Identification of immune responses predictive of clinical protection against Malaria using novel statistical pipelines

André Fonseca[1,2], Przemyslaw Biecek[3], Clara Cordeiro[1,2] and Nuno Sepúlveda[2,3]

[1] Faculdade de Ciências e Tecnologia, Universidade do Algarve, Faro, Portugal

[2] CEAUL – Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa, Portugal

[3] Faculty of Mathematics & Information Science, Warsaw University of Technology, Warsaw, Poland

***Correspondence:**

André Fonseca

a49406@ualg.pt

Keywords: Immunity, antibodies, vaccine, penalized regression, Random Forest

Abstract

Identifying antibody responses associated with natural immunity to malaria is key to advancing antimalarial vaccine development. With the advent of high-throughput serological assay robust pipelines that produce solid and reproducible results are crucial for the identification of the antibody responses that lead to malaria protection.

Here we have developed two pipelines and assessed their predictive performance on published data from IgG antibody responses against 36 antigens derived from Plasmodium falciparum in 121 Kenyan children with ages below 10 years old. The first pipeline relied on parametric approaches while the second represented a more pragmatic approach to data analysis.

The proposed pipelines enabled us to construct classifiers based on few antibodies, whose performances outperformed previous findings based on Random Forest. The best classifier overall was based on antibodies against the *msp2*, *msp4*, *msp7*, *msp10*, *pf11_0373* and *pf113* antigens and reached a predictive performance of 86% (AUC = 0.86; 95% CI = (0.79-0.93)) using the pragmatic approach. Concerning the parametric approach, our best achievement was a classifier against the *h103*, *msp2*, *msp4*, *msp7* and msrp3 antigens with a predictive performance of 82% (AUC = 0.82; 95% CI = (0.75-0.90)). The good performance of our pipelines suggests their applicability in antibody data analysis intending to identify antimalarial vaccine candidates.

In summary, we were able to identify several antibody responses with high predictive ability against clinical malaria. The proposed pipelines also showed promise to improve the statistical analysis of antibody data aiming to identify antimalarial vaccine candidates.

1. Introduction

Malaria is a disease caused by the Plasmodium parasites and it is endemic to tropical and subtropical regions, where the rates of severe symptoms and lethality are higher in children under 5 years old [1–3]. The vulnerability of these children has been mainly attributed to a slow acquisition of neutralizing antibody responses to malaria parasite[4–6].

Given the putative protective effect of antimalarial antibodies, several studies aimed to identify which antigens elicit the generation of such antibodies[7,8]. However, which parasite-related peptides elicit protective antibody responses to clinical malaria remains to be determined[7,9]. Although the Merozoite Surface Protein 1 (MSP1) and the Apical Membrane Antigen 1 have been suggested as candidate targets for resolving *Plamodium falciparum* (Pf) infections, only the vaccine RTS,S/AS01 (RTS,S) based on the Circumsporozoite protein (CSP) reached phase-3 clinical trials[10–14]. The efficacy of this vaccine has been estimated at around 31% in infants[11]. The fundamental question in this context is to understand why we cannot identify the other key antibody responses that lead to malaria protection.

The current understanding of antimalarial immunity is undermined by the lack of reproducibility across studies [7–9]. This lack of reproducibility can be explained by the complex life cycle, constant evolution, and high genomic diversity of malaria parasites[15–18]. Another factor is the adoption of different techniques to analyse data of hundreds or even thousands of antibody responses in the same individuals[7–9,19,20]. Although being tailored to analyse high dimensional datasets, these techniques are often agnostic about the stochastic nature of serological data, in which multiple serological populations can be present [7,8,21]. In this regard, alternative statistical strategies have been proposed for serological data analysis[22–24], however they have never been applied to the analysis of data from multiple antibody responses.

The present paper aims to propose two statistical pipelines to improve the analysis of high-throughput serological data. The first pipeline (hereafter named as the parametric pipeline) includes both classical statistical methods of data transformation and the use of flexible finite mixture models for the serological status of each individual. A less complex and optimistic pipeline (hereafter names as the pragmatic pipeline) was also developed. It relies on identifying the optimal cut-off point in the antibody distributions that maximises the differences between immune and non-immune individuals. The illustration of these two pipelines was done with a published data on IgG antibody responses to *Pf* antigens in Kenyan children[7].

2. Materials and methods

Data

We have analysed serological data from a previously published study on 286 Kenyan children. These data referred to ELISA-based antibody titters against 36 Pf antigens[25]. Children were monitored for clinical episodes of malaria and classified as Susceptible (Sus) if they had at least one recorded episode of symptomatic malaria (clinical disease). Clinical disease was defined as an axillary temperature > 37.5ºC and any parasitaemia for children less than 1 year, and an axillary temperature of > 37.5ºC and parasitaemia > 2500/µl for children older than 1 year of age, during the 6-month follow-up. Children with no clinical episode were classified as Protected (Prt). The analysis was performed solely on the subset of 121 parasite-positive children at baseline. As followed in the original study [25], forty of these children (33%) were considered susceptible (n_s=40) while the remaining ones were classified as protected (n_p=81, 67%).

Overview of the Parametric pipeline

The identification of antibody responses associated with protection to clinical malaria was initially achieved by developing and applying a parametric pipeline to the dataset. This pipeline was broadly divided into four main phases (Figure 1). The first phase consisted of normality assessment of each antibody distribution. If there was evidence for a normal distribution in a given antibody distribution, a two-sample independent t-test with equal variances was used. Otherwise, flexible finite mixture models based on the Skew-Normal distribution or its extension were used. When using finite mixture models, single population distribution models were also fitted to the data. Pearson goodness-of-fit and Likelihood ratio tests (LRT) were used to evaluate the quality of the fitted models. The Akaike's information criterion (AIC) selected the best model for each antibody data[26]. In antibody distributions where there was evidence for a single serological population, we used linear regression models based on skew-elliptical error terms to compare the two groups of children. In antibody distributions in which there was evidence for multiple serological populations, children were classified into either seronegative or seropositive. After this classification, the chi-square test (χ^2) was performed. The next phase was the predictive analysis in which different classifiers for binary (Logistic/Probit, Ridge, LASSO, Elastic-Net regressions, and the Random Forest) were constructed to predict the clinical status of each children as a function of the selected antibodies. The last phase was the evaluation of the performance of each classifier using the area under the curve (AUC) of the Receiving Operating Characteristic (ROC) curve. A detailed description of each phase can be found below.

Normality assumption

The normality assessment was performed using the Anderson-Darling (AD)[27,28] Jarque-Bera (JB)[29,30] and Shapiro-Wilk (SW)[31] tests. Given that these normality tests have different ways of detecting deviations from a Normal distribution, applying them simultaneously excludes any possible bias that could arise when selecting a single test[27,32]. A non-statistically significant result for all these tests at a 5% significance level, after performing a Bonferroni correction[33] for the number of different statistical tests performed (m=3), was defined as a condition to assume antibody data normality.

Data transformation

The Box-cox transformation was used to obtain the best data transformation of each variable in terms of the normal distribution[34,35]. In general, this transformation is defined in terms of a parameter λ. According to Box G, Cox D (1964)[36] a low lambda value (λ) is recommended, therefore, we consider a λ value between -3 and 3 to represent suitable data transformation[35].

Analysis of antibodies based on finite mixture models

As commonly done in sero-epidmiological studies, finite mixture models were utilised to determine the serological status of the children related to a given antibody[37]. The fitness of the two and three-component mixture models were evaluated under the assumption that antibody distributions could represent a mixture of latent seronegative and seropositive populations[37]. Four classes of mixing distributions were used to fit our data: Normal, Generalized T, Skew-normal and Skew-T. Note that these mixing distributions are special cases of each other[24]. During this step, single component distributions were also fitted to understand whether a single probability distribution could describe the data well. To estimate these models, we used maximum likelihood estimation via the expectation-maximisation (EM) algorithm[38–40]. The AIC was used to select the best model[41,42]. According to this criterion, the best model was the one that provided the lowest estimate among all the models tested to the same data.

