USAID
FROM THE AMERICAN PEOPLE

DHS WORKING PAPERS

Analysis of Sickle Cell Genotypes of Young Children in Nigeria Using the 2018 DHS Survey

Thomas W. Pullum

2020 No. 175

DEMOGRAPHIC AND HEALTH SURVEYS

September 2020

DHS Working Papers No. 175

Analysis of Sickle Cell Genotypes of Young Children in Nigeria Using the 2018 DHS Survey

Thomas W. Pullum

The DHS Program
ICF
Rockville, Maryland, USA

September 2020

Corresponding author: Thomas W. Pullum, The DHS Program, ICF, 530 Gaither Road, Suite 500, Rockville, MD 20850, USA; phone: +1 301-407-6500; fax: +1 301-407-6501; email: tom.pullum@icf.com

Acknowledgments: The author thanks Adebowale Adeyemo (National Institutes of Health), as well as Fred Arnold, Joanna Lowell, and Cameron Taylor (DHS), for very helpful comments on an earlier draft.

Editor: Diane Stoy
Document Production: Chris Gramer

This study was conducted with support from the United States Agency for International Development (USAID) through The DHS Program (#720-OAA-18C-00083). The views expressed are those of the authors and do not necessarily reflect the views of USAID or the United States Government.

Recommended citation:

Pullum, Thomas W. 2020. *Analysis of Sickle Cell Genotypes of Young Children in Nigeria Using the 2018 DHS Survey*. DHS Working Papers No. 175. Rockville, Maryland, USA: ICF.

CONTENTS

TABLES

FIGURES

ABSTRACT

A survey conducted in Nigeria in 2018 as part of The Demographic and Health Surveys Program included sickle cell genotyping of a subsample of children age 6-59 months. The survey was the first population-based household survey to include sickle cell genotyping of children at a national level. This working paper provides continued analysis of the data that was not possible in the main survey report. The paper examines the spatial distribution of the problematic genotypes and alleles and their relationship with indicators of child health, and compares with evidence from clinical studies of higher mortality for the SS and SC genotypes, and lower mortality for the AS genotype, which has a known protective effect against malaria. We find serious limitations in the capacity of these data to quantify differences in the mortality risk from sickle cell disease, malaria, or any other cause. However, we found that the siblings of genotyped children with sickle cell disease are about 2.5 times as likely to have died as the siblings of other genotyped children. The only significant relationship with childhood illness is with severe anemia. The main value of the data is the description of the spatial distribution of the genotypes and alleles within Nigeria. The S and C genes are primarily concentrated in most states in the South West Zone, including Lagos. Spatial information such as this will be informative for education and intervention campaigns.

Key words: sickle cell genotype, Nigeria

1 INTRODUCTION

The Demographic and Health Surveys Program (DHS), which has been funded primarily by the United States Agency for International Development (USAID) since 1984, is a major source of estimates of demographic and health indicators in low- and middle-income countries. In recent years, these household surveys have included biomarkers based on the analysis of blood samples. A DHS survey conducted in Nigeria during August–December of 2018 was the first national household survey in any country to include sickle cell genotyping. Children age 6-59 months in a subsample of households were eligible for genotyping.

The main report of that survey (NPC and ICF 2019) included a description of the distribution of the genotypes by background characteristics. The schedule for the release of the report and data files prevented an in-depth analysis. This working paper provides that analysis, although it does not include all possible approaches. The goal is to determine the potential value of the data and to provide guidance for further research and data collection. Future surveys in Nigeria or in other countries affected by sickle cell anemia may benefit from an exploration of the data in this survey.

In Nigeria, DHS surveys were conducted in 1986 (limited to Ondo State), 1990, 2003, 2008, 2013, and 2018. The DHS Program also provided support for a Malaria Indicator Survey (MIS) in 2010 and 2015. Nigeria has the largest population in sub-Saharan Africa, and the Nigeria DHS surveys have had relatively large sample sizes. The 2018 survey included 40,427 households, of which 13,514 were selected with systematic sampling for a longer interview that included hemoglobin and genotype testing. These households included 11,590 children age 6-59 months. A valid genotype was obtained from 11,210 children (unweighted counts) after informed consent for the testing was obtained from a parent or other individual responsible for the child. Although not an MIS, the 2018 survey also included microscopy and rapid diagnostic testing (RDT) for malaria for children in this age range.

The main report (NPC and ICF 2019, p. 265) included the following two paragraphs that are quoted verbatim. The first paragraph provides a justification for the genotyping and the second gives a detailed description of the procedure.

“The 2018 NDHS, for the first time in a DHS survey, collected information on sickle cell disease (SCD) and sickle cell trait (SCT). Various sources have pointed to sickle cell disease being a major public health issue in Nigeria. The prevalence of sickle cell trait ranges between 10% and 45% in various parts of sub-Saharan Africa (WHO AFRO 2013). The National Strategic Plan of Action on Prevention and Control of Non-Communicable Diseases under Nigeria’s Federal Ministry of Health has estimated that approximately 24% of Nigerians have SCT (Federal Ministry of Health 2015a). Also, it is estimated that when the prevalence of SCT is above 20%, SCD can be as high as 2% (Federal Ministry of Health 2015a). According to this estimation, over 3.4 million Nigerians currently have SCD (Federal Ministry of Health 2015a). This disorder manifests early in life and has diverse clinical complications, including cardiovascular and renal diseases, thus fueling major noncommunicable diseases (NCDs). In addition, SCD patients experience different degrees of stigmatization and discrimination in society. Although a policy on universal newborn screening was introduced in 2011 in Nigeria, the policy needs to be updated to accommodate recent knowledge and trends in detection and treatment of the disease.”

"Blood collection for genotype testing was carried out in a subsample of 14,000 households selected for the men's survey. In total, 11,536 (unweighted) children were eligible for the test, of whom 97% were successfully tested. In the 25% of households where genotype testing was done, a confirmatory test was conducted. The test was done in the standard laboratory for high-performance liquid chromatography (HPLC) confirmatory testing at the International Foundation Against Infectious Disease in Nigeria (IFAIN) in Abuja. Test results obtained from SickleSCAN were compared with the HPLC diagnostics. The results of the comparison showed a diagnostic sensitivity of 85%, a specificity of 98%, a positive predictive value of 91%, and a negative predictive value of 96%. These diagnostic results indicate that the estimates obtained from the SickleSCAN are valid."

As a minor correction of the second paragraph quoted above, the subsample for genotyping was one-third, rather than one-fourth, of households. We do not have access to the results of the HPLC confirmatory test. The quoted sensitivity (true positive rate) of 85% and specificity (true negative rate) of 98% suggest that the data tended to underestimate the true prevalence of the S and C alleles.

Table 1 Unweighted and weighted frequency distribution across genotypes. Nigeria DHS 2018.

Genotype	Unweighted n	Weighted n	% (wtd)
Normal (AA)	8,700	8,779.8	77.20
Sickle cell trait (AS)	2,186	2,242.7	19.72
Hb C trait (AC)	155	185.2	1.63
Hb C disease (SC)	34	50.5	0.44
Sickle cell anemia (SS)	102	100.0	0.88
Other	9	14.4	0.13
Total	11,186	11,372.6	100.00

Among the children age 6-59 months with *de facto* household residence who were genotyped in the Nigeria 2018 survey, the distribution across genotypes is given in Table 1.[1]

The data files include sampling weights. For unbiased population estimates, weights should be used. The weights compensate for several aspects of the design and implementation of the survey, including the oversampling of small strata and undersampling of large strata, disparities between the number of households in the sampling frame and the number actually identified in the listing of the cluster, and nonresponse rates. Table 1 shows the differences between the weighted and unweighted numbers of children with different genotypes, which are small.

Children in the age range of 6-59 months (or 0-59 months) are the focus of many indicators and much of the standard data collection in DHS surveys. These children are about 6% of the entire population of Nigeria. The prevalence of the different genotypes among these children cannot be extrapolated to the entire population, primarily because the genotypes are associated with different levels of mortality. We can expect that by the time this cohort of children reaches adulthood, the proportion with high-risk genotypes will be

[1] Some numbers in this report do not exactly match the main report. For example, Table 1 includes 11,373 children (weighted), but Table 11.9 in the main report includes 11,391. We are unable to resolve this discrepancy, which may be due to decisions about inclusion or exclusion of children who were not in the stated age range. Here, we include children who were *de facto* residents, which means that they slept in the household the night before the household interview, and were age 6-59 months. Tests were incorrectly conducted on 33 children under age 6 months (mostly 5 months). Some of these children may have been included in Table 11.9 in the main report, but none of them are included here.

smaller than at the time of the survey. The high-risk genotypes are probably less prevalent in the data than they were at birth.

In this report, the genotypes will generally be referred to with the letter combinations of the alleles A, S, and C. For example, we refer to "sickle cell trait" as "AS". The "other" category will sometimes be identified with "CC", the only remaining combination of A, S, and C, but "other" may also include ambiguous or indeterminate results. The term "sickle cell disease" applies to the SS and SC combinations.

The Nigeria 2018 survey had a stratified two-stage sample design, in which the strata were the combinations of place of residence (urban/rural) and state. Within each stratum, using the most recent national census as the sampling frame, census enumeration areas were sampled with probability proportional to size. These areas are the primary sampling units and are described as clusters. Selected clusters were visited and a listing of all households was prepared. Approximately 30 households in each cluster were then selected with systematic sampling. All members of the households, *de jure* as well as *de facto*, were listed in the household schedule. Some information was obtained from a household respondent about all members of the household, and subsequently all women age 15-49 were interviewed individually. A systematic subsample of about one-third of the households was selected for a longer version of the Women's Questionnaire, including a domestic violence module, interviews with all men age 15-59 in the household, and additional biomarkers for young children, including sickle cell genotyping of children age 6-59 months.

This report is related to other reports produced by DHS, such as a report that analyzed the quality of data on the ages and health of young children (Pullum 2008), and a report on the quality of the hemoglobin data for children and women (Pullum et al. 2017).

This report will explore three topics, with the following research questions:

- What is the geographic distribution of the genotypes?
- How are the genotypes associated with available indicators of child health?
- Is there evidence of different mortality for different genotypes?

These topics will be considered in Sections 2, 3, and 4. Section 5 will provide conclusions and some recommendations for future analysis and data collection.

2 GEOGRAPHIC VARIATION IN THE GENOTYPES

Variation in the distribution of genotypes can be interpreted in two ways. First, the genotypes may be adaptive: the prevalence of different genotypes may be associated with different environments, recently or in the distant past. For example, if the S allele is protective against malaria, it may be more common in geographic areas where malaria is endemic. It is potentially misleading to use survey data to investigate this possibility. Because of internal migration within Nigeria, children's current place of residence may be very different from their ancestors' location. The current distribution of malaria endemicity may not be the same as in the past, when such adaptation was taking place. In recent decades, premarital genetic testing may have led to reductions in marriages between men and women who both have genotype AS. Therefore, we do not assess the geographic correspondence between genotype and endemicity, but only describe the geographic distribution of the genotypes.

Second, in the alternative causal direction, the genotypes may be a source of variation in the risk of negative outcomes. This possibility requires analysis of the individual child as the unit of analysis. The child's genotype will be treated as a predictor of different kinds of child health outcomes in Chapter 3.

In both types of analysis, particularly the latter, the analysis is hampered by the limitation of the data to surviving children. Mortality is the ultimate negative health outcome, and because we do not know the genotype of children who died, we can only describe statistical relationships for the surviving children.

We will describe geographic variation in the genotype profile at three levels of analysis: the zone, the state, and the sample cluster, which is a relatively homogeneous census enumeration area.

2.1 Genotype Variation across Zones

As shown in Figure 1, Nigeria has six zones: North Central, North East, North West, South East, South South, and South West. There are massive health and economic disparities across the zones, which are not described here. In a broad sense, the southern zones have better outcomes for children than the northern zones. South West is the most prosperous zone and includes Lagos, the largest city. The distribution of the genotypes within each zone is shown in Table 2. In all zones, the large majority of genotypes are AA. Overall, 77.2% of children are AA. In all zones, the next largest genotype is AS, which includes 19.7% of children.

Table 2 Percentages of the genotypes in the six zones. Children age 6-59 months, Nigeria DHS 2018. Weighted.

