

AI-Driven Pharmacogenomic Modeling of Gene–Micronutrient Interactions in Pulmonary Tuberculosis: An in-Silico Study Using Synthetic Patient Data

Dhivya S. Rajasekaran

Department of Pharmacology, Aarupadai Veedu Medical College (AVMC) and Hospital, Vinayaka Mission Research Foundation (Deemed to be University), Kirumampakkam, Puducherry 607402

Sivaraj Rengaraj

Department of Pharmacology, Aarupadai Veedu Medical College (AVMC) and Hospital, Vinayaka Mission Research Foundation (Deemed to be University), Kirumampakkam, Puducherry 607402

Satyavani Kaliamurthi

Medical and Bioresearch, Aarupadai Veedu Medical College (AVMC) and Hospital, Vinayaka Mission Research Foundation (Deemed to be University), Kirumampakkam, Puducherry 607402

Gurudeeban Selvaraj

`gurudeeban.selvaraj@avmc.edu.in`

Compute Medicine Lab, Department of Medical Biotechnology, Aarupadai Veedu Medical College (AVMC) and Hospital, Vinayaka Mission Research Foundation (Deemed to be University), Kirumampakkam, Puducherry 607402 <https://orcid.org/0000-0002-7223-3853>

Research Article

Keywords: pharmacogenomics, pulmonary tuberculosis, anti-tubercular therapy, SNP polymorphisms, micronutrients, machine learning, synthetic data, precision medicine

Posted Date: October 15th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11371300/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

AI-Driven Pharmacogenomic Modeling of Gene–Micronutrient Interactions in Pulmonary Tuberculosis: An in-Silico Study Using Synthetic Patient Data

Dhivya S. Rajasekaran¹, Sivaraj Rengaraj¹, Satyavani Kaliamurthi², Gurudeeban Selvaraj³

Department of Pharmacology, Aarupadai Veedu Medical College (AVMC) and Hospital, Vinayaka Mission Research Foundation (Deemed to be University), Kirumampakkam, Puducherry 607402

Medical and BioResearch, Aarupadai Veedu Medical College (AVMC) and Hospital, Vinayaka Mission Research Foundation (Deemed to be University), Kirumampakkam, Puducherry 607402

Compute Medicine Lab, Department of Medical Biotechnology, Aarupadai Veedu Medical College (AVMC) and Hospital, Vinayaka Mission Research Foundation (Deemed to be University),

Kirumampakkam, Puducherry 607402

Email: gurudeeban.selvaraj@avmc.edu.in

Abstract

Background: Treatment outcomes in pulmonary tuberculosis (PTB) vary widely despite standardized anti-tubercular therapy (ATT). Host genetic variation in immune-regulatory genes such as *BCL2* and *TNF*, together with micronutrient status (zinc, vitamin D₃, vitamin C), may modulate immune response and treatment outcome, but gene–micronutrient interactions are difficult to study in small clinical cohorts.

Objective: To develop and interpret an AI-driven pharmacogenomic modeling framework for gene–micronutrient interactions in adult PTB using a fully in silico cohort.

Methods: A synthetic cohort of 1,000 adults (18–65 years, balanced by sex) was generated. Genotypes for *BCL2* rs2279115 and *TNF* rs1800629 were simulated under Hardy–Weinberg equilibrium from 1000 Genomes Project allele frequencies, and serum 25-hydroxyvitamin D, zinc and vitamin C were sampled within clinically relevant ranges. A binary treatment outcome (success = 1; unfavorable = 0) was generated by a probabilistic function containing non-linear gene–micronutrient interaction terms and 10% stochastic noise. A conditional tabular GAN (CTGAN) was used to generate synthetic records, and fidelity was assessed by principal component and correlation analyses. Logistic regression, random forest and XGBoost were trained (70/30 split, 5-fold cross-validation, grid search) and evaluated by accuracy, precision, recall and AUC-ROC; SHAP analysis was used for interpretation.

Results: XGBoost performed best on the hold-out set (accuracy 87.2%, AUC-ROC 0.91, precision 0.85, recall 0.88), followed by random forest (84.5%, 0.88) and logistic regression (76.8%, 0.79). Vitamin D₃ (~35% of predictive weight) and *TNF* genotype (~28%) were the leading predictors, with zinc (~15%) and *BCL2* genotype (~12%) of moderate influence. SHAP analysis showed non-linear conditional effects: higher vitamin D₃ levels were required to reach the same predicted success probability in carriers of the *TNF* A allele, and the influence of *BCL2* genotype depended on zinc level.

Conclusions: The framework recovered non-linear gene–micronutrient structure from noisy mixed-type data and produced interpretable, testable hypotheses. Because outcomes were generated from

assumed rules, these results demonstrate methodological feasibility rather than real-world effects. Validation in real patient cohorts is essential before any clinical or programmatic use.

Keywords: pharmacogenomics; pulmonary tuberculosis; anti-tubercular therapy; SNP polymorphisms; micronutrients; machine learning; synthetic data; precision medicine

1. Introduction

Tuberculosis (TB), caused by *Mycobacterium tuberculosis* (*M. tb*), remains among the deadliest infectious diseases worldwide, with an estimated 10.7 million people falling ill and about 1.23 million deaths in 2024 [1]. Drug-susceptible pulmonary TB (PTB) is treated with a standardized six-month regimen of isoniazid, rifampicin, pyrazinamide and ethambutol [2]. Nevertheless, outcomes are heterogeneous: while many patients are cured, others experience treatment failure, relapse or death, and patient-level analyses of trial data show that the risk of an unfavorable outcome differs substantially between individuals [3]. Bacterial factors such as drug resistance, and clinical factors such as adherence and HIV co-infection, explain only part of this variability, which has focused attention on host-specific determinants.

