npj | Digital Medicine

www.nature.com/npjdigitalmed

ARTICLE OPEN

Precision identification of high-risk phenotypes and progression pathways in severe malaria without requiring longitudinal data

Iain G. Johnston[1,2,3], Till Hoffmann[4], Sam F. Greenbury[2,4], Ornella Cominetti[5], Muminatou Jallow[6], Dominic Kwiatkowski[7], Mauricio Barahona[2,4], Nick S. Jones[2,4] and Climent Casals-Pascual[7,8]

More than 400,000 deaths from severe malaria (SM) are reported every year, mainly in African children. The diversity of clinical presentations associated with SM indicates important differences in disease pathogenesis that require specific treatment, and this clinical heterogeneity of SM remains poorly understood. Here, we apply tools from machine learning and model-based inference to harness large-scale data and dissect the heterogeneity in patterns of clinical features associated with SM in 2904 Gambian children admitted to hospital with malaria. This quantitative analysis reveals features predicting the severity of individual patient outcomes, and the dynamic pathways of SM progression, notably inferred without requiring longitudinal observations. Bayesian inference of these pathways allows us assign quantitative mortality risks to individual patients. By independently surveying expert practitioners, we show that this data-driven approach agrees with and expands the current state of knowledge on malaria progression, while simultaneously providing a data-supported framework for predicting clinical risk.

npj Digital Medicine (2019)2:63 ; https://doi.org/10.1038/s41746-019-0140-y

INTRODUCTION

Severe malaria (SM) is a major public health problem and a complex disease. Worldwide, 3.3 billion people live in areas where malaria is transmitted by infected anopheline mosquitoes. Despite recent improvements in the implementation of effective control measures in some countries, an estimated 216 million clinical malaria cases and 445,000 deaths were reported in 2016, with deaths mainly occurring in sub-Saharan Africa.[1]

The definition of SM proposed by the World Health Organization (WHO) was designed to capture the majority of children at risk of dying, and thus it prioritises sensitivity over specificity. In sub-Saharan Africa, children with coma (cerebral malaria) and/or respiratory distress (RD) are at the highest risk of death. However, these clinical syndromes encapsulate a heterogeneous population, and possibly reflect diverse pathophysiological processes. Critically, the current WHO classification of SM fails to capture this heterogeneity, and so treatment allocation based on this definition may have undesired consequences. Indeed, most adjuvant treatments proposed to date have consistently failed to improve patient outcome, and some of these treatments have been shown to increase mortality in children.[2–4]

The sequence of events leading to SM is poorly understood. A major determinant of death is the time elapsed from the initial symptoms to clinical presentation, with most deaths taking place within 24 h of admission.[5–7] Typically, clinical studies rarely capture the temporal component of the infection; accordingly, the natural history of the disease is inferred from experimental models, even though findings from these models are not always easily translated to human malaria.[8]

The explosion of data generation across the biomedical sciences coupled with advances in mathematical and computational tools provide the unprecedented opportunity to learn about these poorly understood dynamics.[9] Quantitative approaches leveraging these large datasets can be used to reveal progression pathways and learn features correlated with patient outcomes. Such approaches are central to the ongoing goal of "precision medicine", where clinical protocols are optimally tailored to the individuals, or subgroups of individuals, under consideration. However, the heterogeneity and the large scale of biomedical data, including data on malaria progression, poses challenges to quantitative approaches. Typically, the identification of prognostic factors is based on generalised linear models that use a set of features as independent variables. However, the complexity of these models increases dramatically when interactions between factors are important, and generalised linear models lack the ability to naturally dissect dynamic data.

Here, we address these issues by undertaking two complementary analyses to exploit a large dataset on clinical malaria presentation. The first uses mutual information (MI) to learn clinical factors predictive of patient outcomes. The second uses the recently developed HyperTraPS (hypercubic transition path sampling) algorithm,[10] to learn dynamic probabilistic pathways of disease progression. This dual approach provides the advantages of both data-driven analysis and model-based inference. MI

[1]Faculty of Mathematics and Natural Sciences, University of Bergen, Bergen, Norway; [2]EPSRC Centre for the Mathematics of Precision Healthcare, Imperial College London, London, UK; [3]Alan Turing Institute, London, UK; [4]Department of Mathematics, Imperial College London, London, UK; [5]Nestlé Institute of Health Sciences, Lausanne, Switzerland; [6]Medical Research Council Unit, The Gambia, Fajara, P.O. Box 273, Banjul, The Gambia; [7]Wellcome Trust Centre for Human Genetics, Oxford, UK and [8]Hospital Clinic i Provincial de Barcelona, CDB and ISGlobal, Barcelona, Spain
Correspondence: Nick S. Jones (nick.jones@imperial.ac.uk) or Climent Casals-Pascual (ccasals@well.ox.ac.uk)
These authors contributed equally: Iain G. Johnston, Till Hoffmann; Nick S. Jones, Climent Casals-Pascual

Received: 28 February 2019 Accepted: 6 June 2019
Published online: 10 July 2019

approaches are more robust regarding nonlinearities in relationships and do not suffer from the shortcomings of log odds ratios (LORs) associated with linear regression. HyperTraPS allows the Bayesian inference of dynamic pathways describing the successive acquisition of features or symptoms directly from cross-sectional (or longitudinal) observations. This method has been recently applied to elucidate the dynamics and mechanisms underlying the evolution of mtDNA genome structure[10] and efficient photosynthesis.[11]

We proceed by applying independent MI and HyperTraPS approaches to identify prognostic factors and to infer the sequence of events from cross-sectional data in patients with SM. We underline that the approaches we describe are not specific to SM and have general applicability to the study of disease progression. We compare the results of both our data-driven analysis and inference approaches and a survey of expert opinion on malarial progression, and demonstrate how these quantitative methods can be used to make predictions in precision medicine approaches for patient stratification.

RESULTS

The original dataset includes 2915 patients with severe *Plasmodium falciparum* malaria.[12] Of these, we use the 2904 children with available outcome data, of which 387 deaths were recorded (case-fatality rate: 13.3%). Clinical features of disease severity were used to classify patients into three overlapping clinical categories: 1166 children had respiratory distress (RD) (40.2%); 1060 (36.5%) had cerebral malaria (CM); and 659 (22.7%) had severe anaemia.

