




OPEN


Spatiotemporal mapping of malaria prevalence in Madagascar using routine surveillance and health survey data

Rohan Arambepola¹, Suzanne H. Keddie², Emma L. Collins¹, Katherine A. Twohig¹, Punam Amratia¹, Amelia Bertozzi-Villa^{1,8}, Elisabeth G. Chestnutt¹, Joseph Harris², Justin Millar¹, Jennifer Rozier², Susan F. Rumisha¹, Tasmin L. Symons¹, Camilo Vargas-Ruiz¹, Mauricette Andriamananjara^{5,7}, Saraha Rabeherisoa⁵, Arsène C. Ratsimbaoa^{5,6}, Rosalind E. Howes^{1,4}, Daniel J. Weiss^{1,2,3}, Peter W. Gething^{1,2,3} & Ewan Cameron^{1,2,3}

Malaria transmission in Madagascar is highly heterogeneous, exhibiting spatial, seasonal and long-term trends. Previous efforts to map malaria risk in Madagascar used prevalence data from Malaria Indicator Surveys. These cross-sectional surveys, conducted during the high transmission season most recently in 2013 and 2016, provide nationally representative prevalence data but cover relatively short time frames. Conversely, monthly case data are collected at health facilities but suffer from biases, including incomplete reporting and low rates of treatment seeking. We combined survey and case data to make monthly maps of prevalence between 2013 and 2016. Health facility catchment populations were estimated to produce incidence rates from the case data. Smoothed incidence surfaces, environmental and socioeconomic covariates, and survey data informed a Bayesian prevalence model, in which a flexible incidence-to-prevalence relationship was learned. Modelled spatial trends were consistent over time, with highest prevalence in the coastal regions and low prevalence in the highlands and desert south. Prevalence was lowest in 2014 and peaked in 2015 and seasonality was widely observed, including in some lower transmission regions. These trends highlight the utility of monthly prevalence estimates over the four year period. By combining survey and case data using this two-step modelling approach, we were able to take advantage of the relative strengths of each metric while accounting for potential bias in the case data. Similar modelling approaches combining large datasets of different malaria metrics may be applicable across sub-Saharan Africa.

Malaria is a major public health problem in Madagascar, with an estimated 2.16 million cases leading to more than 5000 deaths in the country in 2018¹. Malaria burden decreased in the early 2000s with an increase in control efforts but this progress was largely halted following political turmoil in 2009^{2–4} resulting in a resurgence in endemicity in the last decade^{1,4–6}. Transmission exhibits strong spatial trends, with high endemicity in the coastal regions and lower transmission in the highlands, and seasonal patterns in incidence are widely observed^{4,5}. An accurate understanding of spatiotemporal variation in transmission can facilitate strategic resource allocation and evaluation of control measures^{1,4,7}.

Routine malaria case data are collected through the Health Management Information System (HMIS) from reports from health facilities. These data are collected monthly and have a high spatial coverage (see Fig. 2) but also suffer from a number of potential biases. This passive case detection is unlikely to capture all malaria cases in the community, missing those that do not seek care or do so from informal or private providers, which likely represents a large fraction of the population (treatment-seeking rates in the public sector in 2013 were estimated to be around 45%)^{8–10}. Furthermore, cases seen at health facilities may not be diagnosed or reported to the central

¹Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford, UK. ²Telethon Kids Institute, Perth Children's Hospital, Perth, Australia. ³Curtin University, Perth, Australia. ⁴Foundation for Innovative New Diagnostics, Geneva, Switzerland. ⁵Programme National de Lutte contre le Paludisme, Antananarivo, Madagascar. ⁶University of Fianarantsoa, Fianarantsoa, Madagascar. ⁷Ministère de Santé Publique, Antananarivo, Madagascar. ⁸Institute for Disease Modeling, Bellevue, WA, USA. email: rohan.arambepola@stx.ox.ac.uk

system due to resource constraints (such as malaria rapid diagnostic test (RDT) stock-outs) or weak communication infrastructure⁴. Nevertheless, while these data may not represent all malaria incidence, they are an important source of information on trends in transmission due to their high temporal and spatial coverage^{4,6,11–13}.

Malaria Indicator Surveys (MIS) provide another source of data for understanding the spatiotemporal patterns of malaria endemicity^{9,10,14}. These cross-sectional surveys are designed to be nationally representative and collect data on a number of indicators, including prevalence of *Plasmodium falciparum* infection in individuals between 6 and 59 months of age. They are conducted with a standardised methodology applied over all sites and surveys and, unlike routine case data, are not affected by reporting incompleteness, treatment-seeking behaviour or standards for clinical diagnosis. For these reasons, prevalence information from national health surveys has traditionally been the primary source of data for mapping malaria risk in sub-Saharan Africa^{3,5,15,16}. Kang et al.⁵ used a Bayesian hierarchical model to map prevalence in Madagascar in 2011, 2013 and 2016 using parasite prevalence data from the 2011, 2013 and 2016 MIS reports^{9,10,14}. However, cross-sectional surveys provide limited insight into seasonal patterns of transmission or transmission in years when no survey took place. Variation in the timing of surveys between years may also make it difficult to distinguish changes in prevalence between survey years from seasonal variation. Moreover, survey data is less informative in low burden areas, where sample sizes are likely to be inadequate to accurately assess changes in transmission due to low rates of detectable parasitaemia¹⁷.

In this study, we combined routine case data and prevalence survey data within a formal modelling framework, taking advantage of their relative strengths, in order to provide a more complete understanding of the spatiotemporal heterogeneity of transmission between 2013 and 2016. One method for combining prevalence and incidence data is to use a joint model, a single model which includes the likelihoods of both metrics and the relationship between them^{18–20}. These models allow for sharing of information between metrics and produce predictions of both incidence and prevalence. However, balancing this sharing of information can be challenging, particularly in situations where one response variable has many more observations (and therefore a much greater likelihood contribution) than the other^{20,21}. Moreover, there is limited scope to learn the relationship between response variables without introducing too many degrees of freedom and therefore this relationship is often fixed or highly constrained. Lucas et al.¹⁹, for example, modelled malaria prevalence and incidence jointly using a fixed previously estimated relationship²². To allow us to learn the relationship between incidence and prevalence and control the relative effect of each metric on the final results, we used a two-step process in which we produced spatially smooth monthly incidence surfaces which were then used as a covariate in a Bayesian geostatistical model of prevalence. This method is conceptually simpler than a joint model and allows the incidence-to-prevalence relationship to be learned within the prevalence model. An important benefit of learning this relationship is that systematic biases in routine case data are implicitly accounted for. Modelling the incidence and prevalence processes separately is equivalent to making a ‘cut’ between the incidence and prevalence processes in the model. Modularising the inference in this way prevents over-reliance on the less reliable incidence data in the final prevalence estimates^{20,21}. This approach is similar to the use of modelled surfaces, such as temperature suitability²³ or accessibility²⁴, as inputs when mapping malaria risk^{3,5,15,16}. It is also similar to the work of Lucas et al.²⁵ who used modelled prevalence surfaces as inputs to an incidence model. However, their prevalence model used environmental and socioeconomic covariates and therefore overall this approach was equivalent to a non-linear model of incidence using these covariates. In contrast, due to the high spatiotemporal coverage of the routine case data, our incidence surfaces are a direct smoothing of treatment-seeking adjusted observed incidence rates and therefore represent malaria risk more directly.

We incorporated routine case data to update the estimates of prevalence for 2013 and 2016 made by Kang et al.⁵ and to produce estimates for 2014 and 2015. By producing monthly prevalence maps over all four years, both long term and seasonal trends in prevalence across the country could be assessed. We used the eight ecozones identified by Howes et al.⁴ as a basis for assessing how these trends vary spatially.

Methods

Study area. Malaria transmission is highly heterogeneous across Madagascar, reflecting the diverse ecological landscape of the island. Transmission is highest in the east and west coastal regions, where it follows a seasonal pattern with clinical incidence peaking between February and May depending on location. In the central highlands and the desert south, transmission is lower and annual trends are less consistent, with temporal variation appearing to be driven by outbreak dynamics^{4,11,26}. Howes et al.⁴ identified eight contiguous ecozones representing distinct transmission settings (Central highlands, Highlands fringe west, Highlands fringe east, Northeast, Northwest, Southeast, Southwest, South) which are shown in Fig. 1a. When devising the 2018–2022 malaria National Strategic Plan, the National Malaria Control Programme (NMCP) of Madagascar classified 106 of the 114 districts in Madagascar as control areas, 3 pre-elimination and 5 elimination, based on reported case numbers in 2016⁷. Control strategies are currently stratified by risk level, with intermittent preventative treatment for pregnant women (IPTp) and mass distribution of insecticide-treated bed nets (ITNs) in the 106 control districts and indoor residual spraying (IRS) targeted at high transmission districts in the southeast and southwest. In pre-elimination settings the focus is on outbreak and active case detection^{27,28}.