Pharmacogenomic research in TB has largely concentrated on drug-metabolizing enzymes such as *NAT2*, where genotype-guided dosing reduced isoniazid-induced liver injury and early treatment failure in a randomized trial [4]. Host immune-regulatory variation is also relevant. Human genetic susceptibility to TB is well established, although individual variants usually have small and inconsistent effects [5,6], and work on *LTA4H* in tuberculous meningitis showed that host genotype can determine who benefits from adjunctive anti-inflammatory therapy [7]. Two immune-regulatory genes of particular interest are *TNF* and *BCL2*.

Tumor necrosis factor- α (TNF) is indispensable for granuloma formation and bacterial control: TNF-deficient mice are highly susceptible to *M. tb* [8], and TB reactivation is a recognized complication of TNF-neutralizing therapy [9]. The promoter variant rs1800629 (–308G>A) increases transcriptional activation in reporter assays [10], but associations of *TNF* promoter variants with TB have been inconsistent across studies [5,11]. *BCL2* encodes an anti-apoptotic protein. Apoptosis of infected macrophages is a host-protective mechanism associated with reduced bacterial viability, whereas necrosis favors bacterial spread, and virulent *M. tb* suppresses apoptosis [12], in part by inducing anti-apoptotic Bcl-2 family members such as Mcl-1 [13]. Polymorphisms in *BCL2* have been associated with TB susceptibility [14], and the promoter variant rs2279115 (–938C>A) alters *BCL2* promoter activity and expression [15].

Nutritional status is a long-recognized modulator of TB risk and outcome [16], and WHO recommends nutritional assessment and support for people with TB [17]; a large programmatic cohort in India reported better outcomes with nutritional support [18]. Among micronutrients, vitamin D is the most studied. Its active form induces cathelicidin and is required for the antimicrobial activity of human macrophages against *M. tb* [19,20], and vitamin D deficiency has been associated with active TB [21]. Clinical trials of adjunctive vitamin D, however, have been inconsistent [22,23]: an individual-

participant-data meta-analysis found no overall effect on sputum culture conversion, with a signal confined to multidrug-resistant disease [24], and vitamin D supplementation did not prevent TB infection in Mongolian schoolchildren [25]. Zinc is required for T-cell and macrophage function [26] and for regulating apoptotic pathways [27], although macrophages also exploit zinc toxicity against mycobacteria, which in turn carry efflux systems to resist it [28]; a meta-analysis of randomized trials found that zinc and vitamin A supplementation improved early sputum smear conversion but not treatment success [29]. Vitamin C kills *M. tb* in vitro through an iron-dependent Fenton reaction [30] and potentiates isoniazid and rifampin in mice [31], but human evidence is lacking.

Taken together, these observations suggest that the effect of a micronutrient may depend on host genotype and vice versa, which could explain part of the heterogeneity in trial results. Genetic and nutritional factors are nevertheless usually examined in isolation, and detecting interactions with conventional regression requires large samples and prespecified terms [32]. Tree-based ensembles such as random forests [33] and gradient-boosted trees (XGBoost) [34] can model non-linearities and interactions without prespecification and remain highly competitive on tabular data [35]; machine learning has already been applied in TB to predict treatment failure and drug resistance [36,37]. However, multimodal datasets that combine host genotype, micronutrient status and treatment outcome are scarce, particularly in high-burden, resource-limited settings. Synthetic data can support method development, hypothesis generation and privacy-preserving data sharing [38,39], provided that its limitations are made explicit. Such work is consistent with the goals of precision medicine [40] and the research and innovation pillar of the WHO End TB Strategy [41].

The aim of this study was to develop and interpret an AI-driven pharmacogenomic modeling framework for gene–micronutrient interactions in adult PTB, using a synthetic cohort built from 1000 Genomes Project allele frequencies and clinical micronutrient reference ranges. We specifically asked (i) whether standard machine-learning models can recover non-linear interactions between *BCL2* and *TNF* genotypes and serum zinc, vitamin D₃ and vitamin C levels, (ii) which features dominate prediction, and (iii) how these interactions can be made interpretable using SHAP analysis. Because the outcome was simulated, the study is explicitly a methodological and hypothesis-generating exercise.

2. Methods

2.1 Study design and reporting

This was an in-silico prognostic-modeling study based exclusively on synthetic data. No human participants, biological samples or identifiable data were used. Reporting was guided by the TRIPOD+AI statement where applicable to simulation-based studies [42]. Analyses were implemented in Python using scikit-learn [43], XGBoost [34] and the SHAP library [44,45].

2.2 Synthetic cohort

A cohort of 1,000 adult PTB patients was simulated, with an approximately balanced sex distribution and ages of 18–65 years. Each patient was described by two host genetic variables and three serum micronutrient variables (Table 1).

2.2.1 Pharmacogenomic features

Single-nucleotide polymorphisms (SNPs) were chosen in *BCL2* and *TNF*, two genes with established roles in the antimycobacterial immune response. For *BCL2*, the promoter variant rs2279115 (−938C>A) was modeled [15]. For *TNF*, the promoter variant rs1800629 (−308G>A) was modeled, with the A allele treated as the high-producer, high-inflammatory allele [10]. Allele frequencies were taken from the 1000 Genomes Project, using global averages to keep the model broadly applicable [46]. Each patient was assigned a genotype (homozygous reference, heterozygous or homozygous variant) by sampling under Hardy–Weinberg equilibrium from these allele frequencies.

