From my point of view, doppelganger effects are not unique to biomedical data. Firstly, it is necessary to know data doppelgangers. Data doppelgangers refers to the data where training and validation sets are highly similar because of chance or otherwise. Following doppelganger effect means when a classifier falsely performs well because of the presence of data doppelgangers. It is important to point out that not all data doppelgangers generate doppelganger effects. Thus, data doppelgangers that also generate a doppelganger effect (confounding ML out comes) are termed functional doppelgangers. Since the root of doppelganger exists in the highly similarities of data in both training and validation sets, it is clear that doppelganger relates to the attributes of data sets. However, this kind of attribute exists not only in biomedical data which means doppelganger effects are not unique to biomedical data.

In the original paper, the author gives out some recommendations about how to avoid doppelganger effects. The first one is to perform careful cross-checks using meta-data as a guide. With information from meta-data, we are able to identify potential doppelgangers and assort them all into either training or validation sets, effectively preventing doppelganger effects and allowing a relatively more objective evaluation of ML performance. The second recommendation is to perform data stratification. The key point of the recommendation is to stratify data into strata of different similarities. With a known proportion of similarities, it is still possible to evaluate the real performance of ML models. The third recommendation is to perform extremely robust independently validation checks involving as data sets as possible. It can inform on the objectivity of the classifier. It also informs on the generalizability of the model despite the possible presence of data doppelgängers in the training set.