bioRxiv preprint doi: https://doi.org/10.1101/2021.04.16.440107; this version posted April 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Improving statistical power in severe malaria genetic association studies by augmenting phenotypic precision

James A Watson[1,2,†,*], Carolyne M Ndila[3,†], Sophie Uyoga[3], Alexander W Macharia[3], Gideon Nyutu[3], Mohammed Shebe[3], Caroline Ngetsa[3], Neema Mturi[3], Norbert Peshu[3], Benjamin Tsofa[3], Kirk Rockett[4,5], Stije Leopold[1,2], Hugh Kingston[1,2], Elizabeth C George[6], Kathryn Maitland[3,7], Nicholas P Day[1,2], Arjen Dondorp[1,2], Philip Bejon[2,3], Thomas N Williams[3,7,‡], Chris C Holmes[8,9,‡], Nicholas J White[1,2,‡]

1: Mahidol Oxford Tropical Medicine Research Unit, Faculty of Tropical Medicine, Mahidol University, Bangkok, Thailand
2: Centre for Tropical Medicine and Global Health, Nuffield Department of Medicine, University of Oxford, United Kingdom
3: KEMRI-Wellcome Trust Research Programme, Centre for Geographic Medicine Research-Coast, Kilifi 80108, Kenya
4: The Wellcome Trust Sanger Institute, Cambridge, United Kingdom
5: Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, United Kingdom
6: Medical Research Council Clinical Trials Unit, University College London, United Kingdom
7: Institute of Global Health Innovation, Imperial College, London, United Kingdom
8: Nuffield Department of Medicine, University of Oxford, United Kingdom
9: Department of Statistics, University of Oxford, United Kingdom
†: Joint first authors
‡: Joint senior authors
*Correspondence: jwatowatson@gmail.com

Abstract

Severe falciparum malaria has substantially affected human evolution. Genetic association studies of patients with clinically defined severe malaria and matched population controls have helped characterise human genetic susceptibility to severe malaria, but phenotypic imprecision compromises discovered associations. In areas of high malaria transmission the diagnosis of severe malaria in young children and, in particular, the distinction from bacterial sepsis, is imprecise. We developed a probabilistic diagnostic model of severe malaria using platelet and white count data. Under this model we re-analysed clinical and genetic data from 2,220 Kenyan children with clinically defined severe malaria and 3,940 population controls, adjusting for phenotype mis-labelling. Our model, validated by the distribution of sickle trait, estimated that approximately one third of cases did not have severe malaria. We propose a data-tilting approach for case-control studies with phenotype mis-labelling and show that this reduces false discovery rates and improves statistical power in genome-wide association studies.

Introduction

Severe malaria caused by the parasite *Plasmodium falciparum* kills nearly half a million children each year, mostly in sub-Saharan Africa [1]. By causing death in children before reaching their reproductive age, *P. falciparum* has exerted a substantial selective evolutionary pressure on the human genome [2, 3]. Recent advances in whole genome sequencing and haplotype imputation [4], combined with data gathered prospectively from large patient cohorts has improved our understanding of genetic susceptibility to *P. falciparum* infection and severe disease [5, 6, 7, 8] but many questions remain unanswered [3]. A major limitation of genetic association studies in severe malaria is that the diagnosis of severe falciparum malaria in children is imprecise [9, 10, 11]. This imprecision increases with transmission intensity due to the low positive predictive value of

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.16.440107; this version posted April 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

blood-stage parasitaemia in areas where the background prevalence of microscopy detectable parasitaemia in apparently healthy young children is high (typically around 30% [12] but can exceed 90%[13]).

Severe falciparum malaria has been defined by experts convened by the World Health Organization (WHO) as clinical or laboratory evidence of vital organ dysfunction in the presence of circulating asexual *P. falciparum* parasitaemia [14]. The WHO definition of severe malaria is aimed primarily at clinicians and health care workers managing patients with malaria who appear severely ill. This appropriately prioritises sensitivity over specificity [15]. An inclusive clinical definition ensures that cases are not missed and patients receive the best treatment. In contrast genetic association studies require high specificity [16]. For a given sample size, their statistical power, false-discovery rates and the validity of their interpretation are weakened by phenotypic inaccuracy. Specificity in the severe malaria diagnosis depends on the prevalence of malaria parasitaemia which reflects background transmission intensity. In areas of low or seasonal transmission (e.g. most of endemic Asia and the Americas), clinical and laboratory signs of severity accompanied by a positive blood film for *P. falciparum* are highly specific for severe malaria, which predominantly affects young adults. In contrast in high transmission areas in sub-Saharan Africa and the islands of New Guinea, where severe malaria is largely a disease of young children, the diagnostic criteria for defining severe malaria are less specific because of the high background prevalence of asymptomatic parasitaemia and the lower specificity of the clinical manifestations. Standard case definitions of severe malaria will therefore inevitably include both patients with non-malarial severe illness with concomitant parasitaemia, and with concomitant non-severe malaria.

We developed a probabilistic diagnostic model of severe malaria based on haematological biomarkers using data from $1,704$ adults and children mainly from low transmission settings whose diagnosis of severe malaria is considered to be highly specific. We used this model to demonstrate low phenotypic specificity in a cohort of $2,220$ Kenyan children who were diagnosed clinically with severe malaria. We validated the predictions using a natural experiment, the distribution of sickle cell trait (HbAS), the genetic polymorphism with the strongest known protective effect against all forms of clinical malaria [6]. Building on work on 'data-tilting' [17], we suggest a new method for testing genetic associations in the context of case-control studies in which cases are re-weighted by the probability that the severe malaria diagnosis is correct under the model. As proof-of-concept, we ran a genome-wide association study across 9.6 million bi-allelic variants using the subset of cases with whole-genome sequencing data ($n = 1,297$) and population controls ($n = 1,614$). Adjusting for case mis-classification decreased genome-wide false-discovery rates [18], and increased effect sizes in the top three regions of the human genome most strongly associated with protection from severe malaria in East Africa (*HBB*, *ABO*, and *FREM3* [7]). A re-analysis of 120 directly typed polymorphisms in 70 candidate malaria-protective genes in the 2,220 Kenyan cases and 3,940 population controls, examining differential effects between correctly and incorrectly classified cases, suggests that the protective effect of glucose-6-phosphate dehydrogenase (G6PD) deficiency has been obscured in this population by case mis-classification. Our results show that adding full blood count meta-data - routinely measured in most hospitals in sub-Saharan Africa - to severe malaria cohorts would lead to more accurate quantitative analyses in case-control studies and increased statistical power.

Results

Reference model of severe malaria

We used the joint distribution of platelet counts and white blood cell counts (both on a logarithmic scale) to develop a simple biomarker-based reference model of severe malaria. To fit the reference model (i.e. P[Data | Severe malaria]), we used (i) platelet and white count data from severe malaria patient cohorts enrolled in low transmission areas where severe disease accompanied by a positive blood stage parasitaemia has a high positive predictive value for severe malaria (930 adults from Vietnam [19, 20] and 653 adults and children from Thailand and Bangladesh); and (ii) data from severely ill African children with plasma *Pf*HRP2 concentrations $>$ 1,000 ng/ml and $>$ 1,000 parasites per μL of blood (121 children from Uganda [21]). Severe illness accompanied by a high plasma *Pf*HRP2 concentration makes the diagnosis of severe malaria highly specific [22]. The joint distribution of platelet and white blood cell counts in severe malaria was modelled as a bivariate t-distribution with both blood count variables on the $\log_{10}$ scale.

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.16.440107; this version posted April 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

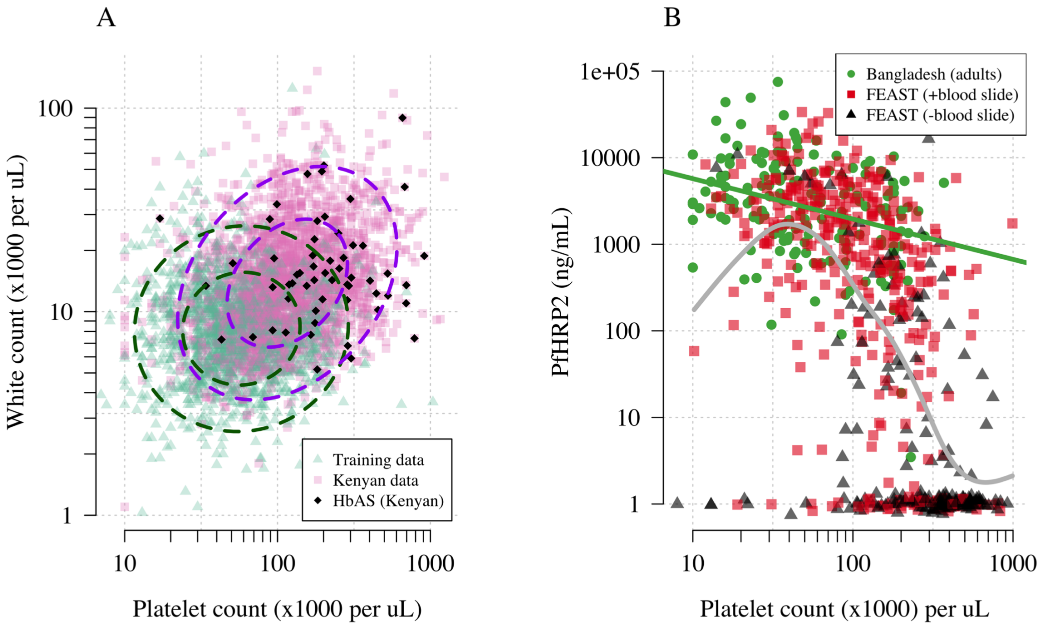

Figure 1: **Platelet counts and white blood cell counts as diagnostic predictors of severe malaria.** Panel A shows the bi-variate marginal distribution in the training data (thought to be highly specific to severe malaria, green triangles, $n = 1,704$) and in the Kenyan case data (pink squares, $n = 2,220$; black diamonds: HbAS). The dashed ellipses show the 50 and 95% bivariate normal probability contours approximating each dataset (dark green: training data; purple: Kenyan data). Panel B shows the relationship between platelet counts and plasma *Pf*HRP2 in adults and children with severe malaria from Bangladesh (green circles, $n = 172$, the green line shows a linear fit) and in the FEAST trial ($n = 566$, not specific to severe malaria) [21]. Red squares: malaria-positive blood slide; black triangles: malaria-negative blood slide. The grey line shows a spline fit to the FEAST data (*smooth.spline* function in R with default parameters). Undetectable plasma *Pf*HRP2 concentrations were set to 1 ng/mL ± random jitter.

Figure 1A shows the training data (green triangles: patients with a highly specific diagnosis of severe malaria) alongside data from a large Kenyan cohort of hospitalised children diagnosed with severe malaria, whose diagnosis had unknown specificity (pink squares). The median platelet count in the training data was 57,000 per μL and the median total white blood cell count was 8,400 per μL. In contrast, the median platelet count in the Kenyan children was 120,000 per μL and the median total white blood cell count was 13,000 per μL. To rule out substantial confounding by geography and age, we demonstrate the discriminatory value of platelet counts alone (Figure 1B). Low platelet counts were highly predictive of blood stage parasitaemia and elevated *Pf*HRP2 in a cohort of 566 severely ill African children enrolled in the FEAST trial [21] (p=10^{-16} for a spline term on the $\log_{10}$ platelet count in a generalised additive logistic regression model predicting *Pf*HRP2 $>$ 1,000 ng/mL, Figure S1). African children enrolled in the FEAST trial who had severe thrombocytopenia ($< 100,000$ platelets per μL) had comparable *Pf*HRP2 concentrations to Asian adults diagnosed with severe falciparum malaria. Total white blood cell counts are age dependent and vary across genetic backgrounds, in particular related to mutations in the *ACKR1* gene that results in the Duffy negative phenotype prevalent in African populations [23]. However, after adjustment for age (see Methods), the marginal distributions of total white counts were comparable between Asian adults and children with severe malaria and African children with high *Pf*HRP2 (Figure S2).

Estimating the proportion of children mis-diagnosed with severe malaria

We can consider the hospitalised Kenyan children in this series as a mixture of two latent subpopulations, 'severe malaria' and 'not severe malaria' (i.e an alternative aetiology for severe illness). To estimate the proportion of each we use the distribution of HbAS, the human polymorphism most protective against all forms of clinical falciparum malaria. HbAS provides at least 90%

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.16.440107; this version posted April 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

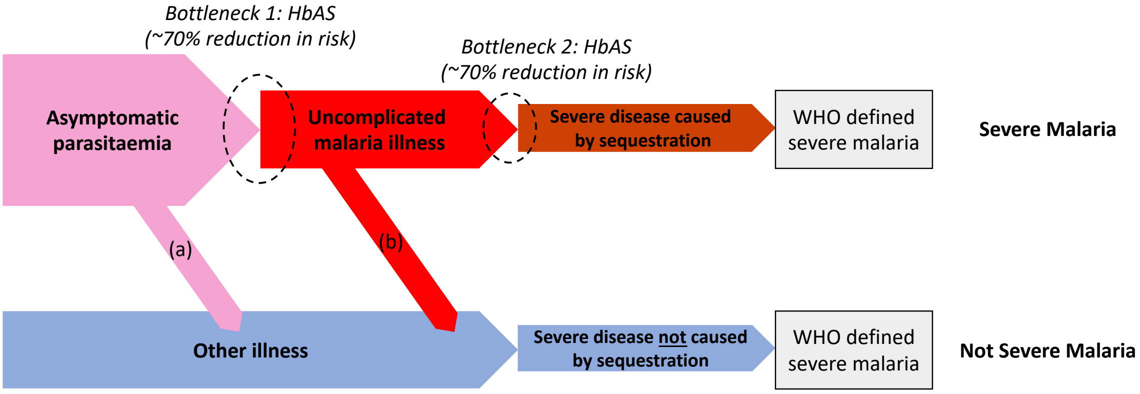
Bottleneck 1: HbAS
(~70% reduction in risk)
Bottleneck 2: HbAS
(~70% reduction in risk)
Asymptomatic parasitaemia
Uncomplicated malaria illness
Severe disease caused by sequestration
WHO defined severe malaria
Severe Malaria
(a)
(b)
Other illness
Severe disease not caused by sequestration
WHO defined severe malaria
Not Severe Malaria

Figure 2: **Theoretical causal pathways that lead to the clinical diagnosis of severe malaria under the current WHO definition [14].** Pathways (a) & (b) represent the two ways patients can be mis-classified as severe malaria. For both pathways (a) & (b), we expect a higher prevalence of HbAS relative to the population with true severe malaria due to the protective bottlenecks. In this causal model we assume that HbAS does not protect against asymptomatic parasitaemia, although this assumption is not strictly necessary. Adapted with permission from [25].

protection against severe malaria [24, 6]. The causal SNP rs334 was genotyped in 2,213 of the Kenyan children, of whom 57 were HbAS. The causal pathways (a) or (b) in Figure 2 (note all children have been selected into the study on the basis of clinical symptoms consistent with severe malaria) show how the distribution of HbAS can be used to infer the marginal probability P(Severe malaria) in the Kenyan cohort as the prevalence of HbAS is expected to differ in the two latent sub-populations.

We assumed that cases with the highest likelihood values P(Data | Severe malaria) under the reference model (a bivariate t-distribution fit to the training data) had a diagnosis of severe malaria that was 100% specific (top 40% of cases, a sensitivity analysis varied this threshold). The cases with lower likelihood values were assumed to be drawn from a mixture of the two latent populations with an unknown mixing proportion; the prevalence of HbAS in the 'not-severe malaria' subgroup was estimated from a cohort of hospitalised children enrolled in the same hospital and who were malaria blood slide positive but were clinically diagnosed as not having severe malaria ($n = 6,748$ of whom 364 were HbAS [26]). We assumed that this diagnosis of 'not-severe malaria' was 100% specific. Under these assumptions, we estimated that P(Severe malaria)=0.64 (95% credible interval (C.I.) 0.46 to 0.8), implying that approximately one third of the 2,200 cases are from the 'not-severe malaria' sub-population (they have malaria parasitaemia in addition to another severe illness - likely to be bacterial sepsis, Figure 2).

