Open Access



Journal of Scientific Insights
E-ISSN: 3062-8571
DOI: https://doi.org/10.69930/jsi.v3i6.1003
Research Article

ScienceTech Group

Vol. 3 (6), 2026
Page: 777-793

Mathematical Modeling of the Effect of Inbreeding on Homozygosity of Recessive Disease-Associated Alleles: A Case Study of Afghanistan

Ghulam Hassan Danishyar [1], Sebghatullah Rahimi [2,*], Ezzatullah Hasanyar [3]

[1] Department of Mathematics &Physic, Faculty of Engineering, Balkh University, Afghanistan
[2] Department of Mathematics, Faculty of Education, Nimroz University , Afghanistan
[3] Department of Mathematics, Faculty of Education, Kunduz University, Kunduz, Afghanistan
*Email (corresponding author): sebghatullahsaaed@gmail.com

Abstract. *Consanguineous marriage can increase the probability of offspring homozygosity for recessive disease-associated alleles. However, parental relatedness, genetic homozygosity, and clinical disease are distinct concepts. This study develops a probabilistic framework to examine their relationships while distinguishing modeled outcomes from individual clinical risk. The single-locus model defines the probability of homozygosity for a specified recessive allele as $H(q,F)=q2+Fq(1-q)$, where (q) is the population frequency of the recessive allele and (F) is the inbreeding coefficient. A weighted model-based index is defined as $R(q,F,\pi)=\pi\ H(q,F)$, where , with $0\leq\pi\leq1$), is a simplified weighting parameter representing penetrance under the model assumptions. The case study draws on published cross-sectional data from 7,140 couples in eight Afghan provinces, where consanguineous marriages accounted for 46.2% of the sample and the reported mean inbreeding coefficient was F=0.0277. For illustrative Monte Carlo analysis, the parameters were assigned the distributions ($q\sim U(0.010,0.020)$), ($F\sim U(0,0.125)$), and ($\pi\sim\beta(2,2)$), with 100,000 iterations. Sensitivity analysis used 200,000 sample pairs. The simulated mean of (R) was approximately (5.76×10-4). Analytically, (R) increases with (F) when (q) and (π) are held constant and ($0<q<1$), ($\pi>0$). These findings describe the behavior of a mathematical model under specified assumptions; (R) is not an estimate of observed disease prevalence or individual clinical risk in Afghanistan.*

Keywords: *Inbreeding Coefficient; Homozygosis; Recessive Alleles; Monte Carlo Simulation; Afghanistan*

1. Introduction

The relationship between parental consanguinity and the genetic outcomes of offspring involves several distinct stages. Parental relatedness and pedigree history determine the expected inbreeding coefficient of the offspring, which influences the probability of homozygosity at a particular genetic locus. When the allele is associated with a recessive disorder, however, the relationship between genotype and phenotype depends on disease-specific characteristics, variant type, penetrance, and environmental factors (Hajnal, 1963). Therefore, genetic homozygosity should not, by itself, be equated with a clinical diagnosis or the probability that an individual will develop a disease.

The inbreeding coefficient has a classical foundation in population genetics and describes the probability that two alleles in an individual are identical by descent (Wright, 1921). In autosomal recessive disorders, homozygosity for a disease-associated allele may have biological significance, but its interpretation depends on the inheritance pattern, allelic variation, and genotype–phenotype relationship. Previous studies of consanguineous

Received: 17 October 2026/ Accepted: 30 October 2026 / Available online: 3 November 2026

marriages have emphasized the importance of distinguishing marriage patterns, genetic structure, and clinical outcomes (Bittles, 2001; Bittles & Black, 2010). In Afghanistan, (Saify & Saadat, 2012) conducted a cross-sectional study involving couples from eight provinces: Badakhshan, Baghlan, Balkh, Bamyan, Kabul, Kunduz, Samangan, and Takhar. They reported a consanguineous-marriage rate of 46.2% and a mean inbreeding coefficient of 0.0277. Regional and ethnic differences were also observed, indicating that the overall sample mean may not represent the variation within the country. These findings are used here as historical demographic context, not as direct evidence of the frequency of a particular disease-associated allele or the prevalence of a clinical disorder (Saadat & Tajbakhsh, 2013). Several methodological distinctions are essential for modeling these relationships. Parental relatedness describes genealogical relationships, whereas the inbreeding coefficient is a genetic quantity derived from pedigree information)Hamamy, 2012). Similarly, the probability of homozygosity for a specified recessive allele is not equivalent to disease risk, which requires information about pathogenicity, penetrance, and the genotype–phenotype relationship. When uncertainty in model parameters is represented by probability distributions, summary statistics and sensitivity measures must be calculated using reproducible methods consistent with the specified assumptions (Modell & Darr, 2002).

The present study also distinguishes mathematical modeling from the statistical analysis of observational data. Measures such as odds ratios and risk ratios are estimated from observational data and depend on the study design, comparison groups, outcome definitions, potential confounding factors, and estimation methods. By contrast, the model developed here uses genetic parameters and mathematical assumptions to calculate the probability of homozygosity under specified scenarios. Accordingly, the modeled index R is not an observational odds ratio and cannot be estimated solely from data on consanguineous marriages (Charlesworth & Willis, 2009). Nevertheless, observational evidence and mathematical modeling can complement each other: empirical data may support model evaluation and parameter calibration, while mathematical analysis helps clarify assumptions and examine the consequences of parameter changes (Nussbaum et al., 2016). The primary aim of this study is to establish a mathematical framework linking allele frequency, the inbreeding coefficient, and the probability of homozygosity, and to examine the behavior of the resulting index under specified numerical assumptions. The model parameters are defined separately, and their uncertainty is represented through explicit scenario distributions. Monte Carlo simulation with a fixed random seed, numerical tables, and an R script is used to support reproducibility (Bittles & Neel, 1994). No simulated value is presented as an actual observation or as a definitive estimate of disease risk in Afghanistan. Converting the modeled index R into a disease-probability estimate would require reliable data on allele or carrier frequencies, genotype-specific penetrance, and clinical outcomes. In the absence of such evidence, R is interpreted exclusively as a modeled index rather than an estimate of clinical risk.