Zone	AA	AS	AC	SC	SS	Other
North Central	79.07	17.82	1.76	0.36	0.94	0.05
North East	78.02	20.47	0.33	0.27	0.91	0.00
North West	77.56	19.97	1.25	0.16	1.01	0.05
South East	79.74	19.11	0.00	0.13	1.02	0.00
South South	80.35	19.09	0.24	0.00	0.32	0.00
South West	70.81	20.96	5.23	1.60	0.82	0.58
Total	77.20	19.72	1.63	0.44	0.88	0.13

Figure 1 Map of Nigeria, showing the zones and states at the time of the 2018 DHS survey

A clearer picture of the composition of each zone is given by the relative prevalence of the A, S, and C alleles that is implied by the frequencies of the six genotypes in Table 1.[2] Use nAA, for example, for the weighted frequency of genotype AA, and n for the total of all genotypes. The estimate of Pr(A) is Pr(A) = [nAA+(nAS+nAC)/2]/n. Similarly, Pr(S) = [nSS+(nAS+nSC)/2]/n and Pr(C) = [nCC+(nAC+nSC)/2]/n. The estimates add to 1. Rounding to two decimal places, these formulas yield Pr(A)=0.88, Pr(S)=0.11, and Pr(C)=0.01.[3] These values are consistent with other national estimates for Nigeria (Piel et al. 2010).

The distributions of the alleles are shown in Table 3 and are also in horizontal bar graphs in Figure 2.[4] This figure includes separate subgraphs for A, S, and C because the alleles vary so much in their respective levels. The horizontal axes range from 0 to 1, 0 to 0.15, and 0 to 0.04, respectively. Each figure includes a red vertical line for the national mean. The prevalence of A is virtually identical across all zones other than South West, where it is slightly lower. The lower prevalence of A in South West is a result of a slightly higher prevalence of S, and primarily by a much higher prevalence of C than is found in any other zone—slightly over 0.04. The prevalence of S is in a very narrow range that rounds to 0.10-0.12. The prevalence of C rounds to 0.01 in North Central and North West, and to 0.00 in North East, South East, and South South.

[2] The small number of cases in the "other" category will be treated as "CC" for this calculation. We use the weighted numbers in Table 2, although it makes virtually no difference whether the distribution is weighted or not.

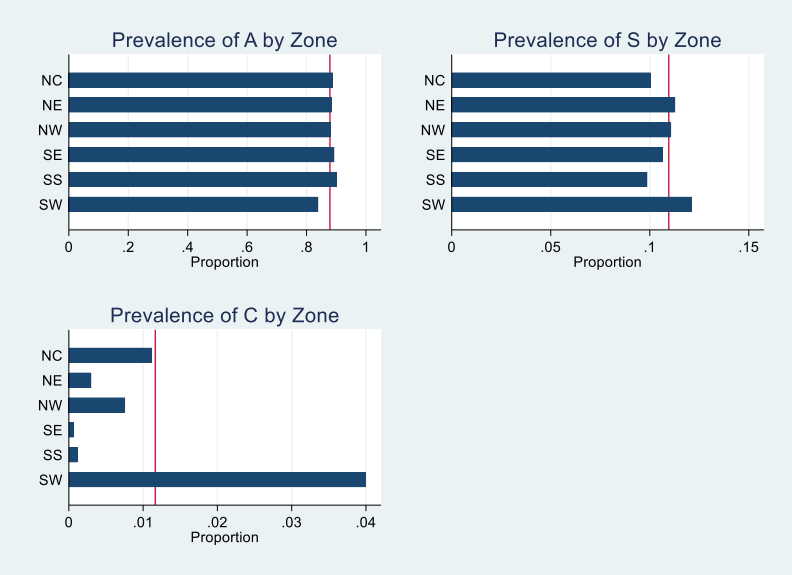
[3] With more decimal places, Pr(A)=0.8788, Pr(S)=0.1096, and Pr(C)=0.0116.

[4] It is conventional for a distribution of genotypes to be shown with percentages and a distribution of alleles with proportions.

Table 3 **Prevalence of the A, S, and C alleles in the six zones (in proportions). Children age 6-59 months, Nigeria DHS 2018. Weighted.**

Region	A	S	C
North Central	0.8886	0.1003	0.0111
North East	0.8842	0.1128	0.0030
North West	0.8817	0.1107	0.0076
South East	0.8930	0.1064	0.0007
South South	0.9001	0.0987	0.0012
South West	0.8390	0.1210	0.0399
Total	0.8788	0.1096	0.0116

Figure 2 **Prevalence of the A, S, and C alleles in the six zones. Children age 6-59 months, Nigeria DHS 2018. Weighted.**



2.2 Genotype Variation across States

Nigeria has 37 states, which include the Federal Capital Territory (FCT). The distribution of genotypes within the states is provided in Appendix Table B1. Here we present figures that show the percentages of the A, S, and C alleles in each state, with the states grouped by zone. There are six figures—Figures 3(a) through 3(f)—one for each zone. Each figure contains three bar graphs, analogous to those in Figure 2, which show the proportions of A, S, and C, respectively, within the state. The horizontal scales are the same in each zone. For A, the horizontal axis ranges from 0 to 1; for S, from 0 to 0.15; and for C, from 0 to .05. In each state, or zone, or at the national level, the three proportions add to 1. Since the variation in A can be attributed entirely to variation in S and C, it is sufficient to focus on the patterns of those two alleles.

Each figure includes a red vertical line that indicates the national mean prevalence and a green vertical line for the mean prevalence for the zone. If the green line is to the left of the red line, the mean for the zone is *less than* the national mean. If the green line is to the right of the red line, the mean for the zone is *greater than* the national mean. The heights of the blue bars can be assessed relative to these horizontal lines.

Figure 2 shows that the proportion of children with the C allele is below 0.01 in three zones: North East, South East, and South South. Figures 3(b), 3(d), and 3(e) show that the proportion is below 0.01 in every

state within those three zones. There are only eight states in Nigeria in which the prevalence of C (rounded to the nearest multiple of 0.01) exceeds 0.01. The prevalence rounds to 0.02 in Sokoto (North West) and Niger (North Central); 0.03 in Kebbi (North West); 0.04 in Ogun (South West); and 0.05 in Lagos, Oyo, and Osun (South West) and Swara (North Central). Of the five states with the highest prevalence of C, four (Ogun, Lagos, Oyo, and Osun) are in the South West Zone. Within that zone, there is a sharp contrast between those four states and Ekiti and Ondo, where the prevalence of C is much lower.

Figure 3(a) Prevalence of the A, S, and C alleles in children age 6-59 months, by state, in the North Central Zone. Nigeria DHS 2018. Weighted.

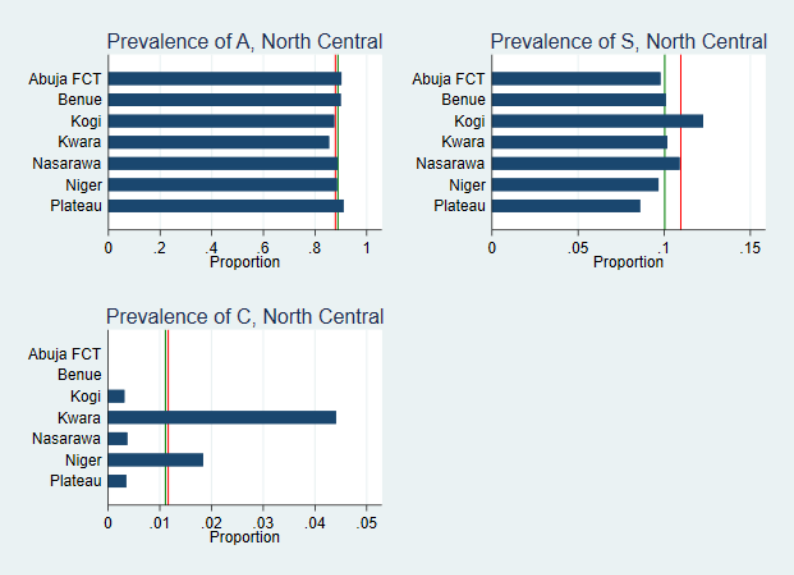


Figure 3(b) Prevalence of the A, S, and C alleles in children age 6-59 months, by state, in the North East Zone. Nigeria DHS 2018. Weighted.

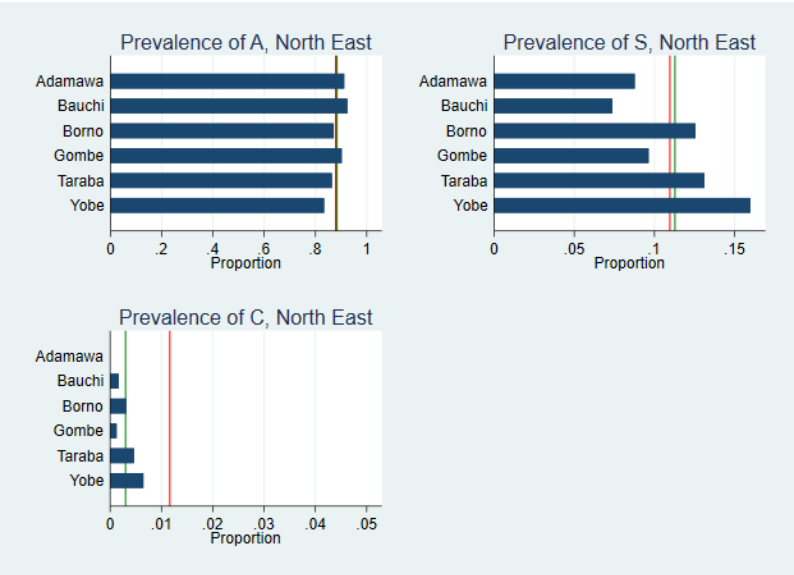


Figure 3(c) Prevalence of the A, S, and C alleles in children age 6-59 months, by state, in the North West Zone. Nigeria DHS 2018. Weighted.

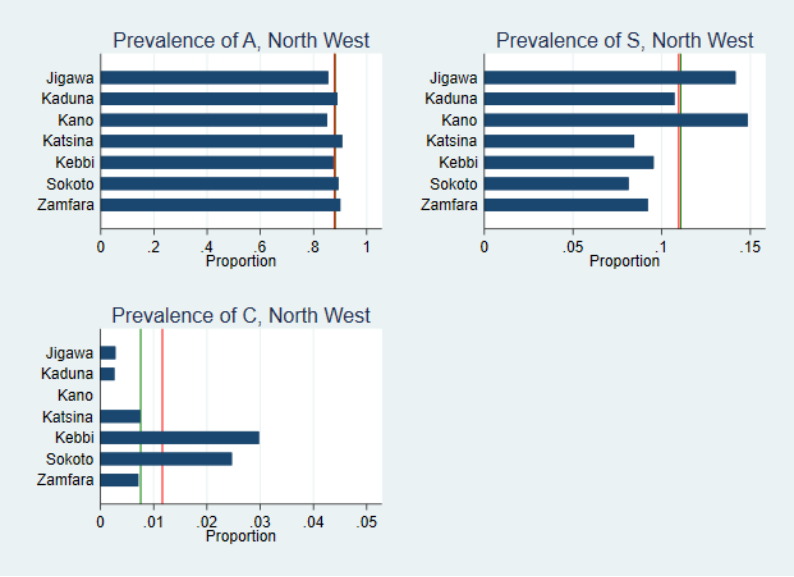


Figure 3(d) Prevalence of the A, S, and C alleles in children age 6-59 months, by state, in the South East Zone. Nigeria DHS 2018. Weighted.

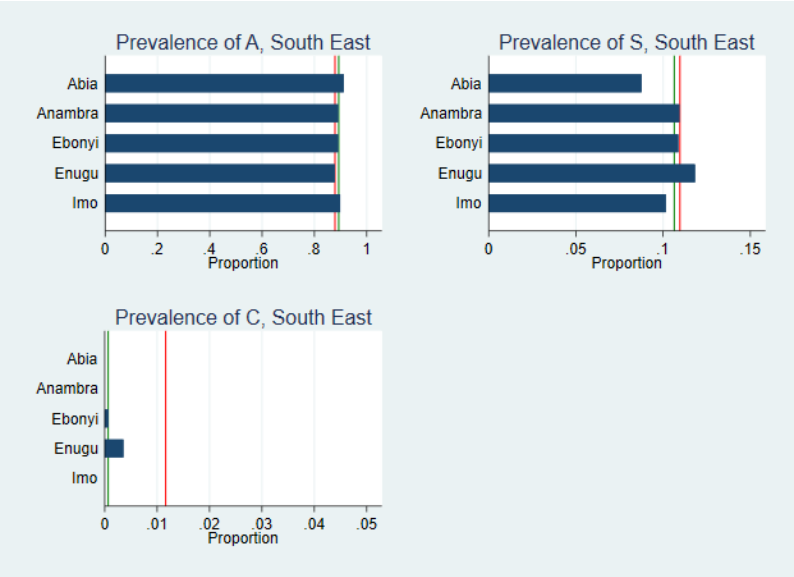


Figure 3(e) Prevalence of the A, S, and C alleles in children age 6-59 months, by state, in the South South Zone. Nigeria DHS 2018. Weighted.

