INSTITUT PASTEUR

le cnam
Conservatoire national
des arts et métiers

PhD proposal entitled: **Applied machine learning to design serological surveillance tools for infectious diseases**

National and International context

There is growing emphasis on the use of Artificial Intelligence (AI) and machine learning applications in the health care domain. AI can produce a profound impact on diagnostics, disease prevention and treatment[1]. In 2018, the FDA (Food and Drug Administration) in the USA has approved the use of a novel AI based diagnostic tool for the early detection of diabetic retinopathy without the intervention of a specialist physician[2]. Despite a growing pace in the transition to AI healthcare, Europe's contribution to AI healthcare has been overshadowed by investments made by the US[3]. Despite boasting world-leading excellence in applied statistics and AI, France is not currently at the forefront of applying "AI for healthcare"[4].

Background

Conventional molecular diagnostic tests typically identify individuals with current infection. However, the biology of *P. vivax* imposes ***substantial diagnostic challenges***. This is because of the ability of the parasite to produce dormant liver forms, called hypnozoites that are undetectable by current diagnostic tools[5]. Hypnozoites in the liver can reactivate to induce regular bouts of blood stage infection, which are responsible for 80% of total blood stage infections[6]. The first relapse happens within nine months of the primary blood stage infection[7]. Thus, identifying persons infected during the past nine months should characterise persons harbouring dormant hypnozoites in their liver. Serological diagnostic tests can identify past exposure given the ability of antibodies to survive for years following an infection[8]. However, conventional serological tests comprise the use of a single biomarker for diagnostic purposes, which makes the validity of these tools highly questionable[9]. First, they are based on the traditional cut-off-based methods which dichotomizes the test measures into seropositive and seronegative. Second, they depend on the discriminatory ability of a single antigen. Individuals typically express different immunological responses, which makes the use of a single serological biomarker insufficient for diagnosing their infection status[10]. However, the measurement of multiple biomarkers can potentially provide a unique and conserved signature pattern for diagnosing infection status across individuals. Machine learning tools can directly tackle the shortcoming of existing serodiagnostic tools by providing a cut-off-free algorithm that allows the incorporation of multiple biomarkers.

The diagnostic innovations required for estimating time since previous *P. vivax* infection given measured antibody responses are also applicable to a range of other infectious diseases. Notably, the same methodology built on multiplex assays and machine learning algorithms can be applied for serological surveillance of SARS-CoV-2. Conventional SARS-CoV-2 serological tests provide a binary seropositivity status depending on whether or not a measured antibody response is greater than a defined cutoff. Using machine learning algorithms applied to data from multiple measured antibody responses it is possible to also provide an estimate of when an individual was previously infected, yielding valuable additional epidemiological information. With estimates of time since previous infection, it is possible to provide a serological reconstruction of past COVID-19 epidemics, estimating the timing and magnitude of past waves.

Serological surveillance systems for the early detection of individuals at risk is indispensable for targeted prevention programmes[11]. Thus, making available low-cost high accuracy point of care testing would address significant gaps in health surveillance. Serological testing (antibody detection) are easier to perform, provide faster results and are significantly less expensive compared to molecular and cultural testing[12]. However, serological tests are often of poor quality when they rely on the results of a single serological biomarker[13]. *AI and in particular machine learning can balance the trade-off between cost and quality*.

Objectives

The core objective of this PhD project is ***to develop algorithms for accurate serological estimation of time since infection*** with applications to *P. vivax* malaria and SARS-CoV-2. The *P. vivax* data is based on measurements of multiple antibody responses using data from Thailand, Brazil and Solomon Islands which represent countries with low endemic malaria transmission. The SARS-CoV-2 data is based on longitudinal follow-up of infected patients and healthcare workers in French hospitals. Secondary objectives will involve the translation of the developed methods to allow serological surveillance of other pathogens. The following objectives are proposed for a doctorate project:

1. To explore a range of machine learning algorithms and assess their performance in different geographical locations to estimate time since *P.* vivax infection and identify hypnozoite carriers.
2. To apply the machine learning predictive models to analyse antibody responses to SARS-CoV-2.
3. To integrate machine learning algorithms into the serological surveillance of other infectious diseases such as corona viruses and neglected tropical diseases (NTDs).

The majority of the methodological development will occur within the first two aims using epidemiologically detailed and well characterised datasets on malaria serology. These advances will rapidly feed into other diagnostic tools for NTDs being developed by the Infectious Disease Epidemiology & Analytics Unit.

Methodology

To answer these research questions, exhaustive ***supervised learning algorithms*** with a wide range of complexity will be evaluated and compared. In order to enhance and extend the utility of the algorithm to other countries, heterogeneous random effects for different geographic locations will be modelled using ***mixed effect machine learning* approaches**. Other sources of heterogeneities such as age and occupation will also be modelled. Simple decision trees and logistic regression classifiers to more sophisticated ensemble approaches such as super learners will be developed and compared. ***Extensive sensitivity*** *analysis* will also be conducted to better understand the impact of each feature on the model's prediction. The literacy of the *P. vivax* and coronavirus serological algorithms will be enhanced by producing ***explainable Artificial intelligence*** models (**X-AI**) targeting high classification performance of serological diagnostic[14-17]. Work accomplished during the master project will contribute to answering the first objective within the first year. The data include *three longitudinal cohorts* funded by the NIH's international Centres of Excellence for Malaria Research (ICEMR) and overseen by Institut Pasteur in three distinct geographical regions: Ethiopia, Cambodia, Papua New Guinea (PNG). Access will be provided to data from additional studies funded by the European Research Council (ERC) and overseen by the Infectious Disease Epidemiology & Analytics Unit. The data is expected to be collected in the upcoming months to include various species of Malaria, coronaviruses, NTDs and other parasitic diseases in Senegal.

