Machine learning-based identification of predictors influencing sub-therapeutic rifampicin concentrations in HIV/TB co-infected patients

Timothy Mwanje Kintu[1], Mugume Twinamatsiko Atwine[1], Ronald Galiwango[1], Christine Sekaggya-Wiltshire[2], Barbara Castelnuovo[2]

[1]African Center of Excellence in Bioinformatics and Data Science, Infectious Diseases Institute

[2] Infectious Diseases Institute, Makerere University

Abstract: Managing Tuberculosis (TB)/HIV co-infected patients poses challenges due to pill burden, compliance, and possible toxic effects. Identifying patients at risk of sub-therapeutic drug concentrations is crucial for guiding interventions. This study used machine learning to identify predictors of sub-therapeutic rifampicin concentrations in 268 TB/HIV co-infected patients from the SOUTH cohort profile. Two datasets were analyzed: the original and synthetic (2000 data points generated). The best-performing model, Random Forest Classifier, was fitted and evaluated through cross-validation. Most participants showed sub-optimal rifampicin concentrations. BMI, age, systolic blood pressure, and baseline culture result emerged as crucial predictors for both datasets. ML demonstrates potential in improving TB/HIV patient management, enabling personalized interventions like drug dosing adjustments and adherence monitoring to optimize treatment outcomes.

Background

Tuberculosis (TB) and HIV continue to pose a significant global burden, with an estimated 1.1 million people living with both infections, 80% of whom are in SSA. Despite being preventable and curable, TB is the leading cause of death among People Living with HIV (PLWH) [1]. Treating TB/HIV co-infected patients requires co-administration of anti-TB and antiretroviral therapy (ART) to be administered concomitantly, which introduces several challenges, including pill burden and patient compliance, overlapping toxic effects, and immune reconstitution inflammatory syndrome. In particular, the pharmacokinetics of anti-TB drugs may be altered in PLWH due to drug interactions, differences in drug metabolism, and changes in the distribution and elimination of drugs, affecting the effectiveness of the drugs and increasing the risk of toxicity.

Among the first-line anti-TB drugs, Rifampicin and isoniazid have been documented to display concentration-dependent killing of mycobacteria, leading to a decrease in bacterial load within the first few days of treatment [2]. As such, sub-therapeutic concentrations of these drugs may carry an increased risk of adverse drug outcomes, affecting adherence, and delayed culture conversion, increasing TB transmission in the communities [3]. Therefore, identifying patients likely to have sub-therapeutic drug concentrations can guide interventions, such as compliance monitoring and adjustment in drug dosing.

While traditional statistical approaches have previously been used to identify predictors of sub-therapeutic drug concentrations [4], machine learning (ML) methods have emerged as a powerful tool for analyzing complex and high-dimensional data. In this study, we propose to develop a ML algorithm to predict the likelihood of sub-optimal rifampicin concentration in a cohort of HIV/TB co-infected patients. Using machine learning methods, we sought to identify non-linear relationships and predictors of sub-therapeutic rifampicin concentrations, which can inform clinical decision-making and improve patient outcomes.

Methods

Study population: Data was extracted from the SOUTH cohort profile comprising 268 TB/HIV co-infected patients. Details on this cohort can be found in a previous paper by Sekaggya-Wiltshire and colleagues [5].

Study outcome: The maximum concentrations of rifampicin were calculated as the highest concentration among the three blood draws on any particular visit. The maximum concentrations were established for each visit (weeks 2, 8, and 24). The outcome of interest was a sub-therapeutic rifampicin concentration on the first clinical visit. A sub-therapeutic rifampicin concentration was defined as less than 8mg/L.

