From bites to bytes: understanding how and why individual malaria risk varies using artificial intelligence and causal inference

Adèle Helena Ribeiro[1], Júlia M. P. Soler[2], Rodrigo M. Corder[3], Marcelo U. Ferreira[3,4]*† and Dominik Heider[1]*†

[1]Institute of Medical Informatics, University of Münster, Münster, Germany, [2]Institute of Mathematics and Statistics, University of São Paulo, São Paulo, Brazil, [3]Institute of Biomedical Sciences, University of São Paulo, São Paulo, Brazil, [4]Institute of Hygiene and Tropical Medicine and Global Health and Tropical Medicine Research Center, NOVA University of Lisbon, Lisbon, Portugal

With an estimated 263 million cases recorded worldwide in 2023, malaria remains a major global health challenge, particularly in tropical regions with limited healthcare access. Beyond its health impact, malaria disrupts education, economic development, and social equality. While traditional research has focused on biological factors underlying human-mosquito interactions, growing evidence highlights the complex interplay of environmental, behavioral, and socioeconomic factors, alongside mobility and both human and parasite genetics, in shaping transmission dynamics, recurrence patterns, and control effectiveness. This work shows how integrating Artificial Intelligence (AI), Machine Learning (ML), and Causal Inference can advance malaria research by identifying context-specific risk factors, uncovering causal mechanisms, and informing more effective, targeted interventions. Drawing on the Mâncio Lima cohort, a longitudinal, multimodal study of malaria risk in Brazil's main urban hotspot, and related studies in the Amazon, we highlight how rigorous, data-driven approaches can address the substantial variability in malaria risk across individuals and communities. AI-driven methods facilitate the integration of diverse high-dimensional datasets to uncover intricate patterns and improve individual risk stratification. Federated learning enables collaborative analysis across regions while preserving data privacy. Meanwhile, causal discovery and effect identification tools further strengthen these approaches by distinguishing genuine causal relationships from spurious associations. Together, these approaches offer a principled, scalable, and privacy-preserving framework that enables researchers to move beyond predictive modeling toward actionable causal insights. This shift supports precision public health strategies tailored to vulnerable populations, fostering more equitable and sustainable malaria control and contributing to the reduction of the global malaria burden.

KEYWORDS

artificial intelligence, causality, causal modelling, malaria, infectious diseases, public health

Introduction

Malaria remains a major health challenge, particularly in tropical and subtropical regions facing poverty, limited healthcare access, and harsh environments, such as the Amazon rainforest. In 2023, an estimated 263 million malaria cases occurred across 83 countries and territories – 37 million more than in 2015 (World Health Organization, 2024). Conflicts, humanitarian crises, climate change, drug and insecticide resistance, and resource constraints are among the threats to malaria control efforts.

P. falciparum predominates in sub-Saharan Africa, causing the most severe form of human malaria (Poespoprodjo et al., 2023). *P. vivax* is the most geographically widespread parasite, responsible for over 80% of infections in the Amazon and causing recurrent infections. Malaria's impact extends beyond health, disrupting education, hindering economic growth, straining healthcare systems, and perpetuating poverty. Effective control is crucial for public health, equity and global prosperity, requiring a shift from the traditional human-mosquito transmission model to a broader understanding of biological, environmental, and socioeconomic factors.

We take as an example the Mâncio Lima cohort study, which focuses on urban malaria in the Brazilian Amazon (Johansen et al., 2021). Approximately 20% of households in Mâncio Lima, Brazil's primary urban hotspot near the Peruvian border, were randomly selected from census data, resulting in 2,774 participants tested for malaria parasites during seven cross-sectional surveys (2018–2021) using conventional microscopy and highly sensitive, species-specific molecular techniques (Rodrigues et al., 2024). The study gathered data on demographics, health, housing conditions, occupation, lifestyle, and mobility, alongside blood samples for human genetics research, including genome-wide association studies.

