Title of the Paper: Predicting over-the-counter antibiotic use in rural Pune, India, using machine learning methods

Author List and Affiliations

Names of Authors: Pravin Arun Sawant[1], Sakshi Shantanu Hiralkar[1], Yogita Purushottam Hulsurkar[1], Mugdha Sharad Phutane[1], Uma Satish Mahajan[1], Abhay Machindra Kudale[1]

Institutional Affiliation: [1]Department of Health Sciences, School of Health Sciences, Savitribai Phule Pune University, Pune, Maharashtra, India

i. Name of the first author: Dr. Pravin Arun Sawant
   - Email: pravinsawant0251@gmail.com

ii. Name of the second author: Sakshi Shantanu Hiralkar
   - Email: hiralkarsakshi@gmail.com

iii. Name of the third author: Yogita Purushottam Hulsurkar
   - Email: Yogitapurushottam95@gmail.com

iv. Name of the fourth author: Mugdha Sharad Phutane
   - Email: phutane.mug@gmail.com

v. Name of the fifth author: Dr Uma Satish Mahajan
   - Email: umasmahajan@gmail.com

vi. Name of the last and corresponding author: Dr Abhay Machindra Kudale
   - Email: abhay.kudale@gmail.com, amkudale@unipune.ac.in

vii. **Corresponding author and contact details:**

viii. Name of the Corresponding Author: Dr. Abhay Machindra Kudale
   - Designation: Head of the Department and Assistant Professor
   - Institute's location and Address: [1]Department of Health Sciences, School of Health Sciences, Savitribai Phule Pune University, Ganeshkhind, Pune - 411007, Maharashtra, INDIA
   - E-mail: abhay.kudale@gmail.com, amkudale@unipune.ac.in,
   - Mobile Phone Number: +91-9881435808, +91-9834774105
   - ORCID ID: https://orcid.org/0000-0003-2887-1636

ix. Running Title: Predicting OTC antibiotic use in India

Predicting over-the-counter antibiotic use in rural Pune, India, using machine learning methods

Abstract

OBJECTIVES: Over-the-counter (OTC) antibiotic use can cause antibiotic resistance, threatening global public health gains. To counter OTC use, this study used machine learning (ML) methods to identify predictors of OTC antibiotic use in rural Pune, India.

METHODS: The features of OTC antibiotic use were selected using stepwise logistic, lasso, random forest, XGBoost, and Boruta algorithms. Regression and tree-based models with all confirmed and tentatively important features were built to predict the use of OTC antibiotics. Five-fold cross-validation was used to tune the models' hyperparameters. The final model was selected based on the highest area under the curve (AUROC) with a 95% confidence interval and the lowest log-loss.

RESULTS: In rural Pune, the prevalence of OTC antibiotic use was 35.9% (95% CI, 31.56%-40.46%). The perception that buying medicines directly from a medicine shop/pharmacy is useful, using antibiotics for eye-related complaints, more household members consuming antibiotics, and longer duration and higher doses of antibiotic consumption in rural blocks and other social groups were confirmed as important features by the Boruta algorithm. The final model was the XGBoost+Boruta model with 7 predictors (AUROC=0.934; 95% CI, 0.8906-0.9782; log-loss=0.2793) log-loss.

CONCLUSIONS: XGBoost+Boruta, with 7 predictors, was the most accurate model for predicting OTC antibiotic use in rural Pune. Using OTC antibiotics for eye-related complaints, higher consumption of antibiotics and the perception that buying antibiotics directly from a medicine shop/pharmacy is useful were identified as key factors for planning interventions to improve awareness about proper antibiotic use.

Keywords: Antibiotic resistance, OTC antibiotic use predictor, Boruta, Lasso, Random forest, XGBoost

Introduction

The emergence of antimicrobial-resistant (AMR) bacterial species that are beyond the reach of medical treatment is a consequence of the over-the-counter (OTC) consumption of antibiotics in human and veterinary medicine (1–4). The misuse and overuse of antibiotics, along with self-medication, have accelerated the rise of AMR in bacteria. According to a WHO report, 50% of antibiotic prescriptions worldwide are inappropriate, with India being one of the largest consumers of these drugs (5–7). The prevalence of OTC antibiotic practices in India can be linked to its highly privatised healthcare infrastructure, informal sectors, and the widespread availability of retail medical stores that sell medicines without valid prescriptions (1). Previous studies have indicated that the high volume of antibiotic consumption in India (8) is associated with a lack of public knowledge, resource limitations in rural areas, the close proximity of retail pharmacies to the population, cultural practices, inadequate formal healthcare services, and a weak regulatory framework and law enforcement (1,2,4,9). In an effort to promote antibiotic stewardship, India has enacted the Drugs and Cosmetics Act (DCA), 1940, the Drugs and Cosmetics Rules (DCR), 1945, Schedule H1 (an amendment to Schedule H, 2014), and has launched a public awareness campaign known as "Medicines with the Red Line" (1,5,9). Despite these measures, the OTC sale of antibiotics continues to be a widespread practice in the country. Recently, Kerala became the first state in India to initiate Operation Amrith ("Antimicrobial Resistance Intervention for Total Health"). This operation involves conducting surprise inspections at retail medical shops to curb the OTC sale of antibiotics. Additionally, a toll-free number (18004253182) has been established for the public to report complaints against medical shops. Upon receiving a complaint, it is forwarded to the relevant zonal office for investigation, and prompt departmental action is taken if any violations are found (10).

As a step forward in antibiotic stewardship, global studies have utilised artificial intelligence (AI) and machine learning (ML) methods to predict AMR across various bacterial strains (11,12) and to assess the susceptibility of bacterial species to AMR, guiding antibiotic prescriptions with personalised antibiograms. After training with whole-genome sequencing data, several machine-learning algorithms, such as support vector machines (SVM), logistic regression (LR) models, and random forests (RF), have demonstrated high accuracy in predicting AMR (12,13). The efficacy of deep learning algorithms in identifying new antibiotics, AMR genes, and AMR peptides has also been recently established (14,15). Studies employing "off-the-shelf" supervised machine learning algorithms to create predictive models for antibiotic prescribing have yielded promising results, indicating that machine learning-

based solutions can offer essential tools to assist in antimicrobial prescribing and contribute to the fight against AMR (16,17,18). Despite these promising results in controlled environments (16,17,18), the current literature indicates that the application of predictive models to support clinical decisions in antibiotic prescribing and antimicrobial management remains limited and has not yet fully leveraged the significant advancements in data and algorithm development (11,16). The research has primarily relied on available secondary datasets for conducting AI and ML analyses, with very few studies situated in low- and middle-income countries, particularly in India.

However, in addition to hospital and laboratory settings, it is essential to implement antibiotic stewardship interventions in community settings. This approach recognises and addresses the behaviours and preferences of both community members and healthcare providers. Against this backdrop, our study sought to identify predictors of OTC antibiotic use in the rural areas of Pune district, India. By employing machine learning methods on a primary dataset, our study contributes to the identification of these predictors of OTC antibiotic use.

Methods

Study design

For primary data collection, a cross-sectional descriptive study was conducted in 2 blocks, Junnar and Mulshi, of Pune district, Maharashtra, to understand antibiotic usage. These blocks were selected based on their proximity to urban settings, with Junnar being distant and tribal, and Mulshi being closer to Pune City and rural.

Sampling

Pune district is divided into 2 rural sub-divisions. The first, Shirur, is relatively more distant from urban Pune and includes the Junnar, Ambegaon, Khed, and Shirur blocks. The second, Maval, is more accessible and comprises the Maval and Mulshi blocks. These 2 sub-divisions, consisting of 6 blocks, served as the sampling frame for our study. From these, 2 blocks—Junnar and Mulshi—were randomly selected. Within these blocks, a total of 23 villages were chosen: 12 from Junnar and 11 from Mulshi. These villages were selected based on their higher human and livestock populations, using a proportionate sampling approach that accounted for both human and animal population sizes.