In relation to Sustainable Development Goal 3 (SDG 3: Good Health and Well-being), this study provides a mathematical framework for exploring how inbreeding may influence the homozygosity of recessive disease-associated alleles. Although the model does not directly estimate disease prevalence or individual clinical risk, it may contribute to the theoretical understanding of genetic factors relevant to public health research, particularly in populations where consanguineous marriage is common.

2. Methods

2.1 Mathematical Model

2.1.1. First Layer: Kinship Relationship and Inbreeding Coefficient

The symbol C denotes the type of kinship relationship between the parents, and F(C) represents the expected inbreeding coefficient of their offspring under that relationship(as shown in Table 1). For standard relationships, assuming no additional ancestral paths, the commonly used theoretical values are as follows: unrelated parents,F = 0; first cousins,F = 1/16; second cousins,F = 1/64; and double first cousins, F = 1/8. In this article, the value F = 1/32 for the first-cousins-once-removed relationship is also included as a theoretical value in the scenario table (Wright, 1921; Hartl & Clark, 2007). For any actual pedigree, the exact value must be calculated based on the ancestral paths and the contribution of each path. Therefore, this table is not a substitute for pedigree analysis.

Table 1. Theoretical Inbreeding Coefficient Values for Standard Kinship Relationships

Kinship relationship	F	Decimal value
Unrelated parents	0	0.000000
Second cousins	1/64	0.015625
First cousins once removed	1/32	0.031250
First cousins	1/16	0.062500
Double first cousins	1/8	0.125000
Mean reported in the Afghan study	0.0277	0.027700

The mean value of 0.0277 was obtained from a published population study and does not represent a specific type of kinship relationship. Rather, it is the weighted mean reflecting the composition of the observed relationships in the study sample. Consequently, comparing this value with the theoretical values above is useful for interpreting the parameter range, but it should not be attributed to any particular kinship relationship.

2.1.2. Second Layer: Probability of Homozygosity at a Single Locus

Let **(a)** denote the recessive allele of interest, and let **(q)** be its population frequency, where **(0≤ q≤1)**. Let **(F)** denote the inbreeding coefficient, with **(0≤ F≤1)**. Under the single-locus biallelic model, the probability of homozygosity for genotype **(aa)** is defined as

$$H\,(q,F) = q^2 + Fq\,(1-q) \qquad (1)$$

Here, (q2) is the baseline homozygosity probability under Hardy–Weinberg equilibrium, while Fq(1-q) represents the additional contribution associated with inbreeding under the model. The formula is used here as a theoretical single-locus model; applying it to a specific population or disease requires an appropriate allele-frequency estimate and consideration of population structure and the relevant genetic assumptions. Differentiating with respect to *F* gives:

$$\frac{\partial H(q,F)}{\partial F} = q(1 - q) \qquad (2)$$

CC BY SA

For $0 < q < 1$, this derivative is positive. Therefore, with q held constant, an increase in F leads to a linear increase in the modeled probability of homozygosity. The second derivative with respect to F is zero, confirming that the relationship is linear in F. This result describes the mathematical structure of the model and does not, by itself, imply that the actual prevalence of the disease in a population changes by the same proportion.

2.1.3. Third Layer: Penetrance and the Modeled Index

To distinguish homozygosity from a disease-related outcome, we introduce) as a simplified weighting parameter representing penetrance under the assumptions of the model, where **(0≤π≤1)**. We define the modeled index as

$$R(q,\ F,\ \pi) = \ \pi H(q,\ F) = \ \pi\left[q^2 + \ Fq(1\ -\ q)\right] \qquad (3)$$

Here, H(q,F) represents the modeled probability of homozygosity for the specified recessive allele, whereas R(q,F,) is the corresponding weighted quantity. Interpreting (R) as a disease probability would require additional assumptions, including that the modeled genotype is the relevant disease-causing genotype and that) appropriately represents its penetrance. In this study, (R) is interpreted as a model-based index and is not treated as an estimate of observed disease prevalence or individual clinical risk in Afghanistan.

Proposition 1. (Monotonicity with Respect to the Inbreeding Coefficient**).** If $0 < q < 1$ and $\pi > 0$, then R(q, F, π) is increasing and linear with respect to F.

Proof. It follows from Equation 3,

$$\frac{\partial R}{\partial F} \ = \ \pi q(1\ -\ q) > \ 0, \qquad \frac{\partial^2 R}{\partial F^2} \ = \ 0.$$

Therefore, the proposition holds.

The derivative with respect to allele frequency is

$$\frac{\partial R}{\partial q} = \ \pi[2q \ + \ F(1\ -\ 2q)] \qquad (4)$$

For the typical ranges of q associated with rare alleles, this derivative is positive; however, its behavior in general must be examined in relation to the parameter domain and the model assumptions. The derivative with respect to is H(q, F), indicating that an increase in π changes the model output proportionally to the probability of homozygosity. Furthermore, for $q > 0$ and $\pi > 0$, comparing the model output at F with its baseline value at $F = 0$ yields the following relative effect of inbreeding:

$$\frac{R(q,F,\pi)}{R(q,0,\pi)} \ = \ 1 \ + \frac{F(1-q)}{q} \qquad (5)$$

This ratio is independent of because the penetrance parameter cancels from the numerator and denominator. For a fixed value of F, the ratio can be larger at lower values of q. However, a large relative increase does not necessarily imply a large absolute value of R, since the baseline value q^2 remains small when the allele is rare. Therefore, both the absolute value of R and its ratio to the baseline value should be reported when presenting the results.