Complementary longitudinal studies across Latin America have investigated the genomic diversity of *P. vivax* and *P. falciparum* (de Oliveira et al., 2020; Cabrera-Sosa et al., 2024; Kattenberg et al., 2024). Conducted in both urban and rural areas around Mâncio Lima (2018–2021) and the Peruvian Amazon (2007–2020), these studies support integrative genomic surveillance to track transmission intensity, imported cases, and drug resistance markers. By linking human and parasite data across diverse settings, these efforts support research on malaria dynamics and the evolution of key traits, such as virulence, resistance, and local adaptation, while accounting for ecological and socio-demographic variation.

The Mâncio Lima cohort has yielded several insights (Corder et al., 2019; Corder et al., 2020b; de Oliveira et al., 2020; Corder et al., 2023; Rodrigues et al., 2024). Of 11,730 samples screened using molecular methods, 4.0% were positive for *P. vivax* and 0.9% for *P. falciparum*, whereas standard microscopy detected much lower rates (0.4% for *P. vivax* and 0.2% for *P. falciparum*) (Rodrigues et al., 2024). Despite the low prevalence, *P. vivax* infections were recurrent (Corder et al., 2020a; Corder et al., 2023), with model simulations indicating that 20% of individuals at highest risk of infection accounted for 86% of the infection burden (Corder et al., 2020b). This highlights that malaria burden is often heterogeneously distributed within communities, following the 20/80 rule, where approximately 20% of individuals carry 80% of infections (Corder et al., 2023). Adult men face the highest risk (Corder et al., 2019), and most laboratory-confirmed infections were asymptomatic

(Rodrigues et al., 2024). Human mobility between urban and rural areas appears to sustain malaria transmission (Johansen et al., 2020). Additionally, genetic analyses of *P. vivax* revealed diverse, spatially and temporally structured lineages, highlighting heterogeneous transmission dynamics across different settings (de Oliveira et al., 2020; Kattenberg et al., 2024). In contrast, *P. falciparum* exhibited lower genetic diversity and stronger temporal clustering, indicating localized and time-limited transmission (Cabrera-Sosa et al., 2024).

Despite these heterogeneities, malaria prevalence in Mancio Lima declined significantly from 2018 to 2021, likely due to extensive control and treatment efforts, including widespread indoor residual spraying, distribution of insecticide-treated bed nets, active case testing, and free treatment programs. Sustaining and advancing this progress requires improved identification of high-risk groups for optimizing resource distribution and implementing tailored interventions. A key challenge is understanding why some individuals repeatedly contract *P. vivax* while others remain uninfected. Clinically, such recurrences can lead to severe complications, including anemia, particularly among vulnerable groups, such as children and pregnant women (Pincelli et al., 2021). Economically, this heterogeneity complicates policy design. The 20/80 rule suggests that targeting high-risk individuals could maximize impact (Corder et al., 2023). Additionally, malaria has emerged as a zoonotic threat. *P. simium*, a parasite of non-human primates, has caused infections in humans in southeastern Brazil, where *P. vivax* is rare (de Oliveira et al., 2021a; b). Distinguishing between human and zoonotic parasites is critical for evaluating interventions and preparing for future outbreaks.

To elucidate the multifaceted dynamics underlying malaria risk, we propose a synergistic integration of AI, ML, and causal inference. This combination enables not only the identification of high-risk groups but also the discovery of causal mechanisms driving individual variability in malaria susceptibility. By leveraging cutting-edge methods, we can move beyond predictive modeling toward causal understanding, thereby informing the development of optimized, targeted interventions. Our approach relies on the integration of high-dimensional, multimodal datasets such as those from the Mancio Lima cohort and other regional studies – including data on malaria episodes, clinical, behavioral, socioeconomic, environmental, and genetic factors. This rich data landscape enables the identification of structured patterns and interpretable representations that explain malaria risk and transmission dynamics. Causal inference methods that account for latent confounding and selection bias are essential to distinguish causal drivers from spurious associations, enabling robust estimation of intervention effects under real-world conditions. Ultimately, this framework will support precision public health by ensuring that prevention, control, and treatment strategies are both timely and tailored to those most at risk, maximizing impact and equity (Khoury et al., 2015).

