Análise Exploratória da Malária na Amazônia Brasileira por Meio da Plataforma de Ciência de Dados aplicada à Saúde*

Lais Ribeiro Baroni[1], Balthazar Paixão[1], Alvaro Chrispino[1], Gustavo Guedes[1]
Christovam Barcellos[2], Marcel Pedroso[2], Eduardo Ogasawara[1]

[1]CEFET/RJ - Centro Federal de Educação Tecnológica Celso Suckow da Fonseca

[2]Fiocruz - Fundação Oswaldo Cruz | Ministério da Saúde

{lais.baroni,balthazar.paixao}@eic.cefet-rj.br

{alvaro.chrispino,gustavo.guedes}@cefet-rj.br

{xris,marcel.pedroso}@icict.fiocruz.br, eogasawara@ieee.org

Abstract. *Malaria is an infectious disease that mainly affects the Legal Amazon. DATASUS includes the Malaria Epidemiological Surveillance Information System. Monitoring this dataset and integrating it with additional data sources, as well as performing proper data preprocessing is crucial to understand the phenomena behind the occurrences and medical care. Therefore, in this paper we make use of the Data Science Platform Applied to Health (PCDaS) as an enabling tool to analyze the evolution of malaria in the Legal Amazon. From its use, we raised research questions that can help in understanding and controlling this disease in Brazil.*

Resumo. *A malária é uma doença infecciosa que atinge principalmente a Amazônia Legal. O Departamento de Informática do Sistema Único de Saúde (DATASUS) hospeda e disponibiliza o Sistema de Informações de Vigilância Epidemiológica da Malária. Acompanhá-lo e integrar seus dados com fontes adicionais, bem como realizar a preparação dos dados é de vital importância para se compreender os fenômenos por trás das ocorrências e dos atendimentos médicos por meio das notificações realizadas no sistema. Para tanto, neste trabalho fazemos uso da Plataforma de Ciência de Dados aplicada à Saúde (PCDaS) como ferramenta para viabilizar a análise da evolução da malária na Amazônia Legal. A partir do seu uso, levantamos perguntas de pesquisas que podem ajudar na compreensão e no combate da malária no Brasil.*

1. Introdução

A malária é uma doença infecciosa causada por parasitas protozoários do gênero Plasmódio (*Plasmodium*) e é transmitida predominantemente a partir da picada do mosquito do gênero *Anopheles*, quando este já está infectado. Os países tropicais e subtropicais constituem a área endêmica da doença por terem estações chuvosas que proporcionam grande disponibilidade de água limpa parada, onde os mosquitos vetores podem depositar seus ovos e se proliferar [WHO, 2018].

*Os autores agradecem à FAPERJ, à CAPES e ao CNPq pelo financiamento parcial do projeto.

A Amazônia Legal - que compreende os estados do Acre, Amapá, Amazonas, Mato Grosso, Pará, Rondônia, Roraima, Tocantins e parte do estado do Maranhão - é a região mais susceptível à malária no Brasil [Recht et al., 2017]. Dentro da Amazônia Legal, a ocorrência da doença não é homogênea, variando de localidade a localidade de acordo com algumas características como fatores naturais, fatores geográficos e condições sociais [Confalonieri et al., 2014].

Embora a taxa de incidência de malária no Brasil não seja tão alta como em alguns países africanos, é importante que esforços sejam feitos para o seu combate no Brasil [Guerin et al., 2002; Griffing et al., 2015]. Desde do início de 2018, a Organização Pan-Americana da Saúde (OPAS) já alertava quanto a um aumento de casos de malária na região das Américas, inclusive no Brasil, e alertou quanto ao risco iminente de surtos. O último Relatório Mundial da Malária, da Organização Mundial de Saúde, confirmou a afirmativa, mostrando um aumento de 25% de casos de malária com relação ao ano anterior [WHO, 2018]. Um dos motivos atribuídos a esse aumento é a negligenciação nos cuidados contra a malária.

