The Battle Against Malaria: A Teachable Moment

Randy K. Schwartz

Schoolcraft College

Follow this and additional works at: <https://scholarship.claremont.edu/jhm>

Part of the [Biostatistics Commons](#), [Disease Modeling Commons](#), [Dynamical Systems Commons](#), [Medical Genetics Commons](#), [Probability Commons](#), [Science and Mathematics Education Commons](#), and the [Statistical Models Commons](#)

Recommended Citation

Randy K. Schwartz, "The Battle Against Malaria: A Teachable Moment," *Journal of Humanistic Mathematics*, Volume 7 Issue 1 (January 2017), pages 204-228. DOI: 10.5642/jhummath.201701.16. Available at: <https://scholarship.claremont.edu/jhm/vol7/iss1/16>

©2017 by the authors. This work is licensed under a Creative Commons License.

JHM is an open access bi-annual journal sponsored by the Claremont Center for the Mathematical Sciences and published by the Claremont Colleges Library | ISSN 2159-8118 | <http://scholarship.claremont.edu/jhm/>

The editorial staff of JHM works hard to make sure the scholarship disseminated in JHM is accurate and upholds professional ethical guidelines. However the views and opinions expressed in each published manuscript belong exclusively to the individual contributor(s). The publisher and the editors do not endorse or accept responsibility for them. See <https://scholarship.claremont.edu/jhm/policies.html> for more information.

The Battle Against Malaria: A Teachable Moment

Randy K. Schwartz

Department of Mathematics, Schoolcraft College, Livonia, Michigan, USA
rschwart@schoolcraft.edu

Synopsis

Malaria has been humanity's worst public health problem throughout recorded history. Mathematical methods are needed to understand which factors are relevant to the disease and to develop counter-measures against it. This article and the accompanying exercises provide examples of those methods for use in lower- or upper-level courses dealing with probability, statistics, or population modeling. These can be used to illustrate such concepts as correlation, causation, conditional probability, and independence. The article explains how the apparent link between sickle cell trait and resistance to malaria was first verified in Uganda using the chi-squared probability distribution. It goes on to explain that the incidence of sickle cell within a given population is an example of an asymptotically stable equilibrium determined by the selective pressure of malaria. It summarizes the impact of malaria on human history in order to explain why this equilibrium has varied over time and space. Finally, the article summarizes how linkage analysis and other statistical modeling techniques are being used as an important step in developing genomic pharmaceuticals to combat such diseases.

1. Introduction

About once per minute on average, somewhere in the world a child dies of malaria [30]. Throughout the broad sweep of recorded human history, this has been the single most deadly and debilitating health problem [11, page 5]. Although progress has been made over the last two centuries in understanding and combating the disease, it remains stubborn—and at certain times and places, resurgent—infesting hundreds of millions of new victims every year. A single outbreak in Ethiopia in 1958, triggered by excessive rainfall, killed 150,000 people [11, page 5, footnote 8].

Figure 1: Malaria clinic in Tanzania. CC-BY photo (Wikimedia Commons) by Olympia Wereko-Brobby for *Saving Lives with SMS for Life*, 23 October 2009.

Today, we know that malaria is an infectious disease caused by microbes spread by certain species of mosquitoes. But to casual observers, this was far from obvious. For millennia, other, more easily-seen factors, such as diet or air quality, were widely blamed for the illness. To identify the actual factors relevant to malaria, within our bodies and within our environments, medical science has had to call upon many tools, from microscopes to mathematics.

In this article, we show that the long-running struggle against malaria provides engaging educational material that can be introduced in lower- or upper-level courses dealing with probability, statistics, or population modeling. Methods of statistical inference have been used to help identify which natural factors are actually involved in susceptibility and outbreaks of malaria, and which factors could potentially be exploited to eradicate the disease from human populations. Population models have been used to understand how the sickle-cell trait and other genetic safeguards against susceptibility to malaria became fixed within certain human populations. Given the horrendous ongoing toll of the disease, these are examples and case studies that carry “life and death” import.

2. Correlation $\neq$ Causation

Before getting into the number-crunching, there are some important underlying concepts for instructors and students to get clear on.

One of the most common mistakes that people can make in trying to understand how the world works is to assume that if condition B arises more commonly when factor A is present, then A must cause B. In statistics, this fallacy is known as mistaking correlation for causation.

For centuries, malaria was called “swamp fever”: people were certain that it was caused by habitually breathing the air found around swamps and marshland. Certainly all evidence pointed to this. Victims of malaria almost invariably resided in swampy regions. As long ago as 100 BCE in the Po River valley of Italy, people found that if they built their huts on stilts high above such stagnant water, instead of closer to the surface, the incidence of swamp fever notably reduced. Aristocratic Italian families who had the means to drain marshes on their huge estates were rarely afflicted with the scourge. It was as recently as 1827 that doctors working on this illness coined the term *malaria*, from Italian words meaning “bad air” [11, pages 5–7].