Study data. Parasite prevalence data was available from Malaria Indicator Surveys that took place in Madagascar in 2013 and 2016. These data consist of geo-located clusters where the prevalence of *P. falciparum* infection was measured, determined by microscopy, in individuals between 6 and 59 months of age ($PfPR_{6-59\text{months}}$). These surveys were conducted with a standard protocol and survey sites were selected to produce nationally representative estimates of prevalence in this age group. The number of positive individuals and the total number tested was recorded at each site. In the 2013 and 2016 surveys, 6323 and 6927 individuals were screened across

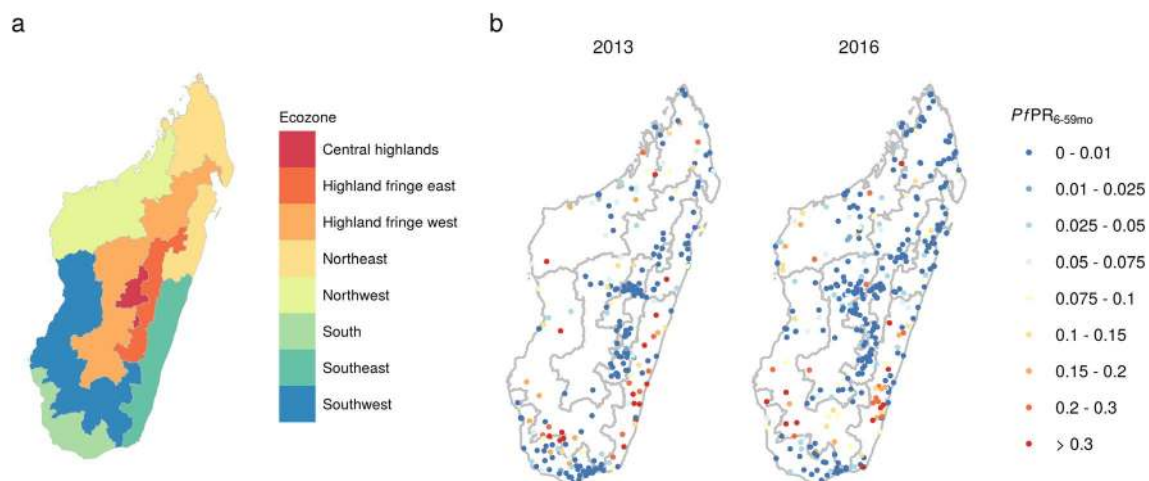


Figure 1. (a) Ecozones defined by Howes et al.⁴ representing contiguous areas with distinct patterns of transmission. (b) Prevalence rates at survey sites in the 2013 and 2016 MIS. These maps were created in R (version 3.6.2, <https://www.r-project.org/>) using the ggplot2²⁹ package.

Covariate	Description	Type	In any causal feature set	In final feature set
Rainfall ³²	Climate hazards group infrared precipitation with station data	Dynamic	Lag 0	Lag 0
LST day ³³	Daytime land surface temperature	Dynamic	No	No
LST night ³³	Night-time land surface temperature	Dynamic	Lag 1	No
TCB ³⁴	Tasselled cap brightness; measure of land reflectance	Dynamic	Lag 2	No
EVI ³⁵	Enhanced vegetation index	Dynamic	No	No
TSI Pf ²³	Temperature suitability index for <i>P. falciparum</i>	Dynamic	Lag 2	No
Accessibility ²⁴	Distance to cities with population > 50,000	Static	Yes	Yes
AI ³⁶	Aridity index	Static	Yes	Yes
Elevation ³⁷	Elevation as measured by the shuttle radar topography mission (SRTM)	Static	Yes	No
PET ³⁶	Potential evapotranspiration	Static	Yes	No
Slope ³⁷	GIS-derived surface calculated from SRTM elevation surface	Static	Yes	No
Night lights ³⁰	Index that measures the presence of lights from towns, cities and other sites with persistent lighting	Static	No	No
Distance to water ³⁸	GIS-derived surface that measures distance to permanent and semi-permanent water based on presence of lakes, wetlands, rivers and streams, and accounting for slope and precipitation	Static	Yes	Yes
TWI ³⁷	Topographic wetness index	Static	No	No

Table 1. List of covariates.

274 and 358 sites, respectively. The survey sites and proportions of positive individuals are shown in Fig. 1b and full details can be found in the original reports^{9,10}.

Monthly health facility case data between January 2013 and December 2016 were provided by the NMCP. These data come from HMIS data reports and represent clinical cases of malaria confirmed by an RDT for all ages, irrespective of parasite species. Data were available from 3342 health facilities in Madagascar, of which 2801 were geo-located and verified using a separate dataset of geo-located health facilities from the Institut Pasteur de Madagascar (as described in¹¹).

A number of covariates were used to inform the prevalence model, which are detailed in Table 1. Most of these variables were environmental, influencing vector abundance, parasite behaviour and environmental suitability. Two of these covariates (accessibility to cities²⁴ and night lights³⁰) relate to the development and urbanicity of a location and therefore are related to vector abundance and to socioeconomic factors, such as access to healthcare. In total there were 26 potential features for the prevalence model—8 static covariates and 6 dynamic covariates, each considered at 0, 1 and 2 month time lags. Causal feature selection³¹ was used to select the variables included in the final model.

Catchment population. In order to calculate incidence rates at each health facility, estimates for catchment populations (the number of people likely to seek treatment at each facility) were needed. A catchment model was used to estimate these (treatment-seeking) catchment populations, based on travel time to health facilities. The country was divided into a grid of approximately 5 km-by-5 km pixels. For each pixel, the travel time to each health facility was calculated using a friction surface (defining travel time through each pixel) developed by Weiss et al.²⁴ and a least cost algorithm³⁹. Given that an individual in pixel *i* seeks treatment, the

probability they seek treatment in health facility j , $p(\text{pixel}_i \rightarrow \text{HF}_j)$, was modelled as inversely proportional to square of the travel time to that health facility. That is,

$$p(\text{pixel}_i \rightarrow \text{HF}_j) = \frac{t(\text{pixel}_i \rightarrow \text{HF}_j)^{-2}}{\sum_{j=1}^{N_{\text{HF}}} t(\text{pixel}_i \rightarrow \text{HF}_j)^{-2}}$$

where $t(\text{pixel}_i \rightarrow \text{HF}_j)$ represents the travel time from pixel i to health facility j and N_{HF} was the total number of health facilities. The catchment population of health facility j was then calculated as

$$\sum_{i=1}^{N_{\text{pixel}}} \text{population}_i \times p(\text{pixel}_i \rightarrow \text{HF}_j)$$

where population_i is the treatment seeking-adjusted population of pixel i and N_{pixel} is the total number of pixels. The proportion of the population in each pixel who would seek treatment at any health facility was also modelled as a function of travel time, which has been identified as an important factor in treatment-seeking behaviour for fever in Madagascar⁴⁰. This proportion was modelled as a logistic function (similar to the functional forms considered by Alegana et al.⁴¹)

$$\frac{\alpha}{1 + \exp(\sigma t)} + \beta$$

where t is the travel time to the nearest health facility in minutes. The parameters values $\alpha = 0.6$, $\sigma = 0.00916$, $\beta = 0.15$ were chosen such that the maximum and minimum possible treatment-seeking proportions were 0.6 and 0.15, and the treatment-seeking proportion at $t = 120$ minutes was 0.3. These parameter values were chosen to produce a similar relationship between treatment-seeking and travel time as that observed in the 2013 and 2016 MISs^{9,10} (see Supplementary Material Fig. 3) and match overall estimated treatment-seeking rates⁸. Catchment populations were calculated for each year between 2013 and 2016 using hybrid population surfaces from the Gridded Population of the World v4⁴² and WorldPop⁴³. In order to assess the sensitivity of the final estimates to these modelling assumptions, three alternative sets of populations were considered: catchment populations supplied by the NMCP (based on the 1993 national census adjusted by a fixed annual growth rate⁴⁴ with no adjustment for treatment-seeking behaviour) and two sets of catchment populations generated by the catchment model under different treatment-seeking parameters (see Supplementary Material). The analysis was repeated using these catchment populations and the resulting prevalence estimates were compared.

Incidence model. Routine case data from health facilities were modelled using a Bayesian geospatial model to produce monthly incidence surfaces which were then used as inputs to the prevalence model. Let c_{it} be the number of cases observed in month t ($t = 1, \dots, 48$) at health facility i ($i = 1, \dots, N$), which is at location s_i and has a treatment-seeking catchment population E_i . The number of cases observed was modelled as a Poisson process

$$c_{it} \sim \text{Pois}(E_i \times \lambda_{it})$$

with mean equal to the product of the catchment population and the underlying incidence rate, λ_{it} . The log incidence rate was modelled as

$$\log \lambda_{it} = \beta_0 + f(s_i, t)$$

where β_0 was an intercept and for each month $f(\cdot, t)$ was a realisation of a Gaussian process over space with zero mean and Matérn covariance structure. The Matérn covariance function is parameterised by the range, ρ , and marginal variance, σ^2 , the values of which were chosen by a search over parameter space to maximise accuracy when predicting incidence in held out locations. These parameters were optimised jointly over all months. The value of smoothness parameter, ν , was fixed at 1.