Uma das iniciativas criadas para apoiar no planejamento, monitoramento e avaliação de políticas públicas e serviços de saúde é a Plataforma de Ciência de Dados aplicada à Saúde Plataforma de Ciência de Dados aplicada à Saúde (PCDaS), desenvolvida pelo Laboratório de Informação em Saúde (LIS), do Instituto de Comunicação e Informação Científica e Tecnologia em Saúde (ICICT), da Fundação Oswaldo Cruz (Fiocruz). A Plataforma, desenvolvida em parceria com o Laboratório Nacional de Computação Científica (LNCC), contempla um conjunto de estratégias, ferramentas e técnicas para coleta, transformação e análise de grandes quantidades de dados (*big-data*) voltados para saúde pública. Ela foi concebida como um serviço público, de modo a viabilizar a análise, o monitoramento, a predição de eventos (casos) e as situações de saúde e doença na população, bem como a associação destes com seus determinantes sociais. O serviço é disponibilizado dentro do próprio site onde qualquer pessoa pode se cadastrar preenchendo uma série de dados pessoais como nome completo, CPF e e-mail. Após o cadastro o usuário tem três opções de acesso: Analise Visual, que disponibiliza através de gráficos interativo os dados já tratados pela equipe da Plataforma; Mineração de Dados e Análise Preditiva, que disponibiliza o R Studio Server para o usuário manipular e trabalhar com os dados disponíveis; e Data Science Lab, onde ele pode conhecer os projetos dos parceiros da Plataforma. O serviço disponibilizado é gratuito e aberto para qualquer pessoa, com a única condição que o código usado seja disponibilizado – mantendo o ideal de ciência aberta para todos.

Neste diapasão, realizar uma análise holística dos dados de malária na Amazônia Legal a partir dos recursos disponíveis na PCDaS pode ajudar no controle da doença. A proposta neste trabalho é de conduzir uma análise exploratória preliminar direcionada para o Sistema de Informações de Vigilância Epidemiológica módulo malária (SIVEP-Malária) entre os anos de 2009 a 2015. A partir dessa análise, levantaram-se perguntas de pesquisas que podem ajudar na compreensão e no combate da malária no Brasil.

2. Trabalhos Relacionados

Muitos autores fazem uso de técnicas de análise exploratória de dados no contexto do estudo da malária. Wiefels et al. [2016] estudaram os dados do SIVEP-Malária no estado do Amazonas entre 2003 a 2014 para avaliar sua qualidade e precisão. Concluíram que existem muitos dados faltantes, dados discrepantes e inconsistências, principalmente naqueles relacionados a informações de pacientes.

Loucoubar et al. [2011] usaram a ferramenta *HyperCube* para analisar um *dataset* de episódios clínicos de malária por *Plasmodium falciparum*. Os autores detalharam a ferramenta de mineração de dados visando encontrar diferentes combinações de variáveis explicativas que afetam a ocorrência da malária clínica. Como um dos resultados, apoiaram a hipótese de que existe imunidade protetora significativa entre as espécies da doença.

Diallo et al. [2017] fizeram uma análise exploratória a partir de amostragem de sangue de crianças em quatro localidades no continente Africano, sendo duas com alta endemicidade de malária em Burkina Faso e duas com baixa endemicidade no Senegal. Os autores fizeram considerações sobre a variabilidade na prevalência de espécies de plasmódio, o número de crianças afetadas por faixa etária e também sobre as ocorrências, considerando a utilização de medidas de controle da doença.

Sweeney et al. [2007] usaram o GARP (algoritmo genético para previsão de conjunto de regras) sobre os dados ambientais e de presença ou ausência de mosquitos vetores da malária, levantados em campo, no norte da Austrália. Eles identificaram a umidade atmosférica como fator crítico na sobrevivência de mosquitos adultos. Analogamente, Labbo et al. [2016] fizeram uma análise exploratória da associação entre a ocorrência de mosquitos vetores urbanos da malária e de fatores ambientais. Usaram Análise de Componentes Principais e teste não-paramétrico de Kruskall-Wallis para confirmação das análises. Observaram que mosquitos vetores da malária são mais propensos a serem encontrados a montante de rios e que são altamente produtivos em lagoas. Ainda nessa mesma linha, Sahle and Meshesha [2014] procuraram inferir a relação entre fatores ambientais e a ocorrência de malária, além das possíveis causas de morte causadas pela doença na Etiópia. A metodologia aplicada foi a mineração de dados para criação de classificadores segundo três diferentes algoritmos (árvore de decisão J48, indução de regra JRip e Rede Neural Multilayer Perceptron (MLP)). Entre as considerações, apontaram a chuva como principal fator para a prevalência da malária e uma probabilidade aumentada de risco de morte para crianças menores de 5 anos de idade.