But this just shows how careful we have to be before jumping to conclusions about cause and effect. We now know that the mosquitoes that spread malaria can only lay their eggs in water; if pools of water are not present, the mosquitoes cannot reproduce. Swampy air (A) and malaria (B) are in fact correlated, because if one is present, then it means the other is more likely to be present, too. But this does not mean that the swampy air is what causes the malaria. Rather, it is swamp water itself (C) that allows mosquitoes to breed, and the mosquitoes spread malaria. So, the swamp water fosters both the swamp vapors and the malaria; the vapors do not cause the malaria. This shows that the correlation between A and B can actually be due to some lurking factor C that promotes the existence of both A and B.

Here is another example. In West Africa, malaria has long been a huge problem, infecting up to 60% of the population in some regions [11, page 17]. Observers thought the disease might have a nutritional cause, since it seemed to be most prevalent where the staple food was yams. As it turned out, yam cultivation traditionally relied on slash-and-burn agriculture, a technique that creates more breeding locations for mosquitoes. It was the mosquitoes that were causing malaria, not the yams [28]. Slash-and-burn yam cultivation and

malaria are in fact correlated, because if one is present, it means that the other is more likely to be present, too. But this does not mean that a diet of yams is what causes the malaria.

In other parts of the world, when cattle were removed to make way for more farming, malaria often became a bigger problem immediately [11, page 34 footnote 7]. It would be easy to fall into the mistake of blaming certain foods that were being eaten more frequently once the cattle land was turned into crop land. In fact, what was happening was that with cattle present, the mosquitoes more often gorged themselves on cattle blood rather than human blood. It was the presence of the cattle that had helped thwart the spread of mosquito-borne illness among humans. In other words, the nutritional change relevant to malaria's spread was the mosquitoes' nutrition, not the humans' nutrition!

Figure 2: In the 1890s it was finally shown that mosquitoes, such as the *Anopheles stephensi* species photographed here, are responsible for spreading malaria parasites from one host's bloodstream to another. Until then, many different factors that were merely correlated with the disease were erroneously blamed as its cause, including swampy air and yam consumption. Public domain photo by Jim Gathany, Public Health Image Library (<https://phil.cdc.gov/phil/details.asp?pid=5814>), United States of America Centers for Disease Control and Prevention.

3. Conditional Probability and Independence

To see whether a given factor A is associated in some way with malaria, we can first check whether the incidence of the illness remains the same whether A is present or not. More precisely, we check whether the incidence of malaria when A is present is the same as the incidence of the disease overall. If so, then the likelihood of contracting malaria is said to be *independent* of factor A .

Suppose that a newly developed anti-malarial drug A is subjected to a controlled field test, with a representative sample of 80 healthy subjects volunteering to participate over a five-year period. The results are summarized in this contingency table:

	malaria	no malaria	totals
treated with A	4	28	32
treated with placebo	6	42	48
totals	10	70	80

Within this group, the overall probability of contracting malaria was:

$$\Pr(M) = 10/80 = 12.5\%.$$

We can compare this with the conditional probability of contracting malaria, given treatment with drug A :

$$\Pr(M|A) = 4/32 = 12.5\%.$$

In medicine, such percentage measures of the incidence of a disease within a population are called *morbidities*. In this case, since the morbidity in the presence of A is the same as the overall morbidity (12.5%), it suggests that the new drug is ineffective, neither increasing nor decreasing the likelihood of contracting malaria. We say that contracting malaria and being treated with the new drug are independent events. So, two events M and A are independent if $\Pr(M) = \Pr(M|A)$.

The term $\Pr(M)$ can be considered as the probability of contracting malaria prior to any assumption about drug treatment, and $\Pr(M|A)$ as the probability of contracting malaria posterior to the assumption that the subject was treated with drug A . (Bayes's Theorem gives a general relation between prior

and posterior probabilities, and is the root of a whole approach to statistical inference called *Bayesian statistics*.)

Notice that we can also express the conditional probability this way:

$$\Pr(M|A) = \frac{4}{32} = \frac{4/80}{32/80}$$

$$\Pr(M|A) = \frac{\Pr(M \cap A)}{\Pr(A)}$$

where $M \cap A$ refers to the conjunction or intersection of the events. Now recall that independence means $\Pr(M) = \Pr(M|A)$, so in that case the above equation becomes:

$$\Pr(M) = \frac{\Pr(M \cap A)}{\Pr(A)}$$

Multiplying to clear the fraction, we arrive at an important new way to characterize independence:

$$\Pr(M \cap A) = \Pr(M)\Pr(A).$$

We can also illustrate this with the following Venn diagram, where we have labeled the events with probabilities found directly from the earlier chart.

The likelihood of the conjunction of events M and A, namely $4/80 = 5\%$, is called the *conjoint probability* and is written in the overlap of the two circles. To check for independence, we use the above characterization as follows.

Representing the fractions as decimals makes it easier to check for equality or inequality.

$$\begin{aligned}\Pr(M \cap A) &= \Pr(M)\Pr(A) \\ 0.05 &= 0.125 \cdot 0.40 \\ 0.05 &= 0.05\end{aligned}$$

The two sides of the equation match if and only if the events are mutually independent.