Prevalence model. Prevalence data from MIS surveys were also modelled using a Bayesian geospatial model, informed by environmental and socioeconomic covariates and the modelled incidence surfaces. Let y_i be the number of infected individuals and N_i be the number of individuals tested in survey i ($i = 1, \dots, M$), taking place in location s_i at time t_i . The results of the survey were modelled as a realisation of a binomial process

$$y_i \sim \text{Binomial}(p_i, N_i)$$

with underlying prevalence p_i at this location and point in time. The logit-transformed prevalence was modelled as

$$\text{logit}(p_i) = \beta_0 + \beta^T X_i + \beta_0^{\text{inc}} g(\lambda(s_i, t_i)) + \beta_1^{\text{inc}} g(\lambda(s_i, t_i + 1)) + f(s_i)$$

where β_0 , β , β_0^{inc} , β_1^{inc} were parameters (with β_0^{inc} and β_1^{inc} non-negative), X_i were covariate values and $\lambda(s, t)$ was the log incidence value from the modelled incidence surfaces at location s and time t . f was modelled as a realisation of a Gaussian process over space with Matérn covariance, parameterised by the range ρ and marginal variance σ^2 , while g is a realisation of a Gaussian process over incidence with a squared exponential kernel, parameterised by the scale κ (with fixed variance 1). This allowed a non-linear effect of incidence on logit-prevalence while limiting model-complexity by assuming a priori that this relationship is smooth (by the choice of squared exponential kernel and placing an appropriate prior on the scale parameter). This flexibility was important as

incidence is likely to be the main driver of prevalence in the model (while the additional covariates are likely to be less informative and therefore were modelled as having linear effects) and assuming a linear effect produces prevalence-incidence relationships that do not match empirical observations²². Incidence in the subsequent month was included in addition to incidence in the current month, as the presence of parasites in the blood could result in a clinical case that is recorded in a health centre several weeks later. The parameters β_0^{inc} and β_1^{inc} allowed the model to learn the relative predictive power of incidence at these two time points.

The Bayesian model was completed by placing appropriate priors on the parameters and hyperparameters. Normal priors were placed on β_0 , β , β_0^{inc} , β_1^{inc} centred around 0 with standard deviation 1 for β_0 and 0.25 for β , β_0^{inc} , β_1^{inc} . The Matérn covariance parameters were given log-normal priors with mean 3 and standard deviation 0.1 for ρ and mean 0 and standard deviation 0.1 for σ , shrinking the spatial term towards a fairly flat and smooth field. A log-normal prior was placed on κ with mean 3 and standard deviation 0.1. Both the incidence and prevalence models were fitted using the Gaussian Markov Random Field approximation⁴⁵ with the **Template Model Builder** package⁴⁶ in **R**⁴⁷ and samples were drawn from a Laplace approximation to the posterior to produce associated uncertainty estimates. Uncertainty was quantified using the interquartile range of the posterior distribution and exceedance and non-exceedance probabilities. For a given prevalence threshold, the exceedance probability at a location is defined as the probability that a prevalence sample from the posterior at this location will exceed the threshold. Similarly, a non-exceedance probability is the probability a prevalence sample will not exceed a given value. The model was validated by fitting the model with data from the 2013 survey and making predictions for 2016 (to evaluate model performance when predicting prevalence in previously unobserved time points) and through k -fold cross validation.

Feature selection. The covariates included in the prevalence model were selected from the potential covariates described in Table 1 using causal feature selection^{31,48}. In total there were 26 potential features, 18 dynamic (6 covariates each at 0, 1 and 2 month time lags) and 8 static. The idea behind causal feature selection is to select features with the most direct causal relationships to the response based on the available data.

We describe the procedure briefly here which is described in full detail by Arambepola et al.³¹. The PC algorithm⁴⁹, a causal discovery algorithm, was used to infer causal relationships between the different features and malaria prevalence. In particular, the output of this algorithm identified all direct causes of prevalence. To quantify the certainty of these direct causes, the algorithm was repeatedly applied to bootstrapped samples of the data. The certainty of a feature being a direct cause of prevalence was then quantified as the proportion of repeats in which it was inferred to be a direct cause. For a given minimum certainty, feature sets were then generated by selecting all direct causes of prevalence and a number of potential feature sets were generated by varying the minimum certainty required between 0 and 1. Out of these potential feature sets, the final set chosen was the feature set which maximised the cross-validated predictive performance of the model. The causal discovery algorithm relies on conditional independence testing. We used the Randomized Conditional Independence Test⁵⁰ to perform scalable non-parametric conditional independence tests.

Selecting causal features may be beneficial for a number of reasons. It is possible that causal selection may lead to smaller feature sets, especially in situations in which many features are associated with the response variable but relatively few are directly causal⁴⁸. Small feature sets may improve computational efficiency and reduce overfitting. Models built on causal feature sets may also be more robust to common problems such as concept drift and covariate shift⁵¹ and therefore make more useful predictions further forward in time or in previously unobserved locations. Arambepola et al.³¹ showed that using causal feature selection resulted in improved performance when forecasting malaria incidence compared to classical feature selection.

Results

Catchment populations. The median size of the modelled catchment populations was 2890 (LQ: 1710, UQ: 4740). The total population served according to the modelled populations increased from 10.15 million in 2013 to 11.02 million in 2016, corresponding to approximately 43% of the Malagasy population each year, which is largely in agreement with estimated treatment-seeking rates⁸.

Incidence. Annual incidence rates at each health facility (calculated using reported cases and modelled catchment populations) are shown in Fig. 2. Spatial patterns of incidence in 2013 and 2016 were similar to observed prevalence (Fig. 1b) with lower rates in the central highlands and the south, and higher rates in the east and west coastal regions. Compared to 2013, incidence in 2016 was generally higher in health facilities in the southeast and southwest but lower in the north. Across the 4 years, rates were generally lowest in 2014 (with the exception of the east coast) and highest in 2015. Figure 3 shows monthly incidence rates aggregated by ecozone. A seasonal pattern of a single peak in incidence between January and April can be seen to some extent in most regions, most clearly in the Southeast, Northeast and (despite low overall incidence) Highland fringe east ecozones. An increase in cases in 2015 was observed in almost all ecozones.

The optimal hyperparameters for the incidence model were $\rho = e^{-0.1}$ and $\sigma = e^{-2}$. Supplementary Material Fig. 4 shows the smooth incidence surfaces produced by this model aggregated annually, which as expected reflect the overall spatiotemporal trend of the observed data. It should be noted that these surfaces were only intended to be a spatial smoothing of the treatment seeking-adjusted case data reflecting relative spatial trends, rather than an enumeration of true incidence rates, and are an intermediate step in the modelling process rather than an output.

Prevalence. The causal discovery algorithm identified 4 dynamic and 4 static feature sets which combined to give 16 potential feature sets for the prevalence model. The features that were present in at least one of these sets are listed in Table 1 and these sets are listed in full in the Supplementary Material. The features used in the

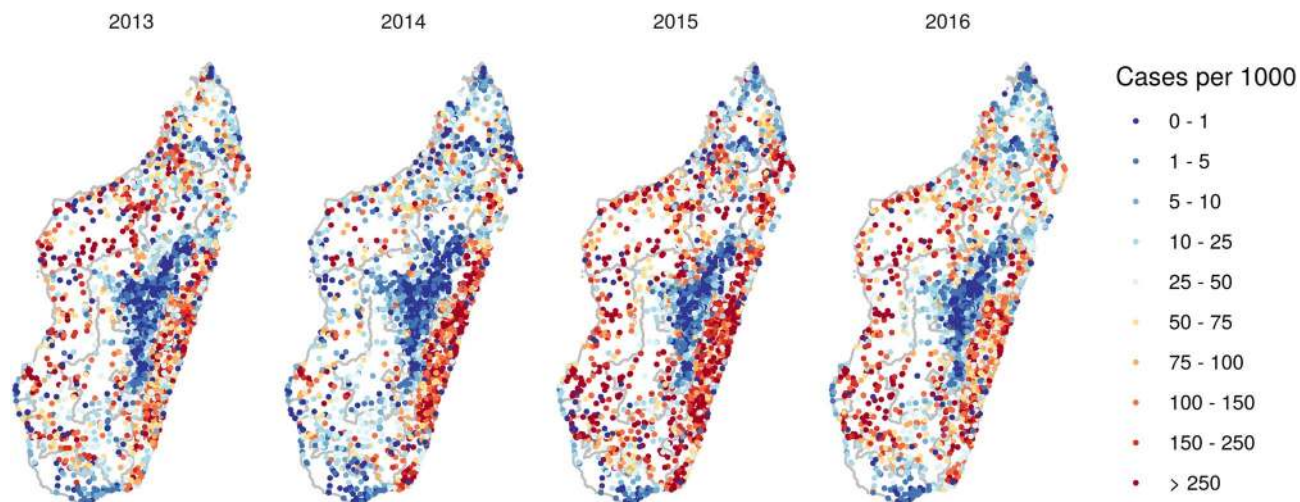


Figure 2. Annual incidence rates at each health facility based on routine case data and modelled catchment populations. These maps were created in R (version 3.6.2, <https://www.r-project.org/>) using the ggplot2 package.

