Classification Model on Antiretroviral Therapy Reaction and Failure Developed on the Unique Records of the Akwa Ibom HIV Database

Nicelle Sernadilla Macaspac

July 2023

Introduction

This undertaking is part of the capstone in the Professional Certificate Program in Data Science of Harvard Online. The corresponding R Markdown and R files are in the GitHub of nsmacaspac.

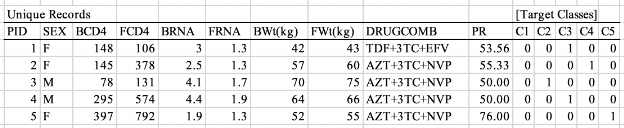
Unique Records of the Akwa Ibom HIV Database

In a previous study on patient response to antiretroviral therapy, Ekpenyong, Etebong, and Jackson (2019, 3) used a database of patients who received treatment for HIV from thirteen health centers in Akwa Ibom, Nigeria, between 2015 and 2018. Two years later, they published the processed dataset (Ekpenyong et al. 2021b, Appendix) with minor oversight in the accompanying article, which were easily reconciled through the 2019 study and were appropriately referenced throughout this project. The processed dataset was composed of an Individual Treatment Change Episodes table with a column for each antiretroviral drug administered and a concatenated Unique Records table with the drugs combined into a column for each antiretroviral therapy of three drugs administered. For the purpose of this project, we utilize only the Unique Records table.

The Unique Records table was imported with the corresponding read_xlsx function in the language R. The dataset was composed of 1,056 patient records, each with 15 variables: patient identification, sex, baseline CD4 count, follow-up CD4 count, baseline RNA load, follow-up RNA load, baseline weight, follow-up weight, drug combination, and patient response and drug reaction classifications 1 to 5 (fig. 1). The immunological marker CD4 counts were given in cells per cubic millimeter (Ekpenyong et al. 2021a, 8). The viral RNA loads were expressed in times 10^2 copies (Ekpenyong, Etebong, and Jackson 2019, 10). The weights ranged from 4.7 to 125 kg on account of the presence of pediatric patients (Ekpenyong, Etebong, and Jackson 2019, 2). The three-drug combinations of antiretroviral therapy were a complementary mix of nucleoside reverse transcriptase inhibitors tenofovir (TDF), lamivudine (3TC) and zidovudine (AZT), and non-nucleoside reverse transcriptase inhibitors efavirenz (EFV) and nivarapine (NVP) given in the first 6 months of treatment (Ekpenyong et al. 2021a, 8). Patient response and reaction to the drugs were quantified and classified with a binary system as very high interaction (C1), high interaction (C2), low interaction (C3), very low interaction (C4), and no interaction (C5) in the 2019 study using the advanced method of interval type-2 fuzzy logic system (Ekpenyong, Etebong, and Jackson 2019, 11). Very high and high interactions signified treatment failure as well (Ekpenyong, Etebong, and Jackson 2019, 10).

Tidy and Preprocessed Dataset

The dataset was prepared for preprocessing by rendering it into tidy and numeric format. The fifteen variables were renamed for consistency with their aforementioned descriptions, with vhi_tf corresponding



Unique Records										[Target Classes]				
PID	SEX	BCD4	FCD4	BRNA	FRNA	BWt(kg)	FWt(kg)	DRUGCOMB	PR	C1	C2	C3	C4	C5
1	F	148	106	3	1.3	42	43	TDF+3TC+EFV	53.56	0	0	1	0	0
2	F	145	378	2.5	1.3	57	60	AZT+3TC+NVP	55.33	0	0	0	1	0
3	M	78	131	4.1	1.7	70	75	AZT+3TC+NVP	50.00	0	1	0	0	0
4	M	295	574	4.4	1.9	64	66	AZT+3TC+NVP	50.00	0	0	1	0	0
5	F	397	792	1.9	1.3	52	55	AZT+3TC+NVP	76.00	0	0	0	0	1

Figure 1: First rows of the Unique Records table.

to very high interaction_treatment failure and ni corresponding to no interaction. Missing values were not detected.

```
head(dataset, n = 5)
##   id sex bcd4 fcd4 brna frna bweight fweight     therapy response vhi_tf hi_tf
## 1  1   F  148  106  3.0  1.3      42      43 TDF+3TC+EFV 53.56199      0     0
## 2  2   F  145  378  2.5  1.3      57      60 AZT+3TC+NVP 55.33422      0     0
## 3  3   M   78  131  4.1  1.7      70      75 AZT+3TC+NVP 50.00000      0     1
## 4  4   M  295  574  4.4  1.9      64      66 AZT+3TC+NVP 50.00000      0     0
## 5  5   F  397  792  1.9  1.3      52      55 AZT+3TC+NVP 76.00000      0     0
##   li vli ni
## 1  1   0  0
## 2  0   1  0
## 3  0   0  0
## 4  1   0  0
## 5  0   0  1
```

