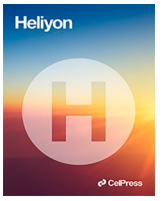
Heliyon 10 (2024) e28534

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

# Machine learning approach to identify malaria risk in travelers using real-world evidence

Pedro Emanuel Fleitas, Leire Balerdi Sarasola, Daniel Camprubi Ferrer, Jose Muñoz [*,1], Paula Petrone [**,1]

*Barcelona Institute for Global Health (ISGlobal) Hospital Clínic - Universitat de Barcelona, Barcelona, Spain*

A B S T R A C T

*Background:* Pre-travel consultation and chemoprophylaxis measures for malaria are a key component in the prevention of imported malaria in travelers. In this study we report a predictive tool for assessing personalized malaria risk in travelers based on the analysis of electronic medical records from travel consultations. The tool aims to guide physicians in the recommendation of appropriate prophylaxis prior to their trip. We also provide best-practice recommendations for pre-processing noisy and highly sparse real world evidence data.

*Methods:* We leveraged a large EMR dataset, containing demographic information about travelers and their destination. The data has been previously preprocessed using various strategies to handle missing and unbalanced data. We compared multiple machine learning approaches to assess the risk of malaria acquisition in travelers during their travels. Additionally, a feature importance analysis was performed using SHAP (SHapley Additive Explanations) values to identify patterns associated with malaria risk.

*Results:* Our study revealed that our XGB models achieved high predictive capacity (AUC >0.80). The most significant features predicting malaria infection during travel included travel destinations with low malaria risk, vaccination history, number of countries visited, age, and trip duration. Remarkably, we were able to obtain a reduced model with only five features. When comparing this model with a population of travelers recommended for malaria chemoprophylaxis, we observed that it was deemed necessary in only 40% of these travelers. This suggests that 60% received chemoprophylaxis despite having a low personalized risk of malaria.

*Conclusion:* We have developed an algorithmic tool that utilizes a concise survey to generate a personalized travel risk assessment, effectively minimizing the prescription of unnecessary malaria chemoprophylaxis. Through the identification of patterns linked to predictions, our model significantly enhances the efficacy of pre-travel consultations.

## 1. Introduction

Malaria is a parasitic disease caused by various species of *Plasmodium* worldwide, transmitted by the bite of Anopheles mosquitoes. Malaria remains one of the main parasitic diseases [1]. Additionally, cases of malaria imported to non-endemic regions and diagnosed among travelers are a challenge for non-endemic countries. In Europe, 12.000–15.000 cases of imported malaria are diagnosed every year [2]. In this context, pre-travel consultation and prophylactic measures for malaria (chemoprophylactics and measures to reduce mosquito bites) play a crucial role in preventing the disease among travelers. This involves assessing the traveler's risks based on their destination, demographic, and health profile [3]. However, despite the existence of guidelines to assist clinicians in assessing malaria risk in travelers [4], there is currently no available tool that assigns a personalized malaria risk score to travelers by considering the interaction of multiple variables. Such a tool would enable clinicians to make an initial assessment of malaria risk in travelers and

* Corresponding author. Address: C/ del Rosselló, 132, 08036 Barcelona, Spain.
** Corresponding author. Address: C/ del Dr. Aiguader, 88, 08003. Barcelona. Spain.
*E-mail addresses:* jose.munoz@isglobal.org (J. Muñoz), paula.petrone@isglobal.org (P. Petrone).
[1] Jose Muñoz and Paula Petrone contributed equally to this study.

https://doi.org/10.1016/j.heliyon.2024.e28534
Received 17 June 2023; Received in revised form 19 March 2024; Accepted 20 March 2024
Available online 22 March 2024

**Table 1**
Characteristics of travelers without malaria chemoprophylaxis (Set_NP) and with malaria chemoprophylaxis (Set_P).

| Set_NP: Travelers without malaria chemoprophylaxis (n = 3490) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Age range** | **Frequency** | **Mean number of vaccines** | **Male** | **Female** | **Attended to pre-travel advice** | **Born in Europe** | **Born in America** | **Born in Africa** | **Travel to America** |
| 0–14 | 1 | 4 | 1 | 0 | 0 | 1 | 0 | 0 | 1 |
| 15–44 | 2785 | 4 | 1172 | 1613 | 1305 | 2538 | 73 | 8 | 1204 |
| 45–64 | 597 | 3 | 295 | 302 | 165 | 529 | 25 | 5 | 269 |
| >64 | 107 | 3 | 59 | 48 | 28 | 99 | 6 | 1 | 55 |
| **Age range** | **Travel to Africa** | **Travel to Asia** | **Travel to Oceania** | **Trip for work** | **Sightseeing trip** | **VFR trip** | **International cooperation trip** | **Travel to a country with a recommendation for malaria chemoprophylaxis** | **Malaria infection** |
| 0–14 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 15–44 | 512 | 1179 | 19 | 405 | 1964 | 196 | 220 | 796 | 90 |
| 45–64 | 147 | 153 | 1 | 124 | 366 | 83 | 24 | 174 | 44 |
| >64 | 25 | 24 | 0 | 10 | 68 | 17 | 12 | 34 | 5 |
| **Set_P: Travelers with malaria chemoprophylaxis (n = 953)** | | | | | | | | | |
| **Age range** | **Frequency** | **Mean number of vaccines** | **Male** | **Female** | **Attended to pre-travel advice** | **Born in Europe** | **Born in America** | **Born in Africa** | **Travel to America** |
| 15–44 | 741 | 5 | 268 | 473 | 676 | 712 | 8 | 0 | 178 |
| 45–64 | 175 | 5 | 78 | 97 | 162 | 167 | 3 | 0 | 41 |
| >64 | 37 | 5 | 16 | 21 | 33 | 34 | 0 | 2 | 7 |
| **Age range** | **Travel to Africa** | **Travel to Asia** | **Travel to Oceania** | **Trip for work** | **Sightseeing trip** | **VFR trip** | **International cooperation trip** | **Travel to a country with a recommendation for malaria chemoprophylaxis** | **Malaria infection** |
| 15–44 | 512 | 90 | 1 | 115 | 424 | 24 | 178 | 625 | 38 |
| 45–64 | 140 | 9 | 0 | 30 | 116 | 5 | 24 | 154 | 5 |
| >64 | 27 | 1 | 0 | 3 | 26 | 2 | 6 | 33 | 2 |