2.1.4. Joint Homozygosity Model

For n genetic loci, let q_i denote the population frequency of the recessive disease-associated allele at locus i, where i = 1, 2, …, n. Let F denote the inbreeding coefficient used in the single-locus model. The probability of homozygosity for the specified allele at locus i is given by:

$$H_i(q_i, F) = q_i^2 + Fq_i(1 - q_i) \quad (6)$$

Assuming that the homozygosity events at the n loci are mutually independent, the probability of simultaneous homozygosity for the specified alleles at all loci is:

$$H_{joint}(F, q) = \prod_{i=1}^{n}\left[q_i^2 + Fq_i(1 - q_i)\right] \quad (7)$$

where q = (q1, q_2, …, q_n). This expression is a theoretical extension of the single-locus model and describes joint homozygosity for the specified alleles, rather than a general model of disease occurrence across multiple genes. The independence assumption requires particular attention. It may be a reasonable approximation when the relevant homozygosity events are independent or when dependence between loci is sufficiently weak for the intended analysis. However, physical linkage between loci can produce statistical dependence between genotypes, particularly when haplotype structure and non-random associations between alleles are present. In addition, population structure, shared ancestry, and other demographic factors may induce dependence between loci. Under these circumstances, the product formula may not provide an accurate estimate of joint homozygosity, and a model incorporating the relevant joint genotype or haplotype probabilities may be required.

A further limitation concerns compound heterozygosity. For some autosomal recessive disorders, an affected individual may carry two different pathogenic variants in the same gene, one on each homologous chromosome. Such an individual is heterozygous at the relevant variant sites but may nevertheless have the recessive disorder. Therefore, the present formulation, which focuses on homozygosity for specified recessive alleles, does not capture all possible disease-causing genotypes. Its application to a particular disorder requires evidence that the modeled homozygous genotype is relevant to the disease mechanism and that clinically important alternative genotypes are either negligible or modeled separately. Accordingly, the joint homozygosity model should be interpreted as a conditional mathematical framework whose applicability depends on the genetic architecture of the condition under investigation.

2.1.5. Alternative Threshold Model

An alternative model is appropriate when the event of interest depends on the number of genetic loci at which an individual is homozygous for the specified recessive allele, rather than requiring simultaneous homozygosity at every locus. For each locus i, define the indicator variable X_i as follows: $X_i = 1$ if the individual is homozygous for the specified allele at locus i; $X_i = 0$ otherwise. Under the single-locus model, the probability that $X_i = 1$ is:

CC BY SA

$$p_i = P\ (X_i = 1) = q_i^2 + Fq_i\ (1 - q_i) \quad (8)$$

If the homozygosity indicators are mutually independent, then X_i follows a Bernoulli distribution with parameter p_i, and the total number of loci exhibiting the specified homozygosity is defined as: $S_n = X_1 + X_2 + \ + X_n$. When the event of interest is defined as homozygosity at no fewer than k loci, the corresponding threshold index is: Rthr(F, q, , k) = P(S_n k) or, equivalently,

$$R_{thr\ (}F, q, \pi, k) = \pi\ P\ (X_1 + X_2 + \ldots + X_n \geq k) \quad (9)$$

Here, k is an integer satisfying $1 \leq k \leq n$, and , with $0 \leq \pi \leq 1$, is the simplified penetrance parameter used in the main framework. If the probabilities p_i differ among loci, the sum S_n follows a Poisson binomial distribution under the independence assumption. Consequently, the threshold probability can be calculated from the specified locus-specific probabilities. The validity of this formulation depends on the independence assumption. Linkage disequilibrium, shared haplotypes, population stratification, or other sources of genetic dependence may cause the indicators X_i to be correlated. In such cases, the Poisson binomial distribution is not generally appropriate, and the threshold probability must instead be calculated using a joint genotype distribution or another model that accounts for the dependence between loci.

Moreover, the number of homozygous loci is not, in general, a sufficient criterion for determining whether an individual has a genetic disorder. Disease expression may depend on the identity of the loci, the specific pathogenic variants, their functional effects, and the inheritance mechanism. In particular, compound heterozygosity may cause an autosomal recessive disorder without homozygosity for the same pathogenic allele. Therefore, the threshold model is most appropriately interpreted as a theoretical index of multiple-locus homozygosity, not as a validated disease-probability model. Applying it to a specific disorder requires biological justification for the threshold k, an appropriate representation of disease-causing genotypes, and disease-specific evidence concerning penetrance. The model is presented here as a theoretical extension; the numerical simulations and principal results of this study continue to rely on the single-locus model defined in Equation (3).

2.1.6. Hardy–Weinberg Equilibrium and Inbreeding

In real populations, deviations from Hardy–Weinberg equilibrium may arise from population structure, non-random mating, migration, selection, sampling error, or other factors. Adding an undefined component to **F** without a clear definition may lead to double-counting these effects. Therefore, when sufficient information is available, pedigree inbreeding, **Fped**, and the population-level component, **Fpop**, can be defined separately:

$$P(aa) = \ q^2 + \ \left(F_{ped} + \ F_{pop}\right)q(1 \ - \ q) \quad (10)$$

This relationship is applicable only when the two components can be distinguished based on the study design and available data. If they overlap conceptually or the necessary information is unavailable, combining them without independent estimation is not

recommended (Hardy, 1908; Hartl & Clark, 2007). The present numerical analysis does not estimate Fpop separately but uses F as an overall scenario parameter. This is an explicit limitation of the model.