Data collection

The study was approved by the Institutional Ethics Committee of Savitribai Phule Pune University (Ref. No. SPPU/IEC/2020/84). Data collection was conducted in 2 phases within the Pune district of Maharashtra State. The first phase included key informant interviews and focus group discussions. Based on the insights gained from the first phase, 3 distinct semi-structured interview schedules were developed for the second phase. This subsequent phase involved gathering both quantitative and qualitative data through semi-structured interviews to understand the perspectives of community members, farmers, and healthcare and veterinary care practitioners on antibiotic use.

Variables and datasets

The analysis utilised quantitative data from semi-structured interviews. The outcome or dependent variable, OTC antibiotic use, was defined as a binary variable. It was coded as 0 when doctors prescribed antibiotics and household members obtained them from the pharmacy, and as 1 when individuals purchased antibiotics from the pharmacy without a doctor's advice. This latter category included instances where antibiotics were self-purchased, used from an old prescription, shared by friends, neighbours, or relatives, or suggested and purchased at the pharmacy.

The analysis included a total of 29 predictor/independent variables, which encompassed (i) sociodemographic characteristics of the households, (ii) help-seeking behaviour, (iii) causes, duration, dosage, and the number of household members who used antibiotics in the past year, and (iv) knowledge, awareness, and perceptions about antibiotics. A detailed description of the predictor/independent variables can be found in Supplementary Material Table S1.

A total of 458 households participated in the survey. Following the exclusion of missing values and non-responses, 443 households remained for inclusion in the analysis. The dataset was randomly split into a training dataset (70% of cases, n=311) and a testing dataset (30% of cases, n=132) for the purpose of selecting predictors and developing machine learning models. We employed 5-fold cross-validation on the training dataset for hyperparameter tuning to minimise prediction error. The performance of the model was assessed using the testing dataset.

Statistical analysis

All analyses were conducted in R Studio using R version 4.2.3 (19). The exploratory data analysis utilised a complete dataset, with categorical variables described in terms of counts and percentages (%). To examine the association between categorical predictor variables and OTC antibiotic use, we applied the chi-square test of independence. We considered results statistically significant at $p \leq 0.05$. We calculated the estimated proportions of OTC antibiotic use and their 95% confidence intervals (CIs) using the method proposed by Agresti-Coull, which was implemented with the "prevalence" package (20). In the Agresti-Coull's CI formula,

$$(1)\ Agresti-Coull\ CI = \tilde{\pi} \pm Z\sqrt{\frac{\tilde{\pi}(1-\tilde{\pi})}{n+Z^2}} \qquad where\ \tilde{\pi} = \frac{y+\left(\frac{Z^2}{2}\right)}{n+Z^2}\ ,$$

$$Z = Z_{\alpha/2}\ \ and\ y = Proportion\ of\ OTC\ antibiotic\ use$$

Selecting predictors

The predictors of OTC antibiotic use were identified by applying logistic regression, the least absolute shrinkage and selection operator (lasso), and Boruta algorithms to the training dataset using the "Caret" package.

Logistic regression employs the Akaike Information Criterion (AIC) for stepwise predictor selection. It eliminates predictors with a p-value greater than 0.10 and compares the AIC of the reduced model at each step to the AIC of the preceding model. The variables that remain in the logistic regression model with the lowest AIC are considered the final predictors. The lasso algorithm, also referred to as L1 penalised/regularised regression, reduces the regression coefficients of unimportant variables to zero (21). The predictors/variables with non-zero coefficients of the lasso regression model were selected as the final predictors.

The Boruta algorithm, which is based on the RF approach, generates dummy, or shadow, variables corresponding to each of the dataset's original predictor or independent variables. It then employs a random forest classifier to compare the original predictors with their shadow counterparts using the mean decrease in accuracy and calculates z-scores. An equality test is used to compare the maximum z-score of the shadow predictors against that of the original predictors. If the z-score of an original predictor exceeds the maximum z-score of its shadow, the predictor is retained in the training dataset; otherwise, both the original and its shadow

predictor are removed from the dataset. This iterative process continues until all predictors are classified as "confirmed," "rejected," or "tentatively important" (25). The predictors identified by the Boruta algorithm as "confirmed important (cnf)" and "tentatively important (tntv)" are collectively referred to as "non-rejected predictors (nonrej)."

RF is an ensemble algorithm based on the "bagging" approach, which stands for "bootstrap averaging." It constructs multiple independent decision tree classifiers (ntree) using a subset of randomly selected variables and two-thirds of bootstrap sample data. The algorithm then validates the predictions with the remaining one-third of the data, known as "out-of-bag" data. RF combines the predictions from all the decision trees, which are trained in parallel, and determines the final predicted class of the outcome variable by the 'majority vote' of all the predictions (22). The extreme gradient boosting tree (XGBtree) algorithm is another ensemble method that enhances prediction accuracy through gradient boosting. Unlike RF, XGBtree builds decision tree classifiers sequentially, learning from the prediction errors of the preceding tree to minimise the error in the subsequent tree. The final prediction is the sum of all individual tree predictions (23,24). Both the RF and XGBtree algorithms utilise all available variables/predictors, and variable importance (VarImp) is crucial for understanding the significance of these variables/predictors in the model. However, to effectively plan targeted program intervention strategies to reduce the OTC use of antibiotics, it is essential to identify the most important predictors. Therefore, 3 sets of predictors were employed to develop the RF and XGBtree models: (i) all 29 predictors, (ii) non-rejected predictors (nonrej) selected using the Boruta algorithm, and (iii) confirmed important predictors (cnf) also selected using Boruta (26).

Developing predictive models

Initially, all 29 variables were included in the comprehensive logistic regression model, and the "glmStepAIC" method was employed for the stepwise selection of predictors. The model that yielded the lowest AIC was deemed the final model, and the predictors that remained were chosen as the final predictors. The hyperparameters of lasso (λ), RF (*mtry* and *ntree*), and XGBtree (*nrounds, max_depth, colsample_bytree, learning rate eta, gamma, min_child_weight,* and *subsample*) were tuned using cross-validation. The regression coefficients of the selected variables of stepwise logistic and lasso regression, the variable importance from RF and XGBtree, and the mean variable importance with decisions about predictors from the Boruta algorithm are reported. The training dataset was used for selecting

predictors, and 5-fold cross-validation was conducted to tune the hyperparameter of the models with selected predictors.

The selected predictors and the best-tuned hyperparameters were used to construct the StepLog and lasso regression models. The RF and XGBtree models were developed using 3 sets of predictors: all 29 predictors for RF and XGBtree; 9 non-rejected predictors for RF+Boruta(nonrej) and XGBtree+Boruta(nonrej); and 7 confirmed important predictors for RF+Boruta(cnf) and XGBtree+Boruta(cnf), each employing the optimally tuned hyperparameters. Model performance was assessed by calculating various metrics: the area under the receiver operating characteristic curve (AUROC) with a 95% confidence interval (CI) using the "pROC" package, log-loss, accuracy, sensitivity, specificity, F1-score, and balanced accuracy, using the "ConfusionTableR" package, all based on the test dataset.

$$(2)\ logloss = -\frac{1}{n}\sum_{i=1}^{n}[y_i * ln(P_i) + (1-y_i) * \ln(1-P_i)]$$

$$where\ y_i\ is\ actual\ OTC\ antibiotic\ use\ and\ P_i\ is\ predicted\ OTC\ antibiotic\ use$$

Confusion matrix

Predicted OTC antibiotic use	Actual OTC antibiotic use		Total
	Yes (1)	No (0)	
Yes (1)	A	b	a + b
No (0)	C	d	c + d
Total	a + c	b + d	n

(3) $$Accuracy = \frac{(a\ +\ d)}{n}$$

(4) $$Sensitivity = \frac{a}{(a\ + c)}$$

(5) $$Specificity = \frac{d}{(b + d)}$$

(6) $$F1 - score = \ 2 * \frac{a}{(a +\ b) + (a + c)}$$

(**7**) $$Balanced\ Accuracy = \frac{Sensitivity + Specificity}{2}$$

Results

The sociodemographic profile, along with knowledge and practices regarding OTC antibiotic use in households, is presented in Table 1.