Bridging AI and causality for targeted malaria interventions

AI and ML have driven significant advancements in medicine and public health (MacEachern and Forkert, 2021) due to their
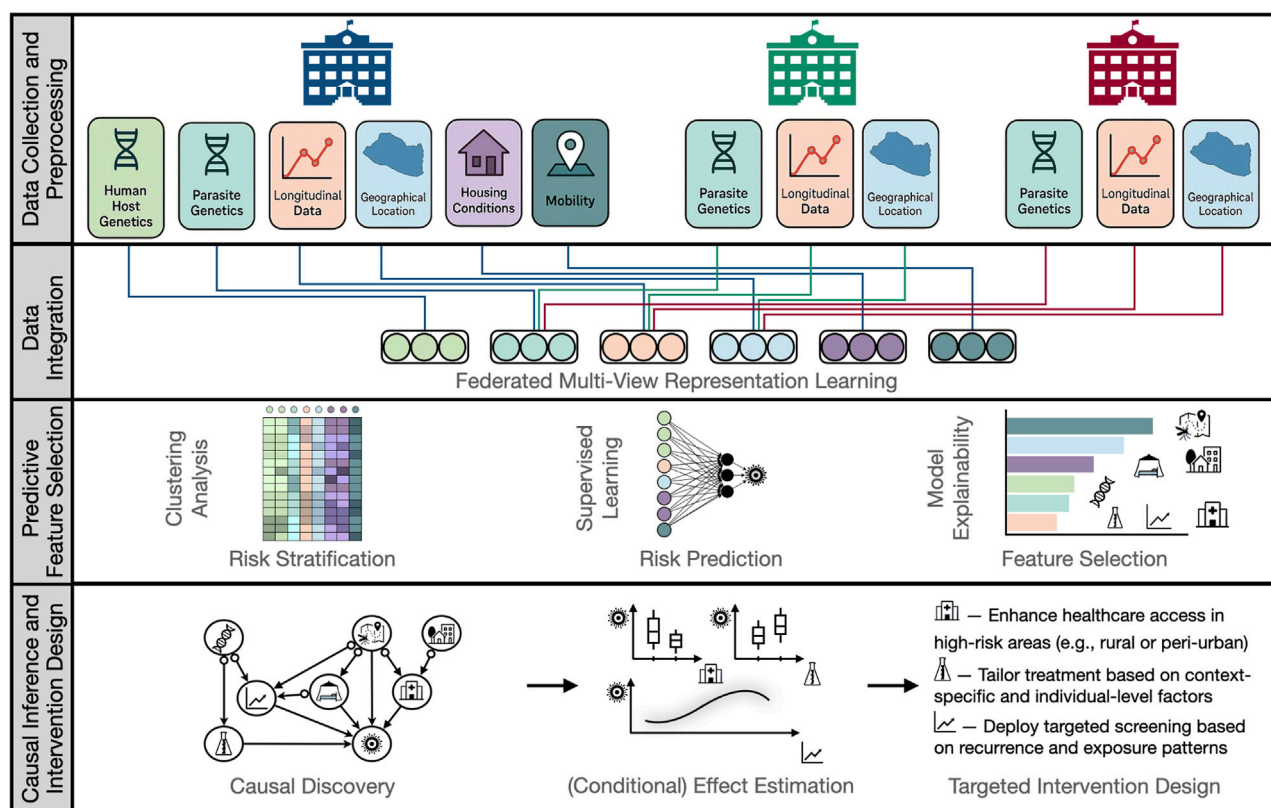
FIGURE 1
AI and Causal Inference Pipeline for Targeted Intervention Design in Malaria Research. The pipeline begins with Data Collection and Preprocessing, including anonymization, harmonization, and normalization of multimodal, multi-center data. In Data Integration, federated multi-view representation learning generates low-dimensional embeddings that capture both within- and cross-modal patterns while maintaining data privacy. Predictive Feature Selection uncovers latent risk profiles and selects interpretable features that predict malaria risk both globally and within specific subgroups. Finally, Causal Inference and Intervention Design applies causal discovery to reveal mechanisms underlying the selected features – e.g., treatment regimens, prior infection history, genetic predispositions, bed net usage, healthcare access, urban vs. rural residence, and proximity to mosquito breeding sites. Causal effect estimation tools then quantify the (conditional) impact of specific interventions (e.g., increasing healthcare access, personalizing treatments, or implementing targeted screening) from observational data, supporting precision public health strategies for effective malaria prevention, treatment, and control.

