarco, H. Garavan, C. A. Orr, T. D. Wager, M. T. Banich, N. K. Speer, M. T. Sutherland, M. C. Riedel, A. S. Dick, J. M. Bjork, K. M. Thomas, B.

Chaarani, M. H. Mejia, D. J. Hagler Jr, M. Daniela Cornejo, C. S. Sicat, M. P. Harms, N. U. F. Dosenbach, M. Rosenberg, E. Earl, H. Bartsch, R. Watts, J. R. Polimeni, J. M. Kuperman, D. A. Fair, A. M. Dale, ABCD Imaging Acquisition Workgroup, The Adolescent Brain Cognitive Development (ABCD) study: Imaging acquisition across 21 sites. *Dev. Cogn. Neurosci.* **32**, 43–54 (2018).

23. L. M. Alexander, J. Escalera, L. Ai, C. Andreotti, K. Febre, A. Mangone, N. Vega-Potler, N. Langer, A. Alexander, M. Kovacs, S. Litke, B. O'Hagan, J. Andersen, B. Bronstein, A. Bui, M. Bushey, H. Butler, V. Castagna, N. Camacho, E. Chan, D. Citera, J. Clucas, S. Cohen, S. Dufek, M. Eaves, B. Fradera, J. Gardner, N. Grant-Villegas, G. Green, C. Gregory, E. Hart, S. Harris, M. Horton, D. Kahn, K. Kabotyanski, B. Karmel, S. P. Kelly, K. Kleinman, B. Koo, E. Kramer, E. Lennon, C. Lord, G. Mantello, A. Margolis, K. R. Merikangas, J. Milham, G. Minniti, R. Neuhaus, A. Levine, Y. Osman, L. C. Parra, K. R. Pugh, A. Racanello, A. Restrepo, T. Saltzman, B. Septimus, R. Tobe, R. Waltz, A. Williams, A. Yeo, F. X. Castellanos, A. Klein, T. Paus, B. L. Leventhal, R. C. Craddock, H. S. Koplewicz, M. P. Milham, An open resource for transdiagnostic research in pediatric mental health and learning disorders. *Sci Data*. **4**, 170181 (2017).

24. L. H. Somerville, S. Y. Bookheimer, R. L. Buckner, G. C. Burgess, S. W. Curtiss, M. Dapretto, J. S. Elam, M. S. Gaffrey, M. P. Harms, C. Hodge, S. Kandala, E. K. Kastman, T. E. Nichols, B. L. Schlaggar, S. M. Smith, K. M. Thomas, E. Yacoub, D. C. Van Essen, D. M. Barch, The Lifespan Human Connectome Project in Development: A large-scale study of brain connectivity development in 5-21 year olds. *Neuroimage*. **183**, 456–468 (2018).

25. T. D. Satterthwaite, M. A. Elliott, K. Ruparel, J. Loughead, K. Prabhakaran, M. E. Calkins, R. Hopson, C. Jackson, J. Keefe, M. Riley, F. D. Mentch, P. Sleiman, R. Verma, C. Davatzikos, H. Hakonarson, R. C. Gur, R. E. Gur, Neuroimaging of the Philadelphia neurodevelopmental cohort. *Neuroimage*. **86**, 544–553 (2014).

26. T. D. Satterthwaite, J. J. Connolly, K. Ruparel, M. E. Calkins, C. Jackson, M. A. Elliott, D. R. Roalf, R. Hopson, K. Prabhakaran, M. Behr, H. Qiu, F. D. Mentch, R. Chiavacci, P. M. A. Sleiman, R. C. Gur, H. Hakonarson, R. E. Gur, The Philadelphia Neurodevelopmental Cohort: A publicly available resource for the study of normal and abnormal brain development in youth. *Neuroimage*. **124**, 1115–1119 (2016).

27. X. Papademetris, M. P. Jackowski, N. Rajeevan, M. DiStasio, H. Okuda, R. T. Constable, L. H. Staib, BioImage Suite: An integrated medical image analysis suite: An update. *Insight J.* **2006**, 209 (2006).

28. X. Shen, F. Tokoglu, X. Papademetris, R. T. Constable, Groupwise whole-brain parcellation from resting-state fMRI data for network node identification. *Neuroimage*. **82**, 403–415 (2013).

29. T. M. Achenbach, T. M. Ruffle, The Child Behavior Checklist and related forms for assessing behavioral/emotional problems and competencies. *Pediatr. Rev.* **21**, 265–271 (2000).

30. D. Wechsler, *WISC-V: Technical and interpretive manual* (NCS Pearson, Incorporated, 2014).

31. T. J. Miller, T. H. McGlashan, J. L. Rosen, K. Cadenhead, J. Ventura, W. McFarlane, D. O. Perkins, G. D. Pearlson, S. W. Woods, Prodromal assessment with the structured interview

for prodromal syndromes and the scale of prodromal symptoms: predictive validity, interrater reliability, and training to reliability. *Schizophr. Bull.* **29**, 703–715 (2003).

32. W. B. Bilker, J. A. Hansen, C. M. Brensinger, J. Richard, R. E. Gur, R. C. Gur, Development of abbreviated nine-item forms of the Raven's standard progressive matrices test. *Assessment*. **19**, 354–369 (2012).

33. T. M. Moore, S. P. Reise, R. E. Gur, H. Hakonarson, R. C. Gur, Psychometric properties of the Penn Computerized Neurocognitive Battery. *Neuropsychology*. **29**, 235–246 (2015).

34. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, Others, Scikit-learn: Machine learning in Python. *The Journal of machine Learning research*. **12**, 2825–2830 (2011).

35. L. Snoek, S. Miletić, H. S. Scholte, How to control for confounds in decoding analyses of neuroimaging data. *Neuroimage*. **184**, 741–760 (2019).

36. D. Chyzhyk, G. Varoquaux, M. Milham, B. Thirion, How to remove or control confounds in predictive models, with applications to brain biomarkers. *Gigascience*. **11** (2022), doi:10.1093/gigascience/giac014.

37. W. E. Johnson, C. Li, A. Rabinovic, Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*. **8**, 118–127 (2007).

38. J.-P. Fortin, D. Parker, B. Tunç, T. Watanabe, M. A. Elliott, K. Ruparel, D. R. Roalf, T. D. Satterthwaite, R. C. Gur, R. E. Gur, R. T. Schultz, R. Verma, R. T. Shinohara, Harmonization of multi-site diffusion tensor imaging data. *Neuroimage*. **161**, 149–170 (2017).

39. J.-P. Fortin, N. Cullen, Y. I. Sheline, W. D. Taylor, I. Aselcioglu, P. A. Cook, P. Adams, C. Cooper, M. Fava, P. J. McGrath, M. McInnis, M. L. Phillips, M. H. Trivedi, M. M. Weissman, R. T. Shinohara, Harmonization of cortical thickness measurements across scanners and sites. *Neuroimage*. **167**, 104–120 (2018).

40. T. Spisak, Statistical quantification of confounding bias in machine learning models. *Gigascience*. **11** (2022), doi:10.1093/gigascience/giac082.

41. R. Whelan, H. Garavan, When optimism hurts: inflated predictions in psychiatric neuroimaging. *Biol. Psychiatry*. **75**, 746–748 (2014).

42. J. W. Koten Jr, G. Wood, P. Hagoort, R. Goebel, P. Propping, K. Willmes, D. I. Boomsma, Genetic contribution to variation in cognitive function: an FMRI study in twins. *Science*. **323**, 1737–1740 (2009).

43. B. M. Adhikari, N. Jahanshad, D. Shukla, D. C. Glahn, J. Blangero, R. C. Reynolds, R. W. Cox, E. Fieremans, J. Veraart, D. S. Novikov, T. E. Nichols, L. E. Hong, P. M. Thompson, P. Kochunov, Heritability estimates on resting state fMRI data using ENIGMA analysis pipeline. *Pac. Symp. Biocomput.* **23**, 307–318 (2018).

44. B. Zhao, J. G. Ibrahim, Y. Li, T. Li, Y. Wang, Y. Shan, Z. Zhu, F. Zhou, J. Zhang, C. Huang, H. Liao, L. Yang, P. M. Thompson, H. Zhu, Heritability of Regional Brain Volumes in Large-Scale Neuroimaging and Genetic Studies. *Cereb. Cortex*. **29**, 2904–2914 (2019).

