Stanford Blueprint Datathon

Yiran Liu
Quenton Bubb
Sai Gourisankar
Jeremy Binagia

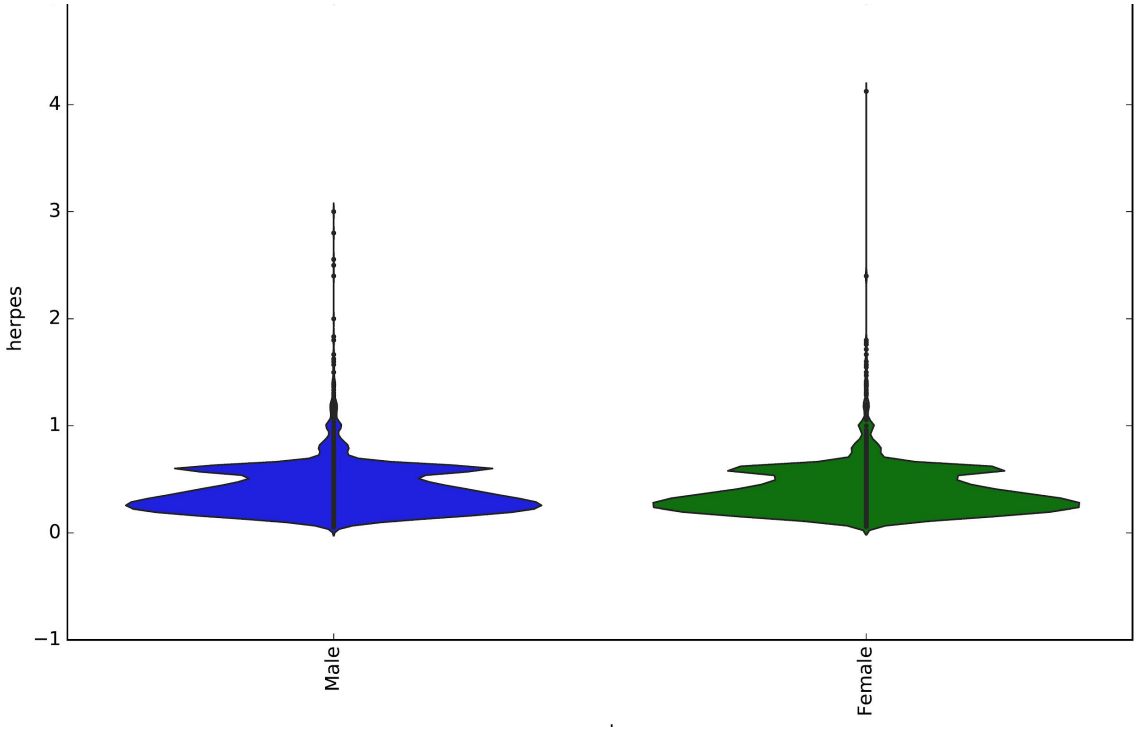
Preface

It is well-known that sexually transmitted diseases are increasing in prevalence within a broad set of unique communities across the United States. A key goal of public health infrastructure is identify and characterize populations that are vulnerable or at higher risk of contracting these sometimes life-threatening diseases in order to minimize harm and maximize positive health outcomes. Analysis of demographic data can often reveal subsets of populations that can benefit from targeted interventions that reflect their needs. Here, we've analyzed a large set of demographic and STD diagnostic data across the United States with the goal of identifying an underserved population that may be at high risk of contracting an STD. Due to statistical concerns regarding the sample sizes of each distinct demographic, we decided to focus on a subset of cohorts (individuals over 65 years of age) in a defined geographical location (California). As the geriatric population rapidly increases in size across the US, medical and public health infrastructure must concurrently evolve to accommodate their needs. Due to literature reports of additional public health risks and concerns regarding the elderly population, we explored these cohorts in order to try to recapitulate or add additional depth to our understanding of sexually transmitted disease risks and demographic factors that pertain to these individuals.

Why we chose to focus on the elderly in California

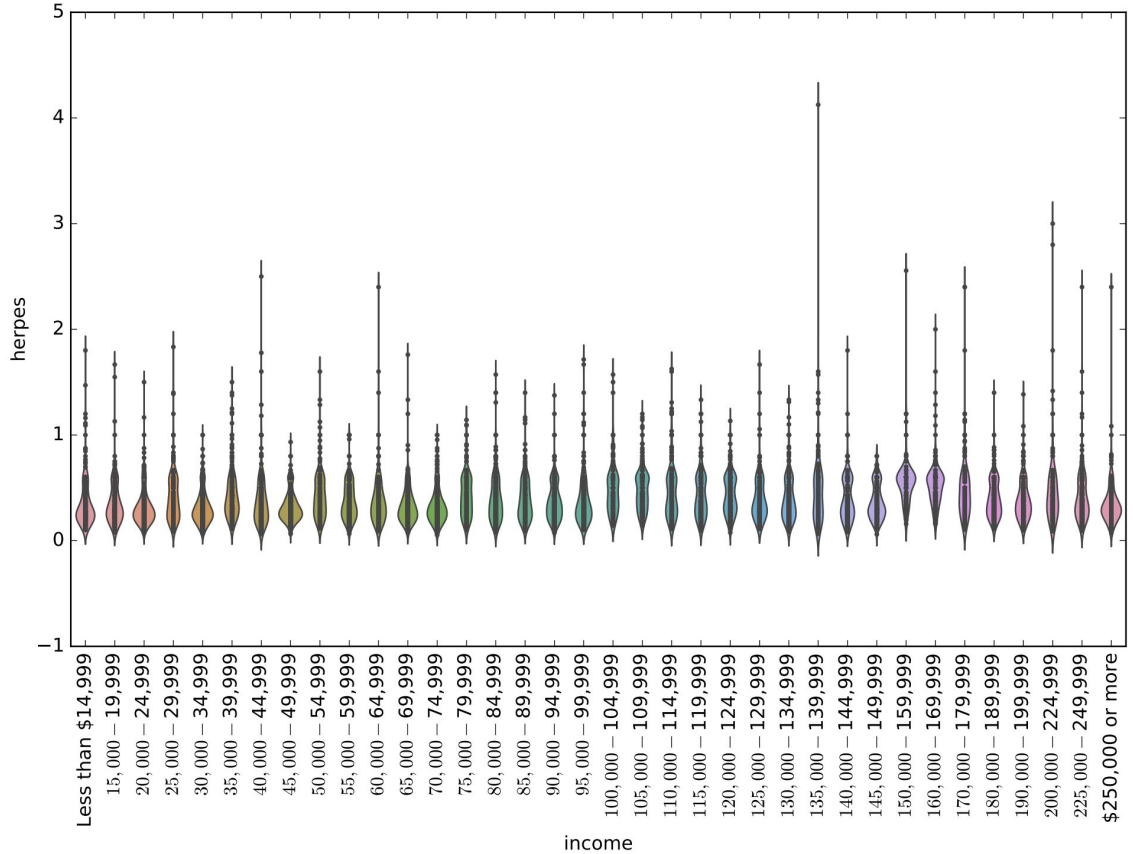
- Old people have sex too!
- Medicare is a huge cost burden
- Older adults represent the largest increase in in-office treatment of STDs from 2014-2017
- Potential factors include:
  - Lack of modern attitudes/knowledge about STDs, safer sex, etc.
  - Lower screening
  - Weaker immune systems

