In this day and age of rapid information development, it is unavoidable that we are bombarded with a large amount of data and information, resulting in duplicate data within a redundant database. For example, when doing platform development, no single sign-on is used, or the same interface is requested twice or more times due to network issues or front-end lag. As a result, two or more identical data points appear in the background. As a result, data duality is not limited to biomedical data. It is difficult to remove the doppelganger effect directly from the data, but there are still options.

For example, the following three methods of testing for the presence of data duality are mentioned in an article published by Nanyang Technological University in Singapore.

1. Double-checking

Data doppelganger effects can be avoided by using metadata as a guide to perform rigorous cross-checks. For example, constructing negative and positive cases from renal cell carcinoma metadata can predict the absence of doppelgangers and the range of PPCC scores where leakage exists. To begin, doppelgangers from credible data come from samples of different patients in the same category, and we can use the identified potential doppelgangers as training and validation sets to effectively prevent doppelganger effects with the help of metadata information. Second, technical replication from the same sample can result in the existence of data doppelgangers

and should be treated similarly. That is, it should be ensured ahead of time that there is no duplication or excessive similarity between the training and test samples.

2. Data classification

The novelty of the second method is that, whereas previously we tested the performance of the evaluation model on the entire data set, method two divides the data into strata of varying degrees of similarity, reducing the data set's dichotomous effect to some extent. Using a world population as an example, and assuming that each stratum represents the same known proportion of the total world population, we can represent the performance of an entire stratum in the total world population by interpreting the performance of the sample in that stratum. A non-ppcc duplex tube, for example, is a papillary renal cell carcinoma sample in the renal cell carcinoma mentioned earlier. We can use the proportion of kidney cancer cells in each tissue (papillary renal cell carcinoma accounts for 10% of kidney cancer cells) to see if the extreme learning model effectively avoids data duality.

3. Validation by dispersion

The text's third approach is to perform as many independent validation checks as possible. Although this approach does not solve the fundamental problem of data duality, various validation techniques can demonstrate objectivity.

Functional doppelganger identification, which does not rely heavily on metadata, can

directly identify doppelganger cells. Finding a subset of the validation set that is a potential functional doppelganger of the training set, for example, will improve the model's accuracy regardless of how we train it.

We know that the doppelganger effect occurs when samples have chance similarity, causing the performance of trained machine learning models to be inflated when the training and validation sets are split. This inflationary effect leads to erroneous confidence in the model's deployability. As a result, there are no tools for split identification or standard practices for dealing with its confounding effects to date.

Cancer specimens are frequently subjected to genome-wide analysis, and researchers frequently share or reuse specimens in subsequent studies. Duplicate expression profiles in public databases, if undetected, will affect reanalysis, a phenomenon known as the "doppelganger" effect. So I went through the literature and discovered an article in the Journal of the National Cancer Institute that proposes a method that should become standard practice for accurately matching duplicate cancer transcriptomes when nucleotide level sequence data is unavailable, even for samples analyzed using different microarray technologies or microarray and RNA sequencing. The effectiveness of this method is demonstrated in a database containing dozens of datasets and thousands of microarray profiles for ovarian, breast, bladder, and colorectal cancers, as well as matching microarrays and RNA sequencing expression profiles from The Cancer Genome Atlas (TCGA). Over 50% of the studies had

potential replicates from different continents, using different technologies, published years apart, and even within the TCGA itself. Finally, the article includes the doppelgangR Bioconductor package, which can be used to search for duplicates in transcriptome databases. Given that unidentifiable repeats may inadvertently improve the predictive accuracy and confidence in differential expression, doppelgänger-checking should become standard procedure when combining multiple genomic datasets.

