The report about doppelganger effects

Doppelganger effects (DEs) occur when samples exhibit chance similarities such that, when split across training and validation sets, inflates the trained machine learning (ML) model performance. Duplicate expression profiles in public databases will impact re-analysis if left undetected, a so-called “doppelgänger” effect. It also means that the “doppelgänger” effect: hidden duplicates can inflate the apparent accuracy of predictive and prognostic models.

Firstly, the DEs are not unique to biomedical data. It is also found in other area of life. For example, the consumer literature established long ago that individuals monitor other people’s consumption behavior and copy it. In fact, many well-documented consumption phenomena are implicitly based on the assumption that consumers mimic other consumers’ behaviors. For example, opinion seekers copy the consumption behavior of opinion leaders (Flynn et al., 1996; Bertrandias and Goldsmith, 2006; StokburgerSauer and Hoyer, 2009), and teenagers mirror their role models’ choice preferences (Lockwood and Kunda, 1997). However, despite this rich literature, the issue of consumers’ intentional mimicry remains unaddressed. Thus, this paper focuses on people’s tendency to intentionally mimic other consumers’ behavior and regard this behavior as the consumer’s doppelganger effect.

In addition, there are several methods to avoid the DEs in the practice and development of machine learning models for health and medical science.

By placing all doppelgängers in the training set, accuracies drop to ~0.5, which is the expected accuracy of a model trained on random signatures. Obviously, when all pairwise Pearson’s correlation coefficient (PPCC) data doppelgängers are placed together in the training set, the doppelgänger effect is eliminated. (Rongwang et al., 2022)

Given the potential for unrecognized duplication to falsely inflate prediction accuracy and confidence in differential expression, doppelgänger-checking should be a part of standard procedure for combining multiple genomic datasets.

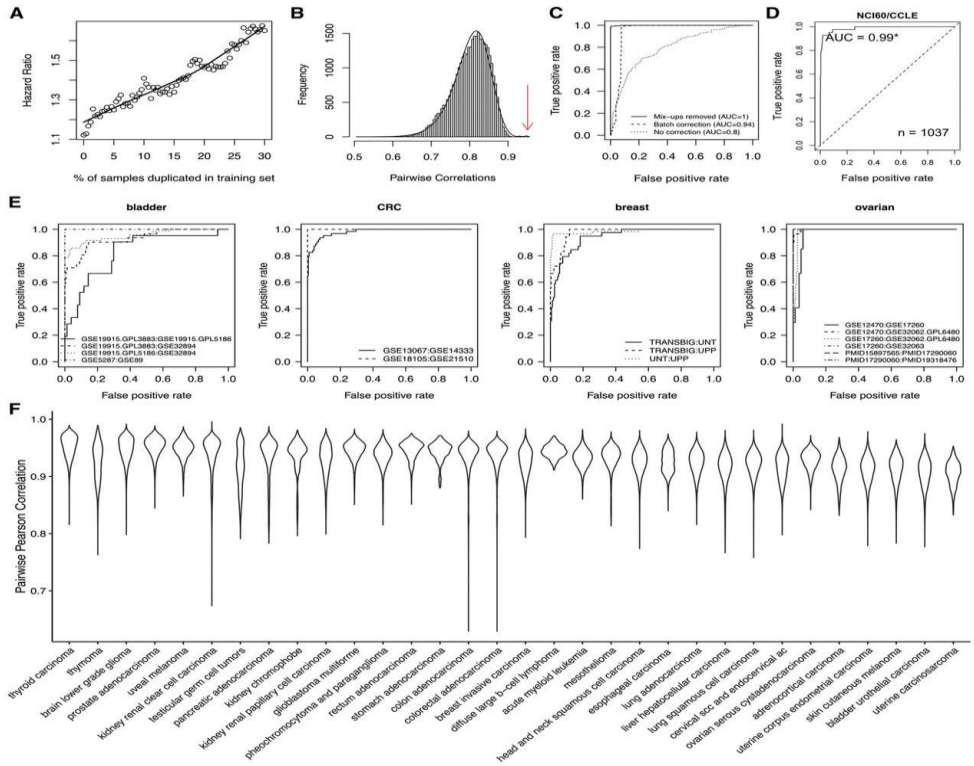
Previously, data doppelgängers were identified within a single data set between two even-sized batches. However, this experimental setup does not account for other compositions of data sets. For instance, a common practice for ML practitioners in the biomedical field is to utilize multiple data sets from different sources in order to increase statistical power and reduce uncertainty. This process is referred to as data integration or mega analysis (Eisenhauer, 2021), which, unfortunately, produces a multitude of problems, the most prominent of which is known as batch effects (BEs) (Goh et al., 2017). BEs are technical sources of variation that can confound statistical feature selection, and mislead ML model

training. The most common BE correction method, ComBat, is very widely used (Zhang et al., 2020) and usually assumed to work correctly. However, ComBat should not be applied carelessly as its efficiency relies on the balance of class distributions across batches (Li et al., 2021). Moreover, if the new source of data is used as a form of external validation (Ho et al., 2020a) i.e., the ML model is trained on a data set and evaluated on another independently-derived data set, the DE may overstate the ML model's performance (Wang et al., 2021). Scientists expect batch effects may confound DEs, especially when not removed properly. Recently, the effects of batch imbalance on proper batch effect removal are becoming a serious concern. Scientists believe it may also affect the sensitivity of our doppelgänger identification algorithm. Thus, our final aim is to see what are the technical barriers that prevent us from correctly estimating and observing DEs.