The sex and therapy variables were changed from character to numeric format using the case_when function. The brna and frna variables were multiplied by 10^2 to align with the unit used for viral RNA load in the WHO definition of HIV (WHO, n.d.). The vhi_tf, hi_tf, li, vli, and ni variables were verified to have only one value per row. Hence, they were merged under a newly defined reaction variable then relabeled as 5 (vhi_tf) to 1 (ni). This brought the number of variables down to eleven.

```
dataset1 <- dataset |>
  mutate(sex = ifelse(sex == "F", 1, 2)) |> # relabels sexes as 1-2
  mutate(brna = brna*10^2) |>
  mutate(frna = frna*10^2) |>
  mutate(therapy = case_when(therapy == "AZT+3TC+EFV" ~ 1,
                             therapy == "AZT+3TC+NVP" ~ 2,
                             therapy == "TDF+3TC+EFV" ~ 3)) |> # relabels antiretroviral therapies as 1
  mutate(reaction = case_when(vhi_tf == 1 ~ 5,
                              hi_tf == 1 ~ 4,
                              li == 1 ~ 3,
                              vli == 1 ~ 2,
                              ni == 1 ~ 1,)) |> # relabels drug reactions as 5-1 then merges them under
  select(-vhi_tf, -hi_tf, -li, -vli, -ni)
head(dataset1, n = 5)
##   id sex bcd4 fcd4 brna frna bweight fweight therapy response reaction
## 1  1   1  148  106  300  130      42      43       3 53.56199        3
## 2  2   1  145  378  250  130      57      60       2 55.33422        2
## 3  3   2   78  131  410  170      70      75       2 50.00000        4
## 4  4   2  295  574  440  190      64      66       2 50.00000        3
## 5  5   1  397  792  190  130      52      55       2 76.00000        1
```

Given that all values were then in numeric format, the correlation of the variables was examined using the corrplot function (fig. 2).

-the distribution summaries -the variance

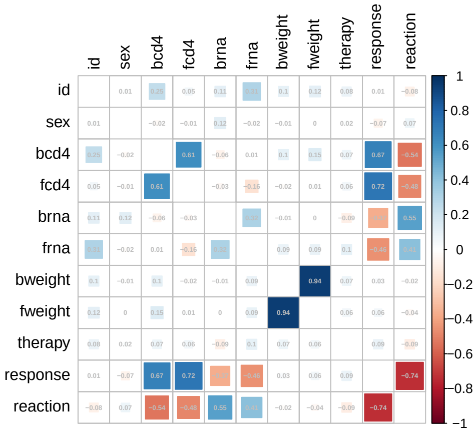


Figure 2: Matrix of the correlation coefficients between variables.

CD4 counts and RNA loads

The dataset was preprocessed to retain only CD4 counts and RNA loads

CD4 Count

RNA Load

Classification Models

- why this partition

k-Nearest Neighbor Model

- why this model

Recursive Partitioning and Regression Trees Model

Rborist Model

Quadratic Discriminant Analysis Model

Predictive Model

Conclusion

-meaningful decisions on antiretroviral therapy administration

References

Ekpenyong, Moses E., Mercy E. Edoho, Ifiok J. Udo, Philip I. Etebong, Nseobong P. Uto, Tenderwealth C. Jackson, and Nkem M. Obiakor. 2021a. "A Transfer Learning Approach to Drug Resistance Classification in Mixed HIV Dataset.” *Informatics in Medicine Unlocked* 24: 100568. https://doi.org/10.1016/j.imu.2021.100568.

Ekpenyong, Moses E., Philip I. Etebong, and Tenderwealth C. Jackson. 2019.”Fuzzy-Multidimensional Deep Learning for Efficient Prediction of Patient Response to Antiretroviral Therapy.” *Heliyon* 5: e02080. https://doi.org/10.1016/j.heliyon.2019.e02080.

Ekpenyong, Moses E., Philip I. Etebong, Tenderwealth C. Jackson, and Edidiong J. Udofa. 2021b.”Processed HIV Prognostic Dataset for Control Experiments.” *Data in Brief* 36: 107147. https://doi.org/10.1016/j.dib.2021.107147.

Irizarry, Rafael A. 2002. *Introduction to Data Science: Data Analysis and Prediction Algorithms with R.* http://rafalab.dfci.harvard.edu/dsbook/.

WHO, n.d. “HIV.” Accessed July 21, 2023. https://www.who.int/health-topics/hiv-aids#tab=tab_1.