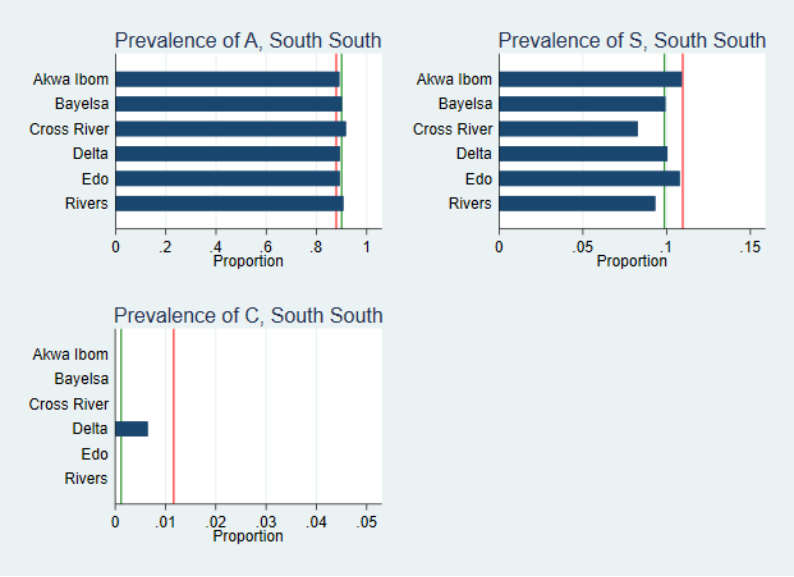
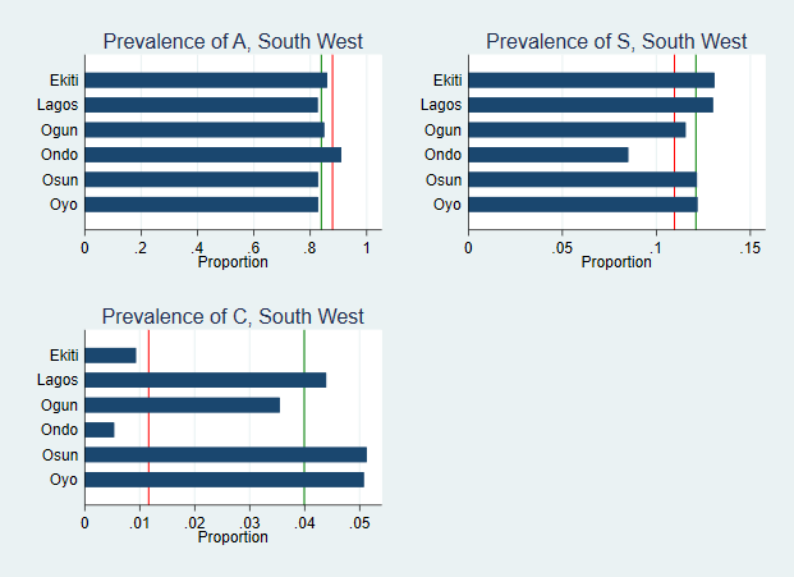


Figure 3(f) Prevalence of the A, S, and C alleles in children age 6-59 months, by state, in the South West Zone. Nigeria DHS 2018. Weighted.



The S allele is much more common than C but it too varies across zones and across states within zones. Like the C allele, the S allele is most prevalent in the South West Zone. Within the South West Zone, Ondo State has the lowest prevalence of both S and C. Figures 4(a) and 4(b) are scatterplots that show the correspondence between S and C. In these figures, each state is represented by a point. (In Figure 4(a), the points and labels for Oyo and Osun, two states in South West Zone, overlap and are virtually indistinguishable.) The vertical axis is the prevalence of C within the state and the horizontal axis is the prevalence of S. If all states were shown in a single scatterplot, the states with the lowest levels of C would be too crowded at the bottom. Therefore, Figure 4(a) shows the states in which the prevalence is estimated to be 0.0100 or greater, and the remaining states are shown with a magnified vertical scale for C between

0.00 and 0.01 in Figure 4(b). Figure 4(a) suggests a strong positive association between S and C in the eight states with the highest levels of C. There is little, if any, evidence of a correspondence between S and C within Figure 4(b).

Figure 4(a) Prevalence of the S and C alleles in children age 6-59 months, for the states in which the prevalence of C is 1.00% or more. Nigeria DHS 2018. Weighted.

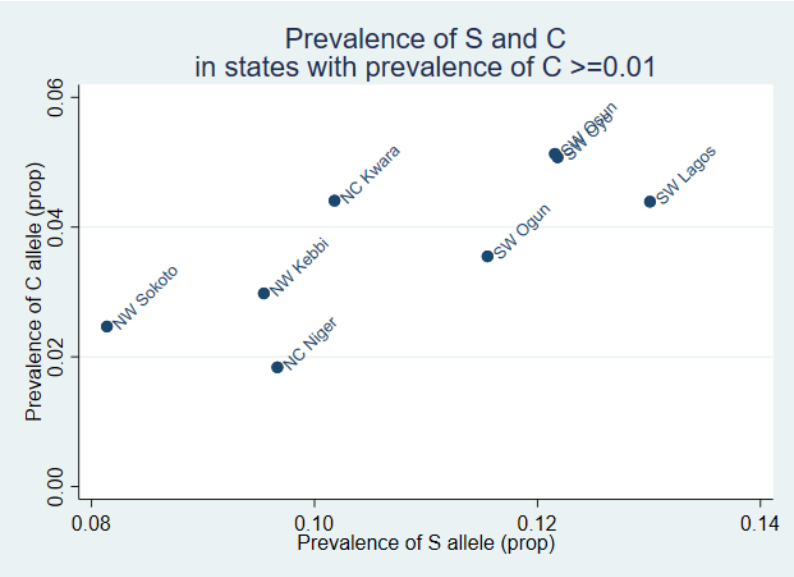
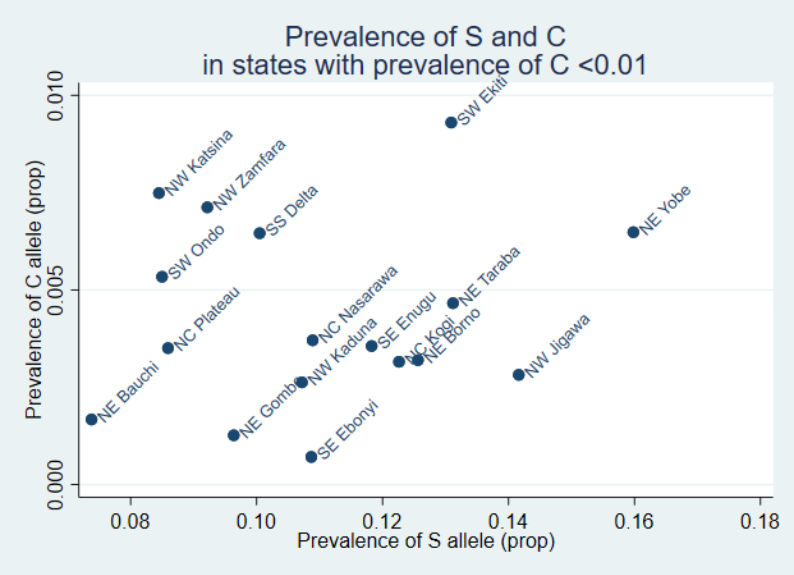


Figure 4(b) Prevalence of the S and C alleles in children age 6-59 months, for the states in which the prevalence of C is less than 1.00%. Nigeria DHS 2018. Weighted.

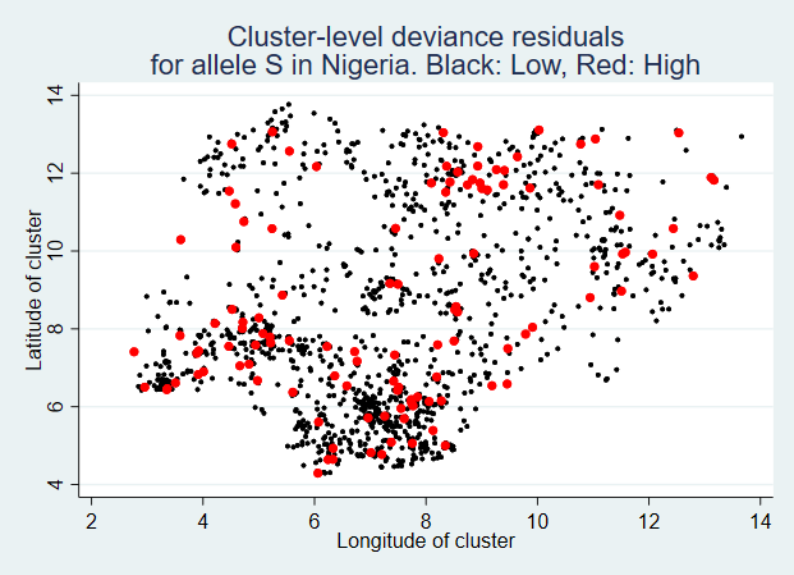


2.3 Genotype Variation across Clusters

In the two-stage sample design of the Nigeria DHS 2018 survey, as in nearly all DHS surveys, the primary sampling units were census enumeration areas (EAs). An EA usually is a village or a neighborhood in a town or city. The survey included 1,376 clusters, with an average of 37.2 clusters per state. The number of clusters per state ranged from 35 to 42, except for two states below this range and two above. Households were sampled within these clusters. The clusters had an average of 7.4 genotyped children age 6-59 months; the number of children per cluster varied greatly.

The small number of children per cluster limits the kinds of inferences that are possible with clusters as units. The prevalence of the genotypes or the alleles is not meaningful at the level of the cluster. At this level, we use a deviance residual, or standardized residual, from statistical models that compare each cluster with all of Nigeria, or with the zone or state in which it is located. The deviance residual takes into account both the cluster-level prevalence of the S or C allele and the number of children in the cluster (by being proportional to the square root of that number). Asymptotically, it has a unit normal distribution under a null hypothesis of homogeneity within all of Nigeria, or within each zone, or within each state. The calculation of deviance residuals is described in Appendix C.

Figure 5(a) Spatial distribution of prevalence of the S allele for children age 6-59 months. Dots are clusters. Red: deviance residual is ≥1. Black: deviance residual is <1. Nigeria DHS 2018. Unweighted.



Figures 5(a)-5(d) are geographic scatterplots that show the spatial distribution of the cluster-level deviance residuals for the S and C alleles. Figure 5(a) shows the national distribution of the S allele. The horizontal and vertical axes are the longitude (East) and latitude (North), respectively, of each cluster, as given in a separate GIS file.[5] The overall shape clearly corresponds with Nigeria. The dot for a cluster is red if the deviance residual is 1.00 or more—that is, at least one standard deviation above the overall prevalence for all of Nigeria. Otherwise, the dot is black. The red dots are larger than the black dots to improve visibility in the figures. There is no obvious spatial concentration of the red dots. Figure 5(b) is analogous to 5(a),

[5] The true coordinates of the clusters are slightly displaced during the data processing of each DHS survey, in order to prevent disclosure of the actual enumeration area and households within it.

but it represents the geographic distribution of the C allele. The figure shows a high concentration of C in South West, smaller concentrations in North West and North Central, and large areas where there are few or no red clusters. Figures 5(a) and 5(b) show a spatial representation of the prevalences that corresponds with the bar graphs for the prevalence of S and C by zones in Figure 3.

Figure 5(b) Spatial distribution of prevalence of the C allele for children age 6-59 months. Dots are clusters. Red: deviance residual is ≥1. Black: deviance residual is <1. Nigeria DHS 2018. Unweighted.

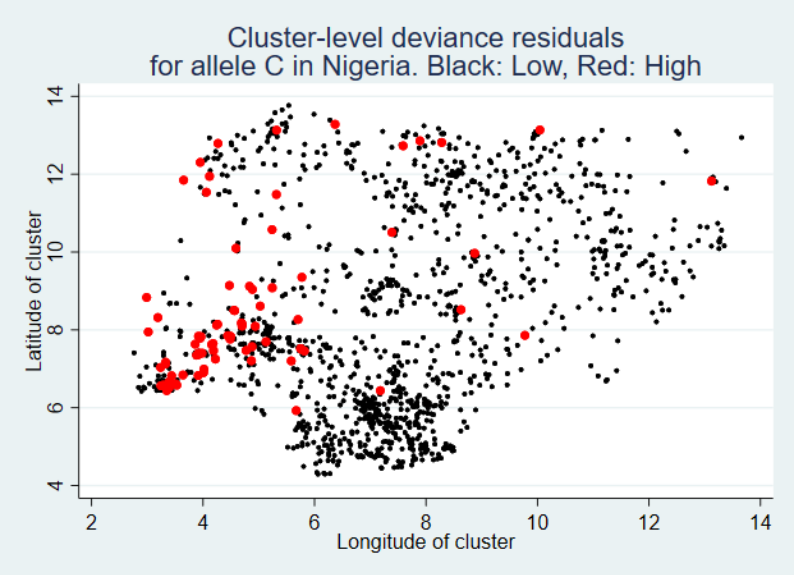


Figure 5(c) Spatial distribution of prevalence of the C allele for children age 6-59 months, within the South West Zone. Dots are clusters. Red: deviance residual is ≥1. Black: deviance residual is <1. Nigeria DHS 2018. Unweighted.