Johansson et al. [2016] usaram árvores de decisão para minerar dados clínicos de unidades de saúde de Malawi em 2013-2014 com o objetivo de estimar a correlação entre a prescrição de antibióticos e o resultado de testes de malária. A hipótese estudada era de que, ao ser medicado para os sintomas da malária com antibióticos, os resultados dos exames para a malária eram mascarados, podendo ser negativos quando os pacientes, na verdade, possuíam a enfermidade. Evidenciaram, por fim, a importância do uso racional de medicamentos antimaláricos e antibióticos, objetivando o compromisso no combate à resistência a doença.

Tendo em vista a pesquisa bibliográfica feita e apresentada, nosso trabalho se destaca ao fazer uso dos serviços de computação científica da PCDaS para realizar um

estudo do comportamento de variáveis clínicas brasileiras de malária, incluindo dados pessoais integrados a demografia e outras informações do Departamento de Informática do Sistema de Saúde (DATASUS) pouco exploradas na literatura.

3. Fonte de Dados

As fontes de dados utilizadas para a análise exploratória englobam um compilado dos dados do SIVEP-Malária e do censo do IBGE. Embora o objetivo do SIVEP-Malária seja para monitoramento e administração da malária, os dados compõem uma excelente fonte para pesquisa científica [Wiefels et al., 2016]. Desde que foi implantado, em 2003, o formulário para seu preenchimento sofreu algumas alterações. O período escolhido (2009-2015) para essa pesquisa, no entanto, apresenta certa regularidade nas variáveis existentes. No total, 30 atributos compõem o banco de dados com 15.764.287 registros. Desses registros, cerca de 12% correspondem a casos positivos de malária.

As demais fontes de dados também constam da Plataforma, incluindo os dados do Sistema de Informações sobre Mortalidade (SIM) do Ministério da Saúde, dados do Sistema de Informações Sobre Nascidos Vivos (SINASC) do Ministério da Saúde e os dados do Censo Demográfico 2010 do Instituto Brasileiro de Geografia e Estatística (IBGE).

4. Metodologia

A integração do dado estudado com dados de outras fontes proporciona adição de novas informações, enriquecendo a análise. Ao se integrar dados é possível obter uma visão diferenciada e mais detalhada de ocorrências, podendo colocar a análise exploratória em um contexto diferenciado. Neste trabalho, por exemplo, usamos os dados do Censo do IBGE para melhor observar o efeito da malária de acordo com as características da população.

Além disto, para a condução da análise exploratória é necessário realizar o pré-processamento dos dados para melhorar a qualidade destes, seja em exatidão, integridade, consistência ou interpretabilidade [Han et al., 2011]. Neste trabalho o dado bruto com as informações do SIVEP-Malária passaram por limpeza de dados (remoção de valores inconsistentes) e por transformação de dados (valores traduzidos de código para texto).

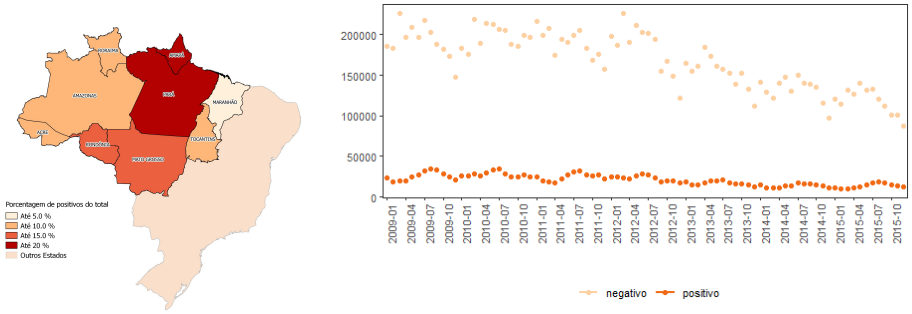
O pré-processamento, assim como a análise exploratória, foi feito via R Studio Server, na PCDaS. Nessa ferramenta é feita a conexão dos pacotes R, utilizando a infraestrutura de processamento e armazenamento da Plataforma. Por meio do R Studio Server é possível trabalhar na interface do R Studio diretamente na página do navegador via nuvem tendo disponível todo o potencial de processamento suportado pela Plataforma. A PCDaS conta com quatro servidores para armazenamento das bases de dados onde cada um deles conta com oito processadores Intel Xeon de oito núcleos, 16 GB de memória RAM e compartilham um total de 40 TB de armazenamento.