determine the appropriate prophylactic measures. In line with this objective, the data collected during pre-travel consultation and recorded in electronic medical records (EMR) can be aggregated and analyzed to obtain a set of real-world data (RWD). This information can then be utilized to develop models that assist in risk assessment, clinical decision making, and health technology assessment among other applications [5].

EMR offers several advantages, including large sample sizes, a wide range of predictor variables, and data that better reflect the real world compared to well-designed prospective cohort studies. However, the analysis of EMR also presents several limitations, such as a high degree of missing data, imbalance between features, and potential bias towards a specific population, such as patients with a particular disease in a specific hospital. This situation is referred to as "health data poverty", which occurs when individuals, groups, or populations cannot benefit from a discovery or innovation due to insufficient data that adequately represent of the general patient population [6]. Consequently, robust imputation models and data processing methods are necessary to overcome health data poverty and generate descriptive and predictive insights.

Machine learning techniques are a powerful tool for designing models and identifying individuals at risk of adverse health events. This allows healthcare providers to plan preventive interventions [5]. Machine learning has proven successful in assessing personalized risk in several chronic medical conditions such as: diabetes, cancer, asthma, and heart disease [7], and also infectious diseases such as COVID-19, dengue, brucellosis, etc. [8].

In the specific context of malaria, machine learning models applied to EMR has been successful in identifying malaria patients in

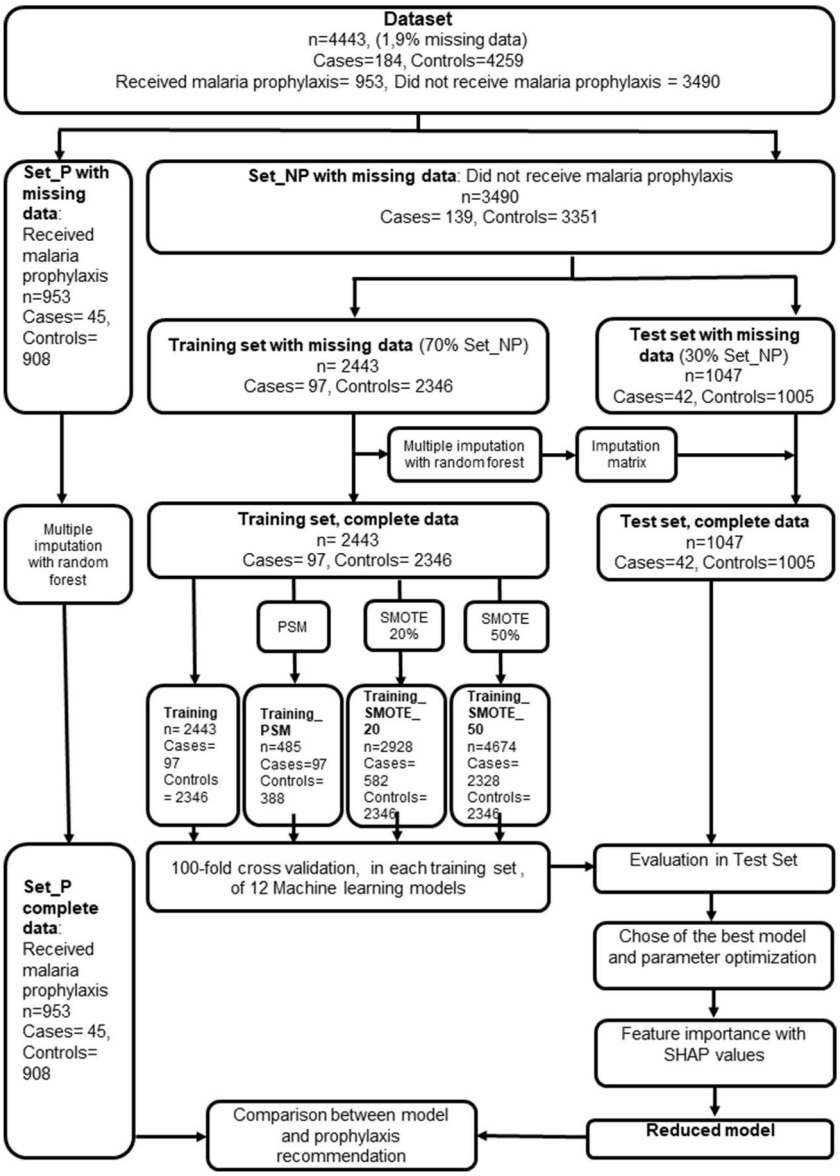



**Fig. 1.** Flowchart of data processing, validation and evaluation of machine learning models.

endemic areas, utilizing symptoms and demographic characteristics of the patients [9,10]. To our knowledge, there is no prior study that attempts to predict the risk of infection in travelers. This topic is also relevant in the context of pandemic preparedness and surveillance. Travelers can serve as potential sources of disease transmission, especially if they visit areas with ongoing malaria outbreaks. By accurately assessing and mitigating the risk of infection in travelers, the spread of malaria to new regions or the reintroduction of malaria in previously controlled areas can be prevented.

