Question 3:

In this article by Wang et el, data doppelgangers are discussed as potential confounders of machine learning validation in the biomedical field. Data doppelgangers are independently derived data with very similar characteristics. Due to their similarities, when each of the doppelgangers are placed in the training and validation sets respectively, they can potentially inflate the performance of the machine learning model because the model will perform well on the similar data regardless of whether the model has been properly trained or not using relevant features. This is coined the doppelganger effect. This effect can give us a false sense of confidence in the accuracy of a model, which can lead to wasted time and resources when we try to apply the model to real-world data for our research because we might find that the predictions are actually wildly inaccurate since the model might not have been suitably trained to identify key features at all. As a result, it is unable to handle the diversity and complexity of real-world data. In the article, proteomics data from renal cell carcinoma was used to demonstrate the doppelganger effect. Using pairwise Pearson's correlation coefficients (PPCC), it was shown that a portion of the data from different patients with the same class (potential valid data doppelganger cases) had high correlation coefficients that exceeded the maximum PPCC of the data from different patients and different classes (negative cases). In fact, the PPCC of these samples neared the values of the positive cases, where data came from the same patient and the same class (i.e technical duplicates). This suggests that these samples are likely to be data doppelgangers and confirms they do exist in biomedical data. Subsequently, these samples were used in different machine learning models in increasing numbers to examine the functional effects of these data doppelgangers on the validation of the machine learning models. It was found that even when using randomly selected features, which should give around ~0.5 accuracy since it was randomly trained, when the number of data doppelgangers in the training and validation sets increased, the accuracy of the model also increased correspondingly. This shows how the performance of a model can be falsely inflated by the presence of data doppelgangers. Interestingly, not all models show a linear relationship between the number of data doppelganger and the model accuracy. This might suggest that some models are more insensitive to the effects of data doppelgangers than others. Regardless, the doppelganger effect can be seen.

Biomedical data is a source of data doppelgangers because the fact is that similar biological systems inevitably contain similar structures and gene/protein expression profiles, especially in the same disease states. This makes it likely for data derived independently (like from different patients) to have high correlation profiles. For example, in another article by Wang et al, they found that data doppelgangers and specifically functional doppelgangers also exist in the Duchenne Muscular Dystrophy (DMD) Microarray DataSet and the Leukemia Microarray DataSet.[1] This shows that data doppelgangers are common across different biomedical datasets, though the extent to which these doppelgangers are present and the extent of their effects on the validation of models may differ. The doppelganger effect is unlikely to be limited to just biomedical data. Just looking at how data in a population is commonly distributed naturally– approximating to a normal distribution or at least a distribution with some kind of peak – we can see how it is possible for independent samples taken from that population to have similar characteristics since some characteristics are just more common. For example, we can imagine that in other datasets like a consumer behavior dataset, consumer profiles taken from different individuals who made the same spending decisions are likely to be correlated, potentially qualifying as data doppelgangers. However, the extent to which these doppelgangers would affect the validation of a machine learning model of consumer behavior prediction requires further study. It should also be noted that if a population consists of samples with largely similar traits, the doppelganger effect would not be such a big problem because the accuracy of the model as validated on doppelgangers would apply to the real-world population too. Of course, it is unlikely for real-world datasets of interest to be so clean cut.

In the article, several solutions were proposed to ameliorate the doppelganger effect. First, careful identification of potential data and functional doppelgangers was recommended before splitting the data for training and validation. This can be done by the PPCC method and testing out potential data doppelgangers in different machine learning models as performed here. After identification, one can classify all data doppelgangers in either the training or validation sets, thereby eliminating the doppelganger effect. However, as noted, this would likely come at the expense of training variety since at extreme ends, we could end up with training data that consists of mainly data doppelgangers or no representation of the data doppelgangers at all. Still, this could be a reasonable solution in cases where the sample size is small and doppelgangers cannot be removed. If the sample size is sufficiently large and the proportion of doppelgangers present is small, perhaps one could also consider just choosing a few samples that have the most doppelgangers to represent this type of data in the training set and remove any doppelgangers from the validation set to avoid any confounding effects. This will allow for variety in the training set while avoiding the doppelganger effect. Secondly, stratifying the validation data into PPCC data doppelgangers and non-PPCC data doppelgangers was also recommended to examine the performance of the model on similar and dissimilar datasets. If these strata coincide with clinically relevant strata, valuable data can be gained on the strengths and weaknesses of the model in different settings. This can be useful in showing us the subset of population that we might still be able to use the model on even if it doesn’t test well on non-PPCC data doppelgangers. It also more importantly shows us the gaps in the model and what we need to improve on. Thirdly, robust independent validation checks involving many data sets were also recommended. This can help to dilute the effects of any doppelgangers present. Lastly, the authors also proposed that we should find ways to identify functional doppelgangers experimentally by finding datasets that test well regardless of what machine learning models are used. This will reduce reliance on metadata to identify doppelgangers but may require a lot of additional work to develop varied models for testing.

In summary, data doppelgangers are a real occurrence in biomedical data and they can have inflationary effects on the evaluated performance of a machine learning model. It is important for scientists to be cognizant of this fact and actively try to identify these doppelgangers before splitting data into training-validation sets and find ways to mitigate the effect so that we can have greater confidence in the accuracy of our model.

Bibliography

1. Wang LR, Choy XY, Goh WW Bin. Doppelgänger spotting in biomedical gene expression data. *iScience*. 2022;25(8):104788. doi:10.1016/j.isci.2022.104788