Goodness of fit of the Models

Although AIC allows to identify the most parsimonious model, it cannot provide evidence for the respective goodness-of-fit[41,42]. To do that, we used the Pearson's goodness of fit[43] test and the LRT[44] by splitting the data into 5%-quantiles or deciles. Simultaneous acceptance of the null hypothesis at a 0.05 significance level in both regardless of the number of bins considered was required for model acceptance.

Analysis of antibodies with evidence of a single serological population

To analyse antibodies in which there was evidence for a single serological population, we constructed two linear regression models with skew-elliptical error terms[45] using antibody values as the response variable. The first model comprised only the intercept (i.e., not including any covariate). The second model comprised the clinical status as a single covariate. The LRT test was then used to compare the two models at 5% significance level. If there was no difference between the two models (i.e., no effect of clinical status on the respective antibody levels), the antibody at hand was discarded from the analysis. Otherwise, the antibody was carried on to the next phase of the analysis.

Defining antibody seropositivity

For antibody distributions in which there was evidence for mixture models with two or more components, four different thresholds for seropositivity were defined by the following conditions: (i) probability of classifying a seropositive individual (sp) was equal to 0.5 (*50:50*); (ii), (iii), and (iv) the first, second, and third quartile (*q1*, *q2*, and *q3*, respectively) of the estimated distribution for the hypothetical seronegative population, respectively. The hypothetical seronegative population was defined by the component with the lowest mean.

Comparison between immune and non-immune individuals

Two independent samples t-tests for equal variances and χ^2 tests of independence were used to identify antibodies in which there were significant differences between immune and non-immune individuals [46,47]. In particular, the t-test was used to compare antibody mean antibody level between the protected and susceptible when data followed a Normal distribution. The χ^2 test was used to determine if the seroprevalence was different between the two groups of individuals. A 5% significance level was stipulated as evidence of statistically significant differences.

Predictive methodologies:

Both logistic and probit regression models were performed to reduce the chance of the results being affected by a single choice of a link function[48,49]. In these models, the clinical status of the individuals was used as the response variable while the respective covariates were antibody responses selected in the previous phase. For each type of model, a classical stepwise model selection was conducted in to select the best subset of antibodies responses for the prediction of clinical malaria. The Hosmer-Lemeshow (HL) goodness of fit test [50] was used to assess the quality of the fitted models at the 5% significance level. When performing the HL test, the number of bins to calculate quantiles was set to 10.

To complement a stepwise model selection, we also applied Ridge, Least Absolute Shrinkage and Selection Operator (LASSO) and Elastic-Net regressions to the data of the same selected antibodies; these techniques were previously applied to predict immune signatures underlying clinical protection to malaria[51–53]. For the estimation of these regression models, the likelihood function is penalized by a function dependent on a parameter λ. These penalisations enable the shrinkage of coefficient estimates towards 0 for covariates with negligible effect on the outcome[54]. To obtain the optimal parameter λ for model prediction, we specified values for this parameter between 0.001 and 1 with a lag of 0.01. A 10-fold cross-validation was used to compute the accuracy for each value of λ, and this process was repeated one hundred times[54,55]. Usually, two distinct λ values are chosen when performing the Ridge and LASSO regressions. However, when using the *glmnet* R package[56], a single λ value can be selected, and a second parameter α can be adjusted. In this case, α was set at 0 and 1 to perform Ridge and the LASSO regression, respectively[56]. To perform the Elastic-Net regression, we sought the optimal value of α between 0 and 1 increasing by using a lag of 0.1.

We also applied Random Forest to predict the clinical malaria outcome using data from selected antibodies[57]. In general, this popular machine technique constructs multiple decision trees applied on different parts of a training dataset by a process called bagging or bootstrap aggregation[58]. The number of trees to grow and the number of predictors randomly sampled as candidates in each split were set by default (number of trees = 500; number of predictors randomly selected = 3, 14 and 17). We performed one hundred iterations of 10-fold cross-validation to assess the uncertainty of the predictions[59].

The area under the Receive Operating Characteristic curves (AUC)[60] was utilized as a measure of the predictive model accuracies (or discrimination performance)[60,61]. In addition, confusion matrices were used describe the performance of a classification model on a set of data for which the true values are known[62.]The confusion matrices are generated by comparing the predicted classes (i.e., the ones with highest classification probability) against the observed ones of the individuals in the dataset[62].

Pragmatic pipeline

To reduce pipeline complexity while increasing computational feasibility in large serological datasets, a more pragmatic pipeline to the same problem (Figure 2) was constructed. In this pipeline, the statistical analysis of a specific antibody begins by sorting individuals according to the respective antibody levels and specifying each observed value as a possible cut-off to divide individuals into two serological groups. For simplicity of interpretation, we considered the resulting groups as indicative of seropositive or seronegative individuals. In this scenario, seropositive individuals had levels above the cut-off point, otherwise, they were

considered seronegative. Two-by-two contingency tables were then constructed where the rows and columns referred to seropositivity and protection status of the individuals, respectively. Chi-square tests of independence were used to determine if the resulting serological classification was associated with clinical malaria. For each antibody, we selected the cut-off that provided the lowest p-value among all to the Chi-square tests performed for a given antibody. This process was repeated for each of the 36 antibodies in the data set. The Logistic/Probit, Ridge, LASSO, Elastic-Net regressions, and the Random Forest and their respective performance were then used to construct different classifiers for clinical malaria in the same way as explained above for the first pipeline.

Statistical Software

Both pipelines were implemented in the R[63] version 4.0.4 using the following packages: AID[64] to perform Box-cox transformation, caret[57] to implement the predictive models using cross-validation, glmnet[56] to perform the Lasso/Ridge/Elastic-net regression, lmtest[65] to perform the LRT test, MASS[66] to perform stepwise feature selection, mixsnsm[67] to perform mixture models, moments[68] to perform the Jarque-Bera test, nortest[69] to perform the Anderson-Darling test, pROC[70] to perform ROC curves and obtain the AUC, randomForest[71] to perform Random Forest, sn[72] to perform linear regression models based on skew-elliptical error terms. The R scripts are freely available upon request to AF.

3. Results

Application of the parametric pipeline to the data

In the first phase of the pipeline, we found that the distributions of the 36 antibody responses departed significantly from a Normal distribution in the original scale of the measurements (Supplementary Figure S1). This basic finding provided the basis to apply the Box-cox transformation to the data. Data of six antibody responses (*gama*, *msrp3*, *h101*, *ron6*, *pf10_0323*, and *tlp*) provided estimated parameters for the Box-Cox transformation outside the eligible interval (λ = -3.78, -3.40, -3.49, -4.66, -5.20, and -9.53, respectively (Supplementary Figure S2). Then, the normality assessment was only performed for the remaining 30 antibody responses. The results suggested that a Normal distribution described data from antibody responses against *asp*, *msrp2*, *mtrap*, *pf11_0373*, *pf14_0344*, *pf34*, and *rh5* (Supplementary Figure S3). The antibody responses against *asp*, *pf11_0373*, *pf34*, and *rh5* showed significant differences between immune and non-immune individuals (p=0.016, 0.013, 0.007, and 0.048, respectively; t-test for two independent groups with equal variances). Therefore, they were selected for the predictive analysis (Supplementary Figure S4). The remaining four antibodies

(*msrp2*, *mtrap* and *pf14_0344*) were excluded from the analysis due to no statistically significant differences between the two groups.

Twenty-three antibody responses did not follow a Normal distribution even after the Box-Cox transformation of the data. The respective data, together with the data for the six antibodies with non-eligible Box-Cox transformation parameters were then log10-transformed and analysed by finite mixture models based on the Normal, Generalized T, Skew-normal, and Skew-T distributions. We also estimated these models under the assumption of a single population. Data of antibody responses against *gama*, *msp3*, *msrp3*, *h101*, *h103*, and *pf10_0323* were described well by a single population distribution (Table 1). Of these antibody responses, only those against *gama*, *h103*, *msp3*, and *msrp3* were associated with clinical protection (p=0.010, 0.004, 0.004, and 0.0006, respectively; LRT). Therefore, these four antibody responses were selected for the predictive analysis.