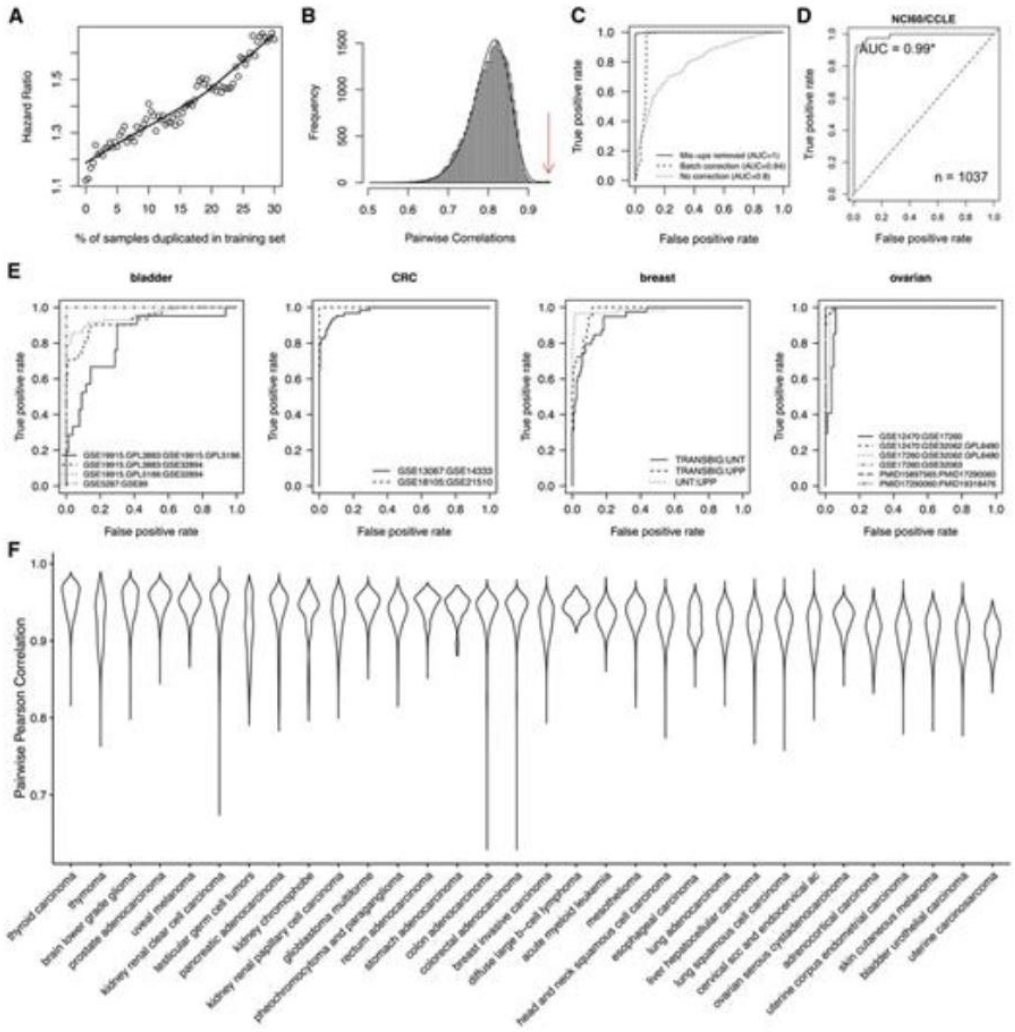

Table 1: Statistical Significance of Genomic Models

Reference

[1]Levi Waldron, Markus Riester, Marcel Ramos, Giovanni Parmigiani, Michael Birrer, The Doppelgänger Effect: Hidden Duplicates in Databases of Transcriptome Profiles, *JNCI: Journal of the National Cancer Institute*, Volume 108, Issue 11, November 2016, djw146, https://doi.org/10.1093/jnci/djw146

[2]Butler JM. Genetics and genomics of core short tandem repeat loci used in human identity testing. *J Forensic Sci*. 2006;51(2):253–265.

[3]Sotiriou C Wirapati P Loi S, et al. . Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *J Natl Cancer Inst.*2006;98(4):262–272.