2.2.2 Nutritional features

Serum levels of 25-hydroxyvitamin D (referred to as vitamin D₃ status), zinc and vitamin C were sampled from normal and skewed distributions parameterized from clinically relevant reference ranges and the deficiencies observed in TB-endemic populations. Vitamin D₃ was sampled between 5 and 60 ng/mL with approximately 40% of the cohort deficient (<20 ng/mL); zinc between 40 and 120 µg/dL; and vitamin C (serum ascorbic acid) between 0.1 and 1.5 mg/dL.

Table 1. Variables and simulation specifications for the synthetic cohort (n = 1,000).

Variable	Type / coding	Simulation specification
Age	Continuous (years)	18–65 years
Sex	Binary	Approximately balanced
<i>BCL2</i> rs2279115 (−938C>A)	Genotype (CC / CA / AA); one-hot encoded	Hardy–Weinberg sampling from 1000 Genomes global allele frequency (variant allele frequency)
<i>TNF</i> rs1800629 (−308G>A)	Genotype (GG / GA / AA); one-hot encoded	Hardy–Weinberg sampling from 1000 Genomes global allele frequency (variant allele frequency)
Serum 25(OH)D (vitamin D ₃ status)	Continuous (ng/mL)	5–60 ng/mL; ≈40% deficient (<20 ng/mL)
Serum zinc	Continuous (µg/dL)	40–120 µg/dL
Serum vitamin C	Continuous (mg/dL)	0.1–1.5 mg/dL
Treatment outcome	Binary (1 = success; 0 = unfavorable)	Probabilistic function of genotype and micronutrient levels with interaction terms; 10% stochastic noise

Abbreviations: 25(OH)D, 25-hydroxyvitamin D.

2.2.3 Outcome definition

The primary outcome was treatment success, modeled as a binary variable (success = 1; unfavorable outcome = 0). Outcome categories were coded as 0 = cured, 1 = death, 2 = failure and 3 = relapse, and the primary analysis dichotomized these into success (cured) versus unfavorable outcome (death, failure or relapse). The outcome was generated by a probabilistic logic function that encoded hypothesized biological effects: the probability of success decreased with vitamin D₃ deficiency and with high-inflammatory *TNF* genotypes, and non-linear interaction terms represented synergistic effects between micronutrient levels and genetic predisposition. A stochastic noise component (10%) was added to mimic unmeasured confounders and inherent variability in clinical data. Because the generating function embeds the hypothesized relationships, the learning task is one of recovering known structure from noisy data; the implications are discussed in Section 4.6.

2.3 Synthetic data generation and fidelity assessment

To expand the cohort and to test whether its statistical structure could be reproduced by a generative model, a conditional tabular generative adversarial network (CTGAN) was used. CTGAN is designed for tabular data with mixed continuous and categorical columns. It uses mode-specific normalization to represent non-Gaussian continuous distributions and a conditional generator with training-by-sampling to handle imbalanced categorical variables, and it learns the joint distribution through adversarial training of a generator and a discriminator [47].

Fidelity of the synthetic data relative to the reference data was assessed in two ways. First, both datasets were projected onto the leading principal components and overlaid, to check whether the synthetic data occupy the same region of feature space and show similar variability (Figure 1). Second, pairwise feature–feature correlation matrices (range -1 to $+1$) were computed for the reference and synthetic data and compared side by side, to check that key relationships among variables are preserved (Figure 2).

2.4 Preprocessing and feature selection

Categorical genetic variables were one-hot encoded so that all algorithms could process them. Continuous micronutrient variables were standardized (zero mean, unit variance) with a standard scaler so that variables with larger numerical ranges did not dominate model fitting. Feature engineering was deliberately minimal so that the models would have to discover interactions among the raw variables. Feature selection was applied to remove uninformative or redundant variables before the final models were trained, and model performance was compared with and without this step.

2.5 Machine-learning models

Three algorithms covering a range of complexity and interpretability were the primary models: (i) logistic regression, as a linear baseline; (ii) random forest, a bagging ensemble of decision trees that captures non-linear patterns [33]; and (iii) XGBoost, a regularized gradient-boosting framework that handles complex feature interactions through additive tree modeling [34]. In an exploratory comparison of the effect of feature selection, decision tree, support vector machine and k-nearest-neighbor classifiers were also evaluated.

2.6 Training and evaluation

The data were divided into training (70%) and hold-out test (30%) sets. A 5-fold cross-validation scheme was used within the training phase to limit overfitting, and hyperparameters of the random forest and XGBoost models (tree depth, number of estimators and, for XGBoost, learning rate) were tuned by grid search. Using cross-validation both for tuning and for error estimation can give optimistic estimates [48]; reported performance is therefore that on the untouched hold-out set. Treatment success was treated as the positive class. Performance was summarized by accuracy, precision (the proportion of patients predicted to succeed who did succeed), recall (the proportion of patients who succeeded that were correctly identified) and the area under the receiver operating characteristic curve (AUC-ROC).

