Development and external validation of a web-based risk prediction tool using machine learning algorithms for an individual's risk of HIV and sexually transmitted infections

Xu X[1,2,3], Yu Z[2,4], Ge Z[4], Bao Y[3], Ong J[1,2,3], Li W[6], Wu J[7], Fairley C[1,2,3], Zhang L[1,2,3,8]

[1]Melbourne Sexual Health Centre, Carlton VIC, Australia, [2]Central Clinical School, Faculty of Medicine, Nursing and Health Sciences, Monash University, Melbourne VIC, Australia, [3]China Australia Joint Research Center for Infectious Diseases, School of Public Health, Xi'an Jiaotong University Health Science Centre, Xi'an Shaanxi, People's Republic of China, [4]Monash e-Research Centre, Faculty of Engineering, Airdoc Research, Nvidia AI Technology Research Centre, Monash University, Melbourne VIC, Australia, [5]Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Melbourne VIC, Australia, [6]School of public health, Southeast University, Nanjing Jiangsu, People's Republic of China, [7]Research Centre for Data Analytics and Cognition, La Trobe University, Bundoora VIC, Australia, [8]Department of Epidemiology and Biostatistics, College of Public Health, Zhengzhou University, Zhengzhou Henan, People's Republic of China

Background: HIV and sexually transmitted infections (STI) are major global public health concerns. Insufficient testing or delayed diagnosis substantially impedes the elimination of HIV/STI transmission. This study aimed to develop an HIV/STI risk prediction tool using machine learning algorithms.

Methods: We used clinic consultations where individuals who were tested for HIV/STI at the Melbourne Sexual Health Centre between March 2, 2015, to December 31, 2018, as the development dataset (training and testing dataset). We also used two external validation datasets, including data in 2019 as the external 'validation data 1' and data during January 2020 and January 2021 as the external 'validation data 2'. We developed 34 machine learning models to assess the risk of acquiring HIV, syphilis, gonorrhoea, and chlamydia. We created an online tool to generate an individual's risk of HIV/STI.

Results: Our ML-based risk prediction tool named MySTIRisk performed at an acceptable or excellent level on testing datasets (area under the curve (AUC) for HIV= 0.78; syphilis = 0.84; gonorrhoea = 0.78; chlamydia = 0.70) which had stable performance on both external validation data in 2019 (AUC for HIV= 0.79; syphilis = 0.85; gonorrhoea = 0.81; chlamydia = 0.69), and data in 2020-2021 (AUC for HIV= 0.71; syphilis= 0.84; gonorrhoea = 0.79; chlamydia = 0.69).

Conclusions: Our online risk prediction tool could accurately predict the risk of HIV/STI in clinic attendees with a simple self-administered questionnaire. MySTIRisk could serve as an HIV/STI screening tool on clinic websites or digital health platforms. The public can use this tool to assess their HIV/STI risk to inform testing. Clinicians or public health workers can use this tool to identify high-risk individuals for further interventions.

Disclosure of Interest Statement: The authors declare no competing interests.