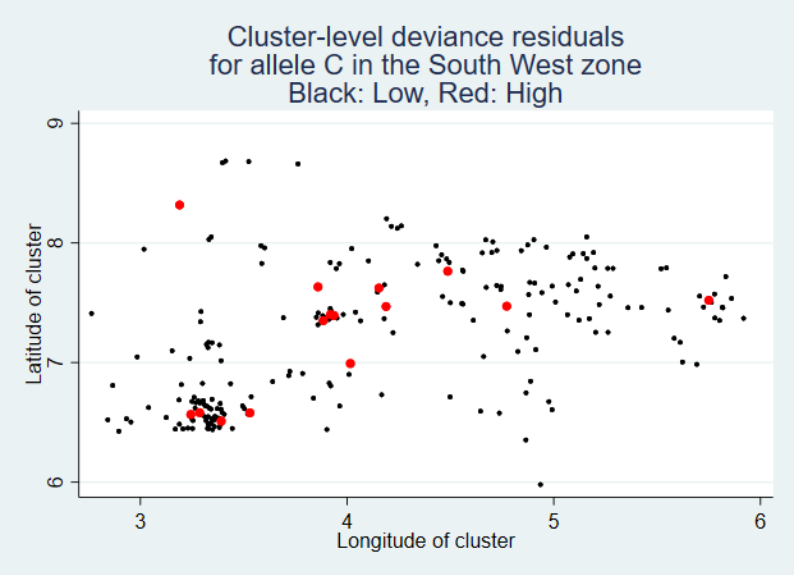
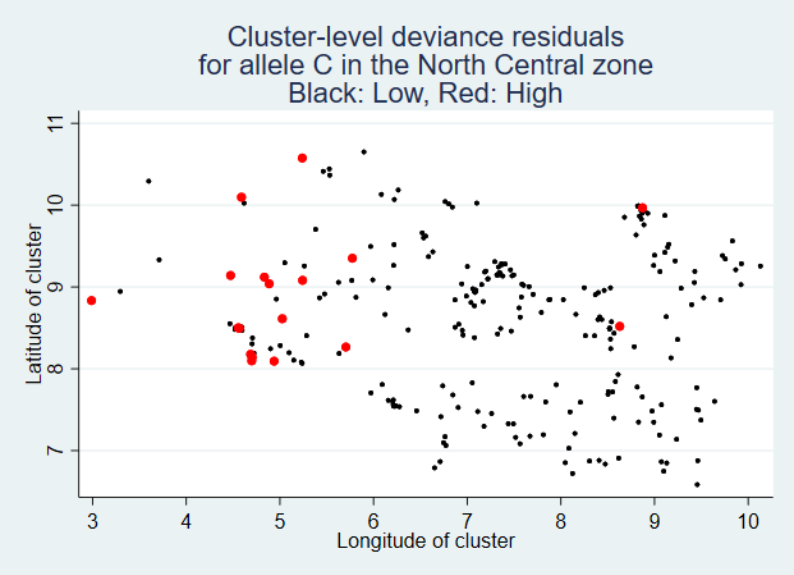


Figure 5(d) Spatial distribution of prevalence of the C allele for children age 6-59 months, within the North Central Zone. Dots are clusters. Red: deviance residual is ≥1. Black: deviance residual is <1. Nigeria DHS 2018. Unweighted.



The main irregularity in the distribution of S and C seen earlier is that the C allele is rare outside of the South West, North West, and North Central zones. Figures 5(c) and 5(d) show the spatial distribution of C within those two zones. In these scatterplots, the deviances describe the deviation from the national mean. The highest prevalence of C, by far, is in the South West Zone, which is shown in Figure 5(c), and is next highest in the North Central Zone, shown in Figure 5(d).

3 THE ASSOCIATION BETWEEN GENOTYPE AND CHILD HEALTH

3.1 Overview for All Anthropometry and Health Outcomes

The DHS surveys collect data on several indicators of child health. For the Nigeria 2018 survey, we can construct the following 11 binary (0/1) indicators for children. Each outcome, one at a time, will be regressed on the six-category covariate, genotype, with logit regression adjusted for survey design effects.

- **Nutrition/anthropometry:** stunted, underweight, overweight, wasted
- **Recent symptoms of illness:** diarrhea, fever, cough
- **Anemia:** any (Hb<11.0), severe (Hb<7.0)
- **Malaria:** microscopy test positive, RDT positive

The four nutrition indicators are based on measurements of height and weight, age in days at the time of the survey, and sex of the child, applied to the 2006 WHO Child Growth Standards. A child is stunted if the height-for-age (HAZ) score is -2 or less. A child is underweight if the weight-for-age (WAZ) score is -2 or less, overweight if the weight-for-height (WHZ) score is +2 or more, and wasted if the weight-for-height score is -2 or less. Nutritionists generally attach more significance to wasting than to underweight. We include all four indicators, although few Nigerian children in the age range 6-59 months are overweight.

In DHS surveys, there are questions on three childhood illnesses: diarrhea, fever, and symptoms of acute respiratory illness (ARI). The reference period for all three is the previous 2 weeks. For each illness, there are follow-up questions about advice or treatment sought for the child and whether the child received appropriate treatment. Although fever can be a symptom of malaria, and short, rapid breathing that is chest-related and/or difficult breathing that is chest-related is potentially a symptom of ARI, no inferences will be drawn here about any underlying illness.

Anemia is inferred from the hemoglobin concentration, Hb, adjusted for altitude. The lower the Hb, the more serious the anemia. The Hb concentration is measured in grams per deciliter (g/dL). The categories of anemia are severe (Hb<7.0), moderate (7.0-9.9), mild (10.0-10.9), and not anemic (11.0+). Three percent of children have severe anemia, 39% have moderate anemia, 27% have mild anemia, and 31% are not anemic. The analysis includes two definitions of anemia: “any” which includes severe, moderate, and mild, and “severe.” A subsequent section will assess the Hb distribution in more depth, going beyond the anemia categorization.

The final child health indicator is malaria, measured with two different tests. The first test is microscopy, which is based on laboratory analysis of a blood smear. The second test, labeled RDT, is a rapid diagnostic test completed during the household interview. The RDT has an important advantage in the context of a household survey because feedback can be given immediately to the caregiver. If the result is positive, the child can be referred to a facility for a more accurate diagnosis and treatment. Some DHS surveys include only one malaria test, which is always the RDT. It is helpful that the Nigeria 2018 survey included microscopy as well.

The (weighted) estimate of malaria prevalence with microscopy is 22%. The prevalence with an RDT is 36%. The two tests disagree for 20% of children. They are correlated but only with r=0.60.[6] It is possible that the higher prevalence with the RDT is partly because of a higher false positive rate for the RDT and/or a higher false negative rate for microscopy. Although theoretically more accurate, there is evidence in other settings (Berzosa et al. 2018) of a higher false negative rate with microscopy. Blood smears are prepared under difficult conditions and may spend days in the field before they can be shipped to a laboratory. Table 4 includes the results for both types of tests.

Children who live in geographic areas where malaria is endemic may have mitigated risk of malaria because of campaigns that promote insecticide-treated nets (ITNs). There is evidence that ITNs greatly reduce the risk of malaria (Pryce, Richardson, and Lengeler 2018). The use of ITNs is widespread in Nigeria and is believed to be responsible for the substantial declines in malaria in recent years. Questions about household use of mosquito nets were included in this survey, but will not be used in the analysis because it is impossible to know whether their use has prevented malaria for a specific child.

Table 4 Results of regressions of child health outcomes on genotype, as a six-category predictor. Children age 6-59 months in the Nigeria 2018 DHS. Weighted.

	Odds ratios for genotypes, relative to AA									
	Genotype AS		Genotype AC		Genotype SC		Genotype SS		Genotype CC	
Outcome	OR	Sig.	OR	Sig.	OR	Sig.	OR	Sig.	OR	Sig.
Stunted	1.07	ns	1.12	ns	3.07	*	1.63	*	0.33	ns
Underweight	1.07	ns	1.02	ns	3.28	ns	1.87	*	0.74	ns
Overweight	0.76	ns		ns		ns	0.97	ns		ns
Wasted	1.07	ns	1.02	ns	3.30	ns	1.85	*	0.48	ns
Diarrhea	1.07	ns	0.71	ns	0.60	ns	1.20	ns		ns
Fever	0.97	ns	0.91	ns	0.74	ns	1.09	ns		ns
ARI symptoms	0.78	ns	1.14	ns	0.77	ns	0.27	ns		ns
Any anemia	1.11	ns	1.24	ns	9.28	**	65.87	***	0.83	ns
Severe anemia	0.53	***	0.13	*	1.85	ns	18.87	***	8.80	*
Malaria microscopy positive	1.14	ns	1.49	ns	0.57	ns	0.89	ns	0.47	ns
Malaria RDT positive	0.96	ns	1.32	ns	0.96	ns	0.57	*	0.30	ns

Note: Statistical significance: * .05; ** .01; *** .001; ns = Not significant.

The rows of Table 4 show the results of separate logit regressions of each outcome on genotype, as a categorical covariate. The regressions include adjustments for sampling weights, clusters, and stratification. The odds of each health outcome in genotypes other than AA are compared with the odds in genotype AA. "OR" is the odds ratio and "Sig" is the level of significance, with a single asterisk for the .05 level, two for the .01 level, and three for the .001 level. We focus on significance at the .01 and .001 levels. For example, an odds ratio of 1.07 for AS and "stunted" means that in the data, the odds that a child with genotype AS is stunted are 1.07 times as large as (are 7% greater than) the odds that a child with genotype AA is stunted; "ns" indicates that the odds ratio is not significantly different from 1, and not even at the .05 level.

[6] A microscopy test checks for the presence or absence of malaria parasites in the blood. In contrast, RDTs detect specific antigens (proteins) produced by malaria parasites. The RDT positive prevalence is typically higher than microscopy positive prevalence in cases when the child has been given malaria medication. Medication will clear parasites from the blood, but antigens will remain in the blood for some time after the parasites have been cleared.

At the .01 or .001 levels, only two outcomes are significantly related to genotype as a categorical covariate: any anemia and severe anemia. There are no other undesirable health outcomes among those available in the survey for which genotype is a statistically significant risk factor with a p-value<.01 criterion. Specifically, in comparisons with the AA genotype,

- Children with the AS genotype are less likely to have severe anemia (OR=0.53, p<.001)
- Children with the SC genotype are more likely to have severe anemia (OR=9.28, p<.01)
- Children with the SS genotype are more likely to have any anemia (OR=65.87, p<.001)
- Children with the SS genotype are more likely to have severe anemia (OR=18.87, p<.001)

The very large estimates of odds ratios for the SC and SS genotypes and anemia have wide confidence intervals because of the small numbers of cases in the data. Some combinations in Table 4, especially involving the Other or CC genotype, cannot be estimated for the same reason.

3.2 Relationship among Genotype, Hemoglobin Concentration, Malaria, and Weight-for-height

There is particular interest in any evidence of a relationship between genotype and malaria. With the RDT, one of the five coefficients for specific odds ratios in Table 4 is significant at the .05 level, although not at a higher level, and none of the microscopy coefficients are significant at even the .05 level. We have also examined the association of the two malaria indicators with known correlates of malaria, such as place of residence and wealth quintile. The relationships are consistently stronger with the RDT than with the microscopy test. For this reason, only the RDT results, and not the microscopy results, will be used as the measure of malaria. This decision is based on the empirical strength of relationships, at least in this survey, although we recognize that it is a data-based decision, and in another survey we might have proceeded with the microscopy results. The stronger relationship with the RDT results is assumed to be due to the difference between what the two tests actually measure.

We now look in more detail at the association among four variables: genotype, which is inherently a categorical variable with six categories; hemoglobin concentration (Hb), which can be grouped into the four categories of anemia but will be used as an interval-level variable; WAZ, which is the basis of the categorization as wasted or not wasted, but can also be treated as an interval-level variable; and the result of the RDT for malaria.

