Anemia and its Associated Factors Among Women of Reproductive Age in Africa

Scarlett He, Nuona Chen, Huizi Yu

2022-12-17

INTRODUCTION

Anemia is defined as the situation that the quantity or quality of circulating red cells is below the normal level. About one third of the population are anemic worldwide. Specifically, approximately 40% of children(0-12 years), 35% of women and 18% of men are anemic(DeMaeyera and Adiels-Tegmanb, n.d.). In Africa, anemia is one of the most common public health problems. More than half of pregnant women and children less than 5 years old have this health problem(Crawley 2004). Our research was based on "Anemia and its associated factors among women of reproductive age in eastern Africa: A multilevel mixed-effects generalized linear model" and we expanded the object of study from eastern Africa to all available countries in Africa. Our research was conducted in two parts. We used inference to identify the associated factors of Anemia among women of reproductive age in Africa and prediction to predict the risk of Anemia among women of reproductive age in Africa.

DATA MANIPULATION & PREPROCESSING

We obtained our data from the Demographic and Health Survey (DHS), which is a nationally representative and large-scale survey of women aged 15-49. The extracted data contains 32 unique countries in Africa and 482,187 responses of participating women's health, socioeconomics, quality-of-life and behavior information. For the purpose of this study, we encoded the binary outcome variable "anemia" based on the women's pregnancy status: pregnant women with hemoglobin value less than 11 g/dL and non-pregnant women with hemoglobin value less than 12 g/dL are considered anemic.

The raw dataset contains 21 individual-level variables, including (1) demographic characteristics (e.g., age, household size, education level) (2) healthcare status (e.g., access to healthcare, distance to closest health-care facility) (2) quality of life variable (e.g., access to improved plumbing facility, media exposure). We also included 2 community-level variables: (1) community literacy rate and community poverty rate. The communities are identified based on the participants' residence.

We conducted further analysis on the selection of variables in an attempt to simplify the model form. Although the raw dataset does not contain an unprocessable amount of variables, we believe this procedure is generally useful for big data explorations. We selected a statistical model "glmmLasso" to perform variable selection. This method is computationally efficient to implement and relies on the LASSO algorithm for classification. We split our dataset into training and testing to avoid overfitting the evaluated performance of the model: the optimal lambda value is selected by evaluating the AUC of the testing set. We selected a subset of 19 individual-level variables and 2 community-level variables to include in the final model. The variable selection procedure and the distribution of the variables are shown in Figure 1 and Appendix Figure 6 and Figure 7. The Figure 7 shows the prevalence of anemia across countries, we can see that the anemia epidemic is especially severe in the Western African countries.

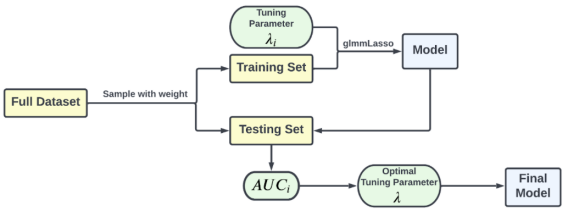

Figure 1: Variable Selection Procedure

METHOD & RESULT

Aiming to investigate the prevalence of anemia among women of reproductive age in Africa, we generated two research objectives: first, we want to identify the associated factors of anemia using an Inference Model, and second, we want to predict the probability of anemia.

For the inference model, we use a multi-level fixed effect model to evaluate the relationship between anemia and other variables. We are particularly interested in the coefficient estimate, which can inform us of the specific relationship (strong or weak, positive or negative) between variables and outcome. For the predictive models, we first split the dataset into training and testing sets. The training set contained 70% of data. It was used to build and train the machine learning model. The testing set contained 30% of the data. It was used to evaluate the performance of the model. As shown in Appendix Figure 6, we highlight the distribution of the outcome variable anemia: it is balanced with approximately 60% of the population categorized as non-anemic and 40% of the population as anemic. Because of the relative balanced distribution, we will use Accuracy as the evaluation metric for the performance, as we do not have to worry about outcome imbalance issues.