45. A. M. Winkler, M. A. Webster, D. Vidaurre, T. E. Nichols, S. M. Smith, Multi-level block permutation. *Neuroimage*. **123**, 253–268 (2015).

46. D. Scheinost, S. Noble, C. Horien, A. S. Greene, E. M. Lake, M. Salehi, S. Gao, X. Shen, D. O'Connor, D. S. Barron, S. W. Yip, M. D. Rosenberg, R. T. Constable, Ten simple rules for predictive modeling of individual differences in neuroimaging. *Neuroimage*. **193**, 35–45 (2019).

47. S. Noble, A. F. Mejia, A. Zalesky, D. Scheinost, Improving power in functional magnetic resonance imaging by moving beyond cluster-level inference. *Proc. Natl. Acad. Sci.* **119**, e2203020119 (2022).

48. T. Verstynen, K. P. Kording, Overfitting to "predict"suicidal ideation. *Nature Human Behaviour*, 1–2 (2023).

49. P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, M. P. Lungren, A. Y. Ng, CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. *arXiv [cs.CV]* (2017), (available at http://arxiv.org/abs/1711.05225).

50. S. More, S. B. Eickhoff, J. Caspers, K. R. Patil, "Confound Removal and Normalization in Practice: A Neuroimaging Based Sex Prediction Case Study" in *Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track* (Springer International Publishing, 2021; http://dx.doi.org/10.1007/978-3-030-67670-4_1), pp. 3–18.

51. S. Hamdan, B. C. Love, G. G. von Polier, S. Weis, H. Schwender, S. B. Eickhoff, K. R. Patil, Confound-leakage: Confound Removal in Machine Learning Leads to Leakage. *arXiv [cs.LG]* (2022), (available at http://arxiv.org/abs/2210.09232).

52. C. Horien, S. Noble, A. S. Greene, K. Lee, D. S. Barron, S. Gao, D. O'Connor, M. Salehi, J. Dadashkarimi, X. Shen, E. M. R. Lake, R. T. Constable, D. Scheinost, A hitchhiker's guide to working with large, open-source neuroimaging datasets. *Nat Hum Behav*. **5**, 185–193 (2021).

53. J. Li, D. Bzdok, J. Chen, A. Tam, L. Q. R. Ooi, A. J. Holmes, T. Ge, K. R. Patil, M. Jabbi, S. B. Eickhoff, B. T. T. Yeo, S. Genon, Cross-ethnicity/race generalization failure of behavioral prediction from resting-state functional connectivity. *Sci Adv*. **8**, eabj1812 (2022).

54. Z. Chen, X. Liu, Q. Yang, Y.-J. Wang, K. Miao, Z. Gong, Y. Yu, A. Leonov, C. Liu, Z. Feng, H. Chuan-Peng, Evaluation of Risk of Bias in Neuroimaging-Based Artificial Intelligence Models for Psychiatric Diagnosis: A Systematic Review. *JAMA Netw Open*. **6**, e231671 (2023).

55. R. Wang, P. Chaudhari, C. Davatzikos, Bias in machine learning models can be significantly mitigated by careful training: Evidence from neuroimaging studies. *Proc. Natl. Acad. Sci. U. S. A.* **120**, e2211613120 (2023).

56. G. Varoquaux, P. R. Raamana, D. A. Engemann, A. Hoyos-Idrobo, Y. Schwartz, B. Thirion, Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. *Neuroimage*. **145**, 166–179 (2017).

57. R. Botvinik-Nezer, F. Holzmeister, C. F. Camerer, A. Dreber, J. Huber, M. Johannesson, M.

Kirchler, R. Iwanir, J. A. Mumford, R. A. Adcock, P. Avesani, B. M. Baczkowski, A. Bajracharya, L. Bakst, S. Ball, M. Barilari, N. Bault, D. Beaton, J. Beitner, R. G. Benoit, R. M. W. J. Berkers, J. P. Bhanji, B. B. Biswal, S. Bobadilla-Suarez, T. Bortolini, K. L. Bottenhorn, A. Bowring, S. Braem, H. R. Brooks, E. G. Brudner, C. B. Calderon, J. A. Camilleri, J. J. Castrellon, L. Cecchetti, E. C. Cieslik, Z. J. Cole, O. Collignon, R. W. Cox, W. A. Cunningham, S. Czoschke, K. Dadi, C. P. Davis, A. D. Luca, M. R. Delgado, L. Demetriou, J. B. Dennison, X. Di, E. W. Dickie, E. Dobryakova, C. L. Donnat, J. Dukart, N. W. Duncan, J. Durnez, A. Eed, S. B. Eickhoff, A. Erhart, L. Fontanesi, G. M. Fricke, S. Fu, A. Galván, R. Gau, S. Genon, T. Glatard, E. Glerean, J. J. Goeman, S. A. E. Golowin, C. González-García, K. J. Gorgolewski, C. L. Grady, M. A. Green, J. F. Guassi Moreira, O. Guest, S. Hakimi, J. P. Hamilton, R. Hancock, G. Handjaras, B. B. Harry, C. Hawco, P. Herholz, G. Herman, S. Heunis, F. Hoffstaedter, J. Hogeveen, S. Holmes, C.-P. Hu, S. A. Huettel, M. E. Hughes, V. Iacovella, A. D. Iordan, P. M. Isager, A. I. Isik, A. Jahn, M. R. Johnson, T. Johnstone, M. J. E. Joseph, A. C. Juliano, J. W. Kable, M. Kassinopoulos, C. Koba, X.-Z. Kong, T. R. Koscik, N. E. Kucukboyaci, B. A. Kuhl, S. Kupek, A. R. Laird, C. Lamm, R. Langner, N. Lauharatanahirun, H. Lee, S. Lee, A. Leemans, A. Leo, E. Lesage, F. Li, M. Y. C. Li, P. C. Lim, E. N. Lintz, S. W. Liphardt, A. B. Losecaat Vermeer, B. C. Love, M. L. Mack, N. Malpica, T. Marins, C. Maumet, K. McDonald, J. T. McGuire, H. Melero, A. S. Méndez Leal, B. Meyer, K. N. Meyer, G. Mihai, G. D. Mitsis, J. Moll, D. M. Nielson, G. Nilsonne, M. P. Notter, E. Olivetti, A. I. Onicas, P. Papale, K. R. Patil, J. E. Peelle, A. Pérez, D. Pischedda, J.-B. Poline, Y. Prystauka, S. Ray, P. A. Reuter-Lorenz, R. C. Reynolds, E. Ricciardi, J. R. Rieck, A. M. Rodriguez-Thompson, A. Romyn, T. Salo, G. R. Samanez-Larkin, E. Sanz-Morales, M. L. Schlichting, D. H. Schultz, Q. Shen, M. A. Sheridan, J. A. Silvers, K. Skagerlund, A. Smith, D. V. Smith, P. Sokol-Hessner, S. R. Steinkamp, S. M. Tashjian, B. Thirion, J. N. Thorp, G. Tinghög, L. Tisdall, S. H. Tompson, C. Toro-Serey, J. J. Torre Tresols, L. Tozzi, V. Truong, L. Turella, A. E. van ’t Veer, T. Verguts, J. M. Vettel, S. Vijayarajah, K. Vo, M. B. Wall, W. D. Weeda, S. Weis, D. J. White, D. Wisniewski, A. Xifra-Porxas, E. A. Yearling, S. Yoon, R. Yuan, K. S. L. Yuen, L. Zhang, X. Zhang, J. E. Zosky, T. E. Nichols, R. A. Poldrack, T. Schonberg, Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*. **582**, 84–88 (2020).

58. J. Dafflon, P. F Da Costa, F. Váša, R. P. Monti, D. Bzdok, P. J. Hellyer, F. Turkheimer, J. Smallwood, E. Jones, R. Leech, A guided multiverse study of neuroimaging analyses. *Nat. Commun.* **13**, 3758 (2022).

59. N. Carlini, C. Liu, Ú. Erlingsson, J. Kos, D. Song, "The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks" in *USENIX Security Symposium* (2019; https://www.usenix.org/system/files/sec19-carlini.pdf), vol. 267.