Estimating individual probabilities of severe malaria

We then estimated P(Severe malaria | Data) for each Kenyan case by fitting a mixture model to the training data and to the Kenyan data jointly. The model assumed that the platelet and white count data for the Kenyan children were drawn from a mixture of P(Data | Severe malaria) and P(Data | Not severe malaria). The training data (Asian adults and children with severe malaria and African children with PfHRP2 $> 1{,}000$ ng/mL) were assumed to be drawn only from P(Data | Severe malaria). P(Data | Not severe malaria) was modelled itself as a mixture of bivariate t-distributions. We used an informative prior on the mixture proportion ('severe malaria' versus 'not severe malaria') in the Kenyan cases, a beta distribution approximating the posterior estimate from the analysis of HbAS prevalence.

Figure 3A shows the bi-modal distribution of the posterior individual estimates of P(Severe malaria | Data). The individual posterior probabilities of severe malaria were highly predictive of HbAS ($p = 10^{-6}$ from a generalised additive logistic regression model fit, Figure 3C) and in-hospital mortality ($p = 10^{-9}$ from a generalised additive model fit; Figure 3D). In the top quintile of patients with the highest estimated P(Severe malaria | Data), the prevalence of HbAS was 0.7% (3 out of 446). In contrast, for patients in the lowest quintile of estimated P(Severe malaria | Data), the prevalence of HbAS was 4.8% (21 out of 444). These patients with a low probability

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.16.440107; this version posted April 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

of severe malaria had a substantially higher case fatality ratio (6.1% mortality for patients in the top quintile of P[Severe malaria | Data] versus 18.8% mortality for the bottom quintile of P[Severe malaria | Data]). This may be explained by the higher case-specific mortality of severe bacterial sepsis (the most likely alternative cause of severe illness). The blood culture positive rate was 2.1% in the top quintile of P(Severe malaria | Data), and 4.4% in the lowest quintile of P(Severe malaria | Data) and the individual probabilities were predictive of blood culture results ($p = 0.004$ under a generalised additive logistic regression model fit).

Accounting for case imprecision in case-control studies

'False-positive' cases reduce statistical power and dilute effect size estimates in case-control studies. We propose a novel approach for case-control studies with phenotypic imprecision based on data tilting [17]. The idea is to 'tilt' the cases towards a pseudo-population with higher specificity for severe malaria. We can do this by re-weighting the data by the probabilities P(Severe malaria | Data), i.e. re-weighting the contribution to the log-likelihood in an association model.

We applied this approach as proof-of-concept to a genome-wide association study using the subset of Kenyan children who had clinical and whole genome data available (after quality control checks $n = 1,297$ cases) and a set of matched population controls ($n = 1,614$), across 9.6 million bi-allelic variants on the autosomal chromosomes. We compared the data-tilting method to the standard non-weighted approach by estimating local false discovery rates (FDR) [18]. Compared to the standard non-weighted GWAS, data-tilting substantially increased the number of significant associations for local FDRs in the range of 1-5% (Figure 4). For example, at an FDR of 2%, the number of significant hits is more than doubled with the additional hits all around known loci associated with protection from severe malaria. We note that if the data weights were not predictive of the true latent phenotype, we would expect fewer significant hits for a given FDR due to the reduction in effective sample size. This is demonstrated by permuting the data weights (for the cases only), which results in 50-75% reduction in the number of significant hits at a FDR of 5% (Figure S3).

Examining the three major genetic regions strongly associated with protection from severe malaria in East Africa (*HBB*: HbAS; *ABO*: O blood group; *FREM3*: Dantu blood group) [7], the data-tilted approach estimated larger effect sizes compared to the non-weighted model in all three regions (effect size increases: 30% around *HBB*, 9% around *ABO*, and 5% around *FREM3*). This resulted in larger -$\log_{10}$ p-values for *HBB* and *ABO*, but slightly smaller for *FREM3* (Figure 5).

Reappraisal of directly typed polymorphisms

We re-analysed 120 polymorphisms on 70 candidate malaria-protective genes which were typed directly in the 2,220 Kenyan children along with 3,940 population controls. In this case-control cohort, 14 polymorphisms had previously been identified as associated with protection or increased risk in severe malaria [27]. A re-analysis of these 14 variants using the same models of association as previously published and down-weighting the likely mis-classified cases replicated the majority of associations, with increased effect sizes and increased -$\log_{10}$ p-values (Figure S4). For the three major genes (*HBB, ABO, FREM3*), effect sizes were increased by 10-30% and associations all had higher significance levels on the -$\log_{10}$ scale (0.25-1.7). The allele frequencies of all three polymorphisms were directly associated with the probability weights, showing increased protection in individuals more likely to have severe malaria (Figure S5). Two polymorphisms on the genes *ARL14* and *LOC727982*, reported previously as associated with protection in severe malaria (neither of which are related to red cells), showed decreased effect sizes and -$\log_{10}$ p-values and are thus potentially spurious hits.

We explored whether there was evidence of differential effects in the Kenyan cases using P[Severe malaria | Data] to assign probabilistically each case to the 'severe malaria' versus 'not severe malaria' sub-populations. We fitted a categorical logistic regression model predicting the latent sub-population label versus control, where the latent case label was estimated from the weights shown in Figure 3A. This resulted in approximately 1,279 cases in the 'severe malaria' sub-population and 941 cases in the 'not severe malaria' sub-population. Differential effects were tested by comparing the estimated log-odds for the two sub-populations. After accounting for multiple testing, two polymorphisms showed significant differential effects: rs334 (derived allele encodes haemoglobin S, $p = 10^{-6}$) and rs1050828 (derived allele encodes *G6PD*+202T, $p = 10^{-3}$ in the model fit

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.16.440107; this version posted April 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

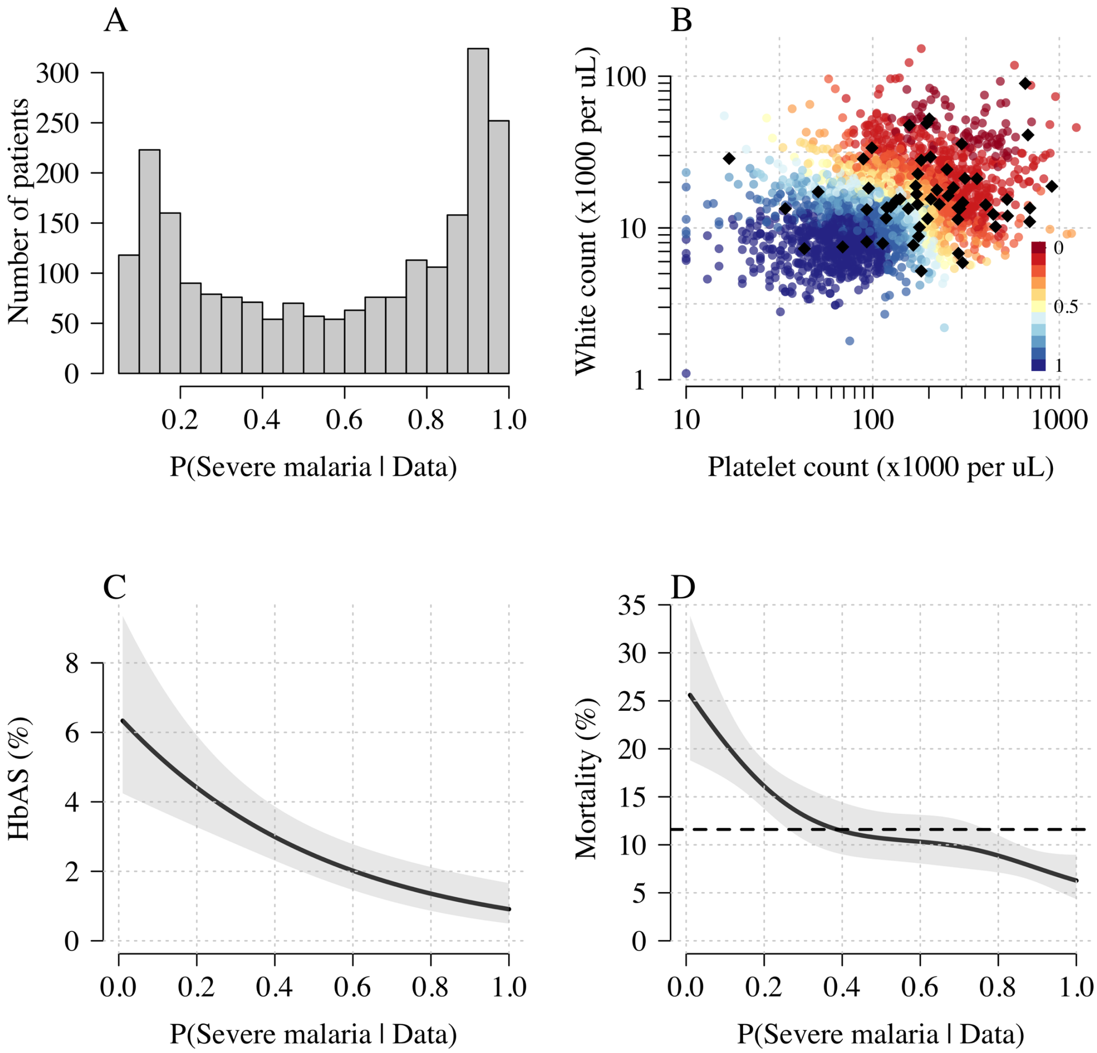

Figure 3: **Model estimates of P(Severe malaria | Data) in 2,220 Kenyan children clinically diagnosed with severe malaria.** Panel A: distribution of posterior probabilities of severe malaria being the correct diagnosis. Panel B shows these probabilities plotted as a function of the platelet and white counts on which they are based. The black diamonds show the HbAS individuals. Panels B & C show the relationship between the estimated probabilities of severe malaria and HbAS and in-hospital mortality. The black lines (shaded areas) show the mean estimated values (95% confidence intervals) from a generalised additive logistic regression model with a smooth spline term for the likelihood (R package *mgcv*).

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.16.440107; this version posted April 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

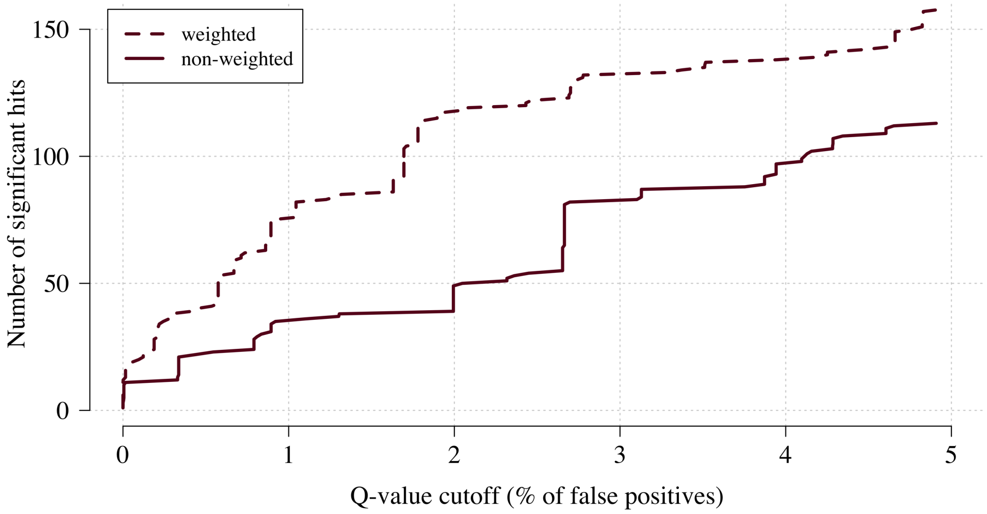

Figure 4: **The number of significant hits as a function of the false discovery rate for the genome-wide association study across 9.6 million bi-allelic variants.** This analysis is based on a subset of the Kenyan children with whole genome data available and passing quality checks $n = 1,297$, and $n = 1,614$ controls. Dashed line: weighted-model; thick line: non-weighted model.

to females only), see Figure 6. As expected, rs334 was associated with protection in both sub-populations [28, 26] but the effect was almost 8 times larger on the log-odds scale in the 'severe malaria' sub-population relative to the 'not severe malaria' sub-population (odds-ratio of 0.029 [95% C.I. 0.0088-0.094] in the 'severe malaria' population versus 0.63 [95% C.I. 0.48-0.83] in the 'not severe malaria' population). For rs1050828 (*G6PD*+202T allele), approximately the same absolute log-odds were estimated for both sub-populations but they had opposite sign. Under an additive model in females, the rs1050828 T allele was associated with protection in the 'severe malaria' sub-population (odds-ratio of 0.71 [95% C.I. 0.57-0.88]) but with increased risk in the 'not severe malaria' sub-population (odds-ratio of 1.30 [95% C.I. 1.00-1.70]). The additive model including both males and females was consistent with these opposing effects but significant only at a nominal threshold ($p = 0.02$). Opposing effects across the two sub-populations is consistent with the hypothesis that G6PD deficiency leads to a greater risk of being erroneously classified as severe malaria due to the severe anaemia criterion [29] (shown in more detail in Figure S5). Investigation of haemoglobin concentrations as a function of P(Severe malaria | Data) indicates that the mis-classified group is very heterogeneous, but with a larger proportion of severe anaemia (<5 g/dL) relative to the correctly classified sub-population (Figure S6).

Discussion

The clinical diagnosis of severe falciparum malaria in African children is imprecise [10, 11, 9]. Even with quantitation of parasite densities, specificity is still imperfect [11]. In children with cerebral malaria (unrouseable coma with malaria parasitaemia), the most specific of the severe malaria clinical syndromes, post-mortem examination revealed another diagnosis in about 25% of cases studied in Blantyre, Malawi [10]. Diagnostic specificity can be improved by visualisation of the obstructed microcirculation in-vivo (e.g. through indirect ophthlamoscopy) or from parasite biomass indicators (quantitation and staging of malaria parasites on thin blood films, counting of neutrophil ingested malaria pigment, measurement of plasma concentrations of *Pf*HRP2 or parasite DNA), but these are still largely research procedures and have not been widely adopted or measured at scale for genetic association studies. Our results suggest that imprecision in clinical phenotyping is more substantial than thought previously. In this cohort of 2,220 Kenyan children diagnosed with severe malaria from an area of moderate transmission, a probabilistic assessment suggests that

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.16.440107; this version posted April 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

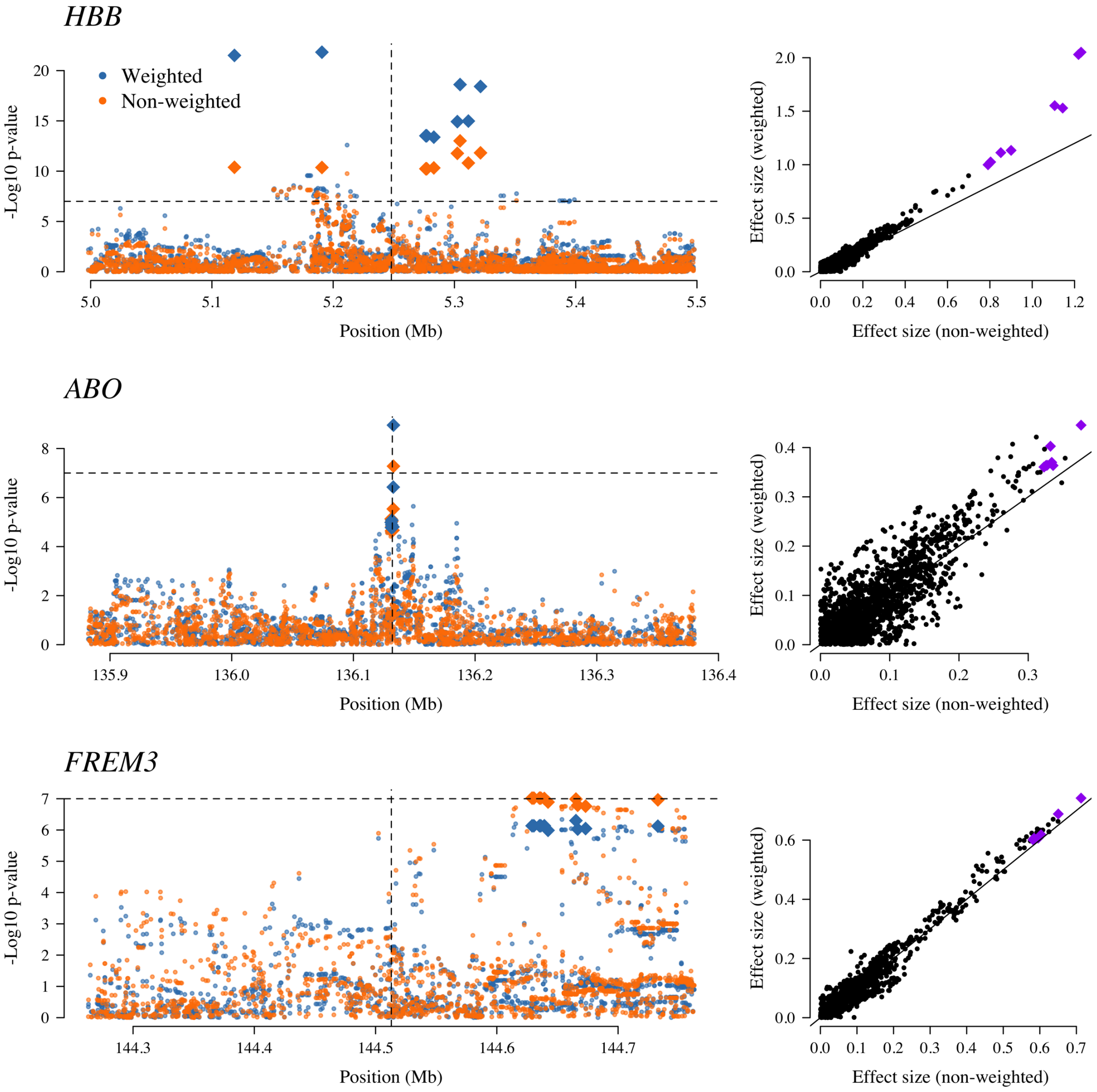

Figure 5: **The three regions in the human genome with the greatest evidence for protection against severe malaria in East Africa (*HBB*, *ABO* and *FREM3*) [7].** The Manhattan plots (left panels) compare p-values from the weighted model (blue) and the non-weighted model (orange). Each Manhattan plot is centred around the known causal position shown by the vertical dashed line (0.5 Mb region). The horizontal dashed line shows $p = 10^{-7}$ (threshold often used for defining genome-wide significance). The 10 positions with the greatest -log$_{10}$ p-values under the non-weighted model are shown as large diamonds. The scatter plots on the right compare absolute effect size estimates under both models with the same top 10 hits shown by the larger purple diamonds. Increases of 30%, 9% and 5% are seen for the ten top hits for *HBB, ABO*, and *FREM3*, respectively.