final model were rainfall³² with no time lag, accessibility to cities²⁴, an aridity index³⁶ and distance to water³⁸. The posterior mean and 95% credible interval for all the model parameters, including coefficients of these features, are shown in Table 2. As expected, incidence was the most important predictor, with incidence in the current month having a greater effect than incidence in the following month. The other covariates in the model had smaller and less consistent effects on prevalence. When fitting the model on 2013 data and making predictions for 2016, there was a correlation of 0.63 between predicted and observed values. For 3, 4 and 5-fold cross-validation, there were average correlations of 0.58, 0.59 and 0.6, respectively. These values are reasonably high given that the observed prevalence rates are themselves noisy point estimates of underlying prevalence; for context, the standard errors of the mean of binomial samples of 20 individuals (the median survey size) with true prevalence values of 0.01, 0.05 and 0.15 are 0.022, 0.049 and 0.080, respectively.

Prevalence estimates for individuals between 6 and 59 months of age are shown aggregated annually in Fig. 4a. Spatial patterns were similar across all four years, with highest estimated prevalence near the southeast coast and other areas of high prevalence along the west coast. Prevalence estimates were consistently low in the far south of the country and in the central highlands. Prevalence in the north varied more between years, although was generally low or moderately low. Population-weighted mean prevalence was lowest in 2014 (6.0%, 95% credible interval (CI) 3.3–8.8), followed by 2013 (6.4%, CI 3.5–9.5) and 2016 (6.6%, CI 3.6–9.6), and was highest in 2015 (8.9%, CI 4.6–12.9). Population-weighted prevalence over time (Fig. 4b) showed a clear seasonal pattern, peaking between February and April each year with the lower prevalence occurring between July and September. Prevalence was highest in 2015, with the peak prevalence across all four years occurring in April 2015 (11.4%, CI 5.8–16.0) and notably high prevalence sustained into the lower transmission season. By the second half of 2016, prevalence appeared to have returned to similar levels to 2013 and 2014.

Figure 5 shows the mean prevalence over time for each ecozone. Prevalence was consistently highest in the Southeast ecozone, with peak monthly prevalence of 0.3 or more in all years, while consistently lowest in the Central highlands and Highland fringe east ecozones. Seasonal patterns of prevalence are present in most of these ecozones, including both higher transmission regions (Northeast, Southeast and Southwest) and the lower transmission highland ecozones. These regional seasonal trends are similar to the pattern observed for mean prevalence across the country, with a single peak occurring between February and April in most areas (though often earlier in the Southeast ecozone) and lowest prevalence around August. The Northwest and South ecozones exhibit some similar seasonal trends but these are less consistent. In most areas, prevalence in 2014 was slightly lower than in 2013, with large decreases in the Northwest and South ecozones. Increased prevalence in 2015 was observed in all areas. In the highland ecozones, prevalence was higher than normal around April but later in the year returned to similar levels to previous years. In the rest of the country, a high peak in April was generally followed by higher than average prevalence throughout the rest of the year and into 2016. The ecozones generally correspond well to these latest prevalence estimates (Supplementary Material Fig. 5 shows the 2016 estimates with ecozone borders highlighted). The greatest heterogeneity within ecozones was in the Highland fringe west ecozone, which contained a region of higher prevalence to the south, and the Northwest ecozone.

Monthly prevalence estimates can also be used to identify areas where transmission is consistently low or high. Two examples are shown in Figure 6 for estimates of prevalence in 2016. Figure 6a shows areas where prevalence was always below 5% or below 5% for at least 9 months and Fig. 6b areas where prevalence was always above 20% or above 20% for at least 6 or 9 months. Despite the seasonal trends in transmission (Fig. 5), in much of the central highlands prevalence was below 5% year-round, while some areas in the western highlands only exceeded 5% for at most 3 months of the year. Transmission was also consistently low in parts of the desert south and far north of the country. The areas of consistently high prevalence were concentrated on the east and southwest coasts, following a similar pattern to overall prevalence.

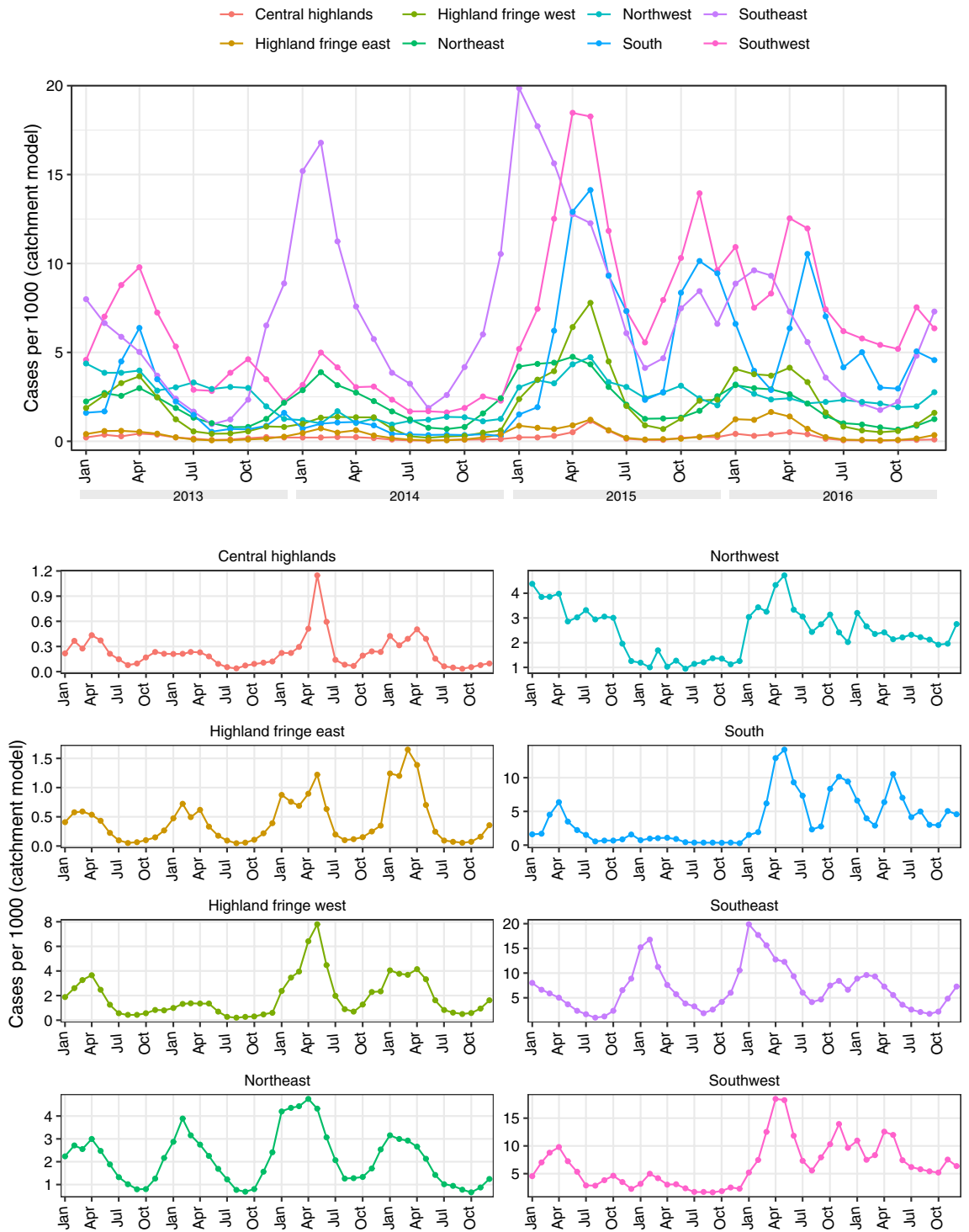


Figure 3. Monthly incidence rates at each health facility based on routine case data and modelled catchment populations stratified by ecozone (2013–2016). These graphs were created in R (version 3.6.2, <https://www.r-project.org/>) using the ggplot2 package.