There is no correlation between gender and STD disease status within this data set, contrary to government reported data (1)

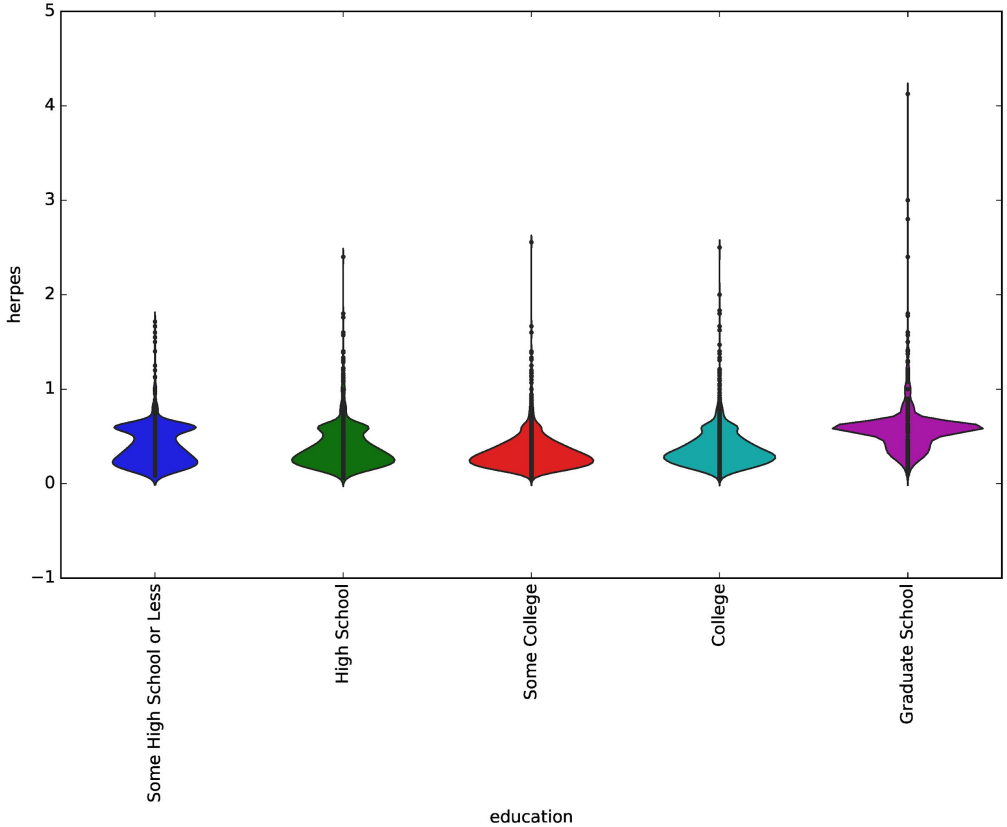


Similar trends found for all other STDs

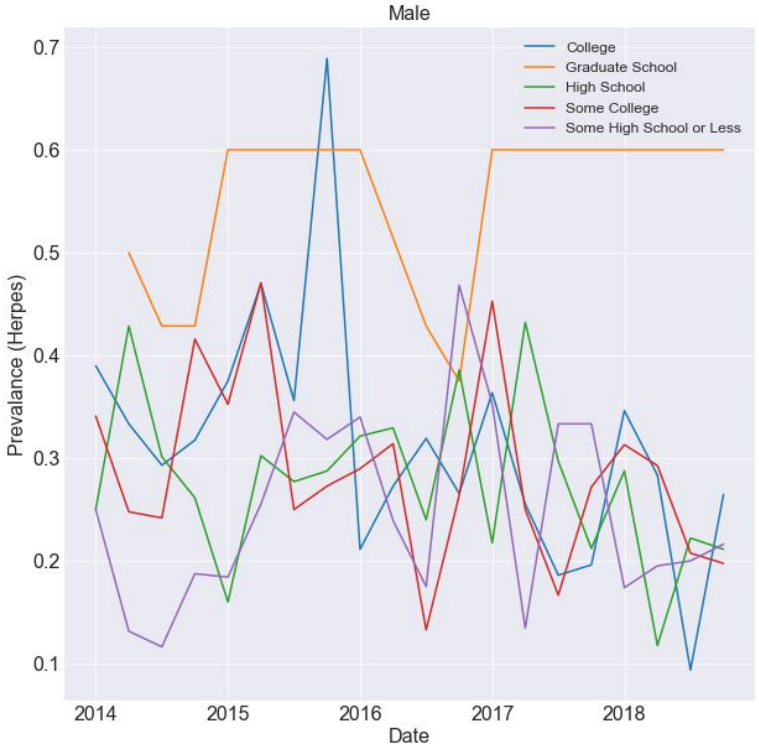
Across income distributions, there is no apparent income bracket alone that is at particular risk