Data analysis: Data was cleaned in RStudio and TableauPrep, with only variables with more than one category and those hypothesized to affect the outcome, based on literature and expert review, retained. The dataset was then imported into Google Colaboratory and Jupyter Notebooks for analysis using Python. A new variable, Body Mass Index (BMI), was created based on the height and weight values in the dataset. In order to identify the most important features in making the prediction, the Python Featurewizz library [6] for feature selection and engineering was used. Featurewizz selects these features through steps: 1) Minimum Redundancy Maximum Relevance, where it selects pairs of highly correlated variables exceeding a selected correlation threshold (in this case, 0.7) and then 2) finding the Mutual Information Score, that quantifies the mutual dependence between each of the selected pair of variables and the outcome variable. Only the variable with a high MIS was retained for each pair of correlated variables. Featurewizz then uses the XGBoost (Extreme Gradient Boosting) machine learning (ML) algorithm to select the most important features from the remaining group of features.

Two routes of analysis were then taken, one utilizing synthetic data, where 2000 data points were created, and another utilizing the original dataset. In both cases, the ML workflow was set up in PyCaret, which automatically divided the data into a training and test set and dealt with any multi-collinearity. Prior to entering the data in the ML workflow, the SMOTE (Synthetic Minority Oversampling Technique) algorithm was used to handle the class imbalance, given that majority of the data points had sub-therapeutic concentrations as the outcome. SMOTE works by identifying the minority class and creating synthetic samples for this class by using the k-nearest neighbors' algorithm [7]. The performance of different machine learning models (including logistic regression, random forests, gradient boosted classifier, Light Gradient Boosting machine, extra trees classifier, among others) on the data was compared based on evaluation metrics such as accuracy, precision, recall, Matthew's Correlation Coefficient (MCC) and Area Under the ROC curve (AUC-ROC).

For both the synthetic and non-synthetic data, the best performing machine learning models were fitted to the data using a popular Python machine learning library, sci-kit learn. Cross validation was done to evaluate the effectiveness and generalizability of the created model. To further improve the performance of the selected model, hyperparameter optimization with Grid Search was done. The pre-trained evaluation metric for hyperparameter optimization was AUC-ROC. The optimal hyperparameters identified were used to train the final model.

Results

Study characteristics: Of the 268 participants, data from the first clinical visit was available for 251 participants. Most of these participants (62%) had sub-optimal rifampicin concentrations on their first visit. Out of the selected 54 original features, 13 were selected by the algorithm to make up the baseline model. These included: being an adult, age, Body Mass Index (BMI), baseline culture positivity, being on ART, taking herbs (yes, no or unknown), systolic blood pressure, having symptoms of (yes/no): weight loss, sputum production, chest pain, difficulty in breathing, and Chest Xray features (pleural effusion and infiltrates).

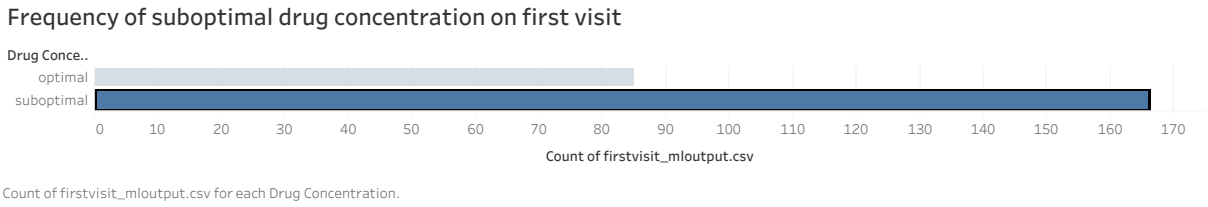

Figure 1: Optimal versus sub-optimal drug concentrations

Comparing performance across different models: Performance was compared across five different machine learning models. For the original dataset, the best performing models were the extra trees classifier, extreme gradient

boosting, random forest classifier, quadratic discriminant analysis and the gradient boosting classifier.

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
et	Extra Trees Classifier	0.6312	0.7338	0.6750	0.6272	0.6321	0.2546	0.2667	0.8130
xgboost	Extreme Gradient Boosting	0.6181	0.6631	0.6375	0.5963	0.5996	0.2285	0.2404	0.7385
rf	Random Forest Classifier	0.6167	0.6716	0.6550	0.5849	0.6070	0.2276	0.2344	0.8330
qda	Quadratic Discriminant Analysis	0.6021	0.5900	0.5650	0.6236	0.5765	0.1952	0.2041	0.7045
gbc	Gradient Boosting Classifier	0.5931	0.6306	0.6050	0.6011	0.5886	0.1855	0.1911	0.7720

Figure 2: Performance of ML models on non-synthetic dataset

For the synthetic dataset, the best performing models were Extreme Gradient Boosting, the Random Forest Classifier, Light Gradient Boosting machine, extra trees classifier and the Gradient Boosting Classifier.