In this study, our focus is on predicting malaria risk in travelers using typical pre-travel consultation questionnaires, and identifying relevant patterns and correlations in malaria detection. The primary objective of this paper is to build a predictive tool for personalized risk of malaria for travelers based on the analysis of EMR. The tool aims to identify travelers at higher risk who would benefit from malaria chemoprophylaxis. Our ultimate goal is for this tool to assist physicians in making recommendations for malaria prophylaxis in high-risk travelers.

## 2. Methods

### *Ethical approval*

This study received approval from the Medical Research Committee of Hospital Clinic, which authorized the review of medical records for the purposes of this study. As the study is based on the analysis of existing electronic medical records, the informed consent of the patients was not required. No additional interventions were conducted, and no new data was obtained from the patients, thus active participation in the study was not requested. To ensure the confidentiality and protection of personal data, anonymized data was utilized to remove any identifiable information, and access to the records was restricted.

Considering the retrospective nature of the study and the anonymity of the data, the study posed minimal risk while offering potential benefits in terms of generating valuable insights.

### *2.1. Dataset*

The dataset comprises 4443 anonymized traveler records from pretravel consultations at the Department of International Health, Hospital Clínic, University of Barcelona, Spain, from October 2014 to January 2022. A total of 5% (184/4443) travelers had been diagnosed with malaria during or immediately after their trip. The EMR for each patient includes data on age, sex, country of birth, administered vaccines, travel destinations, duration of the trip, number of countries visited, type of area visited (i.e., urban, rural or urban and rural), travel reason (i.e., work, tourism, international cooperation, or VFR (Visited friends and relatives)), attendance to pre-travel advice consultation (yes/not), and the outcome of malaria diagnosis. Features with more than two categories were processed with one-hot encoding.

This dataset was divided into two datasets, one where travelers did not receive malaria chemoprophylaxis (n = 3490) called set_NP and another where travelers received malaria chemoprophylaxis (n = 953) called set_P. The latter set is composed of those travelers who declared in the post-travel consultation that they had received malaria chemoprophylaxis during their trip. However, 45 (5%) of travelers in set_P acquired malaria during their trip. The characteristics of these datasets can be observed in Table 1.

The two different datasets were used for different analyses to avoid the potential confounder effect of malaria chemoprophylaxis when assessing the outcome of imported malaria. The Set_NP was used for the building, validation and evaluation of the machine learning models. On the other hand, the set_P was used to compare the model prediction with the chemoprophylaxis recommendation. We used the set_P to retrospectively assess the potential impact of the machine learning model recommendation on the indication of

**Table 2**
Predictive characteristics of the 6 best machine learning models for assessing malaria infection risk in travelers.

| Model | Training | | | | Training_PSM | | | |
|---|---|---|---|---|---|---|---|---|
| | Training set | Test set | | | Training_PSM | Test set | | |
| | *Mean AUC (IC95%) | AUC | Sensitivity | Specificity | *Mean AUC (SD) | AUC | Sensitivity | Specificity |
| QDA | 0.76 (0.71–0.82) | 0.8 | 0.86 | 0.74 | 0.70 (0.67–0.73) | 0.8 | 0.86 | 0.74 |
| GNBC | 0.78 (0.75–0.81) | 0.75 | 0.86 | 0.65 | 0.74 (0.71–0.77) | 0.75 | 0.86 | 0.65 |
| AB | 0.87 (0.85–0.89) | 0.93 | 0.24 | 0.99 | 0.79 (0.75–0.84) | 0.82 | 0.31 | 0.94 |
| DT | 0.59 (0.54–0.65) | 0.63 | 0.17 | 1.00 | 0.66 (0.59–0.69) | 0.74 | 0.21 | 0.99 |
| LR | 0.89 (0.87–0.92) | 0.92 | 0.14 | 1.00 | 0.80 (0.77–0.84) | 0.8 | 0.24 | 0.98 |
| XGB | 0.80 (0.78–0.83) | 0.85 | 0.14 | 1.00 | 0.68 (0.64–0.72) | 0.75 | 0.24 | 0.94 |
| **Model** | **Training_SMOTE_20** | | | | **Training_SMOTE_50** | | | |
| | **Training_SMOTE_20** | **Test set** | | | **Training_SMOTE_50** | **Test set** | | |
| | ***Mean AUC (IC95%)** | **AUC** | **Sensitivity** | **Specificity** | ***Mean AUC (IC95%)** | **AUC** | **Sensitivity** | **Specificity** |
| QDA | 0.97 (0.96–0.98) | 0.81 | 0.86 | 0.76 | 0.99 (0.99–0.99) | 0.81 | 0.86 | 0.77 |
| GNBC | 0.92 (0.91–0.93) | 0.86 | 0.83 | 0.89 | 0.95 (0.94–0.95) | 0.86 | 0.67 | 0.90 |
| AB | 0.98 (0.97–0.99) | 0.89 | 0.36 | 0.97 | 0.99 (0.99–0.99) | 0.92 | 0.52 | 0.95 |
| DT | 0.91 (0.90–0.92) | 0.79 | 0.48 | 0.94 | 0.96 (0.96–0.97) | 0.84 | 0.52 | 0.91 |
| LR | 0.97 (0.96–0.97) | 0.87 | 0.17 | 1.00 | 0.98 (0.98–0.99) | 0.82 | 0.24 | 0.98 |
| XGB | 0.97 (0.97–0.98) | 0.86 | 0.14 | 0.98 | 0.99 (0.99–0.99) | 0.84 | 0.14 | 0.98 |