ability to model complex relationships and uncover subtle patterns in high-dimensional, heterogeneous datasets. These methods have been successfully applied across various medical domains (Theodosiou and Read, 2023), including infectious disease research, such as AMR prediction (Ren et al., 2022), zoonotic disease detection (Ren et al., 2024), and biomarker discovery in malaria (Jung et al., 2023).

In malaria research, AI and ML provide powerful tools to disentangle complex, often hidden dependency structures and enable precise individual risk stratification. The pipeline (Figure 1) begins with data collection and preprocessing, crucial for multimodal, heterogeneous, and sensitive data such as genomic and socio-behavioral information. Ensuring data privacy and quality through anonymization, harmonization, imputation, and normalization – while following FAIR principles (Findable, Accessible, Interoperable, Reusable) (Kush et al., 2020) – is essential for robust model development. In multi-center studies, federated learning supports privacy-preserving collaboration by enabling joint analysis without exchanging raw data (McMahan et al., 2017; Tajabadi et al., 2023; Tajabadi et al., 2024).

Multi-view representation learning approaches, such as multimodal variational autoencoders, enable data integration by generating low-dimensional latent embeddings that retain modality-specific features while capturing cross-modal dependencies (Guo et al., 2019). Clustering these embeddings can reveal subgroups of individuals with shared but not directly observed risk profiles, shaped by common exposures or susceptibilities (Jaeger et al., 2023). This step can be enriched through co-clustering, which jointly identifies groups of individuals and co-varying variables, highlighting context-specific drivers of malaria vulnerability (Govaert and Nadif, 2013). Moreover, federated representation learning and clustering (Zhang et al., 2023; Pedrycz, 2021) support robust and generalizable predictions across distributed, heterogeneous datasets. To enhance interpretability and inform downstream modeling, cluster-aware feature selection (Wang and Allen, 2021) identifies both globally predictive variables and those particularly informative within specific subgroups. These selected features and representations are then used to predict individual malaria risk, forming a cohesive and interpretable AI-driven framework for risk assessment.

While essential, high predictive accuracy alone is not sufficient to uncover the underlying data-generating mechanisms or support meaningful, actionable interventions. This is particularly true in biomedical and epidemiological research, where data are largely observational and vulnerable to multiple sources of bias. In malaria research, for example, unmeasured factors such as socio-economic status, mobility patterns, or environmental exposures can confound associations between risk factors and outcomes. Selection bias is also widespread due to underreporting, especially in remote regions or among asymptomatic individuals. If not properly addressed, these biases can reinforce existing health disparities and lead to interventions that are ineffective or even harmful.

Causal inference provides a principled framework to uncover cause-and-effect relationships and mitigate the impact of bias in observational studies (Pearl, 2009). It enables the estimation of the effect of interventions with a level of rigor comparable to randomized controlled trials. Several approaches exist, including the Potential Outcomes Framework (Rubin, 1974), Causal Machine Learning (van der Laan and Rubin, 2006; Feuerriegel et al., 2024), and Instrumental Variables (Angrist et al., 1996), also known in genetics as Mendelian Randomization (Haycock et al., 2016; Ribeiro et al., 2016). However, these frameworks rely on strong, sometimes unverifiable assumptions – such as the absence of latent confounding or availability of valid instruments – which are often violated in real-world settings.