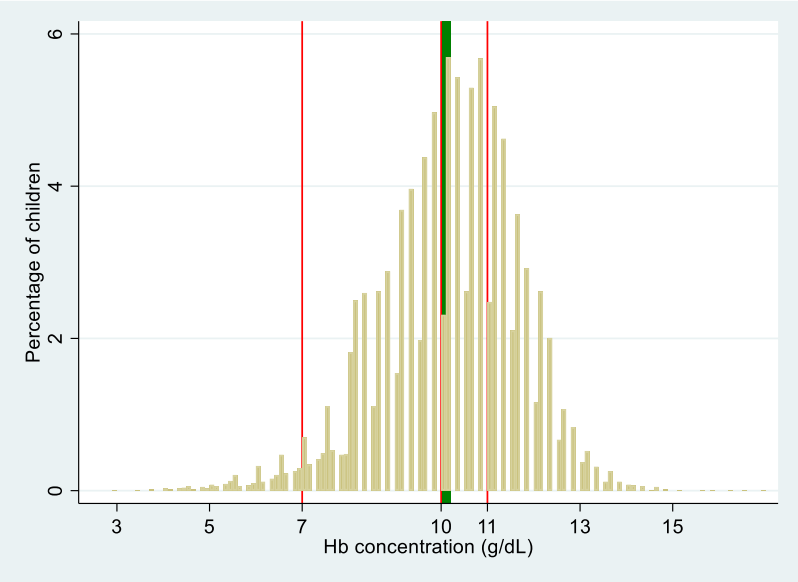
The limitation to surviving children is important. Anemia, malaria, and wasting are potential causes of death. The children who were most seriously affected by these conditions are among the children who died between birth and the date of the survey. Thus, inferences based on surviving children are incomplete. We continue this analysis without any assumptions about how anemia, malaria, and wasting are causes of mortality.

3.2.1 Genotype and hemoglobin concentration

As stated earlier, a majority of children age 6-59 months in the survey (68%) have some degree of anemia. Twenty-seven percent have mild anemia, 38% have moderate anemia, and 3% have severe anemia.

The overall mean of Hb is 10.15, and the median is 10.3. Both numbers are just above the lower limit for mild anemia, which is 10.0. The overall distribution is shown in Figure 6. In this figure, red vertical lines at 7, 10, and 11 identify the boundaries of the anemia categories, and a green vertical line identifies the mean.

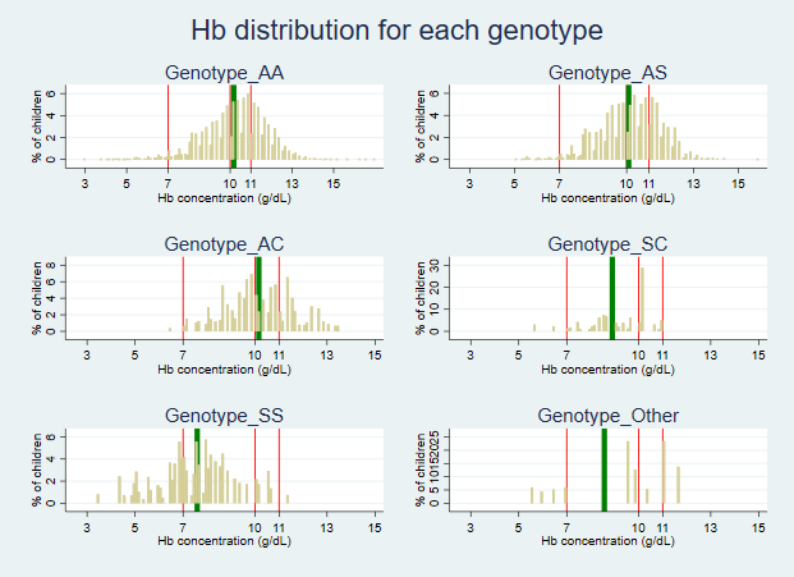
Figure 6 Distribution of hemoglobin concentration (Hb) among all children age 6-59 months in the Nigeria DHS 2018. Weighted.



There is a strong relationship between genotype and Hb, but primarily because children with the SC and SS genotypes have lower levels of Hb. The mean for AA is 10.19, which is only slightly above the overall mean of 10.15. The AA genotype accounts for 77.1% of all children. The mean for SC is lower by 0.97, approximately the full width of the mild anemia category, but SC only includes 0.4% of the children. This difference has a p-value of .005. The mean for SS is 2.69 less than the mean for AA, a difference that is significant at the .001 level, although SS only includes 0.9% of children in the sample. The multiple R-squared for this model is 0.027. That is, 2.7% of the variation in hemoglobin concentration is statistically explained by genotype. To show the relative importance of genotype as a determinant of Hb, we compare it with three other possible predictors. R-squared is 0.057 for a model that predicts Hb with state; it is 0.024 if the predictor is type of place of residence (urban/rural); 0.066 if the predictor is the wealth quintile; and 0.1068 for an additive model with all three predictors.

Figure 7 shows the distribution of Hb within the six genotypes. The vertical green lines show clearly that the means for genotypes SC and SS are displaced to the left.

Figure 7 **Distribution of hemoglobin concentration (Hb) among children age 6-59 months in each genotype in the Nigeria DHS 2018. Red vertical lines at 7, 10, and 11 identify the boundaries of the anemia categories; a green vertical line identifies the mean Hb within each genotype. Weighted.**



3.2.2 Genotype and malaria

As described above, the data include two indicators of malaria—the results of a microscopy test and the results of an RDT. The RDT results have a prevalence that is much higher than microscopy. However, because the RDT results consistently show stronger correspondence with covariates, we infer that they are less affected by measurement error and give priority to the RDT test results in this analysis.

We now add adjustments for the survey design—namely, sampling weights and adjustments for clustering and stratification. The use of sampling weights corrects for bias in estimates due to oversampling or undersampling of the strata (combinations of state and urban/rural residence). Without the adjustment for clustering, the standard errors tend to be too large, and without the adjustment for stratification, the standard errors tend to be too small. The latter two adjustments only affect the standard errors, and therefore, the tests and confidence intervals.

First, we examine the association among genotype, malaria, and Hb. There is convincing evidence that low Hb is not a risk factor for malaria, although it is associated with malaria (White 2018). Therefore, there are three potential pairwise relationships: the effect of genotype on malaria and genotype on Hb, and the association between malaria and Hb.

Anemia has many causes. Globally, only about half of anemia is due to iron deficiency. It can also be caused by vitamin deficiency (b12 or folate), sickle cell, malaria, thalassemia, lead poisoning, hypothyroidism, and many other conditions. A hemoglobin level of <8 g/dL is most often associated with malaria.

As was seen in Table 4, when the RDT results are regressed on genotype with logit regression, and with adjustments that take the design effect into account, the only evidence of a relationship is that the coefficient for the SS category has a p-value less than .05 (specifically, 0.024). When the microscopy results are regressed on genotype, no coefficient is significant at even the .05 level.

If Hb is regressed on genotype, there is a significant relationship with the SC and SS genotypes. When Hb is regressed on the RDT result, there is a strong, highly significant negative relationship. The mean Hb score for children with a negative RDT result is 10.60. Children with a positive result have a mean Hb of 9.36. The difference of 1.24 points is substantial, greater than the 1-point width of the mild anemia category. Among children with a negative result, 28.7% have severe or moderate anemia. Among children with a positive result, more than twice as many, 63.1%, are in those two categories.

When Hb is regressed on genotype and the RDT result together, all coefficients and test statistics are nearly identical with what they are in the separate one-predictor models. This would be expected because genotype and the RDT result have a very weak statistical relationship. Malaria is more important in this model than genotype.

It may be surprising that we have not found a relationship between genotype and malaria. Based on other studies, it could have been expected that children with the AS genotype, the "sickle cell trait," would show a reduced risk of malaria. To better understand the association, or lack of it, with these data, we will stratify the cases by the level of hemoglobin concentration.

We repeat the logit regression of the RDT diagnosis on genotype, as a categorical variable, in the following groups:

- Children in the 1st quartile of the Hb distribution, with Hb≤9.2
- Children in the 2nd quartile of the distribution, with 9.2<Hb≤10.3
- Children in the 3rd quartile of the distribution, with 10.3<Hb≤11.2
- Children in the 4th quartile of the distribution, with Hb>11.2

Table 5 Coefficients and p-values for genotypes AS and SS in the logit regression of malaria (RDT results) on genotype in the quartiles of the distribution of hemoglobin concentration (Hb). Weighted.

Quartile	Coefficient of AS	p	Coefficient of SS	p
1	-0.2648	0.017*	-1.7222	0.000***
2	-0.0782	0.503	-0.2698	0.742
3	-0.1017	0.408	na	na
4	0.3572	0.025*	na	na

Notes: * p<.05; **p<.01; *** p<.001

The results of the four logit regressions are presented in Table 5. There is a separate regression for each row. The table only shows the coefficients and p-values for AS and SS. The coefficients are the logs of the odds ratios for AS or SS relative to AA. In the third and fourth quartiles of the Hb distribution, there are no cases with genotype SS; all children with the SS genotype have an Hb value below the median. Thus, there are no coefficients for SS in those quartiles.

In the first quartile, which has the lowest Hb levels, the coefficients are negative and have p-values of .017 and .001, respectively. The negative coefficient for SS in the first quartile, -1.7222 (corresponding to an odds ratio of 0.18), stands out. Children who are in the 1owest quartile of anemia and have the SS genotype are much less likely to have a positive RDT result than children in the first quartile with the AA genotype.

In the second quartile, children with the AS and SS genotypes have negative coefficients; in the third quartile, children with the AS genotype have a negative coefficient. Negative coefficients imply a reduced risk of malaria, but these three coefficients do not approach significance.

For children in the fourth quartile of the Hb distribution, the bottom row of Table 5, there is a positive coefficient, 0.3572 (equivalent to an odds ratio of 1.43). Although not significant at the .01 level, the coefficient suggests that children with a high Hb level and the AS genotype are *more* likely than children with the AA genotype to have a positive RDT result. Another manifestation of this positive effect will be seen with the uptick in the far right of Figure 8.

Our interpretation of this pattern is that children with the AS or SS genotypes have a reduced risk of plasmodium infection if, but only if, they have relatively low Hb values. Otherwise, there is no evidence of a reduced risk of contracting malaria.

Table 6 shows the percentages of children who are wasted, who test positive for malaria with microscopy, or test positive for malaria with RDT, in each category of anemia. Anemia clearly has a very strong relationship with both wasting and malaria. Children who have severe anemia are much more likely to be wasted or to have malaria than children who are not anemic. Twenty-two percent of children are wasted, but among nonanemic children, the percentage is less at 15%. Thirty-six percent of children test positive for malaria (with the RDT), but among nonanemic children, only 18% test positive. Among children who are severely anemic, 49% are wasted and 80% have a positive RDT result. The correspondences are similar for microscopy results.

Table 6 Percentage of children age 6-59 months who are wasted or have a positive test result, with either microscopy or RDT, in each category of anemia. Nigeria DHS 2018. Weighted.

		Positive result	
Anemia	**Wasted**	**Microscopy**	**RDT**
Severe	49.18	62.54	80.36
Moderate	27.11	35.34	53.46
Mild	20.57	15.12	28.28
Not anemic	15.25	10.11	17.97
Total	22.21	22.65	36.15

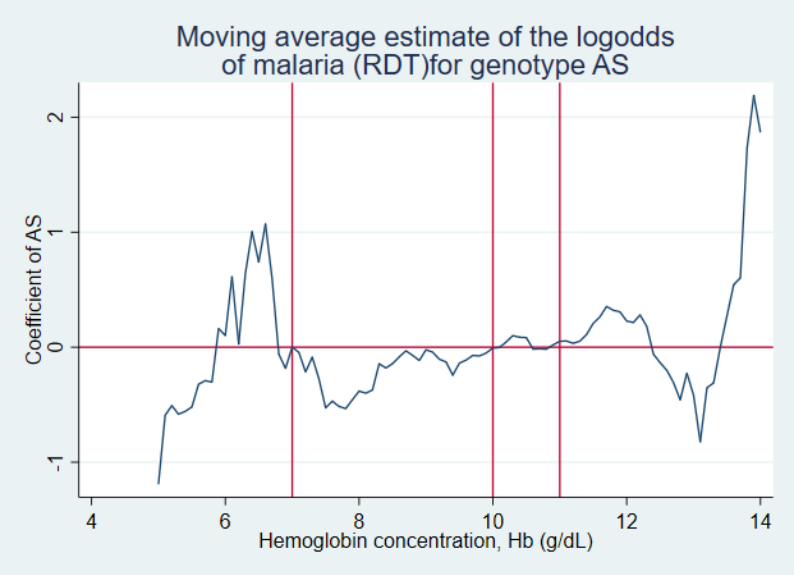
We assessed the Hb distribution in terms of the standard categories of anemia and quartiles. We now focus on Hb concentration as a continuous variable. Ideally, the percentage of children who have malaria for each specific value of Hb would be calculated, but this is impossible because of the small number of cases at each specific value. It is necessary to smooth the data in some way. Figure 8 is based on children on an 11-point moving interval of Hb levels. For example, the value given for Hb=10.0 (g/dL) is based on those children whose measured Hb is 9.5-10.5, inclusive.

