The paper talks about the doppelgänger effect, specifically the functional doppelgänger. Functional doppelgängers are data doppelgängers that can generate a doppelgänger effect. In detail, this paper describes the abundance of data doppelgängers in biological data and explains the identification, confounding effects, and ways to ameliorate data doppelgängers through the application of PPCC on the RCC data set.

In fact, I don't think doppelganger effects are unique to biomedical data. Because in many areas of machine learning, there are a lot of independent but similar data, they will also interfere with the training of machine learning models. For example, in the field of face recognition, an untrained machine learning model may recognize a face photo as a real person's face. However, data doppelgängers in biomedical data do need to be paid attention to. Because through the detailed introduction of this paper, we can know that in the field of biomedicine, the generation of these data doppelgängers is very complicated and large in number.

In this paper, many methods are introduced to avoid doppelganger effects in the practice and development of machine learning models for health and medical science. Specifically，this paper has proved that the presence of PPCC data doppelgängers in both training and validation data inflates ML performance, even if the features are randomly selected. Therefore this paper comes up with two ways to avoid the doppelgänger effect, both of which have disadvantages. One is placing all PPCC data doppelgängers in the training set. In this case, when the size of training set is fixed (thus, each data doppelganger that gets included causes a less similar sample to be excluded from the training set), it leads to models that might not generalize well because the model lacks knowledge, though the doppelgänger effect is eliminated. The other is constraining the PPCC data doppelgängers to either the training or validation set. But you might end up with spectacular winner-takes-all scenarios (the doppelgängers will all either be predicted correctly or wrongly.)

So this paper proposes other attempts. For example, removing PPCC data doppelgängers when doppelgangR was used for the identification of doppelgängers or alleviating doppelgänger effect with methods that would not lead to a significant reduction in sample size or require a high amount of contextual data. But they are not very successful or only applicable in a specific situation. Besides, this paper proposes three recommendations. The first is to perform careful cross-checks using meta-data as a guide. The second is to perform data stratification. The third is to perform extremely robust independent validation checks involving as many data sets as possible.