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.16.440107; this version posted April 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

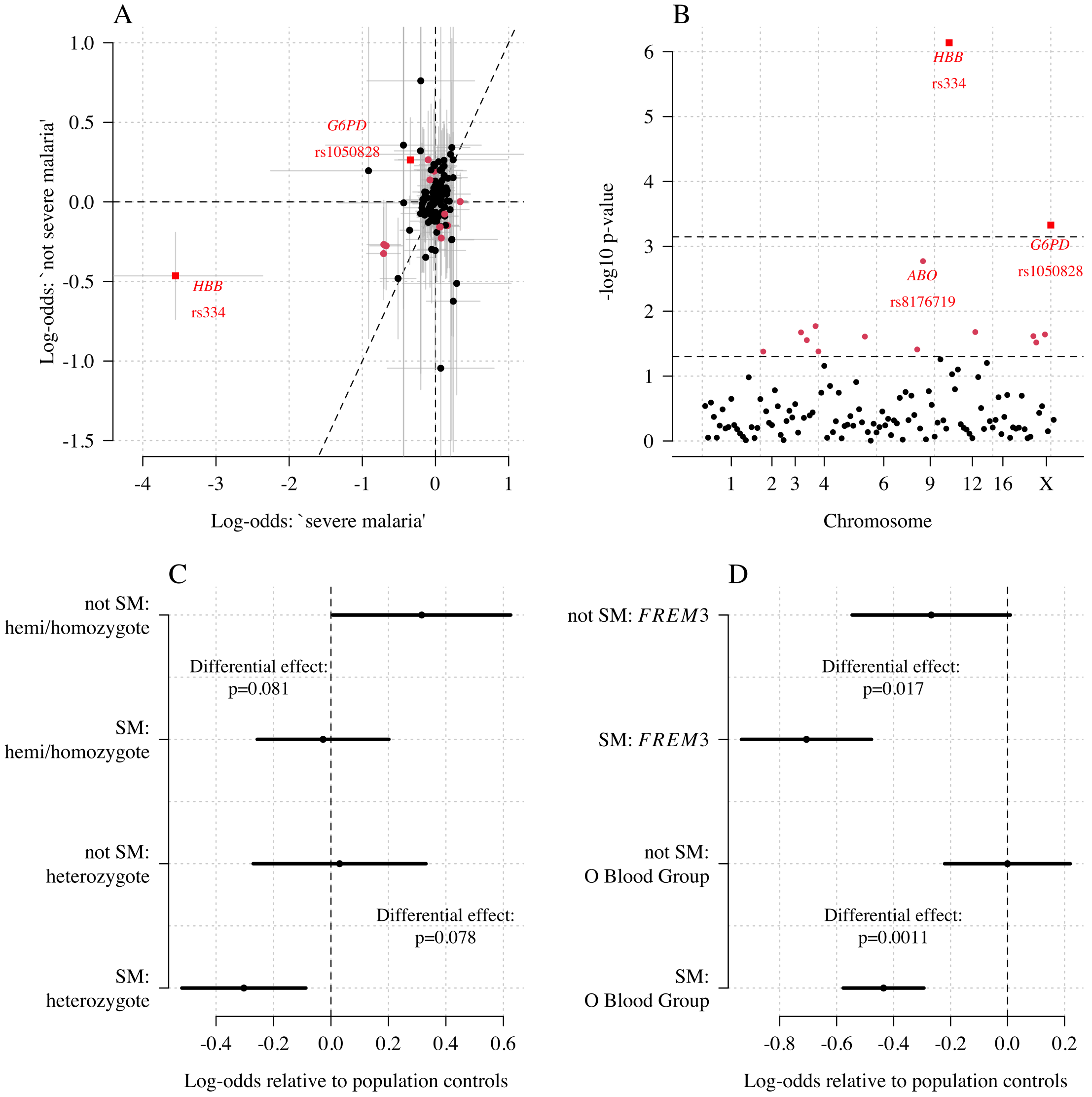

Figure 6: **Exploring differential effects in 120 directly typed polymorphisms across 70 candidate malaria-protecting genes.** Panel A: case-control effect sizes estimated for the 'severe malaria' sub-population versus the 'not severe malaria' sub-population ($n = 3,940$ controls and $n = 2,220$ cases, with approximately 1,279 in the 'severe malaria' sub-population and 941 in the 'not severe malaria' sub-population). The vertical and horizontal grey lines show the 95% credible intervals. Panel B shows the $\log_{10}$ p-values testing the hypothesis that the effects are the same for the two sub-populations relative to controls. The top dashed line shows the Bonferroni corrected $\alpha = 0.05$ significance threshold (assuming 70 independent tests). The bottom dashed line shows the nominal $\alpha = 0.05$ significance threshold. In both panels, red circles denote $p < 0.05$ (nominal significance level), and red squares denote $p < 0.05/70$. Panel C: Analysis of the rs1050828 SNP (encoding G6PD+202T) under a non-additive model (hemi/homozygotes and heterozygotes are distinct categories). This shows that heterozygotes are clearly under-represented in the 'severe malaria' sub-population and hemi/homozygotes are clearly over-represented in the 'not severe malaria' sub-population. Panel D: evidence of differential effects for the O Blood Group (rs8176719, recessive model) and *FREM3* (additive model).

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.16.440107; this version posted April 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

over one third may not have had severe malaria (although malaria may have contributed to their illness [25]). This supports our previous conclusion that differences in treatment effects between Asian adults and African children (i.e the benefits of artesunate over quinine in severe malaria estimated from randomised trials [30, 31]) are predominantly driven by differences in diagnostic specificity [22, 9]. Using HbAS as a natural experiment to validate the biomarker model, we show that the joint distribution of platelet and white blood cell counts is a diagnostic predictor of severe malaria. Complete blood counts are inexpensive and increasingly available in low-resource setting hospitals. An upper threshold of 200,000 platelets per μL would have substantially decreased mis-classification in this large cohort of Kenyan children diagnosed with severe malaria.

Our re-analysis using rich clinical data provides additional evidence for the three major genetic polymorphisms present in East Africa which are the most highly protective against severe malaria. After probabilistic down-weighting of the likely mis-classified cases substantial increases in effect sizes were found. Dilution of effect sizes resulting from mis-classification could explain the large heterogeneity in effects noted in the largest severe malaria GWAS to date [7]. For haemoglobin S (rs334) there was a 4-fold variation in estimated odds-ratios across participating sites. Some of this heterogeneity can be attributed to variations in linkage disequilibrium affecting imputation accuracy [5], but our analysis shows an additional substantial source of heterogeneity which results from diagnostic imprecision. This can be adjusted for if detailed clinical data are available. For example, in the case of rs334 (directly typed), the data-tilting approach results in a 25% increase in effect size on the log-odds scale, corresponding to 35% decrease in estimated odds-ratios (0.1 versus 0.16).

As for the interpretation of genetic effects, one of the most interesting results concerns the *G6PD* gene. G6PD deficiency is the most common enzymopathy of humans, and its role in falciparum malaria has been controversial [32, 29]. A very large multi-country genetic association study with over 11,000 severe malaria cases and 17,000 population controls found no overall protective effect of the *G6PD*+202T allele (the most common mutation in sub-Saharan Africa causing G6PD deficiency), under an additive model [6]. The same pattern is observed in this Kenyan cohort (which is a subset of the larger study). In the Kenyan cohort overall, a previous analysis found no clear evidence of protection for male homozygotes but substantial evidence of protection for female heterozygotes [33]. This suggests a heterogyzote advantage leading to a balancing polymorphism. However, when the Kenyan cases are modelled as two distinct sub-populations, there is evidence of differential effects between the ‘severe malaria’ and ‘not severe malaria’ sub-populations. Hemi and homozygous G6PD deficiency was associated with an increased risk of mis-classification (reflecting an increased risk of severe anaemia), but it is unclear whether or not hemi/homozygous G6PD deficiency was protective in the ‘true severe malaria’ sub-population (Figure 6C). On the other hand, heterozygote deficiency was very clearly protective in the true severe malaria subgroup, consistent with previous findings, and did not appear to lead to an increased risk of mis-classification (consistent with a lower risk of extensive haemolysis and thus false classification in heterozygotes who have both normal and G6PD deficient erythrocytes in their circulation). When examining the ‘severe malaria’ sub-population only, the sample size in this study is too small to discriminate between the heterozygote and additive models of association. In our view, the relationship between G6PD deficiency and severe falciparum malaria remains unanswered. This approach should now be applied to other case-control cohorts for a definitive understanding of the role of this major human polymorphism.

The limitations of our diagnostic model can be summarised as follows. First, the validity and interpretation of the individual probabilities that severe malaria is the correct diagnosis is heavily dependent on the reference model and thus the training data. Our training data were primarily from Asian adults in whom diagnostic specificity for severe malaria is thought to be very high. Diagnostic checks suggested that the marginal distributions of platelet counts were similar between adults and children, and we made age corrections to the white blood cell count, but small deviations could reduce the discriminatory value (e.g. lower white counts associated with the Duffy negative phenotype [23]). Second, it is possible that rare genetic conditions exist in which the probabilities of severe malaria under this model might be biased. One example is sickle cell disease (HbSS, $<$0.5% in the Kenyan cases), which results in chronic inflammation with high white counts and low platelet counts relative to the normal population [34]. The 11 children with HbSS in this cohort were all assigned low probabilities of severe malaria, but this should be interpreted with caution. Whether HbSS is protective against severe malaria or increases the risk of severe malaria remains unclear [35]. For these patients, other biomarkers such as plasma *Pf*HRP2 may

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.16.440107; this version posted April 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

be more appropriate. Third, it is theoretically possible that the joint distribution of the clinical variables used to fit the reference model could be dependent on the underlying distribution of severe malaria sub-phenotypes. For example, if the training data were biased towards cerebral malaria, and the joint distribution of platelet and white cell counts in cerebral malaria differed from those in the other severe malaria syndromes, then the predicted outliers could represent other forms of severe malaria instead of 'not-severe' malaria. This would not impact the estimate of 1 in 3 mis-classification unless the protective effect of HbAS varied substantially amongst the different sub-phenotypes of severe malaria. This variation has not been noted in previous analyses [7].

In summary, under a probabilistic model based on routine blood count data, we have shown that it is possible to estimate mis-classification rates in diagnosed severe childhood malaria in a higher transmission endemic area and compute probabilistic weights that can downweight the contribution of likely mis-classified cases. The well-established protective effect of HbAS provided an independent validation of the model. These data suggest that normal range platelet counts ($>$ 200,000 per μL) could be used as a simple exclusion criterion in severe malaria cohort studies. Based on this analysis we recommend that future studies in severe malaria collect and record complete blood count data. Further studies of platelet and white blood cell counts from a diverse cohort of children with severe malaria, confirmed using high specificity diagnostic techniques such as indirect ophthlamoscopy, plasma *Pf*HRP2, or plasma *P. falciparum* DNA should be conducted to validate this approach.

Methods

Data

Kenyan case-control cohort

The Kenyan case-control cohort has been described in detail previously [27]. Severe malaria cases consisted of all children aged <14 years who were admitted with clinical features of severe falciparum malaria to the high dependency ward of Kilifi County Hospital between June 11th 1999 and June 12th 2008. Severe malaria was defined as a positive blood-film for *P. falciparum* along with: prostration (Blantyre Coma Score of 3 or 4); cerebral malaria (Blantyre Coma Score of <3); respiratory distress (abnormally deep breathing); severe anaemia (haemoglobin $<$ 5 g/dL). Controls were infants aged 3-12 months who were born within the same area as the cases and who were recruited to a cohort study investigating genetic susceptibility to a wide range of childhood diseases. Cases and controls were genotyped for the rs334 SNP and for α^+-thalassaemia along with 120 other SNPs using DNA extracted from fresh or frozen samples of whole blood as described in detail previously [27, 36].

The Fluid Expansion as Supportive Therapy (FEAST) trial

FEAST was a multicentre randomised controlled trial comparing fluid boluses for severely ill children ($n = 3,161$) that was not specific to severe malaria [21]. Platelet counts, white blood cell counts, parasite densities and *Pf*HRP2 were jointly measured for 566 children (for children enrolled in the sites in Mulago, Lacor and Mbale, Uganda). In order to select only those with a very high probability of having severe malaria as the primary cause of illness, we selected the 121 children who had measured *Pf*HRP2 $>$ 1,000 ng/mL and parasitaemia $>$ 1,000 per μL.

AQ Vietnam and AAV randomised controlled trials

The AQ and the AAV studies were two randomised clinical trials in Vietnamese adults diagnosed clinically with severe falciparum malaria recruited to a specialist ward of the Hospital for Tropical Diseases, Ho Chi Minh City, Vietnam, between 1991 and 2003 [19, 20]. AQ Vietnam was a double blind comparison of intramuscular artemether versus intramuscular quinine ($n = 560$); AAV compared intramuscular artesunate and intramuscular artemether ($n = 370$).

Observational studies in Thai and Bangladeshi adults and children

We included data from multiple observational studies in severe falciparum malaria conducted by the Mahidol Oxford Tropical Medicine Research Unit in Thailand and Bangladesh between 1980

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.16.440107; this version posted April 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

and 2019. These pooled data have been described previously [37]. Platelet counts and white blood cell counts were available in 657 patients. We excluded one 30 year old adult from Bangladesh whose recorded platelet count was 1,000 per μL, and three other adults with platelet counts greater than 450,000 per μL as outliers reflecting likely data entry errors. Plasma *Pf*HRP2 concentrations were available in 172 patients from Bangladesh. 55 patients from this series were younger than 15 years of age.

Multiple imputation

In the Kenyan severe malaria cohort ($n = 2,220$), data on platelet counts were missing in 18%, white blood counts were missing in 0.2%, and parasite density was missing in 1.6%. In-hospital outcome (died/survived) was missing for 13 patients. rs334 genotype was missing for 7; α^+-thalassaemia genotype was missing for 101 patients. In the Vietnamese adults, platelet counts were missing in 4%, white counts in 2% and parasitaemia in 0%.

We did multiple imputation using random forests for all available clinical variables using the R package *missForest* (targeted genotyping data was not included for imputation). Supplementary Figures S7 and S8 shows the missing data pattern in the studies in Vietnamese adults and in the Kenyan severe malaria cases, respectively. Ten datasets were imputed for each dataset independently and were used for the subsequent analyses. Analyses using directly typed genetic polymorphisms or the within-hospital outcome as the dependent variables used only the data where these outcomes were recorded, assuming that they were missing at random.