With the exception of the antibody responses against pf92, data of the remaining antibody responses were better described by a mixture of two or more serological populations (Table 1). For these antibody responses, we assumed a seronegative and a seropositive population for the two-component mixture models. In turn, we assumed a seronegative population and two seropositive populations for data with evidence of three serological populations. The overall antibody seroprevalence can be found in Table 2 for the different cut-off methods used.

When the *50:50* cut-off method was applied to the data, antibody responses to eight antibodies were significantly associated with clinical malaria (Table 3). Using these eight antibodies, we constructed another set with the individuals seropositivity probabilities given by the Finite Mixture Models, designated *sp*. Eight, eleven, and twelve antibody responses were found to be associated with clinical malaria using the *q1*, *q2* and *q3* categorization methods, respectively (Table 3). These antibody responses were selected for the predictive analyses. The remaining ones were discarded at this stage of the analysis. The only problematic data was the related to the antibody responses against pf92. In this case, the best finite mixture model could not describe the data well according to the Pearson's goodness-of-fit or LRT tests (Supplementary Table 1). Data from this antibody was then discarded from further analysis.

The predictive analysis proceeded using data of the above antibody responses selected by different models and tests. In this analysis, the predictive performance was conducted by combining two types of variables: numeric antibody responses using appropriate data transformation, and binary antibody responses according to different criteria for the serological classification (seropositive and seronegative).

The best set of antibody responses for predicting protection was obtained using Logistic and Probit regressions after a stepwise variable selection procedure. Given the similarity of these classifiers, we are only presenting the results for the Logistic one. This classifier was based on data of antibody responses against the *h103*, *msp2*, *msp4*, *msp7* and *msrp3* antigens. The

respective AUC was estimated at 0.82 (95% CI = (0.75-0.90)) for the *sp* set of antibodies (Figure 3A and Table 4). To complement this analysis, we included age in this classifier as an additional covariate. The predictive performance improved to AUC=0.86 (95% CI = (0.80-0.94)) (Figure 3A) providing a sensitivity=0.93 and specificity=0.65. Overall, this classifier predicted the correct clinical status of 101 individuals (Figure 3B). Both logistic models including or not age as a covariate seemed to fit the data well (HL test, p =0.323 and 0.937, respectively).

When using cross-validation in the predictive analysis, the best classification model was equal to the one without cross-validation. The AUC for *sp* set of antibodies was estimated at 0.74 and the respective fit to the data was acceptable (p=0.929; HL test).

Notwithstanding the good predictive performance of the regression models, we also decided to perform the predictive analysis using Ridge, LASSO and Elastic-Net regressions. A model based on the Elastic-regression (α=0.4) was the one that performed best among the three regularisation methods utilised. This model reached an AUC close to 77% for the optimal λ and *sp* method for the dichotomization of the antibody data (Figure 4A). The most important covariates for this model were the antibody responses against *msp4*, *msp7*, *h103* and *msp2* (Figure 4B). Of note, these covariates were also the most important ones in the Logistic regression model.

Finally, we applied a Random Forest approach to the same data. The best classifier reached an AUC of 0.72 using data for the antibody responses based on the q2 method (Supplementary figure S5A). The most important antibody responses for predicting protection were largely different to those using the Logistic and Elastic-Net regressions. The antibodies against *h103*, were considered the most important ones followed by antibodies against *gama*, *pf34*, *msrp3* and *msp3*. M*sp4* antibody was in the middle of the table of importance, while *msp7* and *msp2* were the antibodies with the lowest importance overall.

Application of the pragmatic pipeline to the data

We then constructed the pragmatic pipeline to the same data. The basic idea was to sort data of a given antibody response to determine the value that maximized the discrimination between immune and non-immune individuals according to the χ^2 test for two-way contingency tables. This step identified 28 antibody responses that could significantly distinguish these two groups of individuals (Table 5). Data on these antibody responses were carried out to the predictive phase, as done for the parametric pipeline.

In this predictive analysis, the best classifier was based on the logistic regression with the antibodies against the *msp2*, *msp4*, *msp7*, *msp10*, *pf11_0373* and *pf113* as covariates. This model had an estimated AUC of 0.86 (95% CI = (0.79-0.93) (Figure 5A) and it provided good fit to the data (p=0.93; HL test). When we included age as an additional covariate, the respective AUC increased to 0.89 (95% CI = (0.82-0.95)) (Figure 5A). This extended model also provided a good

fit to the data (p=0.35; HL test) and it could correctly classify 101 out of 121 individuals (Figure 5B) providing values of 0.91 and 0.68 for sensitivity and specificity respectively.

Once more, we used cross-validation to assess the performance of the predictive models. Again, the best model was based on antibody responses to *msp2*, *msp4*, *msp7*, *msp10*, *pf11_0373* and *pf113* antigens. However, the accuracy decreased to approximately 72% while maintaining a good fit to the data (p=0.90; HL test).

As done in the parametric pipeline, we also used Ridge, LASSO and Elastic-Net regression. In this regard, Ridge regression led to the best classifier. It reached an AUC of 0.78 using λ within the range 0.94-1.00 (Supplementary figure S6A). As for the traditional logistic regression model, the antibody responses to *msp7*, *msp4*, *msp10*, and *pf11_0373* had the highest importance in the best Regression model (importance = 100.0, 87.0, 78.3, and 68.5, respectively; Supplementary figure S6B). The exception was the antibody response against *msrp3*, which had an importance close to 0 (Supplementary figure 6B).

Concerning the Random Forest strategy, we obtained an AUC of 0.78 (Figure 6A) using all the 28 antibody responses selected in the first stage of the pipeline but msrp3 (Figure 6B). The most important antibody responses were the ones against *msp7*, *eba175*, *msp10*, *h103*, *msp4* and *msp2* (importance = 100.0, 57.8, 56.1, 54.8, 49.5 and 47.4 respectively; Figure 6B). In contrast to Ridge regression, the antibody responses against pf11_0373 had a lower importance using this alternative analytical strategy (importance= 21.8).

4 Discussion:

In summary, we have implemented two pipelines to analyse multiple antibody data. These pipelines were developed under a solid statistical modelling foundation that coupled classical traditional statistical techniques for variable selection with popular machine learning techniques for the predictive analysis.

The application of the parametric pipeline allowed us to construct a good predictor of clinical protection based on antibody responses to *msp2*, *msp4*, *msp7*, *msrp3* and *h103*. The accuracy of this predictor could be improved by using the age of the individuals as covariate. Given that age is a proxy of immunity acquisition to malaria parasites, the classifiers' improved performance suggests that additional but unmeasured antibody responses remain to be identified among natural factors contributing to protection to clinical malaria. Therefore, we can speculate that the predictions from our classifiers could be improved on data with a wider range of antibody responses.

Some reduction in the predictive accuracy was observed when including the cross-validation stage in both pipelines. This result suggested that the initial good predicting could be tightly linked to the dataset under analysis. Given the differences between estimated accuracies using or not cross-

validation stages, we can speculate that the sample size of this dataset was a limiting factor for this analysis. In this regard, increasing the sample size in future studies is essential to increase reproducibility of findings.