In response, data-driven causal discovery methods within Pearl's framework have emerged as robust alternatives. Algorithms such as Fast Causal Inference (FCI) (Zhang, 2008) and its variants can recover causal structures directly from observational data, even in the presence of unmeasured confounding and selection bias. Notably, AnchorFCI (Ribeiro et al., 2024) enhances robustness and discovery power by strategically selecting and integrating reliable anchor variables – such as genetic variants – that are known not to be influenced by the variables of interest (e.g., clinical or sociodemographic factors). These methods infer a Partial Ancestral Graph (PAG) representing causal relationships shared across all models supported by the data, thus revealing the true data-generating processes. This enables the identification of key factors – e.g., use of insecticide-treated bed nets, housing conditions, or access to healthcare – that causally influence malaria risk and can be targeted by interventions. By applying causal effect identification algorithms to the resulting PAG, we can then quantify the isolated or combined impact of specific interventions, based solely on observational data (Perković et al., 2018; Jaber et al., 2022). This fully data-driven causal pipeline supports the development of more robust, transparent, and socially responsible interventions, providing a clearer pathway for addressing malaria risk in diverse settings.

A key strength of constraint-based causal discovery approaches such as FCI and its variants lies in their flexibility to account for mixed-type variables and complex dependency structures by adapting conditional independence tests. This is particularly important for analyzing malaria datasets, which typically comprise a mix of continuous, ordinal, categorical, and count variables, along with non-independent observations arising from genetic relatedness, repeated measures, household clustering, and spatial correlations. Conditional independence tests that account for such complexities can be constructed using generalized mixed models, which incorporate structured covariance and random effects to model known or inferred dependencies (Ribeiro and Soler, 2020). These tests can also be extended to federated learning settings, enabling collaborative, privacy-preserving causal discovery. Furthermore, causal discovery at the level of variable clusters – either predefined or learned through representation learning and clustering – can yield more interpretable insights into the interactions among biological, behavioral, and environmental risk factors for malaria (Anand et al., 2023).

Discussion

Progress toward malaria elimination in regions such as the Amazon requires a deep understanding of the intricate factors driving infection risk and recurrence. The Mâncio Lima cohort and regional studies offer a unique opportunity to uncover malaria dynamics by combining comprehensive data on human hosts, parasites, and their environments. However, the inherent complexity and heterogeneity of these datasets demand analytical frameworks that extend beyond traditional epidemiological or statistical approaches.

By integrating AI, ML, and causal inference, we move toward a more holistic strategy that not only accurately identifies high-risk individuals but also elucidates the causal mechanisms underlying malaria transmission and infection. This shift from descriptive and predictive modeling to causal reasoning enables the development of optimized, targeted interventions and lays the foundation for precision public health strategies that are not only more effective but also more equitable. Federated learning further supports this approach by enabling collaborative analysis across diverse regions without compromising data privacy. Together, these methodologies empower local health systems to respond more precisely and efficiently and contribute meaningfully to global control efforts.

Author contributions

AR: Formal Analysis, Writing – original draft, Methodology, Investigation, Conceptualization, Writing – review and editing. JS: Conceptualization, Investigation, Resources, Writing – review and editing. RC: Resources, Conceptualization, Investigation, Methodology, Writing – review and editing, Formal Analysis, Writing – original draft, Data curation. MF: Resources, Data curation, Supervision, Methodology, Investigation, Conceptualization, Writing – original draft, Funding acquisition, Writing – review and editing. DH: Writing – review and editing, Project administration, Conceptualization, Supervision, Writing – original draft, Funding acquisition, Investigation, Visualization.

