MACHINE LEARNING METHODS AS A TOOL FOR PREDICTING RISK OF SYMPTOMS AND REPRODUCTIVE TRACT COMPLICATIONS OF CHLAMYDIA INFECTION

Authors:
Alexiou ZW[1,2], Hoenderboom BM[1,2], Hoebe CJPA[3,4,5,6], Ouburg S[7], van Benthem BH[1], Morré SA[2,6,8]

[1]Centre for Infectious Disease Control, National Institute for Public Health and the Environment, Bilthoven, The Netherlands, [2]Institute for Public Health Genomics (IPHG), GROW Research Institute for Oncology and Reproduction, Maastricht University, Maastricht, the Netherlands, [3]Department of Social Medicine, Care and Public Health Research Institute (CAPHRI), Maastricht University, Maastricht, The Netherlands, [4]Department Sexual Health, Infectious Diseases and Environmental Health, Public Health Service South Limburg, Heerlen, The Netherlands, [5]Department of Health Promotion, Care and Public Health Research Institute (CAPHRI), Maastricht University, The Netherlands, [6]Dutch Chlamydia trachomatis Reference Laboratory, Department of Medical Microbiology, Infectious Diseases and Infection Prevention, Care and Public Health Research Institute (CAPHRI), Maastricht University Medical Centre (MUMC+), Maastricht, The Netherlands, [7]Microbe&Lab, Amsterdam, The Netherlands, [8]Department of Molecular and Cellular Engineering, Jacob Institute of Biotechnology and Bioengineering, Sam Higginbottom University of Agriculture, Technology and Sciences, Allahabad, India

Background:
Identifying women at the highest risk for Chlamydia trachomatis (Ct) complications is crucial for effective disease management. Most women clear the infection, but in some, Ct ascends the upper genital tract, causing complications. Host genetic factors influence the immunological response; however, the complexity, and high dimensionality of genetic data make traditional regression models less suitable for risk prediction. To address this challenge, we explore the use of random forest decision trees (RFDT), a highly flexible artificial intelligence method, in a large-scale prospective cohort to assess its advantages over conventional prediction models.

Methods:
From women of reproductive age, DNA for single-nucleotide polymorphism (SNP) determination was extracted from buccal swabs, vaginal swabs, or urine and analyzed using kompetitive allele specific PCR sequencing. Ct status was determined by NAAT results, self-report, and antibodies. Outcomes included self-reported Ct with symptoms and complications (pelvic inflammatory disease, ectopic pregnancy, tubal infertility). Prediction models were built using RFDT and conventional multivariable logistic regression (MLR). Performance was assessed by area under the curve (AUC). Models were stratified by Ct status and included socio-demographics, sexual behavior, chlamydia/gonorrhea diagnoses, and 24 SNPs in candidate genes with three variants (wildtype, heterozygous, homozygous). Variable importance was calculated to identify key predictors.

Results:
In total 5094 women were included, 669 (13.2%) reported Ct infection(s) with symptoms and 299 (5.9%) complications. For symptomatic Ct infection RFDT

models showed higher performance (AUC 0.9) than MLR models (AUC 0.8). Different SNPs were identified as most important when comparing RFDT and MLR models. For complications RFDT models and MLR models showed poor prediction accuracy (AUC 0.6).

Conclusion:
Host genetic data, when analyzed using RFDT methods, may enhance the accuracy of predicting immediate host response to Ct infection. For long-term complications, it should be studied whether alternative machine learning algorithms better handle imbalanced data and rare outcomes.

Disclosure of Interest Statement:
Funding was received from the Netherlands Organisation for Health Research and Development (ZonMW Netherlands) and Research Funding from the Ministry of Health, Welfare and Sports.