Besides classical logistic regression models, we decided to include alternative regression models from the field of machine learning. The inclusion of these techniques was motivated by three major reasons. Firstly, we aimed to create general pipelines in which it was possible to select a large number of antibody responses in the first stage. However, in traditional logistic regression, it is recommended to use the rule of having at least 10 events per predictor variable[73,74]. In this regard, the use of Ridge, LASSO or Elastic-Net regression are simple solutions that allows to include a large number of covariates in the models without the necessity of using time-consuming stepwise variable selection procedures. However the number of antibody responses analysed was not so high, and therefore, traditional stepwise variable selection procedures were feasible to perform in the predictive stage. We envision that these alternative regression approaches will be more useful in datasets in which the number of antibody responses exceeds the number of observations, as in a study from Mali[7]. A possible solution to reduce the selected number of antibody responses in the first stage is to use a multiple testing correction. This solution and its impact on variable selection will be specifically investigated in a separate study. Secondly, machine learning techniques in general are expected to deal better with complex interactions between antibody responses and the outcome of interest than traditional logistic regression. However, this expectation could not be confirmed in the dataset analysed. This might be explained by the modest number of antibody responses selected from the first stage. Another possible explanation is that we used the default and not the optimal hyperparameters when applying these techniques. In this regard, it is important to note that hyperparameter optimization is a hot topic in the field of machine learning. In this sense, our pipeline could be improved by using automated machine learning strategies such as autoML[75] and SkLearn[76] in the predictive stage. Thirdly, the use of multiple strategies for constructing a classification model can help to understand whether there is any model dependency in terms of antibody selection. In this regard, our findings suggested that different strategies may lead to a slightly different antibody selection. Although the most relevant antibodies among the different approaches were somehow similar, there were instances where the importance of some antibodies changed more dramatically. These suggest that the antibody panels identified depend on the strategy used.

We proposed a parametric pipeline and a pragmatic one. Given the distinct antibody selection approaches used in their initial stages, it was unsurprising that they did not select the same set of antibody responses for predicting malaria protection. Given its optimistic nature, the pragmatic pipeline, which aimed to maximize the discrimination between immune and non-immune individuals, had unsurprisingly a better predictive performance than the parametric one. Therefore, the pragmatic pipeline can be seen as the optimistic case scenario. However, this

approach is highly dependent on the calculation of the optimal cut-off value for each antibody. Therefore, small perturbations in the data might influence this calculation, which reduces the chance of obtaining the same antibody selection and classifier.

Previously, Valletta and Recker[7] analysed the same dataset using a Random Forest approach. The respective accuracy was 68% using data of all the antibody responses. In this scenario, the pipelines here proposed improved this previous prediction of malaria protection. This result can be partly explained by inclusion an initial antibody selection stage in our pipelines. On the one hand, this stage removed antibody responses with negligible effect on clinical malaria. On the other hand, this stage increased the chance of finding the right data transformation and dichotomization for each antibody response. A reduced predictive performance reported by Valletta and Recker[7] could also be explained by using the default hyperparameters when applying a Random Forest approach to the data. Therefore, a pure Random Forest approach to the data could have a comparable predictive performance to our pipelines if the optimal hyperparameters had been used instead.

We found that the antibody responses against different Merozoite Surface Proteins (MSPs) were consistently selected across our pipelines. These proteins are expressed at the parasite surface, thus, providing promising targets for malaria immunity, because they are repeatedly and directly exposed to the host humoral immune system[77,78]. In particular, *msp2* has been associated with protection from clinical malaria in many studies and even suggested as a vaccine candidate[79–82]. For example, *msp2* has been strongly associated with protection against clinical malaria in two independent cohorts of Kenyan children[8]. *Msp4* has also been reported to have a protective effect in Kenyan children[25,83]. Additionally, high antibody levels against *msp4* constructs have been associated with reduced morbidity in a Senegalese community[84]. *Msp7* protection against malaria has also already been identified in the Kenyan population[83,85]. Moreover, panels of antibodies comprising *msp7* have been associated with clinical protection against malaria in Kilifi, a rural district along the Kenyan coast[25]. In the same article, high antibody levels of antibodies against the Erythrocyte-binding antigen-175 (*eba175*) antigen were also associated with protection from clinical malaria in children[25]. Moreover, *eba175* associated protection from symptomatic malaria has been demonstrated in Papua New Guinean children[86]. Concerning *pf113* antigen, a longitudinal study in a cohort of Kenyan children has identified ten antigens, among which *pf113* were associated with protection against clinical episodes of malaria[25]. Furthermore, several other studies refer to *pf113* as a promising malaria vaccine candidate[87,88].

These findings further corroborate our results. Interestingly however, *msp10* and *h103* have not been associated with clinical malaria protection. This evidence suggests that there are antibodies associated with protection against clinical malaria that have not yet been identified. Nevertheless, further studies are necessary to validate our analysis. Finally, immune responses

commonly associated with malaria protection and often referred to as potential vaccine candidates such as the *msp1* and the *ama1* were not among the best predictors of clinical protection malaria in children. These have already been reported in other studies, where they have been described to show low or no associations with protection to clinical malaria[7]. These inconsistent findings further suggest the need for constructing robust pipelines that could help increasing reproducibility among studies.

In conclusion, these pipelines showed potential for analysing data from multiple antibodies in the context of malaria immunity and vaccine development. To understand their scope and computational feasibility, they should be tested in multiple datasets. Finally, since the pipelines were constructed under general ideas related to serological data analysis, we envision that these pipelines have also the potential to be deployed to high-throughput serological data from infectious diseases other than malaria.

Acknowledgements

André Fonseca has a PhD fellowship by FCT – Fundação para a Ciência e Tecnologia (ref. SFRH/BD/147629/2019). Clara Cordeiro and Nuno Sepúlveda are partially financed by national funds through FCT under the project UIDB/00006/2020. Nuno Sepúlveda is funded by Polish National Agency for Academic Exchange (ref. grant: PPN/ULM/2020/1/00069/U/00001).

Bibliography:

1. Talapko J, Škrlec I, Alebić T, Jukić M, Včev A. Malaria: The Past and the Present. *Microorganisms*. 2019;7(6):179. doi:10.3390/microorganisms7060179

2. Ashley EA, Phyo AP, Woodrow CJ. Malaria. *The Lancet*. 2018;391(10130):1608-1621. doi:10.1016/S0140-6736(18)30324-6

3. Greenwood BM, Fidock DA, Kyle DE, et al. Malaria: progress, perils, and prospects for eradication. *Journal of Clinical Investigation*. 2008;118(4):1266-1276. doi:10.1172/JCI33996

4. Moormann AM. How might infant and paediatric immune responses influence malaria vaccine efficacy? *Parasite Immunol*. 2009;31(9):547-559. doi:10.1111/j.1365-3024.2009.01137.x

5. Doolan DL, Dobaño C, Baird JK. Acquired Immunity to Malaria. *Clin Microbiol Rev*. 2009;22(1):13-36. doi:10.1128/CMR.00025-08

6. Barry A, Hansen D. Naturally acquired immunity to malaria. *Parasitology*. 2016;143(2):125-128. doi:10.1017/S0031182015001778

7. Valletta JJ, Recker M. Identification of immune signatures predictive of clinical protection from malaria. *PLoS Comput Biol*. 2017;13(10):e1005812. doi:10.1371/journal.pcbi.1005812

8. Osier FHA, Fegan G, Polley SD, et al. Breadth and Magnitude of Antibody Responses to Multiple *Plasmodium falciparum* Merozoite Antigens Are Associated with Protection from Clinical Malaria. *Infect Immun*. 2008;76(5):2240-2248. doi:10.1128/IAI.01585-07

9. Proietti C, Krause L, Trieu A, et al. Immune Signature Against Plasmodium falciparum Antigens Predicts Clinical Immunity in Distinct Malaria Endemic Communities. *Molecular & Cellular Proteomics*. 2020;19(1):101-113. doi:10.1074/mcp.RA118.001256

10. Blank A, Fürle K, Jäschke A, et al. Immunization with full-length Plasmodium falciparum merozoite surface protein 1 is safe and elicits functional cytophilic antibodies in a randomized first-in-human trial. *NPJ Vaccines*. 2020;5(1):10. doi:10.1038/s41541-020-0160-2

11. Laurens MB. RTS,S/AS01 vaccine (Mosquirix™): an overview. *Hum Vaccin Immunother*. 2020;16(3):480-489. doi:10.1080/21645515.2019.1669415

12. Thera MA, Doumbo OK, Coulibaly D, et al. Safety and Immunogenicity of an AMA1 Malaria Vaccine in Malian Children: Results of a Phase 1 Randomized Controlled Trial. *PLoS One*. 2010;5(2):e9041. doi:10.1371/journal.pone.0009041