3. Results and Discussion

3.1. Consanguinity Data from Afghanistan

The Afghanistan case study uses published historical survey findings as contextual population data. reported that 46.2% of 7,140 couples surveyed across eight Afghan provinces were in consanguineous marriages, with a mean inbreeding coefficient of (F=0.0277). These findings describe the surveyed sample and should not be interpreted as current national estimates)Saify & Saadat 2012). The allele-frequency range (q~ U(0.010,0.020)) and the penetrance-related weighting parameter π~Beta(2,2)) are hypothetical inputs selected for illustrative uncertainty analysis; they were not estimated from the Afghan survey. Similarly, (F~ U(0,0.125)) is a modeling assumption, whereas the historical value (F=0.0277) is examined separately in a fixed-(F) scenario. Therefore, the simulated results describe model behavior under specified assumptions and do not estimate present-day disease prevalence or national disease burden in Afghanistan(see Table 2).

Table 2. Summary of the Afghan Population Study Data Used in This Article

Indicator	Reported value
Number of couples studied	7,140
Provinces included	Badakhshan, Baghlan, Balkh, Bamyan, Kabul, Kunduz, Samangan, and Takhar
Proportion of consanguineous marriages in the sample	46.2%
Mean inbreeding coefficient	0.0277
Study design	Cross-sectional
Use in the present article	Demographic context, not direct data on allele frequency or clinical disease occurrence

3.2. Beta-Thalassemia and Allele Frequency

Beta-thalassemia is considered here as an example of an autosomal recessive disorder; however, the parameterization used in this article does not constitute an estimate of the actual risk of beta-thalassemia in Afghanistan. Assuming Hardy–Weinberg equilibrium, the carrier frequency c for a rare allele can be approximated by c ≈ 2q. The exact relationship under a biallelic model is given by:

$$c = 2q(1 - q), \qquad q = \frac{(1 - \sqrt{1 - 2c})}{2} \tag{11}$$

As shown in Table 3 For the illustrative scenario c = 0.03, the corresponding allele frequency is q ≈ 0.01523. This value is solely the algebraic result of converting an assumed carrier frequency under Hardy–Weinberg equilibrium. In the absence of a reliable and representative estimate of c for the target population, it should not be reported as the actual allele frequency throughout Afghanistan. Accordingly, rather than fixing q at an asserted value, the main simulation uses the scenario range [0.010, 0.020].

Table 3. Illustrative Conversion of Carrier Frequency to Allele Frequency Under Hardy–Weinberg Equilibrium

Quantity	Value	Interpretation
Assumed carrier frequency c	0.03000	Scenario assumption
Calculated allele frequency q	0.01523	Result of the equation 9
H(q, 0)	Approximately 0.000232	Homozygosity at F = 0
H(q, 1/16)	Approximately 0.00117	Homozygosity under the F = 1/16 scenario

3.3. Parameter Distributions and Simulation Assumptions

For the illustrative uncertainty analysis, the model inputs are assumed to be mutually independent and are assigned the following probability distributions:

- $q \sim U(0.010, 0.020)$
- $F \sim U(0, 0.125)$
- $\pi \sim Beta(2, 2)$

Here, U(a, b) denotes a continuous uniform distribution on the interval [a, b], and Beta(2, 2) is a beta distribution with shape parameters $\alpha = 2$ and $\beta = 2$. Thus, q is sampled between 0.010 and 0.020, F between 0 and 0.125, and between 0 and 1 according to the Beta(2, 2) density. These distributions are illustrative assumptions, not empirically estimated distributions for Afghanistan. The reported Afghan mean inbreeding coefficient, F = 0.0277, is used as a separate contextual value in the fixed-F scenario and is not the upper or mean bound of the uniform distribution above.

In the Monte Carlo implementation, independent values q_i, F_i, and $_i$ are generated from these distributions at each iteration i, and the model output is calculated as $R_i = \pi_i[q_i^2 + F_i q_i(1 - q_i)]$. The baseline simulation uses N = 100,000 iterations. The summary statistics reported in Table 6 are calculated from the resulting simulated R values and therefore describe the model output under these specified assumptions; they are not empirical estimates of disease prevalence or individual clinical risk.

3.4. Uncertainty Analysis Method

3.4.1. Monte Carlo Simulation

For each iteration i, the values of qi, F_i, and $_i$ are generated from the distributions defined in Equations (12)–(14), and the corresponding value of R_i is then calculated:

$$R_i = \pi_i\left[q_i^2 + F_i q_i(1 - q_i)\right],\ i = 1, \ldots, N$$

In the baseline implementation, N = 100,000. To reduce dependence on different versions of R and different implementations of random number generation, the script accompanying the paper uses the Park–Miller linear congruential generator with a fixed seed and inverse cumulative distribution function sampling for Beta(2, 2). This choice facilitates the exact reproduction of the generated samples and does not require any additional packages(Peng, 2011). To describe the output, the mean, sample standard deviation, median, 2.5th and 97.5th percentiles, and Monte Carlo standard error of the mean are reported. The

interval between the 2.5th and 97.5th percentiles represents the simulated uncertainty interval for the output and is not equivalent to a statistical confidence interval for a population parameter. The standard error of the mean is calculated as sR/N and indicates only the numerical precision of the Monte Carlo estimate of the mean; it does not measure the biological uncertainty of the model.