2.7 Interpretability and interaction analysis

Global feature importance was obtained from the Gini importance of the random forest and the gain metric of XGBoost; both can be biased toward variables with many possible split points and are therefore reported as descriptive rankings. To move beyond black-box prediction, SHAP (SHapley Additive exPlanations) values, which derive from cooperative game theory, were computed for the best-performing model to quantify each feature's contribution to individual predictions [44]. For tree ensembles, exact SHAP values and SHAP interaction values can be computed efficiently [45]; these were used to visualize the direction of feature effects and to identify non-linear interactions between gene variants and micronutrient levels. SHAP values describe the behavior of the fitted model and do not by themselves represent causal effects.

3. Results

3.1 Fidelity of the synthetic data

The PCA projection (Figure 1) and the paired correlation matrices (Figure 2) were used to judge how closely the synthetic data reproduce the structure of the reference data: overlapping point clouds indicate a similar underlying distribution, and similar correlation patterns indicate that relationships among variables are preserved, an essential requirement for valid downstream modeling.

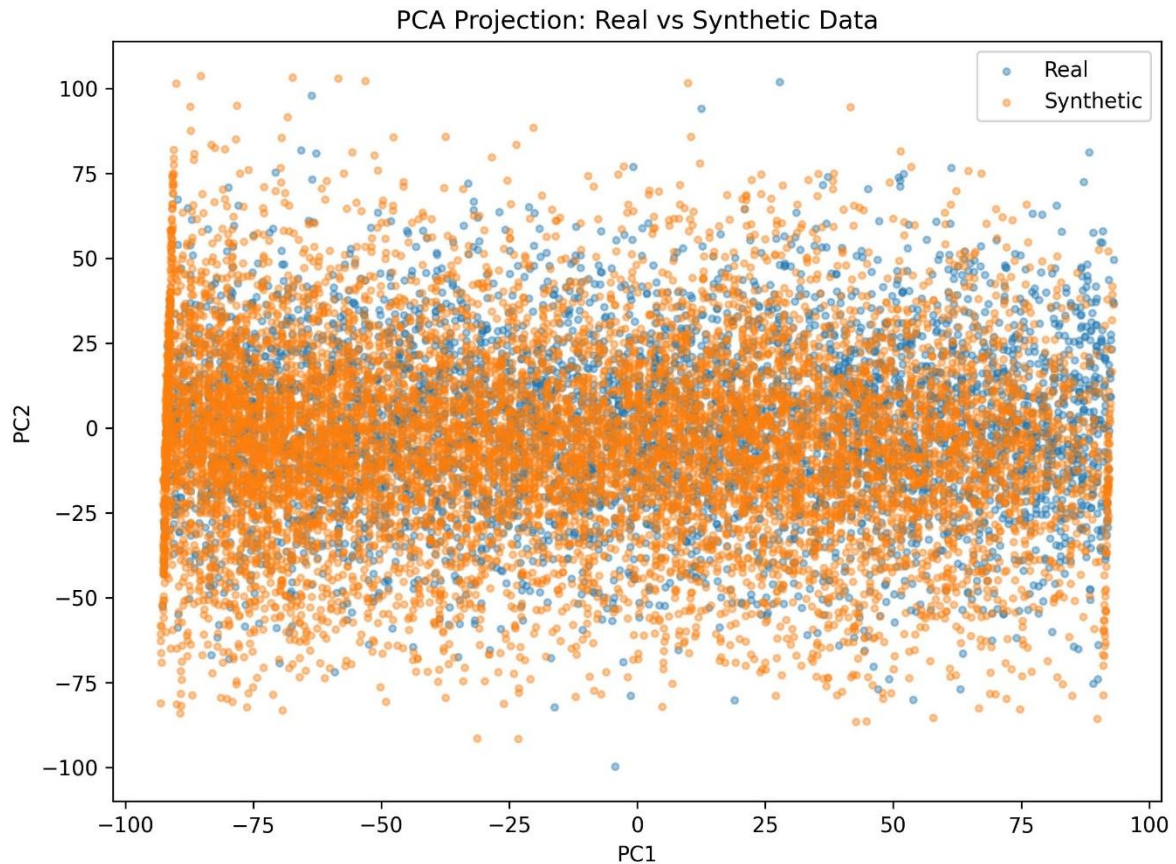


Figure 1. Principal component projection of reference data (blue) and CTGAN-generated synthetic data (orange).

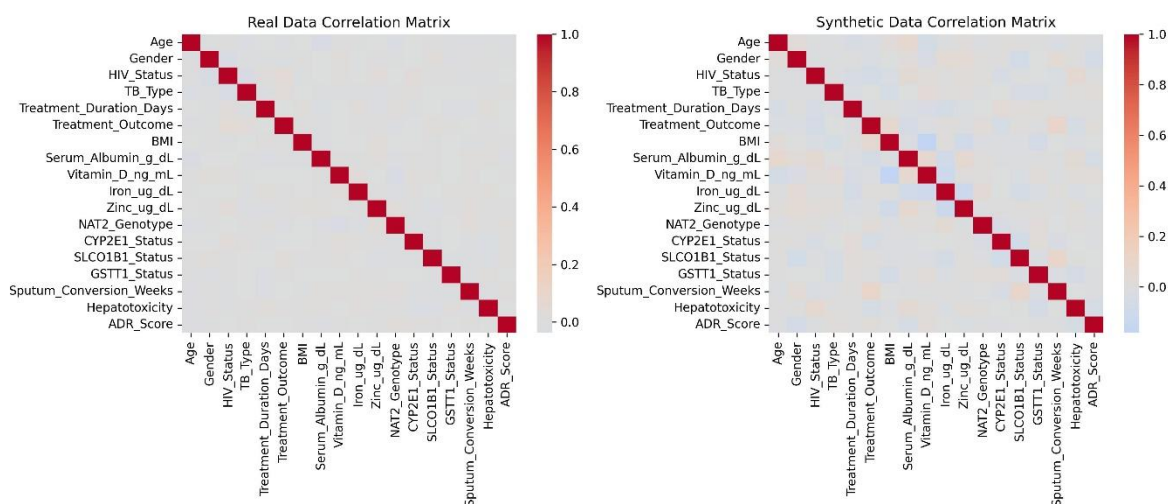


Figure 2. Feature–feature correlation matrices for the reference data (left) and the synthetic data (right). Colors indicate Pearson correlation coefficients from -1 to $+1$; the diagonal shows each feature's correlation with itself.