13. Sirima SB, Durier C, Kara L, et al. Safety and immunogenicity of a recombinant Plasmodium falciparum AMA1-DiCo malaria vaccine adjuvanted with GLA-SE or Alhydrogel® in European and African adults: A phase 1a/1b, randomized, double-blind multi-centre trial. *Vaccine*. 2017;35(45):6218-6227. doi:10.1016/j.vaccine.2017.09.027

14. Chitnis CE, Mukherjee P, Mehta S, et al. Phase I Clinical Trial of a Recombinant Blood Stage Vaccine Candidate for Plasmodium falciparum Malaria Based on MSP1 and EBA175. *PLoS One*. 2015;10(4):e0117820. doi:10.1371/journal.pone.0117820

15. Aly ASI, Vaughan AM, Kappe SHI. Malaria Parasite Development in the Mosquito and Infection of the Mammalian Host. *Annu Rev Microbiol*. 2009;63(1):195-221. doi:10.1146/annurev.micro.091208.073403

16. Siciliano G, Alano P. Enlightening the malaria parasite life cycle: bioluminescent Plasmodium in fundamental and applied research. *Front Microbiol*. 2015;6. doi:10.3389/fmicb.2015.00391

17. Loy DE, Liu W, Li Y, et al. Out of Africa: origins and evolution of the human malaria parasites Plasmodium falciparum and Plasmodium vivax. *Int J Parasitol*. 2017;47(2-3):87-97. doi:10.1016/j.ijpara.2016.05.008

18. Su XZ, Zhang C, Joy DA. Host-Malaria Parasite Interactions and Impacts on Mutual Evolution. *Front Cell Infect Microbiol*. 2020;10. doi:10.3389/fcimb.2020.587933

19. Ahsan H, Ahmad R. Multiplex Technology for Biomarker Immunoassays. In: *Innate Immunity in Health and Disease*. IntechOpen; 2021. doi:10.5772/intechopen.91730

20. Ellington AA, Kullo IJ, Bailey KR, Klee GG. Antibody-Based Protein Multiplex Platforms: Technical and Operational Challenges. *Clin Chem*. 2010;56(2):186-193. doi:10.1373/clinchem.2009.127514

21. França CT, White MT, He WQ, et al. Identification of highly-protective combinations of Plasmodium vivax recombinant proteins for vaccine development. *Elife*. 2017;6. doi:10.7554/eLife.28673

22. Sepúlveda N, Stresman G, White MT, Drakeley CJ. Current Mathematical Models for Analyzing Anti-Malarial Antibody Data with an Eye to Malaria Elimination and Eradication. *J Immunol Res*. 2015;2015:738030. doi:10.1155/2015/738030

23. Kyomuhangi I, Giorgi E. A unified and flexible modelling framework for the analysis of malaria serology data. *Epidemiol Infect*. 2021;149:e99. doi:10.1017/S0950268821000753

24. Domingues TD, Mouriño H, Sepúlveda N. Analysis of antibody data using Finite Mixture Models based on Scale Mixtures of Skew-Normal distributions. Published online 2021. doi:10.1101/2021.03.08.21252807

25. Osier FH, Mackinnon MJ, Crosnier C, et al. New antigens for a multicomponent blood-stage malaria vaccine. *Sci Transl Med*. 2014;6(247). doi:10.1126/scitranslmed.3008705

26. Bozdogan H. Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika*. 1987;52(3):345-370. doi:10.1007/BF02294361

27. Yap BW, Sim CH. Comparisons of various types of normality tests. *J Stat Comput Simul*. 2011;81(12):2141-2155. doi:10.1080/00949655.2010.520163

28. Arshad M, Rasool MT, Ahmad MI. Anderson Darling and Modified Anderson Darling Tests for Generalized Pareto Distribution. *Journal of Applied Sciences*. 2003;3(2):85-88. doi:10.3923/jas.2003.85.88

29. Jarque CM, Bera AK. A Test for Normality of Observations and Regression Residuals. *Int Stat Rev*. 1987;55(2):163. doi:10.2307/1403192

30. Thadewald T, Büning H. Jarque–Bera Test and its Competitors for Testing Normality – A Power Comparison. *J Appl Stat*. 2007;34(1):87-105. doi:10.1080/02664760600994539

31. Shapiro SS, Wilk MB. An Analysis of Variance Test for Normality (Complete Samples). *Biometrika*. 1965;52(3/4):591. doi:10.2307/2333709

32. Mendes M, Pala A. Type I Error Rate and Power of Three Normality Tests. *Information Technology Journal*. 2003;2(2):135-139. doi:10.3923/itj.2003.135.139

33. Armstrong RA. When to use the Bonferroni correction. *Ophthalmic and Physiological Optics*. 2014;34(5):502-508. doi:10.1111/opo.12131

34. Sakia RM. The Box-Cox Transformation Technique: A Review. *The Statistician*. 1992;41(2):169. doi:10.2307/2348250

35. Box GEP, Cox DR. An Analysis of Transformations Revisited, Rebutted. *J Am Stat Assoc*. 1982;77(377):209-210. doi:10.1080/01621459.1982.10477788

36. Box GEP, Cox DR. An Analysis of Transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*. 1964;26(2):211-243. doi:10.1111/j.2517-6161.1964.tb00553.x

37. Sepúlveda N, Stresman G, White MT, Drakeley CJ. Current Mathematical Models for Analyzing Anti-Malarial Antibody Data with an Eye to Malaria Elimination and Eradication. *J Immunol Res*. 2015;2015:1-21. doi:10.1155/2015/738030

38. Fraley C, Raftery AE. Model-Based Clustering, Discriminant Analysis, and Density Estimation. *J Am Stat Assoc*. 2002;97(458):611-631. doi:10.1198/016214502760047131

39. Banfield JD, Raftery AE. Model-Based Gaussian and Non-Gaussian Clustering. *Biometrics*. 1993;49(3):803. doi:10.2307/2532201

40. McLachlan GJ, Lee SX, Rathnayake SI. Finite Mixture Models. *Annu Rev Stat Appl*. 2019;6(1):355-378. doi:10.1146/annurev-statistics-031017-100325

41. Posada D, Buckley TR. Model Selection and Model Averaging in Phylogenetics: Advantages of Akaike Information Criterion and Bayesian Approaches Over Likelihood Ratio Tests. *Syst Biol*. 2004;53(5):793-808. doi:10.1080/10635150490522304

42. Cavanaugh JE, Neath AA. The Akaike information criterion: Background, derivation, properties, application, interpretation, and refinements. *WIREs Computational Statistics*. 2019;11(3). doi:10.1002/wics.1460

43. Smyth GK. Pearson's goodness of fit statistic as a score test statistic. In: ; 2003:115-126. doi:10.1214/lnms/1215091138

44. Jiang W, Zhang CH. Generalized likelihood ratio test for normal mixtures. *Stat Sin*. Published online 2016. doi:10.5705/ss.202015.0086

45. Jeong J, Chung Y. Bayesian inference for an ordered multiple linear regression with skew normal errors. *Commun Stat Appl Methods*. 2020;27(2):189-199. doi:10.29220/CSAM.2020.27.2.189

46. McHugh ML. The Chi-square test of independence. *Biochem Med (Zagreb)*. Published online 2013:143-149. doi:10.11613/BM.2013.018

47. Gronow DGC. Non-normality in two-sample *t* -tests. *Biometrika*. 1953;40(1-2):222-225. doi:10.1093/biomet/40.1-2.222

48. Ruspriyanty DI, Sofro A. Analysis of Hypertension Disease using Logistic and Probit Regression. *J Phys Conf Ser*. 2018;1108:012054. doi:10.1088/1742-6596/1108/1/012054

49. Begg MD. An introduction to categorical data analysis (2nd edn). Alan Agresti, John Wiley & Sons, Inc., Hoboken, New Jersey, 2007. No. of Pages: 400. Price: $100.95. ISBN: 978-0-471-22618-5. *Stat Med*. 2009;28(11):1643-1643. doi:10.1002/sim.3564