3.4.2. Fixed Scenarios for the Inbreeding Coefficient

To distinguish the effect of F from its variation under the baseline distribution, three fixed scenarios are examined: F = 0, F = 1/16, and F = 1/8. In this comparison, q and are sampled from the same baseline scenario distributions, while only F is held constant. In addition, F = 0.0277 is examined as a scenario corresponding to the mean reported in the Afghan study. This comparison does not imply that the historical mean can be directly applied to every couple or every generation.

3.4.3. Global Sensitivity Analysis and Sobol Indices

Sensitivity analysis examines which inputs contribute most to the variability of the output. The first-order index S_i represents the proportion of output variance attributable to the independent effect of input X_i, whereas the total-order index STi includes both the main effect of that input and its interactions with the other inputs. Conceptually:
First-order index:

$$S_i = \frac{Var_{X_i}\left(E\left[X_i\right]\right)}{Var(R)}, \qquad S_{Ti} = 1 - \frac{Var_{X_{\sim i}}\left(E\left[X_{\sim i}\right]\right)}{Var(R)}, \tag{12}$$

For the numerical calculations, the pick-freeze sampling method is used with two independent matrices, A and B. For each input, a hybrid matrix is constructed in which the corresponding column is taken from B and all other columns are taken from A. The first-order and total-order indices are calculated using standard estimators based on these samples. In the script, M = 200,000 is specified as the number of rows in each matrix. The estimated indices depend on the sampling design and Monte Carlo error(Saltelli et al., 2008). Therefore, small differences between the indices, or deviations of the sum of the first-order indices from one, particularly in numerical estimation, should be interpreted in light of sampling error.

3.5. Analytical Results, Numerical Results, and Discussion

3.5.1. Kinship Relationships: Analytical Results

Substituting F(C) into Equation (3) gives:

$$R(C,\ q,\ \pi) = \pi\left[q^2 + F(C)q(1 - q)\right] \tag{13}$$

If $0 < q < 1$ and $\pi > 0$, then whenever $F(C_1) < F(C_2)$, we have $R(C_1, q,) < R(C_2, q,)$. This ordering is a direct consequence of the model. In practical applications, however, if kinship relationships are associated with other differences, such as population structure or allele frequencies, a direct comparison between two groups requires appropriate data and a more suitable model(see Table 4).

For a simple numerical comparison, q = 0.015 and $\pi = 0.5$ are held constant. Table 4 presents the values of H, the index R, and the ratio relative to the case F = 0. This table

CC BY SA

represents an analytical calculation at a specified parameter point and should not be interpreted as an estimate of disease risk in an actual population.

Table 4. Numerical results of the model for standard kinship relationships with **q = 0.015** and **π = 0.5**

Relationship/Scenario	F	H(q, F)	R(q, F, π)	Ratio relative to F = 0
Unrelated parents	0.000000	0.00022500	0.00011250	1.0000
Second cousins	0.015625	0.00045586	0.00022793	2.0260
First cousins once removed	0.031250	0.00068672	0.00034336	3.0521
First cousins	0.062500	0.00114844	0.00057422	5.1042
Double first cousins	0.125000	0.00207188	0.00103594	9.2083
Mean reported in the Afghan study	0.027700	0.00063427	0.00031713	2.8190

The table 4 shows that, at this parameter point, increasing F increases the output. For example, whenF = 1/16, the index R is approximately 5.10 times its baseline value at F = 0; when F = 1/8, this ratio is approximately 9.21. These ratios depend on q and therefore change with the allele frequency. Furthermore, the small value of R in this example demonstrates that even a large ratio relative to the baseline does not necessarily produce a large absolute value.

3.5.2. Afghan Case Study Scenario

The reported mean **F = 0.0277** in the Afghan study lies between certain theoretical values associated with kinship relationships. Under the scenario **q = 0.015** and **π = 0.5**, direct substitution of this value yields **R ≈ 0.00031713**. However, because the allele frequency and penetrance are assumed in this example,(see Table 5) the output is only a conditional result of the model. The source study provides data on types of marriage and the inbreeding coefficient but does not determine the parameters q and π for any specific disease.

Table 5. Distinguishing published data from scenario assumptions in the case study

Quantity	Status in this paper	Basis for interpretation
Number of couples: 7,140	Published data	Cross-sectional study by Saify and Saadat
Proportion of consanguineous marriages: 46.2%	Published data	Characteristic of the source study sample
Mean F = 0.0277	Published data	Sample mean reported in the same study
q ~ U(0.010, 0.020)	Scenario assumption	Used for uncertainty analysis, not an estimate for Afghanistan
π ~ Beta(2, 2)	Scenario assumption	Simplified penetrance distribution
F ~ U(0, 0.125)	Scenario assumption	Computational range covering standard kinship relationships
R and its percentiles	Model output	Dependent on the assumed distributions and Equation 3

3.5.3. Monte Carlo Simulation Results

The results in Table 6 were obtained from **N = 100,000** iterations of the baseline scenario. The mean of **R** is approximately **0.00057555**, and its sample standard deviation is approximately **0.00041899**. The **2.5th** and **97.5th** percentiles are approximately **0.00005868** and **0.00162028**, respectively. As shown in Table 6 The interval between these percentiles summarizes the dispersion of the model output under the selected distributions. Because the distributions are hypothetical, this interval should not be interpreted as a disease-risk interval for the Afghan population or as a confidence interval for an actual parameter.

Table 6. Summary of the Monte Carlo simulation for the baseline scenario (N = 100,000)

Statistic	Simulated value
Mean of R	0.00057555
Sample standard deviation of R	0.00041899
2.5th percentile	0.00005868
Median	0.00047400
97.5th percentile	0.00162028
Monte Carlo standard error of the mean	0.000001325
Number of iterations	100,000
Random number generator seed	20261010

The Monte Carlo standard error of the mean in the table is much smaller than the standard deviation of the output because the mean is calculated from a large number of samples. Nevertheless, this standard error reflects only the numerical error associated with Monte Carlo sampling. It does not account for uncertainty arising from incorrect assumptions, the lack of representativeness of the source data, population variation, or the unknown allele frequency.