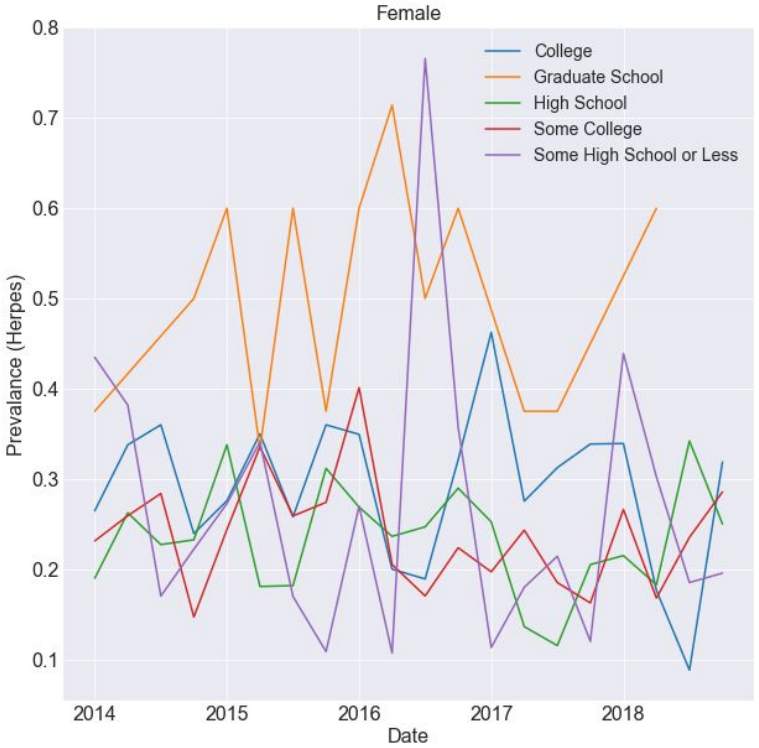


Education level within the 65+ cohort reveals potential risk factors, namely among those who went to graduate school



This is corroborated by plots of prevalence over time for specific cohorts

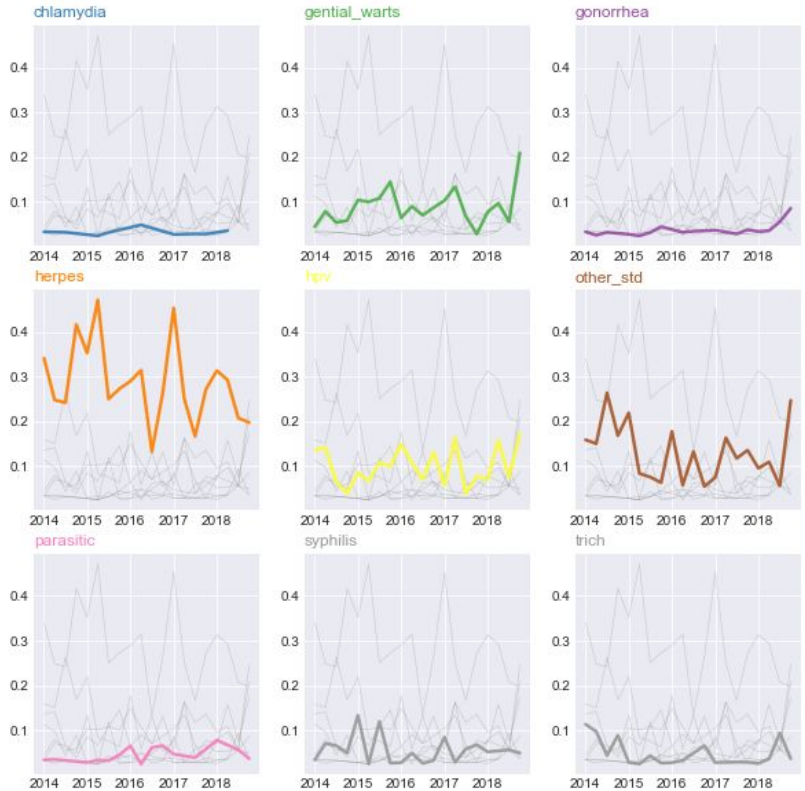




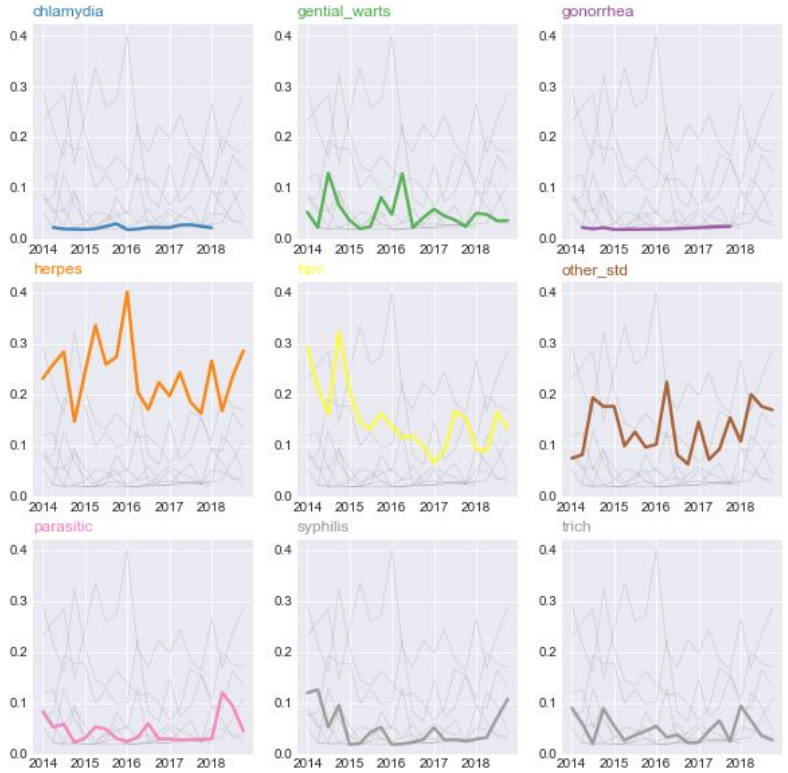
65 - 74 years old in the $70,000 - $74,999 income bracket (cohort with most recorded data)