Features informative of death beyond the WHO classification

We used mutual information (MI) to seek the features that most strongly correlated with mortality in our datasets. We employed the following iterative procedure (see Methods): we identified the feature that best predicts mortality for a set of observations and split the dataset according to that feature; we then sought the next feature that best predicts mortality. The algorithm stops when no further statistically robust connections with mortality can be identified. In this way, the stratification afforded by this iterative approach reveals the features most strongly linked to mortality for any given combination of clinical presentations.

Using this approach we found that presence of CM was the strongest initial predictor of mortality (CM in Fig. 1). This presence is classified in the clinic via the Blantyre coma score (BCS), which ranges from 5 (fully conscious) to 0 (not responsive); if BCS ≤ 2 with any parasitic presence, CM is diagnosed. Splitting the cohort into cerebral versus non-CM cases, we found that the presence of RD was the next most predictive feature of mortality in both cases. Further iterations of the MI evaluation revealed that abnormal posturing, absence of transfusion and lack of splenomegaly were robust predictors of mortality.

Sequence of events leading to severe malaria

To elucidate the pathways by which malaria progresses, we next used HyperTraPS (hypercubic transition path sampling), an algorithm for inferring the dynamics by which traits are acquired or lost over time.[10] HyperTraPS uses a general model of all possible transitions that can occur between states in a system, and uses data to refine estimates of the probability of each transition. This model-based inference method, identifying dynamic orderings and couplings, provides a natural complement to the above data-driven MI-based analysis, which identifies how outcomes depend on features. The clinical features in our dataset comprise a collection of true/false and ordinal features (Supplementary Table 1; Fig 2). In the case of true/false features, "true" reflects feature presence and "false" feature absence (for example, "Is the patient coughing?"). In some cases (see Methods) these

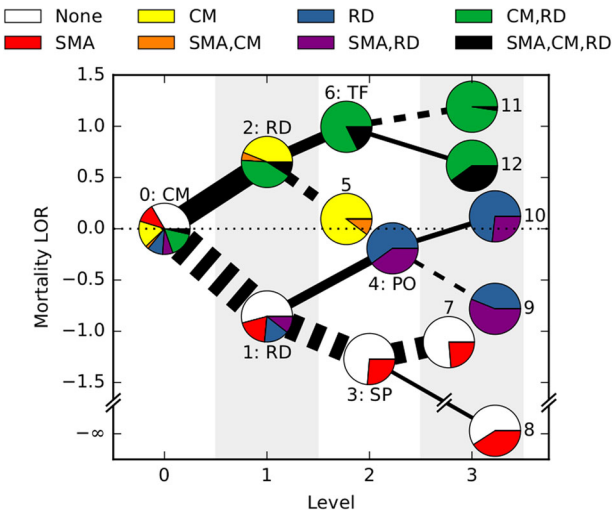
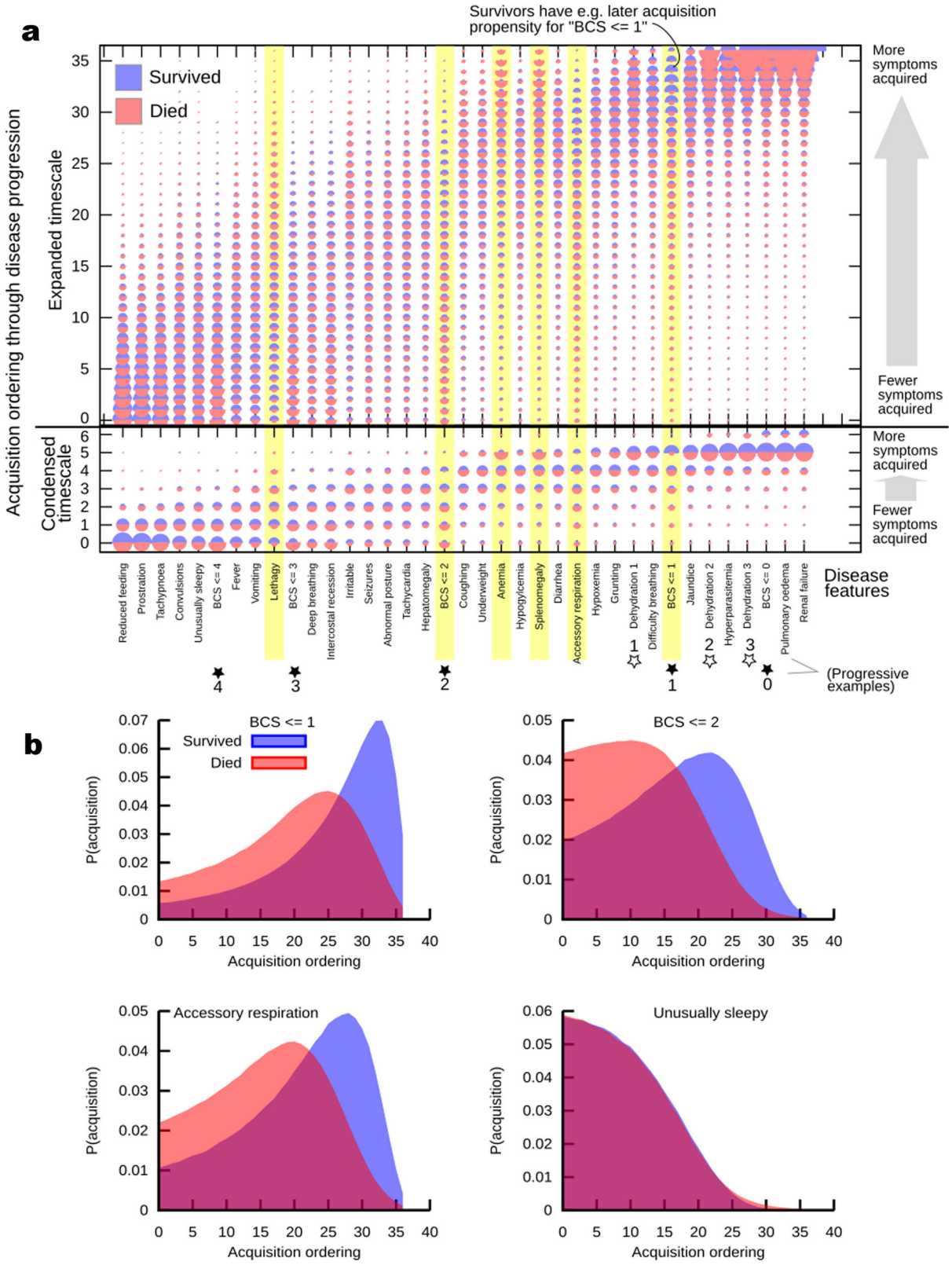