Reference model of severe malaria

Biological rationale

Thrombocytopenia accompanied by a normal white blood count and a normal neutrophil count are typical features of severe malaria [38, 39], but they may also occur in some systemic viral infections and in severe sepsis. Neutrophil leukocytosis may sometimes occur in very severe malaria, but is more characteristic of pyogenic bacterial infections. These indices, whilst individually not very specific, could each have useful discriminatory value. We reasoned therefore that their joint distribution could help discriminate between children with severe malaria versus those severely ill with coincidental parasitaemia. The Kenyan severe malaria cohort did not have differential white count data, so we used platelet counts and total white blood cell counts as the two diagnostic biomarkers in the reference model of severe malaria.

Choice of training data and confounders

The best data for fitting the biomarker model are either from children or adults from low transmission areas (where parasitaemia has a high positive predictive value); or in children or adults with high plasma *Pf*HRP2 measurements indicating a large latent parasite biomass [22]. In the first years of life, white blood cell counts are often much higher than in adults because of lymphocytosis. We used data from 858 children from the FEAST trial, in whom white counts were measured, to estimate the relationship between age and mean white count in severe illness (median age was 24 months). The estimated relationship is shown in Figure S9 (using a generalised additive linear model with the white count on the $\log_{10}$ scale), with mean white counts reaching a plateau around 5 years of age. We used this to correct all white count data in children less than 5 years of age, both in the training data and the Kenyan cohort.

There is also a systematic difference associated with the Duffy negative phenotype which is near fixation in Africa but absent in Asia. Duffy negative individuals have lower neutrophil counts (termed benign ethnic neutropenia) [23]. The use of Asian adults to estimate the reference distribution of white counts in severe malaria could thus falsely include individuals with elevated white counts (relative to the normal ranges). However, a diagnostic quantile-quantile plot (Figure S2, on the log-scale) comparing the white count distribution in Vietnamese adults and in children in the FEAST trial who had *Pf*HRP2 $>$ 1,000 ng/mL did not suggest any major differences. In fact the African children had slightly higher white counts on average even after the correction for age, this may represent imperfect specificity when using a plasma *Pf*HRP2 cutoff of 1,000 mg/mL.

For platelet counts (which have the greatest diagnostic value for severe malaria in our series) age is not a confounder and published data support the hypothesis that thrombocytopenia is highly

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.16.440107; this version posted April 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

specific for severe malaria in children as well as adults (with a diagnostic and a prognostic value). The French national guidelines specifically mention thrombocytopenia ($<$150,000 per μL) for the diagnosis of children who have travelled to a malaria endemic area. In a French paediatric series in travellers, almost half had severe thrombocytopenia ($<$50,000 per μL) [40, 41]. In Dakar, Senegal (one of the lowest transmission areas in Africa) thrombocytopenia was an independent predictor of death and the median platelet count was 100,000 [42, 43]. Comparison of the distributions of platelet counts (on the log scale) between Asian children and Asian adults suggested no major differences (Figure S10), although we had few data for Asian children. In the seminal Blantyre autopsy study [10], platelet counts were substantially different between fatal cases confirmed post-mortem to be severe malaria (62,000 per μL, and 56,000 per μL for the children with sequestration only, and for sequestration + microvascular pathology, respectively) and fatal cases with a mis-diagnosis of severe malaria (no sequestration: 176,000 per μL; the inter-group difference was significant, $p = 0.008$). A larger cohort from the same centre in Malawi reported substantially higher platelet counts in retinopathy negative cerebral malaria (mean count was 161,000 per μL, $n = 288$) compared to retinopathy positive cerebral malaria (mean count was 81,000 per μL, $n = 438$) [25].

We visually checked approximate normality for each marginal distribution using quantile-quantile plots (Figure S11). On the $\log_{10}$ scale, platelet counts and white counts show a good fit to the normal approximation but with some outliers so a t-distribution was used (robust to outliers). For all modelling of the joint distribution of platelet counts and white blood cell counts, we chose bivariate t-distributions with 7 degrees of freedom as the default model. The final reference model used was a bi-variate t-distribution fit to the joint distribution of platelet counts and white counts both on the logarithmic scale. On the $\log_{10}$ scale the mean values (standard deviations) were approximately 1.76 (0.11) and 0.92 (0.055) for platelets and white counts, respectively. The covariance was approximately 0.0035. These values varied very slightly across the ten imputed datasets. Log-likelihood values for each severe malaria case in the Kenyan cohort were calculated for each imputed dataset independently. The median log-likelihoods per case were then used in downstream analyses.

Limitations of the model

The diagnostic model of severe malaria using platelet counts and white blood cell counts cannot be applied to all patients. We summarise here the known and possible limitations. When using this model to estimate the association between a genetic polymorphism and the risk of severe malaria, if the genetic polymorphism of interest affects the complete blood count independently, there will be selection bias (see the directed acyclic graph in Figure S12). One example is HbSS. Children with HbSS have chronic inflammation with white blood cells counts about 2-3 times higher than normal and slightly lower platelet counts [34]. All 11 children in the Kenyan cohort with HbSS were assigned low probabilities of having severe malaria (Figure S13), but these probabilities reflect a deficiency of the model. Including or excluding these children from the analysis had no impact on the results as they represent less than 0.5% of the cases.

The second possible limitation concerns the validation using HbAS. Previous studies have suggested negative epistasis between the malaria-protective effects of HbAS and α^+-thalassaemia [44, 45]. The 3.7 kb deletion across the *HBA1-HBA2* genes (known as α^+-thalassaemia) has an allele frequency of $\sim$ 40% in this population, therefore 16% of HbAS individuals are homozygous for α^+-thalassaemia [46]. Negative epistasis implies that those with both polymorphisms would have less or no protective effect against severe malaria. Of the 2,113 Kenyan cases with both HbS and α^+-thalassaemia genotyped, 13 were HbAS and homozygous α^+-thalassaemia. Figure S14 shows that the majority of those with both polymorphisms had clinical indices pointing away from severe malaria suggesting that the observed number of patients with both HbAS and homozygous α^+-thalassaemia is inflated by 2 to 3 fold.

The final possible problem concerns the use of white blood cell counts in relation to invasive bacterial infections. Bacteraemia could either be the cause of severe illness (with coincidental parasitaemia), or it could be concomitant (which may result from extensive parasitised erythrocyte sequestration in the gut), i.e. a result of severe malaria. The former should be identified as 'not-severe malaria' (as bacteraemia is the main cause of illness), but the latter should be identified as 'severe malaria' and might be mis-classified as 'not-severe malaria' under our model. However, in a series of 845 Vietnamese adults (high diagnostic specificity), only one of eight patients who had concomitant invasive bacterial infections and a white count measured had leukocytosis (median

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.16.440107; this version posted April 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

white count was 8,100; range 3,500 to 14,850) [47].

Estimating the diagnostic specificity in the Kenyan cohort

We assume that the Kenyan cases are a latent mixture of two sub-populations: P_0 is the population 'severe malaria' and P_1 is the population 'not-severe malaria' (mis-classified). For diagnostic biomarkers X, this implies that $X \sim G = \pi f_0 + (1-\pi) f_1$, where f_0, f_1 are the sampling distributions (likelihoods) of each sub-population, respectively.

We can infer the value of π (proportion correctly classified as severe malaria) without making parametric assumptions about f_1 by using the distribution of HbAS (see Figure 2). This done as follows. We first estimate $\widehat{f_0}$ by fitting a bivariate t-distribution to the training data - this approximates the sampling distribution for P_0. We then make three assumptions:

1. Out of the 2,213 Kenyan cases with rs334 genotyped, we assume that cases in the top 40th percentile of the likelihood distribution under $\widehat{f_0}$ are drawn from P_0: $N_0 = 887$, of which $N_0^{sickle} = 9$ are HbAS.

2. For the other cases the proportion drawn from P_0 is unknown and denoted π': $N_G = 1,326$, of which $N_G^{sickle} = 48$ are HbAS.

3. Finally, additional information is incorporated by using data from a cohort of individuals with severe disease from the same hospital who had positive malaria blood slides but whose diagnosis was not severe malaria ($N_1 = 6,748$, of which $N_1^{sickle} = 364$ were HbAS) [26].

Under these assumptions, we can fit a Bayesian binomial mixture model to these data with three parameters: $\{\pi', p_0, p_1\}$. The likelihood is given by: $N_0^{sickle} \sim \text{Binomial}(p_0, N_0)$; $N_G^{sickle} \sim \text{Binomial}(\pi' p_0 + (1-\pi')p_1, N_G)$; $N_1^{sickle} \sim \text{Binomial}(p_1, N_1)$ The priors were: $p_1 \sim \text{Beta}(5, 95)$ (i.e. 5% prior probability with 100 pseudo observations); $p_0 \sim \text{Beta}(1, 99)$ (1% prior probability with 100 pseudo observations). A sensitivity analysis with flat beta priors (Beta[1,1]) did not qualitatively change the result (by one percentage point for the final estimate of π). To check the validity of the use of the external population from [26], we did a sensitivity analysis using the lowest quintile of the likelihood ratio distribution as a population drawn entirely from P_1 (instead of the external data from [26]).

Estimating P(Severe malaria | Data) in the Kenyan cohort

Denote the platelet and white count data from the FEAST trial as $\{X_i^{\text{FEAST}}\}_{i=1}^{121}$; the data from the Vietnamese adults and children as $\{X_i^{\text{Asia}}\}_{i=1}^{1583}$; the data from the Kenyan children as $\{X_i^{\text{Kenya}}\}_{i=1}^{2220}$. We fit the following joint model to the training biomarker data and the Kenyan biomarker data.

$$\begin{aligned}
X_i^{\text{FEAST}} &\sim \text{Student}(\mu_{SM}^1, \Sigma_{SM}^1, 7) \\
X_i^{\text{Asia}} &\sim \text{Student}(\mu_{SM}^2, \Sigma_{SM}^2, 7) \\
X_i^{\text{Kenya}} &\sim \pi f_0 + (1-\pi) f_1 \\
f_0 &= p\,\text{Student}(\mu_{SM}^1, \Sigma_{SM}^1, 7) + (1-p)\,\text{Student}(\mu_{SM}^2, \Sigma_{SM}^2, 7) \\
f_1 &= \sum_{j=1}^{K} \alpha_j\, \text{Student}(\mu_{notSM}^j, \Sigma_{notSM}^j, 7)
\end{aligned}$$

with the following prior distributions and hyperparameters, where $\alpha = \{\alpha_1, .., \alpha_K\}$ such that $\sum_{j=1}^{K} \alpha_j = 1$:

$$\begin{aligned}
\pi &\sim \text{Beta}(40.3, 24.7) \\
p &\sim \text{Beta}(2, 2) \\
\mu_{SM}^{1,2} &\sim \text{Normal}(\{1.8, 0.95\}, 0.1^2) \\
\mu_{notSM}^{1..K} &\sim \text{Normal}(\{2.5, 1.5\}, 0.25^2) \\
\alpha &\sim \text{Dirichlet}(1/K, ..., 1/K)
\end{aligned}$$

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.16.440107; this version posted April 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

The covariance matrices $\Sigma_{SM}^{1,2}$ and $\Sigma_{SM}^{1:6}$ were parameterised as their Cholesky LKJ decomposition, where the L correlation matrices had a uniform prior (i.e. hyperparameter ν=1). The model was implemented in *rstan*.

This models the biomarker data in 'not severe malaria' as a mixture of K t-distributions. We chose $K = 6$ as the default choice (sensitivity analysis increasing this has no impact). The Dirichlet prior with hyperparameter $1/K$ forces sparsity in this mixture model (most of the prior weight is on the vertices of the K-dimensional simplex), see for example [48]. This is a very general and flexible way of modelling the 'not severe malaria' distribution: we are not trying to make inferences about this distribution, we just want the mixture model to be flexible enough to describe it. The model also allows for differences in the joint distribution of platelet counts and white counts between the training datasets (FEAST trial and the Asian studies). The Kenyan cases drawn from the 'severe malaria' sub-population are then modelled as a mix of these two training models.

Reweighted likelihood for case-control analyses

For each $\{X_i^{\text{Kenya}}\}_{i=1}^{2220}$ we estimate the posterior probability of being drawn from the sampling distribution f_0. The mean posterior probability then defines a precision weight w_i which can be used in a standard generalised linear model (glm) with the same interpretation as inverse probability weights. The weighted glm is equivalent to computing the maximum likelihood estimate where the log-likelihood is weighted by w_i. In our case-control analyses all the controls are given weight 1. Nie *et al* [17] give a proof of correctness for this re-weighted log-likelihood (equivalent to 'tilting' the dataset towards the desired distribution $\widehat{f}_0(X)$).

Genome-wide association study

Anonymised whole genome data from the Illumina Omni 2.5M platform for 1,944 severe malaria cases and 1,738 population controls were downloaded from the European Genome-Phenome Archive (dataset accession ID: EGAD00010001742, release date March 2019 [7]). This contained sequencing data on 2,383,648 variants. We used the quality control meta-data provided with the 2019 data release to select SNPs and individuals with high quality data. We first excluded 386 individuals (due to relatedness: 155; missing data or low intensity: 226; gender: 5). We then removed 616,426 SNPs that did not pass quality control, leaving a total of 1,767,222 SNPs. We used plink2 to prune the SNPs (options: –maf 0.01 –indep-pairwise 50 2 0.2) down to a set of 462,120 SNPs in approximate linkage equilibrium. These SNPs were then used to calculated the first 5 principal components (Figure S15), which we subsequently used to control for population structure in the genome-wide association study. We used the Michigan imputation server with the 1000 Genomes Phase 3 (Version 5) as the reference panel to impute 28.6 million polymorphisms across the 22 autosomal chromosomes. This is a web-based service that runs imputation pipelines (phasing is done with Eagle2, imputation with Minimac4). Encrypted results are returned with a one-time password. Of the remaining 3,682 individuals (1,681 cases and 1,615 controls), we had clinical data available for 1,297 cases. We only used the subset of individuals with clinical data available in order for a fair comparison between the weighted and non-weighted genome-wide association studies. We ran subsequent genome wide association studies on all bi-allelic sites with a minor allele frequency $\geq$ 5% (9,615,446 sites in total) assuming an additive model of association. We used the R function *glm* with a binomial link for all tests of association (genetic data are encoded as the number of reference alleles). The supplementary appendix gives the R code for weighted logistic regression. The point estimates from the weighted model estimated by *glm* are correct but it is necessary to transform the standard errors in order to take into account the reduction in effective sample size (see code).

Case-control study in directly typed polymorphisms

We fit a categorical (multinomial) logistic regression model to the case-control status as a function of the directly typed polymorphisms (120 after discarding those that are monomorphic in this population, see [27] for additional details). We modelled the severe malaria cases as two separate sub-populations with a latent variable: 'severe malaria' versus 'not severe malaria', resulting in 3 possible labels (controls, 'severe malaria', 'not severe malaria'). The models adjusted for self-reported ethnicity and sex. The model was coded in *stan* [49] using the log-sum-exp trick to marginalise out the likelihood over the latent variables (see code). Normal(0,5) priors were set on

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.16.440107; this version posted April 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

all parameters and parameter estimates and standard errors were estimated from the maximum a posteriori value (function *optimizing* in *rstan*).

Code availability

Code along with a minimal clinical dataset for reproducibility of the diagnostic phenotyping model is available via a github repository: https://github.com/jwatowatson/Kenyan_phenotypic_accuracy.

Data availability

A curated minimal clinical dataset is currently available alongisde the code on the github repository. This will also be made available at publication via the KEMRI-Wellcome Harvard Dataverse (https://dataverse.harvard.edu/dataverse/kwtrp). Whole genome data are available from European Genome-Phenome Archive (dataset accession ID: EGAD00010001742). Requests for access to appropriately anonymized clinical data and directly typed genetic variants for the Kenyan severe malaria cohort can be made by application to the data access committee at the KEMRI–Wellcome Trust Research Programme by e-mail to mmunene@kemri-wellcome.org. The FEAST trial datasets are available from the principal investigator on reasonable request (k.maitland@imperial.ac.uk). Requests for access to appropriately anonymized clinical data from the AQ and AAV Vietnam study and the Asian paediatric cohort can be made via the Mahidol Oxford Tropical Medicine Research Unit data access committee by emailing the corresponding author JAW (jwatowatson@gmail.com) or Rita Chanviriyavuth (rita@tropmedres.ac).