60. V. Feldman, "Does learning require memorization? a short tale about a long tail" in *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing* (Association for Computing Machinery, New York, NY, USA, 2020; https://doi.org/10.1145/3357713.3384290), *STOC 2020*, pp. 954–959.

61. M. Hosseini, M. Powell, J. Collins, C. Callahan-Flintoft, W. Jones, H. Bowman, B. Wyble, I tried a bunch of things: The dangers of unexpected overfitting in classification of brain data. *Neurosci. Biobehav. Rev.* **119**, 456–467 (2020).

Data and code availability

Data are available through the Adolescent Brain Cognitive Development Study (*22*), the Healthy Brain Network Dataset (*23*), the Human Connectome Project Development Dataset (*24*), and the Philadelphia Neurodevelopmental Cohort Dataset (*25*, *26*). Code for the analyses is available at: https://github.com/mattrosenblatt7/leakage_neuroimaging.

Acknowledgements

This study was supported by the National Institute of Mental Health grant R01MH121095 (obtained by D.S.). M.R. was supported by the National Science Foundation Graduate Research Fellowship under grant DGE2139841. L.T. was supported by the Gruber Science Fellowship. S.N. was supported by the National Institute of Mental Health under grant K00MH122372. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the funding agencies.

The Human Connectome Project Development data was supported by the National Institute Of Mental Health of the National Institutes of Health under Award Number U01MH109589 and by funds provided by the McDonnell Center for Systems Neuroscience at Washington University in St. Louis. The HCP-Development 2.0 Release data used in this report came from DOI: 10.15154/1520708. Additional data used in the preparation of this article were obtained from the Adolescent Brain Cognitive DevelopmentSM (ABCD) Study (https://abcdstudy.org), held in the NIMH Data Archive (NDA). This is a multisite, longitudinal study designed to recruit more than 10,000 children age 9-10 and follow them over 10 years into early adulthood. The ABCD Study® is supported by the National Institutes of Health and additional federal partners under award numbers U01DA041048, U01DA050989, U01DA051016, U01DA041022, U01DA051018, U01DA051037, U01DA050987, U01DA041174, U01DA041106, U01DA041117, U01DA041028, U01DA041134, U01DA050988, U01DA051039, U01DA041156, U01DA041025, U01DA041120, U01DA051038, U01DA041148, U01DA041093, U01DA041089, U24DA041123, U24DA041147. A full list of supporters is available at https://abcdstudy.org/federal-partners.html. A listing of participating sites and a complete listing of the study investigators can be found at https://abcdstudy.org/consortium_members/. ABCD consortium investigators designed and implemented the study and/or provided data but did not necessarily participate in the analysis or writing of this report. This manuscript reflects the views of the authors and may not reflect the opinions or views of the NIH or ABCD consortium investigators. The Healthy Brain Network (http://www.healthybrainnetwork.org) and its initiatives are supported by philanthropic contributions from the following individuals, foundations and organizations: Margaret Bilotti; Brooklyn Nets; Agapi and Bruce Burkard; James Chang; Phyllis Green and Randolph Cōwen; Grieve Family Fund; Susan Miller and Byron Grote; Sarah and Geoff Gund; George Hall; Jonathan M. Harris Family Foundation; Joseph P. Healey; The Hearst Foundations; Eve and Ross Jaffe; Howard & Irene Levine Family Foundation; Rachael and Marshall Levine; George and Nitzia Logothetis; Christine and Richard Mack; Julie Minskoff; Valerie Mnuchin; Morgan Stanley Foundation; Amy and John Phelan; Roberts Family Foundation; Jim and Linda Robinson Foundation, Inc.; The Schaps Family; Zibby Schwarzman; Abigail Pogrebin and David

Shapiro; Stavros Niarchos Foundation; Preethi Krishna and Ram Sundaram; Amy and John Weinberg; Donors to the 2013 Child Advocacy Award Dinner Auction; Donors to the 2012 Brant Art Auction. Additional data were provided by the PNC (principal investigators Hakon Hakonarson and Raquel Gur; phs000607.v1.p1). Support for the collection of these datasets was provided by grant RC2MH089983 awarded to Raquel Gur and RC2MH089924 awarded to Hakon Hakonarson

Supplementary Materials for

The effects of data leakage on neuroimaging predictive models

Matthew Rosenblatt et al.

Present manuscript	**Kapoor and Narayanan** (*17*)
Leaky feature selection	Feature selection jointly on the training and test sets
Leaky z-score	Pre-processing on the training and test sets
Leaky site correction	Pre-processing on the training and test sets
Leaky covariate regression	Pre-processing on the training and test sets
Family leakage	Non-independence between the training and test sets; Sampling bias
Subject leakage	Duplicate data points

Table S1. Mapping of the leakage types used in this study to those used by Kapoor and Narayanan (*17*).

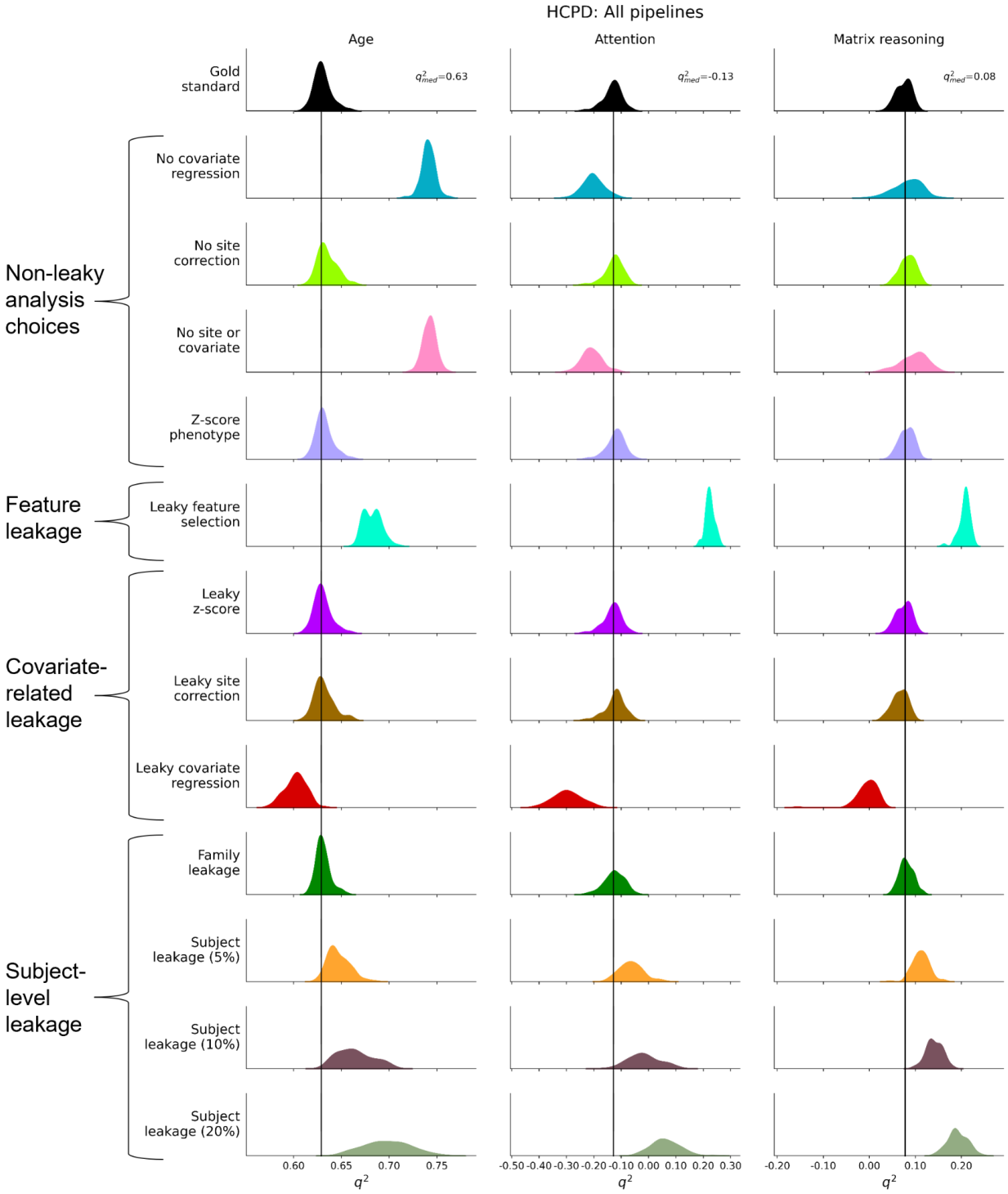

Figure S1. Evaluation of leakage in HCPD with a metric of q^2, related to Figures 2-5. Rows represent different leakage types, and columns show different phenotypes. The black bar represents the median performance of the "gold standard" models across random iterations, and the histograms show prediction performance across 100 iterations of 5-fold cross-validation.