The proposed method relies on exhaustive comparisons of dataset pairs and sample pairs to empirically estimate the distribution of pairwise transcriptome correlations between biological replicates within a dataset or between two datasets where potentially different profiling technologies were used. The key aspects to identifying duplicates in a pair of datasets are 1) using transcript identifiers available in both datasets, 2) batch correction, 3) calculating Pearson's Correlation Coefficient (PCC) between every sample in one dataset against every sample in the other dataset, and 4) duplicate-oriented outlier detection. The background distribution of PPCC values varies depending on the tissue assayed and the technologies used, and must be estimated for every dataset pair. Doppelgängers can be identified as outliers at the high end of the distribution of batch-corrected correlations.

Scientists studied databases of ovarian, breast, bladder, and colorectal cancers and of cell lines and assessed their accuracy against a "gold standard" of duplicated samples generated through further manual inspection of expression data, clinical annotations, and sample identifiers. Confirmed doppelgängers were identified in more than half of all studies. For example, among the 1467 breast cancer gene expression profiles, Doppelgängers identifies 59 samples present in both the Sotiriou et al. and Miller et al. studies. In the ovarian cancer database, which they have inspected in great detail, Scientists identified 17% of records as nonunique, including duplicates in different datasets originating from the same institution, between the TCGA dataset and datasets of institutions that contributed samples to the TCGA project and within the TCGA dataset itself. In approximately 75% of duplicate pairs, samples matched by expression data had identical or compatible clinical and tumor data, but in the other 25% of cases the clinical data were discordant. Previous work on identifying duplicate microarray profiles has been limited to matching identical raw data files, and this would not identify any of these duplicates.

Cancer transcriptomes undergo alterations that are highly distinctive but much more difficult to identify uniquely in summarized form. Re-use of tissue specimens is widespread in clinical genomic studies, creating a “doppelgänger effect” in publicly available datasets: hidden duplicates that, if left undetected, can inflate statistical significance or apparent accuracy of genomic models when combining data from different studies.



Demonstration and benchmarking of the doppelgangR method for identifying expression profiles of the same biological specimen. A) The “doppelgänger” effect: hidden duplicates can inflate the apparent accuracy of predictive and prognostic models. Models of overall survival for high-grade, serous ovarian cancer were trained and then validated in two studies containing duplicates identified by *doppelgangR* . Validation set hazard ratio was calculated with duplicates incrementally removed so that between 0% and 30% cross-study duplication of samples remained. Thirty percent duplication inflates the apparent hazard ratio from 1.1 to 1.7. B) doppelgangR identifies duplicate expression profiles as outliers with unusually high pairwise correlation compared with other pairs of unrelated expression profiles. This histogram is the diagnostic plot produced by doppelgangR software, showing the best fit to the distribution of pairwise correlations, with vertical darklines showing outliers that are probable duplicates in the UNT (Unilateral nevoid telangiectasia) and Miller et al. breast cancer datasets. C) Batch correction allows RNA-seq profiles to be matched accurately to Affymetrix microarray profiles in the The Cancer Genome Atlas (TCGA) ovarian cancer dataset. True positives are tumors whose RNA-seq and microarray profiles are more highly correlated to each other than to any other

profile. Batch correction increases area under the receiver operating characteristic plot (AUC) from 0.79 to AUC = 0.94, and removing 50 microarray profiles incorrectly labeled by TCGA further increases AUC to > 0.995. D and E) Benchmarking. Scientists estimated the accuracy of the *doppelgangR* approach by applying it to pairs of datasets with confirmed duplicates. D) Shows AUC for identifying the 43 cell lines present in two different panels (CCLE [n = 1037] and NCI60 [n = 59]). **E)** The performance on primary tumor data in four cancer types. The AUC averaged across the four cancer types is 0.97. F) Suitability to TCGA cancer types. The doppelgangR approach only works for cancer types in which expression profiles of individual tumors are sufficiently distinct. Violin plots depict distributions of PCCs for all pairs of expression profiles within each TCGA dataset, in order (left to right) of increasing distinctiveness. In cancer types with high pairwise PCCs, such as thyroid carcinoma, patients have very similar expression profiles and are hard to distinguish based on expression data only. In contrast, in cancer types with low PCCs, such as bladder cancer, extensive genomic alterations generate unique expression fingerprints that make doppelgänger identification possible. (Rongwang et al., 2022)

Reference

Flynn LR, Goldsmith R, Eastman J. 1996. Opinion leaders and opinion seekers: Two new measurement scales. Journal of the Academy of Marketing Science 24(2): 137–147.

Bertrandias L, Goldsmith RE. 2006. Some psychological motivations for fashion opinion leadership and fashion opinion seeking. Journal of Fashion Marketing and Management 10(1): 25–40.

Stokburger-Sauer NE, Hoyer WD. 2009. Consumer advisors !revisited: What drives those with market mavenism and opinion leadership tendencies and why? Journal of Consumer Behaviour 8(2): 100–115.

Lockwood P, Kunda Z. 1997. Superstars and me: Predicting the impact of role models on the self. Journal of Personality and Social Psychology 73(1): 91–103.

Johnson WE Li C Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods *Biostatistics*. 2007;8(1):118–127.

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.

Miller LD Smeds J George J, et al. An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proc Natl Acad Sci U S A.* 2005;102(38):13550–13555.

Ganzfried BF Riester M Haibe-Kains B, et al. curatedOvarianData: clinically annotated data for the ovarian cancer transcriptome. *DATABASE*. 2013;2013:bat013.