Figure 7 shows two ways of visualising the uncertainty in prevalence estimates, using annual 2016 estimates as an example. The first (Fig. 7a) maps the interquartile range of the posterior distribution. The spatial patterns of uncertainty are similar to the prevalence estimates, with higher (absolute) uncertainty areas of high prevalence. An alternative quantification of uncertainty is in terms of probability of estimates exceeding (or not exceeding) chosen values. Figure 7b shows the probability of prevalence (sampled from the posterior distribution) exceeding 0.15. This can be interpreted as the probability that the true prevalence in 2016 exceeded 0.15. In the majority of the country there is a very high certainty that prevalence was below 0.15, while on the east and west coasts there

Parameter	Mean	CI
Intercept	-3.956	-6.082, -1.926
Accessibility	0.060	-0.021, 0.150
AI	-0.066	-0.172, 0.042
Distance to water	-0.031	-0.108, 0.042
Rainfall (no lag)	0.019	-0.097, 0.128
$\log \rho$	1.48	1.263, 1.693
$\log \sigma$	-0.762	-0.968, -0.561
$\log \lambda_{inc}$	1.883	0.881, 3.043
β_0^{inc}	0.802	0.407, 1.22
β_1^{inc}	0.497	0.184, 0.879

Table 2. Mean and 95% credible intervals for the prevalence model parameters

is fairly high confidence that prevalence exceeded this value. In the southwest, in particular, the probability of exceeding 0.15 is often above 90%. Similarly, we can visualise the probability of not exceeding a certain value, as shown in Fig. 7c with the probability of not exceeding 0.05. Here we can identify areas in the highlands, south and far north where there is very high confidence of low prevalence. We also see more areas around the north where there is less certainty, with probabilities between 20 and 50% of prevalence not exceeding 0.05. Maps of uncertainty in 2013, 2014 and 2015 using the interquartile range and exceedance probabilities are included in the Supplementary Material (Supplementary Material Figs. 5–7).

The sensitivity analysis performed using different catchment populations showed that our prevalence estimates were fairly robust to different methods of estimating these populations. Despite the systematic differences between these estimates, leading to significantly different incidence estimates (Supplementary Material Fig. 1), the resulting prevalence estimates were very similar (Supplementary Material Fig. 2).

Discussion

Our results highlight the benefits of combining routine case data and cross-sectional survey data to provide a more complete understanding of prevalence over time. While the 2013 and 2016 MIS data show similar levels of prevalence in these 2 years, the routine case data suggests there was substantial variation in transmission between these time points, with decreasing case numbers in 2014 followed by a marked increase in 2015. Incorporating this case data into a formal modelling framework allowed us to quantify these trends in terms of changes in prevalence over time. The robust two-step modelling approach allowed us to learn from these noisy data sources effectively while retaining flexibility in important model components, in particular the incidence-to-prevalence relationship.

By producing monthly risk maps we are able to assess temporal trends of interest in detail. For example, in much of the country the increase in prevalence in 2015 was caused by increased prevalence throughout the year and into early 2016, rather than only higher than average prevalence during the high transmission season. The observed increase in cases in 2015 has been attributed to cyclones and the resulting flooding and lack of supplies in many parts of the country⁵². Our results may help confirm this or identify other factors driving these increases in transmission (and similarly the decreases in 2014). We can also distinguish regions where prevalence appears to have returned to normal levels by the second half of 2016 from regions where increased transmission appears to have been sustained, such as the South and Highland fringe west ecozones. The latter may be potential targets for increased control efforts. Monthly prevalence estimates allowed us to evaluate seasonal trends in prevalence. These trends were largely similar across the country, with one seasonal peak occurring between February and April. This seasonality was clear in many of the high transmission areas but was also observed in highland ecozones, despite the low overall prevalence in these regions. A better understanding of baseline transmission patterns in low transmission settings may improve outbreak detection²⁶ or influence plans for moving towards pre-elimination and elimination in these areas.

Identifying areas of consistently low or high transmission may also influence operational planning. Currently ITN mass distribution campaigns (MDCs) are carried out everywhere except elimination and pre-elimination districts. The criteria for this stratification are annual incidence and test positivity rates (elimination districts are defined by an incidence rate of less than 1 per 1000 people and pre-elimination by an incidence rate between 1–10 and less than 5% test positivity)⁷. Use of monthly prevalence estimates could help to distinguish areas that have consistently low transmission (as identified in Fig. 6a) from low transmission areas with short but significant seasonal peaks which could benefit from more intervention. Similarly, highlighting areas with consistently high prevalence could inform the control methods used in these areas.

We can also consider the long term trends over the four years in the context of the MDCs which have taken place approximately every three years since 2009²⁸. The second MDC took place in November 2012 on the east coast and October 2013 in the rest of the country and the third from September to December 2015^{53,54}. Although continuous ITN distribution was carried out to supplement these campaigns, a decline in net coverage and effectiveness over time could be a factor in the overall increase in estimated burden up to the end of 2015.

When comparing our results to the MIS data, the prevalence estimates from the MIS reports of 9.1% in 2013⁹ and 7.0% in 2016¹⁰ are higher than our annual point estimates (6.4% and 6.6% respectively) but are consistent

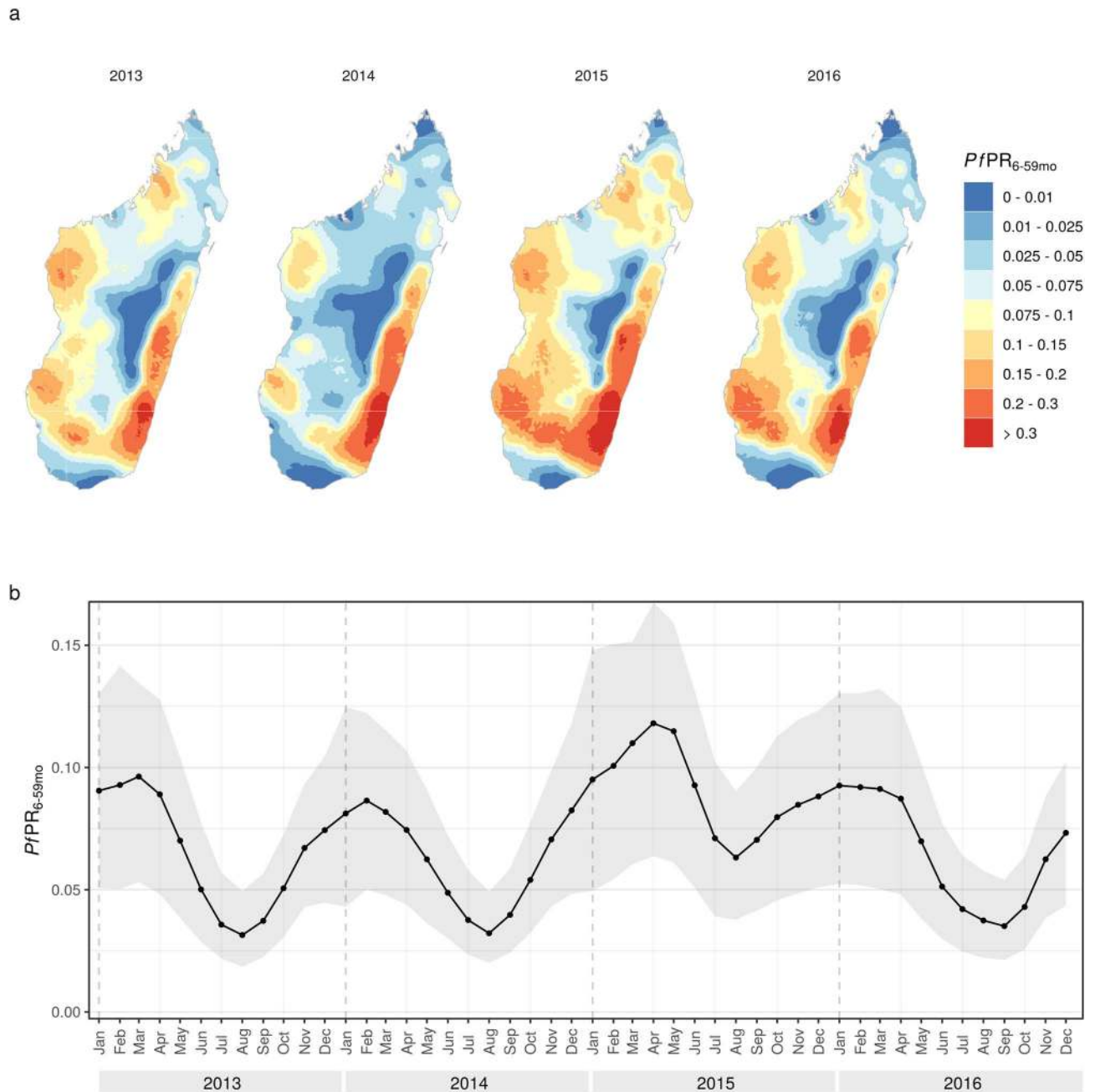


Figure 4. Prevalence estimates for individuals between 6 and 59 months of age, **(a)** aggregated annually and **(b)** population-weighted mean over time with 95% credible intervals. These plots were created in R (version 3.6.2, <https://www.r-project.org/>) using the ggplot2 package.

with our credible intervals. However, our results show limited evidence of a decrease in prevalence in 2016 compared to 2013. It appears likely that the decrease observed in the raw data is a consequence of the timing of the 2016 survey which primarily took place between May and July, around a month later than the 2013 survey and slightly past the seasonal peak in prevalence for the year. The annual 2013 and 2016 risk maps made by Kang et al.⁵ show a similar spatial pattern to our results, with low rates of prevalence in the highlands and south of the country and higher rates in the east and west coasts. However, Kang et al.⁵ estimate high prevalence in the north of the country, whereas in our results areas of higher prevalence are largely in the northeast. Similar to our results, their estimates show very similar prevalence in 2013 and 2016, although in general the overall mean estimated is slightly higher (9.3% in 2016) than our point estimates. We can also compare the monthly estimates of prevalence in each ecozone. While in many regions overall seasonal trends were largely similar, the monthly prevalence estimates made by Kang et al.⁵ show less consistent seasonal patterns with greater uncertainty. We would expect the monthly estimates in our analysis to be more robust, particularly outside of the survey months (April–June in 2013 and May–July in 2016) and in lower transmission areas, as there is more data to inform each month.