{Table 1 Here}

Of the 443 households surveyed, 217 (49.0%) were from the tribal Junnar block and 226 (51.0%) from the rural Mulshi block of Pune district, respectively. In the rural areas of Pune district, the use of OTC antibiotics was 35.9% (95% CI: 31.56%-40.46%). The use of OTC antibiotics was significantly higher for complaints related to the ear, nose, and throat (ENT) at 53.3% (95% CI: 36.14%-69.77%), eyes at 53.6% (95% CI: 41.98%-64.89%), and gastrointestinal system (GIS) at 43.7% (95% CI: 32.74%-55.23%). Additionally, in households where more than 1 person used OTC antibiotics, the usage rate was 46.1% (95% CI: 36.09%-56.37%). A significant 39.9% (95% CI: 33.96%-46.19%) of households spent less than Rs. 200 on purchasing OTC antibiotics. Moreover, 62.5% (95% CI: 46.99%-75.82%) of households perceived that their health condition either did not improve or deteriorated after using antibiotics. Only 23.8% (95% CI: 15.91%-34.00%) of households were aware that not completing the prescribed antibiotic dosage could lead to a deterioration in health.

A strikingly large proportion of households, 97.5% (95% CI: 92.65%-99.47%), believed that the practice of buying antibiotics directly from the pharmacy was useful.

In the tribal block of Junnar, the use of OTC antibiotics was high, with 75.0% (95% CI: 40.09%-93.69%) for ENT complaints, 52.8% (95% CI: 37.00%-68.02%) for GIS issues, and 33.7% (95% CI: 25.28%-43.19%) for respiratory system-related complaints. In the rural block of Mulshi, OTC antibiotics were consumed by more than 1 person per household in 53.5% (95% CI: 38.91%-67.49%) of cases, for more than 10 days in 47.8% (95% CI: 36.25%-59.52%) of cases, and the use was highest at 69.4% (95% CI: 55.40%-80.56%) for eye-related complaints. In Junnar, 41.0% (95% CI: 29.52%-53.51%) of households reported that antibiotic medications were not affordable, and 35.6% (95% CI: 27.52%-44.57%) spent more than Rs 200 on purchasing these medicines. Meanwhile, in Mulshi, only one-fifth of the households reported the unaffordability of OTC antibiotics. In Junnar, 70.0% (95% CI: 47.87%-85.68%) of households perceived that their health condition was not cured or had deteriorated, 57.9% (95%

CI: 36.24%-76.89%) reported problems after consuming the medications, and only 18.5% (95% CI: 7.72%-37.16%) reported that purchasing medicines directly from medicine shops or pharmacies was not useful. However, more than 95% of households in both blocks believed that antibiotics are beneficial for human health.

The regression coefficients and the importance of predictors/features are shown in Table 2.

{Table 2 Here}

The perception that buying antibiotics directly from the pharmacy is useful was the most important predictor/feature across all 9 algorithms. Antibiotics used for eye-related complaints ranked as the second most significant predictor. The third most important predictor, according to regression and RF algorithms, was the greater distance of households from healthcare facilities; however, this was not supported by the Boruta algorithm. Rural blocks and membership in other social groups were deemed important by the Boruta algorithm. Additionally, the Boruta algorithm highlighted the significance of having more than 2 persons in a household consuming antibiotics, taking antibiotics for longer than 10 days, and administering more than 2 doses as important factors. Completing the prescribed antibiotic course was also considered a tentatively important feature by the Boruta argument. The stepwise LR (StepLog) and lasso regression algorithms identified 3 key features: assistance from government healthcare facilities, antibiotics used for respiratory complaints, and the general usefulness of antibiotics for humans as significant predictors. The Boruta algorithm distinguished 7 confirmed and 2 tentatively important features. The variable importance as determined by the Boruta algorithm is depicted in Figure 1.

{Figure 1 Here}

The results from evaluating the models' prediction performance are shown in Table 3.

{Table 3 Here}

The final StepLog model had an AIC of 168.52 and included 14 predictors. Its log-loss was 0.3781, which was higher than that of other prediction models, and it also had the lowest accuracy (0.8636), specificity (0.8526), F1-score (0.7857), and balanced accuracy (0.8723).

For the lasso model, the optimally tuned 'λ' was 0.0212, which utilised 9 predictors and achieved a log-loss of 0.3258. This model also had the highest sensitivity (0.9706) for predicting the use of OTC antibiotics. All RF models were set with ntree = 500. The mtry was 15 for the RF model with all predictors and 2 for the RF+Boruta model, which included 9 non-rejected and 7 confirmed important predictors. The best-tuned hyperparameters for all 3 XGBtree models were: nrounds at 100, max_depth at 20, eta at 0.1, gamma at 0, min_child_weight at 1, and subsample at 1. The hyperparameter colsample_bytree was set at 0.5, 0.7, and 0.8 for the XGBtree, XGBtree+Boruta(nonrej) model with 9 non-rejected predictors, and XGBtree+Boruta(cnf) model with 7 confirmed important predictors, respectively. The RF+Boruta(cnf) and XGBtree+Boruta(cnf) models, both with 7 confirmed important predictors, achieved the highest accuracy (0.9091), specificity (0.9091), and F1-score (0.8636) compared to the other models. The lasso model had the lowest AUROC at 0.902 (95% CI: 0.8326-0.9712). Overall, the StepLog model performed the worst among all the models considered. The XGBtree+Boruta(cnf) model with 7 confirmed important predictors demonstrated the best prediction performance, with the highest AUROC at 0.934 (95% CI: 0.8906-0.9782) and the lowest log-loss at 0.2793. Therefore, the XGBtree+Boruta(cnf) model with 7 confirmed important predictors was selected as the final model. The use of OTC antibiotics was predicted for individual households in the rural Pune district by applying this final model.

Discussion

Our study aimed to identify predictors of OTC antibiotic use in rural communities through the application of machine learning methods. To the best of our knowledge, this is the first study to employ ML methods to investigate predictors of OTC antibiotic use based on a primary dataset. To minimise geographical and demographic biases, we included multiple study sites, with one located near a city and another situated farther away.

Our study findings indicate that the most significant predictor of OTC antibiotic use was the belief that it is useful to purchase antibiotics directly from pharmacies. This behaviour underscores the cultural and socio-demographic closeness of pharmacists to the rural communities they serve, in contrast to medical doctors. The results also emphasise the need for regulatory interventions to curb over-the-counter antibiotic use, as outlined in Kerala State's AMR intervention program, Operation Amrith. Additionally, the use of antibiotics for eye-related and GIS complaints emerged as the second most significant predictor, likely due to the

higher prevalence of these conditions. In our analysis, the XGBtree+Boruta(cnf) model with 7 predictors was identified as the most accurate in terms of prediction performance. This model outperformed other approaches, including regression models (StepLog, lasso), RF (RF), XGBTree, RF+Boruta(nonrej) with 9 non-rejected predictors, and RF+Boruta(cnf) with 7 confirmed important predictors, as well as XGBtree+Boruta(nonrej) with 9 non-rejected predictors.