3.2 Effect of feature selection

In the exploratory comparison of six classifiers (random forest, XGBoost, decision tree, support vector machine, logistic regression and k-nearest neighbors), models trained on the full variable set performed poorly, with accuracy of approximately 0.28–0.45 and precision, recall and F1-scores below 0.30. After feature selection, the same models were retrained and performance improved substantially, with accuracy up to approximately 0.87 and precision, recall and F1-scores up to approximately 0.88. This indicates that removing noisy or redundant variables markedly improved the predictive performance of the models.

3.3 Model performance

On the hold-out test set ($n = 300$), XGBoost was the most accurate model, with an accuracy of 87.2% and an AUC-ROC of 0.91, indicating good separation of patients with and without treatment success (Table 2, Figure 3). The random forest followed closely (accuracy 84.5%, AUC-ROC 0.88), whereas logistic regression performed less well (accuracy 76.8%, AUC-ROC 0.79). The gap between the ensemble methods and the linear baseline is consistent with a non-linear, interaction-rich relationship between features and outcome, which is what the simulation was designed to contain.

For XGBoost, precision was 0.85 and recall was 0.88. With treatment success as the positive class, this means that 85% of patients predicted to succeed did so, and 88% of patients who succeeded were correctly identified. Metrics oriented toward detecting unfavorable outcomes (specificity and negative predictive value) were not the primary focus here and should be reported alongside calibration in future work, because identifying patients at risk of failure is the clinically more actionable task.

Table 2. Performance of the primary models on the hold-out test set ($n = 300$).

Model	Accuracy	AUC-ROC	Precision	Recall
XGBoost	87.2%	0.91	0.85	0.88
Random forest	84.5%	0.88	0.82	0.85

Model	Accuracy	AUC-ROC	Precision	Recall
Logistic regression	76.8%	0.79	0.73	0.76

Success = positive class. AUC-ROC, area under the receiver operating characteristic curve.

Figure 3. Receiver operating characteristic (ROC) curves of the classifiers (random forest, XGBoost, decision tree, support vector machine, logistic regression and k-nearest neighbors), plotting true-positive rate against false-positive rate.

3.4 Feature importance

Gini importance (random forest) and gain (XGBoost) consistently identified serum vitamin D₃ and *TNF* genotype as the most influential predictors (Table 3). Vitamin D₃ accounted for approximately 35% of the predictive weight and *TNF* genotype for 28%. Serum zinc (15%) and *BCL2* genotype (12%) showed moderate influence. Vitamin C (8%), age (2%) and sex (0.5%) had the smallest individual contributions.

Table 3. Relative feature importance in the XGBoost model.

Feature	Relative importance (%)	Influence
Serum vitamin D ₃ (25-hydroxyvitamin D)	≈35	Dominant
<i>TNF</i> rs1800629 genotype	≈28	Dominant
Serum zinc	≈15	Moderate
<i>BCL2</i> rs2279115 genotype	≈12	Moderate
Serum vitamin C	≈8	Low
Age	≈2	Low
Sex	≈0.5	Low

Values are approximate relative contributions to predictive weight.

3.5 SHAP analysis and gene–micronutrient interactions

SHAP summary plots were generated for the random forest, decision tree and XGBoost models (Figure 4). The three models ranked broadly similar features highest; the decision tree produced sharper, more discrete transitions in SHAP values, whereas XGBoost showed the most structured and stable pattern of feature contributions.

In the XGBoost model, higher vitamin D₃ levels were strongly associated with a positive impact on model output (a higher predicted probability of success). Conversely, carrying the *TNF* A allele at rs1800629 (the high-producer variant) was associated with a higher predicted risk of unfavorable outcome, particularly when combined with low micronutrient levels.

SHAP interaction analysis highlighted non-linear synergies between genetics and nutrition (Figure 5). The clearest was between vitamin D₃ and *TNF* genotype. In patients with the GG (wild-type) genotype, moderate vitamin D₃ levels were sufficient to predict a high probability of success, whereas in patients carrying the GA or AA genotype, substantially higher vitamin D₃ levels were needed to reach a similar predicted probability. A second interaction involved *BCL2* and zinc: the influence of *BCL2* genotype depended on serum zinc, with the effect of certain genotypes more pronounced when zinc was low.

Vitamin C showed a smaller, conditional contribution that was most evident in carriers of the *TNF* A allele.

Figure 4. SHAP summary plots for (A) random forest, (B) decision tree and (C) XGBoost. Features are listed on the y-axis and SHAP values on the x-axis; point color indicates feature value (red, high; blue, low).

Figure 5. SHAP interaction plot for serum vitamin D₃ and *TNF* rs1800629 genotype in the XGBoost model.