5. Análise Exploratória dos Dados

Nesta seção é apresentada a análise exploratória sobre SIVEP-Malária. Esse material foi selecionado de modo a apresentar características relevantes a serem discutidas na próxima seção. Todo o estudo exploratório e a confecção dos gráficos foi executado a partir do uso do R Studio Server da PCDaS.

A Figura 1.a apresenta a percentagem de amostras que confirmaram a suspeita de malária. Os estados do Pará e Amapá apresentam as maiores proporções de casos por

atendimento, já que, a cada 100 amostras coletadas, um número entre 15 a 20 confirmam o diagnóstico da doença. O Maranhão é o único estado a apresentar menos de 5% de casos entre as amostras realizadas.

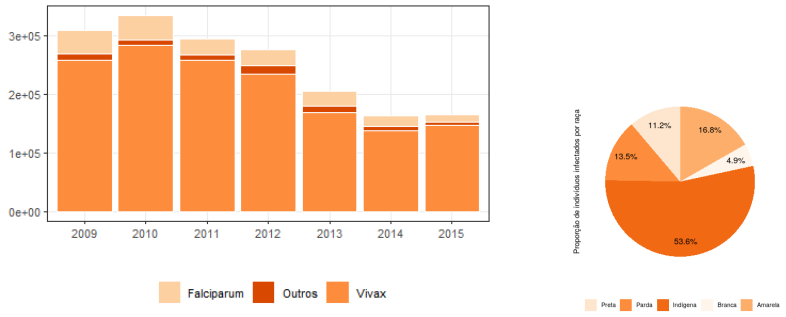


(a) (b)

Figura 1. Perc. de casos entre 2009 e 2015 (a); Evolução de casos da malária (b)

A Figura 1.b apresenta a evolução da quantidade de casos positivos e negativos. Percebe-se um padrão decrescente ao longo dos anos no número de casos positivos e negativos, que nesse caso reflete o número de amostras realizadas para teste de malária. Outra tendência é a sazonalidade percebida principalmente na linha dos casos positivos. É notável o aumento de número de casos no meado do ano (inverno no Brasil) e a diminuição no início e fim do ano.

A Figura 2.a apresenta a distribuição do tipo de malária nos sete anos de estudo. Vivax é claramente o tipo de plasmódio predominante, sendo o tipo Falciparum o segundo mais presente. Outros tipos de Plasmódio ocorrem mais raramente, são esses: F+FG (P. falciparum + gametócitos de P. falciparum), F+V (P. falciparum + P. vivax), V+FG (P. vivax + gametócitos de P. falciparum), FG (gametócitos de P. falciparum), M (P. malariae), F+M (P. falciparum + P. malariae), Ov (P. ovale) e Não F (não falciparum).



(a) (b)

Figura 2. Casos de malária por tipo (a); Proporções de raças afetadas (b)

Analisando os atributos relacionados às informações pessoais dos pacientes infectados por malária foi construída a Tabela 1. Vemos que em todos os estados predomina-se o gênero masculino como o mais afetado, existindo uma variação dessa proporção. Por exemplo, enquanto no Mato Grosso 77% dos casos são em pessoas do gênero masculino, no Acre a disparidade entre homens e mulheres é bem menor, sendo 56% homens e 44% mulheres.

Tabela 1. Proporção de casos na Amazônia Legal agrupados por gênero

Gênero	AC	AP	AM	MA	MT	PA	RO	RR	TO
Masculino	56%	62%	59%	64%	77%	62%	64%	61%	72%
Feminino	44%	38%	41%	36%	23%	38%	36%	39%	28%

A Figura 2.b apresenta as proporções dos índices criados pela divisão da quantidade de indivíduos afetados pela malária segundo o SIVEP-Malária pela quantidade absoluta de indivíduos segundo o censo do IBGE de 2010, para cada raça. Observa-se a população indígena como a mais afetada pela malária nos estados da Amazônia Legal. Isso significa que, dentre toda a população indígena uma porcentagem muito maior de indivíduos é infectada pela malária em comparação com outras raças.

Durante o estudo dos atributos, a partir da análise de matriz de correlação de Pearson entre os atributos, notou-se uma correlação entre os atributos **sintomas** e **tipo de detecção**. Isso motivou a análise combinada desses dois atributos para entender o motivo dessa correlação. A Tabela 2 mostra o resultado dessa análise. Percebe-se que, para a detecção do tipo ativa (quando o paciente é procurado pelo profissional de saúde para fazer o exame) 16% dos pacientes diagnosticados com malária não apresentavam sintomas. No caso da detecção passiva (quando o paciente procura a unidade de saúde notificante para fazer o exame) apenas em 2% dos casos o paciente não apresentava sintomas.