Fig. 1 Mutual information approach to identify features predicting mortality. At each level (horizontal axis), patient data are greedily split into two subsets according to the remaining feature that most strongly predicts mortality. The algorithm stops when no feature is statistically significantly associated with death. The figure shows a tree generated by this algorithm: cerebral malaria (CM), respiratory distress (RD), splenomegaly (SP), abnormal posturing (PO) and transfusion (TF) are selected as informative features. Nodes are shown as pie charts representing the composition of WHO classifications in each cluster. Solid (dashed) edges indicate that the feature was present (absent) and their width is proportional to the number of patients. The vertical axis corresponds to the mortality log odds ratio compared with the average mortality. Partition 8 has infinite log odds ratio (LOR) because all patients survive

observations correspond to a quantitative threshold; for example, hyperparasitaemia is defined as over 2.5×10^5 parasites/μL blood. Regarding ordinal features, we consider a set of inequalities reporting the severity of the condition. For the 'dehydration' feature, the score runs from 0 to 3: we assign separate features describing when this score is ≥1, ≥2, and =3. For consciousness we use the BCS score: we assign separate features describing when this score is ≤4, ≤3, ≤2, ≤1, and =0. Importantly, we take into account that these inequality features are not independent, i.e., a BCS ≤ 2 is also ≤4. In consequence, we expect to see an ordering in any posteriors on these features.

Two features in our dataset were treated differently: death and transfusion. We used death to split the dataset into those patients that survived and those that died: these different sets are kept separate in the analyses to identify differences in the pathways associated with these different outcomes. Since transfusion is an extrinsic intervention rather than an intrinsic feature of disease progression, we removed this feature from the dataset for the HyperTraPS analysis of disease progression (if transfusion is retained in the dataset, inferred disease progression pathways are similar, and transfusion is identified as a significant discriminant between survivor and death pathways, as seen in Fig. S1). All other features in the original dataset were retained: the HyperTraPS analysis was not linked to the MI analysis above and constitutes an independent analysis approach treating all features a priori on the same footing. We computed posterior distributions over disease progression pathways separately for cases where the patient survived and cases in which the patient died. These posteriors are summarised in Fig. 2.

Intuitively, the inferred orderings for those features that are themselves progressive (decreasing BCS and increasing dehydration indices) match the expected disease course for an initially healthy patient whose condition worsens over time. Substantial separation of features is observed, with refusal to feed and



tachypnoea among the earliest-onset features, and renal failure and pulmonary oedema among the latest. The posteriors for some features (dehydration, BCS, refusal to feed, jaundice and hyperparasitaemia) are more tightly constrained around given orderings than other features, suggesting varying flexibility amongst clinical features in the ordering in which they are manifested during disease progression.

Several clinical features notably discriminate the live and dead outcome groups in Fig. 2a. The dead cohort exhibits substantially earlier onset of low BCS, consistent with the predictive power of

Fig. 2 Inferring the pathways of malarial disease progression with HyperTraPS. **a** The HyperTraPS algorithm (see text) was used to infer the ordering with which malarial symptoms are likely acquired across patients. Horizontal axis records symptoms; vertical axis records ordering from low (early acquisition) to high (late acquisition). This ordering axis is grouped into seven longer "ordering windows" in the lower subsection of the figure, to display broader trends in addition to specific features of the dynamics. The size of a semicircle denotes the posterior probability that a given symptom is acquired at a given ordering in progression of malaria. Red semicircles are posteriors from the dataset of patients who died; blue semicircles inferred from patients who lived. Highlighted symptoms display a greater Kolmogorov–Smirnov distance between posteriors from survival and death pathways than between either posterior and the uninformative prior, forming potentially diagnostic features. **b** Posterior distributions on ordering for three features that differentiate between patients that eventually die and those that eventually survive, and for one that does not discriminate

these features identified in the above MI analysis (where CM, defined as BCS ≤ 2, was the strongest initial predictor of mortality). The dead cohort also shows later inferred ordering for anaemia, and earlier accessory muscle use and deep breathing. The posteriors for all these features were notably different between living and dead data classes, with greater Kolmogorov–Smirnov distances between posteriors from survival and death pathways than between either posterior and the uninformative prior (Fig. 2b). The congruence between the features identified through this pathway analysis and those identified through the above MI approach suggest that this novel inference approach reveals robust aspects of disease progression, which we test further below.

Validation of data-driven inference with independent expert survey

To validate our findings on disease progression, we surveyed 11 clinical practitioners in the field of malaria to build a consensus picture of clinical perceptions of malaria progression. We asked respondents to score each of the 25 features from our dataset as 1-2-3 (early; intermediate; late) according to their clinical perception of when a particular symptom was most likely to appear during disease progression. We then compared the mean response computed from the survey data to the mean orderings inferred through HyperTraPS. We found a statistically significant correlation between the clinicians' views and the orderings arising from computational inference from the full dataset (Fig. 3, $R = 0.518$, $p = 0.008$), the strength of which was particularly notable given the range of expert opinions for several features (error bars in Fig. 3). The views from the survey correlated better with inferred results from the subset of patients who lived ($R = 0.521$) than with those from patients who died ($R = 0.399$). The largest discrepancies between the inferred and survey orderings include fever, tachycardia and coughing (which the survey ranks earlier than the inference) and prostration, convulsions and coma (which the survey ranks later than the inference). The raw data for these features, in the form of frequency counts for each feature across patients, tend to support the inference picture: for example, prostration is seen to be an extremely common feature and coughing often occurs only in patients with several other symptoms (Fig. 3).