Funding

2020 and LA-REAL LA/P/0117/2020). The Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Brazil, provides a senior research scholarship to MF We acknowledge support from the Open Access Publication Fund of the University of Münster.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Generative AI statement

The authors declare that no Generative AI was used in the creation of this manuscript.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Anand, T. V., Ribeiro, A. H., Tian, J., and Bareinboim, E. (2023). Causal effect identification in cluster DAGs. *Proc. AAAI Conf. Artif. Intell.* 37 (10), 12172–12179. doi:10.1609/aaai.v37i10.26435

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *J. Am. Stat. Assoc.* 91 (434), 444–455. doi:10.2307/2291629

Bareinboim, E., Correa, J. D., Ibeling, D., and Icard, T. (2022). "On Pearl's hierarchy and the foundations of causal inference," in *Probabilistic and causal inference: the works of Judea Pearl*, 507–556.

Cabrera-Sosa, L., Nolasco, O., Kattenberg, J. H., Fernandez-Miñope, C., Valdivia, H. O., Barazorda, K., et al. (2024). Genomic surveillance of malaria parasites in an indigenous community in the Peruvian Amazon. *Sci. Rep.* 14, 16291. doi:10.1038/s41598-024-66925-x

Corder, R. M., Arez, A. P., and Ferreira, M. U. (2023). Individual variation in *Plasmodium vivax* malaria risk: are repeatedly infected people just unlucky? *PLOS Neglected Trop. Dis.* 17 (1), e0011020. doi:10.1371/journal.pntd.0011020

Corder, R. M., de Lima, A. C. P., Khoury, D. S., Docken, S. S., Davenport, M. P., and Ferreira, M. U. (2020a). Quantifying and preventing *Plasmodium vivax* recurrences in primaquine-untreated pregnant women: an observational and modeling study in Brazil. *PLOS Neglected Trop. Dis.* 14 (7), e0008526. doi:10.1371/journal.pntd.0008526

Corder, R. M., Ferreira, M. U., and Gomes, M. G. M. (2020b). Modelling the epidemiology of residual *Plasmodium vivax* malaria in a heterogeneous host population: a case study in the Amazon Basin. *PLOS Comput. Biol.* 16 (3), e1007377. doi:10.1371/journal.pcbi.1007377

Corder, R. M., Paula, G. A., Pincelli, A., and Ferreira, M. U. (2019). Statistical modeling of surveillance data to identify correlates of urban malaria risk: a population-based study in the Amazon Basin. *PLOS ONE* 14 (8), e0220980. doi:10.1371/journal.pone.0220980

de Oliveira, T. C., Corder, R. M., Early, A., Rodrigues, P. T., Ladeia-Andrade, S., Alves, J. M. P., et al. (2020). Population genomics reveals the expansion of highly inbred *Plasmodium vivax* lineages in the main malaria hotspot of Brazil. *PLOS Neglected Trop. Dis.* 14 (10), e0008808. doi:10.1371/journal.pntd.0008808

de Oliveira, T. C., Rodrigues, P. T., Duarte, A. M. R. C., Rona, L. D. P., and Ferreira, M. U. (2021a). Ongoing host-shift speciation in *Plasmodium simium*. *Trends Parasitol.* 37 (11), 940–942. doi:10.1016/j.pt.2021.08.005

de Oliveira, T. C., Rodrigues, P. T., Early, A. M., Duarte, A. M. R. C., Buery, J. C., Bueno, M. G., et al. (2021b). *Plasmodium simium*: population genomics reveals the origin of a reverse zoonosis. *J. Infect. Dis.* 224 (11), 1950–1961. doi:10.1093/infdis/jiab214

Feuerriegel, S., Frauen, D., Melnychuk, V., Schweisthal, J., Hess, K., Curth, A., et al. (2024). Causal machine learning for predicting treatment outcomes. *Nat. Med.* 30, 958–968. doi:10.1038/s41591-024-02902-1

Govaert, G., and Nadif, M. (2013). *Co-clustering: models, algorithms and applications*. John Wiley and Sons.