When people first learn about dependence and independence of events, they sometimes misconstrue it as a case of “all or nothing.” That is, they think dependence means that if one event occurs then the other one must occur (“all”), and they think independence means that if one event occurs then the other one cannot occur (“nothing”). Neither of these are true. If events M and A cannot both occur, they are called disjoint or mutually exclusive events. In that case, the two circles in the Venn diagram have no overlap, and can be drawn as such. Such a pair of disjoint circles might be called “independent” in everyday lingo, but not in statistics. Run the test $\Pr(M \cap A) = \Pr(M)\Pr(A)$, and you will see that disjoint events are in fact always dependent provided that neither event is impossible.

To keep this straight, I like to tell my students to compare the events M and A to a pair of former best friends who have had a falling-out. To be independent, they cannot deliberately avoid one another (be disjoint) because then, the movements of each one would affect the other one as they make sure not to cross paths. Instead, to be genuinely independent they must simply ignore one another, which means that they will still bump into each other on occasion (overlap), no more nor less than if they had never been best friends!

4. Sickle Cell and Malaria are Dependent Events

By the 1940s, people in the biomedical community had begun to notice an apparent relationship between malaria and sickle-cell trait, a genetic condition that is relatively common in central Africa.

This inherited trait causes many of the red blood cells to have an abnormal shape, like a crescent or sickle. Sickle cell is an example of an autosomal recessive trait: to have the condition in its full-blown form, a person must

inherit the trait from both parents. The sickling of the corpuscles obstructs their free flow in the blood vessels and their ability to carry oxygen. The worst symptoms are anemia, chronic organ damage, and debilitating pain. Without medical intervention, many of these victims do not survive to adulthood [20].

A much larger number of people inherit sickle-cell trait from only one of their parents. They have some sickled blood cells, but almost no outward symptoms of a disorder; the condition is called *sickleemia*. Because they can pass the trait on to their children, they are said to be “carriers” of the trait.¹

Figure 3: A sickled blood corpuscle beside three normally-shaped corpuscles. Public domain photo by Alisa Zapp Machalek for National Institute of General Medical Sciences, National Institutes of Health, 2000.

What was noticed in Africa was that children with sickleemia hardly ever contracted severe cases of malaria. They actually seemed to have more resis-

¹For example, if two individuals with sickleemia have children together, then, on average, a quarter of their offspring will be born with sickle-cell anemia, another quarter will be “normal” (i.e., without the trait at all), and the remaining one-half will have sickleemia.

tance to malaria than either of the other two groups: children with sickle-cell anemia and normal children. (This unusual situation, where heterozygous individuals have greater fitness than homozygous ones, is called *overdominance* or *heterozygous advantage*.) It was also noticed that the frequency of sicklemlia, meaning the probability, or percentage, of the trait among the population, seemed to be highest in precisely the areas of Africa where malaria was most prevalent and where resistance to it would be most important.

To test this apparent correlation between sicklemlia and malaria resistance, in 1953 the young medical researcher Anthony C. Allison took a team to Ganda, a village outside Kampala, Uganda. Allison had himself been stricken with malaria while growing up on a chrysanthemum farm in East Africa, before he eventually went to Oxford to study medicine. In Ganda, his team took blood samples from 290 children aged 5 months to 5 years, and tested these for malaria parasites and for sicklemlia. Their results were as follows [1]:

	malaria	no malaria	totals
sicklemlia	12	31	43
no sicklemlia	113	134	247
totals	125	165	290

To check for a possible relation between the events malaria (M) and sicklemlia (S), we use our criterion for independence:

$$\begin{aligned}\Pr(M \cap S) &= \Pr(M)\Pr(S) \\ \frac{12}{290} &= \frac{125}{290} \cdot \frac{43}{290} \\ 0.041 &= 0.064\end{aligned}$$

Since the two sides of the equation do not agree, the two events are not independent. This suggests that there might be a relation between sicklemlia and malaria; if so, we say that these two factors are *mutually dependent*. In fact, the conjoint probability on the left-hand side of the equation is lower than we would have expected if the two events were independent, so this suggests that the sicklemlia children had a lower chance of contracting malaria than did the other children.

5. Chi-Squared as a Measure of Dependence

Readers have a right to be skeptical about the discrepancy between 0.041 and 0.064 that we just noted: do a couple percentage points create a large enough gap for us to be confident that malaria and sicklemlia have any relation with one another? After all, our sampling of patients will always be subject to some randomness. This can lead to what is called *sampling error*: even if two factors are independent, the numbers that we get from any particular sample might not satisfy the independence equation precisely. If the discrepancy is very small, then it would be foolish to rush off claiming that we have proven that the two factors are mutually dependent.

How do we judge whether a discrepancy is big enough to claim with confidence that two factors are mutually dependent? Over a century ago, in the year 1900, the British mathematician Karl Pearson developed a statistical test for this purpose, as follows.

