**Section E: Attach your structured thesis plan narrative (2-3 pages, single spaced, 11 or 12 point font):**

**Title:** The pattern of risk factors associated with STI testing in local school districts and States in the United States using YRBSS 2019 data.

Student Name: Yu He

**Introduction:** Brief background/context for your study, gaps in relevant knowledge, the significance of the proposed research, the study question you hope to answer

Burden of disease

The health of the American young population aged 15 to 24 is facing great challenges from sexually transmitted infections (STI), which include chlamydia, gonorrhea, genital herpes, human papillomavirus (HPV), syphilis, and HIV, defined by the Centers for Diseases Control and Prevention (CDC). Young people account for at least half of all new sexually transmitted infections (STIs) contracted each year, and a quarter of sexually active adolescent females have STIs(Shannon & Klausner, 2018). In 2021 alone, there were at least 2.5 million reported cases of chlamydia, gonorrhea, syphilis, and congenital syphilis, according to the CDC. From 2020 to 2021, syphilis cases alone have risen by 26%, and the number of syphilis cases last year was the highest since 1948(Liddon et al., 2022). Some believe that the number of cases is underreported and that there are untreated infections so these numbers may be higher.

The consequence or harm of STI among teenagers.

Most STDs cause much suffering in the acute phase and, in some cases, can produce long-term damage with serious consequences. STDs may cause a variety of complications, such as miscarriage, infertility, heart and bone, and even brain damage, seriously affecting health(Nicoll & Hamers, 2002). The increase in the incidence of STDs may also bring many serious consequences, such as increased drug resistance to STDs, poor health quality of the next generation, increased social medical costs, and increased social insecurity(WHO regional office for Europe, 2001).

Considering the serious consequences and burden of disease on the population and individual level of adolescent health, multiple organizations and agencies recommend some level of STI screening for adolescents. In addition, STIs are characterized by un-symptomatic occurrences, which can infect others unknowingly(Samkange-Zeeb et al., 2011).

Different resources of STI testing in different states and local school districts.

According to the National Academies of Sciences, health inequities become the major challenge for the nation, which results in health disparities between states and school districts all around the United States(Weinstein et al., 2017). In addition, the communities encountered threats from disparities in health, sociality, ethnicity, economy, employment, and education that can impede the deployment of STI testing to adolescents.

STI testing is an effective way of reducing STIs.

The implementation of STI testing guided by the WHO **A**ffordable, **S**ensitive, **S**pecific, User-friendly, **R**apid and robust, **E**quipment-free, and **D**eliverable criteria become a benchmark for controlling the rapid spreading of STIs globally(Peeling, Holmes, et al., 2006). The STI screening provided its value in controlling the spreading of STIs by interrupting the transmission chain between the patients and their sexually active partners(Dewart et al., 2018). Most STIs can be treated with a single dose of antibiotics. Therefore, the early and high prevalence of testing is crucial in controlling the rising STIs among adolescents in the United States(Peeling, Mabey, et al., 2006).

Research gap

Though, the low prevalence of STI testing is threatening the populational health of teenagers(St Lawrence et al., 2002).

Previous studies identified barriers, such as limited access to healthcare, parental pressure, social stigma, and limited knowledge of STIs and their consequences, that prevent adolescents from receiving STI testing(Copen et al., 2015-2016 CDC.)(Bronwen Lichtenstein, 2003).

However, limited studies research the pattern of risk factors associated with STI testing and compare the difference between the pattern of risk factors in different states and local school districts. The association between risk factors and STI testing in local school districts and states is not fully understood.

Significance of study

In this study, we used the complex survey design to improve the precision of the sample estimate for the model. The understanding of risk factor patterns between different states and local school districts is important in identifying and providing appropriate aid and support precisely for vulnerable populations among adolescents.

Research question

The purpose of this study is to examine the pattern of risk factors associated with STI testing and compare the difference between the pattern of risk factors in different states and local school districts using the data from the Youth Risk Behavior Surveillance System (YRBSS) 2019 database. We examine the association between age, gender, school grade, race, unintentional injuries and violence, tabacco use, alcohol and other drug use, risky sexual behavior, cognitive, and English speaking with STI testing other than HIV.

**Study Aims:** 1‐2 specific study aims/study questions

Investigate the

Aim 1: To identify the risk factors associated with STI testing other than HIV among middle and high school students.

Aim 2: To compare the patterns of risk factors associated with STI testing other than HIV among middle and high school students between selected states and local school districts.