Inference Model

We fit a multi-level model on the selected variables to account for both individual-level variables and the community-level variables. We also apply a weighting to adjust for non-response or disproportionate sampling such that the samples correspond to the country's true population. We fit a multilevel mixed effect generalized linear model with logit link, and the functional form is shown in Figure 2. We use subscript i for individual level, j for community level, and k for country level and l for year level variables. We similarly label the coefficients. Note that our focus is on the relationship between individual-level and community-level variables. We include the intercept for country level and year level as control variables. Figure 3 shows the individual level coefficient estimates.

$$
\begin{aligned}
\text{logit}(p(Y_{ijkl}=1)) = {} & \beta_0 + \alpha_{0j} + \gamma_{0k} + \eta_{0l} \\
& + \beta_1 \text{ageGroup}_i + \beta_2 \text{edu}_i + \beta_3 \text{marriage}_i + \beta_4 \text{job}_i \\
& + \beta_5 \text{wealth}_i + \beta_6 \text{head}_i + \beta_7 \text{mediaExposure}_i + \beta_8 \text{toilet}_i \\
& + \beta_9 \text{water}_i + \beta_{10} \text{terminatedPregnancy}_i \\
& + \beta_{11} \text{houseHouldSize}_i + \beta_{12} \text{facilityDistance}_i \\
& + \beta_{13} \text{contraceptive}_i + \beta_{14} \text{pregenentNow}_i \\
& + \beta_{15} \text{breastfeeding}_i + \beta_{16} \text{Residence}_i \\
& + \alpha_{1j} \text{poverty}_i + \alpha_{2j} \text{literacy}_i + \epsilon_{ijkl}
\end{aligned}
$$

Figure 2: Multi-Level Mixed Effect Logit

From Figure 3, we were able to draw some conclusions about factors that are associated with higher risk of anemia. For example, being in the older age group was associated with a lower prevalence of anemia as compared to the age group 15-24 years except that the age group 35-44 was 2.6% higher. Additionally, Women who had a terminated pregnancy had 5.4% higher prevalence of anemia as compared with their

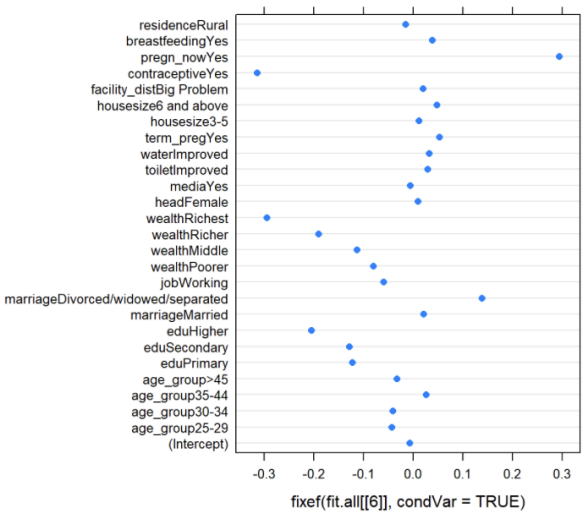

Figure 3: Coefficient Estimates of Mixed Effect Model

counterparts. These relationships are useful for identifying the driving forces behind anemia among women of reproductive age in Africa and can provide insights into policy development. Additionally, using the coefficient estimate, we were able to construct a profile for women in Africa at high risk of anemia shwon in Figure 4.

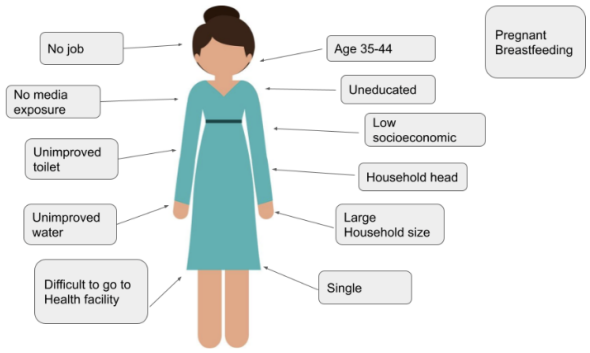

Figure 4: Risk Profile of Anemia

Predictive Model

Naive Bayes Naive Bayes classifier is a supervised probabilistic machine learning model used for classification tasks based on the Bayes theorem. After comparing the outcome of the training and testing phase, objects X and Y were created to store values of predictor variables and response variables. A predictive model was then created by using Naive Bayes Classifier. The accuracy of the model was then checked by using the testing dataset. Confusion matrix was then used to evaluate the accuracy of the model. The final output showed that the Naive Bayes Classifier we built could predict whether an individual has anemia problem or not, with an accuracy of 0.6024, Sensitivity of 0.7606, and Specificity of 0.3689.

	Actual No Anemia	Actual Anemia
Predicted No Anemia	10374	5833
Predicted Anemia	3265	3410

KNN The K-nearest Neighbor model is a supervised machine learning model that classifies a target point's outcome based on its relative relationship to its neighbors. As illustrated in the previous Naive BAyes section, we built a KNN model using the training set. To achieve the best accuracy of the model, we further tuned the algorithm using a grid of hyperparameter (k: # of neighbors) and cross validation. The performance evaluation of a varying number of k is shown in Figure 5. As we can see, the accuracy plateaus after the number of neighbors reaches 50. We chose # of neighbor = 50 as the final hyperparameter. The confusion matrix on the testing data is shown below. The KNN model reached an overall accuracy of 0.6026, sensitivity of 0.7438, and Specificity of 0.395.