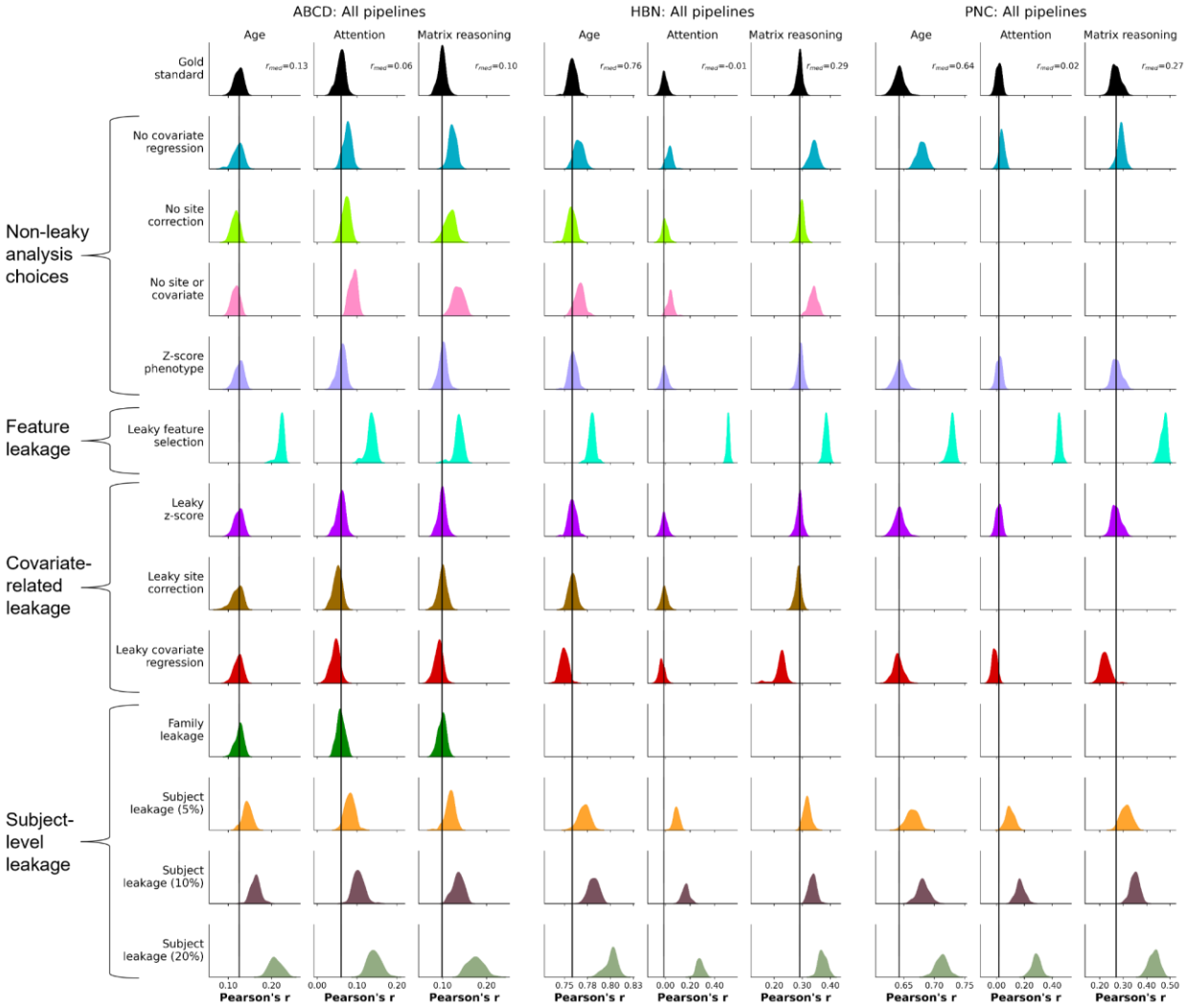

Figure S2. Summary of leakage types in additional three datasets (ABCD: left, HBN: center, PNC: right), related to Figure 6. The rows show different leakage types, and, within each dataset, the columns show different phenotypes. The black bar represents the median performance of the "gold standard" models across random iterations, and the histograms show prediction performance across 100 iterations of 5-fold cross-validation.

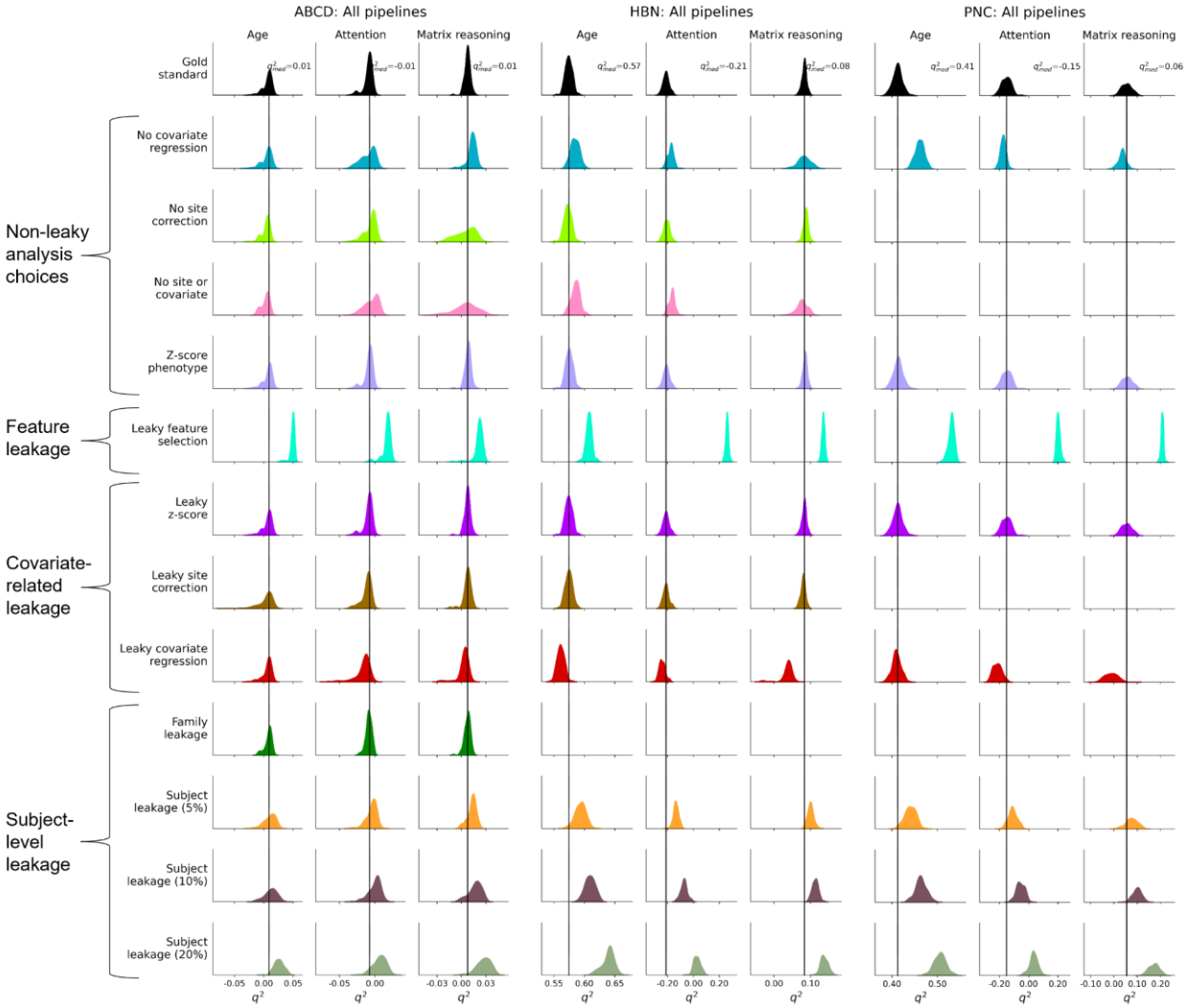

Figure S3. Evaluation of leakage in ABCD, HBN, and PNC with a metric of q^2, related to Figure 6. Rows represent different leakage types, and columns show different phenotypes. The black bar represents the median performance of the "gold standard" models across random iterations, and the histograms show prediction performance across 100 iterations of 5-fold cross-validation.