Acknowledgements

This research was funded by The Wellcome Trust. A CC BY or equivalent licence is applied to the author accepted manuscript arising from this submission, in accordance with the grant's open access conditions. This work was done as part of SMAART (Severe Malaria Africa – A consortium for Research and Trials) funded by a Wellcome Collaborative Award in Science grant (209265/Z/17/Z:) held in part by KM, NPJD and AD. TNW and NJW are senior and principal research fellows respectively funded by the Wellcome Trust (202800/Z/16/Z and 093956/Z/10/C, respectively). ECG acknowledges funding from a core grant to the MRC CTU at UCL from the MRC (MC_UU_12023/26)

This study makes use of data generated by MalariaGEN. A full list of the investigators who contributed to the generation of the data is available from www.MalariaGEN.net. Funding for this project was provided by Wellcome Trust (WT077383/Z/05/Z) and the Bill & Melinda Gates Foundation through the Foundation of the National Institutes of Health (566) as part of the Grand Challenges in Global Health Initiative. This study also makes use of data from the FEAST trial. The FEAST trial was supported by a grant (G0801439) from the Medical Research Council, United Kingdom provided through the (MRC) DFID concordat. KM and ECG were supported by this grant.

References

[1] World Health Organization. World malaria report 2020: 20 years of global progress and challenges (2020).

[2] Carter, R. & Mendis, K. N. Evolutionary and historical aspects of the burden of malaria. *Clinical microbiology reviews* **15**, 564–594 (2002).

[3] Kariuki, S. N. & Williams, T. N. Human genetics and malaria resistance. *Human Genetics* **139**, 801–811 (2020). URL https://doi.org/10.1007/s00439-020-02142-6.

[4] Teo, Y.-Y., Small, K. S. & Kwiatkowski, D. P. Methodological challenges of genome-wide association analysis in Africa. *Nature Reviews Genetics* **11**, 149–160 (2010).

[5] Band, G. *et al.* Imputation-based meta-analysis of severe malaria in three African populations. *PLoS Genetics* **9**, e1003509 (2013).

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.16.440107; this version posted April 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[6] The Malaria Genomic Epidemiology Network. Reappraisal of known malaria resistance loci in a large multicenter study. *Nature Genetics* **46**, 1197 (2014).

[7] Band, G. *et al.* Insights into malaria susceptibility using genome-wide data on 17,000 individuals from Africa, Asia and Oceania. *Nature Communications* **10**, 5732 (2019). URL https://doi.org/10.1038/s41467-019-13480-z.

[8] Leffler, E. M. *et al.* Resistance to malaria through structural variation of red blood cell invasion receptors. *Science* **356** (2017).

[9] White, N. J., Turner, G. D., Day, N. P. & Dondorp, A. M. Lethal malaria: Marchiafava and Bignami were right. *The Journal of Infectious Diseases* **208**, 192–198 (2013).

[10] Taylor, T. E. *et al.* Differentiating the pathologies of cerebral malaria by postmortem parasite counts. *Nature Medicine* **10**, 143–145 (2004).

[11] Bejon, P. *et al.* Defining childhood severe falciparum malaria for intervention studies. *PLoS Medicine* **4**, e251 (2007).

[12] Rodriguez-Barraquer, I. *et al.* Quantification of anti-parasite and anti-disease immunity to malaria as a function of age and exposure. *eLife* **7**, e35832 (2018).

[13] Smith, T., Schellenberg, J. A. & Hayes, R. Attributable fraction estimates and case definitions for malaria in endemic. *Statistics in Medicine* **13**, 2345–2358 (1994).

[14] WHO. Severe malaria. *Tropical Medicine & International Health* **19**, 7–131 (2014). URL https://onlinelibrary.wiley.com/doi/abs/10.1111/tmi.12313_2. https://onlinelibrary.wiley.com/doi/pdf/10.1111/tmi.12313_2.

[15] Anstey, N. M. & Price, R. N. Improving case definitions for severe malaria. *PLoS Medicine* **4**, e267 (2007).

[16] Zondervan, K. T. & Cardon, L. R. Designing candidate gene and genome-wide case-control association studies. *Nature Protocols* **2**, 2492–2501 (2007). URL https://doi.org/10.1038/nprot.2007.366.

[17] Nie, L., Zhang, Z., Rubin, D., Chu, J. *et al.* Likelihood reweighting methods to reduce potential bias in noninferiority trials which rely on historical data to make inference. *The Annals of Applied Statistics* **7**, 1796–1813 (2013).

[18] Storey, J. D. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64**, 479–498 (2002).

[19] Hien, T. T. *et al.* A controlled trial of artemether or quinine in Vietnamese adults with severe falciparum malaria. *New England Journal of Medicine* **335**, 76–83 (1996).

[20] Phu, N. H. *et al.* Randomized controlled trial of artesunate or artemether in Vietnamese adults with severe falciparum malaria. *Malaria Journal* **9**, 97 (2010).

[21] Maitland, K. *et al.* Mortality after fluid bolus in African children with severe infection. *New England Journal of Medicine* **364**, 2483–2495 (2011).

[22] Hendriksen, I. C. *et al.* Diagnosing severe falciparum malaria in parasitaemic African children: a prospective evaluation of plasma PfHRP2 measurement. *PLoS Medicine* **9**, e1001297 (2012).

[23] Reich, D. *et al.* Reduced neutrophil count in people of African descent is due to a regulatory variant in the Duffy antigen receptor for chemokines gene. *PLoS Genetics* **5**, e1000360 (2009).

[24] Taylor, S. M., Parobek, C. M. & Fairhurst, R. M. Haemoglobinopathies and the clinical epidemiology of malaria: a systematic review and meta-analysis. *The Lancet Infectious Diseases* **12**, 457–468 (2012).

[25] Small, D. S. *et al.* Evidence from a natural experiment that malaria parasitemia is pathogenic in retinopathy-negative cerebral malaria. *eLife* **6**, e23699 (2017).

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.16.440107; this version posted April 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[26] Uyoga, S. *et al.* The indirect health effects of malaria estimated from health advantages of the sickle cell trait. *Nature Communications* **10**, 856 (2019). URL https://doi.org/10.1038/s41467-019-08775-0.

[27] Ndila, C. M. *et al.* Human candidate gene polymorphisms and risk of severe malaria in children in Kilifi, Kenya: a case-control association study. *The Lancet Haematology* **5**, e333–e345 (2018).

[28] Scott, J. A. G. *et al.* Relation between falciparum malaria and bacteraemia in Kenyan children: A population-based, case-control study and a longitudinal study. *The Lancet* **378**, 1316–1323 (2011). URL http://dx.doi.org/10.1016/S0140-6736(11)60888-X.

[29] Watson, J. A. *et al.* Collider bias and the apparent protective effect of glucose-6-phosphate dehydrogenase deficiency on cerebral malaria. *eLife* **8**, e43154 (2019).

[30] Dondorp, A. M., Nosten, F., Stepniewska, K., Day, N. & White, N. Artesunate versus quinine for treatment of severe falciparum malaria: a randomised trial. *The Lancet* **366**, 717–725 (2005).

[31] Dondorp, A. M. *et al.* Artesunate versus quinine in the treatment of severe falciparum malaria in African children (AQUAMAT): an open-label, randomised trial. *The Lancet* **376**, 1647–1657 (2010).

[32] Clarke, G. M. *et al.* Characterisation of the opposing effects of G6PD deficiency on cerebral malaria and severe malarial anaemia. *eLife* **6**, e15085 (2017).

[33] Uyoga, S. *et al.* Glucose-6-phosphate dehydrogenase deficiency and the risk of malaria and other diseases in children in Kenya: a case-control and a cohort study. *The Lancet Haematology* **2**, e437–e444 (2015).

[34] Sadarangani, M. *et al.* An observational study of children with sickle cell disease in Kilifi, Kenya. *British Journal of Haematology* **146**, 675–682 (2009).

[35] Williams, T. N. & Obaro, S. K. Sickle cell disease and malaria morbidity: a tale with two tails. *Trends in parasitology* **27**, 315–320 (2011).

[36] Wambua, S. *et al.* The effect of α^+-thalassaemia on the incidence of malaria and other diseases in children living on the coast of Kenya. *PLoS Medicine* **3**, e158 (2006).

[37] Leopold, S. J. *et al.* Investigating causal pathways in severe falciparum malaria: A pooled retrospective analysis of clinical studies. *PLoS Medicine* **16** (2019).

[38] Hanson, J. *et al.* The clinical implications of thrombocytopenia in adults with severe falciparum malaria: a retrospective analysis. *BMC medicine* **13**, 1–9 (2015).

[39] Leblanc, C. *et al.* Management and prevention of imported malaria in children. update of the french guidelines. *Medecine et maladies infectieuses* **50**, 127–140 (2020).

[40] Lanneaux, J. *et al.* Retrospective study of imported falciparum malaria in french paediatric intensive care units. *Archives of Disease in Childhood* **101**, 1004–1009 (2016).

[41] Mornand, P. *et al.* Severe imported malaria in children in France. A national retrospective study from 1996 to 2005. *PLoS ONE* **12**, e0180758 (2017).

[42] Gérardin, P. *et al.* Outcome of life-threatening malaria in African children requiring endotracheal intubation. *Malaria Journal* **6**, 51 (2007).

[43] Gérardin, P. *et al.* Prognostic value of thrombocytopenia in African children with falciparum malaria. *The American Journal of Tropical Medicine and Hygiene* **66**, 686–691 (2002).

[44] Williams, T. N. *et al.* Negative epistasis between the malaria-protective effects of $\alpha+$-thalassemia and the sickle cell trait. *Nature Genetics* **37**, 1253–1257 (2005).

[45] Opi, D. H. *et al.* Mechanistic studies of the negative epistatic malaria-protective interaction between sickle cell trait and $\alpha+$ thalassemia. *EBioMedicine* **1**, 29–36 (2014).

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.16.440107; this version posted April 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[46] Ndila, C. *et al.* Haplotype heterogeneity and low linkage disequilibrium reduce reliable prediction of genotypes for the $\alpha^{3.7I}$ form of α-thalassaemia using genome-wide microarray data [version 1; peer review: awaiting peer review]. *Wellcome Open Research* **5** (2020).

[47] Phu, N. H. *et al.* Concomitant Bacteremia in Adults With Severe Falciparum Malaria. *Clinical Infectious Diseases* (2020). URL https://doi.org/10.1093/cid/ciaa191. Ciaa191, https://academic.oup.com/cid/advance-article-pdf/doi/10.1093/cid/ciaa191/33095946/ciaa191.pdf.

[48] Frühwirth-Schnatter, S. & Malsiner-Walli, G. From here to infinity: sparse finite versus Dirichlet process mixtures in model-based clustering. *Advances in Data Analysis and Classification* **13**, 33–64 (2019). URL https://doi.org/10.1007/s11634-018-0329-y.

[49] Stan Development Team. RStan: the R interface to Stan (2020). URL http://mc-stan.org/. R package version 2.21.2.

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.16.440107; this version posted April 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

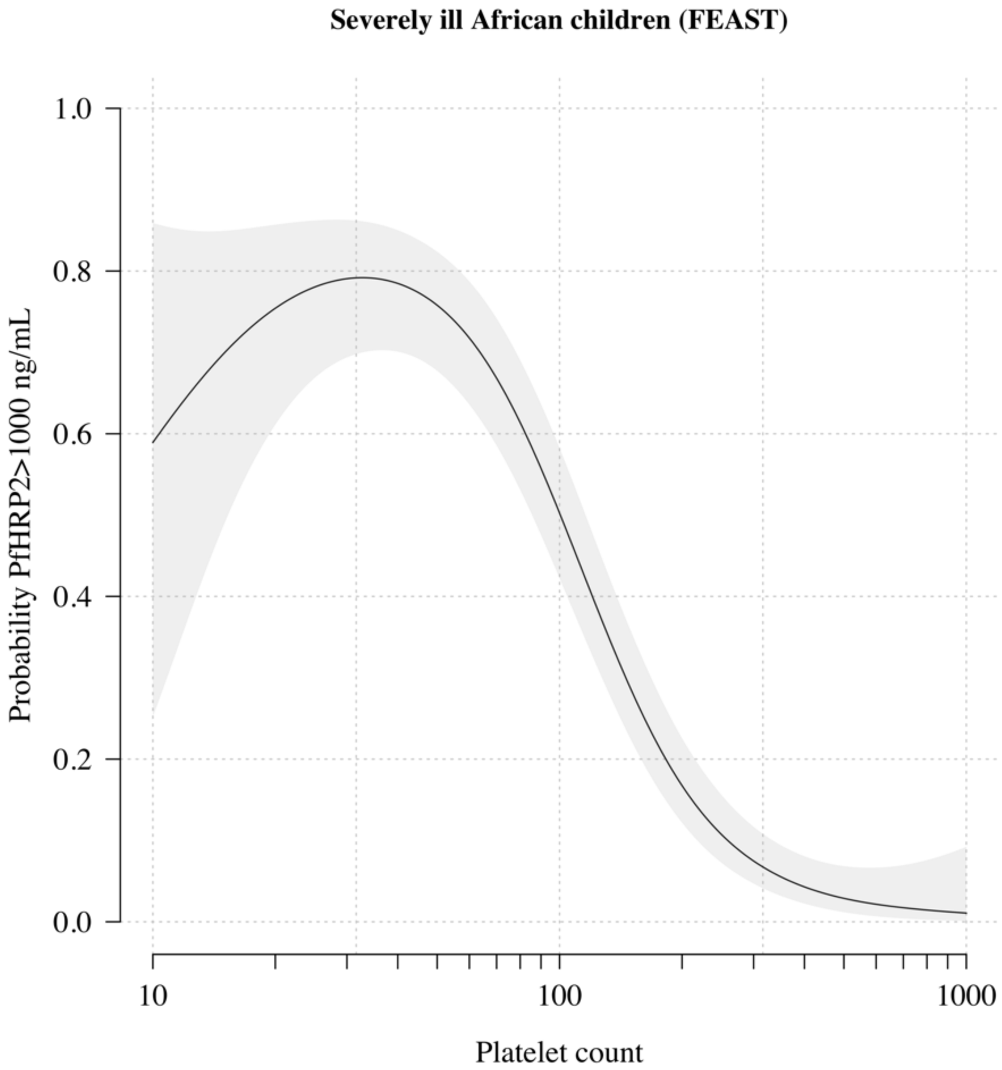

Figure S1: **The relationship between platelet counts and plasma *Pf*HRP2 in severely ill African children.** The black line (shaded area) shows the estimated probability (95% confidence interval), derived from a generalised additive logistic regression model ($p < 10^{-16}$ for the spline term, fit using the R package *mgcv*), that the plasma *Pf*HRP2 $>$ 1,000 ng/mL as a function of $\log_{10}$ platelet count. The generalised additive model was fit to data from 566 African children enrolled in the FEAST trial [21] (all the children who had both platelet counts and *Pf*HRP2 data available). Plasma *Pf*HRP2 $>$ 1,000 ng/mL is highly discriminatory for severe malaria [22].