	Actual No Anemia	Actual Anemia
Predicted No Anemia	10015	5555
Predicted Anemia	3450	3640

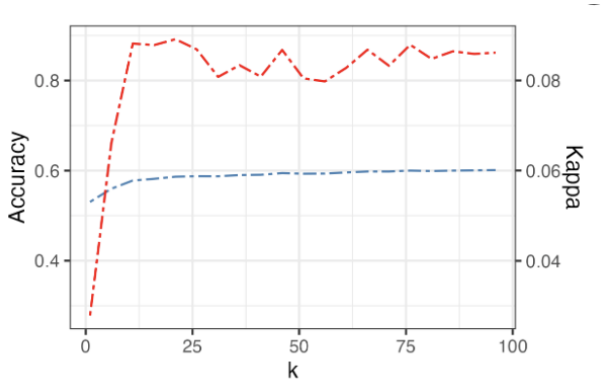

Figure 5: Cross Validated Accuracy (blue) & Kappa (red) by # of Neighbors

SVM Support Vector Machine is a supervised learning technique that is aimed to drive a hyperplane to separate and classify data points. The ideal hyperplane has the maximal marginal distance between different classes. SVM models were fitted using the "e1071" library, with and without using parallel computing. We used the "doParallel" package to conduct parallel computing for our SVM models. There are 15,951 observations in the training data. With parallel computing, it took 52.481seconds to finish building the model. Without parallel computing, it took 55.129 seconds to complete the model building process, with accuracy of 0.5144, sensitivity of 0.910 and specificity of 0.133.

	Actual No Anemia	Actual Anemia
Predicted No Anemia	10192	10041
Predicted Anemia	1009	1514

Random Forest The random forest is a supervised machine learning algorithm that uses bootstrap aggregation to fit multiple decision trees; the final prediction outcome is calculated by averaging the predictions from each tree. We built a random forest model using the caret package in R. We further tuned the hyperparameters of number of trees, max depth and max features. The resulting confusion matrix on the testing dataset is shown below. The overall accuracy is 0.6289, sensitivity is 0.9765, and specificity is 0.1161

	Actual No Anemia	Actual Anemia
Predicted No Anemia	13318	8160
Predicted Anemia	321	1073

Overall Performance Evaluation As we observe from the previous confusion matrix, Random Forest showed the best overall performance among all models. It reached a 62.89% accuracy with 97.65% sensitivity. However, as we have observed, the random forest model had a specificity of 0.1161. Since we are particularly interested in identifying the negative classes (those with anemia), we might prefer a model that had higher specificity and similar accuracy. In this case, we might alternatively choose KNN as the evaluation model, which had similar accuracy (60.26%) and much higher specificity (39.5%). In the future, we might train the model by optimizing the AUC or F1 score, which is a balance of precision and recall. We would also like to design some additional metrics that will weigh correctly identifying the negative class (anemia) higher than identifying the positive class (no anemia).

SQL Database Construction Structured Query Language (SQL) is a programming language used to communicate with and manipulate databases. It has several advantages. Large amounts of data can be retrieved quickly and efficiently. Operations like insertion, deletion, manipulation of data is also done fast. Users do not need professional training in programming. It integrates easily with other programming languages such as Python and R. We established a local SQL database and uploaded our data used on this research on the established database(Appendix1). It can be utilized for future research with faster data retrieval.

CONCLUSION

The prevalence of anemia in Africa was relatively high. Both individual level and community level factors were associated with the development of anemia in women. Special consideration should be given to those groups of women who had a higher risk of anemia such as uneducated women, not currently working women, divorced/widowed/separated women and pregnant women, those who are from households with low socioeconomic status, unimproved toilet facility and source of drinking water is recommended.

Additionally, we utilize big data algorithms and tools to predict the risk of Anemia. By training ML models, we achieved a prediction accuracy of 0.62 using Random Forest. We also constructed a SQL database that can be used for easy data access and explorations in the future.

Author Contribution Statements

Huizi Yu performed data manipulation & processing, inference model, KNN and random forest.Scarlett He worked on Naive Bayes and SQL database establishment. Nuona Chen carried out SVM and performance evaluation and attempted using parallel computing on the models. However, parallel computing didn't improve the model efficiency so we omitted it in the final report. We aim to futher investigate using parallel computing in the future.All authors provided critical feedback and helped shape the research, analysis and manuscript.