Let us take another look at the calculation we just made:

$$\frac{12}{290} = \frac{125}{290} \cdot \frac{43}{290}.$$

Multiplying both sides by the sample size converts the probabilities into counts of children:

$$12 = \frac{125 \cdot 43}{290}.$$

The left side represents the 12 children whom Allison's team observed had both malaria and sicklemlia. On the right side, the term $125/290$ is the fraction of children with malaria, and 43 is the number of children with sicklemlia. So if these two events were independent, then we would expect $125/290$ of the 43 children with sicklemlia to also have malaria. (Of course, we can also look at it the other way: $43/290$ have sicklemlia, so if the two events were independent then we would expect $43/290$ of the 125 children with malaria to also have sicklemlia.) Thus, the right side, which is about 18.534, represents the number of children (having both malaria and sicklemlia) that would have been expected if the two events were independent.

Pearson called the 12 and the 18.534 the *observed* (O) and *expected* (E) counts, respectively. Notice that E is calculated by multiplying the row total (43) by the column total (125), then dividing by the grand total number

of children (290). This can be done for each of the four cells of observed data, and we write the resulting expected values in parentheses next to the corresponding observed values:

	malaria	no malaria	totals
sickleemia	12 (18.534)	31 (24.466)	43
no sickleemia	113 (106.466)	134 (140.534)	247
totals	125	165	290

(Notice that the expected values, in parentheses, still obey the same row and column totals as did the original observed figures.)

In Pearson's method, we aggregate the relative sizes of the discrepancies within the cells according to the formula,

$$\chi^2 = \sum \frac{(O - E)^2}{E},$$

where the sum is taken over all four cells. The result is a statistic called *chi-squared*:

$$\chi^2 = \frac{(12 - 18.534)^2}{18.534} + \dots + \frac{(134 - 140.534)^2}{140.534} \approx 4.754.$$

Let us examine the rationale underlying this statistic:

- The differences are squared because we are interested only in the magnitude of each discrepancy, not in its direction, positive or negative (which indicates which one is larger, O or E).
- Division is used in order to state the relative magnitude of a given discrepancy.
- The term appearing in the denominator is E (rather than O) because we are testing the observed rather than the expected values. In other words, since, we want to know how well the observed values compare to those that we expected under the assumption of independence, we express the discrepancy relative to what we expected and assumed.

- And of course, addition is used in order to aggregate these relative magnitudes of discrepancy. Thus, the larger the value of χ^2 , the greater the aggregate relative discrepancy between what was actually observed in our random sample and what would have been expected under the assumption that the factors being considered are mutually independent.

6. Using a Probability Distribution

To complete Pearson's method, a statistical function called the *cumulative chi-squared distribution* is used to calculate the probability that a given value of χ^2 is so large that it must have arisen from a dependence among the factors, instead of merely from chance (statistical error). This function can be evaluated with calculus techniques or by using either a graph (as shown below), a published table, or a statistical software package.

The curve shown in the graph expresses and quantifies the concept that the larger the computed value of χ^2 , the greater the discrepancy between the observed numbers and the numbers expected under the assumption of independence—and therefore the greater the likelihood that that assumption is incorrect. Based on the above curve, a χ^2 value of 8 or more implies near-certainty (a probability of virtually 100%) that the two factors are mutually dependent. Using instead the value $\chi^2 = 4.754$ obtained from Allison's 2×2 contingency table, we find that there is approximately a 97.08% probability

that the two factors are dependent. Provided that Allison's sample of 290 children was genuinely a representative one, this means that we can be 97.08% confident in claiming that malaria and sickle cell anemia are related (dependent) factors; the other 2.92% is our level of doubt about this conclusion.

Probability distributions such as the one we just used are the central tools of *statistical inference*. They allow us to quantify how likely or unlikely a given assumption about a population is, based on a random sample. Each distribution is a function that maps the sample space (the set of all possible values of the statistic, or variable) to the closed interval $[0,1]$ of probabilities.

The χ^2 distribution is defined as a sum of squares, but actually in a more general way than we did above. It is defined by the probabilities associated with the sum of the squares of k independent variables, assuming that each individual variable has a probability distribution of the standard normal type (the most important type of bell-shaped curve). The exact shape and formula for the χ^2 curve depends on k , the number of variables whose squares are added to obtain it.

What Pearson proved in his path-breaking paper of 1900 [21] is that for any two independent phenomena, if a sufficiently large random sample of observations is summarized in an $m \times n$ contingency table (m possible outcomes for the first phenomenon and n for the second), then the probabilities associated with each fraction $(O - E)^2/E$ (one fraction for each cell) approximate the square of a normally distributed random variable. Thus, Pearson showed, the sum $\sum (O - E)^2/E$ will have a distribution that is well-approximated by the χ^2 distribution.

But in a contingency table, the numbers in the last row and the last column are not independent of the others: they are constrained by the fact that all of the row and column totals for the expected values must match those for the observed values. Thus, in an $m \times n$ contingency table, the independent (unconstrained) cells form a rectangular sub-grid with $(m - 1)$ rows and $(n - 1)$ columns, and the number of independent fractions $(O - E)^2/E$ is only $k = (m - 1)(n - 1)$, called the number of *degrees of freedom*. For a 2×2 grid such as the one Allison recorded, we get $k = 1$. So the χ^2 curve that we selected earlier corresponds to $k = 1$ and therefore applies to any 2×2 contingency table. For other table sizes, the curve will vary according to the parameter k . For more information about either the normal or the chi-squared distribution, readers can consult any standard statistics textbook.