This study demonstrated the potential use of ML models for predicting OTC antibiotic use. ML models have proven to be helpful in the medical and health sciences, particularly in the areas of diagnosis and outcome prediction (27). Previous research has suggested that the application of ML models in the healthcare industry, although still in the early stages, is primarily focused on the early diagnosis of chronic diseases, predicting future disease incidence, conducting epidemiological studies, and facilitating evidence-based decision-making (27-32). There is also evidence supporting the use of AI and ML models to predict AMR among bacterial species based on whole genome sequencing (12,13,34-38). As part of antibiotic stewardship efforts, AI and ML have been employed to guide targeted empiric antibiotic prescribing (14, 39-41), profile and analyse drug resistance, and design targeted drug therapies (42,43) in pharmacometrics (44), and antibiotic discovery (45). Previously conducted studies in the health and medicine domains have employed several methods, including recursive decision tree-based models, XGBoost (46,47), a fuzzy logic model (48), ADABoost, RF, convolutional neural networks, SVM, logistic regression, lasso regression, and classification and regression trees (49-50).

As this study represents one of the initial attempts of its kind, we contend that employing AI and ML models can assist in the planning and enhancement of public health interventions in other states. This approach could mirror the successes of Operation Amrith in Kerala State (10), potentially increasing the novelty and impact of our study. Additionally, our findings highlight the imperative for more research into the patterns of OTC antibiotic usage that contribute to AMR. Such research should leverage AI and ML to inform targeted antibiotic therapies. Building on the results of our study, we advocate for further investigations that could guide the development of structured health interventions in rural Pune. There is also a pressing need for community-level health education interventions that focus on antibiotic stewardship and the broader implications of AMR.

Conclusions

Households that found the practice of purchasing medications directly from a pharmacy to be useful were more likely to consume antibiotics for eye-related complaints, engage in longer durations of antibiotic use, take higher doses of antibiotic medications, and have more household members using antibiotics in rural blocks and other social groups. These factors were confirmed as significant predictors of OTC antibiotic use. The XGBtree ML algorithm in conjunction with the Boruta feature selection method, which identified 7 significant predictors, emerged as the best model with the lowest prediction error. Predictions of OTC antibiotic use for individual households can be instrumental in devising intervention strategies aimed at curbing the non-prescription use of antibiotics in the rural areas of Pune district, Maharashtra.

Acknowledgements: We gratefully acknowledge our institute, the Department of Health Sciences under the School of Health Sciences at Savitribai Phule Pune University, for allowing us to conduct this study. We remain grateful to the general community members and farmers from the Junnar and Mulshi Blocks of Pune District who participated in this study.

Source(s) of Support and Funding: This study was funded by the Indian Council of Medical Research, New Delhi, India's Epidemiology and Communicable Diseases Division (Ad hoc Project Proposal ID 240/2020-ECD-II).

Conflict of Interest Statement: None declared

References

1. Nafade V, Huddart S, Sulis G, Daftary A, Miraj SS, Saravu K, et al. Over-the-counter antibiotic dispensing by pharmacies: A standardised patient study in Udupi district, India. BMJ Glob Heal. 2019;4(6):1–9.
2. Adhikari B, Pokharel S, Raut S, Adhikari J, Thapa S, Paudel K, et al. Why do people purchase antibiotics over-the-counter? A qualitative study with patients, clinicians and dispensers in central, eastern and western Nepal. BMJ Glob Heal. 2021;6(5).
3. Jacobs TG, Robertson J, Van Den Ham HA, Iwamoto K, Bak Pedersen H, Mantel-Teeuwisse AK. Assessing the impact of law enforcement to reduce over-the-counter (OTC) sales of antibiotics in low- And middle-income countries; A systematic literature review. BMC Health Serv Res. 2019;19(1):1–15.
4. Porter G, Kotwani A, Bhullar L, Joshi J. Over-the-counter sales of antibiotics for human use in India: The challenges and opportunities for regulation. Med Law Int.

2021;21(2):147–73.

5. Kotwani A, Joshi J, Lamkang AS, Sharma A, Kaloni D. Knowledge and behaviour of consumers towards the non-prescription purchase of antibiotics: An insight from a qualitative study from New Delhi, India. Pharm Pract (Granada). 2021;19(1):1–11.
6. Blondeau LD, Blondeau JM. Antimicrobial Resistance. Diagnostics Ther Vet Dermatology. 2021;163–74.
7. Franco BE, Mart ́nez MA, S ́anchez Rodr ́guez MA, Wertheimer AI. The determinants of the antibiotic resistance process. Infect Drug Resist. 2009;2(1):1–11.
8. Nayiga S, Kayendeke M, Nabirye C, Willis LD, Chandler CIR, Staedke SG. Use of antibiotics to treat humans and animals in Uganda: A cross-sectional survey of households and farmers in rural, urban and peri-urban settings. JAC-Antimicrobial Resist. 2020;2(4):1–11.
9. Kotwani A, Joshi J, Lamkang AS. Over-the-counter sale of antibiotics in India: A qualitative study of providers' perspectives across 2 states. Antibiotics. 2021;10(9):1–19.
10. Arora G. Kerala comes up with Operation AMRITH to tackle AMR [Internet]. 2024 [cited 2024 Jan 26]. Available from: https://www.downtoearth.org.in/news/health/kerala-comes-up-with-operation-amrith-to-tackle-amr-93865.
11. Rabaan, A.A.; Alhumaid, S.; Mutair, A.A.; Garout, M.; Abulhamayel, Y.; Halwani, M.A.; Alestad, J.H.; Bshabshe, A.A.; Sulaiman, T.; AlFonaisan, M.K.; et al. Application of Artificial Intelligence in Combating High Antimicrobial Resistance Rates. Antibiotics 2022, 11, 784. https://doi.org/10.3390/antibiotics11060784
12. Liu, Z., Deng, D., Lu, H., Sun, J., Lv, L., Li, S., et al. (2020). Evaluation of machine learning models for predicting antimicrobial resistance of Actinobacillus pleuropneumoniae from whole genome sequences. Front. Microbiol. 11:48. doi: 10.3389/FMICB.2020.00048/FULL
13. Yang, Y., Niehaus, K., Walker, T., Iqbal, Z., Walker, S., Wilson, D., et al. (2018). Machine learning for classifying tuberculosis drug-resistance from DNA sequencing data. Bioinformatics 34, 1666–1671. doi: 10.1093/bioinformatics/btx801
14. Arango-Argoty, G., Garner, E., Pruden, A., Heath, L. S., Vikesland, P., and Zhang, L. (2018). Deep ARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. Microbiome 6:23. doi: 10.1186/S40168-018-0401-Z
15. Stokes, J., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N. M., et al.

(2020). A deep learning approach to antibiotic discovery. Cells 180, 688–702.e13. doi: 10.1016/j.cell.2020.01.021

