==================================
SEARCH - UCSF - Application
==================================

==================================
The Challenge
==================================

Thank you for your interest in SEARCH!
http://www.searchendaids.com/

We have provided a dataset that includes demographic information and HIV diagnoses, treatment
and lab results. The dataset resembles the data we work with, although they are simulated and
have been simplified.

To reflect the data we work with, the dataset may have data quality problems. If you see a
data quality problem, make your own decision about how to address it, but document your choice
(see [2] below).

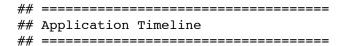There are three questions to address:
[1] For each of the 32 communities in the data and for each of 4 time points, report the
proportion of patients who have an unsuppressed viral load. This quantity is defined below.

[2] Pretending these were real data, are there any data quality problems in the dataset that
the team would need to investigate? What decisions did you make in addressing these problems?

[3] (bonus question - the ability to answer this question is not a requirement for the
position, but an ideal candidate would be able to do this)
Suppose we changed our data simulation so that all patients who are HIV positive at time 1, 2,
or 3 are immediately treated with antiretroviral therapy (ART). The data generating process is
otherwise unchanged (including treatment at time 0). In the resulting data, what would be the
total population proportion of patients with an unsuppressed viral load at time 3? Provide a
single estimate and a 95% confidence interval. Very briefly describe your methodology.

We primarily use R, but you are welcome to use any programming language for this challenge.

==================================
Background
==================================

In this dataset, there are 32 communities in Kenya and Uganda. At each of 4 time periods
(t=0,1,2,3) there is a community health clinic (CHC) where we attempt to measure the HIV
status of all individuals. If a person tests HIV positive, their viral load will be measured
and they may be given antiretroviral therapy (ART). If a person does not attend the CHC, a
tracking team will attempt to find the person and perform the same functions (test for HIV,
measure viral load, possibly provide ART). Viral loads are also measured at other times if a
patient comes in for care.

==================================
Data
==================================

CHC and tracking data: Community_t.csv
searchid - patient identifier, constant over time
HIV - 1 if tested HIV positive at time t
ART - 1 if assigned antiretroviral therapy at time t
chcdate - date patient attended the Community Health Clinic (NA if did not attend)
trdate - date patient was tracked (NA if not tracked)

braceletid – identifier attached to viral load test (distinct for each test)
age – age in years
male – 1 if male

viral load tests: ViralLoads.csv
braceletid – identifier attached to viral load test (distinct for each test)
VL – viral load (HIV RNA copies/ml), lower level of detection = 40
date – date of viral load test

===================================
Definitions
===================================

chcstart_t = community level – beginning of CHC for time t

trkend_t = community level – end of tracking for time t

VL_t = patient level – viral load measured between chcstart_t and trkend_t (inclusive); if more than one viral load is measured in this window, use the one with the date closest to chcstart_t

unsupp_t = patient level – unsuppressed viral load – 1 if HIV positive at time t and VL_t > 500

prop_unsupp_t = community level – mean of unsupp_t

===================================
Output
===================================

1) unsupp.csv – a csv file with 32 rows and 5 columns:
community name, prop_unsupp_0, prop_unsupp_1, prop_unsupp_2, prop_unsupp_3

2) writeup.pdf – Your answers to [2] and [3].

3) The code you used for this challenge.

===================================
Submission
===================================

Please submit your output along with your CV and a cover letter to search.ucsf.hiring@gmail.com, including your name and contact email address in the body of the email.

===================================
Updates
===================================

Any updates will be posted here.
5/8/18 Initial version
5/9/18 for [3] : "In the resulting data, what would be the total population proportion of patients with an unsuppressed viral load AT TIME 3?" and "Suppose we changed our data simulation so that all patients who are HIV positive AT TIME 1, 2, or 3 are immediately treated with antiretroviral therapy (ART). The data generating process is otherwise unchanged (INCLUDING TREATMENT AT TIME 0)"

====================================
Application Timeline
====================================

We will accept applications on a rolling basis. Once the position has been filled, the Dropbox folder will be taken down.

====================================
Questions
====================================

If you have questions about the data, please make a reasonable assumption about how to handle it and document in your answer to [2].

If you have questions about the hiring process or the position, please email search.ucsf.hiring@gmail.com