50. Fagerland MW, Hosmer DW. A Generalized Hosmer–Lemeshow Goodness-of-Fit Test for Multinomial Logistic Regression Models. *The Stata Journal: Promoting communications on statistics and Stata*. 2012;12(3):447-453. doi:10.1177/1536867X1201200307

51. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Series B Stat Methodol*. 2005;67(2):301-320. doi:10.1111/j.1467-9868.2005.00503.x

52. Tibshirani R. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*. 1996;58(1):267-288. doi:10.1111/j.2517-6161.1996.tb02080.x

53. Hoerl AE, Kennard RW. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*. 2000;42(1):80. doi:10.2307/1271436

54. McNeish DM. Using Lasso for Predictor Selection and to Assuage Overfitting: A Method Long Overlooked in Behavioral Sciences. *Multivariate Behav Res*. 2015;50(5):471-484. doi:10.1080/00273171.2015.1036965

55. Melkumova LE, Shatskikh SYa. Comparing Ridge and LASSO estimators for data analysis. *Procedia Eng*. 2017;201:746-755. doi:10.1016/j.proeng.2017.09.615

56. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw*. 2010;33(1). doi:10.18637/jss.v033.i01

57. Kuhn M. caret: Classification and Regression Training. Published online 2022. Accessed May 27, 2022. https://CRAN.R-project.org/package=caret

58. Sarica A, Cerasa A, Quattrone A. Random Forest Algorithm for the Classification of Neuroimaging Data in Alzheimer's Disease: A Systematic Review. *Front Aging Neurosci*. 2017;9. doi:10.3389/fnagi.2017.00329

59. Ahmed FYH, Ali YH, Shamsuddin SM. Using K-Fold Cross Validation Proposed Models for Spikeprop Learning Enhancements. *International Journal of Engineering & Technology*. 2018;7(4.11):145. doi:10.14419/ijet.v7i4.11.20790

60. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit*. 1997;30(7):1145-1159. doi:10.1016/S0031-3203(96)00142-2

61. Tripepi G, Jager KJ, Dekker FW, Zoccali C. Diagnostic methods 2: receiver operating characteristic (ROC) curves. *Kidney Int*. 2009;76(3):252-256. doi:10.1038/ki.2009.171

62. Düntsch I, Gediga G. Confusion Matrices and Rough Set Data Analysis. *J Phys Conf Ser*. 2019;1229(1):012055. doi:10.1088/1742-6596/1229/1/012055

63. R Core Team. R: A Language and Environment for Statistical Computing. Published online 2022. Accessed February 27, 2022. https://www.R-project.org/

64. Asar Ö, Ilk O, Dag O. Estimating Box-Cox power transformation parameter via goodness-of-fit tests. *Commun Stat Simul Comput*. 2017;46(1):91-105. doi:10.1080/03610918.2014.957839

65. Zeileis A, Hotborn T. Diagnostic Checking in Regression Relationships. *R news*. 2002;2:7-10. Accessed April 21, 2022. https://CRAN.R-project.org/doc/Rnews/

66. Venables WB, Ripley BD. *Modern Applied Statistics with S*. Fourth.; 2002. Accessed April 24, 2022. https://www.stats.ox.ac.uk/pub/MASS4/

67. Prates MO, Cabral CRB, Lachos VH. mixsmsn : Fitting Finite Mixture of Scale Mixture of Skew-Normal Distributions. *J Stat Softw*. 2013;54:1-20. Accessed April 29, 2022. https://doi.org/10.18637/jss.v054.i12

68. Komsta L, Novomestky F. moments: Moments, cumulants, skewness, kurtosis and related tests. Published online 2015. Accessed March 19, 2022. https://CRAN.R-project.org/package=moments

69. Gross J, Ligges U. nortest: Tests for Normality. Published online 2015. Accessed April 29, 2022. https://CRAN.R-project.org/package=nortest

70. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*. 2011;12:77.

71. Breiman L. Random Forests. *Mach Learn*. 2001;45(1):5-32. doi:10.1023/A:1010933404324

72. Azzalini A. The R package “sn”: The Skew-Normal and Related Distributions such as the Skew-t and the SUN (version 2.0.2). Published online 2021. Accessed May 1, 2022. https://cran.r-project.org/package=sn

73. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol*. 1996;49(12):1373-1379. doi:10.1016/S0895-4356(96)00236-3

74. Peduzzi P, Concato J, Feinstein AR, Holford TR. Importance of events per independent variable in proportional hazards regression analysis II. Accuracy and precision of regression estimates. *J Clin Epidemiol*. 1995;48(12):1503-1510. doi:10.1016/0895-4356(95)00048-8

75. Feurer M, Eggensperger K, Falkner S, Lindauer M, Hutter F. Auto-Sklearn 2.0: Hands-free AutoML via Meta-Learning. Published online July 8, 2020. http://arxiv.org/abs/2007.04074

76. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*. 2011;12:2825-2830.

77. Lin CS, Uboldi AD, Epp C, et al. Multiple Plasmodium falciparum Merozoite Surface Protein 1 Complexes Mediate Merozoite Binding to Human Erythrocytes. *Journal of Biological Chemistry*. 2016;291(14):7703-7715. doi:10.1074/jbc.M115.698282

78. Jäschke A, Coulibaly B, Remarque EJ, Bujard H, Epp C. Merozoite Surface Protein 1 from Plasmodium falciparum Is a Major Target of Opsonizing Antibodies in Individuals with Acquired Immunity against Malaria. *Clinical and Vaccine Immunology*. 2017;24(11). doi:10.1128/CVI.00155-17

79. Cortés A, Mellombo M, Benet A, Lorry K, Rare L, Reeder JC. Plasmodium falciparum: distribution of msp2 genotypes among symptomatic and asymptomatic individuals from the Wosera region of Papua New Guinea. *Exp Parasitol*. 2004;106(1-2):22-29. doi:10.1016/j.exppara.2004.01.010

80. McCarthy JS, Marjason J, Elliott S, et al. A Phase 1 Trial of MSP2-C1, a Blood-Stage Malaria Vaccine Containing 2 Isoforms of MSP2 Formulated with Montanide® ISA 720. *PLoS One*. 2011;6(9):e24413. doi:10.1371/journal.pone.0024413

81. Engelbrecht F, Felger I, Genton B, Alpers M, Beck HP. Plasmodium falciparum: Malaria Morbidity Is Associated with Specific Merozoite Surface Antigen 2 Genotypes. *Exp Parasitol*. 1995;81(1):90-96. doi:10.1006/expr.1995.1096

82. Ariey F, Hommel D, Le Scanf C, et al. Association of Severe Malaria with a Specific *Plasmodium falciparum* Genotype in French Guiana. *J Infect Dis*. 2001;184(2):237-241. doi:10.1086/322012

83. Dent AE, Nakajima R, Liang L, et al. *Plasmodium falciparum* Protein Microarray Antibody Profiles Correlate With Protection From Symptomatic Malaria in Kenya. *Journal of Infectious Diseases*. 2015;212(9):1429-1438. doi:10.1093/infdis/jiv224

84. Perraut R, Varela ML, Joos C, et al. Association of antibodies to Plasmodium falciparum merozoite surface protein-4 with protection against clinical malaria. *Vaccine*. 2017;35(48):6720-6726. doi:10.1016/j.vaccine.2017.10.012

85. Medeiros MM, Fotoran WL, dalla Martha RC, Katsuragawa TH, Pereira da Silva LH, Wunderlich G. Natural antibody response to Plasmodium falciparum merozoite antigens MSP5, MSP9 and EBA175 is associated to clinical protection in the Brazilian Amazon. *BMC Infect Dis*. 2013;13(1):608. doi:10.1186/1471-2334-13-608

86. Richards JS, Stanisic DI, Fowkes FJI, et al. Association between Naturally Acquired Antibodies to Erythrocyte-Binding Antigens of *Plasmodium falciparum* and Protection from Malaria and High-Density Parasitemia. *Clinical Infectious Diseases*. 2010;51(8):e50-e60. doi:10.1086/656413

87. Imboumy-Limoukou RK, Oyegue-Liabagui SL, Ndidi S, et al. Comparative antibody responses against three antimalarial vaccine candidate antigens from urban and rural exposed individuals in Gabon. *Eur J Microbiol Immunol (Bp)*. 2016;6(4):287-297. doi:10.1556/1886.2016.00027

88. Chia WN, Goh YS, Rénia L. Novel approaches to identify protective malaria vaccine candidates. *Front Microbiol*. 2014;5. doi:10.3389/fmicb.2014.00586

Figure Legends:

Figure 1. Parametric Pipeline. The different analysis steps are displayed on the workflow using distinct grey-scale coloured shapes. Dark grey colour identifies the Normality assessment step. White shapes with outlines indicate the two-sample t-test performed on the antibodies found to follow a Normal distribution. Light grey shapes with no outline refer to the Finite Mixture models used to identify the underlying distributions of then non-normal antibodies. Light grey with outline refers to the predictive performance step, and black indicates antibodies removed from the analysis. The dotted white shapes provide additional information.