For the children in that interval, the RDT result is regressed (with logit regression) on the six-category genotype, with AA as the reference category. The coefficient of AS in that regression, on a log odds scale, is given on the vertical axis of Figure 8. The exponential of this coefficient is not presented, but it would be the observed odds that a child in this range of Hb has a positive diagnosis of malaria for genotype AS compared with genotype AA. If the coefficient is negative, or below the horizontal red line in the figure, then the AS genotype is protective against malaria. If it is positive, then the genotype is a risk factor for

malaria. We emphasize that this analysis looks at association with the AS genotype and does not attribute causality.

The procedure just described for Hb=10.0 was conducted for each possible value of Hb in the full range of the data. The figure includes vertical red lines at 11 (the lower end of the normal range), 10 (the lower end of the mild anemia range), and 7 (the lower end of the moderate anemia range). Values to the left of Hb=7 describe severe anemia. The moving interval crosses these arbitrary boundaries, such as in the regression for Hb=10 that includes children in both the moderate and mild categories of anemia. The line is relatively smooth in the range from 7 to 12, which includes most children. Only about 3% are below 7 and about 9% are above 12. The line is erratic to the left of Hb=7, but has very low values of the log odds of a positive RDT result for the very lowest values of Hb. The line is also erratic to the right of Hb=12, but has very high values of the log odds of a positive RDT result for the very highest values of Hb.

Figure 8 **Log odds that a child has malaria (with the RDT test), for the AS genotype versus the AA genotype, as a function of hemoglobin concentration (Hb, g/dL). Calculated for children whose Hb is the index level ±0.5 g/dL.**



The log odds in Figure 8 are consistently negative in the range for moderate anemia, which in the Nigeria survey includes about 40% of children age 6-59 months. For these children, the AS genotype appears to be somewhat protective against malaria.

The association among genotype, malaria, and Hb is complex, and may be affected by characteristics that are not available in the data. The association is certainly affected by higher mortality prior to the survey that is the result of the SC and SS genotypes, malaria, and severe anemia. As stated earlier, the sample is selected for survival. We must be very cautious and avoid overinterpreting the data.

A major strength of the Nigeria DHS survey is that it is representative of the entire household population. It is possible that other data used to investigate the relationships among these variables have not been representative and have been biased toward children who are less healthy, such as children who have moderate or even severe anemia. Over-representation of sick children in a study population would tend to indicate that the AS genotype is protective against malaria. Without that compositional bias, there is no

evidence in these data of a protective effect. Further analysis, perhaps with more explicit consideration of the C allele, may help to clarify the relationships. However, given that the C allele is relatively rare, it is unlikely to be important at the population level.

We have also investigated the association that wasting, as a binary outcome, and the WHZ, as an interval-level outcome, have with Hb, malaria, and genotype. We find that children who have a lower hemoglobin concentration and/or a positive RDT result are significantly more likely to have a lower WHZ score. In a regression of WHZ on Hb and the RDT result, the coefficients for both Hb and RDT are significant at the .001 level. However, genotype is not associated with the WHZ, either in a bivariate relationship or in models that include Hb and/or malaria. Genotype is not associated with the likelihood of being overweight or wasted, the two extremes of the WHZ distribution. Genotype only affects Hb and the anemia categories that are constructed from Hb.

4 EVIDENCE OF A RELATIONSHIP BETWEEN GENOTYPE AND MORTALITY

4.1 Survival Model

It is known *a priori* that children with genotypes SC and SS have higher mortality than children with genotypes AA, AS, and AC. Children with genotype CC, to the extent that they are in the "other" category, are known to have mild disease and, as a group, do not have higher mortality. For simplification, we can dichotomize the genotypes into a variable that is coded Y=1 for children with genotypes SC or SS and Y=0 for children with genotypes AA, AS, or AC. The children are then classified into six-month intervals of age (time since birth), 6-11 months through 54-59 months. These age intervals do not describe a cohort followed over time. However, because of higher mortality if Y=1, we might expect the data to show a monotonically declining proportion of children with Y=1 in the successive age groups. Is it possible to infer actual differences in survivorship for the two groups of genotypes?

The short answer is no. For readers who may doubt this, Appendix 1 offers an analogy with another binary variable, defined as Y=1 if the child is from a multiple birth (twin) and Y=0 if a single birth. We know whether each child recorded in the Nigeria 2018 survey's birth histories is a twin or a singleton, for decedents as well as survivors. About 3% of children are twins. The appendix shows that the relative risk of death is about three times as great for a twin as for a singleton. However, if we remove the information about status at birth and only have access to the status of survivors, we are unable to estimate the relative risk even with a plausible statistical and demographic model. Any estimate of relative risk is very sensitive to the percentage of children who are twins *at birth*.

The percentage of children with a problematic genotype is less than half of the percentage who are twins. Since genotyping was done only on a subsample of children in the household survey, the genotyping data are even more statistically unstable than the data on twins. This strategy for estimating mortality can be eliminated.

There is strong evidence from longitudinal studies of cohorts of children genotyped soon after birth that children with the AS genotype have *lower* mortality than children with the AA genotype. Aidoo et al. (2002) (see a summary at https://www.cdc.gov/malaria/about/biology/#tabs-1-4) identified a protective effect in a cohort study in Kenya. A protective effect against death from malaria was especially evident at age 2-16 months. The DHS data are unable to identify the protective effect of AS for the same reason that we cannot identify the increased risk associated with SC and SS. We do not know the genotype of the children in the birth histories who have died.

4.2 Consistency of the Genotypes

With the estimated prevalences of the A, S, and C alleles calculated in Chapter 2, we can calculate the fitted or expected frequencies of the six combinations in Table 1. For each combination, we can also calculate the contribution to the Pearson chi-square test statistic, which assesses the overall fit. The fit is excellent for genotypes AA and AS (the expected frequencies, weighted, are 8804.3 and 2196.4), and the differences between the observed and expected frequencies are not significant. The observed frequencies for the last

four combinations are all statistically significantly different from the expected frequencies (207.7, 25.9, 137.0, and just 1.2). The overall Pearson chi-square is 178.7, with three degrees of freedom, which is highly significant ($p<.001$), and which indicates that the simple combinatorial model for genotypes does not fit the data. The lack of fit is primarily due to the relative frequencies of the SC, SS, and CC combinations.[7]

As a minor modification, we have also estimated Pr(A), Pr(S), and Pr(C) using just the first three genotypes (AA, AS, and AC). These estimates of the three probabilities give an exact fit to the observed numbers of cases in the first three genotypes. The fitted frequencies of SC, SS, and CC with these probabilities (and a coefficient for a modified sample size) are almost exactly the same as in the previous paragraph and the interpretation is the same. The observed frequencies of SC, SS, and CC are not consistent with the observed frequencies of AA, AS, and AC. It is likely that the deviations result from different mortality risks for the different genotypes between birth and the date of genotyping.[8] However, we are unable to identify a systematic pattern to the deviations that could be interpreted as a difference in mortality risk. It is possible that the combinatorial model would fit better if it were estimated within disaggregated subpopulations, and the subpopulations were then aggregated. As seen in Chapter 2, Nigeria is far from homogeneous in the distribution of genotypes.

It is possible that this approach could be improved by borrowing information about the genotype distribution in other populations. A serious limitation is that the data include a total of only 145 children (unweighted) in the last three genotypes—34, 102, and 9 for SC, SS, and CC, respectively—and only 2 or 3 degrees of freedom.[9]

4.3 Indirect Comparison with Parents

The genotypes of the children, as noted above, are derived from the parents' genotypes. If we know the genotypes of the parents, we can then calculate the probabilities of the genotypes of the children. If the mother's genotype is ab, and the father's is cd, then the combinations ac, ad, bc, and bd are equally likely and have probabilities ¼ = 0.25. Thus, for example, if one parent is AA and the other is AS, then the four equally likely combinations are AA, AS, AA, and AS, so that Pr(AA)=0.5 and Pr(AS)=0.5.

If there are six possible combinations for the mother and six for the father, there are 6*6*4=144 possible outcomes for a child, although they collapse to the same six combinations. The probabilities of A, S, and C and the different combinations are plausibly similar for mothers and fathers, but they can be quite different for parents and children. For example, it would be possible for AA, AS, and AC to be the only observed genotypes in the parental generation, while some children have combinations SC, SS, and CC. Such a pattern would be consistent with a pattern of mortality risk in which children with those genotypes do not survive to adulthood.

We have developed a procedure to simulate the inheritance of the genotypes. It is difficult to be conclusive because we have no data about the actual genotypes of the parents. However, the procedure implies that the

[7] With more decimal places, Pr(A)=.8803, Pr(S)=.1116, and Pr(C)=.0082.

[8] Some of the discrepancy between observed and expected frequencies may be due to the uncertain composition of the "other" category. However, the results are similar for a model that excludes "other".

[9] There are three available df if the sum of the fitted frequencies is not required to be the sum of the observed frequencies. That requirement must be dropped in a model that attempts to estimate different levels of mortality.

SC, SS, and CC genotypes are less prevalent in the parental generation than they are in the observed cohort of children. Higher mortality risk for those genotypes is the likely explanation. This approach could benefit from information borrowed from other populations and genetic modeling.

4.4 Pairs of Genotyped Siblings

The data include 2,841 (unweighted) pairs of children age 6-59 months who were siblings and were genotyped. These are children who were genotyped as part of the household survey, were identified in the birth histories of the interviewed women, and have the same mother. Table 7 shows the frequencies of the different possible pairs. The rows refer to Child 1, who is most recent child in the pair listed in the birth history, and the columns to Child 2, who is the next-most-recent child listed in the pair.[10] In a few cases, three siblings were genotyped. Those three siblings would contribute three pairs to the table.

The majority of pairs are AA+AA (1914 pairs); AA+AS pairs are next most common (266+275=541 pairs); and then AS+AS (277 pairs). All other pairings are much less common. There are only six pairs of genotyped siblings in which both children are SC or SS or CC.

Table 7 Among pairs of siblings age 6-59 months who were genotyped, the observed frequency of each combination. Nigeria DHS 2018. Unweighted.

Genotype of Child 1	Genotype of Child 2						
	AA	AS	AC	SC	SS	Other	Total
Normal (AA)	1,914	266	18	2	5	1	2,206
Sickle cell trait (AS)	275	277	7	2	8	0	569
Hb C trait (AC)	10	4	16	2	0	0	32
Hb C disease (SC)	3	2	2	3	0	0	10
Sickle cell anemia (SS)	8	13	0	0	2	0	23
Other	0	0	0	0	0	1	1
Total	2,210	562	43	9	15	2	2,841

It was initially expected that there would be a high correspondence between the genotypes of siblings, so that if one child has a problematic genotype, the other children with the same parents (full siblings or even half-siblings) would have a high probability of a problematic genotype. With these data, it is possible to estimate the conditional probabilities of all siblings' genotypes in the entire birth history, and then, using the survivorship of siblings in the birth history, to estimate the genotype-specific probabilities of a child death.

Unfortunately, the number of siblings who share high-risk genotypes is much too small to permit the application of such a procedure. There are insufficient pairs of genotyped siblings in which both siblings have sickle cell disease. It is possible that genetic modeling or borrowing of information from other data sets with genotyped siblings would be helpful.

4.5 Pairs of Genotyped and Non-genotyped Siblings

The second strategy with siblings is based on the survivorship of the non-genotyped siblings of the genotyped children. We assume that the siblings have an elevated probability of having the same genotype

[10] In DHS birth histories, children are indexed in the reverse order of birth, with "1" for the most recent child. Within multiple births, the sequence is what the mother reported and may not be chronological.

as the tested children. We expect an indirect, attenuated correspondence with the mortality risk of the different genotypes. We identified all children in the household survey who were eligible for genotyping, who had any genotype code, including "not tested," and whose mothers were in the same household and were eligible for the survey of women. Children whose mother was not living or did not reside in the same household at the time of the survey are not included because those children cannot appear in a birth history. We refer to the children in the household survey who had a genotype code as "index children" and to other children in the same birth history as "siblings of index children." An important distinction here is that all the index children were known to be alive at the time of the survey, and their siblings could have been either living or dead.