PhD supervisors

Mounia N. HOCINE, PhD HDR, will act as the director of the doctoral project, and Michael White, PhD, will act as co-director. Mounia N. HOCINE, associate professor in biostatistics at laboratoire MESuRS at Cnam will provide extensive machine learning advisory and support in developing surveillance algorithms. She demonstrated experience in supervising PhD students using machine learning tools for tackling public health issues[17.] Michael White, a renowned researcher at Institut Pasteur, team "Infectious Disease Epidemiology & Analytics" Unit, will bring his international expertise in epidemiology and mathematical modelling of infectious diseases into the project.

PhD student skills

High skills in statistics and programming. An experience in health data science will be appreciated.

Timetable

1st year	- ***Apply for a PhD funding* over the period march-June 2021** - Literature review on machine learning and XAI tools for diagnosis & surveillance
2nd year	- Consolidating the new database collected in 2020-2021. - Develop ML predictive algorithms for detecting recent exposure to malaria species
3rd year	- Extend the ML algorithm to other infectious diseases such as covid-19. - An R Shiny App will be delivered for the local on-site use of PvSeroTAT

Completion of these objectives of the doctorate project will result in 2-3 papers published in peer-reviewed journals within three years.

References

1. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare [Internet]. Future healthcare journal. Royal College of Physicians; 2019
2. Padhy SK, Takkar B, Chawla R, Kumar A. Artificial intelligence in diabetic retinopathy: A natural step to the future [Internet]. Indian journal of ophthalmology. Wolters Kluwer - Medknow; 2019
3. Körner K. (How) will the EU become an AI superstar?Germany: Deutsche Bank Research 2020
4. Laï M-C, Brian M, Mamzer M-F. Perceptions of artificial intelligence in healthcare: findings from a qualitative survey study among actors in France. J Translational Medicine 2020; 18(1).
5. Hulden L, Hulden L. Activation of the hypnozoite: a part of Plasmodium vivax life cycle and survival. Malaria Journal. 2011;10(1).
6. Battle KE, Lucas TCD, Nguyen M, Howes RE, Nandi AK, Twohig KA, et al. Mapping the global endemicity and clinical burden of Plasmodium vivax, 2000–17: a spatial and temporal modelling study. The Lancet. 2019;394(10195):332–43.
7. White MT, Shirreff G, Karl S, Ghani AC, Mueller I. Variation in relapse frequency and the transmission potential of Plasmodium vivax malaria. Proceedings of the Royal Society B: Biological Sciences. 2016;283(1827):20160048.
8. Wipasa J, Suphavilai C, Okell LC, Cook J, Corran PH, Thaikla K, et al. Long-Lived Antibody and B Cell Memory Responses to the Human Malaria Parasites, Plasmodium falciparum and Plasmodium vivax. PLoS Pathogens. 2010;6(2).
9. Weber GE, White MT, Babakhanyan A, Sumba PO, Vulule J, Ely D, et al. Sero-catalytic and Antibody Acquisition Models to Estimate Differing Malaria Transmission Intensities in Western Kenya. Scientific Reports. 2017;7(1).
10. Moreno YR, Donato ST, Nogueira F, Silva MS. Comparative Analysis of the Serological Reactivity of Individuals with Clinical History of Malaria using Two Different ELISA Tests. Diagnostics. 2019;9(4):168.
11. Chapter 53. Public Health Surveillance: A Tool for Targeting and Monitoring Interventions. Disease Control Priorities in Developing Countries (2nd Edition). 2006Feb;:997–1016.
12. Dowdy DW, Steingart KR, Pai M. Serological Testing Versus Other Strategies for Diagnosis of Active Tuberculosis in India: A Cost-Effectiveness Analysis. PLoS Medicine. 2011Sep;8(8).
13. Longley RJ, White MT, Takashima E, Brewster J, Morita M, Harbers M, et al. Development and validation of serological markers for detecting recent Plasmodium vivax infection [Internet]. Nature News. Nature Publishing Group; 2020.
14. Miller T. Explanation in artificial intelligence: Insights from the social sciences [Internet]. Artificial Intelligence. Elsevier; 2018.
15. Duchemin T, Bar-Hen A, Lounissi R, Dab W, Hocine MN. Hierarchizing Determinants of Sick Leave: Insights From a Survey on Health and Well-being at the Workplace [Internet]. Journal of occupational and environmental medicine. U.S. National Library of Medicine; 2019.
16. Rosado J, Pelleau S, Cockram C, Merkling SH, Demeret C, et al. Multiplex assays for the identification of serological signatures of SARS-CoV-2 infection: an antibody-based diagnostic and machine learning study. Lancet Microbe. 2020; 2(2): E60-E69
17. Pelleau et al. Serological reconstruction of COVID-19 epidemics through analysis of antibody kinetics to SARS-CoV-2 proteins. medRxiv. 2021; doi: https://doi.org/10.1101/2021.03.04.21252532

Contact: Mounia N. Send your CV and motivation letter to: mounia.hocine@cnam.fr and michael.white@pasteur.fr