7. Genetic Diversity as a Dynamic Equilibrium

“But wait a minute,” you might object, “just because we’re confident that sicklemlia and resistance to malaria are dependent events, that does not mean that sicklemlia causes the resistance to malaria.” Again, you are exactly right to be skeptical! As was emphasized above, even a strong correlation between two factors does not necessarily imply causation. In fact, statistical probabilities alone can never prove causation; they can only alert us to factors that merit further investigation.

The only way to prove causation with certainty is to discover and observe an actual mechanism whereby one event is causing the other to occur. And that is exactly what happened in 2011, when biomedical scientists discovered the mechanisms by which the body of an individual with sicklemlia is able to tolerate the effects of an invasion by malaria parasites and thwart the disease from engulfing the body [6, 7].

Allison’s statistical field surveys opened the way to investigating other such genetic correlates to health. Just as with sicklemlia, several other genetic disorders of the red blood cell (including thalassemia, Cooley’s anemia, ovalocytosis, G6PD deficiency, Duffy antigen deficiency, HbC, and HbE), all widespread in historically malaria-ridden regions, were shown to be statistically correlated with heightened resistance to malaria [17]. Collectively, these red-corpuscle abnormalities represent the most common type of genetic disease in the human population. Other dependencies in which a genetic disorder protects against an infectious disease have also now been demonstrated, such as Tay-Sachs/tuberculosis, cystic fibrosis/cholera, and phenylketonuria/ochratoxin infection.

A lesson here is that a gene that is “bad” in one way can also be “good” in other ways. What makes a gene beneficial or harmful for survival depends on what other genes it is combined with in the same genome, and depends especially on the environment in which people or other organisms find themselves. Having a human gene pool that is very diverse, as opposed to one that is uniform and seemingly “perfect”, increases the probability that our species can survive and thrive in the changing circumstances that we are sure to face in the future.

Often, when a genetic mutation arises, it confers an unambiguous fitness advantage or disadvantage. This introduces to the gene pool an unstable diversity in which the less-favored gene slides toward extinction. By contrast, overdominance is an instance in which genetic diversity approaches a limit, also called an equilibrium or steady state. At the equilibrium point the gene variants coexist at respective frequencies that are perfectly balanced, in the sense that, taking into account the fitness advantages and disadvantages of each one, the weighted-average fitness of the population as a whole is maximized.

In particular, even though the sickle-cell trait causes life-shortening anemia in those who inherit it from both parents, it confers life-prolonging protection against malaria among the much larger numbers of carriers who inherit the trait from one parent only. This is why the trait's frequency reached a stable, positive level in conditions where malaria was endemic. Further, this type of equilibrium is asymptotically stable: if there is a perturbation (a relatively short-term event that knocks the gene frequencies a small distance further away from their equilibrium, such as a local human famine or a temporary plunge in the mosquito population), then gradually over succeeding generations the frequencies will shift back toward their steady state as if pulled by a force of attraction.

Researchers can mathematically predict these equilibria using techniques of population genetics. The gene-pool frequencies found at the steady state depend entirely on the relative fitnesses of the gene variants, and these fitnesses are affected by the selective pressure exerted by diseases and other environmental factors. For example, when the conditions that spread malaria recede for a long period of time, the equilibrium point tends to shift in the direction of a smaller fraction of the population carrying the sickle trait. (For details on the mathematics, see [9, pages 135–140].)

8. How Malaria Shaped World History

The rise of the sickle-cell trait to a stable equilibrium in the human population had two major impacts on world history before humans even understood what was happening.

First, the trait allowed humans to adapt to parts of the Earth where swamps, marshland, and therefore malaria were common, first in Africa where the human species was born, then in the Mediterranean basin, the Middle East, and South and East Asia. The genetic mutation that causes the sickling of the corpuscles—a simple transposition of two nucleotides on a single strand of human DNA, known as a *single-nucleotide polymorphism* (SNP)—arose and took hold in five different world regions, four of them in central Africa. Statistical modeling was used to determine that the five mutations arose independently of one another [18].

The second major global impact of the relation between malaria and sickle-cell anemia was that it hastened the enslavement of millions of Africans and the transplantation of their people and culture across the Atlantic. With the arrival of the first European colonists in the New World, malaria and other Old World diseases were introduced, decimating native populations. Malaria became endemic wherever there was enough marshland to sustain the disease, notably in the American South, Brazil, and the West Indies. To work the rice and sugarcane fields of these regions, the colonial plantation-owning class favored slaves from West Africa, among whom immunity to malaria was high. It is estimated that of the African captives arriving in what is now the United States of America, at least 22% possessed the gene for sickle cell trait, about 20% possessed the gene for G6PD deficiency, and about 70% possessed the gene for Duffy antigen deficiency—all of which confer resistance to one or another form of malaria [24]; also see [23, page 27]. One historian who studied the rice-growing slave plantations in colonial South Carolina concluded:

Sickle-cell trait, the negative consequences of which are only now [in the early 1970s] being studied seriously, may in the seventeenth and eighteenth centuries have had a positive influence in warding off malaria which gave its bearers an obvious if highly dubious advantage in the cultivation of rice [29, page 89].