Tabela 2. Associação entre o tipo de detecção e a percepção dos sintomas.

Sintoma	Detecção Ativa	Detecção Passiva
Sim	84%	98%
Não	16%	2%

6. Discussão

Nesta seção os estudos conduzidos na análise exploratória da malária a partir da PCDaS são discutidos. A discussão leva ao levantamento de perguntas de pesquisas que, uma vez respondidas, podem ajudar na compreensão e no combate da malária no Brasil. O cartograma da Figura 1.a apresenta certa disparidade entre os estados na proporção número de casos de malária por número de amostras. Uma pergunta a ser feita é se os estados que apresentam proporções mais altas não estariam fazendo pouco exame ou se os estados que apresentam proporções baixas não estariam atingindo os grupos de risco e, portanto, não encontrando muitos pacientes contaminados em suas amostras.

Pelo fato de ter o mosquito como vetor, a prevalência da malária está relacionada a eventos de chuva [Sahle and Meshesha, 2014]. Sabemos que o verão, por ser mais quente, é a época do ano em que o índice de precipitação é mais elevado. Isso nos leva a supor que é neste período que haverá maior quantidade de casos de malária. No entanto, ao observar o gráfico da Figura 1.b, percebemos que há tanto mais exames feitos (refletidos nos pontos plotados para amostras com resultado negativo) quanto mais ocorrências de casos positivos de malária durante os meses do inverno.

A Figura 2.a mostra a predominância do tipo Vivax entre todas as ocorrências de malária na Região da Amazônia Legal, em todos os anos pesquisados. É sabido que o

tipo Vivax é predominante no território brasileiro (74.1% dos casos [WHO, 2018]), mas há algum padrão envolvido na ocorrência de outros tipos? Uma hipótese é que os casos de malária para os outros tipos de plasmódio que não Vivax podem estar localizados em regiões específicas ou serem provenientes de casos importados para o Brasil.

A predominância da malária em homens é uma tendência observada em todos os estados, segundo mostrado na Tabela 1. As hipóteses relacionadas a essa tendência é que homens, talvez por sua ocupação, estejam mais expostos às áreas com alta incidência do mosquito vetor. Exemplos disso são atividades de garimpo e exploração vegetal. As proporções de homem/mulher variam de estado para estado. Tal variação pode estar ligada também à disponibilidade de trabalho em cada estado.

Figura 2.b é relevante para enfatizar a importância da integração de dados pois, caso a quantidade de indivíduos de cada raça não fosse considerada segundo os dados do IBGE, o gráfico representaria as porcentagens dos números absolutos de cada raça na amostra e, dessa forma, a raça parda predominaria (já que em 70% das amostras o paciente se declara da raça parda) e perderíamos a noção da maior incidência entre a população indígena.

Sobre o gráfico da Tabela 2 observa-se que para detecção ativa existe mais pessoas sem sintomas que no caso da detecção passiva. Isto é esperado, uma vez que o paciente normalmente busca por atendimento em decorrência do fato de estar sentindo algum sintoma. A análise também mostra que a detecção ativa pode ser importante para encontrar (e, posteriormente, diagnosticar e tratar) indivíduos que apresentam casos assintomáticos e que, muito provavelmente, não procurariam uma unidade de saúde.

Ainda nesta tabela é curioso o fato de terem pessoas (2%) que mesmo que assintomáticas vão à procura de atendimento médico por demanda espontânea. Isto pode estar relacionado à procura por prevenção que pode acontecer, por exemplo, em casos onde o paciente vive em uma região endêmica. Outra possibilidade é que pode ter sido exigido a essas pessoas que atestassem que não estavam com malária, ou por questões trabalhistas ou por condições médicas, por exemplo exame pré-natal.

Todas as considerações levantadas nessa seção com o objetivo dar explicações para as ocorrências observadas na análise exploratória tratam-se de conjecturas ou hipóteses. A comprovação pode se dar por uma análise mais aprofundada ou integrada desses dados.