malaria chemoprophylaxis. For which, travelers categorized as having a high risk of acquiring malaria by the model would be the ones to whom malaria chemoprophylaxis would be recommended.

## 3. Data pre-processing

### 3.1. Missing data handling

The main dataset has 1.9% of total missing data split among the features. To overcome it, multiple imputation, excluding the outcome, was performed using random forest using the mice library of R software [11,12]. In set_P the imputation was made on the whole set, while Set_NP was first separated into a training set (70% of the Set_NP, n = 2443, cases = 97, controls = 2346) and a test set (30% of the dataset, n = 1047, cases = 42, controls = 1005). Multiple imputation was performed in the training set. Then, the prediction matrix created based on the training set was applied for multiple imputation in the test set (Fig. 1).

As a result of each multiple imputation, five complete data sets were created for each data set (Set_P, Set_NP Training, and Set_NP Test). The distribution of the variables with missing data from the imputation sets did not present significant differences with the original data set (Quantitative variables $p = 0.99$ Kolmogorov-Smirnov test, categorical variables $p > 0.05$ chi square test) (Supplementary files , Fig. S1). Finally, an imputed Set_P, a training set called Training and a testing set called Test were created by computing the mean (for quantitative variables) or selecting the most likely imputed value (for categorical variables) between the different five sets.

### 3.2. Handling of imbalanced data

The issue of data imbalance can impact on the effectiveness of the machine learning classification model, because the variability of the frequent class is much better represented by a large number of cases compared to the scarce class. Therefore, the model tends to provides more accurate classification for the more abundant class. In our dataset malaria cases accounted for only 4% of the main dataset.

Two approaches were taken to reduce the imbalance in the training set.

- Propensity Score Matching (PSM): This method enables the generation of a smaller dataset where the patterns of one class resembles those of the other [13]. Each malaria case was matched with four healthy individuals who had similar scores. Logistic regression was employed for the calculation of the scores, including variables age, sex, type of area visited, and reason for the trip, as features, and malaria infection as the outcome. Matching was performed by comparing the scores by nearest neighbor using the MatchIt library of R software [14]. Finally, a reduced training set called Training_PSM was obtained, consisting of 97 cases of malaria from the Training set and 388 control cases (n = 485) (Fig. 1).

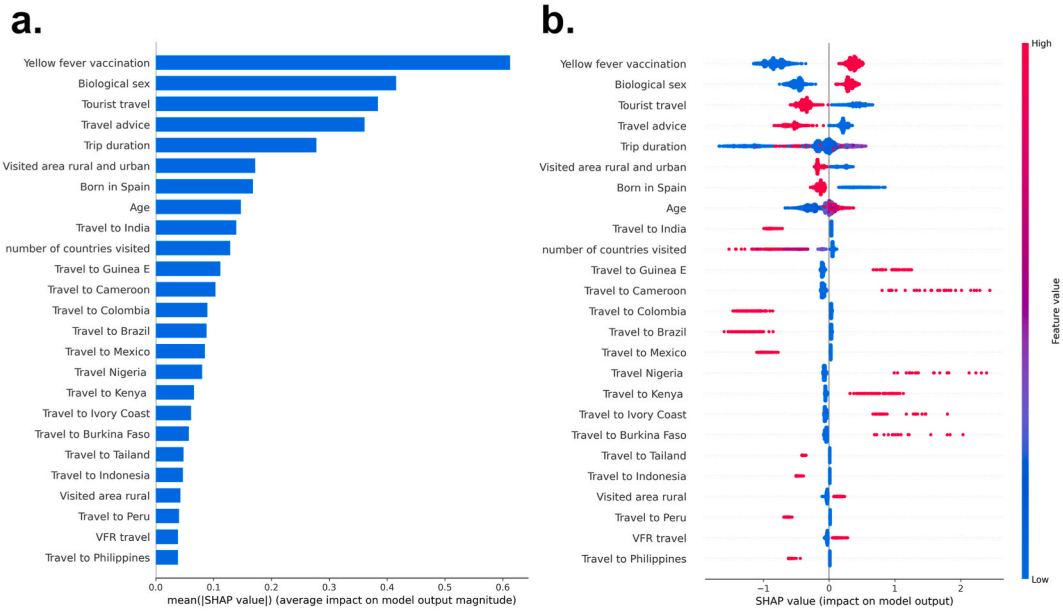



**Fig. 2.** SHAP values computed for malaria infection cases of the 25 most significant features obtained with the optimized XGB model in the Test set. (a.) Mean SHAP values. **(b.)** SHAP values. Each point represents an individual of the Test set. High values of the feature are represented in red, and low values in blue, and the corresponding SHAP value is observed on the x-axis. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

- Synthetic Minority Oversampling Technique (SMOTE): This method allows the creation of a larger data set in which synthetic data are generated for the minority class by an oversampling technique. Two new training datasets were generated: the Training_SMOTE_20 dataset, giving a proportion of 20% malaria infection (n = 2928); and the Training_SMOTE_50 giving a proportion of 50% malaria (n = 4674) (Fig. 1). This was done with the smotefamily library of R software [15]. No modifications were made in the Test set.

