Report for <How doppelganger effects in biomedical data confound machine learning>

(Name: Yaqi Qian)

The doppelganger effect in biomedical data refers to the phenomenon where two or more individuals have very similar or identical features in a dataset, such as their age, gender, and medical history. This can cause problems for machine learning algorithms, as they may struggle to distinguish between the different individuals and may produce inaccurate or biased results.

The doppelganger effect biased training data, machine learning models may generate erroneous or biased results if the training data used to construct the model was skewed. This may happen if the training set contains doppelganger records—records that are similar to the records of the real persons in the dataset but are not an exact match—or if the training set is not representative of the population it is meant to serve.

The doppelganger effect can also lead to overfitting, which happens when an algorithm has been trained too closely on a particular set of data and is unable to generalize effectively to new data. This can cause problems for machine learning. Overfitting can cause the algorithm to be overly specialized to the characteristics of the particular individuals in the training dataset, which can lead to poor performance on unknown data as well as biased findings.

In addition, the class imbalance is also caused by the doppelganger effect, when there are disproportionately more examples of one class (e. g. people with a certain medical condition) than another, can also have an impact on machine learning. As a result, the algorithm may give the more prevalent class higher priority, which might result in incorrect predictions for the minority class. What's more, doppelganger effects can also lead to misclassification, where the machine learning model incorrectly classifies individuals or records. This can occur if the model is not able to distinguish between similar but distinct records, or if it is overly influenced by biases present in the training data.

Moreover, doppelganger effects emerge in biomedical data because of following reasons: First is the human error, as with other types of data, human error can result in duplicate records or incorrect data being entered into biomedical databases. Second is the system errors, duplicate records or inaccurate data being captured are only two examples of the doppelganger effects that can result from glitches or flaws in electronic medical record systems. Third is the data entry errors, when data is transmitted across systems, such as when a paper record is typed into an electronic medical record, there can occasionally be doppelganger effects. The data may therefore have errors or duplicate entries as a result of this. Fourth is multiple identifiers, due to receiving care at many hospitals or clinics, patients can have numerous identifiers (such as multiple medical record numbers). Doppelganger effects might result from the same patient having several records created for them using different identifiers.

As for whether the doppelganger effect is unique to biomedical data, from my point of view, I think the doppelganger effect is not unique to biomedical data. Because the doppelganger effect, or the phenomenon of having very similar or identical examples in a dataset, which can occur in any type of dataset where there are multiple examples with similar or identical features.

For example, if there are numerous clients with the same name, age, and address in a collection of customer data for a retail business, the doppelganger effect could take place. If numerous individuals have the same name and profile image, the doppelganger effect may appear in a collection of social media accounts. Or in a financial system, a doppelganger effect could occur if a transaction is recorded twice, resulting in an incorrect balance or overpayment.

However, there are several ways to avoid doppelganger effects in the practice and development of machine learning models for health and medical science.

First, use a diverse dataset, making sure that the dataset used to train the machine learning model is varied and appropriate for the target audience is crucial. Due to the model's exposure to a wide range of input data and less likelihood of producing biased or erroneous findings, this can help lower the probability of doppelganger effects.

Second, use data from multiple sources, using data from multiple sources can also

help reduce the risk of doppelganger effects, as it allows the model to learn from a wider range of data and can help mitigate biases that may be present in any single source of data. Also, data quality checks can be utilized, which could help identify errors or inconsistencies in the data. For instance, researchers might look for data anomalies like strange or improbable results or conflicts across several data sources.

Third, use data linking techniques and centralized database. Data linking techniques like probabilistic or deterministic matching can be used to find records from the same person in many data sources. This can assist in locating and fixing any doppelganger errors. A centralized database can help avoid doppelganger effects by providing a single source of truth for participant data. This can make it easier to identify and correct errors or inconsistencies in the data.

In addition, monitor model performance can also help avoid doppelganger effects in an effective way. Monitoring the machine learning model's performance regularly can aid in finding any potential doppelganger effects. To make sure the model is delivering accurate and objective results, this might entail employing several performance measures and assessing the model using diverse datasets.

Moreover, use ethical considerations in model development is another approach. When creating machine learning models for use in health and medical research, it's crucial to take ethical issues into account. This may entail making sure the model is applied responsibly and transparently, and that it is not biased against any particular groups.

Last but not least, incorporating domain knowledge into the development and use of the machine learning model can help reduce the risk of doppelganger effects. This can involve working with experts in the field to understand the specific characteristics and needs of the population the model will serve.

In summary, doppelganger effects can have serious consequences in biomedical research, as they can lead to inaccurate or biased results. To avoid doppelganger effects in biomedical data, it is important to use unique identifiers for each participant, confirm the identity of participants using multiple methods, and implement data quality checks to identify and correct errors.

References:

[1] Li W, et al. Doppelganger spotting in biomedical gene expression data. *STAR Protocols*, 2022, vol 3.

[2] Waldron, L. The Doppelganger effect: Hidden duplicates in databases of transcriptome profiles. *JNCI: Journal of the National Cancer Institute*, 2016.

[3] Xue H, et al. Data considerations for predictive modeling applied to the discovery of bioactive natural products. *Drug Discovery Today*, 2022.