7. Conclusão

A análise exploratória feita na Plataforma viabilizou a elaboração de perguntas de pesquisa. Algumas dessas pergunta talvez possam ser respondidas por especialistas em malária, em epidemiologia ou por cientistas da área da saúde. Outras perguntas demandam mais integração de dados para serem pesquisadas, como por exemplo a análise conjunta de dados de malária e dados ambientais, climáticos ou socioeconômicos. Também existem aquelas perguntas que são ainda mais complexas e, para buscar respostas, seria necessário o apoio de métodos de mineração de dados. Muitas das perguntas levantadas trazem aspectos espaço-temporais como *proxy* para um melhor entendimento da malária no Brasil. Tal fato requer a integração com fontes adicionais de dados, como, por exemplo, informações de regiões de saúde.

Adicionalmente, de modo a apoiar os especialistas da área, pretendemos fazer uso no PCDaS de técnicas de mineração de dados como, por exemplo, padrões frequentes. Pretendemos analisar as regras de associação e verificar se a distribuição das regras formadas é aderente aos dados observados na análise exploratória. Caso haja divergências, as regras divergentes serão avaliadas, uma vez que aquelas fogem ao senso comum da análise exploratória.

Referências

Confalonieri, U., Margonari, C., and Quintão, A. (2014). Environmental change and the dynamics of parasitic diseases in the Amazon. *Acta Tropica*, 129(1):33–41.

Diallo, A., Sié, A., Sirima, S., Sylla, K., Ndiaye, M., Bountogo, M., Ouedraogo, E., Tine, R., Ndiaye, A., Coulibaly, B., and others (2017). An epidemiological study to assess Plasmodium falciparum parasite prevalence and malaria control measures in Burkina Faso and Senegal. *Malaria journal*, 16(1):63.

Griffing, S. M., Tauil, P. L., Udhayakumar, V., and Silva-Flannery, L. (2015). A historical perspective on malaria control in Brazil. *Memórias do Instituto Oswaldo Cruz*, 110(6):701–718.

Guerin, P. J., Olliaro, P., Nosten, F., Druilhe, P., Laxminarayan, R., Binka, F., Kilama, W. L., Ford, N., and White, N. J. (2002). Malaria: current status of control, diagnosis, treatment, and a proposed agenda for research and development. *The Lancet infectious diseases*, 2(9):564–573.

Han, J., Kamber, M., and Pei, J. (2011). *Data Mining: Concepts and Techniques*. Morgan Kaufmann, Haryana, India; Burlington, MA, 3 edition.

Johansson, E., Selling, K., Nsona, H., Mappin, B., Gething, P., Petzold, M., Peterson, S., and Hildenwall, H. (2016). Integrated paediatric fever management and antibiotic over-treatment in Malawi health facilities: Data mining a national facility census. *Malaria Journal*, 15(1).

Labbo, R., Fandeur, T., Jeanne, I., Czeher, C., Williams, E., Arzika, I., Soumana, A., Lazoumar, R., and Duchemin, J.-B. (2016). Ecology of urban malaria vectors in Niamey, Republic of Niger. *Malaria journal*, 15(1):314.

Loucoubar, C., Paul, R., Bar-Hen, A., Huret, A., Tall, A., Sokhna, C., Trape, J.-F., Ly, A., Faye, J., Badiane, A., Diakhaby, G., Sarr, F., Diop, A., Sakuntabhai, A., and Bureau, J.-F. (2011). An exhaustive, non-euclidean, non-parametric data mining tool for Unraveling the complexity of biological systems - novel insights into malaria. *PLoS ONE*, 6(9).

Recht, J., Siqueira, A. M., Monteiro, W. M., Herrera, S. M., Herrera, S., and Lacerda, M. V. (2017). Malaria in Brazil, Colombia, Peru and Venezuela: current challenges in malaria control and elimination. *Malaria journal*, 16(1):273.

Sahle, G. and Meshesha, M. (2014). Uncovering knowledge that supports malaria prevention and control intervention program in ethiopia. *Electronic Journal of Health Informatics*, 8(1).

Sweeney, A., Beebe, N., and Cooper, R. (2007). Analysis of environmental factors influencing the range of anopheline mosquitoes in northern Australia using a genetic algorithm and data mining methods. *Ecological Modelling*, 203(3-4):375–386.

WHO (2018). *World malaria report 2018*. World Health Organization.

Wiefels, A., Wolfarth-Couto, B., Filizola, N., Durieux, L., and Mangeas, M. (2016). Accuracy of the malaria epidemiological surveillance system data in the state of Amazonas. *Acta Amazonica*, 46(4):383–390.