### 3.3. Machine learning models

Four training sets (Training, Training_PSM, Training_SMOTE_20, Training_SMOTE_50) were used for training 12 machine learning classification methods (Nearest Neighbors (NN), Support vector machines (SVM), Radial Basis Function SVM (RBF-SVM), Gaussian Process Classifier (GPC), Decision Tree (DT), Random Forest (RF), Multi-layer Perceptron classifier (MLPC), AdaBoost (AB), Gaussian Naive Bayes classifier (GNBC), Quadratic Discriminant Analysis (QDA), XGBoost (XGB), and Logistic Regression (LR)) (Fig. 1). For each model, a 100-fold cross-validation (70% train, 30% test) was applied on each training set and the mean and standard deviation of the area under the receiving-operating curve (ROC AUC) were calculated. In addition, each model was validated on the same Test set, the AUC, sensitivity (positive recall) and specificity (negative recall) for malaria infection were calculated. The best models were chosen for optimization of parameters using the parameter F1 as the reference performance metric (Fig. 1). F1 is calculated as the harmonic mean of both precision and recall for the minority positive class, and is a robust performance metric for imbalanced scenarios.

In addition, the probability of the model was adjusted to obtain a sensitivity greater than 90% in the training set, since the model acts as a first screening stage for the physician in the travel consultation visit. These analyzes were performed with various Python libraries, the data was structured with the Pandas library [16], the machine learning algorithms were sourced from the scikit-learn library [17], and the figures were made with the matplotlib library [18].

### 3.4. Feature importance and feature selection mechanism

In the best-performing machine learning classification model, we analyzed the feature importance using SHAP (Shapley Additive exPlanations) values. SHAP values provide an explanation for each prediction by quantifying the contribution of dataset features to the model's output [19]. In addition, we employed SHAP values as a feature selection mechanism, where the features were ranked according to their contribution, and classification models were evaluated using different percentages of the features [20]. For this, the SHAP library of Python was used [19].

## 4. Results

### 4.1. Machine learning modeling to assess malaria risk

The machine learning approach allowed us to obtain models with high predictive capacity for malaria infection in travelers (Table 2). These models demonstrated notable AUC values exceeding 80%. The QDA and GNBC models presented the highest sensitivity values, successfully detecting more than 85% of the malaria cases. On the other hand, the AB, XGB and LR models presented the highest specificity. Subsequently, an ensemble model was generated with the combination of QDA (higher Sensitivity) and XGB (highest specificity). However, this ensemble model did not exhibit superior characteristics to the individual models (Supplementary files Fig. S2).

Parameter optimization of the aforementioned models was performed and the best results were obtained with the XGB model, obtaining an AUC in the cross-validation of 0.90 (95%CI: 0.84–0.96) and an AUC of 0.80 in the evaluation with the Test set. Therefore, adjusting the sensitivity to 90%, a specificity of 81% was obtained. This implies the high capacity of the model to identify travelers at


**Fig. 3.** Performance of the XGB model for the risk of infection with malaria in travelers, with different percentage of features.

**Table 3**
Comparison between the XGB (five question) model and the Set_P.

| | | | | Set_P (Malaria diagnosis) | |
|---|---|---|---|---|---|
| | | Set_P chemoprophylaxis recommendation (n = 953) | | Positive | Negative |
| **XGB Model** | **High risk of malaria** | 373 (40%) | **High risk of malaria** | 32 | 341 |
| | **No risk of malaria** | 580 (60%) | **No risk of malaria** | 13 | 567 |

risk of acquiring malaria during their trip.

### 4.2. Feature importance

Of the 101 features, the importance analysis with SHAP values revealed that only 63 features (63%) were relevant for the XGB model (Supplementary files, Fig. S3). The ten most important features were being vaccinated with yellow fever vaccine, biological sex, tourist travel, have received travel advice, duration of the trip, age, have visited rural and urban areas, travel to India, and the number of countries visited (Fig. 2a.). From the above-mentioned characteristics, those that increased the risk of malaria in travelers were: having been vaccinated with the yellow fever vaccine, being male, and age. On the other hand, tourist trips, attending the traveler's consultation, having being born in Spain and traveling to India were associated with a lower risk of malaria in travelers (Fig. 2b.). It should be noted that in our population of travelers the destinations associated with a higher risk of malaria were Equatorial Guinea, Nigeria, Kenya, the Ivory Coast and Burkina Faso (Fig. 2). While countries such as Colombia, Brazil and Mexico presented a low risk of malaria for travelers.

### 4.3. Reduced model based on feature selection

The XGB model was optimized in the Training set based on feature importance, considering different percentages of features: 100%, 60% 50%, 40%, 20% 10% and 5% of features (Fig. 3). The optimal number of features was 60% (63 features), resulting in the following performance metrics: Cross validation in the Training set AUC = 0.90 (95% CI: 0.78–0.96), evaluation in Test set AUC = 0.80, sensitivity = 0.90 (95% CI: 0.78–0.96) and specificity = 0.83 (95% CI:0.81–0.86).