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.16.440107; this version posted April 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

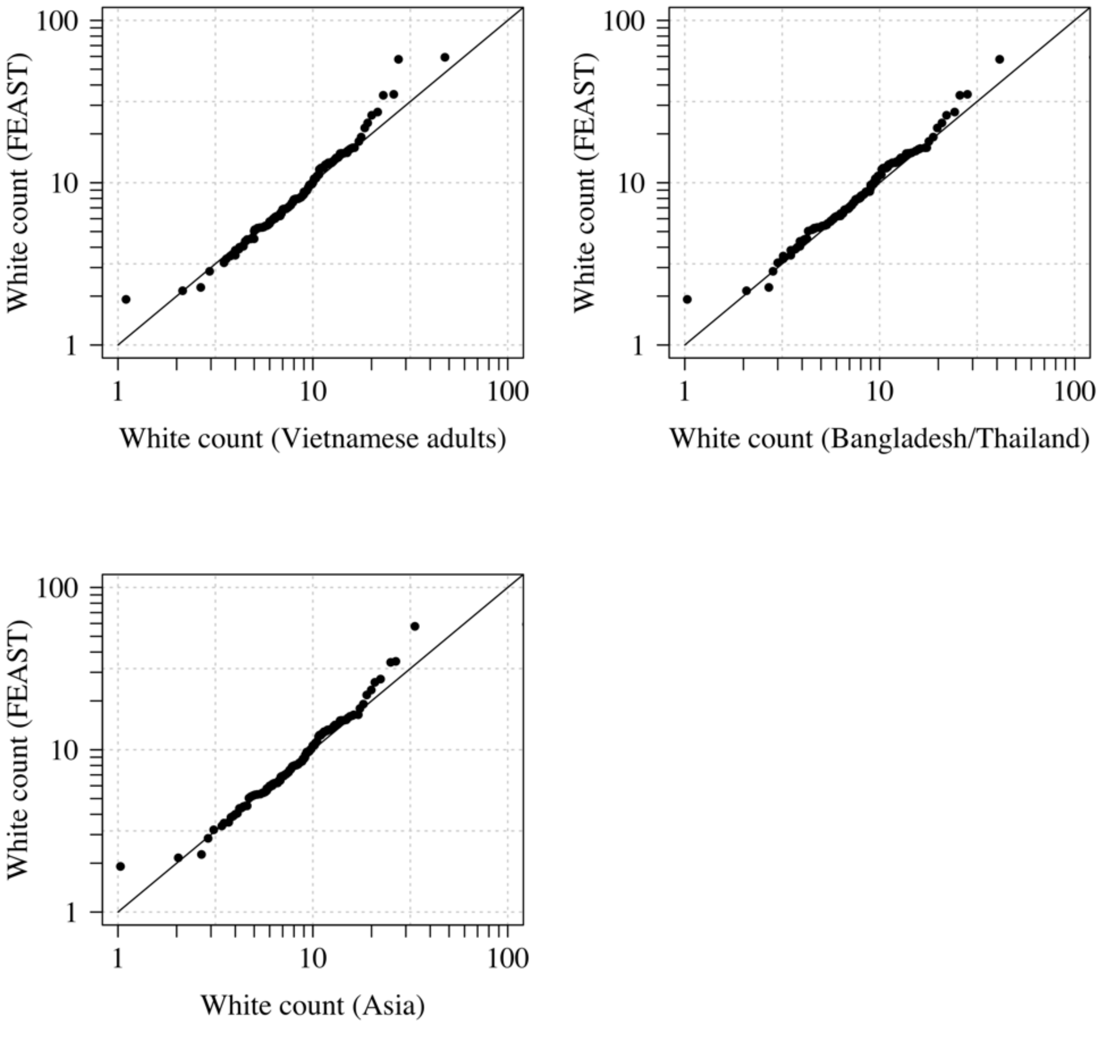

Figure S2: **Comparison of the marginal distributions of white blood cell counts between Asian adults and children with severe malaria and African children with severe malaria.** FEAST: 121 severely ill Ugandan children with *Pf*HRP2 $>$ 1,000 ng/mL [21]. Vietnamese adults: 930 adults from two large randomised trials in severe malaria [20, 19]. Bangladesh/Thailand: 653 adults and children from observational studies of severe malaria [37].

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.16.440107; this version posted April 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

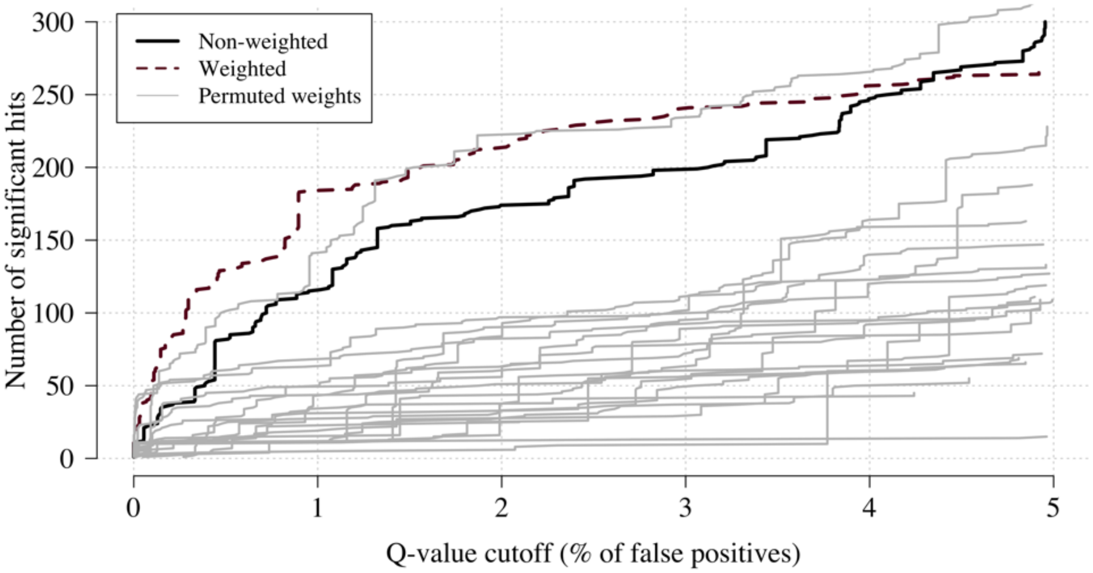

Figure S3: **Effect of permuting the weights in the re-weighted (data-tilting) GWAS.** Here we show the results of 20 random permutations of the weights, applied to the Kenyan case-control GWAS using only chromosomes 4, 9 and 11 (where the top hits are - we limit it to these 3 chromosomes for computational reasons). The random permutations (grey) decrease the number of significant hits compared to the non-weighted (thick black) and the non-permuted re-weighted model (dashed purple).

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.16.440107; this version posted April 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

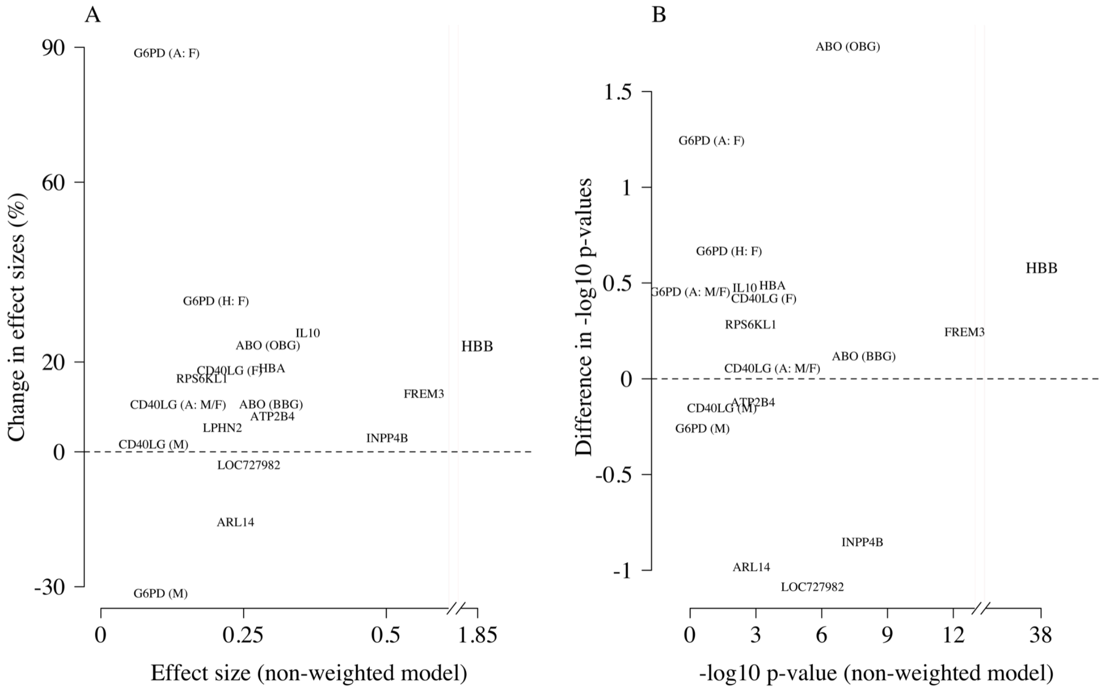

Figure S4: **Comparison of the non-weighted and weighted models of association for directly typed polymorphisms previously reported as associated with severe malaria [27].** Panel A: estimated effect sizes under the non-weighted model versus the difference in effect sizes between the weighted and non-weighted models (absolute effects on the log-odds scale). Differences > 0 imply that the absolute effect size is estimated to be larger under the weighted model. Panel B: -log$_{10}$ p-values under the non-weighted model versus the differences in -log$_{10}$ p-values under the weighted and non-weighted models, again differences >0 represent larger -log$_{10}$ p-values for the weighted model. Each point is represented by the gene name. In each case we use the model that best fit the data in the original analysis [27]. For the X-linked polymorphisms (*G6PD, CD40LG*), multiple models were reported and so the association model is also shown: H (heterozygote); A (additive); M (males only); F (females only); M/F (all).

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.16.440107; this version posted April 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

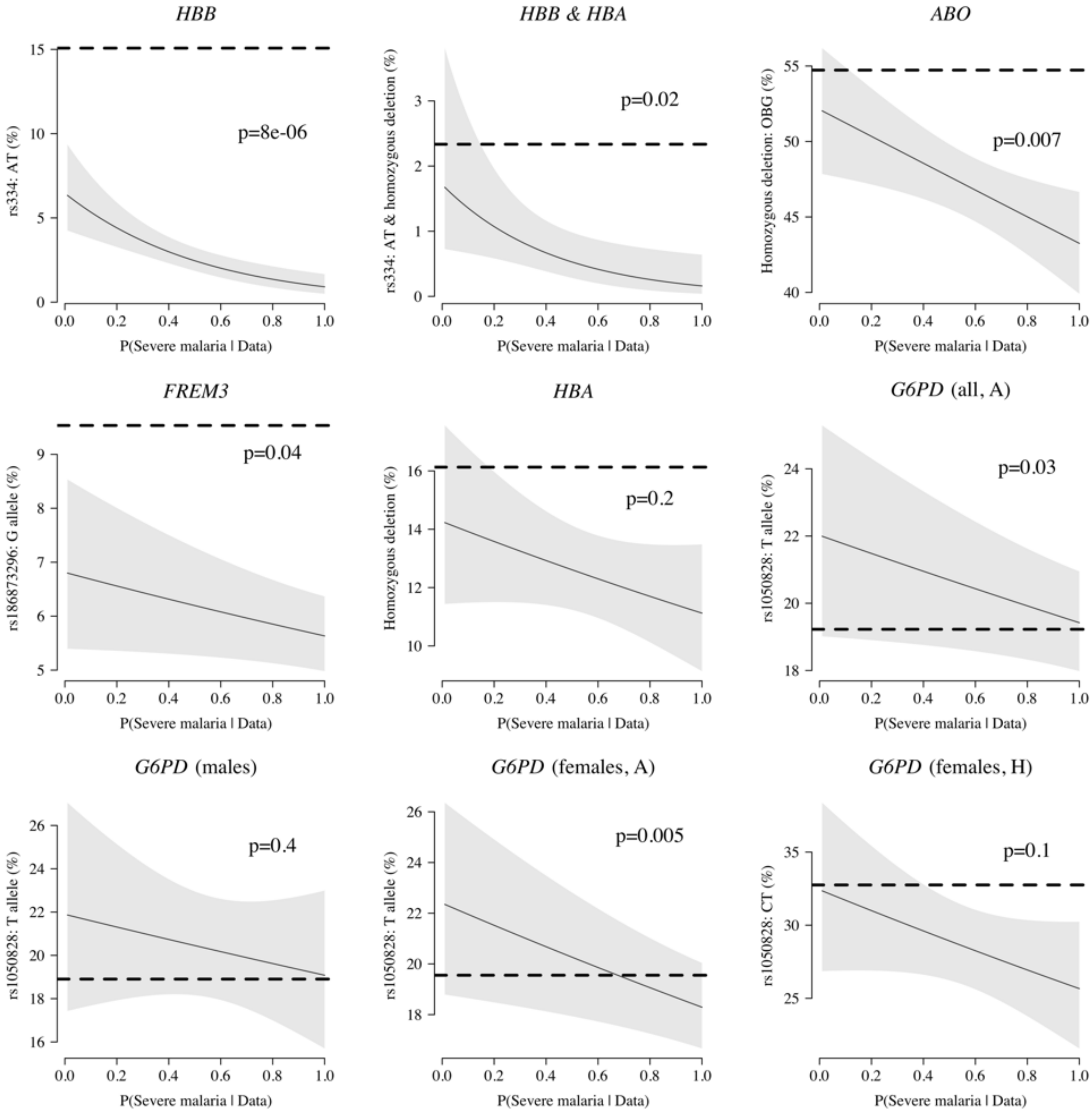

Figure S5: **Case-only analysis of five key polymorphisms effecting red cells, reported in [27] under additive, recessive or heterozygote models.** The horizontal dashed lines show the estimated frequency in the controls (for additive models this is the frequency of the derived allele, for the heterozygote or recessive models this is the frequency of the genotype thought to confer protection). The line (shaded area) show logistic regression fits with P(Severe malaria | Data) as the predictor (95% confidence interval of the fit). The p-value corresponds to the test that the predictor P(Severe malaria | Data) is not associated with the genotype in the cases only. OBG: O Blood Group

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.16.440107; this version posted April 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

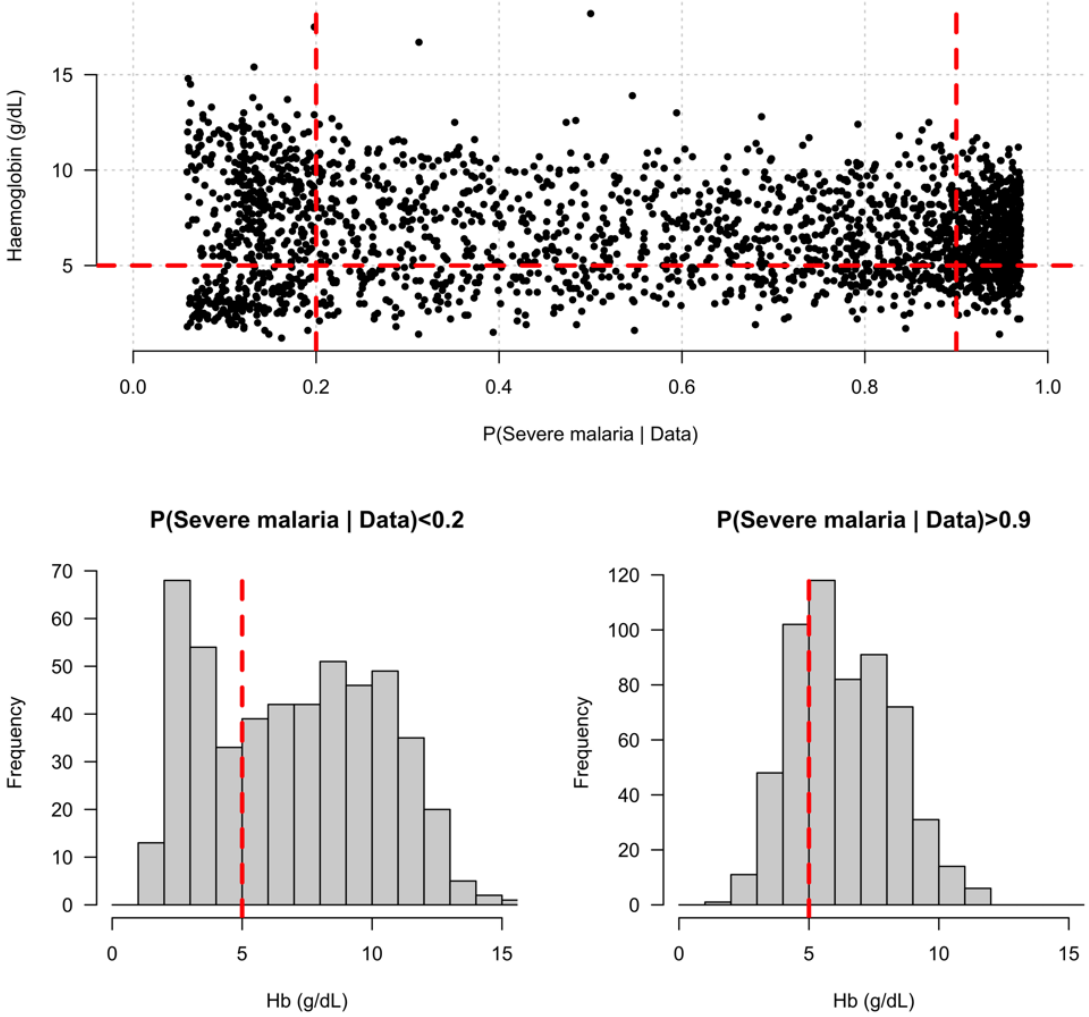

Figure S6: **Distribution of admission haemoglobin concentrations as a function of P(Severe malaria | Data).** Severe anaemia is generally defined as a haemoglobin less than 5 g/dL in African children diagnosed with severe malaria, shown by the horizontal dashed red line in the top panel and the vertical dashed red lines in the bottom panels. The vertical dashed red lines in the top panel show the top and bottom quintiles of the probability distribution (0.9 and 0.2, respectively). Patients in the bottom quintile of the probability distribution had a markedly bi-modal distribution in haemoglobin concentrations with a substantial proportion meeting the severe anaemia criterion and a substantial proportion with relatively high haemoglobin concentrations (> 10 g/dL), suggesting two patients subgroups. Patients in the top quintile had a uni-modal distribution of haemoglobin.