Important since herpes appears to be the most prevalent STD

Male



Female



65 - 74 years old in the $70,000 - $74,999 income bracket with some college

Statistical Model: Multiple Linear Regression

- We ran a multiple linear regression over all the gender, income, and education categories within the 65+ cohorts in California. We encoded each value as a 0 or 1 (one-hot encoding; for example, gender was split into two features, gender_Male and gender_Female, and, for example, a male group was given a value of 1 in the gender_Male column and a 0 in the gender_Female column).

- After converting the categories to numerical data, we ran a simple linear regression where the dependent variable, *y*, was the prevalence of each STD, and **X** was a vector of all the other independent *demographic* variables (for example, **X** would be = [x_gender_Female, x_gender_Male, x_education_College, …]).

- We tested the null hypothesis, for each independent variable in **X**, that it's coefficient was 0, and obtained a p-value for that test. In addition, we obtained a Bonferroni-corrected significance level accounting for multiple-hypothesis testing by dividing a threshold of **0.05** by the number of independent demographic variables, and compared the p-values of each coefficient to that threshold.

Statistical Model, pt. 2

This analysis led us to exclude, for each STD, any variable associated with a coefficient not significantly different than zero; *i.e.* the variation in that STD could not be explained by the independent variable.

From this analysis, we observed that:

- For most STDs, either low or high incomes did not significantly explain prevalence, whereas middle incomes did.
    - This could let us, or other researchers, high and low incomes into one or two buckets, e.g. <$40,000 and >$75,000.
        - However, this might reflect the challenges of collecting data from low-income and very high-income people, rather than inherent variation,
    - Some STDs, such as trich and parasitic, had too much missing data to allow any of the demographic variables by themselves to explain much of the variance.
- Gender did not matter statistically, which is contrary to government-reported data (not included in this analysis) and probably reflects the internal biases of the data construction.

Issues/assumptions with data set:

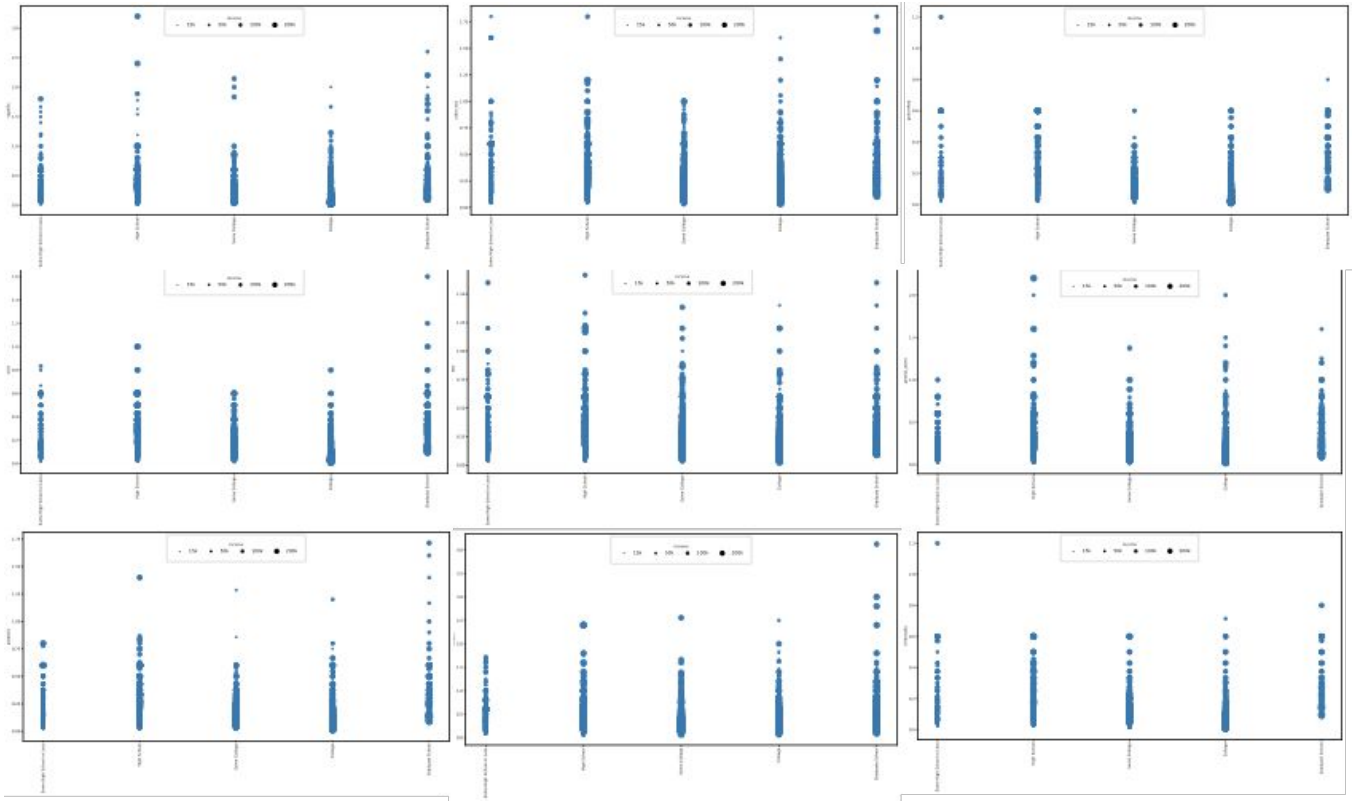
1) *The sample sizes pertaining to each unique set of demographic indicators is unknown.*
    a) *It is well-known that there is not a flat distribution of age, income, education, etc. across the United States. Without additional data on the population numbers within each unique set of demographic indicators, we cannot sufficiently power traditional statistical/epidemiological methods.*
    b) *Applying census data to the dataset in question is not appropriate with any additional information on how the survey was run.*
    c) *Furthermore, the statistical weights between each unique set of demographic indicators that pertain to STD prevalences will consequently be over or underestimated, which can mask underserved or underrepresented populations.*

