Question 3

Doppelganger effects are not unique to biomedical data, it can also appear in machine learning tasks performed on other data. Especially when the data set is too small or the sampling range of the data is too small so that the data feature values are too similar, doppelganger effects will be very likely to appear.

This article gives three suggestions for avoiding the doppelganger effect in machine learning tasks, cross-checks, data stratification and independent validation checks. In addition, I think better perform data collection and data preprocessing can also effectively avoid doppelganger effect. For example, if you can cover more subjects and ensure the breadth of the data when collecting the data, then the distribution of the final data will be large enough. This will lead to better results in the subsequent division of the training set, validation set, and test set. Each data set has its own characteristic value relatively, and there will not be too many repetitions.