Report

By reading the following papers, I first make my point: I do not believe that the doppelganger effect is unique to biomedical data. The doppelganger effect can simply be understood as duplicate records or duplicate data. It can happen in any large database or dataset, so I don't think it is necessary to consider the application domain.

Medical data is not collected from one hospital or party, it is often collected from multiple sources, such as different hospitals or clinics. These data are then combined into a large dataset for computational biologists to work on data analysis. The collection of data must be accompanied by the important dependent variable of temporal extrapolation, where the same person's identity is recorded multiple times, the person's medication is recorded multiple times, and so on.

There are many ways in which this can be expressed in biomedical molecular data. Some examples include:

Repeated genes or protein sequences. In molecular biology, the same gene or protein sequence can be identified and recorded multiple times, resulting in duplicate entries in the database. This can lead to inaccurate mathematical analysis by machine learning or analysis by classifiers, and make it difficult to identify unique genes or proteins.

Duplicate chemical structures. In some drug compositions, the same compound will be synthesized by different chemical components under different temporal conditions, which again can lead to duplicate entries in the database.

Whether it is a biochemical molecule at the micro level or a sample or patient at the macro level, it is inevitable that the dupe effect will lead to inaccurate data analysis. This can have a significant impact on the quality and integrity of biomedical data, leading to a range of uncontrollable problems occurring. It is therefore imperative that researchers or operators take the necessary steps to identify and remove duplicate records to ensure accurate and reliable data analysis.

In terms of how to effectively avoid the doppelganger effect in the practical and development of health and medical machine learning models, I feel that this can be attempted through the following approaches.

Data integration: The main reason for the doppelganger effect to occur stems from too much duplicate data. One of the most important clear strategies for parties to merge data from multiple sources is to use unique identifiers and thus link and represent records from different sources, thus minimizing the risk of duplicate records.

Data cleansing and pre-processing: We need to find ways to thoroughly clean and pre-process data to better identify and remove duplicate records from large data sets. Pre-processing can be done by referring to the unique identifiers (e.g. patient ID numbers) mentioned in the first scenario.

Data quality monitoring: continuously monitor the quality of important data in the dataset and periodically review the data for duplicate records so that duplicate records can be technically addressed.

Use of unique identifiers: e.g. patient ID numbers, numbering of unique gene sequences, etc.

In general, the above approaches need to be used in combination so as to avoid the doppelganger effect that occurs in the development and event of machine learning models.