Among all respondents to the women's survey, the birth histories included a (weighted) total of 126,538.2 births. Of this number, 18,471.9 children had died. The crude death rate is 146.0 deaths per 1,000 births. These children could have died at any age. By international standards, this is a very high rate. The non-genotyped siblings consist largely[11] of 95,774.7 children who were born outside the age range of 6-59 months before the interview. Of that number, 15,320 had died, giving a crude death rate of 160.0 deaths per 1,000 births. We calculate this rate as a check because the overall death rate of siblings of genotyped children should be approximately this amount.

We then merged each genotyped child with the mother's number of births ("births") and deaths ("deaths"). For each index child, we calculate the number of siblings ("sibling births") and the number of those siblings who died ("sibling deaths"). The number of births in the birth history is reduced by one, the child in the sibship who was genotyped, because that child is known not to have died. That is, "sibling births = births – 1" and "sibling deaths = deaths." Index children who had no siblings are dropped.

Using index children as cases, we then estimate the sibling death rate overall and for each genotype. This uses a generalized linear model with binomial error and logit link, and with "sibling deaths" as the numerator and "sibling births" as the denominator for each child.

The design effect (weights, clustering, and stratification) is taken into account with a modification of the weight. As noted in the previous section, there were many cases of siblings being genotyped. If S siblings from the same sibship were eligible for genotyping and were genotyped, we do not alter the "sibling births," but we divide the sampling weight by S. The treatment of duplicates should only have a weak influence on the results, depending on the relationship—if any—between the fertility of the mothers and the genotypes of their children. Otherwise, duplication is effectively random with respect to genotype.

With this procedure, the crude death rate for all siblings is 157.7 deaths per 1,000 births, which is very close to the crude death rate for all children (160.0 deaths per 1,000 births). The rates for the different genotypes of the index children are not statistically different from the crude rate, with two exceptions.

The death rate for siblings of SC index children is estimated to be 355.0, which is significantly different with p=.0024. The death rate for siblings of SS index children is estimated to be 293.4, which is significantly

[11] The non-genotyped siblings could include children who were born 6-59 months before the survey but died before the survey and could not be genotyped. The matching of children with mothers is also affected by residency status. If a child age 0-17 and mother are in the same household, the record for the child includes the line number of the mother, which is crucial for merging with the mother's data.

different with p<.0001. The risks of dying for the siblings of SC and SC children, relative to siblings of AS children, are 3.01 and 2.27, respectively.

In another logit regression, we pool the SC and SS genotypes, the two types of sickle cell disease, into a category SCD=1, and compare with the children in the AA, AS, and AC categories, as SCD=0. The relative risk for SCD=1 is 2.49 times the relative risk for SCD=0. The difference is significant with a p-value less than .0001.

The "other" genotype also has high relative risk of death (3.18). However, because it includes so few index children and few siblings, the p-value is 0.0740 and not significant. As noted above, in clinical studies, genotype CC does not have high mortality, but we do not know the correspondence between "other" and CC, and this section concerns siblings of children with genotype "other", not the "other" children themselves.

The siblings of index children with genotypes AS and AC, although numerous, do not provide any evidence of a difference from siblings of children with genotype AA. There is no evidence of reduced mortality for siblings of AS index children.

Approximately half of the deaths to siblings of children with genotypes SC, SS, and "other" can be attributed to the index child's genotype. Fortunately, however, only about 1% of children are in such sibships.

5 CONCLUSIONS

This report explored the uses and implications of sickle cell genotyping of children age 6-59 months in the Nigeria 2018 DHS survey. It is important to determine whether this biomarker should be included in future surveys in Nigeria or in other countries with a relatively high prevalence of the S and C alleles.

An important conclusion of this analysis is that these data cannot be used to estimate either the increased or decreased risk of death known to be associated with the sickle cell genotype. Clinical and longitudinal studies have conclusively established that young children who are SC or SS have an increased risk of dying, and children who are AS have a degree of protection from severe malaria and death from malaria. In this report, we have shown that cross-sectional information about the genotype of surviving children is insufficient for making inferences about the association between genotype and survival. This methodological limitation could have been anticipated with the analysis shown in Appendix A. We suggested some indirect strategies to identify mortality differences, and found that the siblings of genotyped children who are SC or SS have higher mortality than the siblings of children who are AA, AS, or AC. This strategy could be pursued further, although only about 1% of the genotyped children were SC or SS, and any mortality estimates that could ultimately be generated with that strategy would have wide confidence intervals.

An exploration of associations between genotype and child health identified only one strong relationship, between genotype and anemia, and especially severe anemia. There was no evidence of a relationship with the result of the microscopy test for malaria. There was a statistical relationship with some genotypes and the RDT test, although it was weak ($p<.05$). We attempted to clarify the relationship among genotype, hemoglobin concentration, and a positive malaria RDT result by using moving averages and logit regression. We found that the AS gene (compared with AA) tended to increase the odds of having a positive RDT result if the Hb concentration was higher than about 10 g/dL, and tended to decrease the odds if the Hb concentration was lower. However, this pattern (shown in Figure 8) is unclear at the very lowest and very highest concentrations. We emphasize that a positive RDT result is not equivalent to a diagnosis of malaria, and among children who do have malaria, the severity of the infection can vary widely.

The inability of these data to demonstrate clear relationships between genotype and survival, or between genotype and illness, simply means that these data are not a substitute for clinical studies. It is difficult to establish a correspondence with clinical findings.

The most useful information from the data may relate to the spatial distribution of the genotypes and alleles. Both the S and C alleles are concentrated primarily in the South West Zone, and secondarily in the North Central Zone. The C allele has relatively high prevalence, with a near-linear positive correspondence with S, in eight states, as shown in Figure 4(a). Four of the five highest combinations of S and C are in states located in the South West Zone.

Knowing the locations of the children with highest prevalence of S and C is helpful for estimating the burden of risk and for prioritizing interventions and education related to the problematic genotypes. For this purpose, genotyping in future surveys and expansion to older children and/or adults would be desirable.

REFERENCES

Aidoo, M., D. J. Terlouw, M. S. Kolczak, P. D. McElroy, F. O. Ter Kuile, S. Kariuki, B. L. Nahlen, A. A. La, and V. Udhayakumar. 2002. "Protective Effects of the Sickle Cell Gene Against Malaria Morbidity and Mortality." *Lancet* 359 (9314): 1311-1312. https://doi.org/10.1016/S0140-6736(02)08273-9.

Berzosa, P., A. de Lucio, M. Romay-Barja, Z. Herrador, V. González, L. García, A. Fernández-Martínez, M. Santana-Morales, P. Ncogo, B. Valladares, et al. 2018. "Comparison of Three Diagnostic Methods (Microscopy, RDT, and PCR) for the Detection of Malaria Parasites in Representative Samples from Equatorial Guinea." *Malaria Journal* 17 (1): 333. https://doi.org/10.1186/s12936-018-2481-4.

Federal Ministry of Health [Nigeria]. 2015a. *National Strategic Plan of Action on Prevention and Control of Non-Communicable Diseases*. Abuja, Nigeria: Federal Ministry of Health. https://extranet.who.int/ncdccs/data/nga_b3_ncd policy and strategic plan of action.pdf.

Luzzatto, L. 2012. "Sickle Cell Anaemia and Malaria." *Mediterranean Journal of Hematology and Infectious Diseases*. 4 (1): e2012065. https://doi.org/10.4084/mjhid.2012.065.

National Population Commission (NPC) [Nigeria] and ICF. 2019. *Nigeria Demographic and Health Survey 2018*. Abuja, Nigeria, and Rockville, MD, USA: NPC and ICF. https://dhsprogram.com/publications/publication-fr359-dhs-final-reports.cfm.

Piel, F. B., A. P. Patil, R. E. Howes, O. A. Nyangiri, P. W. Gething, T. N. Williams, D. J. Weatherall, and S. I. Hay. 2010. "Global Distribution of the Sickle Cell Gene and Geographical Confirmation of the Malaria Hypothesis." *Nature Communications* 2010. 1: 104-110. https://doi.org/10.1038/ncomms1104.

Pryce, J., M. Richardson, and C. Lengeler. 2018. "Insecticide-treated Nets for Preventing Malaria." *Cochrane Database of Systematic Reviews*. https://doi.org/10.1002/14651858.CD000363.pub3.

Pullum, T. W. 2008. *An Assessment of the Quality of Data on Health and Nutrition in the DHS Surveys, 1993-2003*. DHS Methodological Reports No. 6. Calverton, MD, USA: Macro International Inc. https://dhsprogram.com/pubs/pdf/MR6/MR6.pdf.

Pullum, T., D. K. Collison, S. Namaste, and D. Garrett. 2017. *Hemoglobin Data in DHS Surveys: Intrinsic Variation and Measurement Error*. DHS Methodological Reports No. 18. Rockville, MD, USA: ICF. https://dhsprogram.com/pubs/pdf/MR18/MR18.pdf.

White, N. J. 2018. "Anaemia and Malaria." *Malaria Journal* 17, 371. https://doi.org/10.1186/s12936-018-2509-9.

World Health Organization (WHO) Regional Office for Africa. 2013. *Sickle Cell Disease Prevention and Control*. Geneva, Switzerland: WHO.

APPENDIX A ANALOGY WITH THE MORTALITY OF TWINS

This appendix explores an analogy with multiple births as a risk factor. The data come from the KR file for the Nigeria DHS 2018 survey, which has as units all children born within the 60 months (labelled 0-59, inclusive) before the date of the mother's interview. This file includes all births, regardless of whether the child survived to the date of interview. For each child, the information includes the date of birth; a code for a multiple birth[12] (); sex of the child; whether still alive; and, if dead, the age at death. The file includes 33,924 births. Among these births, 32,661 were singletons, 1,236 were twins, and 27 were triplets. Although there were 9 sets of triplets, we refer to the children of multiple births as twins. Overall, the death rate was 95 deaths per 1,000 children, which separates into 260 per 1,000 twins and 88 per 1,000 singletons.[13] Twins have much higher mortality than singletons.

Appendix Table A1 Births, deaths, death rates, and relative risk for twins, compared with singletons, Nigeria DHS 2018

	All births			Twin births			Single births			
Time in months	Births	Deaths	Death rate	Births	Deaths	Death rate	Births	Deaths	Death rate	Relative risk
0-5	3,409	159	46.6	124	16	129.0	3,285	143	43.5	2.96
6-11	3,350	201	60.0	143	31	216.8	3,207	170	53.0	4.09
12-17	3,675	263	71.6	111	31	279.3	3,564	232	65.1	4.29
18-23	2,868	221	77.1	85	19	223.5	2,783	202	72.6	3.08
24-29	3,617	372	102.8	124	37	298.4	3,493	335	95.9	3.11
30-35	2,923	334	114.3	116	29	250.0	2,807	305	108.7	2.30
36-41	3,911	436	111.5	138	44	318.8	3,773	392	103.9	3.07
42-47	3,100	407	131.3	117	42	359.0	2,983	365	122.4	2.93
48-53	3,840	465	121.1	169	54	319.5	3,671	411	112.0	2.85
54-59	3,231	353	109.3	136	25	183.8	3,095	328	106.0	1.73
Total	33,924	3,211	94.7	1,263	328	259.7	32,661	2,883	88.3	2.94

Table A1 provides more detailed information on the survivorship of singletons and twins born in the 5 years before the survey, derived from the birth histories and aggregated by 6-month time intervals since the time of the birth. The first three columns show the number of births, number of deaths, and the death rate (1,000*deaths/births) for all children. For children who are alive, time describes their current age; for children who have died, time is the elapsed time since birth and not the age at death. For example, the time interval 6-11 refers to children who were born 6-11 months before the survey. A total of 3,350 children were born in that interval. Among those children, 201 had died, and the death rate was 60.0 per 1,000 births. In Table A1, each time interval identifies a 6-month birth cohort, which is followed to the date of interview.

The next group of three columns applies to children who were part of a multiple birth, while the following group of three columns refers to children who were single births. The final column is the ratio of the death rate for twins to the death rate for singletons, interpreted as the relative risk of dying for a twin, relative to a singleton.[14] The relative risk varies greatly, from a minimum of 1.73 to a maximum of 4.29. The overall

[12] The codes are 0 if a singleton, 1 if the first birth in a multiple birth, 2 if the second birth in a multiple birth, etc.; 1, 2, etc. refer to the sequence on a list, not necessarily the sequence during delivery.

[13] These cumulative death rates must not be confused with standard mortality rates, which involve more elaborate calculations to adjust for exposure to risk.