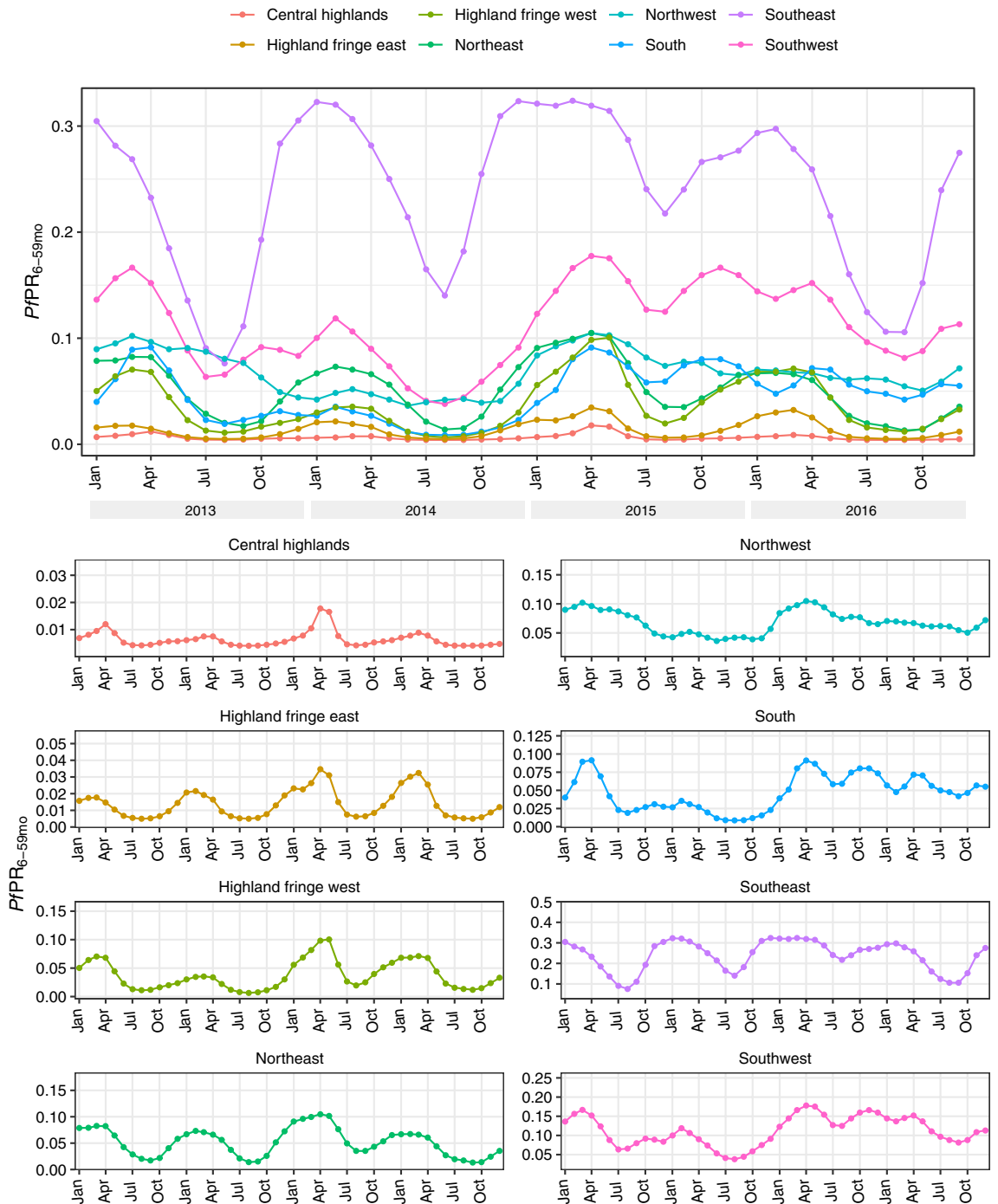


Figure 5. Population-weighted mean prevalence over time stratified by ecozone with 95% credible interval. These graphs were created in R (version 3.6.2, <https://www.r-project.org/>) using the ggplot2 package.

Comparing prevalence estimates to the routine case data, we can see that the temporal patterns in prevalence are typically smoothed trends from the case data (due to the spatial smoothing when producing the incidence surfaces and the temporal smoothing from using incidence at two time points to the predict prevalence). However, we can see distinct spatial trends in the prevalence estimates, for example prevalence was consistently highest in the Southeast ecozone in the whole study period despite similar or higher observed incidence rates in the Southwest ecozone from April 2015 onwards.

As well as visualising uncertainty using the interquartile range, we have used maps of exceedance (and non-exceedance) probabilities. In practice, the latter may be more interpretable and therefore more useful for communicating uncertainty. By choosing relevant thresholds, exceedance surfaces may be useful for identifying areas that are estimated to be high transmission with high confidence or areas that are most likely to be considered pre-elimination.

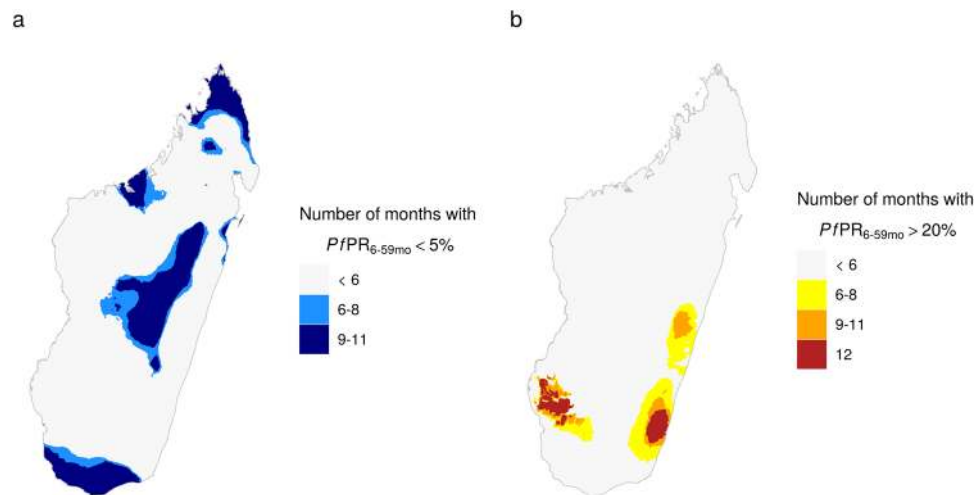


Figure 6. Number of months in 2016 with estimated prevalence (a) below 5% and (b) above 20%. These maps were created in R (version 3.6.2, <https://www.r-project.org/>) using the ggplot2 package.

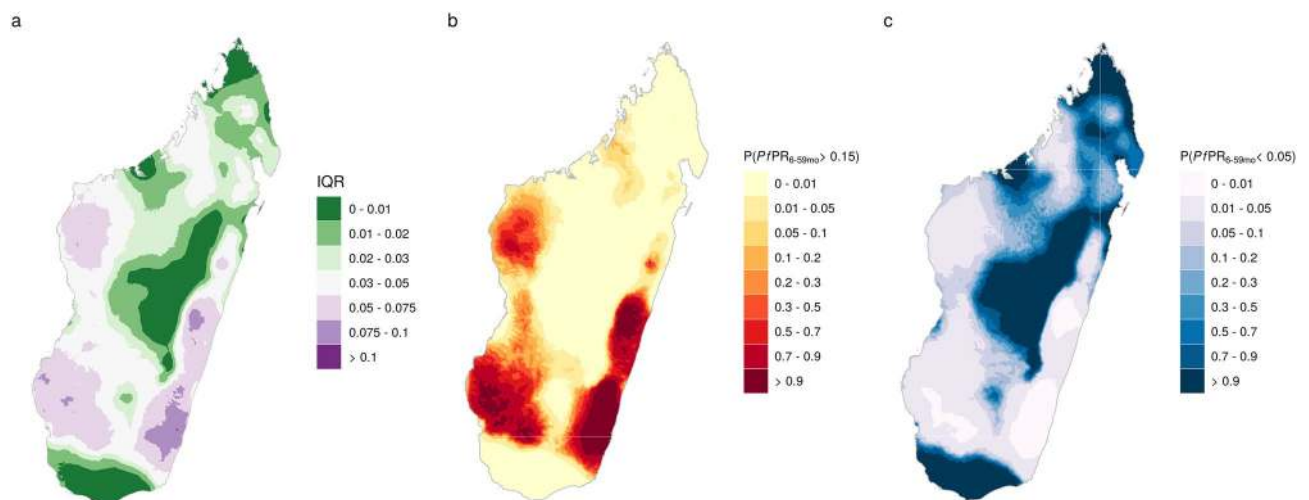


Figure 7. Uncertainty in 2016 annual prevalence estimates expressed with (a) interquartile range, (b) probability of prevalence exceeding 0.15 and (c) probability of prevalence not exceeding 0.05. These maps were created in R (version 3.6.2, <https://www.r-project.org/>) using the ggplot2 package.