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
xgboost	Extreme Gradient Boosting	0.7299	0.8050	0.7241	0.7368	0.7285	0.4598	0.4621	0.7835
rf	Random Forest Classifier	0.7248	0.8035	0.7280	0.7273	0.7258	0.4496	0.4518	0.7575
lightgbm	Light Gradient Boosting Machine	0.7234	0.7976	0.7150	0.7313	0.7208	0.4467	0.4494	0.8865
et	Extra Trees Classifier	0.7120	0.7903	0.6907	0.7231	0.7043	0.4240	0.4267	0.8555
gbc	Gradient Boosting Classifier	0.6913	0.7631	0.6624	0.7042	0.6807	0.3827	0.3852	0.7125

Figure 3: Performance of ML models on synthetic dataset

Using the Random Forest Classifier, a baseline model was built with cross-validation for both the synthetic and non-synthetic datasets utilizing 12 and seven folds respectively. The performance was then compared across four different evaluation metrics (Matthew's Correlation Coefficient, ROC-AUC, Precision and Recall) (**Table 1**).

Table 1: Comparison of evaluation metrics for the synthetic and non-synthetic datasets

Dataset	AUC	MCC	Precision	Recall
Non-synthetic (n=251)	0.77	0.37	0.67	0.72
Synthetic (n=2000)	0.83	0.50	0.75	0.75

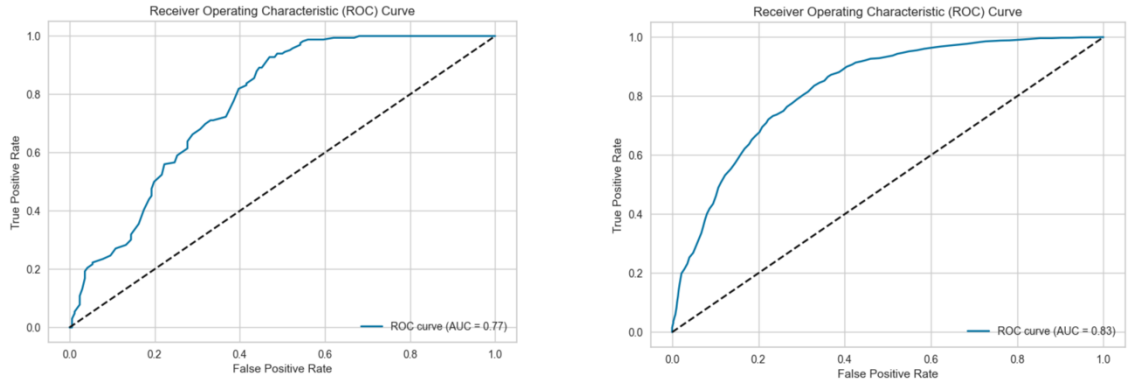

Figure 4: AUC-ROC Comparison for the non-synthetic (left) and synthetic datasets

Feature importance: For the non-synthetic dataset, the five most important features were BMI, age, systolic blood pressure, use of herbs and the baseline culture result; whereas for the synthetic dataset, the five most important features were age, BMI, systolic blood pressure, baseline culture result and the presence of pleural effusion on Chest Xray (**Figure 5 and 6**).

Conclusions and limitations

These findings suggest that machine learning models can effectively predict the likelihood of sub-therapeutic drug concentrations in HIV/TB co-infected patients. BMI and age were essential features in predicting the likelihood of sub-therapeutic drug concentrations in the different datasets. These findings are consistent with previous studies that have shown the influence of these identified factors on anti-TB drug metabolism and distribution [8, 9]. This study highlights the potential of ML in improving the management of TB/HIV co-infected patients by aiding in tailoring personalized interventions, such as drug dosing adjustments and adherence monitoring, to optimize treatment outcomes and reduce the risk of adverse drug effects.