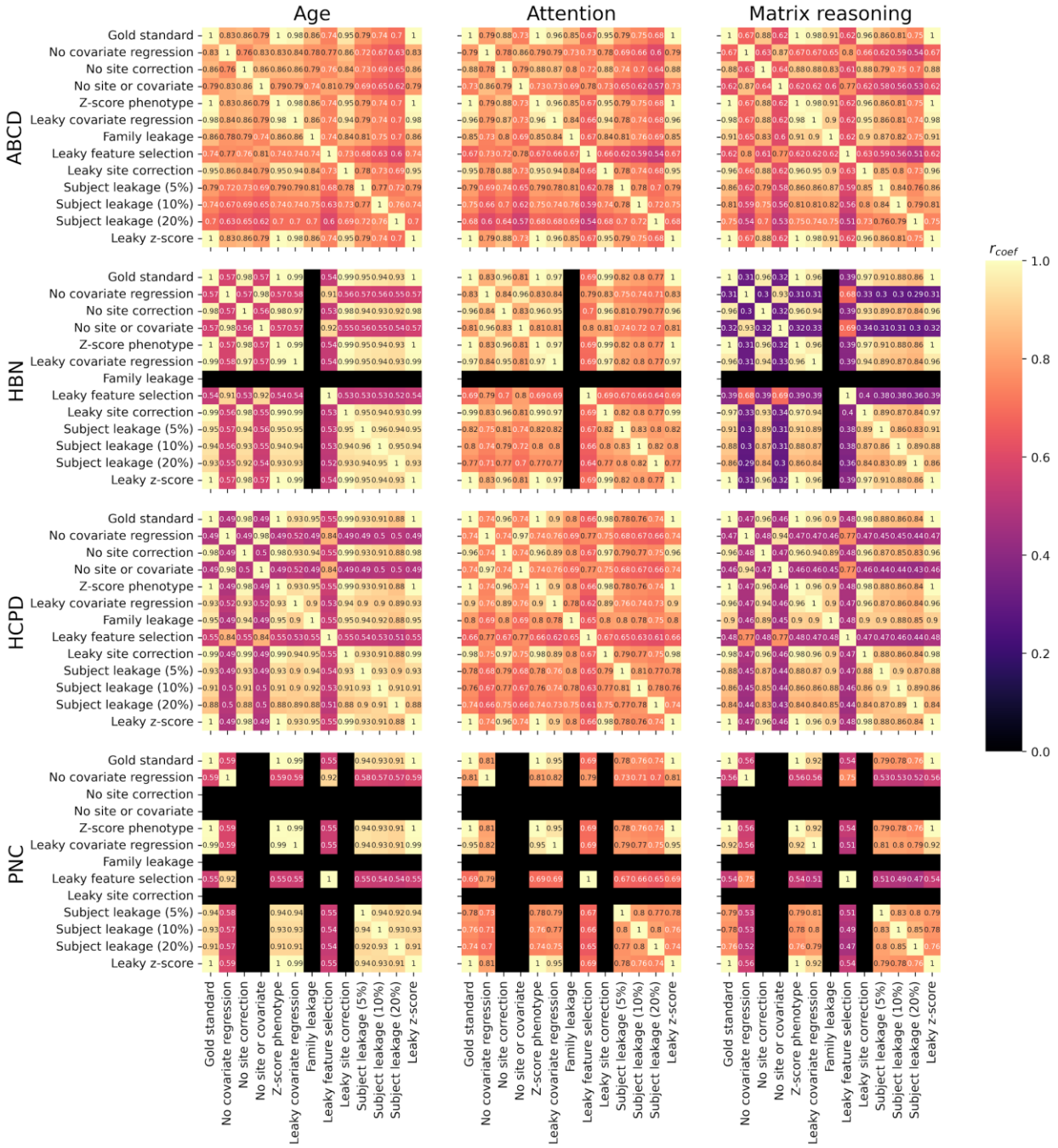

Figure S4. Similarity of coefficients across all pipelines, averaged over 100 random seeds, related to Figure 7. For each pair of the 13 pipelines, we computed the correlation between their coefficients. Missing values (i.e., no site information in PNC) are shown in black.

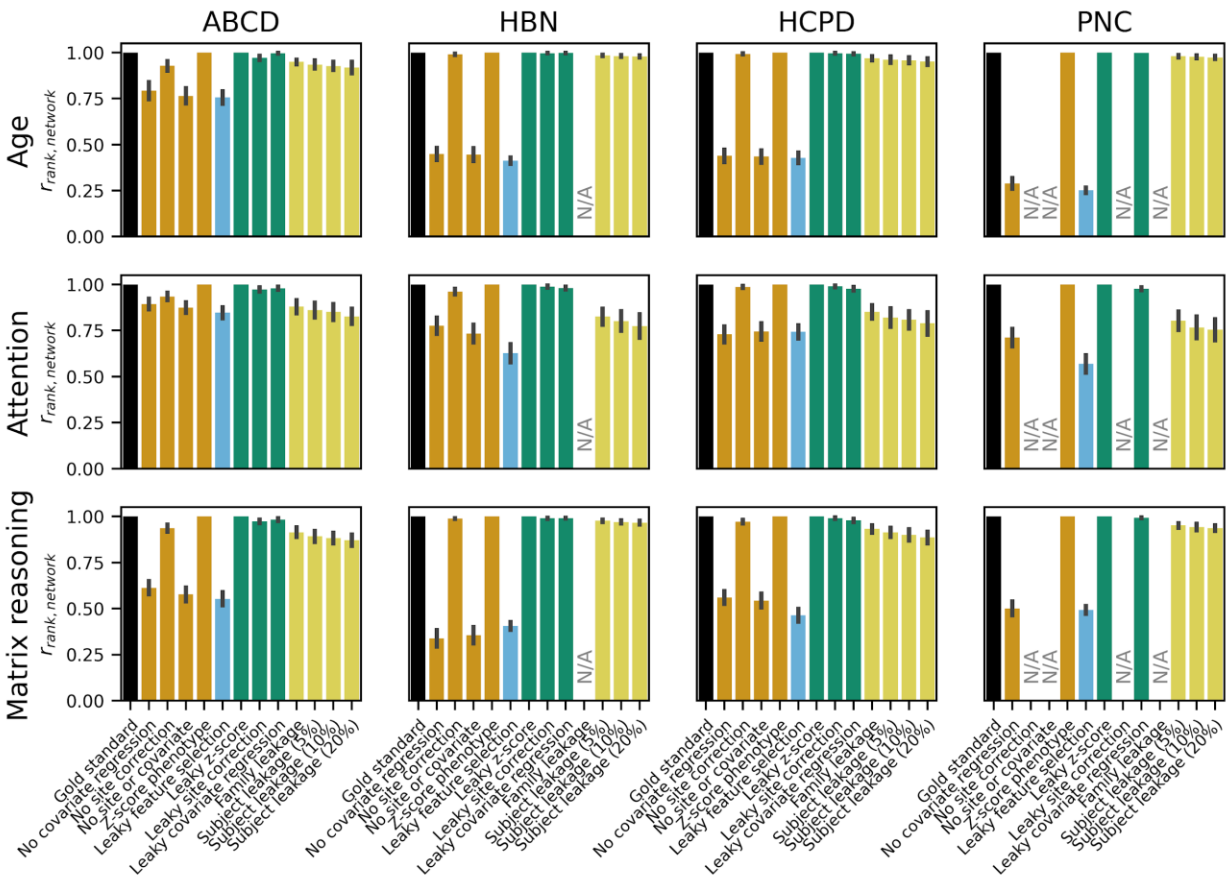

Figure S5. Subnetwork-level rank correlations for number of selected features, related to Figure 7. The bars are colored by leakage family: black (gold standard), orange (non-leaky analysis choices), blue (feature leakage), green (covariate-related leakage), yellow (subject-level leakage). To compare the gold standard features to features from the other pipelines, we calculated the rank correlation between the number of features in each subnetwork (adjusted for subnetwork size). Each bar represents the rank correlation of the specific pipeline with the gold standard, and error bars show the standard deviation across 100 iterations.