16. M. Oonsivilai, Y. Mo, N. Luangasanatip, Y. Lubell, T. Miliya, P. Tan, L. Loeuk, P. Turner, and B. S. Cooper, ``Using machine learning to guide targeted and locally-tailored empiric antibiotic prescribing in a children's hospital in Cambodia," Wellcome Open Res., vol. 3, p. 131, Oct. 2018.
17. S. Mart ínez-Ag üero, I. Mora-Jim énez, J. L érida-Garc á, J. Álvarez-Rodr íguez, and C. Soguero-Ruiz, ``Machine learning techniques to identify antimicrobial resistance in the intensive care unit," Entropy, vol. 21, no. 6, p. 603, Jun. 2019. [Online]. Available: https://www.mdpi.com/1099-4300/21/6/603
18. G. Feretzakis, E. Loupelis, A. Sakagianni, D. Kalles, M. Martsoukou, M. Lada, N. Skarmoutsou, C. Christopoulos, K. Valakis, A. Velentza, S. Petropoulou, S. Michelidou, and K. Alexiou, ``Using machine learning techniques to aid empirical antibiotic therapy decisions in the intensive care unit of a general hospital in Greece," Antibiotics, vol. 9, no. 2, p. 50, Jan. 2020.
19. R: The R Project for Statistical Computing [Internet]. [cited 2023 Aug 5]. Available from: https://www.r-project.org/
20. Brown LD, Cai TT, Das Gupta A. Interval Estimation for a Binomial Proportion. https://doi.org/101214/ss/1009213286 [Internet]. 2001 May 1 [cited 2023 Aug 5];16(2):101–33. Available from: https://projecteuclid.org/journals/statistical-science/volume-16/issue-2/Interval-Estimation-for-a-Binomial-Proportion/10.1214/ss/1009213286.full
21. Tibshirani R. Regression Shrinkage and Selection Via the Lasso. J R Stat Soc Ser B [Internet]. 1996 Jan 1 [cited 2023 Aug 5];58(1):267–88. Available from: https://onlinelibrary.wiley.com/doi/full/10.1111/j.2517-6161.1996.tb02080.x
22. Breiman L. Random Forests. 2001;45:5–32.
23. Friedman JH. Greedy function approximation: A gradient boosting machine. Ann Stat. 2001;29(5):1189–232.
24. Viljanen M, Meijerink L, Zwakhals L, van de Kassteele J. A machine learning approach to small area estimation: predicting the health, housing and well-being of the population of Netherlands. Int J Health Geogr [Internet]. 2022;21(1):1–18. Available from: https://doi.org/10.1186/s12942-022-00304-5
25. Kursa MB, Rudnicki WR. Feature selection with the boruta package. J Stat Softw. 2010;36(11):1–13.

26. Mahajan, U., Krishnan, A., Malhotra, V., Sharma, D., Gore, S. (2021). Machine Learning Feature Selection in Archery Performance. In Advances in Signal and Data Processing (pp. 561-573). Springer, Singapore. https://doi.org/10.1007/978-981-15-8391-9_41
27. Fernández-Delgado M, Cernadas E, Barro S, Amorim D. Do we need hundreds of classifiers to solve real-world classification problems? J Mach Learn Res. 2014;15:3133–81.
28. Harrison CJ, Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction to natural language processing. BMC Med Res Methodol. 2021;21(1):1–18.
29. Wiemken TL, Kelley RR. Machine learning in epidemiology and health outcomes research. Annu Rev Public Health. 2019;41:21–36.
30. Toubiana L, Griffon N. Some Innovative Approaches for Public Health and Epidemiology Informatics. Yearb Med Inform. 2016;(1):247–50.
31. Diallo G, Bordea G. Novelty in Public Health and Epidemiology Informatics. Yearb Med Inform. 2022;31(1):273–5.
32. Poudel GR, Barnett A, Akram M, Martino E, Knibbs LD, Anstey KJ, et al. Machine Learning for Prediction of Cognitive Health in Adults Using Sociodemographic, Neighbourhood Environmental, and Lifestyle Factors. Int J Environ Res Public Health. 2022;19(17).
33. Chen J, Zhou J, Yang F. Analysis of Characteristic Factors of Nursing Safety Incidents in ENT Surgery by Deep Learning-Based Medical Data Association Rules Method. Comput Math Methods Med. 2022;2022.
34. Peng Z, Maciel-Guerra A, Baker M, Zhang X, Hu Y, Wang W, et al. Whole-genome sequencing and gene sharing network analysis powered by machine learning identifies antibiotic resistance sharing between animals, humans and environment in livestock farming [Internet]. Vol. 18, PLoS Computational Biology. 2022. 1–38 p. Available from: http://dx.doi.org/10.1371/journal.pcbi.1010018
35. Ren Y, Chakraborty T, Doijad S, Falgenhauer L, Falgenhauer J, Goesmann A, et al. Prediction of antimicrobial resistance based on whole-genome sequencing and machine learning. Bioinformatics. 2022;38(2):325–34.
36. Santerre JW, Davis JJ, Xia F, Stevens R. Machine Learning for Antimicrobial Resistance. 2016;35(3):1–22. Available from: http://arxiv.org/abs/1607.01224
37. Richards, et al. 乳鼠心肌提取 HHS Public Access. Physiol Behav. 2018;176(5):139–

48.

38. Imchen M, Moopantakath J, Kumavath R, Barh D, Tiwari S, Ghosh P, et al. Current Trends in Experimental and Computational Approaches to Combat Antimicrobial Resistance. Front Genet. 2020;11(November).
39. Oonsivilai M, Mo Y, Luangasanatip N, Lubell Y, Miliya T, Tan P, et al. Using machine learning to guide targeted and locally-tailored empiric antibiotic prescribing in a children's hospital in Cambodia [version 1; referees: 2 approved]. Wellcome Open Res. 2018;3(0):1–18.
40. Corbin CK, Sung L, Chattopadhyay A, Noshad M, Chang A, Deresinksi S, et al. Personalised antibiograms for machine learning driven antibiotic selection. Commun Med. 2022;2(1).
41. Feretzakis G, Loupelis E, Sakagianni A, Kalles D, Martsoukou M, Lada M, et al. Using machine learning techniques to aid empirical antibiotic therapy decisions in the intensive care unit of a general hospital in Greece. Antibiotics. 2020;9(2).
42. Kouchaki S, Yang YY, Walker TM, Walker AS, Wilson DJ, Peto TEA, et al. Application of machine learning techniques to tuberculosis drug resistance analysis. Bioinformatics. 2019;35(13):2276–82.
43. Sharma A, Machado E, Lima KVB, Suffys PN, Concei ção EC. Tuberculosis drug resistance profiling based on machine learning: A literature review. Brazilian J Infect Dis. 2022;26(1):1–9.
44. Wilkins JJ, Svensson EM, Ernest JP, Savic RM, Simonsson USH, McIlleron H. Pharmacometrics in tuberculosis: progress and opportunities. Int J Antimicrob Agents [Internet]. 2022;60(3):106620. Available from: https://doi.org/10.1016/j.ijantimicag.2022.106620
45. David L, Brata AM, Mogosan C, Pop C, Czako Z, Muresan L, et al. Artificial intelligence and antibiotic discovery. Antibiotics. 2021;10(11):1–13.
46. Sakagianni A, Koufopoulou C, Feretzakis G, Kalles D, Verykios VS, Myrianthefs P, et al. Using Machine Learning to Predict Antimicrobial Resistance—A Literature Review. Antibiotics. 2023;12(3):1–18.
47. Yasir M, Karim AM, Malik SK, Bajaffer AA, Azhar EI. Application of Decision-Tree-Based Machine Learning Algorithms for Prediction of Antimicrobial Resistance. Antibiotics. 2022;11(11).
48. Maviglia R, Michi T, Passaro D, Raggi V, Bocci MG, Piervincenzi E, et al. Machine Learning and Antibiotic Management. Antibiotics. 2022;11(3):1–15.

49. Anahtar MN, Yang JH, Kanjilal S. Applications of Machine Learning to the Problem of Antimicrobial Resistance: an Emerging Model for Translational Research. J Clin Microbiol. 2021;59(7):1–14.
50. Wong JG, Aung AH, Lian W, Lye DC, Ooi CK, Chow A. Risk prediction models to guide antibiotic prescribing: a study on adult patients with uncomplicated upper respiratory tract infections in an emergency department. Antimicrob Resist Infect Control. 2020;9(1):1–11.

.