Using inferred pathway data to estimate hidden features

Our computed probabilistic description of the pathways underlying malaria progression allows us to predict unobserved features in particular patients. More precisely, by learning the structure of, and variability in, probabilistic pathways of disease progression, we can, in effect, build a set of probabilistic statements reporting the chance that an unobserved symptom is present or absent in a given patient, contingent on the presentation of other symptoms in light of the inferred ordering and relationship between symptoms. For example, the symptom 'fever' is inferred to occur consistently earlier than 'hyperparasitaemia'; therefore, a patient whose fever status has not been observed, but who has been observed to present hyperparasitaemia, will be predicted to likely

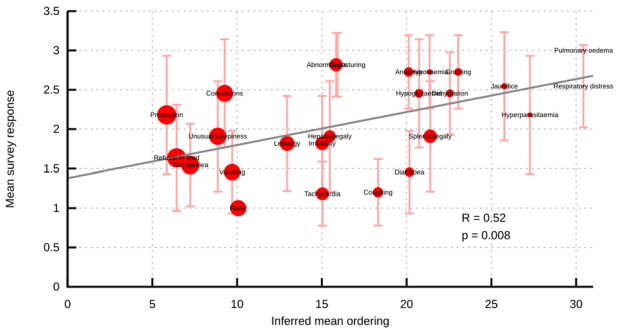


Fig. 3 Comparison of inferred disease progression results with expert survey. Horizontal axis gives the mean ordering of a symptom's acquisition from HyperTraPS inference results; vertical axis gives the mean ordering of that symptom resulting from a survey of expert opinions (see text). The size of each circle is proportional to the frequency with which that feature is "present" when observed in the dataset: small circles are rarely observed, large circles commonly so. Vertical error bars correspond to the standard deviation of expert responses for a given feature, illustrating the substantial range of opinions across our surveyed experts

(but not certainly) have already acquired a fever. The pathways inferred by HyperTraPS generalise this simple two-feature picture to allow predictions based on the presence or absence of all features.

To demonstrate this predictive capability, we tested our algorithm on a randomly sampled subset of half of the patients from our dataset, learning the posterior distributions describing ordering of disease progression as in Fig. 4. We then artificially obscured 10% of the features in a random subset of 1000 patients from the remaining (unseen) dataset and attempted to predict the values of these obscured features in different patients (Fig. 4a). Of 3318 features artificially obscured in the test dataset, we obtained predictions with 75% confidence for 1330 features (40%). Overall 83% of these predictions were correct, reflecting successes in predicting both presence and absence of obscured features. By comparison, a predictor only using the frequency counts of feature incidence in the dataset resulted in a 68% success rate using the same protocol.

Using inferred pathway data to predict future progression

Another, potentially more clinically valuable, mode of prediction facilitated by our pathway analysis is the likely next step in a disease progression pathway associated with an individual patient. Given an inferred set of likely progression pathways and a (possibly incomplete) observation of patient symptoms, we can integrate over all the possible states that could give rise to that observation and compute the probabilities with which each unacquired symptom will be the next step in a progression trajectory. Importantly, this approach may help target therapies to address the next most important stage in progression of the disease in individuals. To demonstrate this process, we used the learned pathways of disease progression to predict the likely next symptoms to become manifest as the disease progresses for a

