**Discussion**

In this study, we examined the pattern of variable importance for each State and each local school district with the outcome - STI testing other than HIV using the best-performed ML method trained on the whole dataset of State and local school districts. We 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 State and local school districts dataset. Based on the results and plot of “resamples,” we were able to know that the best-performed algorithm for State data is ridge regression (mean accuracy = 0.7447), and the best-performed algorithm for local school district data is lasso regression (mean accuracy = 0.7079). Considering the nature of the training data, the lasso and ridge algorithms have the best performance due to their ability to handle multicollinearity, feature scaling, and interpretability. Based on the ridge regression trained on the whole State data, we were able to reveal the pattern of variable importance for each State, which was different. The lasso regression trained on the whole local school data also demonstrated the distinct pattern of variable importance for each local school district. However, most of the States and local school districts share variables q17 about physical fights, q21 about dating forcing sex, q41 about alcohol drinking, q47 about marijuana, q52 about heroin use, q53 about methamphetamine use, and q89 school grade in common. It is not surprising for us to see the distinct patterns of variable importance because we have observed inequities across the United States in multiple fields, such as healthcare quality, social economics, health status, disease prevalence, and healthcare accessibility(Holowatyj et al., 2020; Merkt et al., 2021; Rowley, 2022; Wang et al., 2021). To address the statistical issues that we encountered in applying the lasso and ridge regression algorithms to each State and local school district, we used the random forest method (3rd in accuracy, see Figure. 1 (a) and (b)) to analyze the variable importance for each State and local school district. The random forest algorithm demonstrated different results from the lasso and ridge regression. The top 3 important variables for the State and local school district data include BMI, q41, q47, and q89, which were about alcohol drinking, marijuana use, and school grade. BMI is the most important variable for all States and all local school districts except Shelby County (SC). Although the random forest algorithm had slightly lower accuracy when compared with lasso and ridge regression trained on the whole dataset, the random forest provided more meaningful results.

Machine learning algorithms are widely used in explaining, predicting, or creating risk scores for many diseases or events, such as cardiovascular disease risk prediction, HIV/STI testing clinic attendance, Covid-19 severity, and diabetic kidney diseases(Alaa et al., 2019; Chan et al., n.d.; Chen et al., 2021; Xu et al., 2022). Previous studies only compared the performance of a few ML algorithms and risk factors on data collected at national levels. And limited studies accessed the importance of risk factors associated with STI testing. Our study has the advantage of comparing the performance of multiple ML algorithms on States and local school district data and applying the two different ML algorithms to each State and local school district to ensure accurate results. In addition, we assessed the performance of ML methods using accuracy and trained the ML algorithms on two datasets to enhance the model selection, validation, and generalizability. These approaches were particularly beneficial when analyzing imbalanced datasets with limited sample sizes and complex feature types, which may be less accurate when analyzed using traditional statistical methods. By using the ML algorithms, we provide a comprehensive understanding of important variables associated with the outcome – Q85 STI testing other than HIV for healthcare, education, social work, and government facilities. This information can help public health workers enhance STI testing by addressing important risk factors and identifying vulnerable populations. Ultimately, our findings can serve as a valuable resource for healthcare professionals, educators, social workers, and policymakers to improve STI prevention and treatment efforts.

The study has a few limitations. First, the ML algorithms were trained on the entire dataset and then applied to a subset of the data. This approach can lead to bias in the results, as the selected ML algorithm might not perform optimally on the specific subset. To address this problem, two ML methods were applied to different subsets of the data. One reason for using this approach is that we have limited computer performance. By training and applying multiple algorithms on smaller subsets, we can reduce the computational resources required while also mitigating the potential for bias by using more tailored models for each subset. Second, training time can depend on sample size, algorithms, computer performance, and the complexity of the feature. Limited training time can restrict model selection and tuning options, potentially affecting the overall performance of the chosen models. Third, the cross-sectional study design lacks the ability to establish a causal relationship between the outcome – STI testing other than HIV and risk factors. The future study can include longitudinal studies to enhance the understanding of causality better. Fourth, the measurement of ML algorithms’ performance is based on a single performance metric – accuracy, which may affect the viability of the ML method in a given situation. Additional performance metrics, such as precision, recall, kappa, and F1 score, could provide a more comprehensive evaluation. Fifth, the accuracies for the selected ML algorithms were all around 0.7, which indicated a strong performance. However, there is still significant room for improvement before these models can be deemed appropriate for public health use. Enhancing the performance of these models is crucial for practical applications. Sixty, even though the selected ML model provided meaningful insight into the pattern of important risk factors associated with STI testing across States and local school districts around the United States, there is potential for further improvement by incorporating additional data features. By refining the ML models and including more relevant information, we can achieve better performance and a more comprehensive understanding of the factors affecting STI testing.

Above all, the study successfully demonstrated that BMI, physical fights, dating sexual violence, alcohol drinking, marijuana use, heroin use, methamphetamine use, and school grade were the most important risk factors associated with STI testing other than HIV. Public health professionals can improve STI testing by addressing key risk factors and identifying susceptible populations with the use of this knowledge. Finally, our findings can help to improve STI prevention and treatment efforts by healthcare professionals, educators, social workers, and politicians. Furthermore, we compared the performance of the multiple ML algorithms on two datasets and showed the advantage of ML algorithms in solving complex public health challenges. Future research can benefit from using longitudinal data and cohort study designs. As we move into the post-Covid-19 pandemic stage, public health faces increasing threats and challenges from various aspects. As a powerful tool, the ML approach can aid in addressing these threats and challenges, ultimately improving public health outcomes and promoting overall well-being.