3.6 Illustrative model-derived scenarios

To illustrate the practical meaning of the interaction, two hypothetical patients with the same baseline vitamin D₃ level (15 ng/mL) but different *TNF* genotypes were compared (Table 4). Patient A (AA genotype) had a predicted probability of unfavorable outcome of 72%, which fell to 18% when vitamin D₃ was set to 40 ng/mL, an absolute reduction of 54 percentage points. Patient B (GG genotype) had a baseline predicted probability of 30%, falling to 12% at the same vitamin D₃ level, an absolute reduction of 18 percentage points. These figures are predictions from the model logic applied to hypothetical inputs; they are not estimates of the causal effect of supplementation.

Table 4. Model-derived predicted probability of unfavorable outcome for two hypothetical patients.

Patient	<i>TNF</i> genotype	Vitamin D ₃ (ng/mL)	Predicted probability of unfavorable outcome	Absolute reduction
A	AA	15 → 40	72% → 18%	54 points
B	GG	15 → 40	30% → 12%	18 points

Hypothetical scenarios computed from the model logic; not causal estimates.

4. Discussion

4.1 Principal findings and how to read them

In a fully simulated cohort of 1,000 adults with PTB, tree-based ensembles outperformed logistic regression, with XGBoost reaching 87.2% accuracy and an AUC-ROC of 0.91 on the hold-out set. Vitamin D₃ and *TNF* genotype dominated prediction, zinc and *BCL2* genotype contributed moderately, and SHAP analysis indicated conditional, non-linear relationships between genotype and micronutrient level. These results need to be read in light of the study design. The outcome was generated from a function that explicitly encodes reduced success with vitamin D₃ deficiency and high-inflammatory *TNF* genotypes, together with interaction terms. The superiority of non-linear models over the linear baseline, and the prominence of vitamin D₃ and *TNF*, are therefore expected, and the findings show that the pipeline can recover embedded structure from noisy, mixed-type data. They do not show that such interactions exist, or have these magnitudes, in patients.

The contribution of the work lies elsewhere: it provides a reproducible and interpretable workflow, a benchmark of how well standard learners recover interactions at a realistic sample size and noise level, and a set of explicit hypotheses that can be tested in real cohorts.

4.2 Vitamin D₃ and TNF

The modeled prominence of vitamin D₃ is biologically motivated, since vitamin D signaling induces cathelicidin and supports macrophage antimycobacterial activity [19,20]. TNF is a double-edged mediator: it is required for granuloma formation and bacterial control [8,9], yet high-producer alleles such as -308A, which increase promoter activity [10], might also favor inflammatory pathology. One plausible reading of the simulated interaction, which would require experimental and clinical testing, is that adequate vitamin D₃ status matters more for patients with a pro-inflammatory *TNF* genotype.

This conditional structure is relevant to the mixed clinical evidence on vitamin D. Trials and pooled analyses have generally shown no overall benefit on sputum culture conversion in drug-susceptible disease, with a signal in multidrug-resistant TB [22–24]. If benefit were confined to genetically or nutritionally defined subgroups, an average treatment effect could be diluted toward zero even though some patients benefit. The scenarios in Table 4 illustrate this logic in a simulated setting: the predicted absolute gain from raising vitamin D₃ was three times larger for the AA genotype than for the GG genotype. If such patterns were confirmed in real data, interventions could be prioritized for patients most likely to benefit, and trials could enrich for them. At present, however, this is a hypothesis, not a recommendation for genotype-based supplementation.

4.3 The apoptosis–nutrition axis: BCL2 and zinc

Apoptosis of infected macrophages restricts bacterial growth, and virulent *M. tb* works to suppress it [12,13]. Zinc influences apoptotic signaling [27] and immune-cell function [26], which provides a rationale for modeling an interaction between a *BCL2* promoter variant and zinc status. The role of zinc in TB is, however, not one-directional, because zinc is also deployed against mycobacteria by macrophages [28], and zinc supplementation trials have not shown improved treatment success [29]. The moderate importance of zinc and *BCL2* in the simulated model should be read in that context. Serum zinc, like other micronutrient levels, can also fall as part of the acute-phase response, so that in real data low levels may be a marker of disease severity rather than a cause of poor outcome.

4.4 Vitamin C

Vitamin C had low individual importance in the model. Evidence for its role in TB is largely preclinical: it kills *M. tb* in vitro through an iron-dependent Fenton reaction [30] and potentiates first-line drugs in mice [31]. The conditional contribution seen in carriers of the *TNF* A allele is a model-derived pattern and should not be taken as support for a redox–genetic mechanism; testing such an idea would require measures of oxidative stress and relevant genes, such as *SOD2*, in real cohorts.

4.5 Implications for precision TB care and research

If gene–micronutrient interactions of this kind were confirmed in real data, they could inform three activities. First, trial design: adjunctive-nutrient trials could stratify or enrich by host genotype and baseline micronutrient status instead of testing supplementation in all patients, and simulation of the kind presented here can help to estimate the sample sizes needed to detect interactions. Second, nutritional programs: the findings are consistent with existing recommendations to assess and support

nutritional status in TB care [17,18], and could eventually help to target limited resources. Third, risk stratification: a validated model could in principle contribute to point-of-care risk scores, although this would require validation, calibration and an assessment of cost and feasibility of genotyping and micronutrient testing. These aims align with the precision-medicine agenda [40] and the research and innovation pillar of the End TB Strategy [41]. More generally, the study illustrates how synthetic data can enable early hypothesis generation where multimodal clinical datasets are unavailable [38,39].

