UCSF homework

Step 1: Load the data and libraries

In [1]:
```
import pandas as pd
import numpy as np
import matplotlib, os, glob
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline

path = "../UCSF/UCSF_SEARCH_hiring/data/*_*.csv"
all_files = glob.glob(path)
test_results = pd.read_csv("../UCSF/UCSF_SEARCH_hiring/data/ViralLoads.csv")
```

Step 2: Main body

In [2]:
```
community_names_list = []
time_periods_list = []
proportions_list = []

for file in all_files:
    df = pd.read_csv(file)
    df = pd.merge(df, test_results, how = 'inner', on='braceletid')
    #min & max time frame
    df['chcdate'] = pd.to_datetime(df['chcdate'])
    df['trdate'] = pd.to_datetime(df['trdate'])
    now = pd.to_datetime('today')
    df = df[((df['chcdate'] < now) & (df['chcdate'] > '2000-01-01')) | (df['c
hcdate'].isnull())]
    df = df[((df['trdate'] < now) & (df['trdate'] > '2000-01-01')) | (df['trd
ate'].isnull())]

    trkend_t = df['trdate'].max()
    chcstart_t = df['chcdate'].min()
    #finding VL @time frames
    df['date'] = pd.to_datetime(df['date'])
    df = df[((df['date'] >= chcstart_t) & (df['date'] <= trkend_t))]
    df = df.sort_values(by = ['braceletid','date'],ascending = [True, True])\
            .drop_duplicates(subset = ['braceletid'],keep = 'first')
    df['unsupp_t'] = 0
    df.loc[(df['HIV'] == 1) & (df['VL'] > 500), 'unsupp_t'] = 1
    #calculation of unsupp
    name = os.path.splitext(os.path.split(file)[1])[0]
    name = name.split('_', 1)
    community_names_list.append(name[0])
    time_periods_list.append(name[1])
    proportions_list.append(df['unsupp_t'].mean())

unsupp = pd.DataFrame({'community': community_names_list,
                       'time_period': time_periods_list,
                       'prop_unsupp': proportions_list})

unsupp_csv = unsupp.pivot(index='community', columns='time_period', values='p
rop_unsupp').reset_index()
unsupp_csv = unsupp_csv.rename(columns = {'0': 'prop_unsupp_0', '1': 'prop_un
supp_1', '2': 'prop_unsupp_2', '3': 'prop_unsupp_3'})
unsupp_csv.to_csv("../UCSF/Results/unsupp.csv",index = False)
unsupp_csv
```

Out[2]:

time_period	community	prop_unsupp_0	prop_unsupp_1	prop_unsupp_2	prop_unsupp_3
0	Bugamba	0.378840	0.363426	0.365134	0.383966
1	Bugono	0.397059	0.395248	0.418796	0.442590
2	Bware	0.360100	0.352090	0.388788	0.425201
3	Kadama	0.306488	0.354839	0.422932	0.493040
4	Kameke	0.332727	0.394144	0.424497	0.462626
5	Kamuge	0.442675	0.393740	0.327260	0.302339
6	Kazo	0.345483	0.374233	0.401961	0.456469
7	Kisegi	0.405697	0.352586	0.353488	0.364187
8	Kitare	0.366712	0.346287	0.415375	0.414494
9	Kitwe	0.364286	0.345083	0.378472	0.429705
10	Kiyeyi	0.327160	0.354601	0.444342	0.477290
11	Kiyunga	0.359375	0.366401	0.389258	0.409029
12	Magunga	0.382952	0.372742	0.401506	0.412285
13	Merikit	0.384321	0.405672	0.424696	0.452065
14	Mitooma	0.447840	0.346495	0.295652	0.283237
15	Muyembe	0.391304	0.385350	0.420209	0.426036
16	Nankoma	0.396171	0.372864	0.358679	0.350723
17	Nsiika	0.383534	0.365161	0.379000	0.409468
18	Nsiinze	0.316867	0.357977	0.386156	0.441040
19	Nyamrisra	0.375439	0.370283	0.391497	0.408925
20	Nyamuyanja	0.338681	0.347380	0.424088	0.466300
21	Nyatoto	0.366801	0.378151	0.420411	0.458725
22	Ogongo	0.476724	0.375969	0.324926	0.287754
23	Ongo	0.342857	0.377734	0.414307	0.452550
24	Othoro	0.386258	0.359945	0.326509	0.325939
25	Rubaare	0.375000	0.375935	0.378323	0.379544
26	Rugazi	0.345953	0.355044	0.394550	0.448143
27	Ruhoko	0.343137	0.377692	0.414361	0.433786
28	Rwashamaire	0.339019	0.380886	0.430834	0.443576
29	Sena	0.399123	0.347507	0.368889	0.389680
30	Sibuoche	0.355956	0.366816	0.412763	0.451754
31	Tom Mboya	0.400804	0.350145	0.341408	0.341940

Step 3 : Writeup

Missing data:

- df = pd.merge(df, test_results, how = 'inner', on='braceletid') This is where the most data is lost. 80% of bracelet_id's in Community_t.csv are missing from ViralLoads.csv are missing, preventing us from evaluating the unsuppressed viral load status of these patients.
- chcstart_t & trkend_t calculation: Some dates were input incorrectly (mistakes in data entry?), so I assumed the study was done after year 2000 and before now and removed all observations with dates before year 2000 or after today (May 15 , 2018). <1% data was lost
- df = df[((df['date'] >= chcstart_t) & (df['date'] <= trkend_t))] I removed all observations before chcstart_t or after trkend_t (<1% data lost).

In [3]:
```
#Data lost calculations

Number = []
Braceletid_left = []
Braceletid_left_2 = []
Braceletid_left_3 = []

for file in all_files:
    df = pd.read_csv(file)
    original = df['braceletid'].count()
    Number.append(original)
    df = pd.merge(df, test_results, how = 'inner', on='braceletid')

    Braceletid_left.append(df['braceletid'].count())

    df['chcdate'] = pd.to_datetime(df['chcdate'])
    df['trdate'] = pd.to_datetime(df['trdate'])
    now = pd.to_datetime('today')
    df = df[((df['chcdate'] < now) & (df['chcdate'] > '2000-01-01')) | (df['c
hcdate'].isnull())] # TODO: how many are removed?
    df = df[((df['trdate'] < now) & (df['trdate'] > '2000-01-01')) | (df['trd
ate'].isnull())]

    Braceletid_left_2.append(df['braceletid'].count())

    trkend_t = df['trdate'].max()
    chcstart_t = df['chcdate'].min()
    #finding VL @time frames
    df['date'] = pd.to_datetime(df['date'])
    df = df[((df['date'] >= chcstart_t) & (df['date'] <= trkend_t))]

    Braceletid_left_3.append(df['braceletid'].count())

    df = df.sort_values(by = ['braceletid','date'],ascending = [True, True])\
            .drop_duplicates(subset = ['braceletid'],keep = 'first')

sum(Number), sum(Braceletid_left)/sum(Number), sum(Braceletid_left_2)/sum(Bra
celetid_left), sum(Braceletid_left_3)/sum(Braceletid_left_2)
```

Out[3]: (926496, 0.1910337443442821, 0.9989434550714157, 0.9923418455360425)

Step 4 : Question 3

We don't have the information on how the ART influnces the viral loads and the HIV status.