Later, whole global industries arose based on the production of anti-malarial drugs. A plant substance found in the New World, an Andean tree bark named *cinchona*, turned out to be very effective in relieving malaria symptoms. Quinine, a pharmaceutical first isolated from cinchona bark in 1820, became a key tool in facilitating the imperial expansion of Western powers in the 19th Century [22]. The global availability of quinine also helped

reduce the economic pressures for African servitude. Instead, contract laborers, such as Asians in the American West, were often administered daily doses of quinine by their bosses if they worked in malaria-ridden areas. The quinine cartel that emerged in 1913 to integrate the global industry, from overseas cinchona plantations to industrial factories, was the world's first pharmaceutical cartel [8]. The first generation of synthetic pharmaceuticals for use against malaria symptoms was based on mimicking the effects of quinine. The efforts to design these anti-malaria drugs prompted the first foray into coal-tar chemistry, which became the foundation for the rise of modern chemistry as a whole in the late 1800s [11, pages 23–28].

Due to such advances in the pharmaceutical industry and in the public health infrastructure, the incidence of malaria gradually decreased in the United States, Italy, and other industrialized countries. Malaria has been the single largest known factor for natural selection among humans in recorded history [17], yet the disease has lost most of its selective pressure in regions where it has ceased to be endemic for the past century or more. In those places, malaria has not been prominent enough for sickle-cell carriers to maintain a significant fitness advantage.² This goes a long way toward explaining why the incidence of sickle-cell anemia among African-Americans has fallen to about 8% today [2]. The rise of the sickle-cell trait toward a stable equilibrium in Africa in the past 10,000 years, and the collapse of that equilibrium in the United States after only a century of malaria-free conditions, are vivid examples of the importance of natural selection in evolution.

9. Locating a Cure in the Human Genome

Based on what has been summarized here, it should not surprise you if, in the not too distant future, certain diseases are eradicated after teams of researchers mine databases of genetic and public health information to identify favorable mutations, and then engineer pharmaceuticals to mimic their effects. CCR5 inhibitors, an important new class of anti-retroviral drugs for treatment of HIV infection, were developed based on a natural genetic mutation that affords protection to homozygous individuals [12]. Currently

²In conditions in which the sickle-cell trait has no selective advantage, interbreeding among carriers can lead to as much as a 50% decrease in the carrier fraction of the population in each generation (see Footnote 1).

under development are drugs, also inspired by natural mutations, to help prevent heart disease by lowering blood levels of triglycerides [14, 15, 27] or LDL cholesterol [16, 26]. Statistical inference is playing a crucial role in this emerging genomics-based strategy for public health, called *genomic epidemiology*.

Because of the association with sickle cell anemia and other protective mutations, malaria might well become the first communicable disease to be brought under control by drugs that mimic a heterozygous trait. In 1949, sickle-cell trait became the first “molecular disease” ever identified as such—the first disorder, in any species, attributed to a mutation that causes a single amino-acid substitution [2]. And now that cellular and molecular biologists understand the mechanism by which that mutation protects the body from the effects of a malaria infection, they believe such knowledge could point to a molecular pathway that could be exploited in developing a much more effective drug or even a vaccine. That would be “right on time” for the planet, given that:

- Every year, on the order of 200 million new human victims are infected with malaria. Hundreds of thousands are dying annually, about 90% of them in Africa [30].
- The current generation of anti-malarial drugs is being compromised as parasites evolve resistance to them [4].
- Mosquitoes are evolving to become resistant to the pyrethroids and other insecticides currently being used against them [5].
- Climate change may be making additional swaths of the planet—not just the tropical and subtropical regions—hospitable to the *Anopheles* mosquitoes that spread malaria [13].

The first step in the genomic strategy to defeat malaria, or any disease, is to identify which physical traits have spread most rapidly and achieved a stable equilibrium in historically disease-ridden areas. To do so requires statistical models, of which the chi-squared test discussed earlier is a simple example. Mining the human genome—with its millions of nucleotides—for protective mutations requires having data sets large enough to detect subtle differences in morbidity. This was much of the motivation behind the

2008 founding of the Malaria Genomic Epidemiology Network (MalariaGEN, <http://www.malariagen.net/>), a data-sharing collaboration involving researchers in more than twenty countries.

Once a physical trait is identified as significant in relation to a disease, locating where the gene or genes for it are situated on the chromosomes is important, because the sequence of nucleotides reveals the corresponding sequence of amino acids in the protein(s) or enzyme(s) responsible for the trait. But pinpointing one or a few genes among the thousands that comprise a genome is like looking for a needle in a haystack. A statistical modeling technique called *linkage analysis* or *association analysis* has been essential. It exploits the fact that, due to the mechanics of genetic recombination, two genes on the same chromosome tend to be co-inherited by a single individual with a probability that increases with their proximity on the chromosome. If we already know the chromosome position of one gene, called a *marker*, then linkage analysis narrows down where the second gene might be located. Linkage disequilibrium mapping is the analogous tool used when the physical trait results from two or more genes located on different chromosomes [10].