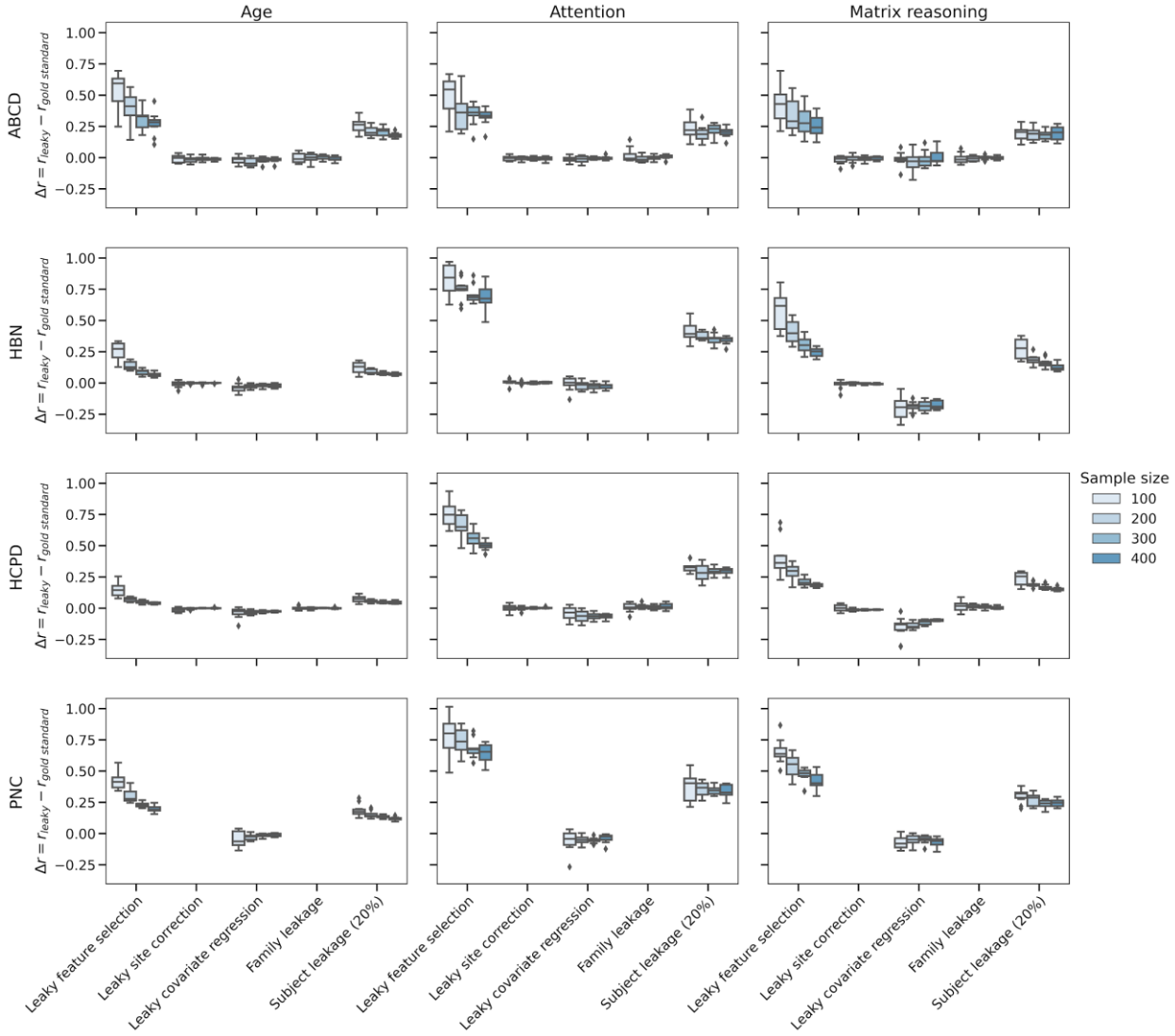

Figure S6. Difference in median performances across 10 random seeds of 5-fold cross-validation, related to Figure 8. Rows represent the dataset, and columns show the phenotype. For each leakage type (x-axis), there are four results (N=100, 200, 300, 400). For each sample size, we repeated 10 random seeds of resampling for 10 iterations of 5-fold cross-validation. The boxplots show the range of performance of the median across iterations of 5-fold cross-validation for a given subsample.

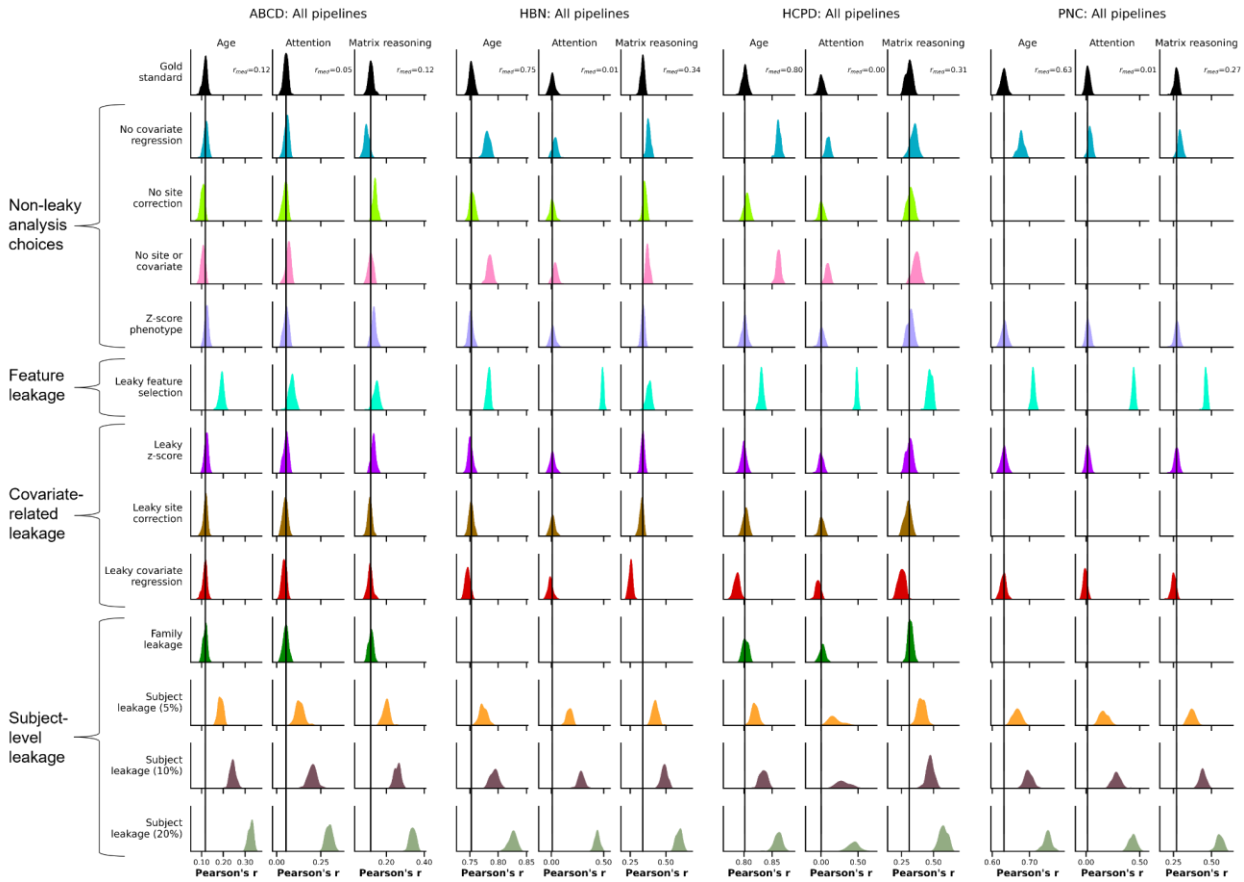

Figure S7. Evaluation of leakage types with *r* and SVR (ABCD: left, HBN: center left, HCPD: center right, PNC: right). The rows show different leakage types, and, within each dataset, the columns show different phenotypes. The black bar represents the median performance of the "gold standard" models across random iterations, and the histograms show prediction performance across 100 iterations of 5-fold cross-validation.