**Hypotheses**: As appropriate, 1‐2 independent hypotheses, include direction of effect/association

**Methods:** Data set name(s), dates of data collection, demographic description and size of the study population, inclusion/exclusion criteria, study design, exposures/interventions, covariates, outcomes, and statistical approach including power calculations

Study Population and Design

The study participants were collected from the Youth Risk Behavior Surveillance System (YRBSS) high school 2019 dataset. We included participants who were enrolled in high schools of selected states and local school districts aged from 12 years to 18 years or older without missing data. The local high school district data were collected from 14 local high school districts selected by the YRBSS, including, Broward County, FL, Chicago, IL, Eaton Consortium, MI, Fort Worth, TX, Genesee Consortium, MI, Hillsborough County, FL, Los Angeles, CA, Newark, NJ, Orange County, FL, Palm Beach County, FL, Pasco County, FL, Philadelphia, PA, Portland, OR, Shelby County, TN. **The local school district data included 12,444 participants.**12 states were selected from the YRBSS state data after removing all the missing values, including Alabama, Arkansas, Illinois, Iowa, Kentucky, Michigan, Mississippi, Nebraska, Oklahoma, Pennsylvania, South Carolina, West Virginia. **The state data included 16,115 participants.**

Research question

The purpose of this study is to examine the pattern of risk factors associated with STI testing and compare the difference between the pattern of risk factors in different states and local school districts using the data from the Youth Risk Behavior Surveillance System (YRBSS) 2019 database. We examine the association between age, gender, school grade, race, unintentional injuries and violence, tabacco use, alcohol and other drug use, risky sexual behavior, cognitive, and English speaking with STI testing other than HIV.

**Outcome variable:**

Q85. During the past 12 months, have you been tested for a sexually transmitted disease (STD) other than HIV, such as

chlamydia or gonorrhea?

A. Yes

B. No

C. Not sure

Variable label: STD testing

**Exposure variable:**

age, sex, grade, race4,

bmi,

bmi percentile,

qnothhpl- Used birth control pills; an IUD (such as Mirena or

ParaGard) or implant (such as Implanon or

Nexplanon); or a shot (such as Depo-Provera), patch

(such as OrthoEvra), or birth control ring (such as

NuvaRing) before last sexual intercourse,

Q17. During the past 12 months, how many times were you in a physical fight?

A. 0 times B. 1 time C. 2 or 3 times D. 4 or 5 times E. 6 or 7 times F. 8 or 9 times G. 10 or 11 times H. 12 or more times

QN17

Q20. During the past 12 months, how many times did anyone force you to do sexual things that you did not want to do? (Count such things as kissing, touching, or being physically forced to have sexual intercourse.)

A. 0 times B. 1 time C. 2 or 3 times D. 4 or 5 times E. 6 or more times

Q21. During the past 12 months, how many times did someone you were dating or going out with force you to do sexual things that you did not want to do? (Count such things as kissing, touching, or being physically forced to have sexual intercourse.)

A. I did not date or go out with anyone during the past 12 months B. 0 times C. 1 time D. 2 or 3 times E. 4 or 5 times F. 6 or more times

Q23. During the past 12 months, have you ever been bullied on school property? A. Yes B. No

Q26. During the past 12 months, did you ever seriously consider attempting suicide? A. Yes B. No

Q30. Have you ever tried cigarette smoking, even one or two puffs? A. Yes B. No

Q41. During the past 30 days, on how many days did you have at least one drink of alcohol?

A. 0 days B. 1 or 2 days – light drinker

C. 3 to 5 days D. 6 to 9 days E. 10 to 19 days - moderate drinker

F. 20 to 29 days G. All 30 days - heavy drinker

Q47. During the past 30 days, how many times did you use marijuana?

A. 0 times B. 1 or 2 times

C. 3 to 9 times D. 10 to 19 times

E. 20 to 39 times F. 40 or more times

Q49. During your life, how many times have you taken prescription pain medicine without a doctor's prescription or differently than how a doctor told you to use it? A. 0 times B. 1 or 2 times C. 3 to 9 times D. 10 to 19 times E. 20 to 39 times F. 40 or more times

Q50. During your life, how many times have you used any form of cocaine, including powder, crack, or freebase? A. 0 times B. 1 or 2 times C. 3 to 9 times D. 10 to 19 times E. 20 to 39 times F. 40 or more times

Q52. During your life, how many times have you used heroin (also called smack, junk, or China White)? A. 0 times B. 1 or 2 times C. 3 to 9 times D. 10 to 19 times E. 20 to 39 times F. 40 or more times

Q53. During your life, how many times have you used methamphetamines (also called speed, crystal meth, crank, ice, or meth)? A. 0 times B. 1 or 2 times C. 3 to 9 times D. 10 to 19 times E. 20 to 39 times F. 40 or more times

Q56. During your life, how many times have you used a needle to inject any illegal drug into your body? A. 0 times B. 1 time C. 2 or more times

