Q1: Whether you think doppelganger effects are unique to biomedical data?
A1: I don't think so. I think this effect can also be applied to other fields, such as the financial field. If someone wants to predict some funds' future performances, instead of trying to map a fund's performance onto a generic financial trajectory curve, it would be better to find the past funds which were statistically most similar to the fund in question. These similar funds could be called doppelgangers. What's more, when we want to classify known funds into different risk classes, if the training and validation sets are highly similar because of chance or otherwise, data doppelganger will occur and affect the assessment of the performance of the classifier. It is generally believed that the risk levels of funds with the same type of holdings may be similar. However, compared with past eras, the current market development direction has changed, so the robustness of classifiers with data doppelgangers in different eras is not high.

Q2: How you think it can be avoided in the practice and development of machine learning models for health and medical science?
A2: From the paper attached, the first method that was mentioned is to perform data stratification. Instead of evaluating model performance on whole test data, we can stratify data into strata of different similarities (e.g., PPCC data doppelgängers and non-PPCC data doppelgängers, and evaluate model performance on each stratum separately). The second one is to perform extremely robust independent validation checks involving as many data sets as possible (divergent validation). Although not a direct hedge against data doppelgängers, divergent validation techniques can inform on the objectivity of the classifier. It also informs on the generalizability of the model (in terms of real-world usage) despite the possible presence of data doppelgängers in the training set.