Machine Learning in Preclinical Drug Proarrhythmic Assessment

Nan Miles Xi [a], Yu-Yi Hsu [b], Qianyu Dang [b], and Dalong Patrick Huang [b]

[a] Department of Mathematics and Statistics, Loyola University Chicago, Chicago, IL 60660, USA
[b] Office of Biostatistics, Office of Translational Science, Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, MD 20993, USA
Disclaimer: This work reflects the views of the authors and should not be construed to represent FDA's views or policies

Introduction

Torsades de pointes (TdP) is a rare but potentially fatal ventricular arrhythmia largely caused by electrolyte imbalance and cardiomyopathies after drug treatment. The identification of TdP is a crucial step in the assessment of safety before a drug reaches the market. The regulatory agency, pharmaceutical industry, and academia proposed serval new paradigms to better predict the drug-induced TdP risk in preclinical studies. The Comprehensive In Vitro Proarrhythmia Assay (CiPA), initiated by the US Food and Drug Administration (FDA), is an essential public-private collaboration among these attempts. Additionally, the rabbit ventricular wedge assay (RVWA), an established in vitro paradigm for detecting drug-induced QT prolongation and arrhythmia, has been adapted for the assessment of drug-induced TdP risk.

In this study, we proposed two statistical learning models, ordinal logistic regression and ordinal random forest, to accurately predict drug-induced TdP risk on datasets generated under CiPA and RVWA paradigms. Our predictive models utilized the ordinal information in low-, intermediate-, and high-risk levels instead of treating them as independent categories. The unbiased model performance on new drugs was estimated by leave-one-drug-out cross-validation (LODO-CV). The uncertainty of model performance was further quantified by stratified bootstrap. We identified the potential outlier drugs using the asymptotic prediction accuracy obtained from stratified bootstrap. Sensitivity analysis was then conducted to investigate the impact of potential outlier drugs on the model performance. To further validate and improve model prediction, we conducted control analysis, a common practice in in vitro studies, by selecting one control drug with mechanistically understood TdP risk. Finally, we examined the model interpretability through the analysis of normalized permutation predictor importance.

In summary, our work is the first attempt to construct multivariate statistical learning models that can accurately predict the drug-induced TdP risk from in vitro data. It satisfies the principles of statistical learning, highlighted by its comprehensive uncertainty measurements and strong interpretability. The proposed modeling and evaluation process can be extended easily to new datasets generated by other experimental protocols. The result of model prediction will serve as supplemental evidence in the drug safety assessment.

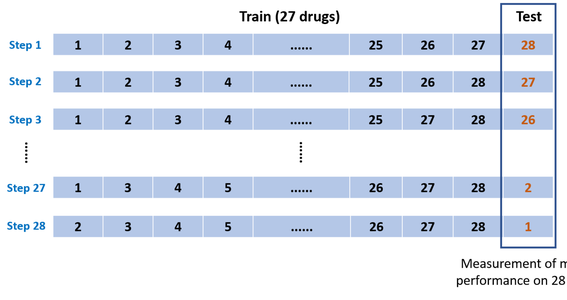

Figure 1. Leave-one-drug-out cross-validation. In each iteration, we trained the predictive model on 27 training drugs and predicted one left-out drug. The same process was repeated until each drug was predicted.

Methods

Leave-one-drug-out cross validation

Let N be the number of drugs in the dataset ($N = 28$ in both datasets) and J be the number of observations per drug ($J = 15$ in the stem cell dataset; $J = 4$ in the wedge dataset). Denote $\hat{f}^{-k}$ as the predictive model trained on the dataset *without* the observations of drug k. Let (x_j^k, y_j^k) be the jth observation of drug k, where x and y refer to the predictor vector and risk category, respectively. Then the three-category prediction accuracy under LODO-CV, $acc_{LODO-CV}$, is calculated as

$$acc_{LODO-CV} = \frac{1}{N}\sum_{k=1}^{N}\frac{1}{J}\sum_{j=1}^{J} I(y_j^k = \hat{f}^{-k}(x_j^k))$$

where $I(x)$ is the indicator function (Figure 1).

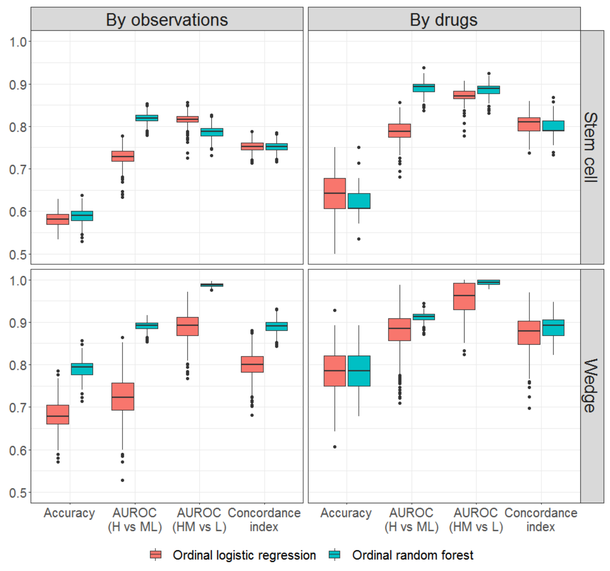

Figure 2. The empirical distributions of model performance under stratified bootstrap. .