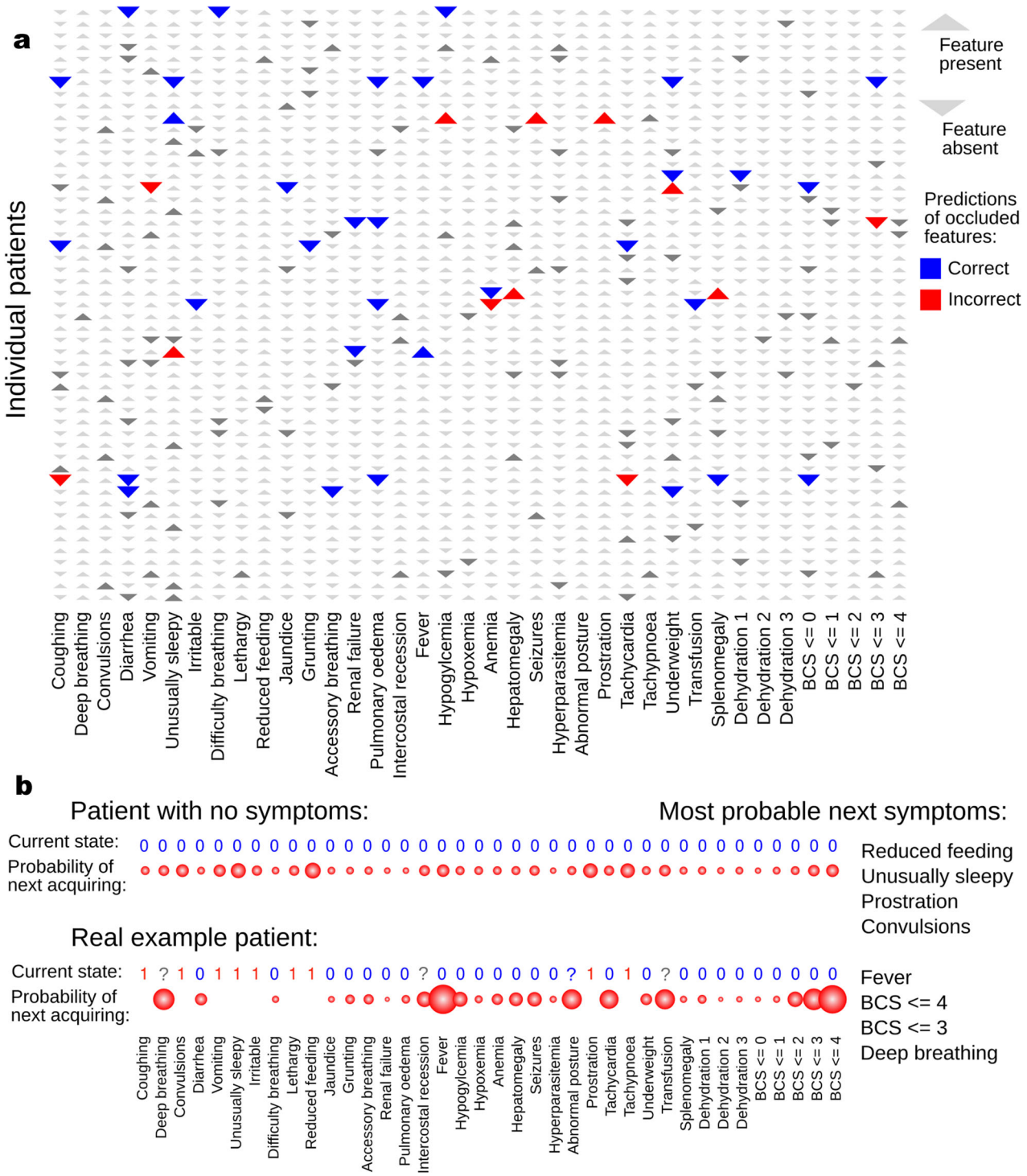


Fig. 4 Prediction and validation of hidden patient symptoms using HyperTraPS. **a** Rows correspond to an illustrative subset of individual patients; columns give different observed symptoms. Upwards triangles denote feature presence, downwards triangles denote absence. A random subset of features was artificially hidden, and the prediction algorithm using HyperTraPS posteriors described in the text was then applied to predict the presence or absence of these features given the remaining features (small grey triangles). Blue triangles denote correct predictions; red denote incorrect predictions; large grey triangles give instance where no strong prediction was available. Overall 83% of predictions (1104 of 1330) were accurate. **b** Illustration of prediction of likely next steps in disease progression for a given patient. Starting from any given patient state, HyperTraPS posteriors give the probability that any symptom is the next to be acquired by that patient. Circles represent the probability that each symptom will be acquired next, in two cases: a patient with no symptoms, illustrating the agreement with Fig. 2, and a real patient taken from the dataset. In both cases the four most likely next symptoms are given on the right

given sampled patient observation (Fig. 4b). Due to the single-point nature of our dataset, we cannot use it to verify these predictions but we include this approach here as a demonstration for validation and a testable suite of hypotheses for future clinical studies.

Using inferred pathway data to classify high-risk patients

HyperTraPS returns posterior distributions on the orderings of a progressive process, distinguishing events likely to occur earlier from those likely to occur later. The posteriors can be used to compute the probability that a patient with a given set of symptoms is on a high-risk disease progression pathway predicted to end in mortality, or on a lower-risk pathway predicted to end in survival.

To do so, we use Bayes' theorem:

$$P(\text{survivor pathway}|\text{patient data}) = P(\text{patient data}|\text{survivor pathway}) \\ P(\text{survivor pathway}) \,/\, P(\text{patient data}) \quad (1)$$

to compute the likelihood associated with observing a given patient's symptoms from the lower-risk (survivor) posterior parameterisations. There is an equivalent computation for the high-risk (dead) posterior (Fig. 5a).

We tested this approach by first assigning equal prior weight to both pathways (P(survivor pathway) = P(dead pathway) = 0.5) and then computing estimates of P(patient data|survivor pathway) and P(patient data|dead pathway) for 50 randomly chosen patients from each of the living and dead datasets (Fig. 5b). If one predicted pathway was at least 1.5 times more likely than the other, we recorded this prediction as the more likely pathway for that patient. Of the 50 randomly chosen patients from the dead dataset, 23 had associated predictions: 22 were correctly predicted as high risk and 1 was incorrectly predicted as low risk. Of the 50 randomly chosen patients from the living dataset, 25 had associated predictions: 15 were correctly classified as low risk and 10 were misclassified as high risk. The risk of false negative high-risk classification from this approach is thus low, although a risk of false positive high-risk classification remains.

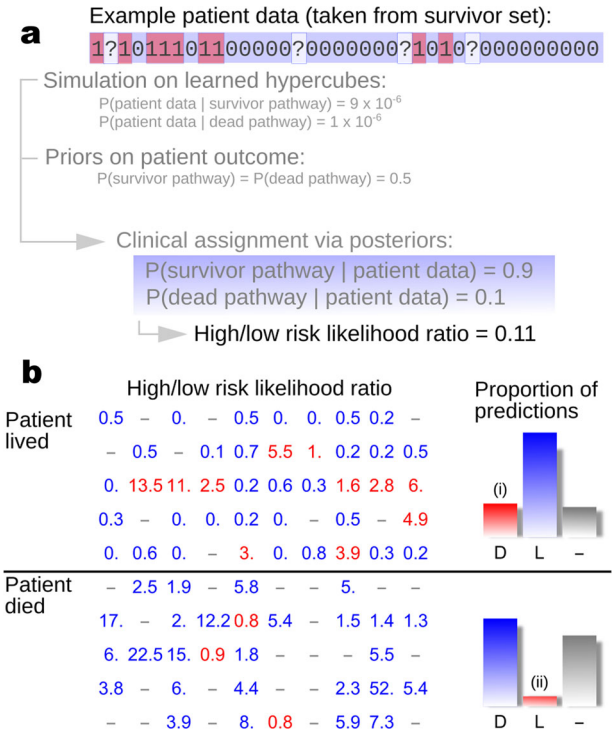


Fig. 5 Bayesian classification of patient risk. **a** Pipeline for applying Bayes' theorem and simulation on the learned dynamics of surviving and dead patients to classify risk of new patients. **b** A test dataset of 50 patients that died and 50 patients that survived was analysed using posterior distributions for disease progression pathways derived from a separate training dataset. Figures give the likelihood ratio of a given patient being on a high-risk trajectory to that patient being on a low-risk trajectory, used to classify patients into high and low risk classes. Blue figures show where this classification aligns with the true patient outcome; red figures show where this classification does not align with patient outcome; dashes indicate cases where a classification was not available. Bars show the proportions of correct (blue) to incorrect (red) classifications. Overall 81% of classifications (57 of 70) were successful; false positive identification rate of high-risk patients (i) is 20%, and false negative identification of high-risk patients (ii) is 6%