Figure 2. Pragmatic pipeline. The different steps of the analysis are displayed on the workflow using distinct grey-scale coloured shapes. Dark grey colour identifies the beginning of the pipeline. Light grey with no outline indicates computational steps prior and after the loop for obtaining the χ^2 test p-value for each potential cut-off (light grey with no outline). Light grey with outline refers to the predictive performance step, and black indicates antibodies removed from the analysis. The dotted white shapes provide additional information.

Figure 3. Predictive performance of the best classifiers for the comprehensive pipeline. (A) ROC curve for the best Logistic regression models including antibody responses to *h103*, *msp2*, *msp4*, *msp7*, and *msrp3* (black), only age (light grey), and the best antibody responses together with age (dark grey) as covariates. (B) Confusion matrices of the best Logistic regression model including data of Abs and Age shown in A. The estimated probability cut-off used was 0.5

Figure 4. Elastic-net regression for the predicting stage of Comprehensive pipeline regularisation strategy results. (A) The mean accuracy for each regularisation parameter (λ) for an $\alpha = 0.4$ after one hundred runs of 10-fold cross-validation are given by a grey circle. (B) Importance of each antibody in the model.

Figure 5. Predictive performance of the best classifiers using the pragmatic pipeline. (A) ROC curve for the best Logistic regression models including antibody responses to *msp2*, *msp4*, *msp7*, *msp10*, *pf11_0373*, and *pf113* (black), only age (light grey), and the best antibody responses together with age (dark grey) as covariates. (B) Confusion matrices of the best Logistic regression model including data of Abs and Age shown in A. The estimated probability cut-off used was 0.5.

Figure 6. Predictive performance of the Random Forest results using the pragmatic pipeline. (A) The mean accuracy for each value of randomly selected predictors when performing each tree after one hundred runs of 10-fold cross-validation are given by a grey circle. (B) Importance of each antibody in the final model.

Tables:

Table 1

Mixture-Model and goodness of fit results. The BIC and AIC values for each model, according to the type of distribution (Model) and the number components (Num.comp). The Pearson goodness-of-fit and Likelihood ratio test *p-values* according to the number of bins are provided. P.10 denotes the data was split in 10 equal size bins containing 10% of the data, while P.5 indicates the data was split in 20 bins of equal size (i.e., containing 5% of the data).

Antibody	Model	Num.Comp	BIC	AIC	Pearson P.10	Pearson P.5	LRT P.10	LRT P.5
ama1	mixNormal	3	101.91	79.55	0.01	0.08	<0.01	0.15
eba140	mixSkewNormal	2	43.49	23.92	0.32	0.88	0.32	0.89
eba175	mixSkewNormal	2	52.86	33.29	0.03	0.17	0.04	0.29
eba181	mixSkewNormal	2	62.47	42.9	0.03	0.20	0.05	0.23
gama	SkewT	1	-261	-272.19	0.24	0.75	0.23	0.78
h101	SkewT	1	-219.73	-230.91	0.33	0.62	0.34	0.66
h103	SkewNormal	1	-33.52	-41.91	0.72	<0.01	0.71	0.06
msp1	mixSkewNormal	2	44.92	25.35	0.26	0.08	0.26	0.19
msp10	mixNormal	2	85.5	71.52	0.07	<0.01	0.09	0.02
msp2	mixSkewNormal	2	-4.52	-24.09	0.43	0.70	0.42	0.75
msp3	SkewNormal	1	9.84	1.46	0.32	0.15	0.32	0.26
msp4	mixNormal	3	98.33	75.96	0.03	0.73	0.02	0.71
msp5	mixNormal	2	-57.27	-71.25	0.33	0.07	0.33	0.10
msp6	mixNormal	2	-154.04	-168.02	0.35	0.49	0.34	0.54
msp7	mixSkewNormal	2	65.68	46.11	0.16	0.22	0.16	0.42
msp9	mixSkewNormal	2	8.82	-10.75	0.53	0.91	0.53	0.93
msrp1	mixSkewNormal	2	-69.23	-88.8	0.09	0.30	0.10	0.38
msrp3	SkewT	1	-272.65	-283.83	0.02	0.05	0.01	0.08
pf10_0323	SkewT	1	-333.33	-344.51	0.62	0.51	0.62	0.54
pf113	mixNormal	2	-125.52	-139.5	0.73	0.80	0.74	0.85
pf12	mixSkewNormal	2	-13.72	-33.29	0.24	0.88	0.23	0.87
pf38	mixSkewNormal	2	118.98	99.41	0.05	0.04	0.05	0.08
pf41	mixSkewNormal	2	55.53	35.96	0.10	0.15	0.09	0.21
pff0335c	mixSkewNormal	2	22.36	2.79	0.04	0.49	0.04	0.51
rama	mixSkewNormal	2	-133.97	-153.54	0.32	0.15	0.32	0.25
rhoph3	mixSkewNormal	2	-133.16	-152.73	0.02	0.24	0.03	0.33
ron6	mixSkewNormal	2	-305.59	-325.16	0.01	0.09	0.02	0.13
tlp	mixSkewNormal	2	-407.36	-426.93	0.02	0.49	0.02	0.54

Table 2

Patient Seroprevalence. The patient seroprevalence for the 22 antibodies best fitted by two or more subpopulation distributions for the different categorization methods: i) the probability of being sampled from the seropositive (sp) population higher than 0.5 (*50:50*), ii) antibody levels higher than the seronegative population value correspondent to the probability given by the first quantile of the Z density distribution function (*q1)*; iii) the second quantile of the Z density distribution function (*q2)* and iv) the third quantile of the Z density distribution function (*q3)*. The total patient, the protected individuals (Prt) and the susceptible individuals (Sus) prevalence are shown.