Other sophisticated modeling tools being used in this effort include:

- *stochastic models and simulation*, used to track the evolution of gene frequencies over time;
- *Bayesian models*, mentioned earlier, used to estimate the age and natural selection rate of a known genetic mutation;
- *selection equations*, used to calculate genetic equilibria and the rates of change around them; and
- *regression models*, used to quantify the association between genes and morbidity.

For more details about this work, see [10, 25, 28].

Engineering a pharmaceutical based on understanding mechanisms of natural genetic immunity to malaria could rid the planet of mankind's worst disease in history. That would also be a stunning harbinger of the new world being opened up by combining statistical modeling and biomolecular analysis in the service of public health.

10. Exercises

Students may be directed to these exercises after an in-class discussion of some of the material in the earlier parts of the article, or the whole article together with the exercises may be assigned for out-of-class work.

1. In each part, do the data provided give evidence that the two events A and B are dependent, or that they are independent?

(a) Consider data:

	B	not B
A	302	2737
not A	253	1108

(b) For Venn diagram:

(c) For Venn diagram:

(d) For Venn diagram:

2. In 1946 E. A. Beet, a British medical officer stationed in Northern Rhodesia (now Zimbabwe) reported findings from blood tests that he administered to a group of native-African residents of the Balovale district [1]:

	malaria	no malaria
sickleemia	10	92
no sickleemia	75	416

- Estimate the respective probabilities of malaria, of sickleemia, and of their conjunction.
 - Estimate the conditional probability of contracting malaria, given that the subject had sickleemia.
 - Use the results of parts (a) and/or (b) to determine whether malaria and sickleemia are dependent or independent events.
 - Calculate the chi-squared statistic, χ^2 .
 - Use the chi-squared probability distribution to estimate the confidence that malaria and sickleemia are related, based on this being a representative sample.
3. The use of insecticide-treated bed nets (ITNs) in mosquito-infested regions has proven to be an effective measure for reducing the transmission of malaria among children. These are fine meshes of polyester or other synthetic fiber treated with a pyrethroid chemical that repels or kills mosquitoes and other insects. An early randomized controlled trial, involving children aged 1-59 months in over 200 villages of western Kenya, yielded the following results during the first phase [19]:

	malaria	no malaria
used ITNs	15	328
no ITNs	27	267

- By what proportion was the morbidity lower with ITN use compared to the control?
- Estimate the respective probabilities of malaria, of ITN use, and of their conjunction.
- Estimate the conditional probability of a child contracting malaria, given that the child used an ITN.

- (d) Use the results of parts (b) and/or (c) to determine whether malaria and ITN use are dependent or independent events.
 - (e) Calculate the chi-squared statistic, χ^2 .
 - (f) Use the chi-squared probability distribution to estimate the confidence that malaria and ITN use are related, based on this being a representative sample.
4. The following chart classifies each of the years 1870 through 1945 according to whether there was a malaria epidemic in Sri Lanka and whether there was an El Niño event, when sea surface temperatures in the Southern hemisphere are unusually warm [3]:

	epidemic	no epidemic
El Niño year	9	11
non-El Niño year	7	49

- (a) Estimate the respective probabilities that a given year brings a malaria epidemic, an El Niño event, or both.
- (b) Estimate the conditional probability of an epidemic in an El Niño year.
- (c) Use the results of parts (a) and/or (b) to determine whether epidemics and El Niño events are dependent or independent.
- (d) Calculate the chi-squared statistic, χ^2 .
- (e) Use the chi-squared probability distribution to estimate the confidence that epidemics and El Niño events are related, based on this being a representative sample.
- (f) Research the association between El Niño events and malaria epidemics, and propose a mechanism that plausibly accounts for it.

References

- [1] Anthony C. Allison, “Protection afforded by sickle-cell trait against subtertian malarial infection,” *British Medical Journal*, Volume **1954:1** (4857) (February 6, 1954), pages 290–294.

- [2] Anthony C. Allison, “The discovery of resistance to malaria of sickle-cell heterozygotes,” *Biochemistry and Molecular Biology Education*, Volume **30** Number 5 (2002), pages 279–287.
- [3] Menno J. Bouma, and Hugo J. van der Kaay, “The El Niño Southern Oscillation and the historic malaria epidemics on the Indian subcontinent and Sri Lanka: An early warning system for future epidemics?,” *Tropical Medicine and International Health*, Volume **1** Number 1 (February 1996), pages 86–96.
- [4] A. Taylor Bright and Elizabeth A. Winzeler, “Resistance mapping in malaria,” *Nature*, Volume **498** Number 7455 (June 27, 2013), pages 446–447.
- [5] Centers for Disease Control and Prevention, “Insecticide-treated bed nets,” online resource at http://www.cdc.gov/malaria/malaria_worldwide/reduction/itn.html, accessed on January 28, 2017.
- [6] Cyrklaff, Marek, et al., “Hemoglobins S and C interfere with actin remodeling in Plasmodium falciparum-infected erythrocytes,” *Science*, Volume **334** Number 6060 (December 2, 2011), pages 1283–1286.
- [7] Ana Ferreira et al., “Sickle hemoglobin confers tolerance to Plasmodium infection,” *Cell*, Volume **145** Number 3 (April 29, 2011), pages 398–409.
- [8] Andrew Goss, “Building the world’s supply of quinine: Dutch colonialism and the origins of a global pharmaceutical industry,” *Endeavour*, Volume **38** Number 1 (March 2014), pages 8–18.
- [9] Philip W. Hedrick, *Genetics of Populations*, 2nd ed., Jones and Bartlett, Sudbury MA, 2011.
- [10] Philip W. Hedrick, “Population genetics of malaria resistance in humans,” *Heredity*, Volume **107** (2011), pages 283–304.
- [11] Henry Hobhouse, *Seeds of Change: Five Plants that Transformed Mankind*, Harper & Row, New York NY, 1985, Chap. 1, “Quinine and the White Man’s Burden,” pages 3–40.
- [12] Gero Hütter et al., “Long-Term control of HIV by CCR5 Delta32/Delta32 stem-cell transplantation,” *New England Journal of Medicine*, Volume **360** Number 7 (February 12, 2009), pages 692–698.