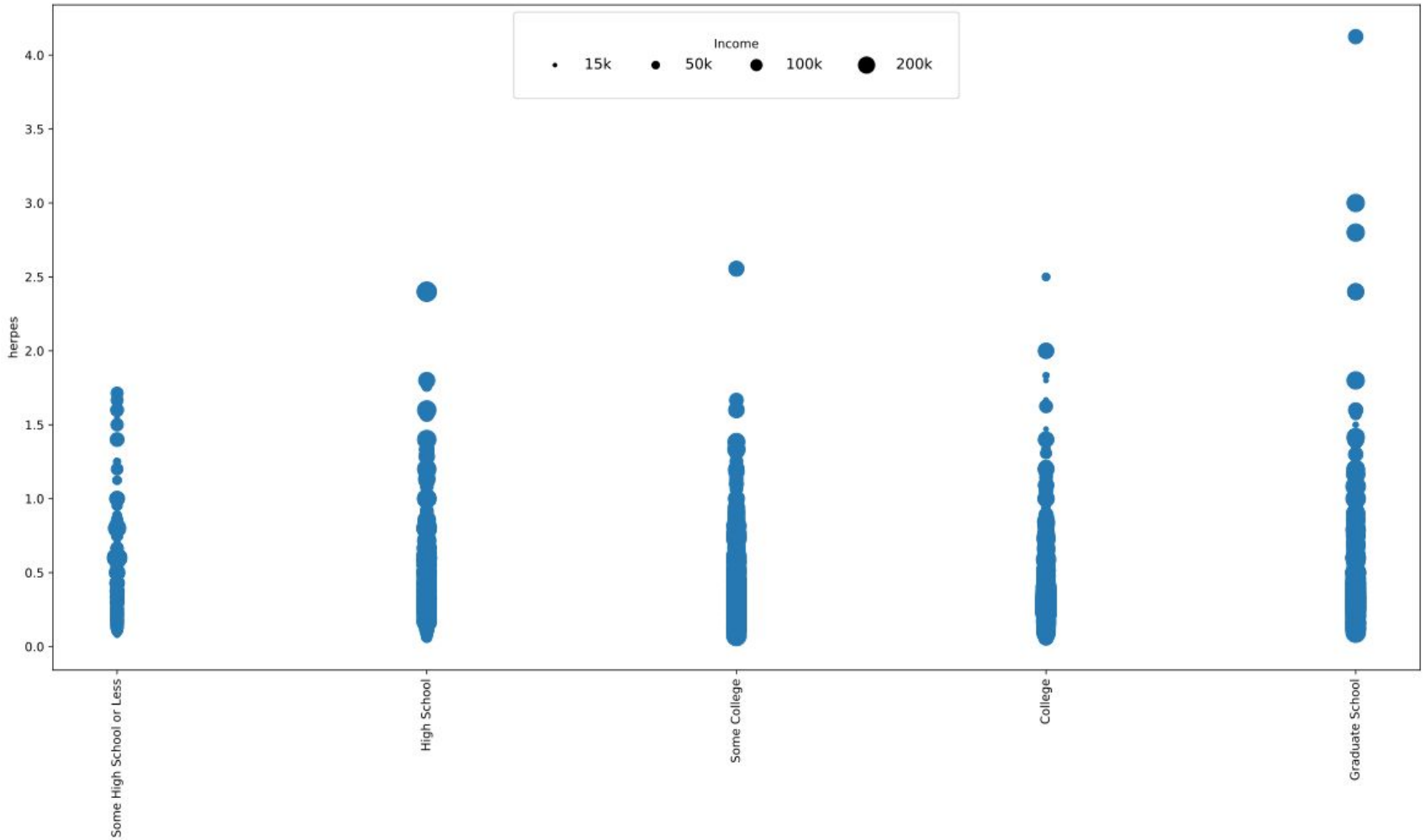
    ***For the purposes of this task, we assume there is an equal number of individuals within each cohort.*

2) *It is unknown whether or not the cohort within a unique set of demographic indicators changes in size over time (for example, loss of follow up, transference between demographic groups, etc.)*
3) *Missing data within the prevalence columns for STD's masks increases or decreases over the available time span, adding an additional layer of uncertainty within the trends.* *****This warranted identification of cohorts with large amounts of prevalence data, identified via the algorithm described in the analyses section.***