Supplementary material

Table S1: List of predictor/independent variables

Variable name	Type of variable	Variable description
f1_Block2	Binary	Rural
f2_Others	Binary	Social group - Others
f3_2gt10.000	Binary	Monthly family income > Rs. 10.000
f4_Collective.decision	Binary	Healthcare decision - Collective decision
f4_Self	Binary	Healthcare decision - Self
f4_Spouse	Binary	Healthcare decision - Spouse
f5_Govt	Binary	Help from Government healthcare facilities
f5_Pvt	Binary	Help from Private healthcare facilities
f6_2gt5.km	Binary	Distance of healthcare facility > 5 km
ent_Yes	Binary	Antibiotics used for ENT
eyes_Yes	Binary	Antibiotics used for Eyes
gis_Yes	Binary	Antibiotics used for Gastro-intestine system
injuryaccident_Yes	Binary	Antibiotics used for Injury or accident
musculoskeletal_Yes	Binary	Antibiotics used for Musculoskeletal
respiratory_Yes	Binary	Antibiotics used for Respiratory system (RTI/URTI)
surgery_Yes	Binary	Antibiotics used for Surgery
persons_2gt1persons	Binary	Total no. of persons consumed antibiotics >1 person
days_6to10days	Binary	Total no of days antibiotics consumed: 6 to 10 days
days_2gt10days	Binary	Total no of days antibiotics consumed > 10 days
dose_2gt2doses	Binary	Total no of tablets/syrups of antibiotics consumed > 2 doses
f7_Yes	Binary	Antibiotics medicines were affordable - Yes
f8_2gt200	Binary	Overall money spent on purchasing antibiotic medicines > Rs. 200
f9_Notcured_Deteriorated	Binary	Perceived effect of antibiotic medicines on health outcome - Not cured/ deteriorated
f10_Yes	Binary	Problems after consuming medicines - Yes
f11_Yes	Binary	Completed dose of antibiotic medicine prescribed by doctor - Yes
f12_Notaware	Binary	Effects/consequences for an incomplete dose of antibiotic medicines - Not aware
f12_Notfullyrecovered	Binary	Effects/consequences for incomplete dose of antibiotic medicines - Not fully recovered
f13_Useful	Binary	Practice of buying medicines directly from medicine shop/pharmacy - Useful
f14_Useful	Binary	Antibiotics are useful for human beings - Useful

Table 1: Socio-demographic characteristics, reasons for antibiotic consumption, and knowledge and awareness about OTC antibiotic use

Predictors	OTC-antibiotics use									
	Total participants				Junnar (distant/tribal)			Mulshi (nearer/rural)		
	N	n	% (95% CI)	p-value	N	n	% (95% CI)	N	n	% (95% CI)
Total	443	159	35.9 (31.56, 40.46)							
Block				0.568						
Junnar (distant/tribal)	217	75	34.6 (28.55, 41.11)		217	75	34.6 (28.55, 41.11)			
Mulshi (nearer/rural)	226	84	37.2 (31.13, 43.64)					226	84	37.2 (31.13, 43.64)
Community group				0.624						
General	216	80	37.0 (30.87, 43.65)		17	6	35.3 (17.17, 58.84)	199	74	37.2 (30.77, 44.09)
Others	227	79	34.8 (28.90, 41.21)		200	69	34.5 (28.25, 41.33)	27	10	37.0 (21.47, 55.84)
Monthly family income (Rs)				0.413						
Up to Rs 10,000	352	123	34.9 (30.15, 40.06)		183	61	33.3 (26.90, 40.45)	169	62	36.7 (29.79, 44.18)
More than Rs 10,000	91	36	39.6 (30.12, 49.84)		34	14	41.2 (26.34, 57.80)	57	22	38.6 (27.04, 51.59)
Responsibility for healthcare decisions				0.188						
Self	150	47	31.3 (24.44, 39.16)		85	25	29.4 (20.75, 39.86)	65	22	33.8 (23.50, 46.00)
Spouse	128	43	33.6 (25.98, 42.16)		60	20	33.3 (22.69, 45.98)	68	23	33.8 (23.68, 45.69)
Close family members	113	45	39.8 (31.27, 49.05)		53	20	37.7 (25.91, 51.22)	60	25	41.7 (30.05, 54.28)
Collective decision	52	24	46.2 (33.34, 59.50)		19	10	52.6 (31.70, 72.67)	33	14	42.4 (27.22, 59.21)
Help from healthcare facilities				0.240						
Government	52	14	26.9 (16.67, 40.35)		29	6	20.7 (9.49, 38.75)	23	8	34.8 (18.70, 55.22)
Private	293	105	35.8 (30.56, 41.48)		115	38	33.0 (25.10, 42.08)	178	67	37.6 (30.85, 44.95)
Both government and private	98	40	40.8 (31.60, 50.72)		73	31	42.5 (31.78, 53.91)	25	9	36 (20.16, 55.57)
Distance of healthcare facility				0.295						
Up to 5 km	145	57	39.3 (31.73, 47.44)		54	18	33.3 (22.19, 46.69)	91	39	42.9 (33.18, 53.11)
More than 5 km	298	102	34.2 (29.07, 39.79)		163	57	35.0 (28.06, 42.57)	135	45	33.3 (25.93, 41.66)

*Note: Responses of "yes" for OTC antibiotic use are shown in the table. Percentages are calculated as n*100/N.*
OTC, over the counter.

Table 1: Socio-demographic characteristics, reasons for antibiotic consumption, and knowledge and awareness about OTC antibiotic use (cont.)

Predictors	OTC-antibiotics use									
	Total participants				Junnar (distant/tribal)			Mulshi (nearer/rural)		
	N	n	% (95% CI)	p-value	N	n	% (95% CI)	N	n	% (95% CI)
Antibiotics used for complaints										
Ear-nose-throat	30	16	53.3 (36.14, 69.77)	0.039	8	6	75.0 (40.09, 93.69)	22	10	45.5 (26.91, 65.35)
Eyes	69	37	53.6 (41.98, 64.89)	0.001	20	3	15.0 (4.39, 36.88)	49	34	69.4 (55.40, 80.56)
Gastro-intestinal system	71	31	43.7 (32.74, 55.23)	0.136	36	19	52.8 (37.00, 68.02)	35	12	34.3 (20.76, 50.92)
Injury or accident	19	8	42.1 (23.11, 63.76)	0.564	11	6	54.5 (27.99, 78.75)	8	2	25.0 (6.31, 59.91)
Musculoskeletal disorders	24	6	25.0 (11.69, 45.21)	0.253	14	4	28.6 (11.34, 55.03)	10	2	20.0 (4.59, 52.06)
Respiratory system	181	56	30.9 (24.65, 38.02)	0.071	104	35	33.7 (25.28, 43.19)	77	21	27.3 (18.53, 38.18)
Surgery	50	14	28.0 (17.38, 41.76)	0.217	20	4	20.0 (7.49, 42.18)	30	10	33.3 (19.13, 51.32)
Total no. of persons who consumed antibiotics				0.025						
Only 1 person	354	118	33.3 (28.62, 38.40)		171	57	33.3 (26.69, 40.71)	183	61	33.3 (26.90, 40.45)
More than 1 person	89	41	46.1 (36.09, 56.37)		46	18	39.1 (26.37, 53.57)	43	23	53.5 (38.91, 67.49)
Total no. of days antibiotics consumed				0.527						
Up to 5 days	239	84	35.1 (29.37, 41.40)		117	47	40.2 (31.73, 49.24)	122	37	30.3 (22.85, 39.00)
6 to 10 days	88	36	40.9 (31.22, 51.36)		51	21	41.2 (28.74, 54.85)	37	15	40.5 (26.32, 56.54)
More than 10 days	116	39	33.6 (25.66, 42.64)		49	7	14.3 (6.78, 26.98)	67	32	47.8 (36.25, 59.52)
Total no. of tablets/ syrups of antibiotics consumed				0.375						
1 to 2 doses per day	315	109	34.6 (29.56, 40.02)		147	52	35.4 (28.10, 43.39)	168	57	33.9 (27.19, 41.38)
More than 2 doses per day	128	50	39.1 (31.04, 47.72)		70	23	32.9 (22.96, 44.53)	58	27	46.6 (34.33, 59.20)
Antibiotic medicines are affordable				0.258						
No	120	38	31.7 (24.00, 40.47)		61	25	41.0 (29.52, 53.51)	59	13	22.0 (13.22, 34.26)
Yes	323	121	37.5 (32.36, 42.86)		156	50	32.1 (25.22, 39.74)	167	71	42.5 (35.27, 50.10)
Overall money spent on purchasing antibiotic medicines				0.052						
Up to Rs 200/-	243	97	39.9 (33.96, 46.19)		99	33	33.3 (24.80, 43.11)	144	64	44.4 (36.58, 52.60)
More than Rs 200/-	200	62	31.0 (24.99, 37.73)		118	42	35.6 (27.52, 44.57)	82	20	24.4 (16.31, 34.76)

*Note: Responses of "yes" for OTC antibiotic use are shown in the table. Percentages are calculated as n*100/N.*

Table 1: Socio-demographic characteristics, reasons for antibiotic consumption, and knowledge and awareness about OTC antibiotic use (cont.)