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.16.440107; this version posted April 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

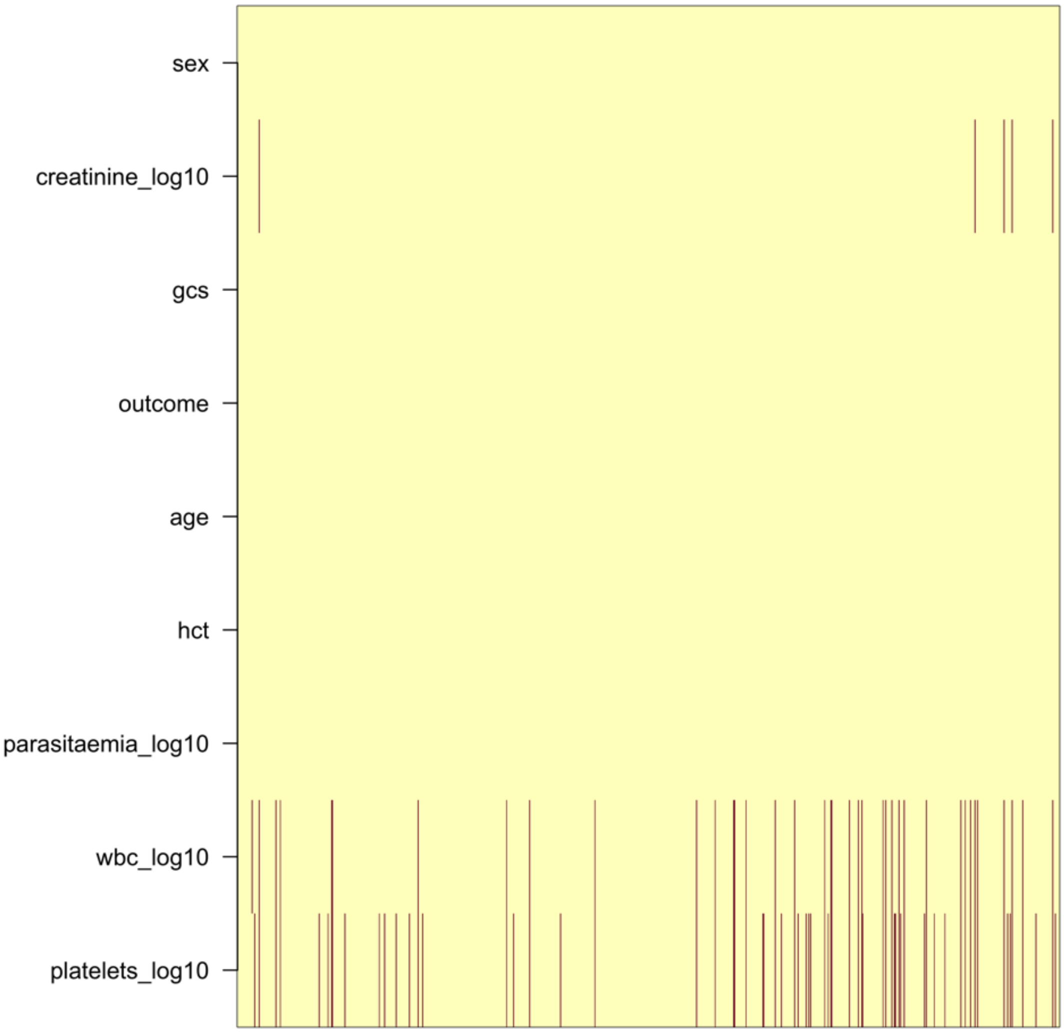

Figure S7: **Pattern of missing clinical data in the 930 Vietnamese adults**. These data pool the AQ Vietnam severe malaria study [19] and the AAV severe malaria study [20] (red: missing; yellow: recorded).

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.16.440107; this version posted April 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

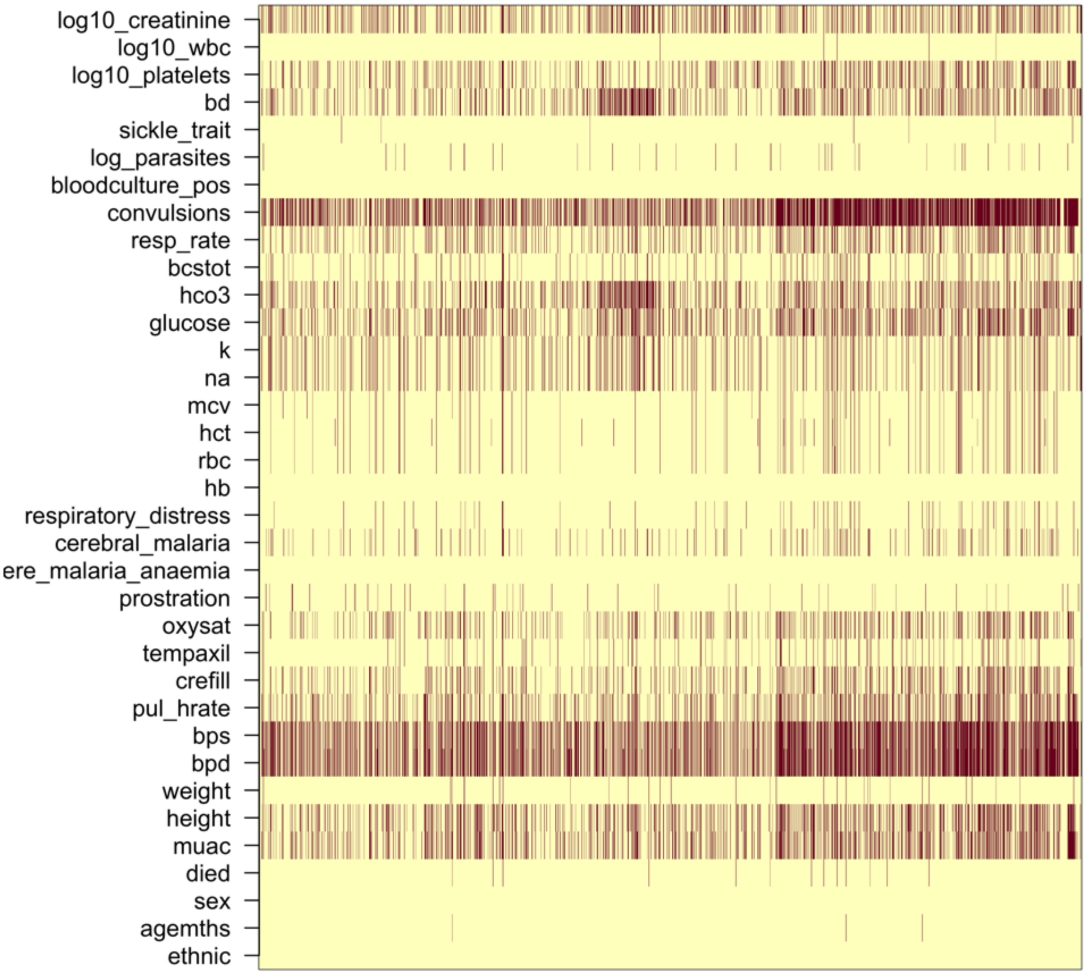

Figure S8: **Missing clinical data in the 2,220 Kenyan children diagnosed with severe malaria** . (red: missing; yellow: recorded).

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.16.440107; this version posted April 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

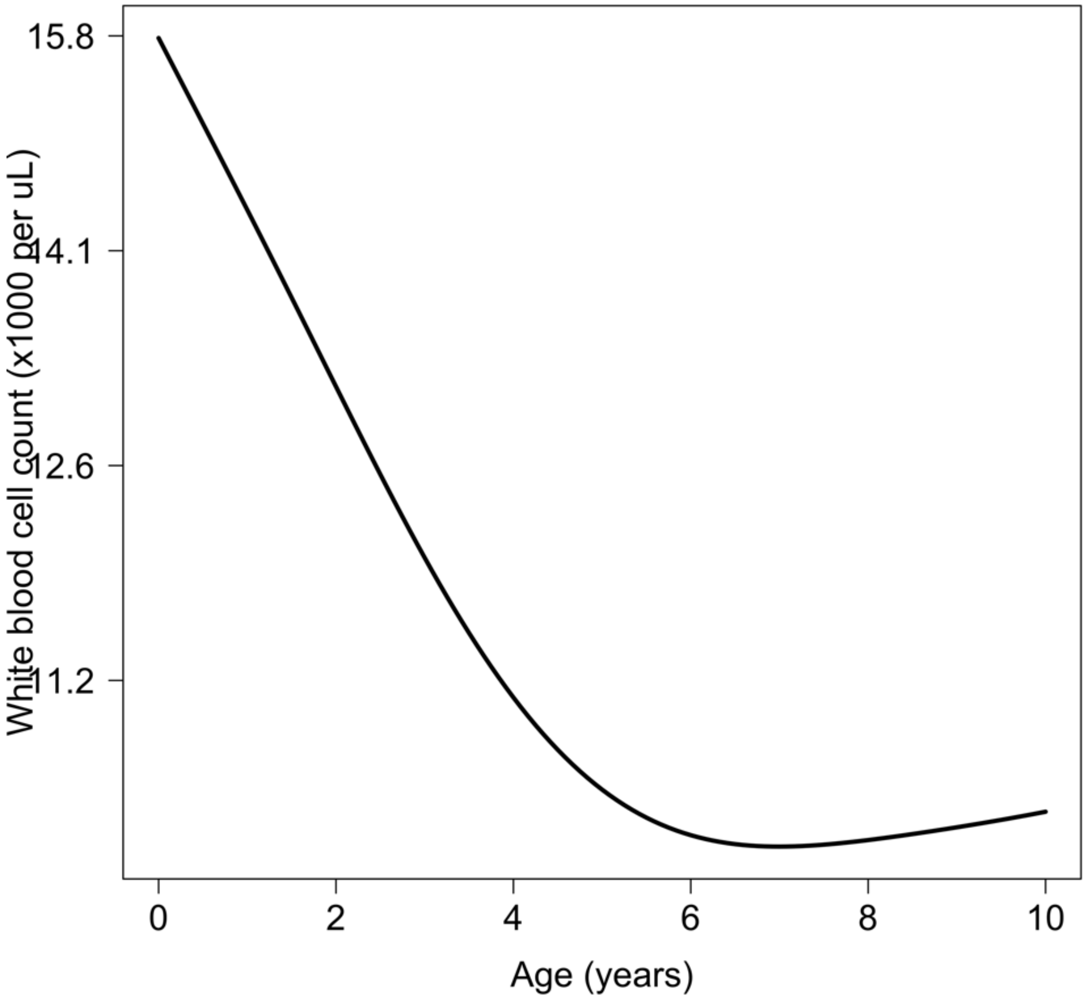

Figure S9: **Relationship between age and mean white count (modelled on the $\log_{10}$ scale).** This is estimated from 858 children in the FEAST trial who had white counts available using a additive linear model ($p = 10^{-8}$ for the smooth spline term). We used this model to adjust observed $\log_{10}$ white counts in all children less than 5 years of age in the training and testing datasets.

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.16.440107; this version posted April 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

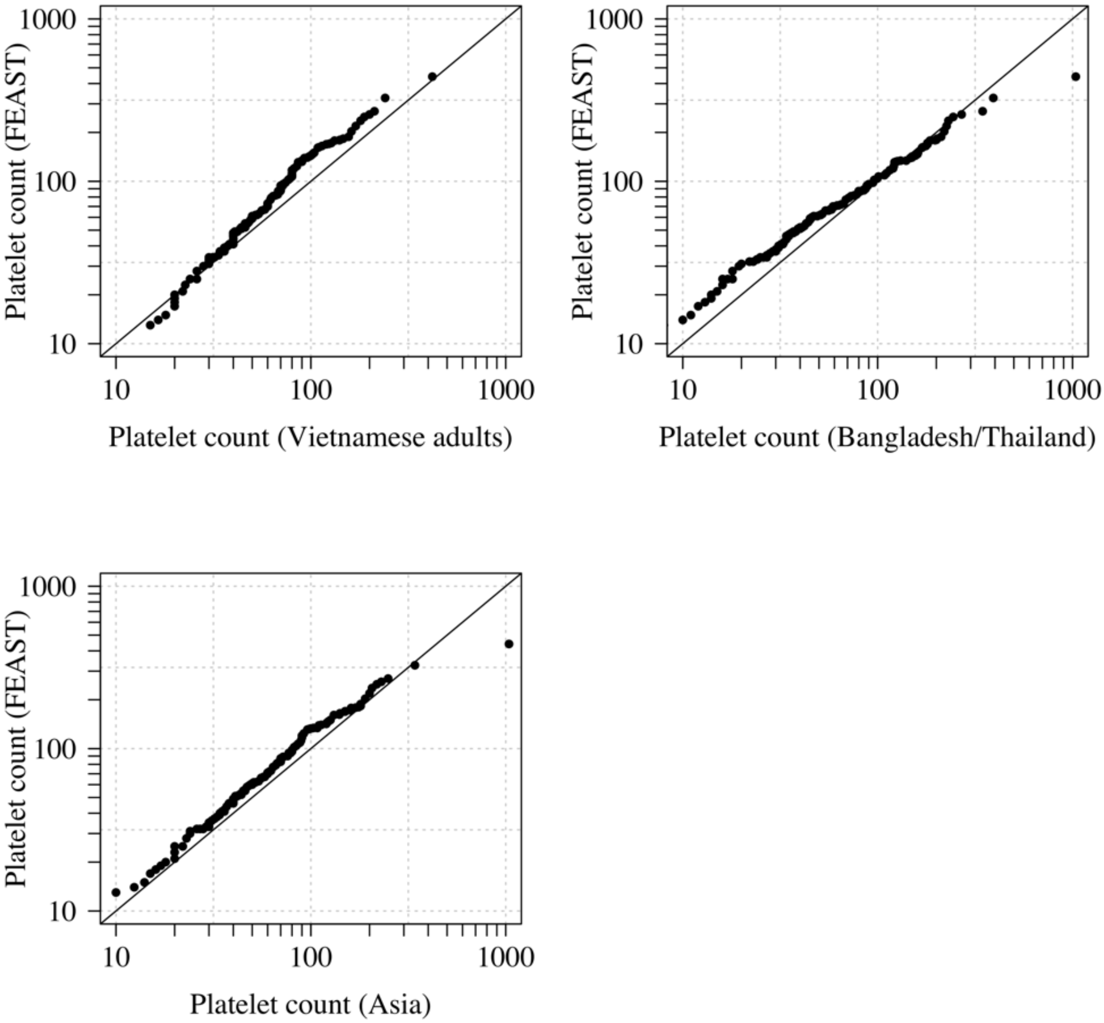

Figure S10: **Comparison of the marginal distributions of platelet counts between Asian adults and children with severe malaria and African children with severe malaria.** FEAST: 121 severely ill Ugandan children with *Pf*HRP2 $>$ 1,000 ng/mL [21]. Vietnamese adults: 930 adults from two large randomised trials in severe malaria [20, 19]. Bangladesh/Thailand: 653 adults and children from observational studies of severe malaria [37].