Remarkably, even when the number of features was reduced to 5%, the model's predictive ability was high, with an AUC = 0.83 (95% CI: 0.74–0.91) during cross-validation with the Training set, and AUC = 0.70 in the evaluation with the Test set, with a sensitivity = 0.90 (95% CI: 0.77–0.96), and a specificity = 0.71 (95% CI: 0.68–0.74). Consequently, a minimal model consisting of only 5 questions was developed for the convenient clinical assessment of malaria risk in travelers in the pre-travel consultation. These questions include: 1. Have you been vaccinated with the yellow fever vaccine? (Yes/No). 2. What is your biological sex? (Female/Male). 3. Is the purpose of your trip tourism? (Yes/No). 4. Did you attend a pre-travel consultation visit? (Yes/No). 5) How long will you be traveling? (Number of days).

### 4.4. Clinical relevance and applicability of the XGB model

In order to assess the practicality of the model for assessing risk in travel consultation visits, the minimum XGB model consisting of five questions was evaluated in Set_P which consisted of unseen data from travelers who had received chemoprophylaxis prior to their travel. The results showed that the model successfully identified 71% of the travelers who were later diagnosed with malaria (Table 3). Moreover, the model predicted that chemoprophylaxis was necessary for only 373 (40%) travelers of Set_P, implying that 60% of travelers received chemoprophylaxis despite having a low risk of malaria.

## 5. Discussion

In this study, we developed machine learning models with high predictive capacity to assess individualized malaria risk and personalize travel risk assessment. Traditionally, the risk of malaria among international travelers is evaluated in pre-travel consultations based on the incidence of malaria in the destination country as the primary indicator. While machine learning has been previously been utilized with clinical data and malaria images (Giemsa-stained blood films) to assist in diagnosis [9,21,22], to the best of our knowledge, this is the first study in which traveler's EMR have been analyzed using this methodology to assess personalized malaria risk for travelers [23].

Our best model with 63 features presented an AUC = 0.80 in the evaluation with the Test set. Furthermore, we successfully obtained a reduced model based on a minimal 5-question survey (Test set evaluation: AUC = 0.70, sensitivity = 0.90 (95% CI: 0.77–0.96), specificity = 0.71 (95% CI: 0.68–0.74)), aiming to assist the physicians during the travel consultation visit. It is essential to consider the specific use case and applicability of the model, as one model may be more advantageous than the other depending on the context.

It is important to note that feature importance does not necessarily indicate a causal relationship with malaria infection and can often point to biases in the data, which should be carefully considered. However, there are notable patterns that emerge, such as the lower risk associated with tourist trips, as many tourist destinations in Southeast Asia, the Caribbean and Latin America have no risk of malaria [24]. Another noteworthy pattern is the higher risk in VFR travelers, and lower risk of malaria in India, which is classified as a

low-risk country, where malaria chemoprophylaxis is not recommended except when visiting specific states like Assam or Orissa [23].

The most informative feature of the XGB model was the vaccination against yellow fever, which was found to be associated with an increased risk of acquiring malaria. However, it is important to interpret this result with caution due to the confounding effect of travel to high-risk countries with malaria. Yellow fever vaccination serves as an example of a feature that is useful for the predictive model, but its association with the outcome is indirect. In many cases, yellow fever vaccination is administered to travelers visiting yellow fever endemic areas, which often overlap with malaria endemic areas [25]. Additionally, the higher risk of malaria among male travelers in our population can be attributed to the higher number of reported cases among male travelers compared to female travelers.

In some cases, recommending malaria chemoprophylaxis when the risk is low may be unnecessary, and could increase cost and the rate of adverse events. It is important to note that malaria chemoprophylaxis alone does not provide complete protection and should be accompanied by measures to reduce mosquito bites [26]. Additionally, non-adherence to malaria chemoprophylaxis is common, particularly among VFR travelers, long-term travelers and those who have difficulty adhering to a daily schedule [27].

Within the set_P population where all travelers received malaria chemoprophylaxis, 5% (45/953 travelers) still acquired malaria during their trip. However, it is challenging to determine how many cases could have acquired malaria if chemoprophylaxis had not been recommended. Nevertheless, in this scenario, the XGB model was able to identify 71% (32/45) of the total malaria cases in Set_P, and classified 40% as high risk of that population. This approach represents a more conservative prescription of antimalarial chemoprophylaxis, with only 40% receiving chemoprophylaxis while emphasizing measures to prevent mosquito bites for the remaining 60%. The validity of this approach should be further explored with new data. However, our model can be easily implemented in an application for use during the pre-travel consultation visit, enabling a quick assessment of malaria risk, and providing initial information to the physician for safer travel measures.

The analysis of EMR, presented several challenges in terms of data quality and completeness, which had to be addressed analytically. In our database, we encountered 1.9% empty data with some features having up to 20% missing data. This was primarily due to ambiguity in documentation of medical data, where positive data is often recorded while negative data is left blank. It is important to note that, in EMRs, missing data is often informative and carries meaning [28]. In this study, we took a different approach to handling missing data, compared to the common practice of removing all incomplete cases and performing complete cases analysis. Instead, we performed multiple imputation using random forest, allowing us to generate imputed datasets with low variability. The high performance of our models suggest that our imputation approach is robust and efficient.