Q58. Have you ever had sexual intercourse? A. Yes B. No

Q61. During the past 3 months, with how many people did you have sexual intercourse? A. I have never had sexual intercourse B. I have had sexual intercourse, but not during the past 3 months C. 1 person D. 2 people E. 3 people F. 4 people G. 5 people H. 6 or more people

Q63. The last time you had sexual intercourse, did you or your partner use a condom? A. I have never had sexual intercourse B. Yes C. No

Q65. During your life, with whom have you had sexual contact? A. I have never had sexual contact B. Females C. Males D. Females and males

Q66. Which of the following best describes you? A. Heterosexual (straight) B. Gay or lesbian C. Bisexual D. Not sure

Q84. Have you ever been tested for HIV, the virus that causes AIDS? (Do not count tests done if you donated blood.) A. Yes B. No C. Not sure

Q87. Has a doctor or nurse ever told you that you have asthma? A. Yes B. No C. Not sure

Q89. During the past 12 months, how would you describe your grades in school? A. Mostly A's B. Mostly B's C. Mostly C's D. Mostly D's E. Mostly F's F. None of these grades G. Not sure

**References to date:** (support for your introduction, methods (include 1‐2 references for the data set if available)

A Directed Acyclic Graph (DAG) is a visualization tool used to represent causal relationships between variables in a study. Here, we provide a textual representation of a DAG for the research question, outcome variable, and exposure variables mentioned above:

Demographic Factors:

a. Age

b. Sex

c. Grade

d. Race (race4)

Anthropometric Factors:

a. Height

b. BMI

c. BMI Percentile

d. Obesity Indicator (qnobese)

Sexual Factors:

a. Sexual Identity (sexed)

b. Birth Control Use (qnothhpl)

c. Sexual Orientation (Q66)

Food Allergy:

a. Food Allergy (Qfoodallergy)

Unintentional Injuries and Violence:

a. Physical fights (Q17)

b. Forced sexual activities (Q20)

c. Dating-related forced sexual activities (Q21)

d. Bullying on school property (Q23)

e. Suicidal thoughts (Q26)

Tobacco Use:

a. Tried cigarette smoking (Q30)

Alcohol and Other Drug Use:

a. Alcohol consumption (Q41)

b. Marijuana use (Q47)

c. Prescription pain medicine misuse (Q49)

d. Cocaine use (Q50)

e. Heroin use (Q52)

f. Methamphetamine use (Q53)

g. Needle use for injecting illegal drugs (Q56)

Risky Sexual Behavior:

a. Sexual intercourse (Q58)

b. Number of sexual partners (Q61)

c. Condom use (Q63)

d. Sexual contact with different genders (Q65)

Cognitive and Health Factors:

a. HIV testing (Q84)

b. Asthma diagnosis (Q87)

c. School grades (Q89)

Outcome Variable:

a. STI testing other than HIV (Q85)

DAG Representation:

1 --> 10

2 --> 10

3 --> 10

4 --> 10

5 --> 10

6 --> 10

7 --> 10

8 --> 10

9 --> 10

This representation implies that each of the nine categories of exposure variables (demographic factors, anthropometric factors, sexual factors, food allergy, unintentional injuries and violence, tobacco use, alcohol and other drug use, risky sexual behavior, and cognitive and health factors) is directly connected to the outcome variable, STI testing other than HIV (Q85). The arrows indicate the direction of causality from each exposure variable to the outcome variable. Note that the DAG does not include specific connections between individual exposure variables within each category, as this would make the representation overly complex.

Research question

The purpose of this study is to examine the pattern of risk factors associated with STI testing and compare the difference between the pattern of risk factors in different states and local school districts using the data from the Youth Risk Behavior Surveillance System (YRBSS) 2019 database. We examine the association between age, gender, school grade, race, unintentional injuries and violence, tabacco use, alcohol and other drug use, risky sexual behavior, cognitive, and English speaking with STI testing other than HIV.

Outcome variable is:

Q85. During the past 12 months, have you been tested for a sexually transmitted disease (STD) other than HIV, such as

chlamydia or gonorrhea?