DISCUSSION

In this study, we have independently used mutual information (MI) and Bayesian inference of dynamic disease pathways with HyperTraPS to identify the most informative clinical features associated with death in Gambian children with SM, and the sequence of appearance of these features. This inference process reflects a novel way of analysing disease datasets, using cross-sectional data to learn the likely dynamics of disease progression. This is possible because the underlying model represents individual patient data as being sampled from that patient's disease progression trajectory from a healthy initial state through the progressive acquisition of disease symptoms. By analysing many patient samples together and placing them all in the same probabilistic framework for disease progression, we can thus use single-point observations to characterise the structure of, and variability in, progression pathways across a population. The value of this powerful approach is clear: we can simultaneously learn the dynamic pathways of disease progression, identify key predictors of clinical outcome, and use this unprecedented elucidation of disease dynamics to facilitate novel and clinically informative classification of the clinical risk associated with individual patients.

MI identified the Blantyre coma score (BCS) as the most informative clinical feature associated with death (a coma score of 2 or less defines cerebral malaria [CM],[1,5–7,13] as in Fig. 1). A major prognostic feature of SM is impaired consciousness, a feature which is adequately captured by the BCS. We and others have previously reported the correlation of the coma score and the odds of death. The absence of CM reduced the odds of death significantly. The next feature identified, respiratory distress (RD), increased the odds of death in patients with and without CM. In this model, both CM alone and CM in combination with RD accounted for the majority of cases with increased mortality. The administration of a blood transfusion was identified as the next informative prognostic feature in patients with CM and RD. Blood transfusion appeared to reduce mortality in those patients presenting with the three SM syndromes (severe malarial anaemia [SMA], CM and RD). Interestingly, although the reduced odds of death were observed in patients with SMA, the majority of this group did not present with SMA. The potential benefit of transfusing patients with haemoglobin concentrations greater than 50 g/L is unclear and the WHO recommends blood transfusion for patients with haemoglobin concentrations up to 60 g/L only in presence of RD or impaired consciousness.[12] Our data suggest that patients with higher haemoglobin concentrations could benefit from blood transfusion and this seems particularly beneficial for patients presenting with CM and RD. Indeed, patients with CM and RD that received blood transfusions presented lower mortality than those not transfused, although their mean haemoglobin concentration (68.7 g/L) was above the recommended threshold for transfusion. Abnormal posturing is not frequently observed in non-comatose patients with SM.

However, we identified abnormal posturing as an informative feature in patients with SMA and RD. A plausible explanation is that these patients presented with advanced RD and oxygen deprivation, exhibiting thus symptoms associated with hypoxic brain dysfunction.

We also observed that clinical features associated with a better clinical outcome (below average mortality), showed a good correlation with the WHO classification. In particular, we found that presence of an enlarged spleen (splenomegaly) was associated with a better outcome in patients with SMA. This is likely related to the spleen's attempt to clear *P. falciparum* infected erythrocytes from blood, which probably reflects an adequate immune response.

In addition to, and independently of, identifying the most informative features that predict death, we used a Bayesian approach to infer the sequence of appearance of events that lead to death, providing the first posterior distributions on the progression dynamics of clinical symptoms in SM. This data-driven approach was validated by an independent survey of 11 experienced clinicians. Most clinicians who agreed to participate noted that they were uncertain about disease progression, as they commonly encountered only late-stage children in the clinical setting. There was general agreement among clinicians to score non-specific features such as fever, loss of appetite or prostration as early stage, whereas features of extreme severity, like deep coma, renal failure or jaundice were consistently placed as late features. Of note, some of the discrepancies observed between the clinical perception (survey) and the data-driven prediction can be reconciled. For example, the presence of fever or "reported fever" is scored as an early feature by clinicians since fever is the most likely guiding feature to suspect malaria. However, the data-driven algorithm may fail to capture this event since presence of fever is based on actual temperature recorded. Indeed, prior self-treatment with antipyretics (mostly paracetamol) can be as high as 50%.[14] On the other hand, neurological features such as coma and convulsions are commonly scored by clinicians as "late", since these life-threatening features usually appear when the infection has not been identified and treated promptly.[15] Conversely, data-driven inference tends to score neurological features as "early" since the algorithm is blind to the time elapsed between the onset of symptoms and presentation to hospital. Also, this study was biased towards recruitment of more severe cases of *P. falciparum* malaria. An apparent limitation of the validation survey was that clinicians were asked to score features as early, intermediate or late. However, despite clinician uncertainty and the limited sample size of the survey, the level of agreement between expert clinical intuition and our prediction was remarkable.

Notably, we can use the differences in inferred dynamics of fatal and non-fatal SM cases to classify new patients according to their inferred risk. This novel approach captures not just a snapshot of individual risk factors but the full probabilistic information about the learned pathways of disease progression, allowing the histories of previous patients to inform the clinical analysis of new patients. This approach, validated with a test dataset, aligns with the goals of precision medicine and makes full use of available biomedical data; we anticipate that it may also find use in numerous other diseases and clinical contexts.

The relevance of the analysis we present in this research is two-fold. Firstly, the prediction of clinical features associated with poor clinical outcome using MI validated prior findings and identified novel features. One of these features has potential translational impact and suggests the potential benefit of transfusing patients with higher haemoglobin concentrations beyond what is currently recommended by the WHO. These findings, however, must be validated in larger datasets and longitudinal studies. Secondly, we believe that the inferred sequence of events from a cross-sectional analysis is a novel and important approach. For ethical reasons, clinical studies are characteristically not suited to study or describe the natural course of a disease, since treatment must be administered as soon as the diagnosis has been made and a treatment option is available. By inferring the sequence of events from cross-sectional data, this approach provides new insights into the natural history of disease in the absence of longitudinal data.

