Report of "How doppelagänger effects in biomedical data confound machine learning"

January 15, 2023

1 Introduction

The doppelagänger effect, which refers to duplicate expression profiles in public databases, will affect re-analysis if they are not found. It is most commonly seen in biomedical imaging. However, they can also occur in other areas, such as facial recognition technologies, remote sensing data and text analysis.
To expedite medication development, machine learning (ML) models are being employed more and more in drug discovery. ML improves the effectiveness of drug discovery in a variety of ways, including: ML models can shortlist better drug candidates (targets) more quickly, cutting down on the amount of time needed for research and testing. The efficacy of medication development is also increased by classification models based on ML and AI. Classifiers have been used to forecast potential adverse medication responses and new drug-disease interactions. It is crucial that these classififiers are appropriately trained and tested in order to discover suitable medication candidates given the pricey drug-testing procedure. Unreliable validation findings could nevertheless be obtained from training and test sets that were independently produced. There is an observed doppelganger effect when a classifier mistakenly outperforms itself due to the existence of data doppelagängers.

The circumstance in which a machine learning model performs well on a validation set regardless of how it was trained is known as the "doppelganger effect". Doppelagänger effect is troublesome since it may overstate the ML model's performance on actual data and may make model selection procedures that are only focused on validation accuracy more difficult. Therefore, prior model validation, ML practitioners must be aware of the existence of any doppelagängers. This article illustrate here the prevalence of functional doppelagängers given this biomedical data, the implications of data doppelagängers on ML, and ways to mitigate the doppelagänger effect.

2 Methods to avoid Doppelganger effects

2.1 Identification

2.1.1 Data doppelagängers

Given the potential of doppelagänger effects to confound, it is crucial to be able to identify the presence of data doppelagängers between training and validation sets before validation. There are two key definitions – data doppelagängers (DDs) and functional doppelagängers (FDs). DDs are sample pairs that exhibit very high mutual correlations or similarities. For example, we may use pairwise Pearson's correlation coefficient (PPCC) to identify DDs such that sample pairs with high PPCCs are also referred to as PPCC DDs. On the other hand, FDs are sample pairs that, when split across training and validation data, results in inflated ML performance, i.e., the ML will be accurate regardless of how it was trained (It can be assumed that such models have not truly "learnt").[WCG22] The pairwise Pearson's correlation coefficient (PPCC) can be used to identify potential functional doppelagängers as the basic design of PPCC as a quantitation measure is reasonable methodologically.

2.1.2 PPCC data doppelagängers

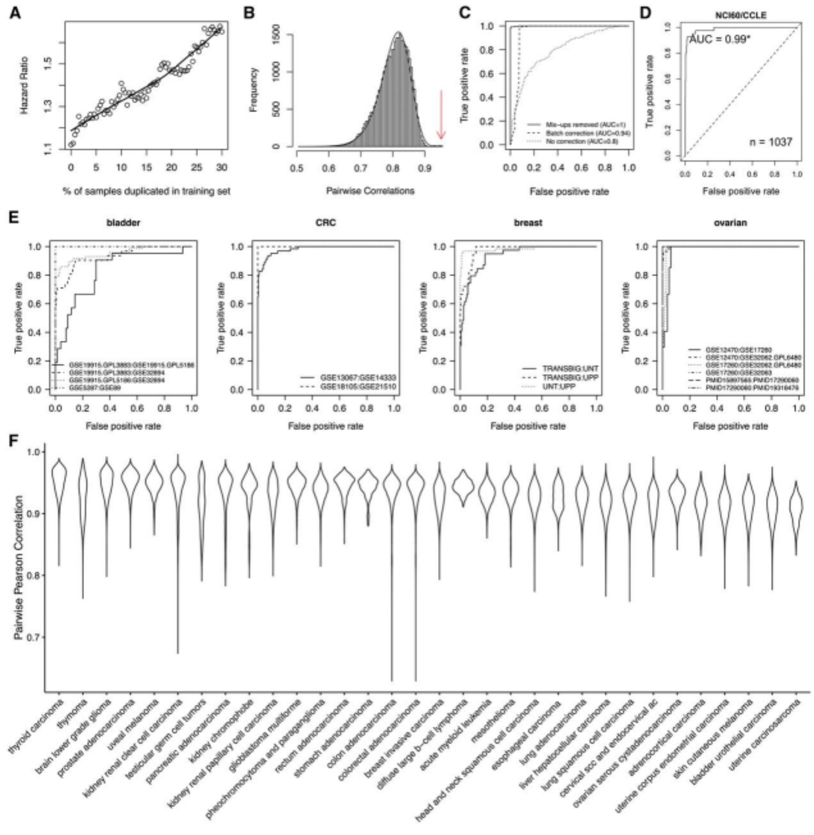
For each dataset pair, the background distribution of PPCC values must be approximated because it changes based on the tissue assessed and the technologies utilized. At the top end of the distribution of batch-corrected correlations, doubles can be recognized as outliers.

2.2 Avoidance of doppelagänger effect

The detailed methodology of package development and validation can be found in the supplement data. See from `https://academic.oup.com/jnci/article/108/11/djw146/2576926#supplementary-data`.

Our approach of duplicate identification reliably works when individual tumors have distinctive expression profiles, as is the case for cell line panels (Figure 1D) and for primary tumors from breast, ovarian, bladder, and colorectal cancers (Figure 1E). We expected it to be more prone to false positives for less differentiated expression profiles such as low-grade and early-stage tumors, and, generally observed, this where sufficient numbers ofnannotated samples were available: Samples falsely identified as duplicates were enriched for low-grade (CRC: 95% confidence interval [CI] = 1.2 to 2.2; ovarian: 95% CI= 1.0 to 1.5) and earlystage (bladder: 95% CI = 1.3 to 4.2; CRC: 95% CI = 1.6 to 2.6). The exception was early-stage ovarian cancer samples, for which doppelgangR was extremely effective. These samples have distinctive profiles, and their rate of sharing was high, possibly because of the rarity of early-stage ovarian cancer and the high importance of specimens.[WRR+16]

To avoid doppelganger effects, it is important to ensure that machine learning algorithms are developed in accordance with ethical and legal frameworks specific to the medical field. Additionally, building a large, diverse dataset of samples can help reduce bias and improve the accuracy of machine learning models.



3 conclusion

I think batch correction across datasets is critical. Analysis of duplicate samples is a serious concern because it may affect the development of specific gene signatures or the ability to identify subsets of patients with clinical differences. While this method can detect duplicate profiles even in the absence of germ-line sequences.

References

[WCG22] Li Rong Wang, Xin Yun Choy, and Wilson Wen Bin Goh. Doppelgänger spotting in biomedical gene expression data. *Iscience*, 25(8):104788, 2022.

[WRR$^+$16] Levi Waldron, Markus Riester, Marcel Ramos, Giovanni Parmigiani, and Michael Birrer. The doppelgänger effect: Hidden duplicates in databases of transcriptome profiles. *JNCI: Journal of the National Cancer Institute*, 108(11), 2016.