This analysis was greatly affected by the small sample size, which affected the validity and accuracy of these findings. Performing this analysis on a larger dataset may prove beneficial, as evidenced by the improved performance of the models on the synthetic dataset. As such, we recommend analyzing data on larger cohorts of patients as needed to confirm these findings and determine the generalizability of the machine learning models. The integration of ML in precision medicine holds significant promise for enhancing outcomes of TB/HIV co-infected patients.

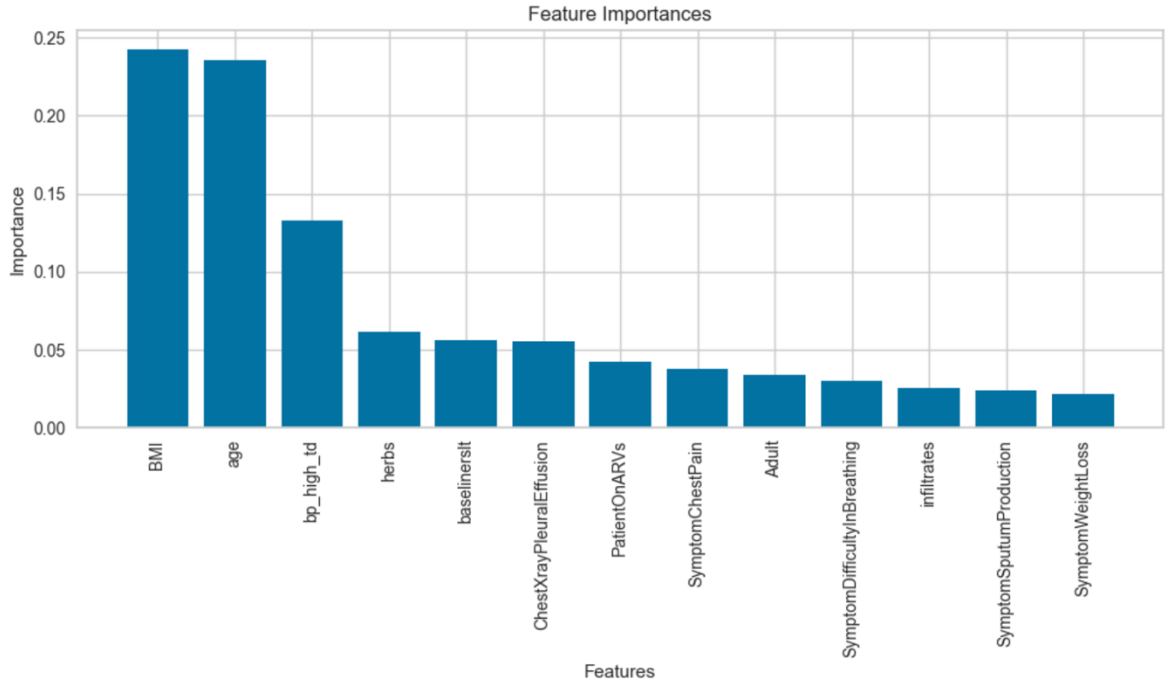

Figure 5: Feature Importance for non-synthetic dataset

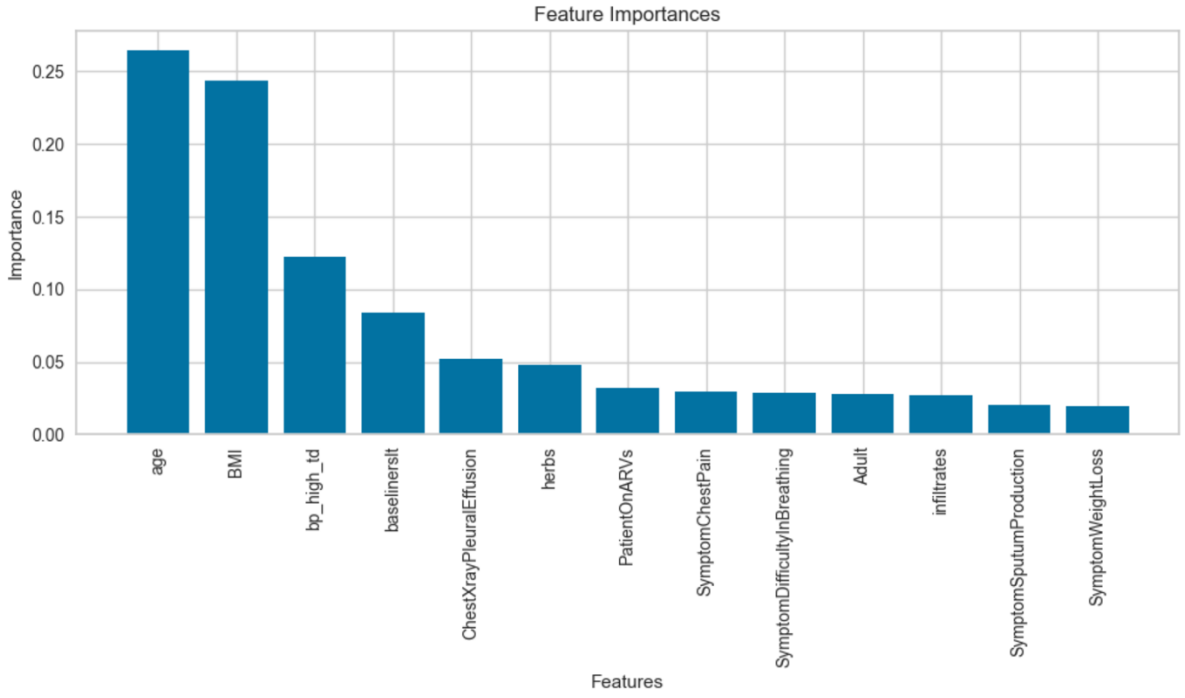

Figure 6: Feature Importance for synthetic data

References

1. Swaminathan S, Narendran G. HIV and tuberculosis in India. J Biosci. 2008;33:527–37.
2. Gumbo T, Louie A, Deziel MR, Liu W, Parsons LM, Salfinger M, et al. Concentration-Dependent *Mycobacterium tuberculosis* Killing and Prevention of Resistance by Rifampin. Antimicrob Agents Chemother. 2007;51:3781–8.
3. Sekaggya-Wiltshire C, von Braun A, Lamorde M, Ledergerber B, Buzibye A, Henning L, et al. Delayed Sputum Culture Conversion in Tuberculosis-Human Immunodeficiency Virus-Coinfected Patients With Low Isoniazid and Rifampicin Concentrations. Clin Infect Dis. 2018;67:708–16.