APPENDIX

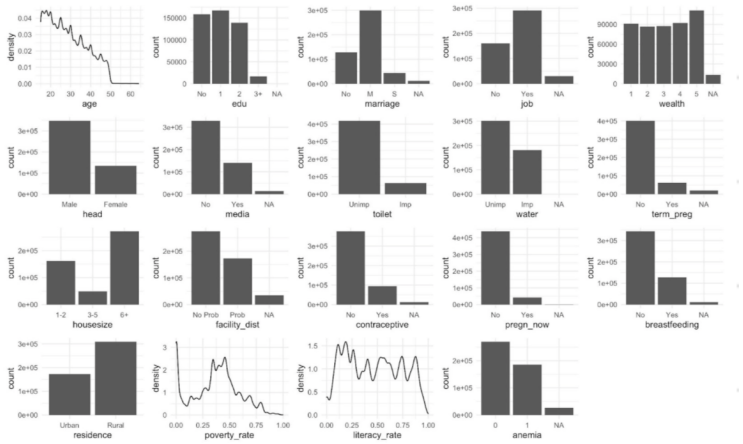

Figure 6: Distribution of Variables

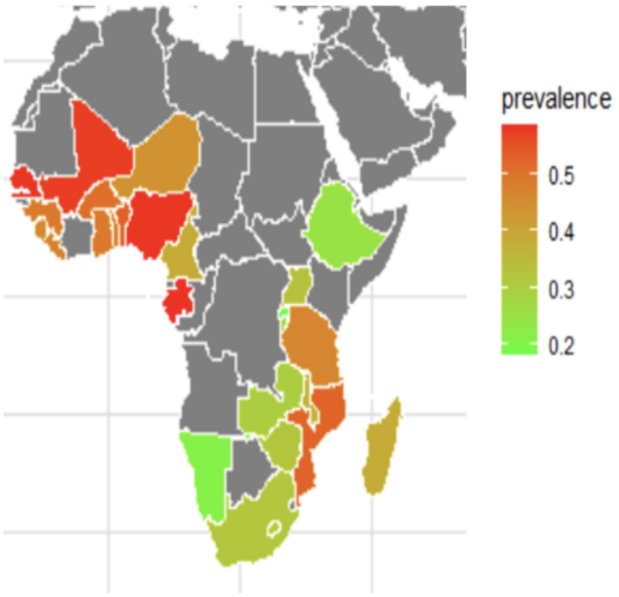

Figure 7: Prevalance of Anemia in Africa by Country

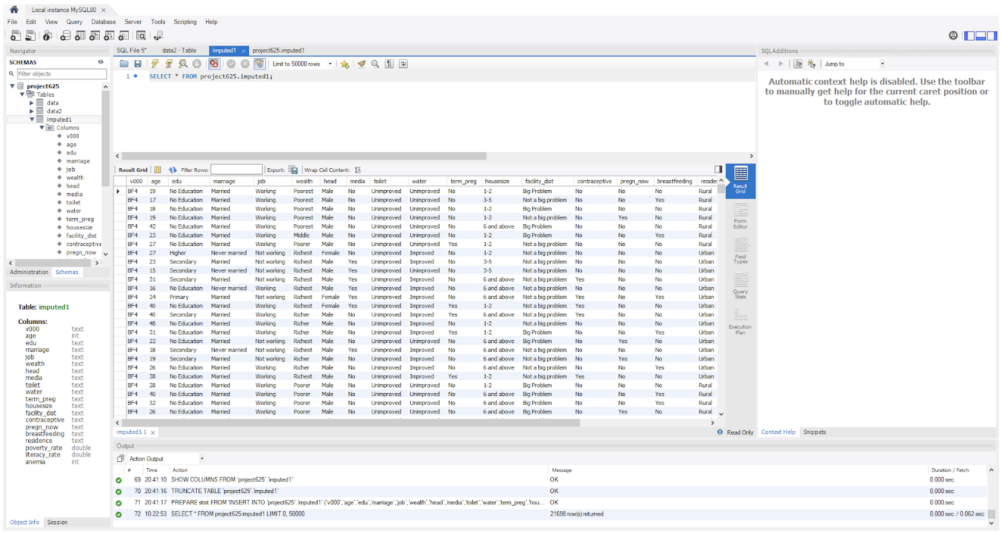

Figure 8: SQL Database Interface

References

Crawley, Jane. 2004. "REDUCING THE BURDEN OF ANEMIA IN INFANTS AND YOUNG CHILDREN IN MALARIA-ENDEMIC COUNTRIES OF AFRICA: FROM EVIDENCE TO ACTION." *The American Journal of Tropical Medicine and Hygiene* 71 (2_suppl): 25–34. https://doi.org/10.4269/ajtmh.2004.71.25.

DeMaeyera, E, and M Adiels-Tegmanb. n.d. "THE PREVALENCE OF ANAEMIA IN THE WORLD."