Predictors	OTC-antibiotics use									
	Total participants				Junnar (distant/tribal)			Mulshi (nearer/rural)		
	N	n	% (95% CI)	p-value	N	n	% (95% CI)	N	n	% (95% CI)
Perceived effect of antibiotic medicines on health outcomes				<0.0001						
Cured	403	134	33.3 (28.83, 37.99)		197	61	31.0 (24.91, 37.74)	206	73	35.4 (29.22, 42.19)
Not cured/deteriorated	40	25	62.5 (46.99, 75.82)		20	14	70.0 (47.87, 85.68)	20	11	55.0 (34.19, 74.19)
Problems after consuming medicines				0.208						
No	403	141	35.0 (30.49, 39.77)		198	64	32.3 (26.19, 39.13)	205	77	37.6 (31.21, 44.37)
Yes	40	18	45.0 (30.70, 60.18)		19	11	57.9 (36.24, 76.89)	21	7	33.3 (17.05, 54.78)
Completed dose of medicine prescribed by doctor				0.213						
No	156	62	39.7 (32.40, 47.58)		57	24	42.1 (30.18, 55.03)	99	38	38.4 (29.40, 48.24)
Yes	287	97	33.8 (28.57, 39.46)		160	51	31.9 (25.14, 39.46)	127	46	36.2 (28.37, 44.88)
Effects/consequences of not completing dose of medicine				0.034						
Incomplete recovery	132	53	40.2 (32.18, 48.68)		72	28	38.9 (28.45, 50.45)	60	25	41.7 (30.05, 54.28)
Health deterioration, partially effective, Antibiotic resistance	84	20	23.8 (15.91, 34.00)		50	12	24.0 (14.16, 37.55)	34	8	23.5 (12.20, 40.23)
Not aware	227	86	37.9 (31.82, 44.35)		95	35	36.8 (27.82, 46.89)	132	51	38.6 (30.76, 47.16)
Perception of buying medicines directly from medicine shop/pharmacy				<0.0001						
Not useful	322	41	12.7 (9.50, 16.84)		27	5	18.5 (7.72, 37.16)	46	17	37.0 (24.49, 51.43)
Useful	121	118	97.5 (92.65, 99.47)		190	70	36.8 (30.30, 43.90)	180	67	37.2 (30.49, 44.49)
Antibiotics are beneficial for human beings				0.262						
Not beneficial	73	22	30.1 (20.78, 41.48)		153	14	9.2 (5.42, 14.88)	169	27	16.0 (11.17, 22.30)
Beneficial	370	137	37.0 (32.26, 42.06)		64	61	95.3 (86.57, 98.92)	57	57	100.0 (94.57, 100.0)

*Note: Responses of "yes" for OTC antibiotic use are shown in the table. Percentages are calculated as n*100/N.*

Table 2: Predictor/feature importance by various machine learning methods for predicting OTC antibiotic use in rural Pune, India

Features	Full logistic	Step wise logistic	Lasso	Boruta		Random forest			XGBtree		
	Regression coefficients			Mean Imp	Decision	RF (All variables)	RF + Boruta (9 nonrej variables)	RF + Boruta (7 cnf variables)	XGBtree (All variables)	XGBtree + Boruta (9 nonrej variables)	XGBtree + Boruta (7cnf variables)
(Intercept)	-1.30	-0.55	-1.55								
Mulshi (nearer/rural)	0.67			4.17	Cnf	8.22	2.07	4.69	3.49	1.99	1.23
Social group - Others	0.75			4.33	Cnf	6.10	4.88	1.69	3.17	2.73	0.09
Monthly family income > Rs. 10.000	0.45			1.06	Rej	3.57			2.06		
Healthcare decision - Collective decision	-0.09			-1.13	Rej	0.65			0.72		
Healthcare decision - Self	0.05			-0.80	Rej	3.68			2.52		
Healthcare decision - Spouse	-0.76	-0.73		-0.09	Rej	1.25			1.10		
Help from government healthcare facilities	-2.03	-2.01	-0.12	0.40	Rej	3.90			0.54		
Help from private healthcare facilities	-0.84	-0.92		0.66	Rej	3.22			2.88		
Distance of healthcare facility > 5 km	0.87	0.85	0.01	0.81	Rej	8.04			3.45		
Antibiotics used for ear-nose-throat	-1.02		0.00	0.61	Rej	1.23			0.39		
Antibiotics used for eyes	1.72	1.79	0.85	12.00	Cnf	12.86	14.67	17.42	3.55	1.94	1.53
Antibiotics used for gastro-intestinal system	0.64	1.13	0.20	3.35	Tntv	6.66	2.43		2.36	1.44	
Antibiotics used for injury or accident	-2.46	-2.34	0.00	-0.62	Rej	2.49			0.00		
Antibiotics used for musculoskeletal disorders	-1.40		0.00	0.88	Rej	1.28			0.18		
Antibiotics used for respiratory system	-1.49	-1.00	-0.29	0.69	Rej	6.17			1.60		
Antibiotics used for surgery	0.15		0.00	-0.95	Rej	1.09			0.73		
Total no. of persons who consumed antibiotics – >1 person	0.49		0.37	5.61	Cnf	3.84	2.77	0.00	4.16	1.06	0.00
Total no. of days antibiotics consumed - 6 to 10 days	0.84	1.20	0.00	-0.14	Rej	2.35			0.96		
Total no. of days antibiotics consumed > 10 days	0.58	1.16	0.00	5.39	Cnf	5.04	0.04	3.61	1.94	0.00	0.70
Total no. of tablets/syrups of antibiotics consumed > 2 doses	0.53		0.00	5.03	Cnf	6.75	0.00	3.24	2.21	1.90	1.23

OTC, over the counter; RF, random forest; XGBtree, extreme gradient boosting tree; Cnf, confirmed important; Rej, rejected; Tntv, tentatively important, nonrej, non-rejected (7 cnf + 2 Tntv variables = 9 variables).

Nine nonrej variables were selected by Boruta: f1_Block2, f2_Others, eyes_Yes, persons_2gt1persons, days_2gt10days, dose_2gt2doses, f13_Useful, f11_Yes, gis_Yes.

Seven Cnf variables were selected by Boruta: f1_Block2, f2_Others, eyes_Yes, persons_2gt1persons, days_2gt10days, dose_2gt2doses, f13_Useful.

Table 2: Predictor/ feature importance by various machine learning models for predicting OTC antibiotic use in rural Pune, India (cont.)