4.6 Limitations

Several limitations must be stated clearly. (i) **Circularity.** The outcome was generated from rules that embed the hypothesized effects, so feature importance and SHAP patterns reflect the simulation assumptions as much as any biology. Real-world effect sizes, and whether the interactions exist at all, cannot be inferred from this study. (ii) **No external validation.** Performance was estimated on a single hold-out set of 300 simulated patients, without confidence intervals, calibration assessment or decision-analytic evaluation; in real data, performance would be expected to be lower. (iii) **Simplified biology.** Only two genes and three micronutrients were modeled. Global allele frequencies ignore population structure and linkage disequilibrium, and the model omits HIV and other co-infections, diabetes, smoking, drug resistance, adherence, disease extent and drug-metabolizing genotypes such as *NAT2*. (iv) **Temporal and causal issues.** Micronutrient levels measured at baseline may reflect inflammation or disease severity (reverse causation), and model-based scenarios such as those in Table 4 are not causal estimates of supplementation effects. (v) **Untested algorithmic claims.** Properties of XGBoost, such as its handling of sparse features and missing values, are attributes of the algorithm that are relevant to future clinical application, but robustness to incomplete patient profiles was not evaluated here. (vi) **Interpretation of explanations.** SHAP values describe model behavior and are sensitive to correlated features and model choice. Future work should follow reporting guidance for prediction models [42], use nested resampling for tuning and evaluation [48], and consider formal sample-size calculations for model development [49].

4.7 Future directions

The essential next step is validation in real patient cohorts with genotype, micronutrient and standardized outcome data, ideally with external validation and prospective evaluation. Subsequent models should extend the feature set to include *VDR* and *NAT2* variants, *SOD2* and other oxidative-stress genes, additional micronutrients such as vitamin A and selenium, and key clinical covariates (HIV status, diabetes, smoking, drug resistance and adherence). Causal-inference approaches and randomized trials stratified by genotype would then be needed to test whether any identified interaction is actionable.

5. Conclusions

This study shows that AI-driven pharmacogenomic modeling, combining gradient-boosted trees with SHAP interpretation, can recover and visualize non-linear interactions between host genotype and micronutrient status in a simulated PTB cohort, with XGBoost reaching an accuracy of 87.2% and an AUC-ROC of 0.91. In the simulation, vitamin D₃ status and *TNF* genotype were the central predictors,

with zinc and *BCL2* genotype as secondary contributors. Because outcomes were generated from assumed rules, the results establish feasibility and generate hypotheses, not clinical evidence. Validation in real cohorts is required before gene–micronutrient modeling can inform personalized TB care.

Declarations

Ethics approval: Not applicable. The study used only synthetic data and involved no human participants or biological samples.

Funding: No

Conflicts of interest: No

References

1. World Health Organization. Global tuberculosis report 2025. Geneva: World Health Organization; 2025.
2. Nahid P, et al. Official American Thoracic Society/Centers for Disease Control and Prevention/Infectious Diseases Society of America clinical practice guidelines: treatment of drug-susceptible tuberculosis. *Clin Infect Dis*. 2016;63(7):e147–e195.
3. Imperial MZ, et al. A patient-level pooled analysis of treatment-shortening regimens for drug-susceptible pulmonary tuberculosis. *Nat Med*. 2018;24(11):1708–1715.
4. Azuma J, et al. NAT2 genotype guided regimen reduces isoniazid-induced liver injury and early treatment failure in the 6-month four-drug standard treatment of tuberculosis: a randomized controlled trial for pharmacogenetics-based therapy. *Eur J Clin Pharmacol*. 2013;69(5):1091–1101.
5. Möller M, Hoal EG. Current findings, challenges and novel approaches in human genetic susceptibility to tuberculosis. *Tuberculosis (Edinb)*. 2010;90(2):71–83.
6. Abel L, et al. Human genetics of tuberculosis: a long and winding road. *Philos Trans R Soc Lond B Biol Sci*. 2014;369(1645):20130428.
7. Tobin DM, et al. Host genotype-specific therapies can optimize the inflammatory response to mycobacterial infections. *Cell*. 2012;148(3):434–446.
8. Flynn JL, et al. Tumor necrosis factor- α is required in the protective immune response against *Mycobacterium tuberculosis* in mice. *Immunity*. 1995;2(6):561–572.
9. Keane J, et al. Tuberculosis associated with infliximab, a tumor necrosis factor α -neutralizing agent. *N Engl J Med*. 2001;345(15):1098–1104.
10. Wilson AG, et al. Effects of a polymorphism in the human tumor necrosis factor α promoter on transcriptional activation. *Proc Natl Acad Sci U S A*. 1997;94(7):3195–3199.
11. Zhang Z, et al. Association between tumor necrosis factor alpha –238G/A polymorphism and tuberculosis susceptibility: a meta-analysis study. *BMC Infect Dis*. 2012;12:328.
12. Behar SM, et al. Apoptosis is an innate defense function of macrophages against *Mycobacterium tuberculosis*. *Mucosal Immunol*. 2011;4(3):279–287.
13. Sly LM, et al. Survival of *Mycobacterium tuberculosis* in host macrophages involves resistance to apoptosis dependent upon induction of antiapoptotic Bcl-2 family member Mcl-1. *J Immunol*. 2003;170(1):430–437.
14. He J, et al. Polymorphisms of the BCL2 gene associated with susceptibility to tuberculosis. *Rev Inst Med Trop Sao Paulo*. 2022;64:e59.
15. Nückel H, et al. Association of a novel regulatory polymorphism (–938C>A) in the BCL2 gene promoter with disease progression and survival in chronic lymphocytic leukemia. *Blood*. 2007;109(1):290–297.
16. Cegielski JP, McMurray DN. The relationship between malnutrition and tuberculosis: evidence from studies in humans and experimental animals. *Int J Tuberc Lung Dis*. 2004;8(3):286–298.