Guo, W., Wang, J., and Wang, S. (2019). Deep multimodal representation learning: a survey. *IEEE Access* 7, 63373–63394. doi:10.1109/access.2019.2916887

Haycock, P. C., Burgess, S., Wade, K. H., Bowden, J., Relton, C., and Davey Smith, G. (2016). Best (but oft-forgotten) practices: the design, analysis, and interpretation of Mendelian randomization studies. *Am. J. Clin. Nutr.* 103 (4), 965–978. doi:10.3945/ajcn.115.118216

Jaber, A., Ribeiro, A. H., Zhang, J., and Bareinboim, E. (2022). Causal identification under Markov equivalence: calculus, algorithm, and completeness. *Adv. Neural Inf. Process. Syst.* 35, 3679–3690.

Jaeger, A., and Banks, D. (2023). Cluster analysis: a modern statistical review. *Wiley Interdiscip. Rev. Comput. Stat.* 15 (3), e1597. doi:10.1002/wics.1597

Johansen, I. C., Rodrigues, P. T., and Ferreira, M. U. (2020). Human mobility and urban malaria risk in the main transmission hotspot of Amazonian Brazil. *PLOS ONE* 15 (11), e0242357. doi:10.1371/journal.pone.0242357

Johansen, I. C., Rodrigues, P. T., Tonini, J., Vinetz, J., Castro, M. C., and Ferreira, M. U. (2021). Cohort profile: the Mâncio Lima cohort study of urban malaria in Amazonian Brazil. *BMJ Open* 11 (11), e048073. doi:10.1136/bmjopen-2020-048073

Jung, A. L., Møller Jørgensen, M., Bæk, R., Artho, M., Griss, K., et al. (2023). Surface proteome of plasma extracellular vesicles as mechanistic and clinical biomarkers for malaria. *Infection* 51 (5), 1491–1501. doi:10.1007/s15010-023-02022-x

Kattenberg, J. H., Monsieurs, P., De Meyer, J., De Meulenaere, K., Sauve, E., de Oliveira, T. C., et al. (2024). Population genomic evidence of structured and connected *Plasmodium vivax* populations under host selection in Latin America. *Ecol. Evol.* 14, e11103. doi:10.1002/ece3.11103

Khoury, M. J., Iademarco, M. F., and Riley, W. T. (2015). Precision public health for the era of precision medicine. *Am. J. Prev. Med.* 50 (3), 398–401. doi:10.1016/j.amepre.2015.08.031

Kush, R. D., Warzel, D., Kush, M. A., Sherman, A., Navarro, E. A., Fitzmartin, R., et al. (2020). FAIR data sharing: the roles of common data elements and harmonization. *J. Biomed. Inf.* 107, 103421. doi:10.1016/j.jbi.2020.103421

MacEachern, S. J., and Forkert, N. D. (2021). Machine learning for precision medicine. *Genome* 64 (4), 416–425. doi:10.1139/gen-2020-0131

McMahan, B., Moore, E., Ramage, D., Hampson, S., and Arcas, B. A. y. (2017). "Communication-efficient learning of deep networks from decentralized data," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics* (Proceedings of Machine Learning Research), 1273–1282. Available online at: https://proceedings.mlr.press/v54/mcmahan17a.html.

Pearl, J. (2009). *Causality: models, reasoning, and inference*. 2nd ed. Cambridge University Press.

Pedrycz, W. (2021). Federated FCM: clustering under privacy requirements. *IEEE Trans. Fuzzy Syst.* 30 (8), 3384–3388. doi:10.1109/tfuzz.2021.3105193

Perković, E., Textor, J., Kalisch, M., and Maathuis, M. H. (2018). Complete graphical characterization and construction of adjustment sets in Markov equivalence classes of ancestral graphs. *J. Mach. Learn. Res.* 18, 1–62.