Another challenge we encountered was data imbalance in EMRs. In our database, we found that the two strategies to solve this problem did not have a significant impact on the performance of the machine learning models. The best performing models maintained their predictive capacity regardless of the data balancing methodology used. Conversely, the models that exhibited poor performance did not demonstrate improvements and even showed signs of overfitting when applying the SMOTE approach. Additionally, no significant in performance was observed between the Training_SMOTE_20 and Training_SMOTE_50 sets, indicating that it was not necessary to equalize the percentage of cases and controls, which is typically done when using SMOTE [29,30]. It worth noting that numerous studies reporting improved performance after balancing the data with SMOTE may misinterpret the model metrics (higher AUC, precision, recall, or F1) by adjusting and evaluating the model using the same data set [31], rather than testing the model with an independent dataset. This practice is not recommended as it can lead to the contamination of the test set with some of the training set data, resulting in overestimated values of the model metrics [31]. Based on our results, we conclude that it is more effective to identify the best machine learning model and optimize it accordingly, rather than solely relying on data balancing techniques to favor any specific model. Furthermore, while the EMRs allowed us to design machine learning models, it is important to acknowledge that the predictive models obtained are specific to the population of travelers within our hospital. Nevertheless, this study demonstrates that models with high predictive capacity can be generated based on relatively simple questionnaires requested from travelers. More accurate models could be obtained with the incorporation of more precise localizations, complemented by up-to-date malaria epidemiology.

In this study, we present guidelines for data processing and model development based on real world data from clinical practice. Our successful achievement was the development a five-question machine learning model capable of identifying travelers at high risk of acquiring malaria during their trip, utilizing information gathered during the traveler consultation visit.

Our model holds the potential to personalize travel risk assessment and minimize unnecessary prescription of malaria chemoprophylaxis. We aim to implement this model to assist in the traveler's consultation of Hospital Clínic. By doing so, we can enhance personalized malaria risk assessment for travelers and promote better preventive measures. This, in turn, can lead to increased adherence to malaria chemoprophylaxis and ultimately save lives.

## Authors' contributions

Conception and design of study: Pedro E. Fleitas, Jose Muñoz and Paula Petrone. Acquisition of data: Jose Muñoz. Statistical analysis and interpretation of data: Pedro E Fleitas, Jose Muñoz and Paula Petrone. Drafting and revision of the manuscript: Pedro E Fleitas, Leire Balerdi Sarasola, Daniel Camprubi Ferrer, Jose Muñoz, Paula Petrone. All authors approved the version of the manuscript to be published.

## Statement on conflicts of interest

We have no conflicts of interest to declare.

## Sources of funding

We acknowledge support from the Spanish Ministry of Science and Innovation through the "Centro de Excelencia Severo Ochoa 2019–2023" Program (CEX2018-000806-S) and support from the Generalitat de Catalunya through the CERCA Program.

## Data availability statement

The study deidentified participant data and data dictionary will be made available upon request to the corresponding author after approval of a proposal and signed data access agreement.

## CRediT authorship contribution statement

**Pedro Emanuel Fleitas:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Leire Balerdi Sarasola:** Writing – review & editing, Investigation. **Daniel Camprubi Ferrer:** Writing – review & editing, Investigation. **Jose Muñoz:** Writing – review & editing, Supervision, Methodology, Funding acquisition, Conceptualization. **Paula Petrone:** Writing – review & editing, Supervision, Methodology, Funding acquisition, Formal analysis, Conceptualization.

## Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Paula Petrone reports financial support was provided by Spanish Ministry of Science and Innovation (Centro de Excelencia Severo Ochoa. Program CEX2018-000806-S), Spain;and Generalitat de Catalunya (CERCA Program), Spain. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e28534.