METHODS

Study population

The study population consisted of 2915 children aged 4 months to 15 years diagnosed with SM according to the WHO definition. Children were admitted to the Royal Victoria Teaching Hospital (RVTH), Banjul, The Gambia from January 1997 to December 2009. The study was originally designed to study genetic variants associated with SM.[13] The initial set of variables used for feature selection included those present in the case report form. The list of the variables included is described in supplementary Table 1. A detailed description of the study population and clinical features associated with death has been published elsewhere.[12]

Clinical definitions

Children aged 4 months to 15 years were eligible for enrolment if they had a blood smear positive for asexual *P. falciparum* parasites and met one or more WHO criteria for SM[14]): Coma (assessed by the BCS [BCS][6]), severe anaemia (haemoglobin [Hb] < 50 g/L or packed cell volume [PCV] < 15), RD (costal indrawing, use of accessory muscles, nasal flaring, deep breathing), hypoglycaemia (<2.2 mM), decompensated shock (systolic blood pressure less than 70 mmHg), repeated convulsions (>3 during a 24-h period), acidosis (plasma bicarbonate <15 mmol/L) and hyperlactatemia (plasma lactate >5 mmol/L). CM was defined as a BCS of 2 or less with any *P. falciparum* parasite density. SMA was defined as haemoglobin under 50 g/L. Hepatomegaly was defined as > 2 cm of palpable liver below the right costal margin. Patients were enrolled in the study if written informed consent was given by the parent or guardian. The study protocol was approved by the Joint Gambia Government/MRC Ethical Committee (protocol numbers 630 and 670).

Laboratory measurements

Haemoglobin was measured with a haematology analyser (Coulter®MD II, Coulter Corporation, USA), and parasite density was counted on Giemsa-stained thick and thin films.

Data management and statistical analyses

The data were collected on standardised forms, double entered into a database and verified against the original.

Mutual information analysis

Data curation. We removed 11 patients from the dataset whose clinical outcome (death) was not known and a further 21 patients for which more than half the features were unobserved or missing. The missing features of the remaining patients were imputed using the k nearest neighbour algorithm. Firstly, we computed the Hamming distance between all pairs of patients, i.e. the distance between patients i and j is

$$Y_{ij} = \frac{1}{m_{ij}} \sum_{l=1}^{p} \left| X_{il} - X_{jl} \right|, \tag{2}$$

where X_{il} denotes feature l of patient i, p is the number of features, and m_{ij} is the number of features that are not missing for both patients. The clinical outcome was not included in the calculation of the distance matrix to avoid inducing artificial correlations between death and clinical features. Secondly, we set the missing features of patients equal to the median of the corresponding feature amongst the k nearest neighbours, i.e.

$$X_{il} = \mathrm{med}_{j \in N_{il}} X_{jl}, \tag{3}$$

where N_{il} is the set of k nearest neighbours of i for which feature l is not missing. We set $k = 13$, which is the square root of the number of complete cases.[15] Because k is odd, imputed values will also be binary.

Greedy, hierarchical partitioning of patients. We wanted to partition the patients into meaningful subsets to be able to predict clinical outcomes

and identify different presentations of SM. It was not feasible to consider a cross-tabulation of all available features because the number of cells would far exceed the number of patients. We thus considered a greedy algorithm that partitioned the patients into two subsets by the feature that was most predictive of death. The same step was recursively applied to the subsets until there were no more features that were predictive of death at a significance level $\alpha = 0.05$ after multiple hypothesis correction using the Holm–Bonferroni method.[16]

We used MI between death and one of the features as a measure of predictive power because MI naturally quantifies the reduction in uncertainty about death achieved by observing said feature. MI has distinct advantages over the log-odds ratio (LOR), which selects features with high sensitivity but low specificity: maximising the LOR is likely to identify rare features that have a major impact on mortality, whereas maximizing MI will minimise uncertainty about death. We used the plugin estimator

$$\hat{I}(A, B) = \sum_{a,b=0}^{1} \frac{n_{ab}}{n} \log\left(\frac{n_{ab} n}{n_a n_b}\right) \quad (4)$$

where n is the number of patients, n_{ab} is the number of patients in the cell ab of the contingency table, and n_a, n_b are the number of patients in the ath row and bth column, respectively. This is a biased estimator;[17] correction strategies exist but we ignored the bias of the plugin estimator because we were more interested in which feature is most predictive than in the exact value of the MI.

We assessed statistical significance using a bootstrap algorithm. Firstly, we computed the unbiased estimates of the marginal distributions of the feature under consideration and mortality. Under the null hypothesis that the feature is not predictive of death, the joint distribution q_{ab} is the product of the marginal distributions.

We drew 10^6 independent samples of the same size as the dataset from the distribution under the null hypothesis and computed the MI for each synthetic dataset. The probability that the MI was larger than the observed value was equal to the proportion of simulated values that exceed the empirical one.

HyperTraPS analyses

The HyperTraPS algorithm estimates the probability of observing a transition between two nodes a and b on a hypercubic transition graph, where edges are parameterised according to transition probabilities. The transitions we consider are between the state with no symptoms (assuming patients start healthy) and an observed set of symptoms in the dataset, corresponding to a particular time sample of a patient's trajectory through the space of possible symptom patterns. We use an $L \times L$ matrix to encode the transition probabilities of the L^2*(L-1)* edges on the hypercube, assuming that each feature has a base rate of acquisition (L parameters) and may influence the rate of acquisition of all other features (*L(L-1)* parameters). We learned posterior distributions on these parameters by assigning them uninformative uniform priors and embedding HyperTraPS in an MCMC auxiliary pseudo-marginal (APM) algorithm.[18] Simulation of trajectories using parameterisations from these posteriors give then directly posterior distributions on orderings of feature acquisitions. To make predictions of unobserved features, we recorded points where these simulated trajectories matched the known features of a given sample and recorded the value(s) of the unobserved trait(s) at each of these points, building a tally of presence vs absence. To make predictions of future behaviour, we simply report the probabilities of subsequent steps from a given point on transition networks parameterised by these posteriors.