Pincelli, A., Cardoso, M. A., Malta, M. B., Johansen, I. C., Corder, R. M., Nicolete, V. C., et al. (2021). Low-level *Plasmodium vivax* exposure, maternal antibodies, and anemia in early childhood: population-based birth cohort study in Amazonian Brazil. *PLOS Neglected Trop. Dis.* 15 (7), e0009568. doi:10.1371/journal.pntd.0009568

Poespoprodjo, J. R., Douglas, N. M., Ansong, D., Kho, S., and Anstey, N. M. (2023). Malaria. *Lancet* 402 (10419), 2328–2345. doi:10.1016/S0140-6736(23)01249-7

Ren, Y., Chakraborty, T., Doijad, S., Falgenhauer, L., Falgenhauer, J., Goesmann, A., et al. (2022). Prediction of antimicrobial resistance based on whole-genome sequencing and machine learning. *Bioinformatics* 38 (2), 325–334. doi:10.1093/bioinformatics/btab681

Ren, Y., Li, C., Nanayakkara Sapugahawatte, D., Zhu, C., Spänig, S., Jamrozy, D., et al. (2024). Predicting hosts and cross-species transmission of *Streptococcus agalactiae* by interpretable machine learning. *Comput. Biol. Med.* 171, 108185. doi:10.1016/j.compbiomed.2024.108185

Ribeiro, A. H., Crnkovic, M., Pereira, J. L., Fisberg, R. M., Sarti, F. M., Rogero, M. M., et al. (2024). AnchorFCI: harnessing genetic anchors for enhanced causal discovery of cardiometabolic disease pathways. *Front. Genet.* 15, 1436947. doi:10.3389/fgene.2024.1436947

Ribeiro, A. H., and Soler, J. M. P. (2020). Learning genetic and environmental graphical models from family data. *Statistics Med.* 39 (15), 2403–2422. doi:10.1002/sim. 8545

Ribeiro, A. H., Soler, J. M. P., Neto, E. C., and Fujita, A. (2016). "Causal inference and structure learning of genotype–phenotype networks using genetic variation," in *Big data analytics in genomics* (Springer), 89–143.

Rodrigues, P. T., Johansen, I. C., Ladeia, W. A., Esquivel, F. D., Corder, R. M., Tonini, J., et al. (2024). Lower microscopy sensitivity with decreasing malaria prevalence in the urban Amazon region, Brazil, 2018–2021. *Emerg. Infect. Dis.* 30 (9), 1884–1894. doi:10.3201/eid3009.240378

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* 66 (5), 688–701. doi:10.1037/h0037350

Tajabadi, M., Grabenhenrich, L., Ribeiro, A., Leyer, M., and Heider, D. (2023). Sharing data with shared benefits: Artificial intelligence perspective. *J. Med. Internet Res.* 25, e47540. doi:10.2196/47540

Tajabadi, M., Martin, R., and Heider, D. (2024). Privacy-preserving decentralized learning methods for biomedical applications. *Comput. Struct. Biotechnol. J.* 23, 3281–3287. doi:10.1016/j.csbj.2024.08.024

Theodosiou, A. A., and Read, R. C. (2023). Artificial intelligence, machine learning and deep learning: potential resources for the infection clinician. *J. Infect.* 87 (4), 287–294. doi:10.1016/j.jinf.2023.07.006

van der Laan, M. J., and Rubin, D. (2006). Targeted maximum likelihood learning. *Int. J. Biostat.* 2 (1). Article 11. doi:10.2202/1557-4679.1043

Wang, M., and Allen, G. I. (2021). Integrative generalized convex clustering optimization and feature selection for mixed multi-view data. *J. Mach. Learn. Res.* 22 (55), 55–73.

World Health Organization (2024). *World malaria report 2024. World Health Organization. Available online at: https://www.who.int/teams/global-malaria-programme/reports/world-malaria-report-2024.*

Zhang, F., Kuang, K., Chen, L., You, Z., Shen, T., Xiao, J., et al. (2023). Federated unsupervised representation learning. *Front. Inf. Technol. and Electron. Eng.* 24 (8), 1181–1193. doi:10.1631/FITEE.2200268

Zhang, J. (2008). On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artif. Intell.* 172 (16–17), 1873–1896. doi:10.1016/j.artint.2008.08.001