Documentation for HVTN 505 T cell and Binding Antibody Data for Janes et al. (2017, JID) and Fong, Shen et al. (2018, JID)

April 6, 2018

Immune response biomarkers are measured at the Week 26 visit in vaccine group subjects, for studying their association with HIV infection through to the Month 24 study visit and prior to the date the study was unblinded on April 22, 2013.

1. The file primary505\_for\_sharing.csv contains the main data for the 2,504 subjects enrolled in the HVTN 505 trial (reported upon in Hammer et al., 2013, NEJM)

Key variables in this data file for a survival analysis are as follows:

* **HIVwk28preunbl** is the indicator of whether the failure event occurs before right-censoring and before the unblinding date April 22, 2013 (the Delta)
* The time from enrollment to HIV infection diagnosis is **HIVwk28preunblfu**, for participants with **HIVwk28preunbl**=1. For right-censored subjects (with **HIVwk28preunbl**==0), the right-censoring time is **HIVwk28preunblfu**, which is the date of last study visit or April 22, 2013, whichever occurs first. Thus this variable **HIVwk28preunblfu** is the standard X = min(T,C) in survival analysis.
* **[If interval-censored methods are used]** For subjects with failure events (with **HIVwk28preunbl**==1), the L and the U straddling their true infection time are defined by the date variables **lastnegdt** and **dxdt** (last HIV negative testing date and diagnosis of HIV infection date), which should be compared to the time origin (which is the enrollment date, variable **enrdt**) to compute L and U as numbers of days since enrollment.
* **bhvrisk** is the baseline behavioral risk score defined in Hammer et al. (2013, NEJM) that is the best available predictor of the amount of exposure to HIV infection. It was controlled for in all models studied in Janes et al. (2017, JID) and Fong, Shen et al. (2018, JID).

Participants to include in the data analysis (under a complete-case inverse probability weighting approach) are participants that satisfy all of the following criteria based on variables in primary505\_for\_sharing.csv:

* **trt**==1 (subset on vaccine group subjects)
* **week28**==1 (in the primary cohort that did not get diagnosed with HIV

infection prior to the week 28 visit. Not being infected before

week 28 is necessary to be included in the case-control cohort

for measuring Week 26 biomarkers.)

* **cc\_cohort**==1 (Selected for inclusion in the case-control cohort and thus

potentially has Week 26 biomarkers measured and can be

included in the IPW complete-case analysis.)

For an analysis plan consistent with previous correlates of risk analyses (Janes et al., 2017, JID; Fong, Shen et al., 2018, JID), the analysis would restrict to vaccine recipients (with **trt**==1). Moreover, all analyses would adjust for the 4 variables **age, race, BMI,** and **bhvrisk**, which would help make the hazard ratio coefficients for the immune response markers to be interpretable in terms of biological susceptibility to acquisition of HIV-1 infection.

The file v505\_tcell\_correlates\_data\_for\_sharing.csv contains the Week 26 T cell response biomarkers for the case-control cohort. The file v505\_tcell\_correlates\_data\_documentation.xlsx documents the variables in this file. This file only contains data for the case-control sample of 189 subjects: 150 case-control vaccine recipients (25 HIV infected cases, 125 HIV uninfected controls) and 39 case-control placebo recipients. Only the 150 vaccine recipients are used in an IPW complete-case data analysis (the biomarkers are not interesting in the placebo group because they generally measure HIV-specific immune responses, and placebo recipients do not have exposure to HIV or to an HIV vaccine). The Week 26 biomarkers in this file can be merged with the main data set file primary505\_for\_sharing.csv based on the **pubid** variable.

A needed variable in v505\_tcell\_correlates\_data\_for\_sharing.csv is **wt**, which is the estimated inverse probability weight for each given subject in the case-control sample included in the IPW complete-case analysis.

A large number of the variables in v505\_tcell\_correlates\_data\_for\_sharing.csv are not needed for the analysis, as they pertain to the post-infection duration of follow-up for subjects who became HIV infected. These variables are listed in the tab “variables for HIV-inf subjects” in v505\_tcell\_correlates\_data\_documentation.xlsx.

Each Week 26 T cell biomarker variable is defined by any unique combination of values for the three variables **tcellsub** x **cytokine** x **antigen**. For each of these variables, there are several different readouts that could be analyzed (columns AG−AL). In addition, there are four different score variable types that combine information across individual variables that could be analyzed (columns AM−AP).

Now, all of the variables defined by **tcellsub** x **cytokine** x **antigen** are of scientific interest. So, one could potentially conduct a correlates of risk analysis for each of these variable types. Or, as an example of a specific analysis of interest, one could focus on the **tcellsub** level CD8+ (because much more variability than CD4+) and only consider the **cytokine** level IL2/ifngamma (as the most common choice), and then repeat the analysis across each of the different antigen types in separate analyses.

For the different readouts in columns AG−AL, one approach would include as input variables the two readout types **logpctpost\_scaled** and **pctpos\_adj\_ind,** because these were pre-specified in the primary analysis of Janes et al. (as noted in the documentation file). And then, for columns AM−AP, to include the two score variable types **score\_scaled** and **score\_ind**, again because they were pre-specified in the primary analysis of Janes et al. (as noted in the documentation file). Alternatively, if the methods can handle high-dimensional data, it is interesting to include all the input marker variables, as we do not know which variables are the most important.

The file bama.m\_for\_sharing.csv contains the Week 26 antibody response biomarkers measured by BAMA that are baseline-subtracted and Week 26 mock-subtracted. This file also only contains data for the case-control sample of 189 subjects: 150 case-control vaccine recipients (25 HIV infected cases, 125 HIV uninfected controls) and 39 case-control placebo recipients. This file also has a **pubid** variable that must be used to merge the data with the data in primary505\_for\_sharing and v505\_tcell\_correlates\_data\_for\_sharing.csv

In this file, the pre-specified primary variables analyzed and described in Fong, Shen et al. are (variable names in the data file on left, long labels on the right):

IgG\_env IgG Binding to gp120/140

IgG\_V2 IgG Binding to V1V2

IgG\_V3 IgG Binding to V3

IgG\_gp41 IgG Binding to gp41

IgG\_BioC4\_427B IgG Binding to C4\_427B

IgA\_env IgA Binding to 140

The pre-specified secondary variables analyzed and described in Fong, Shen et al. are:

IgG\_AEA244V1V2Tags293F

IgG\_C1086C\_V1\_V2Tags

IgG\_gp70\_BCaseA2V1V2169K

IgG\_VRC\_A\_gp70V1V2\_avi

IgG\_BioV3B, IgG\_BioV3M

IgG\_BioV3CRF2

IgG\_A1conenv03140CF

IgG\_Bconenv03140CF

IgG\_Cconenv03140CF\_avi

IgG\_Con6gp120B

IgG\_ConSgp140CFI

IgG\_VRCA\_avi

IgG\_VRCBgp140

IgG\_VRCC\_avi

(plus the same set of 14 secondary variables with IgA\_ instead of IgG\_)

One analysis of interest would include as input features all of the primary and secondary variables. Another analysis of interest would include all 105 of the input features. The file bama.m.for\_sharing\_documentation.doc provides documentation for these 105 IgG and IgA variables studied in Fong, Shen et al. (2018, JID).