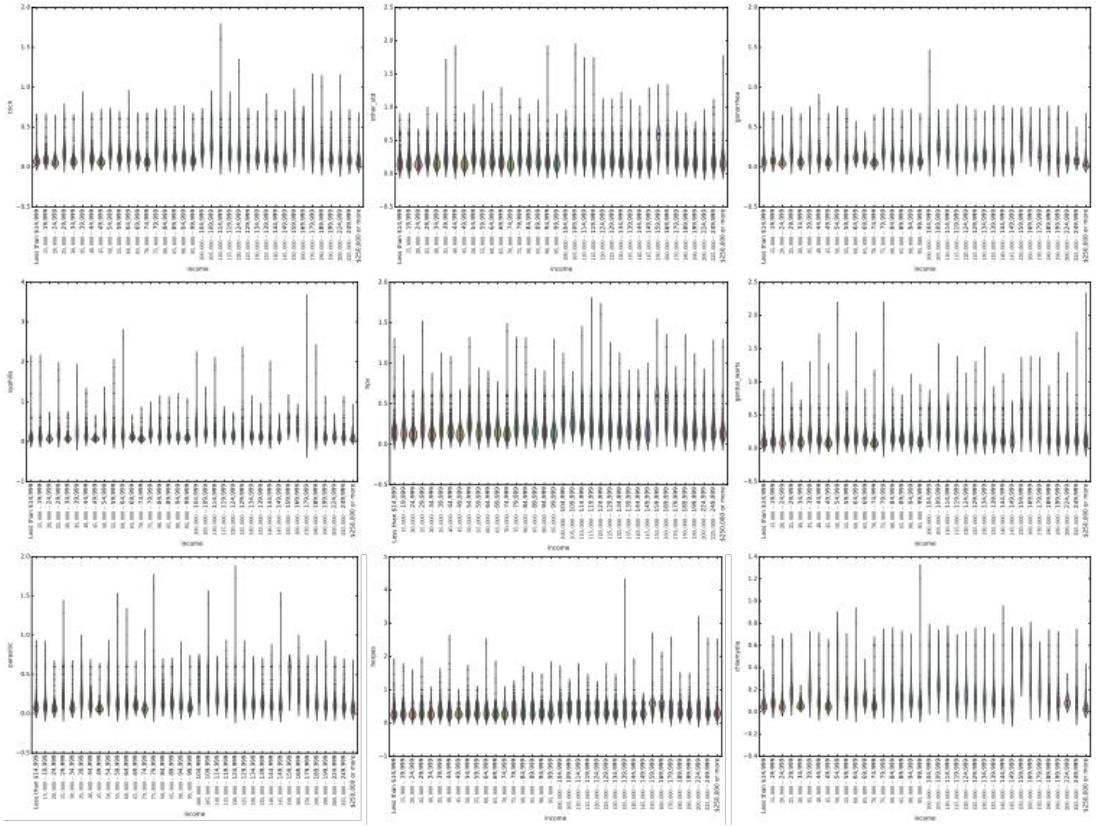
Backup Slides

Previous graduate school or high school education had a wider distribution of disease prevalence across various disease types .

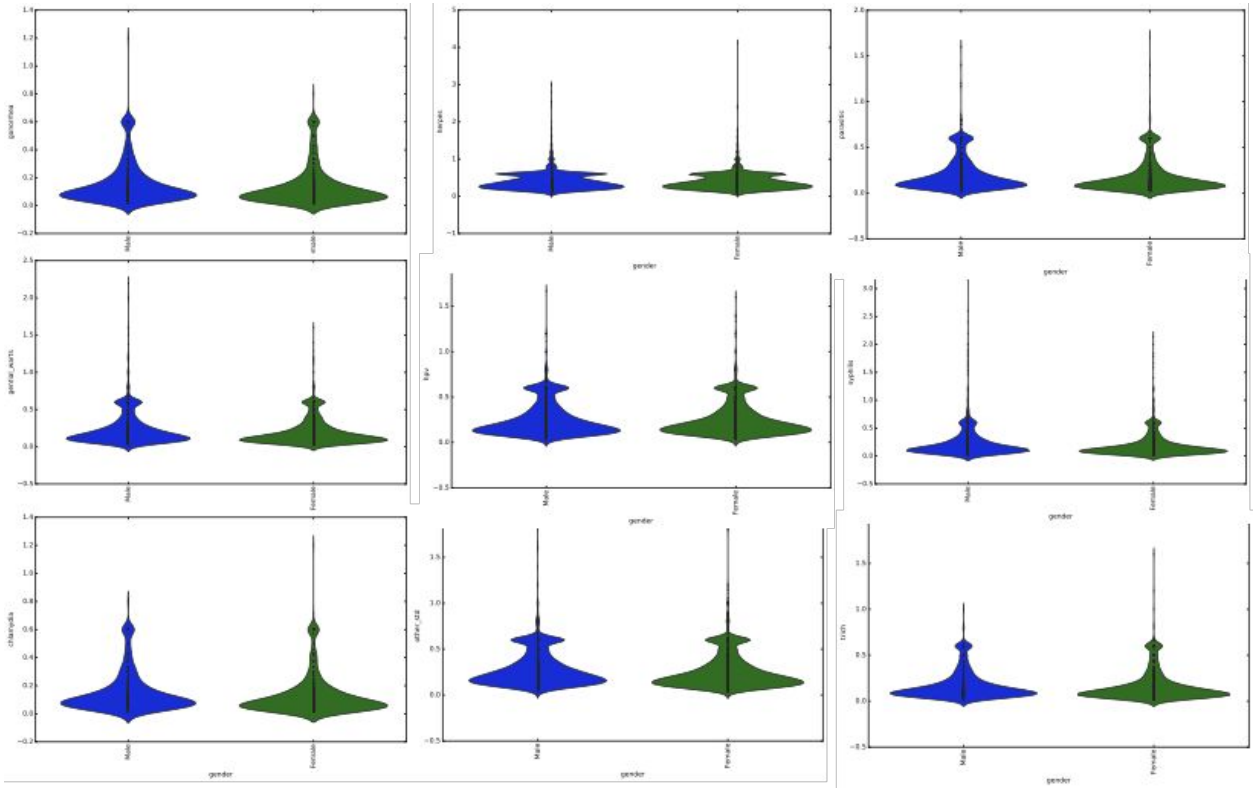



Income
15k
50k
100k
200k
herpes
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Some High School or Less
High School
Some College
College
Graduate School