3.5.4. Comparison of Fixed-F Scenarios

To compare the effect of the inbreeding coefficient, the same samples of q and are used in every row of Table 7, while only F is fixed at the specified value. Thus, the differences in the outputs are primarily attributable to changes in F in the model, avoiding simultaneous changes in the sampled input values.

Table 7. Simulation comparison for fixed values of the inbreeding coefficient

Scenario	F	Mean	Standard deviation	2.5th percentile	97.5th percentile	Standard error of the mean
No inbreeding	0.000000	0.00011651	0.00007067	0.00001754	0.00028385	0.0000002235
F = 1/64	0.015625	0.00023166	0.00012606	0.00003898	0.00051744	0.0000003986
F = 1/32	0.031250	0.00034681	0.00018236	0.00005962	0.00075213	0.0000005767
F = 1/16	0.062500	0.00057710	0.00029562	0.00010127	0.00122190	0.0000009348
F = 1/8	0.125000	0.00103770	0.00052277	0.00018432	0.00216208	0.000001653
Mean reported in	0.027700	0.00032064	0.00016953	0.00005497	0.00069866	0.0000005361

Comparison of the means shows that, under the assumed model, the mean of R at F = 1/8 is greater than its mean at F = 1/16. The ratio of the two simulated means is approximately 1.80. This ratio is not necessarily identical to the ratio calculated at a fixed value of q, because q and π vary in the simulation and their distributions are taken into account when calculating the means. The methodological point is that every ratio should be reported together with a precise definition of the scenario and statistic.

3.5.5. Sensitivity Analysis

According to the Table 8 the sensitivity analysis was conducted using a pick-freeze estimator with M = 200,000 sample pairs. The first-order Sobol index (S_i) quantifies the contribution of an input parameter to the variance of the model output considered independently of the other inputs, whereas the total-order Sobol index (STi) accounts for that parameter's contribution, including interactions with the other inputs. The three model inputs are the allele frequency (q), the penetrance parameter (π), and the inbreeding coefficient (F). The resulting estimates are reported in Table 8.

Table 8. First-order and total-order Sobol sensitivity indices for the model inputs (M = 200,000 sample pairs).

Model input	First-order index (S_i)	Total-order index (S_{ti})
Allele frequency (q)	0.0992	0.1339
Penetrance parameter (π)	0.3791	0.4836
Inbreeding coefficient (F)	0.4056	0.5023

The first-order indices sum to approximately 0.884, whereas the total-order indices sum to approximately 1.120. The first-order indices need not sum to one when interaction effects are present; likewise, total-order indices can sum to more than one because interaction contributions may be included in the total-order index of more than one input. These sums alone, however, do not establish the magnitude of interaction effects or confirm numerical convergence. The indices should therefore be interpreted conditionally on the independent input distributions specified in Section 3.3, rather than as universal measures of biological importance.

Numerical-stability check: To assess numerical stability, the pick-freeze calculation should be repeated at increasing sample sizes (for example, M = 50,000, 100,000, and 200,000), using the same random-number generator, estimator, input distributions, and variance-estimation procedure as in the reported analysis. The estimates at successive sample sizes should then be compared to determine whether the index values and the ranking of the inputs remain sufficiently stable. Numerical-stability results should be reported only after this comparison has been performed.

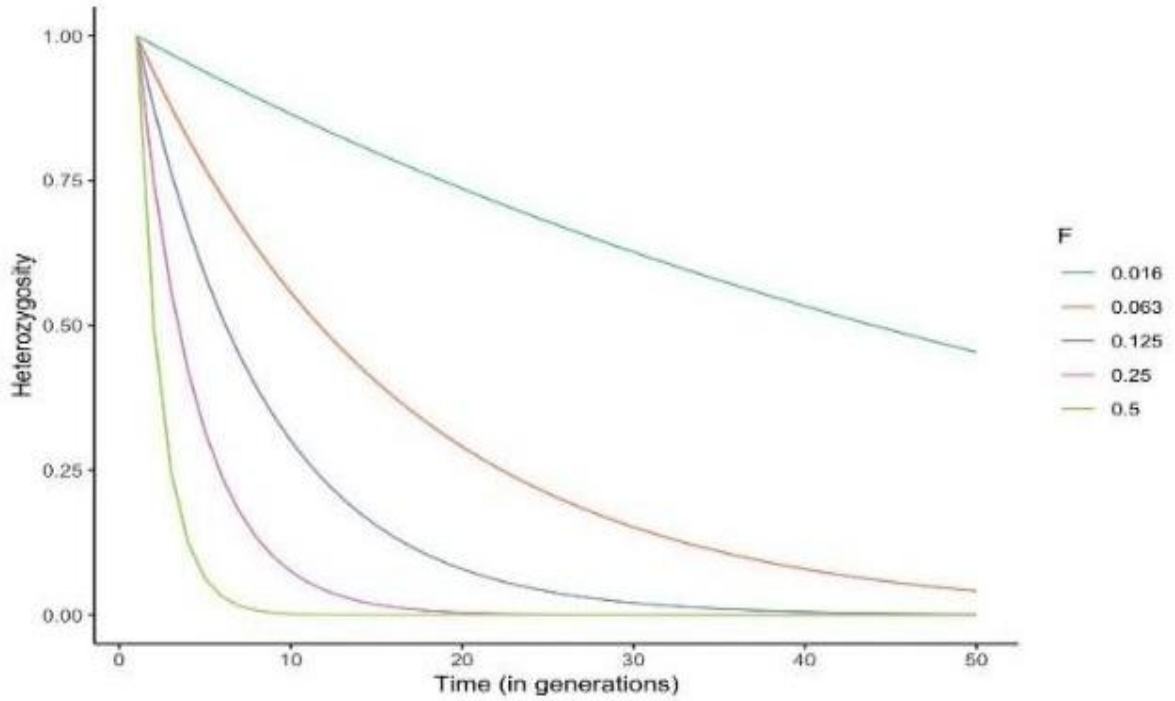


Figure 1. Effect of inbreeding coefficient on R.

