Typically, the performance of a machine learning model is evaluated by validating the model's accuracy using validation data. And only if validation data are independent of training data is the model valid. However, training and validation data may not be independent in the presence of doppelganger effects; hence, it is essential to check for potential doppelgangers in data before to training and validation data assortment (Wang et al., 2022).

The doppelganger phenomenon is not exclusive to biological data; in fact, it occurs in other academic disciplines of study. Take finance cases as an example: certain investment professionals utilise machine learning while assessing the market trend for the upcoming week, and the training data they employ is frequently indicative of historical economic data. The current data of the same stock is likely to have a high degree of resemblance to the historical data. Consequently, there may be a large degree of overlap between the training data utilised by experts and the validation data. This high likelihood of duplicates may render the validation inaccurate and result in a biased forecast.

In addition to the biological and financial areas, the doppelganger effect occurs most frequently in the domains of biostatistics and economics when examining vast quantities of data. Consider EQ-5D, a research project in which I participated. After ensuring that there is a proper relationship between EQ-5D data and economic worth, designers will pick training data for machine learning to utilise. The training data and validation data may, however, have significant doppelganger effects; several distinct research may utilise the same collection of experimental data, resulting in publications with nearly identical statistics. n addition, the EQ-5D questionnaire comprises various interrelated variables – the EQ-VAS is derived from the EQ-5D index. So, if training data and validation data contain the same set of EQ-5d index and EQ-VAS, they are also duplicates, despite the fact that their underlying data are distinct (Figure 1). More comparable doppelganger scenarios might potentially occur from similar medicine classes, similar illness situations (Figure 2), and similar patient demographics, hence bringing substantial ambiguity into the overall research outcomes.

5	602	Coulton-2017		Y	6 month,12 month		utility
6	603	Covelli-2005		Y	12 months		EQ VAS, domain
7	604	Cowger-2018		Y	6 month		sum of response
8	607	Cox-2010		Y	4 weeks,12 weeks		utility
9	609	Cramer-2018		Y	Cycle 1, 3, 5, 7, 10, end of treatment, every 1		utility, EQ VAS
10	610	Crawford-2021		Y	1 month, 3 month		utility
11	611	Crawford-2021		Y	1 month, 3 month		utility
12	612	Creamer-2018		Y	3 month, 6 month		utility, EQ VAS
13	614	Criner-2019		Y	180 days		utility
14	618	Cross-2010	62 was used to est	Y	6 weeks, 6 months		utility, EQ VAS

Figure 1 EQ-5D data used in different papers*. This figure contains number of paper (the first column) and its corresponding EQ-5D data type (the last column) used in paper.*

	ENDNOTE ID	Study name ("First author last name"-"Publication Year". e.g. Xie-2021.)		Indication/ Disease/Diagnosis
5	602	Coulton-2017		Older Alcohol Use
6	603	Covelli-2005		chronic obstructive pulmonary disease (COPD)
7	604	Cowger-2018		left ventricular assist device
8	607	Cox-2010		chronic low back pain
9	609	Cramer-2018		long course of disease
10	610	Crawford-2021		partial and total knee arthroplasty
11	611	Crawford-2021		total hip arthroplasty
12	612	Creamer-2018	4	Poststroke Spasticity
13	614	Criner-2019		chronic obstructive pulmonary disease (COPD)
14	618	Cross-2010		chronic obstructive pulmonary disease

Figure 2 the same disease types studied in different papers*. This figure contains number of paper (the first column) and corresponding indication/disease/diagnosis (the second column) it focused on. Within this figure, COPD has been mentioned 3 times, indicating a high possibility of data overlapping and data doppelganger.*

To resolve this problem in the EQ-5D research, designers should implement the three recommendations suggested in the 2022 study by Wang et al (Wang et al., 2022). First, cross-check the data using meta-data and re-filter the dataset with reference to PPCC values, ensuring that training and test samples are not supplicants or samples with a high degree of resemblance. Second, stratify the data by classifying data, people, and objects into discrete groups or layers, reducing the usage of repeated data, and establishing a biased model. Third, conduct extremely strong and diverse validation tests on as many datasets as feasible, therefore further reducing the prevalence of data duplication in the training set. These three procedures may result in more precise and organised machine learning datasets. As shown by the study (Wang et al., 2022), removing all doppelganger data redundancies is difficult, but we can at least eliminate positive cases and data leakage. As researchers continuously construct more accurate doppelgangerIdentifiers (Wang et al., 2022), we may have the opportunity to receive the most optimum training and validation data set in the near future.