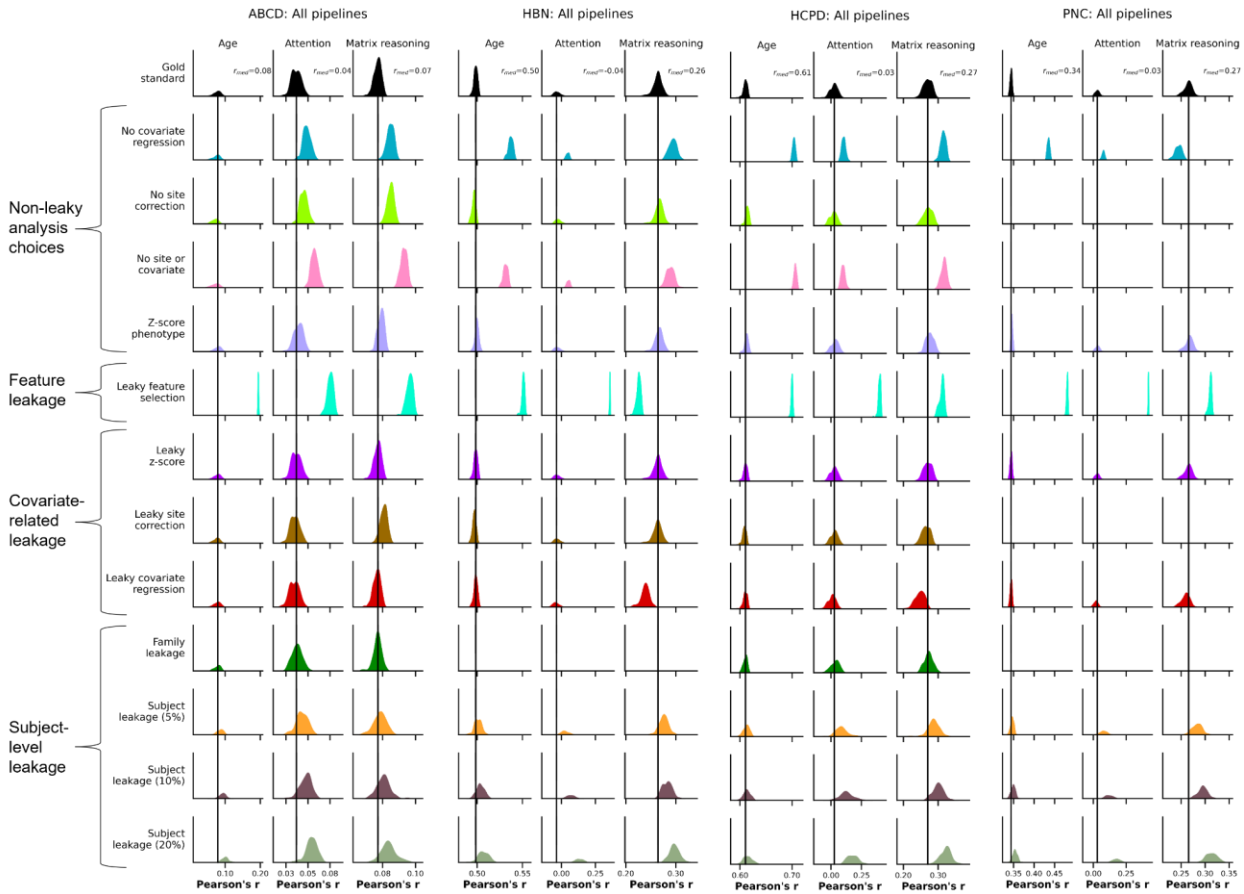

Figure S8. Evaluation of leakage types with *r* and CPM (ABCD: left, HBN: center left, HCPD: center right, PNC: right). The rows show different leakage types, and, within each dataset, the columns show different phenotypes. The black bar represents the median performance of the "gold standard" models across random iterations, and the histograms show prediction performance across 100 iterations of 5-fold cross-validation.

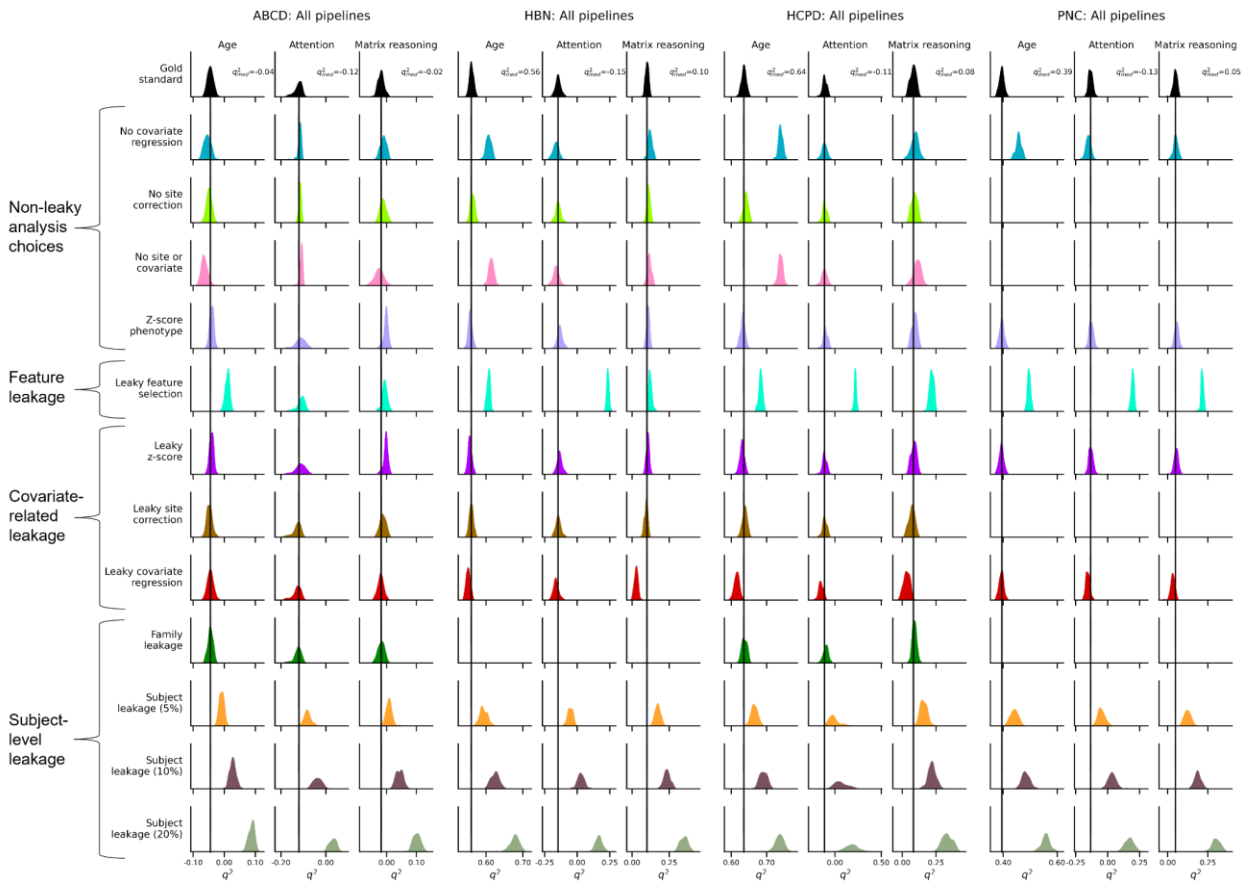

Figure S9. Evaluation of leakage types with q^2 and SVR (ABCD: left, HBN: center left, HCPD: center right, PNC: right). The rows show different leakage types, and, within each dataset, the columns show different phenotypes. The black bar represents the median performance of the "gold standard" models across random iterations, and the histograms show prediction performance across 100 iterations of 5-fold cross-validation.

