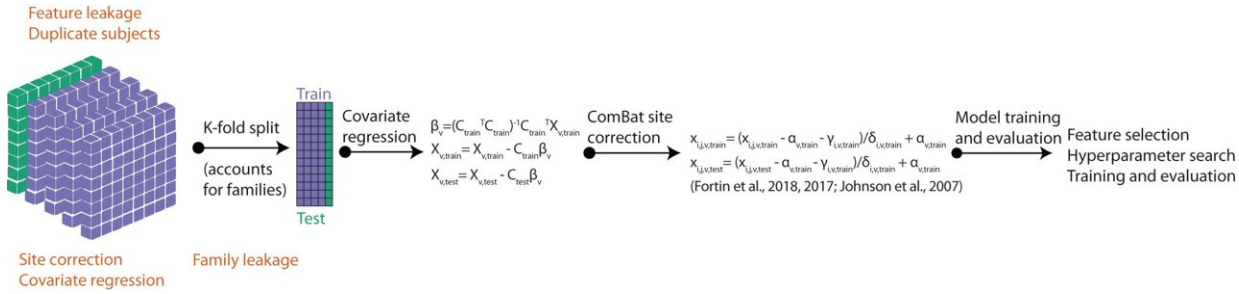
Data leaks can sink machine learning models

February 28 2024, by Mallory Locklear



Summary of the prediction pipelines used in this study. The various forms of leakage that may occur are shown in orange. 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. Credit: *Nature Communications* (2024). DOI: 10.1038/s41467-024-46150-w

When developing machine learning models to find patterns in data, researchers across fields typically use separate data sets for model training and testing, which allows them to measure how well their trained models do with new, unseen data. But, due to human error, that line sometimes is inadvertently blurred and data used to test how well the model performs bleeds into data used to train it.

In a new study, Yale researchers have assessed how data leakage affects the performance of neuroimaging-based models in particular, finding it can both artificially inflate or flatten results.

The study was published Feb. 28 in *Nature Communications*.

Biomedical researchers are evaluating the use of machine learning for all sorts of tasks, from diagnosing illnesses to identifying molecules that could become treatments for disease. In the field of neuroscience, scientists are using machine learning to better understand the relationship between brain and behavior.

To train a model to predict, for example, a person's age based on functional neuroimaging data, researchers provide the model with fMRI data and the ages of the individuals scanned. The model will then begin to associate patterns in the fMRI data with age and if those patterns are strong enough, the model should be able to predict an individual's age from new neuroimaging data it has not yet seen.

When data leakage occurs, part of that "unseen" data has indeed already been seen by the model in some way during the training phase, meaning researchers can't be sure if the model's predictions are really predictions or simply recognition of information it has already analyzed.

Researchers widely acknowledge that data leakage should be avoided, but it happens often, said Dustin Scheinost, an associate professor of radiology and biomedical imaging at Yale School of Medicine and senior author of the study.

"Leaking data is surprisingly easy to do," he said. "And there are a number of ways it can happen."

To better understand how data leakage affects machine learning performance, the researchers first trained a machine learning model using fMRI data not affected by leakage and then tested how well the model could predict age, an individual's ability to perform a type of problem solving known as matrix reasoning, and attention problems

from unseen neuroimaging data. They then introduced different types of leakage into training data and compared the model's predictions to those based on untainted training data.

Two types of leakage drastically inflated the model's prediction performance, the researchers found. The first, known as "feature selection" leakage, occurs when researchers select brain areas of interest from the entire pool of data rather than from the training data only. In the second, called "repeated subject" leakage, data from an individual appears in both the training and testing sets.

"One of our findings was that feature selection leakage inflated the model's predictions for attention problems," said Matthew Rosenblatt, a graduate student in Scheinost's lab and lead author of the study. "With feature leakage, the model's predictions were strong, producing what would be a significant result. But in reality, without data leakage, prediction performance is poor for attention problems."

That kind of false inflation can make it look like the model is performing well when in fact it may not be able to predict much at all with truly unseen data, which could affect how researchers interpret models and reduce the ability for other researchers to replicate published findings that are based on the model.

After introducing another type of leakage in which statistical analyses are performed across the entire data set rather than just the training data, the researchers found it artificially weakened the model's performance.

Leakage effects were also more variable, and therefore more unpredictable, in smaller sample sizes compared to larger datasets.

"And effects are not limited to model performance," said Rosenblatt. "A lot of times we look at our models to get some neurobiological

interpretation and data leakages can also affect that, which is important in terms of trying to establish brain-behavior relationships."

While not every type of leakage strongly affected the model's performance, the researchers say avoiding leakage of all sorts is the best practice. Sharing programming code is one way to prevent mishaps, as others can see if leakage may have inadvertently happened. Using well-established coding packages is another route, which could help prevent errors that may arise when writing code from scratch. Additionally, there are worksheets available that prompt researchers to reflect on potential problem areas.

"Having healthy skepticism about your results is key as well," said Rosenblatt. "If you see something that looks off, it's good to double check your results and try to validate them in another way."

More information: Matthew Rosenblatt et al, Data leakage inflates prediction performance in connectome-based machine learning models, *Nature Communications* (2024). DOI: 10.1038/s41467-024-46150-w

Provided by Yale University