**Supervised machine learning** gradually learns a function that maps an input (e.g., speech features) to a known output (e.g., depression score from PHQ-9). It is a classification problem when the outputs are categorical (e.g., +PTSD or -PTSD) and a regression problem when outputs are continuous values (e.g., PTSD Checklist score). Machine learning is particularly useful in finding structure in big, complex, and multidimensional data, where humans cannot visualize the structure on their own. Certain algorithms learn by repeatedly seeing examples of inputs and outputs; the algorithm first predicts the output of an input, then sees the label, and over iterations corrects its configuration if mistaken in such a way that it would have guessed correctly. These labels can be obtained from clinical diagnosis or self-assessments. *Feeding a bad label will create a bad model.* This creates an open problem to be evaluated when gathering data for a given disorder: clinical diagnoses tend to be considered the gold-standard but certain disorders may have a low inter-rater reliability [17] and be more episodic and thus self-assessments may be preferred. Finally, model performance is measured by learning on a training set of samples (sets of inputs and outputs) and testing on new samples not used for learning.

**Unsupervised machine learning** is used when the labels (e.g., depression, control) for each data point (e.g., speech features of a single participant) are unknown. Therefore, unsupervised algorithms find structure in the data by, for example, clustering similar data points together. This can be useful for observing which features relate samples to each other. Moreover, such unsupervised approaches (e.g., PCA, tSNE, UMAP) may help uncover participants with similar symptoms and on that basis may reveal disorder subtypes.

**Testing performance.** Data can be split into a training and test set (e.g., splitting the data into a random 80-20% split), and the training set can also be split into a training and development set. This way, different models and configurations could be tried on the development to choose the one that increases performance to later evaluate on the test set not used in training. However, an issue in small datasets, which are common in the medical field, is that these random subsets (both the development or the test sets) may not be representative of the general population. Therefore the performance on these small subsets cannot be expected to generalize to the general population. A better estimate than using a single, small subset is taking multiple subsamples through k-fold cross validation, which is splitting the training data into k segments (e.g., k=10) and iteratively training the model on k-1 segments and validating on the left-out segment and then averaging performance. When there are very few participants (e.g., $n < 40$), leave-one-out cross validation is often used, which maximizes the amount of data seen during training at the cost of increasing the variance of the predictive model (see Figure 5 for an alternative).

**Overfitting.** When studies only report performance on small development sets, it is likely their models will not perform as well on unseen data because they will likely *overfit*. Overfitting consists in learning model configurations that increase performance on the development set or folds without being able to generalize performance to an unseen test set, which is the main purpose of building a predictive model. In the figure below, the data distribution is generated by a line with some noise. The polynomial model (orange) can learn to fit the input samples perfectly; however, it is more likely that the linear model (black) will make better predictions on future unseen samples. In other words, overfitting is finding an illusory pattern in the training set that does not exist in the general population. This is why larger, representative samples are needed with held-out test sets. It is also why testing should be done only once; otherwise trying different model configurations will overfit to the test set as well. For further reading, see [34,119].