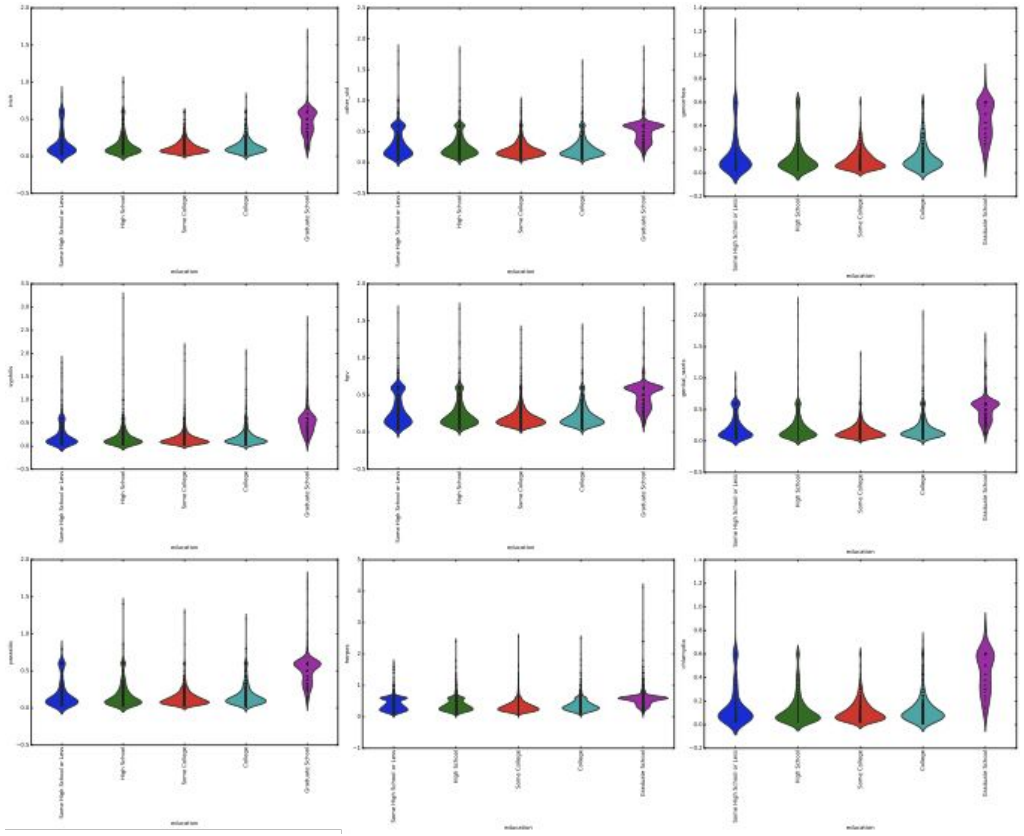
Across income distributions, there is no apparent income bracket alone that is at particular risk:



There is no correlation between gender and STD disease status within this data set, contrary to government reported data (1):



Education level within the 65+ cohort reveals potential risk factors, namely among those who went to graduate school



Statistical Model, pt 3: P-values for all variables with coefficients significantly > 0 in regression model