## References

[1] World Health Organization, Word Malaria Report 2021, 2021. Geneva, Switzerland.
[2] S. Odolini, P. Gautret, P. Parola, Epidemiology of imported malaria in the Mediterranean region, Mediterr. J. Hematol, Inf. Disp. 4 (2012) e2012031, https://doi.org/10.4084/mjhid.2012.031.
[3] N.I. Agudelo Higuita, B.P. White, C. Franco-Paredes, M.A. McGhee, An update on prevention of malaria in travelers, Ther. Adv. Infect. Dis. 8 (2021) 204993612110406, https://doi.org/10.1177/20499361211040690.
[4] R. Morales, N. Rodriguez, S. Otero, L. Cabanas, F. Agüero, I. Oliveira, Guía de recomendaciones para la prevención de la malaria en viajeros, Barceloma Esmon Publicidad, SA, 2019.
[5] W.H. Crown, Real-world evidence, causal Inference, and machine learning, Value Heal 22 (2019) 587–592, https://doi.org/10.1016/j.jval.2019.03.001.
[6] H. Ibrahim, X. Liu, N. Zariffa, A.D. Morris, A.K. Denniston, Health data poverty: an assailable barrier to equitable digital health care, Lancet Digit. Heal. 3 (2021) e260–e265, https://doi.org/10.1016/S2589-7500(20)30317-4.
[7] S. Uddin, A. Khan, M.E. Hossain, M.A. Moni, Comparing different supervised machine learning algorithms for disease prediction, BMC Med. Inform. Decis. Mak. 19 (2019) 281, https://doi.org/10.1186/s12911-019-1004-8.
[8] O.E. Santangelo, V. Gentile, S. Pizzo, D. Giordano, F. Cedrone, Machine learning and prediction of infectious diseases: a Systematic review, Mach. Learn. Knowl. Extr. 5 (2023) 175–198, https://doi.org/10.3390/make5010013.
[9] Y.W. Lee, J.W. Choi, E.-H. Shin, Machine learning model for predicting malaria using clinical information, Comput. Biol. Med. 129 (2021) 104151, https://doi.org/10.1016/j.compbiomed.2020.104151.
[10] M. Mariki, E. Mkoba, N. Mduma, Combining clinical symptoms and patient features for malaria diagnosis: machine learning approach, Appl. Artif. Intell. 36 (2022), https://doi.org/10.1080/08839514.2022.2031826.
[11] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, 2017. https://www.r-project.org/.
[12] S. van Buuren, K. Groothuis-Oudshoorn, Mice : Multivariate imputation by chained equations in R, J. Stat. Softw. 45 (2011) 1–67, https://doi.org/10.18637/jss.v045.i03.
[13] P.C. Austin, An introduction to propensity score methods for reducing the effects of confounding in observational studies, Multivariate Behav. Res. 46 (2011) 399–424, https://doi.org/10.1080/00273171.2011.568786.
[14] D.E. Ho, K. Imai, G. King, E.A. Stuart, Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference, Polit. Anal. 15 (2007) 199–236, https://doi.org/10.1093/pan/mpl013.
[15] M.W. Siriseriwan, A Collection of Oversampling Techniques for Class Imbalance Problem Based on SMOTE, 2019.
[16] W. McKinney, Data structures for statistical computing in Python, in: Proc. 9th Python Sci. Conf., 2010, pp. 56–61, https://doi.org/10.25080/Majora-92bf1922-00a.
[17] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2012) 2825–2830, https://doi.org/10.48550/arxiv.1201.0490.
[18] J.D. Hunter, Matplotlib: a 2D graphics environment, Comput. Sci. Eng. 9 (2007) 90–95, https://doi.org/10.1109/MCSE.2007.55.
[19] S.M. Lundberg, B. Nair, M.S. Vavilala, M. Horibe, M.J. Eisses, T. Adams, D.E. Liston, D.K.-W. Low, S.-F. Newman, J. Kim, S.-I. Lee, Explainable machine-learning predictions for the prevention of hypoxaemia during surgery, Nat. Biomed. Eng. 2 (2018) 749–760, https://doi.org/10.1038/s41551-018-0304-0.

[20] W.E. Marcilio, D.M. Eler, From explanations to feature selection: assessing SHAP values as feature selection mechanism, in: 2020 33rd SIBGRAPI Conf. Graph. Patterns Images, IEEE, 2020, pp. 340–347, https://doi.org/10.1109/SIBGRAPI51738.2020.00053.
[21] R.R. Rees-Channer, C.M. Bachman, L. Grignard, M.L. Gatton, S. Burkot Global, H. Labs, M.P. Horning, C.M. Thompson, K. Woods, P. Lansdell, S. Shah, Evaluation of an automated microscope using machine learning for the detection of malaria in travelers returned to the UK. https://doi.org/10.21203/rs.3.rs-1622168/v1, 2022.
[22] A. Olugboja, Z. Wang, Malaria parasite detection using different machine learning classifier, in: 2017 Int. Conf. Mach. Learn. Cybern., IEEE, 2017, pp. 246–250, https://doi.org/10.1109/ICMLC.2017.8107772.
[23] UK Health Security Agency, Guidelines for malaria prevention in travellers from the UK 2021. https://www.gov.uk/government/publications/malaria-prevention-guidelines-for-travellers-from-the-uk, 2021.
[24] Z. Herrador, B. Fernández-Martinez, V. Quesada-Cubo, O. Diaz-Garcia, R. Cano, A. Benito, D. Gómez-Barroso, Imported cases of malaria in Spain: observational study using nationally reported statistics and surveillance data, 2002–2015, Malar. J. 18 (2019) 230, https://doi.org/10.1186/s12936-019-2863-2.
[25] O.A. Okunlola, O.T. Oyeyemi, Malaria transmission in Africa: its relationship with yellow fever and measles, PLoS One 17 (2022) e0268080, https://doi.org/10.1371/journal.pone.0268080.
[26] F. Castelli, S. Odolini, B. Autino, E. Foca, R. Russo, Malaria prophylaxis: a comprehensive review, Pharmaceuticals 3 (2010) 3212–3239, https://doi.org/10.3390/ph3103212.
[27] B. Genton, V. D'Acremont, Malaria prevention in travelers, Infect. Dis. Clin. North Am. 26 (2012) 637–654, https://doi.org/10.1016/j.idc.2012.05.003.
[28] R.H.H. Groenwold, Informative missingness in electronic health record systems: the curse of knowing, Diagnostic Progn, Res. 4 (2020) 8, https://doi.org/10.1186/s41512-020-00077-0.
[29] E. Garrafa, M. Vezzoli, M. Ravanelli, D. Farina, A. Borghesi, S. Calza, R. Maroldi, Early prediction of in-hospital death of COVID-19 patients: a machine-learning model based on age, blood analyses, and chest x-ray score, Elife 10 (2021) 1–20, https://doi.org/10.7554/eLife.70640.
[30] A. Kishor, C. Chakraborty, Early and accurate prediction of diabetics based on FCBF feature selection and SMOTE, Int. J. Syst. Assur. Eng. Manag. (2021), https://doi.org/10.1007/s13198-021-01174-z.
[31] R. Blagus, L. Lusa, Joint use of over- and under-sampling techniques and cross-validation for the development and assessment of prediction models, BMC Bioinf. 16 (2015) 363, https://doi.org/10.1186/s12859-015-0784-9.