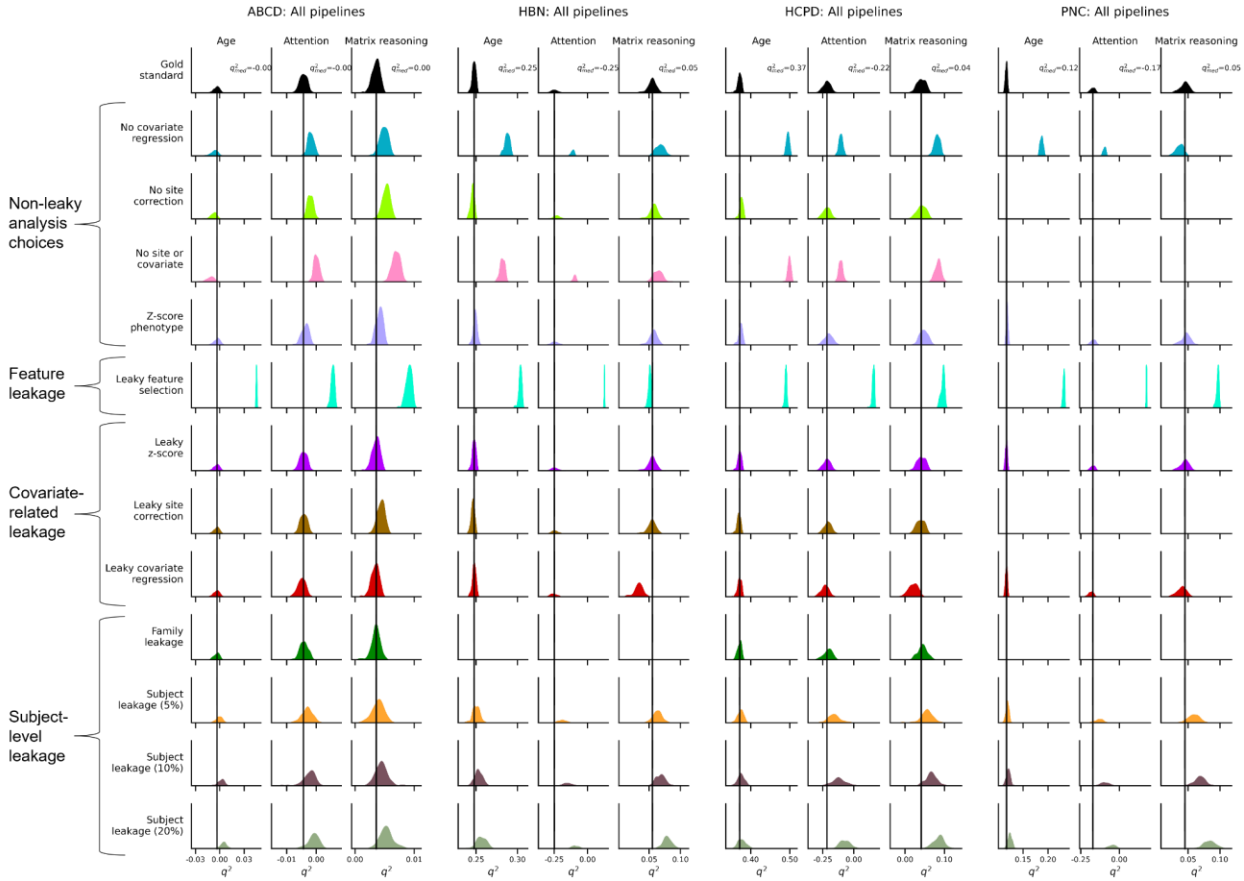

Figure S10. Evaluation of leakage types with q^2 and CPM (ABCD: left, HBN: center left, HCPD: center right, PNC: right). The rows show different leakage types, and, within each dataset, the columns show different phenotypes. The black bar represents the median performance of the "gold standard" models across random iterations, and the histograms show prediction performance across 100 iterations of 5-fold cross-validation.

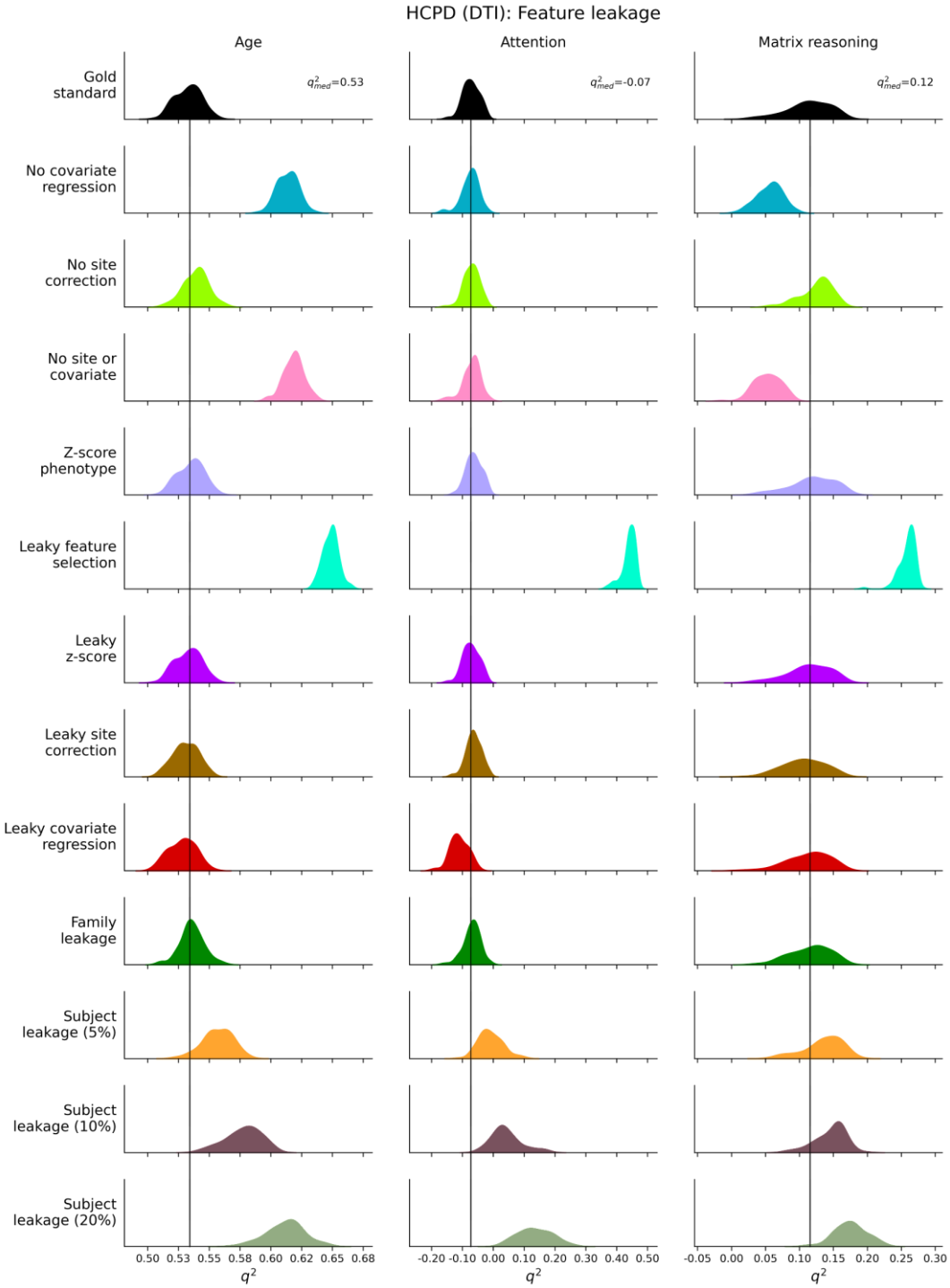

Figure S11. Evaluation of leakage types for matrix reasoning, attention, and age prediction in structural connectomes using q^2, related to Figure 9. The rows show different leakage types, and the columns show different phenotypes. The black bar represents the median performance of the "gold standard" models across random iterations, and the histograms show prediction performance across 100 iterations of 5-fold cross-validation.