Survey

The survey consisted of the question: "When do you expect to observe the following symptoms in the disease progression of malaria?" and the following features were listed: prostration; refusal to feed; convulsions; tachypnoea; abnormal posturing (tonic seizures); fever; vomiting; unusual sleepiness; lethargy; irritability; coma (BCS < = 2); tachycardia; hepatomegaly (>2 cm palpable liver); coughing; diarrhoea; grunting; anaemia (haemoglobin <50 g/L); splenomegaly (>2 cm palpable spleen); hypoglycaemia; jaundice; hyperparasitaemia (>250,000/μL); dehydration; hypoxaemia (oxygen saturation < 90%); respiratory distress (intercostal recession, lower chest indrawing, use of accessory respiratory muscles, nasal flaring, deep breathing); pulmonary oedema. Responses were returned between 16/01/2016 and 21/01/2016.

Reporting Summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

DATA AVAILABILITY

This study did not generate any new data; the source data used is referenced in the text.

CODE AVAILABILITY

HyperTraPS code is freely available on GitHub at https://github.com/sgreenbury/HyperTraPS/.

ACKNOWLEDGEMENTS

To the study participants and their parents/guardians. To the Royal Victoria Teaching Hospital (RVTH) nurses and fieldworkers: Yaya Dibba, Anthony Mendy and Abdoulie Camara; Senior laboratory technicians: Janet Riddle-Fullah and Abdou Bah; and Jalimory Njie (data entry clerk), Emmanuel Onykwelu (clinician), Augustine Ebonyi (clinician) and sister Haddy Njie. To MRC Laboratory technicians and assistants: Idrissa Sambou, Simon Correa, Madi Njie and Omar Janha. To Haddy Kanyi (data supervisor) and Mamkumba Sanneh (MRC Malaria Programme administrator). MalariaGEN's primary funding is from the Wellcome Trust (grant number 077383/Z/05/Z) and from the Bill & Melinda Gates Foundation, through the Foundation for the National Institutes of Health (grant number 566) as part of the Grand Challenges in Global Health initiative. C.C.P is supported by the Medical Research Council (Clinician Scientist Fellowship: G0701885). I.J., S.F.G., M.B., N.J. acknowledge grant EP/N014529/1. I.J. acknowledges funding from the Alan Turing Institute via a Turing Fellowship.

AUTHOR CONTRIBUTIONS

C.C.P. and N.S.J. had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. Study concept and design: I.G.J., T.H., N.S.J. and C.C.P. Acquisition of data: M.J., D.K. and CCP. Analysis and interpretation of data: T.H., I.G.J., S.F.G., O.C., M.B., N.S.J. and C.C.P. Critical revision of the manuscript for important intellectual content: T.H., I.J., M.B., N.S.J. and C.C.P. Data analysis: T.H., I.G.J., and S.F.G. Administrative, technical or material support: M.J. and D.K. Study supervision: I.G.J., N.S.J. and C.C.P.

ADDITIONAL INFORMATION

Supplementary information accompanies the paper on the *npj Digital Medicine* website (https://doi.org/10.1038/s41746-019-0140-y).

Competing interests: The authors declare no competing interests.

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

REFERENCES

1. World Health Organisation. *World Malaria Report 2017*. (World Health Organisation, Luxembourg, 2017).
2. Maitland, K. et al. Mortality after fluid bolus in African children with severe infection. *N. Engl. J. Med* **364**, 2483–2495 (2011).
3. Warrell, D. A., White, N. J. & Warrell, M. J. Dexamethasone deleterious in cerebral malaria. *Br. Med. J.* **285**, 1652 (1982).
4. White, N. J. Not much progress in treatment of cerebral malaria. *Lancet* **352**, 594–595 (1998).
5. Marsh, K. et al. Indicators of life-threatening malaria in African children. *N. Engl. J. Med.* **332**, 1399–1404 (1995).
6. Molyneux, M. E., Taylor, T. E., Wirima, J. J. & Borgstein, A. Clinical features and prognostic indicators in paediatric cerebral malaria: a study of 131 comatose Malawian children. *Q. J. Med.* **71**, 441–459 (1989).
7. Waller, D. et al. Clinical features and outcome of severe malaria in Gambian children. *Clin. Infect. Dis.* **21**, 577–587 (1995).
8. White, N. J., Turner, G. D., Medana, I. M., Dondorp, A. M. & Day, N. P. The murine cerebral malaria phenomenon. *Trends Parasitol.* **26**, 11–15 (2010).
9. Colijn, C., Jones, N. S., Johnston, I. G., Yaliraki, S. & Barahona, M. Toward precision healthcare: context and mathematical challenges. *Front Physiol.* **8**, 136 (2017).

10. Johnston, I. G. & Williams, B. P. Evolutionary inference across eukaryotes identifies specific pressures favoring mitochondrial gene retention. *Cell Syst.* **2**, 101–111 (2016).
11. Williams, B. P., Johnston, I. G., Covshoff, S. & Hibberd, J. M. Phenotypic landscape inference reveals multiple evolutionary paths to C4 photosynthesis. *eLife* **2**, e00961 (2013).
12. Jallow, M. et al. Clinical features of severe malaria associated with death: a 13-year observational study in the Gambia. *PLoS ONE* **7**, e45645 (2012).
13. Jallow, M. et al. Genome-wide and fine-resolution association analysis of malaria in West Africa. *Nat. Genet.* **41**, 657–665 (2009).
14. World Health Organization, Communicable Diseases Cluster. Severe falciparum malaria. *Trans. R. Soc. Trop. Med. Hyg.* **94**(Suppl 1), S1–S90 (2000).
15. Duda, R. O., Hart, P. E. & Stork, D. G. *Pattern Classification and Scene Analysis.* (Wiley, New York, 1973).
16. Holm, S. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **6**, 65–70 (1979).
17. Panzeri, S., Senatore, R., Montemurro, M. A. & Petersen, R. S. Correcting for the sampling bias problem in spike train information measures. *J. Neurophysiol.* **98**, 1064–1072 (2007).
18. Murray, I. & Graham, M. Pseudo-marginal slice sampling. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 911–9 (2016).