Antibody	50 : 50			q1			q2			q3		
	Total	Prt	Sus	Total	Prt	Sus	Total	Prt	Sus	Total	Prt	Sus
ama1	0.40	0.46	0.30	0.55	0.64	0.35	0.50	0.59	0.30	0.46	0.54	0.30
eba140	0.17	0.21	0.10	0.73	0.79	0.60	0.55	0.62	0.40	0.42	0.48	0.30
eba175	0.36	0.43	0.20	0.84	0.91	0.70	0.71	0.84	0.45	0.63	0.75	0.38
eba181	0.26	0.28	0.20	0.78	0.83	0.68	0.64	0.70	0.53	0.55	0.62	0.40
msp1	0.83	0.89	0.73	0.93	0.93	0.93	0.89	0.93	0.83	0.86	0.91	0.75
msp10	0.56	0.63	0.43	0.75	0.85	0.55	0.70	0.79	0.53	0.61	0.70	0.43
msp2	0.23	0.31	0.08	0.83	0.88	0.75	0.73	0.80	0.58	0.64	0.73	0.45
msp4	0.82	0.93	0.60	0.93	0.99	0.80	0.93	0.99	0.80	0.90	0.98	0.75
msp5	0.62	0.65	0.55	0.79	0.83	0.70	0.76	0.80	0.68	0.74	0.79	0.65
msp6	0.34	0.40	0.23	0.64	0.69	0.55	0.64	0.69	0.55	0.64	0.69	0.55
msp7	0.76	0.89	0.50	0.93	0.98	0.83	0.92	0.98	0.80	0.91	0.96	0.80
msp9	0.31	0.30	0.33	0.79	0.77	0.83	0.68	0.68	0.68	0.56	0.56	0.58
msrp1	0.32	0.37	0.23	0.85	0.91	0.73	0.80	0.88	0.65	0.74	0.80	0.60
pf113	0.55	0.59	0.48	0.76	0.83	0.63	0.76	0.83	0.63	0.74	0.81	0.60
pf12	0.49	0.57	0.33	0.93	0.95	0.88	0.88	0.91	0.80	0.83	0.88	0.75
pf38	0.51	0.56	0.43	0.83	0.83	0.83	0.74	0.77	0.70	0.69	0.70	0.68
pf41	0.40	0.49	0.23	0.88	0.91	0.80	0.78	0.85	0.63	0.68	0.74	0.55
pff0335c	0.50	0.58	0.35	0.91	0.95	0.83	0.88	0.95	0.75	0.81	0.89	0.65
rama	0.43	0.43	0.43	0.92	0.95	0.85	0.88	0.93	0.80	0.86	0.90	0.78
rhoph3	0.28	0.27	0.30	0.93	0.93	0.93	0.85	0.88	0.80	0.80	0.83	0.75
ron6	0.06	0.06	0.05	0.90	0.93	0.85	0.89	0.91	0.85	0.81	0.88	0.68
tlp	0.21	0.22	0.20	0.95	0.98	0.90	0.95	0.98	0.90	0.93	0.95	0.88

Table 3

Chi-squared test results using different methods for defining seropositivity. For each categorization method, the antibodies statistically significant for the χ^2 test are highlighted in bold.

Antibody	50:50	q1	q2	q3
ama1	0.15	***<0.01***	***<0.01***	***0.02***
eba140	0.21	***0.04***	***0.04***	0.09
eba175	***0.02***	***0.01***	***<0.01***	***<0.01***
eba181	0.44	0.10	0.08	***0.04***
msp1	***0.04***	1.00	0.17	***0.03***
msp10	0.05	***<0.01***	***0.01***	***0.01***
msp2	***0.01***	0.13	***0.02***	***0.01***
msp4	***<0.01***	***<0.01***	***<0.01***	***<0.01***
msp5	0.36	0.17	0.19	0.15
msp6	0.10	0.18	0.18	0.18
msp7	***<0.01***	***0.01***	***<0.01***	***0.01***
msp9	0.91	0.61	1.00	0.99
msrp1	0.25	***0.01***	***0.01***	***0.03***
pf113	0.30	***0.03***	***0.03***	***0.02***
pf12	***0.02***	0.26	0.14	0.13
pf38	0.25	1.00	0.58	0.91
pf41	***0.01***	0.14	***0.01***	0.06
pff0335c	***0.03***	0.05	***<0.01***	***<0.01***
rama	1.00	0.12	0.08	0.11
rhoph3	0.91	1.00	0.40	0.45
ron6	1.00	0.32	0.45	***0.02***
tlp	0.96	0.18	0.18	0.26

Table 4

Logistic regression stepwise selection. AIC and AUC values for the multiple Logistic regression analysis prior and after stepwise selection are shown together with the final set of antibodies obtained following Stepwise selection.

Stratification	Context	AIC	AUC	Antibodies
50:50	Logistic regression	148.2	0.82	-
50:50	Stepwise logistic regression	130.0	0.80	h103 - msp2 - msp4 - msp7
seropositive probability	Logistic regression	143.3	0.84	-
seropositive probability	Stepwise logistic regression	126.2	0.82	h103 - msp2 - msp4 - msp7 - msrp3
q1	Logistic regression	148.5	0.80	-
q1	Stepwise logistic regression	130.6	0.76	h103 - msp4 - msp10 - eba175
q2	Logistic regression	147.7	0.81	-
q2	Stepwise logistic regression	126.7	0.78	eba175 - h103 - msp4
q3	Logistic regression	152.1	0.83	-
q3	Stepwise logistic regression	131.3	0.80	eba175 - h103 - msp1 - msrp3

Table 5

Patient Seroprevalence. The statistically significant results of the $\chi 2$ test for the 28 antibodies. The antibody levels that maximized the separation between the susceptible and protected group of individuals (Cut-off) and the proportion of seropositive individuals for all (Total), Protected (Prt) and susceptible (Sus) children, respectively

Antibody	P-value	Cut-off	Total	Prt	Sus
msp1	0.01	0.15	0.85	0.91	0.73
msp2	<0.01	0.07	0.45	0.57	0.20
msp4	<0.01	0.13	0.86	0.96	0.65
msp5	0.02	0.09	0.56	0.64	0.40
msp10	<0.01	0.25	0.79	0.90	0.58
pf12	<0.01	0.10	0.65	0.75	0.45
pf92	<0.01	0.11	0.83	0.91	0.65
pf34	<0.01	0.07	0.61	0.72	0.40
pf113	0.02	0.05	0.74	0.81	0.60
gama	<0.01	0.05	0.61	0.72	0.40
ama1	<0.01	0.16	0.74	0.84	0.53
eba175	<0.01	0.14	0.71	0.84	0.45
eba140	<0.01	0.11	0.96	1.00	0.88
eba181	<0.01	0.11	0.90	0.96	0.78
mtrap	0.01	0.05	0.85	0.91	0.73
asp	<0.01	0.08	0.70	0.79	0.53
msp3	0.01	0.08	0.48	0.57	0.30
msp6	<0.01	0.12	0.78	0.86	0.60
msp7	<0.01	0.24	0.71	0.86	0.40
msrp1	<0.01	0.05	0.79	0.88	0.63
msrp3	<0.01	0.04	0.96	1.00	0.88
h101	0.03	0.05	0.74	0.80	0.60
h103	<0.01	0.07	0.50	0.60	0.28
pf41	<0.01	0.12	0.38	0.48	0.18
pff0335c	<0.01	0.05	0.88	0.95	0.75
rh5	0.04	0.16	0.39	0.46	0.25
ron6	0.02	0.04	0.81	0.88	0.68
pf11_0373	<0.01	0.08	0.21	0.28	0.05

Supplementary Files:

Supplementary figure S1. Normality assessment. The *p-values* of the three statistical tests, Anderson-Darling (AD) in black, Jarque-Bera (JB) in dark grey and Shapiro-Wilk (SW) in light grey for all the antibodies available in our data set.

Supplementary figure S2. Box-cox transformation parameters (λ). Values lower than -3 and higher than 3 (outside the range) were highlighted with grey areas.

Supplementary figure S3. Normality assessment after Box-cox transformation. The p-values of the three tests, Anderson-Darling (AD) in black, Jarque-Bera (JB) in dark grey and Shapiro-Wilk (SW) in light grey.

Supplementary figure S4. Two sample T-test. Box plots of 4 statistically significant antibodies for the two-sample t-tests comparing the protected (dark grey) and susceptible (light grey) antibody levels. Values are Box-Cox transformed according to the λ indicated.

Supplementary figure S5. Comprehensive pipeline - Random Forest results. (A) The mean accuracy for each value of randomly selected predictors when performing each tree after one hundred runs of 10-fold cross-validation are given by a grey circle. (B) Importance of each antibody in the model

Supplementary figure S6. Ridge regression regularisation strategy results. (A) The mean accuracy for each regularisation parameter (λ) after one hundred runs of 10-fold cross-validation are given by a grey circle. (B) Importance of each antibody in the model.

Supplementary table I. Mixture-Model and goodness of fit results for pf92. The BIC and AIC values for each model, according to the type of distribution (Model) and the number components (Num.comp) are shown. The Pearson goodness-of-fit and Likelihood ratio test *p-values* according to the number of bins are provided. P.10 denotes the data was split in 10 equal size bins containing 10% of the data, while P.5 indicates the data was split in 20 equal size bins containing 5% of the data.