Utilization of ordinal information

First, we trained a binary classifier $\hat{f}_1$ which differentiated observations between low risk and intermediate-or-high risk. Similarly, we trained another binary classifier $\hat{f}_2$ which differentiated observations between high risk and low-or-intermediate risk. Second, for any observation x in the test set (i.e., the left-out drug), we predicted its probability of low risk $p_x(L)$, intermediate-or-high risk $p_x(MH)$, high risk $p_x(H)$, and low-or-intermediate risk $p_x(LM)$ as:

$$p_x(L) = \hat{f}_1(x),$$
$$p_x(MH) = 1 - p_x(L),$$
$$p_x(H) = \hat{f}_2(x),$$
$$p_x(LM) = 1 - p_x(H).$$

Then the probability of intermediate risk $p_x(M)$ was calculated as

$$p_x(M) = p_x(MH) - p_x(H).$$

Finally, the risk of *observation* x, $\widehat{risk}(x)$, was predicted as

$$\widehat{risk}(x) = argmax_r\{p_x(r)\}$$

where $r \in \{L, M, H\}$.

Ordinal logistic regression model

Formally, for any observation x, the two binary classifiers $\hat{f}_1$ and $\hat{f}_2$ in the ordinal framework are defined as:

$$\hat{f}_1\text{: } log\frac{p_x(L)}{1-p_x(L)} = X\hat{\beta}_1$$
$$\hat{f}_2\text{: } log\frac{p_x(H)}{1-p_x(H)} = X\hat{\beta}_2$$

where X is the predictor vector of observation x and $\hat{\beta}_1$ and $\hat{\beta}_2$ are two model parameter vectors (including interceptions). $X\hat{\beta}_1$ and $X\hat{\beta}_2$ are inner products between predictor vectors and parameter vectors. We estimated $\hat{f}_1$ and $\hat{f}_2$ by maximum likelihood estimation. After obtaining $\hat{f}_1$ and $\hat{f}_2$, the prediction of risk categories for each observation and drug were calculated as described in the previous section.

Ordinal random forest model

Random forest is the ensemble of multiple decision trees and can capture the nonlinear relationship in the dataset. A decision tree T is a predictive model that assigns each observation to a certain category based on split rules defined on the predictor space. Formally, suppose that there are P predictors $X_1, X_2, \ldots, X_P$ in the dataset, and we split the predictor space into two regions, R_1 and R_2, according to predictor X_t and threshold s:

$$R_1(x,t,s) = \{x|X_t \le s\}$$
$$R_2(x,t,s) = \{x|X_t > s\}$$

where x denotes observation. Then for any region R_m with N_m observations, let $\hat{p}_{mr}$ be the proportion of category r in region R_m:

$$\hat{p}_{mr} = \frac{1}{N_m}\sum_{x_i \in R_m} I(y_i = r)$$

where x_i and y_i are the predictor vector and risk category of observation i, respectively. $I(x)$ is the indicator function. The risk of any observation x in region R_m is predicted as:

$$\widehat{risk}(x) = argmax_r\hat{p}_{mr},$$

where $r \in \{L, M, H\}$. In each split generating regions R_1 and R_2, we seek the predictor X_t and threshold s by minimizing the misclassification error, Gini index, or cross-entropy.

Results

Overall model performance

Figure 2 compares the model uncertainty of ordinal logistic regression and ordinal random forest calculated by stratified bootstrap. The ordinal logistic regression and ordinal random forest exhibit similar prediction performance on the stem cell dataset. Ordinal random forest consistently outperforms ordinal logistic regression on the wedge dataset.

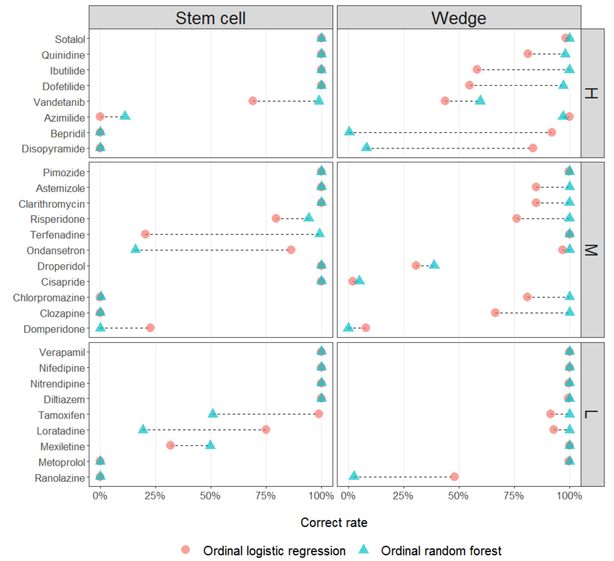

Figure 3. The correct rate of drug prediction calculated by stratified bootstrap. For each drug, we connect the correct rates of two models. In each risk category, drugs are sorted from high to low based on their average correct rates across two models.

Drug prediction analysis

We calculated the proportion of correct predictions (correct rate) for each drug in the 1000 stratified bootstrap predictions. Figure 3 shows the correct rate of predicting each drug across different model-dataset combinations. In the stem cell dataset, there are eight drugs on which both models resulted in less than 25% correct rates. In the wedge dataset, however, only two drugs of intermediate risk are difficult for both models to predict. Many of the drugs with close-to-zero correct rates have been reported to have abnormal observations in their original experiments.

Discussion

The prediction accuracy of two models on the wedge dataset is consistently higher than the stem cell dataset. Such discrepancy is largely due to the different experimental designs of the two datasets. Observations in the stem cell dataset were generated at 10 experimental sites, using two hiPSC-CM lines and five EP platforms. All observations in the wedge dataset were generated at one laboratory using the same type of biological sample and EP platform. Although the multisite experiments were supposed to follow the same protocols, the batch effects caused by site-to-site variability introduced a higher degree of noise in the stem cell dataset. The signal-noise ratio in the stem cell dataset is thus lower than the wedge dataset, resulting in lower prediction accuracy. One potential solution for this issue is to estimate the effects of site, cell line, and EP platform and include these factors into the modeling process. The accurate estimation of such effects requires special experimental designs with enough power to describe and explain the variations from those factors.