This figure shows that, with q and held constant, R increases linearly as F increases. The values q=0.015 and ɪɪ=0.5 can be used to ensure that the plot is directly consistent with Table 4.

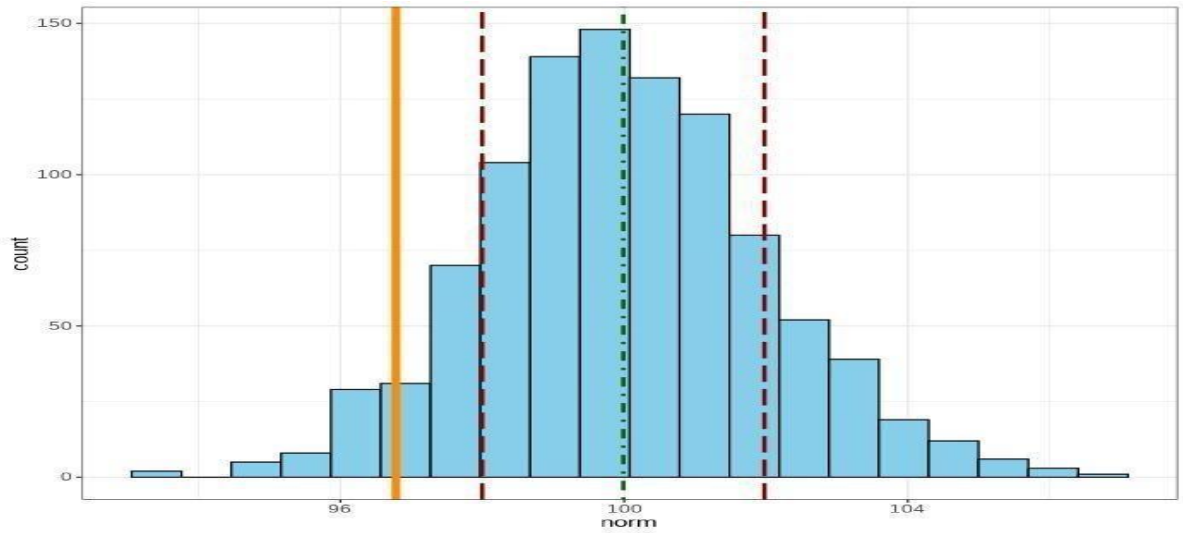


Figure 2. Simulated R across scenarios.

The histogram or density plot shows the distribution of the simulated values of R. The mean and the 2.5th and 97.5th percentiles can also be indicated by vertical lines. This figure helps readers better understand the dispersion and potential asymmetry of the output distribution than they could from the numerical values in the table alone.

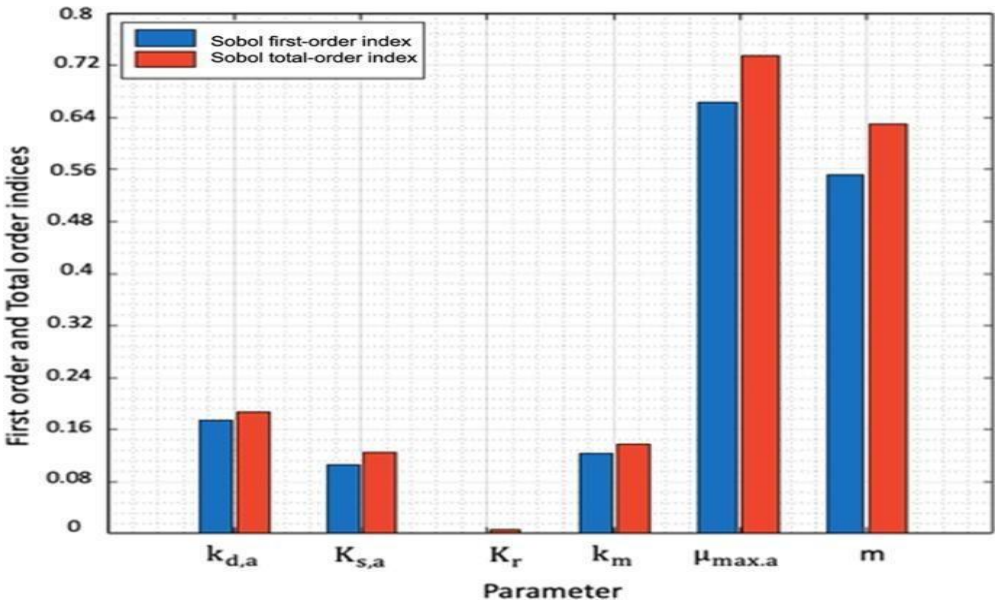


Figure 3. Sensitivity analysis of model parameters.

For each parameter, q, , and F, two bars are plotted, representing the first-order index and the total-order index.

3.6. Discussion

The main analytical result is that, in the single-locus model, holding the allele frequency and penetrance parameter constant, an increase in F raises the probability of homozygosity for the recessive allele and, consequently, the index R. This result follows from Equation (1) and its derivative and does not depend on simulation. Monte Carlo simulation serves a different purpose: it shows the mean and dispersion of the model output when the parameters vary within the assumed ranges.