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.16.440107; this version posted April 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

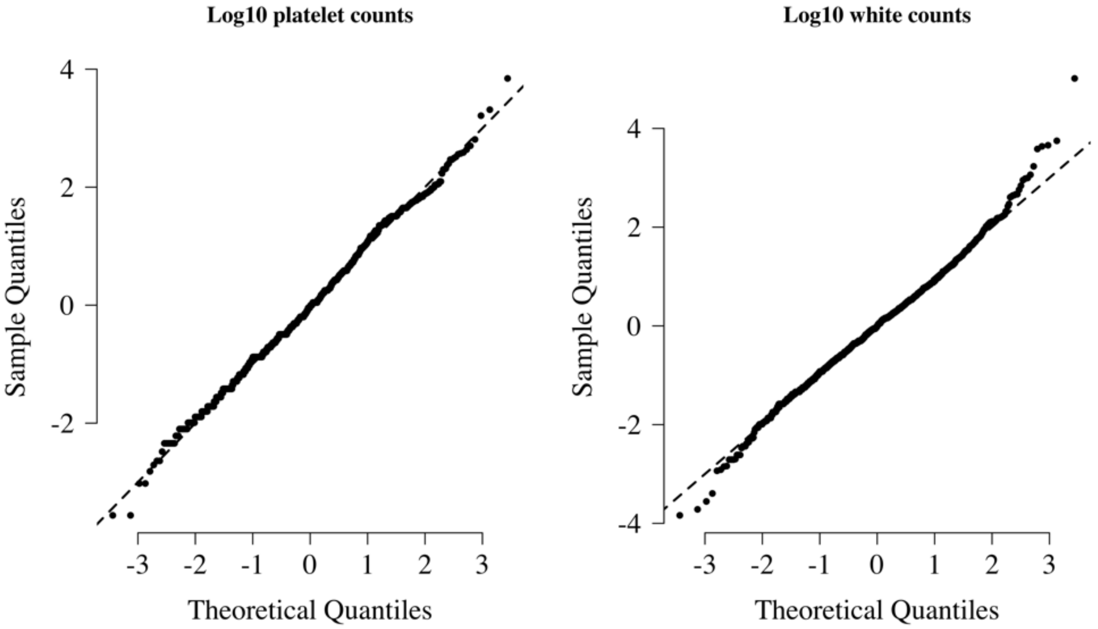

Figure S11: **Normal-quantile plots for platelet counts and white blood cell counts in the training data**. Both were standardised to have mean 0 and standard deviation of 1 on the $\log_{10}$ scale. The diagonal lines shows the identity line.

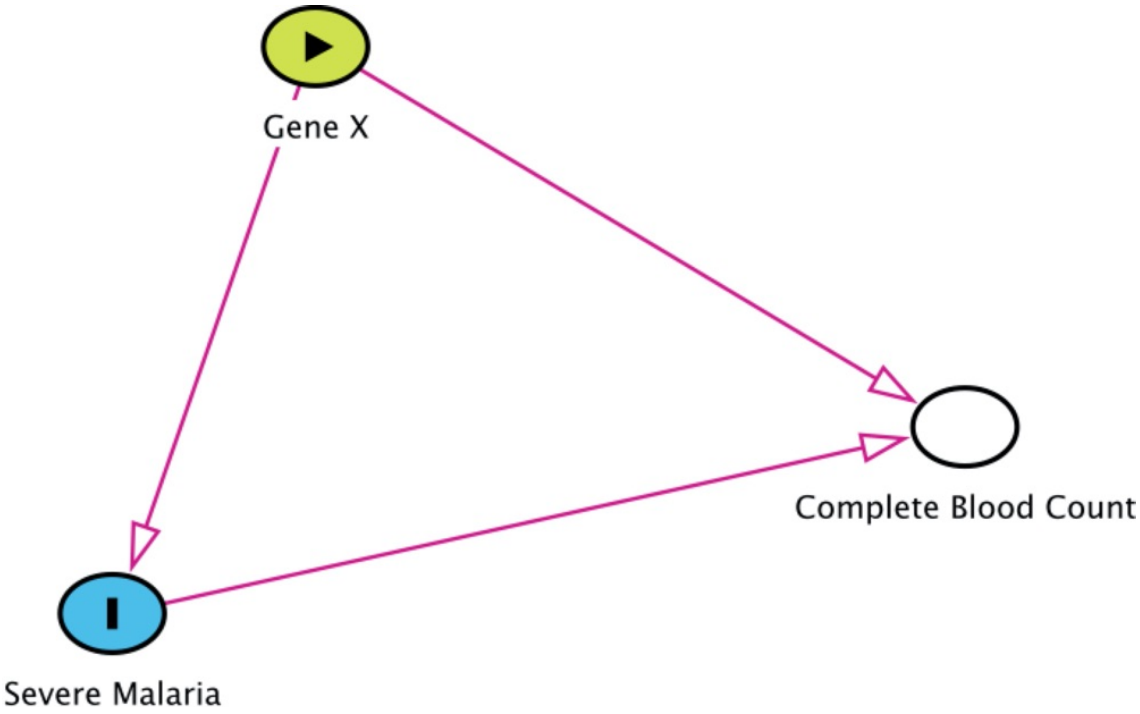

Figure S12: **Collider bias in the diagnostic model of severe malaria based on complete blood count data.** *HBB* in its homozygous S form (HbSS, <1% prevalence in this Kenyan population) is a rare example of how this can occur. Children with HbSS have white counts above 2-3 times higher than the normal population and slightly lower platelet counts [34]. Under the probabilistic model, all 11 children with HbSS were classified as having a low probability of severe malaria, based on their high white counts (mean 40,000 per μL). These probabilities cannot be taken at face value and it remains an unanswered question whether children with HbSS are more or less susceptible than their wild-type counterparts [35]

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.16.440107; this version posted April 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

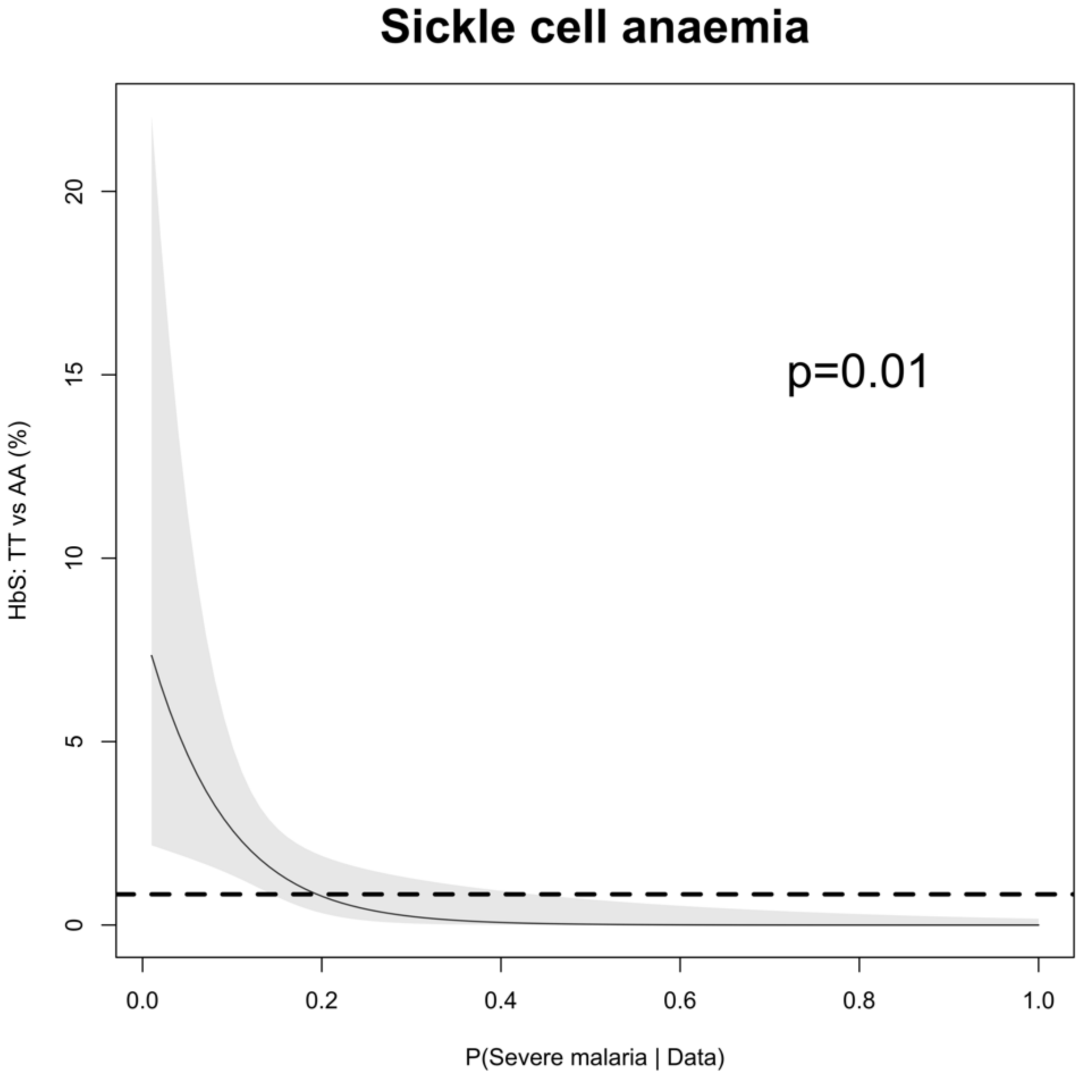

Figure S13: **The relationship between HbSS and the estimated probabilities of severe malaria under the diagnostic model.** There were 11 children with HbSS and they all had low probabilities of severe malaria but this is biased as these children have chronic inflammation with white counts 2-3 higher than the general population [34] (see Figure S12 for the causal diagram showing collider bias).

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.16.440107; this version posted April 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

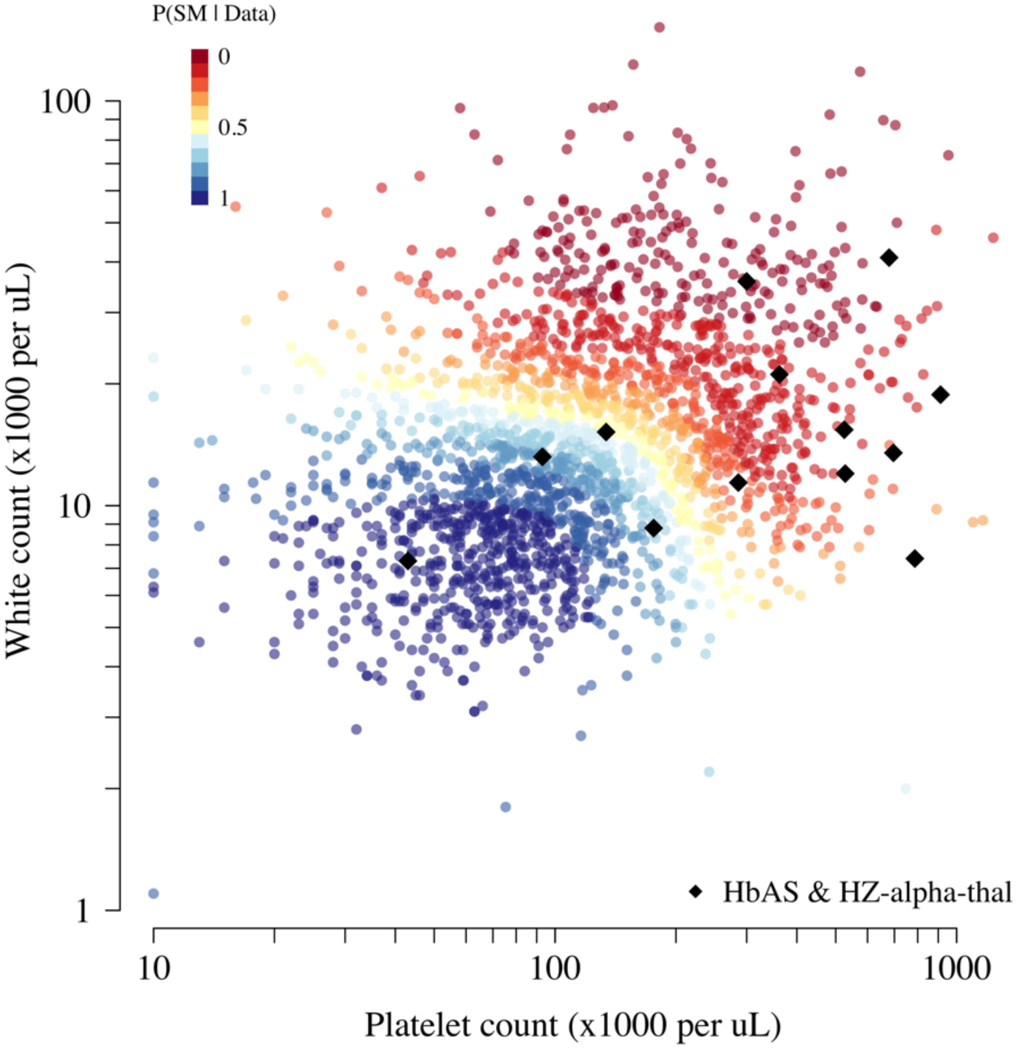

Figure S14: Scatter plots of platelet counts versus white blood cell counts for the Kenyan cohort, showing the 13 individuals with the double mutation HbAS & homozygous α^{+}-thalassaemia as large black diamonds (HZ-alpha-thal)). The red-yellow-blue colour scheme is proportional to the P(Severe malaria | Data) as given by the legend in the top left corner.

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.16.440107; this version posted April 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

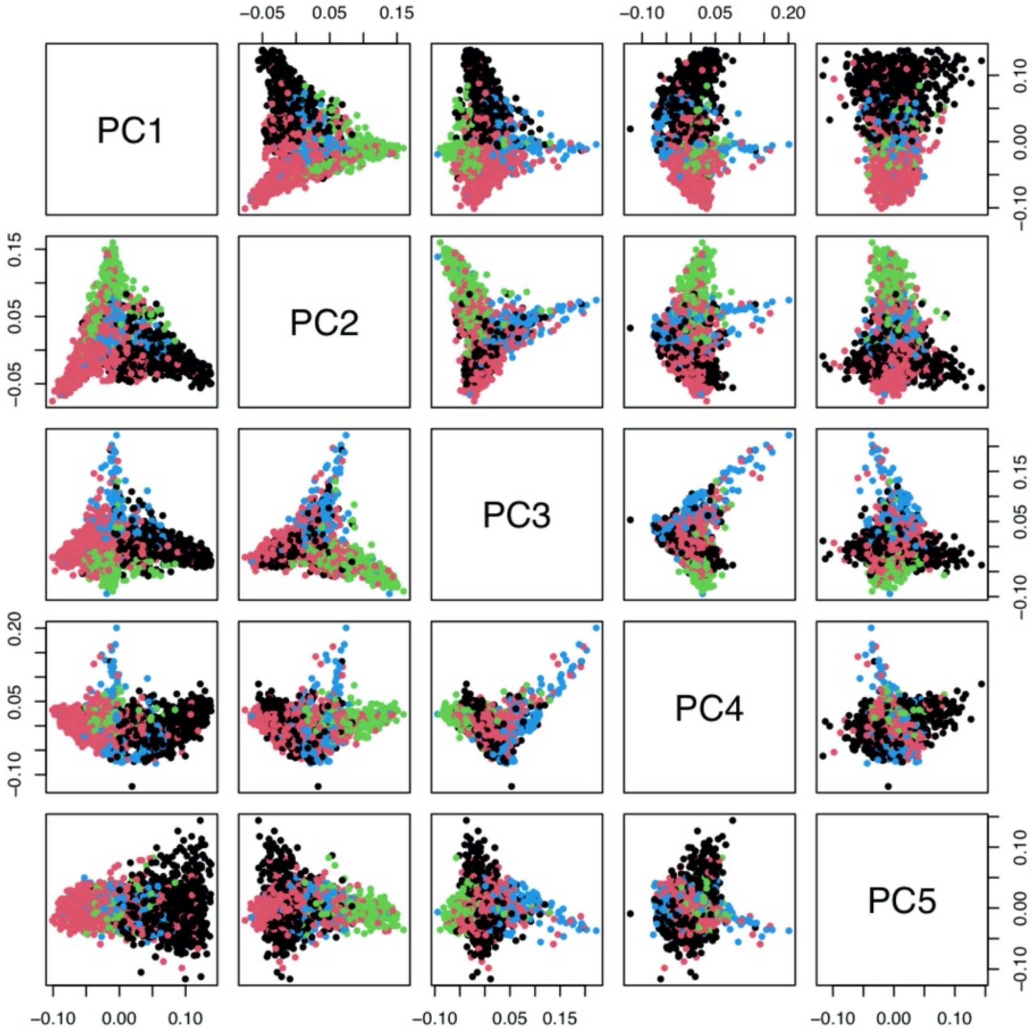

Figure S15: Principal components analysis of 1,666 Kenyan cases and 1,606 population controls. The colours show the main self-reported ethnicities (black: Chonyi; red: Giriama; green: Kauma; blue: other). The first 5 principal components were used to stratify for population structure in the GWAS analyses.