Exposure variable is:

Age, sex, race, grade, bmi, bmi percentile, sex identification,

Based on the information provided, the primary goal of the analysis is to examine the pattern of risk factors associated with STI testing and compare the difference between the pattern of risk factors in different states and local school districts. Since the outcome variable (STD testing) is a categorical variable with three possible responses (Yes, No, Not sure), you may consider using classification algorithms for this analysis. Here are some appropriate machine learning methods:

1. Logistic Regression: A statistical method that works well for binary or ordinal categorical outcome variables. You can use multinomial logistic regression for more than two categories. It can handle multiple predictor variables and estimate the relationship between the predictors and the outcome.
2. Decision Trees: Decision trees can handle both categorical and continuous variables, making them suitable for this type of analysis. They are easily interpretable and can help you identify the most important variables for predicting STI testing.
3. Random Forest: An ensemble learning method that constructs multiple decision trees and combines their output. Random forests are known for their robustness and ability to handle a large number of predictor variables.
4. Gradient Boosting Machines (GBM): Similar to random forests, GBMs are also ensemble methods that construct multiple decision trees. However, GBMs iteratively improve the model by focusing on instances that were misclassified in previous iterations.
5. Support Vector Machines (SVM): A powerful classification algorithm that works well with high-dimensional data and complex relationships between variables. For multi-class problems, you can use one-vs-one or one-vs-all strategies.
6. K-Nearest Neighbors (KNN): A simple, distance-based classification algorithm that works well with small datasets and can handle multi-class problems. The algorithm predicts the outcome based on the majority class of its k nearest neighbors.
7. Neural Networks: Artificial neural networks can be used for complex classification problems. They consist of layers of interconnected nodes and can approximate complex, non-linear relationships between variables.

Before selecting an algorithm, it's crucial to preprocess the data, such as handling missing values, encoding categorical variables, and scaling or normalizing numerical features. You should also split the dataset into training and testing sets, and perform cross-validation to estimate the model's performance.

Finally, to determine the most appropriate method for your analysis, compare the performance of the different algorithms using metrics such as accuracy, precision, recall, F1-score, or the area under the receiver operating characteristic (ROC) curve. The algorithm that performs best based on these metrics can be considered the most suitable for this analysis.

I have listed the variables in the question and grouped them under their respective categories.

1. Demographics:

* Age
* Sex
* Grade
* Race4

1. Health-related factors:

* BMI
* BMI percentile
* Q87: Asthma

1. Unintentional injuries and violence:

* Q17: Physical fights in the past 12 months
* Q20: Forced sexual encounters in the past 12 months
* Q21: Forced sexual encounters by someone dated in the past 12 months
* Q23: Bullied on school property in the past 12 months
* Q26: Seriously considered attempting suicide in the past 12 months

1. Tobacco use:

* Q30: Ever tried cigarette smoking

1. Alcohol and other drug use:

* Q41: Alcohol consumption in the past 30 days
* Q47: Marijuana use in the past 30 days
* Q49: Prescription pain medicine misuse
* Q50: Cocaine use
* Q52: Heroin use
* Q53: Methamphetamine use
* Q56: Illegal drug injection

1. Risky sexual behavior:

* Q58: Ever had sexual intercourse
* Q61: Number of sexual partners in the past 3 months
* Q63: Condom use during last sexual intercourse
* Q65: Sexual contact with partners (female, male, or both)
* Q66: Sexual orientation
* Q84: Ever been tested for HIV

1. Cognitive factors:

* Q89: Grades in school

1. English-speaking status: (Not provided in the question)

Now let's create a DAG (Directed Acyclic Graph) based on these variables. Please note that a DAG would be better represented visually. However, I will describe the connections between the variables to the best of my ability.

In a DAG, we can start with the demographics as the root nodes, as they influence other variables. The paths in the DAG can be described as follows:

1. Demographics (Age, Sex, Grade, Race4) -> Health-related factors (BMI, BMI percentile, Q87: Asthma)
2. Demographics -> Unintentional injuries and violence (Q17, Q20, Q21, Q23, Q26)
3. Demographics -> Tobacco use (Q30)
4. Demographics -> Alcohol and other drug use (Q41, Q47, Q49, Q50, Q52, Q53, Q56)
5. Demographics -> Risky sexual behavior (Q58, Q61, Q63, Q65, Q66, Q84)
6. Demographics -> Cognitive factors (Q89)

Additionally, some variables might have direct connections, such as:

1. Unintentional injuries and violence -> Alcohol and other drug use
2. Tobacco use -> Alcohol and other drug use
3. Alcohol and other drug use -> Risky sexual behavior

Based on the DAG, the following variables should be considered for inclusion in a regression analysis model:

1. Demographics (Age, Sex, Grade, Race4)
2. Health-related factors (BMI, BMI percentile, Q87: Asthma)
3. Unintentional injuries and violence (Q17, Q20, Q21, Q23, Q26)
4. Tobacco use (Q30)
5. Alcohol and other drug use (Q41, Q47, Q49, Q50, Q52, Q53, Q56)
6. Risky sexual behavior (Q58, Q61, Q63, Q65, Q66, Q84)
7. Cognitive factors (Q89)

Please note that further domain knowledge might be needed to refine the DAG and the selection of variables for the regression model