17. World Health Organization. Guideline: nutritional care and support for patients with tuberculosis. Geneva: World Health Organization; 2013.
18. Bhargava A, et al. Nutritional support for adult patients with microbiologically confirmed pulmonary tuberculosis: outcomes in a programmatic cohort nested within the RATIONS trial in Jharkhand, India. *Lancet Glob Health*. 2023;11(9):e1402–e1411.
19. Liu PT, et al. Toll-like receptor triggering of a vitamin D-mediated human antimicrobial response. *Science*. 2006;311(5768):1770–1773.
20. Liu PT, et al. Cutting edge: vitamin D-mediated human antimicrobial activity against *Mycobacterium tuberculosis* is dependent on the induction of cathelicidin. *J Immunol*. 2007;179(4):2060–2063.
21. Wilkinson RJ, et al. Influence of vitamin D deficiency and vitamin D receptor polymorphisms on tuberculosis among Gujarati Asians in west London: a case-control study. *Lancet*. 2000;355(9204):618–621.
22. Wejse C, et al. Vitamin D as supplementary treatment for tuberculosis: a double-blind, randomized, placebo-controlled trial. *Am J Respir Crit Care Med*. 2009;179(9):843–850.
23. Martineau AR, et al. High-dose vitamin D3 during intensive-phase antimicrobial treatment of pulmonary tuberculosis: a double-blind randomised controlled trial. *Lancet*. 2011;377(9761):242–250.
24. Jolliffe DA, et al. Adjunctive vitamin D in tuberculosis treatment: meta-analysis of individual participant data. *Eur Respir J*. 2019;53(3):1802003.
25. Ganmaa D, et al. Vitamin D supplements for prevention of tuberculosis infection and disease. *N Engl J Med*. 2020;383(4):359–368.
26. Prasad AS. Zinc in human health: effect of zinc on immune cells. *Mol Med*. 2008;14(5–6):353–357.
27. Truong-Tran AQ, et al. The role of zinc in caspase activation and apoptotic cell death. *Biometals*. 2001;14(3–4):315–330.
28. Botella H, et al. Mycobacterial P1-type ATPases mediate resistance to zinc poisoning in human macrophages. *Cell Host Microbe*. 2011;10(3):248–259.
29. Wagnew F, et al. Effects of zinc and vitamin A supplementation on prognostic markers and treatment outcomes of adults with pulmonary tuberculosis: a systematic review and meta-analysis. *BMJ Glob Health*. 2022;7(9):e008625.
30. Vilchèze C, et al. *Mycobacterium tuberculosis* is extraordinarily sensitive to killing by a vitamin C-induced Fenton reaction. *Nat Commun*. 2013;4:1881.
31. Vilchèze C, Kim J, Jacobs WR Jr. Vitamin C potentiates the killing of *Mycobacterium tuberculosis* by the first-line tuberculosis drugs isoniazid and rifampin in mice. *Antimicrob Agents Chemother*. 2018;62(3):e02165-17.
32. Cordell HJ. Detecting gene–gene interactions that underlie human diseases. *Nat Rev Genet*. 2009;10(6):392–404.
33. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5–32.
34. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: ACM; 2016. p. 785–794.
35. Grinsztajn L, Oyallon E, Varoquaux G. Why do tree-based models still outperform deep learning on typical tabular data? *Adv Neural Inf Process Syst (Datasets and Benchmarks Track)*. 2022;35.
36. Sauer CM, et al. Feature selection and prediction of treatment failure in tuberculosis. *PLoS One*. 2018;13(11):e0207491.
37. Kouchaki S, et al. Application of machine learning techniques to tuberculosis drug resistance analysis. *Bioinformatics*. 2019;35(13):2276–2282.
38. Chen RJ, et al. Synthetic data in machine learning for medicine and healthcare. *Nat Biomed Eng*. 2021;5(6):493–497.
39. Gonzales A, Guruswamy G, Smith SR. Synthetic data in health care: a narrative review. *PLOS Digit Health*. 2023;2(1):e0000082.

40. Collins FS, Varmus H. A new initiative on precision medicine. *N Engl J Med*. 2015;372(9):793–795.
41. Uplekar M, et al. WHO's new End TB Strategy. *Lancet*. 2015;385(9979):1799–1801.
42. Collins GS, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;385:e078378.
43. Pedregosa F, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 2011;12:2825–2830.
44. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst*. 2017;30:4765–4774.
45. Lundberg SM, et al. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell*. 2020;2(1):56–67.
46. 1000 Genomes Project Consortium; Auton A, et al. A global reference for human genetic variation. *Nature*. 2015;526(7571):68–74.
47. Xu L, et al. Modeling tabular data using conditional GAN. *Adv Neural Inf Process Syst*. 2019;32:7335–7345.
48. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*. 2006;7:91.
49. Riley RD, et al. Calculating the sample size required for developing a clinical prediction model. *BMJ*. 2020;368:m441.