Features	Full logistic	Step wise logistic	Lasso	Boruta		Random forest			XGBtree		
	Regression coefficients			Mean Imp	Decision	RF (All variables)	RF + Boruta (9 nonrej variables)	RF + Boruta (7 cnf variables)	XGBtree (All variables)	XGBtree + Boruta (9 nonrej variables)	XGBtree + Boruta (7 cnf variables)
Antibiotics medicines were affordable - Yes	0.01			0.54	Rej	0.00			1.95		
Overall money spent on purchasing antibiotic medicines > Rs. 200	-0.07			1.35	Rej	1.25			3.20		
Perceived effect of antibiotic medicines on health outcome - Not cured/deteriorated	-1.48	-1.54		1.15	Rej	1.11			0.15		
Problems after consuming antibiotic medicines - Yes	-0.03		0.00	0.56	Rej	2.39			0.39		
Completed dose of antibiotic medicine prescribed by doctor - Yes	-0.97	-0.88	-0.02	3.55	Tntv	6.71	5.21		2.58	1.64	
Effects/consequences of incomplete dose of antibiotic medicines - Not aware	0.40		0.00	-0.49	Rej	3.84			2.89		
Effects/consequences of incomplete dose of antibiotic medicines – Incomplete recovery	0.24		0.00	-0.21	Rej	3.84			1.40		
Perception of buying medicines directly from medicine shop/pharmacy - Useful	8.19	7.99	4.85	77.92	Cnf	100.00	100.00	100.00	100.00	100.00	100.00
Antibiotics are beneficial for human beings - Beneficial	-2.02	-2.03	-0.59	-0.28	Rej	0.71			2.86		

OTC, over the counter; RF, random forest; XGBtree, extreme gradient boosting tree; Cnf, confirmed important; Rej, rejected; Tntv, tentatively important, nonrej, non-rejected (7 cnf + 2 Tntv variables = 9 variables).

Nine nonrej variables were selected by Boruta: f1_Block2, f2_Others, eyes_Yes, persons_2gt1persons, days_2gt10days, dose_2gt2doses, f13_Useful, f11_Yes, gis_Yes.

Seven Cnf variables were selected by Boruta: f1_Block2, f2_Others, eyes_Yes, persons_2gt1persons, days_2gt10days, dose_2gt2doses, f13_Useful..

Table 3: Evaluation of machine learning models using test data

Prediction models	AUC (95% CI)	Log-loss	Accuracy	Sensitivity	Specificity	F1-score	Balanced accuracy
Full logistic regression	0.904 (0.8430, 0.9648)	0.3860	0.8788	0.9189	0.8632	0.8095	0.8910
Stepwise logistic regression	0.905 (0.8450, 0.9640)	0.3781	0.8636	0.8919	0.8526	0.7857	0.8723
Lasso regression	0.902 (0.8326, 0.9712)	0.3258	0.8864	0.9706	0.8571	0.8148	0.9139
RF (29 predictors)	0.919 (0.8632, 0.9754)	0.2915	0.9091	0.9487	0.8925	0.8605	0.9206
RF + Boruta (nonrej) (9 predictors)	0.918 (0.8586, 0.9764)	0.2835	0.8864	0.9706	0.8571	0.8148	0.9139
RF+ Boruta (cnf) (7 predictors)	0.928 (0.8777, 0.9782)	0.3597	0.9091	0.9268	0.9011	0.8636	0.9140
XGB tree (29 predictors)	0.918 (0.8591, 0.9772)	0.2983	0.9091	0.9487	0.8925	0.8605	0.9206
XGB tree + Boruta (nonrej) (9 predictors)	0.930 (0.8831, 0.9760)	0.3030	0.8939	0.8837	0.8989	0.8444	0.8913
XGB tree + Boruta (cnf) (7 predictors)	0.934 (0.8906, 0.9782)	0.2793	0.9091	0.9268	0.9011	0.8636	0.9140

cnf, confirmed important (7 predictors); nonrej, non-rejected (7 cnf + 2 Tntv variables) (9 predictors); RF, random forest; XGB, extreme gradient boosting...

9 nonrej predictors: f1_Block2, f2_Others, eyes_Yes, persons_2gt1persons, days_2gt10days, dose_2gt2doses, f13_Useful, f11_Yes, gis_Yes.

7 Cnf predictors: f1_Block2, f2_Others, eyes_Yes, persons_2gt1persons, days_2gt10days, dose_2gt2doses, f13_Useful.

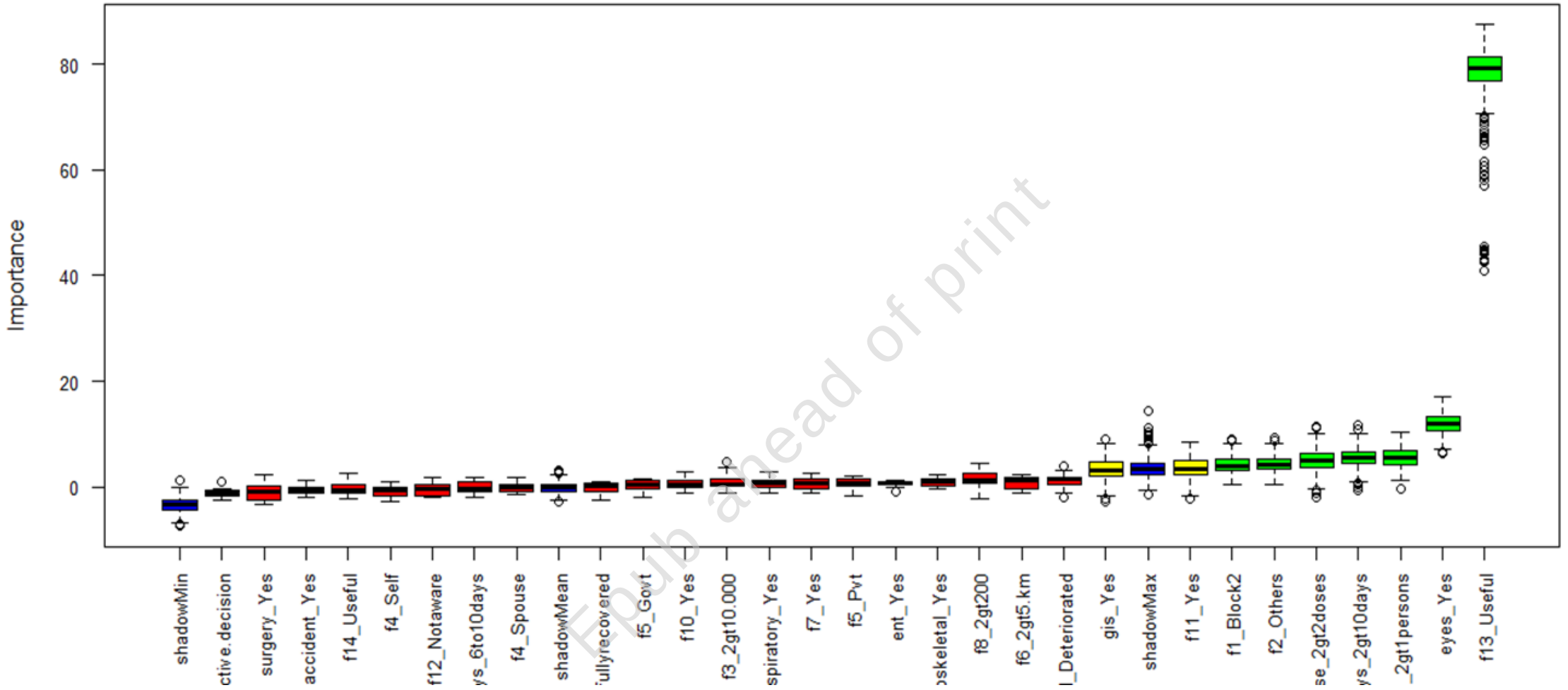

Figure 1: Predictor selection by random forest based Boruta algorithm for predicting OTC-antibiotic use in Rural Pune, India