4. Udy AA, Varghese JM, Altukroni M, Briscoe S, McWhinney BC, Ungerer JP, et al. Subtherapeutic Initial β-Lactam Concentrations in Select Critically Ill Patients. Chest. 2012;142:30–9.
5. Sekaggya-Wiltshire C, Castelnuovo B, Von Braun A, Musaazi J, Muller D, Buzibye A, et al. Cohort profile of a study on outcomes related to tuberculosis and antiretroviral drug concentrations in Uganda: design, methods and patient characteristics of the SOUTH study. BMJ Open. 2017;7:e014679.
6. featurewiz H of A AutoViML and. featurewiz. 2023.
7. IBM. What is the k-nearest neighbors algorithm? https://www.ibm.com/topics/knn. Accessed 20 Jul 2023.
8. Tostmann A, Mtabho CM, Semvua HH, van den Boogaard J, Kibiki GS, Boeree MJ, et al. Pharmacokinetics of First-Line Tuberculosis Drugs in Tanzanian Patients. Antimicrob Agents Chemother. 2013;57:3208–13.
9. Ramachandran G, Hemanth Kumar AK, Bhavani PK, Poorana Gangadevi N, Sekar L, Vijayasekaran D, et al. Age, nutritional status and INH acetylator status affect pharmacokinetics of anti-tuberculosis drugs in children. int j tuberc lung dis. 2013;17:800–6.