**Title:** Comparing the Performance of Machine Learning Methods to Identify Patterns of Risk Factors Associated with STI Testing in Local School Districts and Selected 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 young American population aged 15 to 24 is facing significant 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 yearly, 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). Scientists believe that the number of cases is underreported and 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 severe consequences. STDs may cause various 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 and medical costs, and increased social insecurity(WHO regional office for Europe, 2001). Considering the severe consequences and disease burden on the population and individual level of adolescent health, multiple organizations and agencies recommend some 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 have become a major challenge for the nation, resulting in health disparities between states and school districts around the United States(Weinstein et al., 2017). In addition, the communities encountered threats from inequalities 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).

The complex YRBSS dataset and limitations of traditional statistic analysis methods, and the ML method is an effective ways of dealing with the challenge.

The explanation of STI testing requires complex and high-dimensional datasets such as the YRBSS dataset. However, Traditional statistical methods may have limited ability to handle such data. Machine learning mainly focuses on prediction, and epidemiology mainly needs to know causal effects (causal/etiologic inference), requires background knowledge, and is biased towards parametric/semiparametric estimation, so it is not easy to combine the two. However, in recent years an increasing number of studies on machine learning have been published in frontier epidemiological methods, with a primary focus on public health, explaining chronic diseases, risk factors for chronic diseases, and infectious, parasitic, and communicable diseases(dos Santos et al., 2019). Machine learning (ML) algorithms provide an effective tool for analyzing these datasets and can find hidden links and patterns in the data that may be difficult to spot using more traditional approaches(Mooney & Pejaver, 2018). These methods have the potential to considerably advance our knowledge of STI testing risk factors and make it possible to identify vulnerable individuals more precisely.

Research gap

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, STI-related attitudes, and limited knowledge of STIs and their consequences, that prevent adolescents from receiving STI testing(Copen et al., 2015-2016 CDC.)(Bronwen Lichtenstein, 2003)(Shepherd & Harwood, 2017). 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 has yet to be fully understood.

Significance of study

In this study, we used machine learning methods to improve the accuracy of the model’s ability to explain the association based on the YRBSS dataset. The understanding of risk factor patterns between different states and local school districts is essential 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 employ machine learning techniques to analyze the association between age, gender, school grade, race, unintentional injuries and violence, tobacco use, alcohol and other drug use, and risky sexual behavior with STI testing other than HIV.

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

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

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

Aim 3: Compare the performance of multinomial logistic regression, random forest, support vector machines (SVM), Elastic Net regression, ridge regression, lasso regression, and classification tree methods on the analysis of YRBSS dataset.

**Reference**

Bronwen Lichtenstein. (2003). Stigma as a barrier to treatment of sexually transmitted infection in the American deep south- issues of race, gender and poverty. *Social Science & Medicine*, *57*(12), 2435–2445. https://doi.org/https://doi.org/10.1016/j.socscimed.2003.08.002.

Copen, C. E., Dittus, P. J., & Leichliter, J. S. (n.d.). Confidentiality Concerns and Sexual and Reproductive Health Care Among Adolescents and Young Adults Aged 15-25 Key findings Data from the National Survey of Family Growth Percent. In *SOURCE: NCHS, National Survey of Family Growth*. http://www.cdc.gov/nchs/data/databriefs/db266\_table.pdf#3.

Dewart, C. M., Bernstein, K. T., Degroote, N. P., Romaguera, R., & Turner, A. N. (2018). Prevalence of Rectal Chlamydial and Gonococcal Infections: A Systematic Review. In *Sexually Transmitted Diseases* (Vol. 45, Issue 5, pp. 287–293). Lippincott Williams and Wilkins. https://doi.org/10.1097/OLQ.0000000000000754

dos Santos, B. S., Steiner, M. T. A., Fenerich, A. T., & Lima, R. H. P. (2019). Data mining and machine learning techniques applied to public health problems: A bibliometric analysis from 2009 to 2018. *Computers and Industrial Engineering*, *138*. https://doi.org/10.1016/j.cie.2019.106120

Liddon, N., Pampati, S., Dunville, R., Kilmer, G., & Steiner, R. J. (2022). Annual STI Testing Among Sexually Active Adolescents. *Pediatrics*, *149*(5). https://doi.org/10.1542/peds.2021-051893

Mooney, S. J., & Pejaver, V. (2018). *Big Data in Public Health: Terminology, Machine Learning, and Privacy*. https://doi.org/10.1146/annurev-publhealth

Nicoll, A., & Hamers, F. F. (2002). Are trends in HIV, gonorrhoea, and syphilis worsening in western Europe? In *British Medical Journal* (Vol. 324, Issue 7349, pp. 1324–1327). BMJ Publishing Group. https://doi.org/10.1136/bmj.324.7349.1324

Peeling, R. W., Holmes, K. K., Mabey, D., & Ronald, A. (2006). Rapid tests for sexually transmitted infections (STIs): The way forward. In *Sexually Transmitted Infections* (Vol. 82, Issue SUPPL. 5). https://doi.org/10.1136/sti.2006.024265

Peeling, R. W., Mabey, D., Herring, A., & Hook, E. W. (2006). Why do we need quality-assured diagnostic tests for sexually transmitted infections? In *Nature Reviews Microbiology* (Vol. 4, Issue 12, pp. 909–921). https://doi.org/10.1038/nrmicro1555

Samkange-Zeeb, F. N., Spallek, L., & Zeeb, H. (2011). Awareness and knowledge of sexually transmitted diseases (STDs) among school-going adolescents in Europe: A systematic review of published literature. In *BMC Public Health* (Vol. 11). https://doi.org/10.1186/1471-2458-11-727

Shannon, C. L., & Klausner, J. D. (2018). The growing epidemic of sexually transmitted infections in adolescents: A neglected population. In *Current Opinion in Pediatrics* (Vol. 30, Issue 1, pp. 137–143). Lippincott Williams and Wilkins. https://doi.org/10.1097/MOP.0000000000000578

Shepherd, L., & Harwood, H. (2017). The role of STI-related attitudes on screening attendance in young adults. *Psychology, Health and Medicine*, *22*(6), 753–758. https://doi.org/10.1080/13548506.2016.1234715

St Lawrence, J. S., Montaño, D. E., Kasprzyk, D., Phillips, W. R., Armstrong, K., & Leichliter, J. S. (2002). Screening, Testing, Case Reporting, and Clinical and Partner Notification Practices: A National Survey of US Physicians. In *American Journal of Public Health* (Vol. 92, Issue 11). https://doi.org/10.2105/ajph.92.11.1784

Weinstein, J. N., Geller, A., Negussie, Y., & Baciu, A. (2017). National Academies of Sciences, Engineering, and Medicine; Health and Medicine Division; Board on Population Health and Public Health Practice; Committee on Community-Based Solutions to Promote Health Equity in the United States. In *Communities in Action: Pathways to Health Equity*. National Academies Press. https://doi.org/10.17226/24624

WHO regional office for Europe. (2001). *REGIONAL OFFICE FOR EUROPE WHO REGIONAL STRATEGY ON SEXUAL AND REPRODUCTIVE HEALTH*. https://www.euro.who.int/\_\_data/assets/pdf\_file/0004/69529/e74558.pdf