Limitations. A key modelling input in this study was monthly incidence rates at each health facility, which depend on the estimated catchment populations. Our model for estimating these catchment populations is based only on travel time to each health facility and does not take into account other factors that may influence treatment-seeking or choice of facility, such as type of facility^{40,55}. However, the results of our analysis when using catchment population estimates from the NMCP and when varying treatment-seeking behaviour (see Supplementary Material Figs. 1 and 2) demonstrate that our modelling approach is fairly robust to differences in catchment population estimates. This is likely due to the spatial smoothing of incidence data and learned incidence-to-prevalence relationship. Treatment-seeking behaviour may also vary throughout the year, which is not accounted for in this analysis and may bias monthly estimates of prevalence. Modelling these temporal dynamics is extremely challenging, with limited treatment-seeking data available and behaviour that is likely to vary on a small spatial scale based on local infrastructure and geography.

A strength of our approach is the ability to account for bias in reporting of case information due to the learned relationship between prevalence and incidence. However, differences in reporting across the country cannot be accounted for in this relationship (which is constant across time and space). Howes et al.⁴ investigated the spatial variation in reporting in 2014 by looking at RDT stock-outs and proportion of distributed RDTs for which any result was reported to the HMIS by district. The proportion of RDT results reported was generally higher on the west coast, and therefore prevalence in this region may be somewhat overestimated, but across the rest of the country there were no strong spatial trends in reporting. Similarly, increases in reported cases over time may be partly due to increased access to healthcare or availability of RDTs, leading to overestimates of prevalence in more recent years, although the use of survey data in the first and last years of our study period should mitigate this.

Although the case data used in this study was for a different age range (all ages) to the estimates of prevalence (6–59 months), we believe this is unlikely to bias our estimates. There was a strong relationship between the number of cases in individuals under 5 and all ages in the routine case data (correlation of 0.91 across all health facilities and months) and the relationship between incidence and prevalence was learned within the model. Therefore we would expect the predictive power of incidence in either age range to be similar. Incidence for all ages was chosen due to the larger sample sizes in the data, which should produce more reliable estimates of incidence.

The aim of this study was to estimate prevalence, however estimates of incidence may be of more use for public health purposes (for example, district-level risk stratification by the NMCP is based on estimated incidence⁷). Future work could therefore focus on mapping incidence. A common approach for generating incidence estimates is to use prevalence estimates and an established prevalence–incidence relationship^{3,16}. This approach is generally used where prevalence estimates are informed only by prevalence survey data (where routine case data is unavailable or too unreliable to be used) but a similar transformation could be applied to the prevalence estimates produced here. A more direct way of combining routine case data and survey data to estimate incidence would be to use a joint model¹⁹. To do so successfully would likely require a more complete understanding of the spatiotemporal trends in treatment-seeking and reporting completeness in order to account for bias in the routine case data due to these factors.

Conclusion. In this study, we used routine case data and survey data to produce monthly estimates of malaria prevalence between 2013 and 2016. Our results suggest that while malaria endemicity was similar in 2013 and 2016, there was considerable variation in the intervening period, with a small decrease in 2014 followed by a substantial increase in 2015. In many areas, this increase in prevalence was sustained throughout 2015 and early 2016, and in the Northwest and South ecozones prevalence had yet to return to 2014 levels by the end of 2016. Considering these temporal trends at specific locations in relation to the timing of control efforts (such as ITN distribution) and climatic events may help to identify drivers of transmission and assess intervention effectiveness. Areas where prevalence remains higher than pre-2015 levels could be targets for increased control measures. Seasonality of transmission (with a single peak around March) was widely observed in high transmission areas and in some low transmission areas. The relative spatial patterns of prevalence were mostly consistent over time, with highest prevalence in the southwest and high prevalence in the southeast. Prevalence was consistently low in the highlands and in parts of the south of the country.

The Bayesian modelling approach applied here allowed us to make prevalence estimates with associated measures of uncertainty by learning temporal trends in transmission from the routine case data and calibrating these trends to prevalence observations from the survey data. These risk estimates could be an important tool for assessing the impact of control measures and the progress made towards the goals set out by the NMCP^{27,28} and for better understanding the drivers of changes in transmission. Our results demonstrate the considerable additional information that can be gained by combining data sources in this way. While previous risk mapping efforts produced monthly prevalence maps for 2011, 2013 and 2016⁵, these were informed by a relatively small number of prevalence observations from surveys which only covered 3 months of each year. We were able to make monthly maps of prevalence over a four year period (including years in which no surveys took place), informed by a large amount of monthly case data in addition to the survey data. By using a two-step modelling approach, rather than a joint likelihood model, we were able to learn the relationship between incidence and prevalence within the model (thereby accounting for systematic under reporting in the case data) and ensure that the model was informed by the more reliable prevalence data, despite the relatively small number of observations compared to the case data. Our sensitivity analysis demonstrated that these estimates are robust to varying assumptions about treatment-seeking behaviour and reporting incompleteness. A similar modelling strategy may be effective in other multi-metric disease modelling settings, especially where there is an imbalance between the number of observations or reliability of different metrics.

Incomplete reporting, treatment-seeking behaviour, and inconsistent standardisation of clinical diagnoses affect the quality of routine case data in many countries of sub-Saharan Africa⁵⁶. Consequently, the World Health Organization estimates for malaria burden are often based on community prevalence surveys alone^{1,24}. The methods presented here may be a useful starting point for incorporating routine case data into estimates of malaria burden more widely.

Data availability

Prevalence datasets, sample case data, and code are available at <https://github.com/rarambepola/Prevalence-Madagascar>. The raw case data that support the findings of this study are available from the Programme National de Lutte Contre le Paludisme de Madagascar and the Institut Pasteur de Madagascar (IPM).

Received: 25 August 2020; Accepted: 12 October 2020

Published online: 22 October 2020

References

1. WHO. *World Malaria Report 2019* (World Health Organization, Geneva, 2019).
2. Barmania, S. Madagascar's health challenges. *The Lancet* **386**, 729–730 (2015).
3. Bhatt, S. *et al.* The effect of malaria control on *Plasmodium falciparum* in Africa between 2000 and 2015. *Nature* **526**, 207–211 (2015).
4. Howes, R. E. *et al.* Contemporary epidemiological overview of malaria in Madagascar: operational utility of reported routine case data for malaria control planning. *Malaria J.* **15**, 502 (2016).