[14] This is a crude measure of relative risk for a simplified presentation.

pooled value is 2.94, which indicates that a twin is approximately three times as likely as a singleton to die before age 5.

We then remove all information about whether a child was a twin or a singleton at birth. The resulting data for the 33,924 births is shown in Table A2. Columns 1, 2, and 4 in this table match with columns 1, 2, and 3 in Table A1. The third column shows the total number of survivors. For example, of the 3,350 children born 6-11 months before the survey, the number alive at the time of the survey was 3,350-201=3,149. Those surviving children can be subdivided into 112 twins and 3,037 singletons, as given in columns 5 and 6. The overall percentage of survivors born 6-11 months before the survey who are twins is 100*112/3,350=3.56%.

Appendix Table A2 Data in Table A1, reduced to information about surviving twins and singletons, Nigeria DHS 2018

	All children				Survivors		
Time in months	Births	Deaths	Survivors	Death rate	Twins	Single-tons	Percent twins
0-5	3,409	159	3,250	46.6	108	3,142	3.32
6-11	3,350	201	3,149	60.0	112	3,037	3.56
12-17	3,675	263	3,412	71.6	80	3,332	2.34
18-23	2,868	221	2,647	77.1	66	2,581	2.49
24-29	3,617	372	3,245	102.8	87	3,158	2.68
30-35	2,923	334	2,589	114.3	87	2,502	3.36
36-41	3,911	436	3,475	111.5	94	3,381	2.71
42-47	3,100	407	2,693	131.3	75	2,618	2.78
48-53	3,840	465	3,375	121.1	115	3,260	3.41
54-59	3,231	353	2,878	109.3	111	2,767	3.86
Total	33,924	3,211	30,713	94.7	935	29,778	3.04

The information about the survivorship of twins and singletons in Table A2 is essentially equivalent to the information about genotyped children presented in Table A3. The question is whether Table A2 contains information about the relative risk of deaths for twins compared with singletons.

Define t to be the cohort (with t=1 for births 0-5 months before the survey, etc.), B_t to be the number of births in that cohort, S_t to be the number of survivors at the time of the survey, P to be the proportion of births that are twins, q to be the conditional probability of dying between time t and time t+1, and r to be the relative risk for twins. Assume that q and rare constants.[15] A visual inspection of Table A1 suggests that the greater risk of death for a twin is approximately constant during the first 5 years, not just in the neonatal and infant interval.

A plausible model would then be the following:

- For twins: $S_t=B_t(P)\exp(1-rq)^t$
- For singletons: $S_t=B_t(1-P)\exp(1-q)^t$

This model uses the numbers in columns 1, 5, and 6 of Table A2 to estimate three parameters: P, q, and r. Our interest is primarily in the estimate of r, which we know from the last column of Table A1 should be approximately 2.94.

[15] The estimation procedure could exclude the interval 0-5 months. Mortality is much greater in that interval. Also, we are constructing an analogy with the genotyping data, and children age 0-5 months were not genotyped.

Efforts to apply this model to Table A2 using the framework of generalized linear models lead to the conclusion that it is impossible to obtain estimates that are plausible and statistically stable with these data. The percentage of children who are twins is small and statistically unstable from one interval to the next. The highest percentage of twins, 3.86%, is found in interval 54-59 months and the third highest percentage is in interval 48-53 months. Both percentages are higher than the percentage in 0-5 months, which is 3.32%. Cohort data would show a monotonically declining percentage who are surviving twins as the time since birth increases.

Table A3 shows the relevance of Table A2 to the sickle cell analysis. The sickle cell genotyping was intended for all children age 6-59 months in the household regardless of whether their mother was in the household. Some children age 6-59 months in the household schedule were not tested. Table A3 shows, within each 6-month interval, the numbers of children who had a problematic genotype (SC, SS, CC), the number who had a nonproblematic genotype (AA, AS, AC), and the number not tested. The last column gives the percentage with a problematic genotype of all those tested. For example, in the interval 6-11 months, 21 children, or 1.69%, of the 1,240 who were genotyped, had a problematic genotype. The percentage fluctuates across the intervals, but averages 1.29%.

Appendix Table A3 Numbers of surviving children with or without problematic genotypes, Nigeria DHS 2018.

	Genotype			
Time in months	SC, SS, CC	AA, AS, AC	Not tested	Percent SC,SS,CC
0-5	na	na	1,277	na
6-11	21	1,219	57	1.69
12-17	16	1,392	41	1.14
18-23	14	1,099	38	1.26
24-29	15	1,332	47	1.11
30-35	14	1,048	34	1.32
36-41	23	1,369	43	1.65
42-47	14	1,085	37	1.27
48-53	14	1,377	47	1.01
54-59	14	1,144	36	1.21
Total	145	11,065	1,657	1.29

Table A3 does not include the all birth and survivor columns that were in Table A2. Those columns would be different for Table A3 because the Nigeria DHS 2018 survey had a relatively complex design, with long and short questionnaires. Genotyping was not included for all the surviving children in Table A2. We have not identified the mothers whose children were in households eligible for genotyping. For a complete analysis, we would also remove some of the children in Table A3 whose mother was not in the household or was not interviewed.

The point of Table A3 is clear. In Table A2, it is not possible with these numbers to identify something that we know is happening, that is, the higher mortality of children with genotypes SC and SS compared with children with genotypes AA, AS, AC. The difficulty is even greater than in Table A2, because the problematic genotypes are less than half as prevalent as twinning, and only about one-third of children were genotyped.

APPENDIX B GENOTYPES AND ALLELES WITHIN STATES

Appendix Table B1 Prevalence of the genotypes within states, for children age 6-59 months. Nigeria 2018 DHS. Weighted.

State within zone	AA (%)	AS (%)	AC (%)	SC (%)	SS (%)	Other (%)
NC Abuja FCT	80.45	19.55	0.00	0.00	0.00	0.00
NC Benue	80.94	17.95	0.00	0.00	1.11	0.00
NC Kogi	77.25	20.35	0.00	0.63	1.77	0.00
NC Kwara	73.24	17.62	6.72	1.29	0.73	0.41
NC Nasarawa	79.30	18.50	0.37	0.37	1.46	0.00
NC Niger	78.00	17.76	3.22	0.45	0.56	0.00
NC Plateau	83.06	15.29	0.70	0.00	0.95	0.00
NE Adamawa	83.28	15.91	0.00	0.00	0.82	0.00
NE Bauchi	85.74	13.08	0.33	0.00	0.84	0.00
NE Borno	75.27	23.05	0.64	0.00	1.03	0.00
NE Gombe	81.14	17.95	0.25	0.00	0.66	0.00
NE Taraba	75.04	22.30	0.45	0.48	1.73	0.00
NE Yobe	68.38	29.76	0.22	1.08	0.56	0.00
NW Jigawa	72.67	25.20	0.56	0.00	1.56	0.00
NW Kaduna	78.03	21.44	0.53	0.00	0.00	0.00
NW Kano	72.71	24.86	0.00	0.00	2.42	0.00
NW Katsina	82.68	15.21	1.04	0.46	0.62	0.00
NW Kebbi	76.19	17.28	5.29	0.66	0.57	0.00
NW Sokoto	80.59	14.18	3.42	0.00	1.05	0.76
NW Zamfara	80.64	17.43	1.42	0.00	0.50	0.00
SE Abia	84.52	13.44	0.00	0.00	2.04	0.00
SE Anambra	79.02	20.11	0.00	0.00	0.88	0.00
SE Ebonyi	78.97	20.17	0.00	0.14	0.71	0.00
SE Enugu	76.82	22.00	0.00	0.71	0.47	0.00
SE Imo	81.08	17.53	0.00	0.00	1.38	0.00
SS Akwa Ibom	79.48	19.25	0.00	0.00	1.28	0.00
SS Bayelsa	81.27	17.62	0.00	0.00	1.11	0.00
SS Cross River	83.44	16.56	0.00	0.00	0.00	0.00
SS Delta	78.61	20.10	1.29	0.00	0.00	0.00
SS Edo	78.44	21.56	0.00	0.00	0.00	0.00
SS Rivers	81.36	18.64	0.00	0.00	0.00	0.00
SW Ekiti	73.75	23.40	1.06	0.80	0.99	0.00
SW Lagos	68.99	23.20	4.03	2.55	0.14	1.10
SW Ogun	72.44	20.75	4.17	0.99	0.68	0.97
SW Ondo	83.73	13.84	0.64	0.43	1.36	0.00
SW Osun	68.01	20.41	8.99	1.27	1.31	0.00
SW Oyo	68.71	20.28	7.78	1.62	1.23	0.38
Total	77.20	19.72	1.63	0.44	0.88	0.13

Appendix Table B2 Prevalence of the A, S, and C alleles within states, for children age 6-59 months. Nigeria 2018 DHS. Weighted.

State within zone	A	S	C
NC Abuja FCT	0.9022	0.0978	0.0000
NC Benue	0.8992	0.1008	0.0000
NC Kogi	0.8742	0.1226	0.0032
NC Kwara	0.8541	0.1018	0.0441
NC Nasarawa	0.8874	0.1089	0.0037
NC Niger	0.8849	0.0967	0.0184
NC Plateau	0.9106	0.0859	0.0035
NE Adamawa	0.9123	0.0877	0.0000
NE Bauchi	0.9245	0.0738	0.0017
NE Borno	0.8712	0.1256	0.0032
NE Gombe	0.9024	0.0964	0.0013
NE Taraba	0.8642	0.1312	0.0047
NE Yobe	0.8337	0.1598	0.0065
NW Jigawa	0.8556	0.1416	0.0028
NW Kaduna	0.8902	0.1072	0.0026
NW Kano	0.8515	0.1485	0.0000
NW Katsina	0.9080	0.0845	0.0075
NW Kebbi	0.8748	0.0955	0.0298
NW Sokoto	0.8939	0.0814	0.0247
NW Zamfara	0.9007	0.0922	0.0071
SE Abia	0.9124	0.0876	0.0000
SE Anambra	0.8907	0.1093	0.0000
SE Ebonyi	0.8906	0.1087	0.0007
SE Enugu	0.8782	0.1183	0.0036
SE Imo	0.8985	0.1015	0.0000
SS Akwa Ibom	0.8910	0.1090	0.0000
SS Bayelsa	0.9008	0.0992	0.0000
SS Cross River	0.9172	0.0828	0.0000
SS Delta	0.8931	0.1005	0.0065
SS Edo	0.8922	0.1078	0.0000
SS Rivers	0.9068	0.0932	0.0000
SW Ekiti	0.8598	0.1309	0.0093
SW Lagos	0.8260	0.1301	0.0439
SW Ogun	0.8490	0.1155	0.0355
SW Ondo	0.9097	0.0850	0.0053
SW Osun	0.8271	0.1216	0.0513
SW Oyo	0.8274	0.1218	0.0508
Total	0.8788	0.1096	0.0116

APPENDIX C CALCULATION OF DEVIANCE RESIDUALS

The cluster-level deviance residuals are produced within Stata as follows. The child-level data file is collapsed to a cluster-level file, with n as the number of genotyped children age 6-59 months in a cluster and n1 as the number of those children who have some characteristic such as a specific allele. For example, for the analysis of the S allele, a child with genotype SS will contribute "1" toward n1 and a child who is AS or SC will contribute "0.5." Otherwise, the child contributes "0." The collapsed or aggregated data file has one record for each cluster in the entire survey.

The glm (generalized linear models) command for a logit regression with aggregated data is "glm n1, family(binomial n) link(logit)." In the generalized framework, it is not necessary for n1 to be an integer. The output from this command will include a single coefficient, b0, and the total deviance, D. The overall prevalence will match with the antilogit of b0, i.e., P1=exp(b0)/[1+exp(b0)]. If there are K clusters, then D has a chi-square distribution with K-1 degrees of freedom. The output will include the degrees of freedom and the p-value for the chi-square test of the null hypothesis that all clusters have the same probability of the allele. D can be interpreted as a test statistic for the null hypothesis of homogeneity, or a test of the null hypothesis that the probability of the allele is uniform across clusters.

The post-estimation command in Stata to obtain deviance residuals, which we label "dr", is "predict dr, deviance." Define P0=1-P1 and the cluster-specific prevalence p1=n1/n (and p0=1-p1). For a specific cluster define dr2 = 2n[p0 log(p0/P0) + p1 log(p1/P1)]. The dr2 terms are all positive, proportional to n, and have chi-square distributions with one degree of freedom. The sum of the terms across all clusters is the deviance D. The dr terms produced by Stata are the square root of the dr2 terms (and therefore are proportional to the square root of n, with a positive sign if p1>P1 and a negative sign if p1<P1). The dr terms have approximately a unit normal distribution.