**DHS HIV Abstract Analysis Outline**

**abstract deadline:** Sep 27, 2018 - Conference on Retroviruses and Opportunistic Infections (CROI), 1 table/figure, 2500 characters.

**background:** It would be nice to know how to better target public health programs for HIV identification.

**study framework:** Selection of variables/features that predict HIV status from demographic & health surveys.

**outcome:** measures of HIV status prediction (sensitivity, specificity, PPV, NPV)

**population:**

* treat each country-survey as an observation (max n=296)
* Countries with generalized epidemic (ie: heterosexual transmission) as identified as having prevalence above 1%

**objective/goals:**

1. What are important features for predicting HIV status from DHS?
2. how well can HIV status be predicted?

**Methods:**

- Random forest to predict HIV status (1,0), out-of-bag validation

- Identify features as “important predictors” using Boruta method

- tuned for:

n-vars (how many variables to sample)

n-trees (how many trees to sample)

leaf-size (minimum size of terminal node)

mu (threshold of missingness to include variable)

rho (decision variable for continuous/categorical treatment)

pVal (the threshold that needs to be achieved for feature selection)

**Table 1: Frequency of feature importance for HIV prediction across surveys (varying importance thresholds)**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Importance threshold** | | |
| **Feature** | **1%** | **0.1%** | **0.01%** |
| Has goats | 34 (10%) | 28 (9%) | 22 (5%) |
| Female | 16 (3%) | 15 (3%) | 12 (2%) |
| **…** |  |  |  |

**sample conclusion/summary:** “we identified X, Y, and Z (mostly social) factors that are importantly predictive of HIV status across # surveys in countries with generalized HIV epidemic.”