5. Kang, S. Y. *et al.* Spatio-temporal mapping of Madagascar's Malaria Indicator Survey results to assess *Plasmodium falciparum* endemicity trends between 2011 and 2016. *BMC Med.* **16**, 71 (2018).
6. Ihantamalala, F. A. *et al.* Spatial and temporal dynamics of malaria in madagascar. *Malaria J.* **17**, 58 (2018).
7. National Malaria Control Programme of Madagascar. Plan strategique national de lutte contre le paludisme: elimination progressive du paludisme à madagascar (2018).
8. Battle, K. E. *et al.* Treatment-seeking rates in malaria endemic countries. *Malaria J.* **15**, 20 (2016).
9. Institut National de la Statistique (INSTAT), Programme Nationale de lutte contre le Paludisme (PNLP), Institut Pasteur de Madagascar (IPM), and ICF International. *Madagascar Malaria Indicator Survey 2013 [Enquête sur les Indicateurs du Paludisme (EIPM)]* (INSTAT, PNL, IPM and ICF International, Calverton, 2013).
10. Institut National de la Statistique (INSTAT), Programme National de lutte contre le Paludisme (PNLP), Institut Pasteur de Madagascar (IPM), and ICF International. *Madagascar Malaria Indicator Survey 2016 [Enquête sur les Indicateurs du Paludisme (EIPM)]* (INSTAT, PNL, IPM and ICF International, Calverton, 2016).
11. Nguyen, M. *et al.* Mapping malaria seasonality in Madagascar using health facility data. *BMC Med.* **18**, 1–11 (2020).
12. Bennett, A. *et al.* A methodological framework for the improved use of routine health system data to evaluate national malaria control programs: evidence from Zambia. *Popul. Health Metrics* **12**, 30 (2014).
13. Chanda, E. *et al.* Impact assessment of malaria vector control using routine surveillance data in Zambia: implications for monitoring and evaluation. *Malaria J.* **11**, 437 (2012).
14. Institut National de la Statistique (INSTAT), Programme National de lutte contre le Paludisme (PNLP), and ICF International. *Madagascar Malaria Indicator Survey 2011 [Enquête sur les Indicateurs du Paludisme (EIPM)]* (INSTAT, PNL, IPM and ICF International, Calverton, 2012).
15. Battle, K. E. *et al.* Mapping the global endemicity and clinical burden of *Plasmodium vivax*, 2000–17: a spatial and temporal modelling study. *The Lancet* **394**, 332–343 (2019).
16. Weiss, D. J. *et al.* Mapping the global prevalence, incidence, and mortality of *Plasmodium falciparum*, 2000–17: a spatial and temporal modelling study. *The Lancet* **394**, 322–331 (2019).
17. Sturrock, H. J. *et al.* Mapping malaria risk in low transmission settings: challenges and opportunities. *Trends Parasitol.* **32**, 635–645 (2016).
18. Wang, C., Puhan, M. A., Furrer, R., Group, S. S. *et al.* Generalized spatial fusion model framework for joint analysis of point and areal data. *Spat. Stat.* **23**, 72–90 (2018).
19. Lucas, T. C. *et al.* Mapping malaria by sharing spatial information between incidence and prevalence datasets. *medRxiv* (2020).
20. Jacob, P. E., Murray, L. M., Holmes, C. C. & Robert, C. P. Better together? statistical learning in models made of modules. [arXiv:1708.08719](https://arxiv.org/abs/1708.08719) (2017).
21. Liu, F., Bayarri, M. & Berger, J. Modularization in bayesian analysis, with emphasis on analysis of computer models. *Bayesian Anal.* **4**, 119–150 (2009).
22. Cameron, E. *et al.* Defining the relationship between infection prevalence and clinical incidence of *Plasmodium falciparum* malaria. *Nat. Commun.* **6**, 1–10 (2015).
23. Weiss, D. J. *et al.* Air temperature suitability for *Plasmodium falciparum* malaria transmission in Africa 2000–2012: a high-resolution spatiotemporal prediction. *Malaria J.* **13**, 171 (2014).
24. Weiss, D. J. *et al.* A global map of travel time to cities to assess inequalities in accessibility in 2015. *Nature* **553**, 333–336 (2018).
25. Lucas, T. C. *et al.* Improving disaggregation models of malaria incidence by ensembling non-linear models of prevalence. *Spat. Spatio-temp. Epidemiol.* **100357**, (2020).
26. Randrianasolo, L. *et al.* Sentinel surveillance system for early outbreak detection in Madagascar. *BMC Public Health* **10**, 31 (2010).
27. President's Malaria Initiative. Madagascar malaria operational plan financial year 2019 (2018).
28. President's Malaria Initiative. Madagascar malaria operational plan financial year 2018 (2017).
29. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer-Verlag, New York, 2016).
30. Elvidge, C. D., Baugh, K., Zhizhin, M., Hsu, F. C. & Ghosh, T. VIIRS night-time lights. *Int. J. Remote Sens.* **38**, 5860–5879 (2017).
31. Arambepola, R., Gething, P. & Cameron, E. Nonparametric causal feature selection for spatiotemporal risk mapping of malaria incidence in Madagascar. [arXiv:2001.07745](https://arxiv.org/abs/2001.07745) (2020).
32. Funk, C. C. *et al.* A quasi-global precipitation time series for drought monitoring. *US Geol. Surv. Data Ser.* **832**, 1–12 (2014).
33. NASA Earth Observations. Average land surface temperature. http://neo.sci.gsfc.nasa.gov/view.php?datasetId=MOD_LSTD_CLIM_M (2017). (Accessed Sept 2017).
34. NASA Earth Data. Land processes distributed active archive center. https://lpdaac.usgs.gov/dataset_discovery/modis/modis_products_table/mcd43b5 (2017). (Accessed Sept 2017).
35. NASA Earth Data. MODIS (MOD 13): Gridded vegetation indices (NDVI and EVI). http://modis.gsfc.nasa.gov/data/dataproduct/dataproducts.php?MOD_NUMBER=13 (2017). (Accessed Sept 2017).
36. Trabucco, A. & Zomer, R. J. Global aridity index (global-aridity) and global potential evapo-transpiration (global-pet) geospatial database. *CGIAR Consortium for Spatial Information* (2009).
37. Farr, T. G. *et al.* The shuttle radar topography mission. *Rev. Geophys.* **45**, (2007).
38. Lehner, B. & Döll, P. Global lakes and wetlands database glwd. *GLWD Documentation* (2004).
39. Dijkstra, E. W. A note on two problems in connexion with graphs. *Numer. Math.* **1**, 269–271 (1959).
40. Pach, A. *et al.* A qualitative study investigating experiences, perceptions, and healthcare system performance in relation to the surveillance of typhoid fever in Madagascar. *Clin. Infect. Dis.* **62**, S69–S75 (2016).
41. Alegana, V. A. *et al.* Spatial modelling of healthcare utilisation for treatment of fever in Namibia. *Int. J. Health Geogr.* **11**, 6 (2012).
42. Center for International Earth Science Information Network CIESIN Columbia University, N. S. D. & (SEDAC), A. C. Gridded population of the world, version 4 (gpwv4): Administrative unit center points with population estimates (2018).
43. Tatem, A. J. Worldpop, open data for spatial demography. *Sci. Data* **4**, 1–4. <https://doi.org/10.1038/sdata.2017.4> (2017).
44. Ministry of Health of Madagascar. Reference manual of principal health system indicators (in french). antananarivo: Ministry of health of madagascar (2014).
45. Lindgren, F., Rue, H. & Lindström, J. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *J. R. Stat. Soc. B* **73**, 423–498 (2011).
46. Kristensen, K., Nielsen, A., Berg, C., Skaug, H. & Bell, B. Template model builder TMB. *J. Stat. Softw.* **70**, 1–21 (2015).
47. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2019).
48. Guyon, I. & Aliferis, C. Causal feature selection. In *Computational Methods of Feature Selection*, 79–102 (Chapman and Hall/CRC, London, 2007).
49. Spirtes, P., Glymour, C. N., Scheines, R. & Heckerman, D. *Causation, Prediction, and Search* (MIT press, Cambridge, 2000).
50. Strobl, E. V., Zhang, K. & Visweswaran, S. Approximate kernel-based conditional independence tests for fast non-parametric causal discovery. *J. Causal Inference* **7**, (2019).
51. Schölkopf, B. *et al.* On causal and anticausal learning. [arXiv:1206.6471](https://arxiv.org/abs/1206.6471) (2012).
52. President's Malaria Initiative. Madagascar malaria operational plan financial year 2017 (2016).
53. President's Malaria Initiative. Madagascar malaria operational plan financial year 2013 (2012).
54. President's Malaria Initiative. Madagascar malaria operational plan financial year 2014 (2013).

55. Do, M. *et al.* Associations between malaria-related ideational factors and care-seeking behavior for fever among children under five in Mali, Nigeria, and Madagascar. *PLoS ONE* **13**, (2018).
56. Alegana, V. A., Okiro, E. A. & Snow, R. W. Routine data for malaria morbidity estimation in Africa: challenges and prospects. *BMC Med.* **18**, 1–13 (2020).

Acknowledgements

The authors thank Fanjasoa Rakotomanana and her team at Institut Pasteur de Madagascar for sharing their database of health facility geo-location data for validation of the NMCP data as previously described¹¹. The first author was supported in this work through an Engineering and Physical Sciences Research Council (EPSRC) (<https://epsrc.ukri.org/>) Systems Biology studentship award (EP/G03706X/1). Work by the Malaria Atlas Project on methods development for Malaria Eradication Metrics including this work is supported by a grant from the Bill and Melinda Gates Foundation (OPP1197730).

Author contributions

R.A., E.C. and P.W.G. conceived the study. R.A., E.L.C., S.H.K. and K.A.T. designed and carried out the analysis. E.L.C., S.H.K., M.A., S.R. and A.C.R. prepared the datasets. E.C., P.W.G. and R.E.H. advised on the analysis. J.H., J.R., C.V. and E.G.C. prepared the covariate data. R.A. wrote the manuscript. All authors (all previously mentioned and P.A., A.B., J.M., S.F.R., T.L.S., D.J.W.) contributed to the interpretation of results. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41598-020-75189-0>.

Correspondence and requests for materials should be addressed to R.A.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020