The main advantage of the proposed framework is that it distinguishes four concepts: kinship relationship, inbreeding coefficient, homozygous genotype, and a penetrance-related index. This distinction prevents the proportion of consanguineous marriages in a population study from being interpreted directly as disease prevalence. Accordingly, the Afghan data are useful for providing context for the parameter F, but they are insufficient to estimate q and . The joint homozygosity model and the threshold model provide opportunities for theoretical extensions to multiple loci, but they require additional assumptions about locus independence and the relationship between genotypes and phenotypes. In actual diseases, multiple variants within a gene, compound heterozygosity, genetic linkage, and variable penetrance may play important roles. Therefore, extending the model requires disease-specific data, and adding parameters without supporting data should not be mistaken for greater scientific validity merely because it increases the apparent complexity of the model.

Conclusions

This study presents a mathematical framework for examining the relationship between inbreeding and the homozygosity of recessive disease-associated alleles, using Afghanistan as a contextual case study. The analytical result establishes that, for fixed allele frequency q and weighting parameter , the modeled index R(q,F,π)=π[q2+Fq(1-q)] is non-decreasing in the inbreeding coefficient F. Specifically, its derivative with respect to F is π q(1-

q), which is positive under the usual conditions $0<q<1$ and $\pi>0$. This monotonic relationship follows directly from the mathematical structure of the model and does not depend on the simulation results.

The numerical findings, including the simulated distribution of R, its summary statistics, and the sensitivity-analysis indices, depend on the parameter ranges, probability distributions, and computational procedures specified in this study. They should therefore be interpreted as results under the stated modeling assumptions rather than as universal empirical estimates. Their numerical reliability also depends on adequate computational validation and stability checks. Importantly, R is a model-based index of the weighted homozygosity of a recessive disease-associated allele. It is not an estimate of disease prevalence, an individual's probability of developing a disease, or a directly observed clinical risk in Afghanistan. The findings provide a theoretical basis for understanding how inbreeding may influence homozygosity under specified assumptions. Further studies using representative population data, disease-specific genetic information, and appropriate clinical validation would be required before drawing conclusions about actual disease burden or clinical risk.

References

Bennett, R. L., Malleda, N. R., Byers, P. H., Steiner, R. D., & Barr, K. M. (2021). Genetic counseling and screening of consanguineous couples and their offspring practice resource: Focused revision. Journal of Genetic Counseling, 30(5), 1354–1357. https://doi.org/10.1002/jgc4.1477

Bittles, A. H. (2001). Consanguinity and its relevance to clinical genetics. Clinical Genetics, 60(2), 89–98. https://doi.org/10.1034/j.1399-0004.2001.600201.x

Bittles, A. H., & Black, M. L. (2010). Consanguinity, human evolution, and complex diseases. Proceedings of the National Academy of Sciences, 107(Suppl. 1), 1779–1786. https://doi.org/10.1073/pnas.0906079106

Charlesworth, D., & Willis, J. H. (2009). The genetics of inbreeding depression. Nature Reviews Genetics, 10, 783–796. https://doi.org/10.1038/nrg2664

Hajnal, J. (1963). Concepts of random mating and the frequency of consanguineous marriages. Proceedings of the Royal Society of London. Series B, Biological Sciences, 159(974), 125–177. https://doi.org/10.1098/rspb.1963.0071

Hamamy, H. (2012). Consanguineous marriages: Preconception consultation in primary health care settings. Journal of Community Genetics, 3(3), 185–192. https://doi.org/10.1007/s12687-011-0072-y

Hardy, G. H. (1908). Mendelian proportions in a mixed population. Science, 28(706), 49–50. https://doi.org/10.1126/science.28.706.49

Hartl, D. L., & Clark, A. G. (2007). Principles of population genetics (4th ed.). Sinauer Associates.

Modell, B., & Darr, A. (2002). Genetic counselling and customary consanguineous marriage. Nature Reviews Genetics, 3, 225–229. https://doi.org/10.1038/nrg754

Nussbaum, R. L., McInnes, R. R., & Willard, H. F. (2016). Thompson & Thompson genetics in medicine (8th ed.). Elsevier.

Peng, R. D. (2011). Reproducible research in computational science. Science, 334(6060), 1226–1227. https://doi.org/10.1126/science.1213847

Saadat, M., & Tajbakhsh, K. (2013). Prevalence of consanguineous marriages in west and south of Afghanistan. Journal of Biosocial Science, 45(6), 799–805. https://doi.org/10.1017/S0021932012000661

Saify, K., & Saadat, M. (2012). Consanguineous marriages in Afghanistan. Journal of Biosocial Science, 44(1), 73–81. https://doi.org/10.1017/S0021932011000253

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., & Tarantola, S. (2008). Global sensitivity analysis: The primer. Wiley. https://doi.org/10.1002/9780470725184

Sobol', I. M. (2001). Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. Mathematics and Computers in Simulation, 55(1–3), 271–280. https://doi.org/10.1016/S0378-4754(00)00270-6

Wright, S. (1921). Systems of mating. I. The biometric relations between parent and offspring. Genetics, 6(2), 111–123. https://doi.org/10.1093/genetics/6.2.111

CC BY-SA 4.0 (Attribution-ShareAlike 4.0 International).
This license allows users to share and adapt an article, even commercially, as long as appropriate credit is given and the distribution of derivative works is under the same license as the original. That is, this license lets others copy, distribute, modify and reproduce the Article, provided the original source and Authors are credited under the same license as the original.