- [13] Atul A. Khasnis and Mary D. Nettleman, “Global warming and infectious disease,” *Archives of Medical Research*, Volume **36** (2005), pages 689–696.
- [14] Gina Kolata, “Gene mutation is tied to fast fat breakdown,” *New York Times*, December 11, 2008. Available online at <http://www.nytimes.com/2008/12/12/health/research/12heart.html>, accessed on January 28, 2017.
- [15] Gina Kolata, “In single gene, a path to fight heart attacks,” *New York Times*, June 19, 2014. Available online at <https://www.nytimes.com/2014/06/19/health/scientists-identify-mutations-that-protect-against-heart-attacks.html>, accessed on January 28, 2017.
- [16] Gina Kolata, “Drug to fight heart attack clears panel,” *New York Times*, June 10, 2015. Available online at <https://www.nytimes.com/2015/06/10/health/cholesterol-levels-ldl-drugs-heart-attack-s-fda-panel.html>, accessed on January 28, 2017.
- [17] Dominic P. Kwiatkowski, “How malaria has affected the human genome and what human genetics can teach us about malaria,” *American Journal of Human Genetics*, Volume **77** Number 2 (August 2005), pages 171–192.
- [18] Claudine Lapoumroulie et al., “A novel sickle cell mutation of yet another origin in Africa: The Cameroon type,” *Human Genetics*, Volume **89** Number 3 (May 1992), pages 333–337.
- [19] Kim A. Lindblade et al., “Sustainability of reductions in malaria transmission and infant mortality in western Kenya with use of insecticide-treated bednets,” *Journal of the American Medical Association*, Volume **291** Number 21 (June 2, 2004), pages 2571–2580.
- [20] National Heart, Lung, and Blood Institute, “What Is Sickle Cell Disease?,” online resource available at <https://www.nhlbi.nih.gov/health/health-topics/topics/sca>, accessed on January 28, 2017.
- [21] Karl Pearson, “On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling,”

Philosophical Magazine Series 5, Volume **50** Number 302 (1900), pages 157–175.

- [22] Arjo Roersch van der Hoogte and Toine Pieters, “Science in the service of colonial agro-industrialism: The case of cinchona cultivation in the Dutch and British East Indies, 1852-1900,” *Studies in History and Philosophy of Biological and Biomedical Sciences*, Volume **47** Part A (September 2014), pages 12–22.
- [23] Todd Lee Savitt, *Medicine and Slavery: The Diseases and Health Care of Blacks in Antebellum Virginia*, University of Illinois Press, Urbana IL, 1978.
- [24] Todd Lee Savitt, “The invisible malady: sickle cell anemia in America, 1910-1970,” *Journal of the National Medical Association*, Volume **73** Number 8 (1981), pages 739–746.
- [25] Montgomery Slatkin, “A Bayesian method for jointly estimating allele age and selection intensity,” *Genetics Research (Cambridge)*, Volume **90** Number 1 (February 2008), pages 129–137.
- [26] Evan A. Stein et al., “Effect of a monoclonal antibody to PCSK9 on LDL cholesterol,” *New England Journal of Medicine*, Volume **366** Number 12 (March 22, 2012), pages 1108–1118.
- [27] TG and HDL Working Group of the Exome Sequencing Project, National Heart, Lung, and Blood Institute, “Loss-of-function mutations in APOC3, triglycerides, and coronary disease,” *New England Journal of Medicine*, Volume **371** Number 1 (July 3, 2014), pages 22–31.
- [28] Stephen L. Wiesenfeld, “Sickle-cell trait in human biological and cultural evolution,” *Science*, Volume **157** Number 3793 (September 8, 1967), pages 1134–1140.
- [29] Peter H. Wood, *Black Majority: Negroes in Colonial South Carolina from 1670 through the Stono Rebellion*, Norton, New York NY, 1974.
- [30] World Health Organization (WHO), World Malaria Report 2014, available at http://www.who.int/malaria/publications/world_malaria_report_2014/report/en/, accessed on January 28, 2017.