STD	Variable	P-value	STD	Variable	P-value	STD	Variable	P-value	STD	Variable	P-value	STD	Variable	P-value	STD	Variable	P-value	STD	Variable	P-value
chlamydia	income $40,000 - $44,999	0	herpes	income $190,000 - $199,999	0	gential warts	income $190,000 - $199,999	0	hpv	income $190,000 - $199,999	0	gonorrhea	income $190,000 - $199,99	0	other std	income $190,000 - $199,999	0	syphilis	income $190,000 - $199,99	0
chlamydia	income $95,000 - $99,999	0.0001355	herpes	income $40,000 - $44,999	0	gential warts	income $40,000 - $44,999	0	hpv	income $40,000 - $44,999	0	gonorrhea	income $40,000 - $44,999	0	other std	income $40,000 - $44,999	0	syphilis	income $40,000 - $44,999	0
chlamydia	income $135,000 - $139,99	2.23E-06	herpes	income $95,000 - $99,999	1.37E-06	gential warts	income $95,000 - $99,999	######	hpv	income $170,000 - $179,999	4.18E-07	gonorrhea	income $170,000 - $179,99	0.0009477	other std	income $95,000 - $99,999	9.45E-07	syphilis	income $95,000 - $99,999	2.76E-15
chlamydia	income $50,000 - $54,999	2.95E-35	herpes	income $45,000 - $49,999	1.89E-11	gential warts	income $45,000 - $49,999	######	hpv	income $95,000 - $99,999	9.60E-11	gonorrhea	income $135,000 - $139,99	1.48E-09	other std	income $45,000 - $49,999	2.30E-12	syphilis	income $45,000 - $49,999	2.69E-13
chlamydia	income $65,000 - $69,999	2.64E-22	herpes	income $50,000 - $54,999	3.22E-11	gential warts	income $135,000 - $139,999	######	hpv	income $45,000 - $49,999	2.96E-14	gonorrhea	income $50,000 - $54,999	6.38E-28	other std	income $135,000 - $139,999	9.90E-11	syphilis	income $135,000 - $139,99	9.47E-09
chlamydia	income Less than $14,999	2.48E-10	herpes	income $105,000 - $109,999	5.45E-09	gential warts	income $50,000 - $54,999	######	hpv	income $135,000 - $139,999	3.16E-11	gonorrhea	income $65,000 - $69,999	7.95E-14	other std	income $50,000 - $54,999	1.26E-32	syphilis	income $50,000 - $54,999	2.44E-26
chlamydia	income $85,000 - $89,999	3.23E-27	herpes	income $65,000 - $69,999	1.83E-06	gential warts	income $65,000 - $69,999	######	hpv	income $50,000 - $54,999	3.78E-33	gonorrhea	income Less than $14,999	2.86E-11	other std	income $65,000 - $69,999	1.52E-17	syphilis	income $65,000 - $69,999	7.89E-09
chlamydia	income $90,000 - $94,999	1.25E-16	herpes	income Less than $14,999	8.06E-14	gential warts	income Less than $14,999	######	hpv	income $65,000 - $69,999	4.28E-18	gonorrhea	income $85,000 - $89,999	3.79E-13	other std	income Less than $14,999	4.13E-23	syphilis	income Less than $14,999	1.26E-05
chlamydia	income $250,000 or more	9.78E-05	herpes	income $55,000 - $59,999	1.43E-08	gential warts	income $55,000 - $59,999	######	hpv	income Less than $14,999	2.67E-09	gonorrhea	income $90,000 - $94,999	5.64E-23	other std	income $55,000 - $59,999	1.15E-10	syphilis	income $55,000 - $59,999	4.20E-08
chlamydia	income $125,000 - $129,99	7.43E-07	herpes	income $90,000 - $94,999	1.24E-32	gential warts	income $85,000 - $89,999	######	hpv	income $55,000 - $59,999	1.68E-12	gonorrhea	income $180,000 - $189,99	0.0004673	other std	income $85,000 - $89,999	1.35E-11	syphilis	income $85,000 - $89,999	2.74E-11
chlamydia	income $70,000 - $74,999	4.19E-24	herpes	income $250,000 or more	3.50E-06	gential warts	income $90,000 - $94,999	######	hpv	income $85,000 - $89,999	1.05E-09	gonorrhea	income $125,000 - $129,99	3.28E-09	other std	income $90,000 - $94,999	6.62E-72	syphilis	income $90,000 - $94,999	9.30E-17
chlamydia	income $60,000 - $64,999	1.21E-30	herpes	income $125,000 - $129,999	1.22E-05	gential warts	income $125,000 - $129,999	######	hpv	income $90,000 - $94,999	7.00E-42	gonorrhea	income $70,000 - $74,999	8.75E-25	other std	income $125,000 - $129,999	3.73E-16	syphilis	income $125,000 - $129,99	7.26E-06
chlamydia	income $100,000 - $104,99	5.33E-14	herpes	income $70,000 - $74,999	2.76E-06	gential warts	income $70,000 - $74,999	######	hpv	income $125,000 - $129,999	2.81E-09	gonorrhea	income $60,000 - $64,999	6.25E-30	other std	income $70,000 - $74,999	3.35E-17	syphilis	income $70,000 - $74,999	4.46E-24
chlamydia	income $130,000 - $134,99	1.10E-08	herpes	income $60,000 - $64,999	1.38E-09	gential warts	income $60,000 - $64,999	######	hpv	income $70,000 - $74,999	2.96E-19	gonorrhea	income $100,000 - $104,99	6.92E-09	other std	income $60,000 - $64,999	6.80E-34	syphilis	income $145,000 - $149,99	1.53E-06
chlamydia	income $15,000 - $19,999	3.74E-32	herpes	income $25,000 - $29,999	0.0002885	gential warts	income $25,000 - $29,999	######	hpv	income $145,000 - $149,999	7.23E-11	gonorrhea	income $20,000 - $24,999	8.14E-05	other std	income $25,000 - $29,999	1.16E-06	syphilis	income $60,000 - $64,999	1.46E-27
chlamydia	income $30,000 - $34,999	1.03E-08	herpes	income $100,000 - $104,999	3.00E-27	gential warts	income $100,000 - $104,999	######	hpv	income $60,000 - $64,999	5.69E-33	gonorrhea	income $130,000 - $134,99	2.70E-16	other std	income $100,000 - $104,999	4.43E-43	syphilis	income $115,000 - $119,99	0.0006573
chlamydia	education High School	5.15E-26	herpes	income $130,000 - $134,999	1.64E-17	gential warts	income $130,000 - $134,999	######	hpv	income $25,000 - $29,999	2.92E-11	gonorrhea	income $15,000 - $19,999	3.31E-34	other std	income $130,000 - $134,999	7.61E-37	syphilis	income $100,000 - $104,99	4.00E-10
chlamydia	education Some College	1.39E-11	herpes	income $15,000 - $19,999	2.11E-12	gential warts	income $15,000 - $19,999	######	hpv	income $100,000 - $104,999	9.27E-54	gonorrhea	income $30,000 - $34,999	2.68E-08	other std	income $15,000 - $19,999	3.32E-28	syphilis	income $130,000 - $134,99	5.83E-15
chlamydia	education College	2.83E-07	herpes	income $110,000 - $114,999	0.0007466	gential warts	income $30,000 - $34,999	######	hpv	income $20,000 - $24,999	5.52E-05	gonorrhea	education High School	4.77E-21	other std	income $30,000 - $34,999	1.37E-06	syphilis	income $15,000 - $19,999	1.46E-23
chlamydia	education Graduate School	7.31E-26	herpes	education High School	2.21E-12	gential warts	education High School	######	hpv	income $130,000 - $134,999	2.56E-44	gonorrhea	education Some College	0.0003407	other std	education High School	3.46E-25	syphilis	income $30,000 - $34,999	0.000171
chlamydia	gender Male	4.34E-137	herpes	education Some College	1.72E-08	gential warts	education Some College	######	hpv	income $15,000 - $19,999	7.33E-39	gonorrhea	education College	2.38E-12	other std	education Some College	2.88E-15	syphilis	education High School	3.57E-24
			herpes	education Some High School or Les	6.16E-111	gential warts	education Some High School or Les	######	hpv	income $30,000 - $34,999	1.30E-08	gonorrhea	education Graduate School	4.07E-19	other std	education Some High School or Les	2.33E-53	syphilis	education Some College	1.12E-07
			herpes	education College	2.25E-14	gential warts	education College	######	hpv	education High School	1.15E-30	gonorrhea	gender Male	3.56E-178	other std	education College	1.47E-05	syphilis	education College	7.49E-10
			herpes	education Graduate School	3.92E-140	gential warts	education Graduate School	######	hpv	education Some College	1.16E-11				other std	education Graduate School	#######	syphilis	education Graduate School	2.83E-29
			herpes	gender Female	4.78E-80	gential warts	gender Male	0	hpv	education Some High School or Les	3.93E-27				other std	gender Male	0	syphilis	gender Male	8.62E-269
			herpes	gender Male	0				hpv	education College	1.27E-11									
									hpv	education Graduate School	#######									
									hpv	gender Male	0									