Steve Selvin,[1] *Ph.D.; D. M. Black,*[1] *M.A.;*
B. W. Grunbaum,[1] *Ph.D., M.Crim.; and Nello Pace,*[1] *Ph.D.*

Racial Classifications Based on Blood Group Protein Systems

A large series of blood group protein systems has been analyzed as part of a project funded by the California Office of Criminal Justice Planning. These blood group protein systems potentially form the basis for establishing useful probabilities associated with the occurrence of particular sets of phenotypes. Two facts must be established concerning these systems in order to calculate most types of probabilities. First, the proportion of each phenotype must be accurately estimated for the population of interest. Second, statistical independence must be established among all systems considered.

The phenotype frequencies for eleven blood group protein systems (Table 1) have been estimated from the data collected by Grunbaum et al [*1*] for each of four ethnic groups. These determinations were performed for both men and women, but interest for this paper is focused on men (5043 individuals). A parallel description for females would be redundant since the female phenotypic frequencies are essentially the same as males for all systems except the sex-linked glucose-6-phosphate dehydrogenase (G-6-PD) system.

Data

The eleven genetic systems show no statistical association. The largest product-moment correlation coefficient between any two systems for the 55 paired combinations is 0.053 (see also Ref *1*). This degree of association is observed between the ABO system and the 6-phosphogluconate dehydrogenase system. Statistical independence between all systems is expected since the sources of association among human phenotypes are in general either nonexistent or too small to measure. That is, such sources of dependence as inbreeding, selection, mutation, and so forth, if they exist, will not act as a detectable level in most samples from human populations.

Once the criterion of independence is met, two types of probabilistic calculations are possible. They are the probability that two randomly selected individuals match with respect to a set of phenotypes, and the probability of matching a set of predetermined phenotypes when a single individual is randomly selected from some population. For purposes of clearly defining these two probabilities, a hypothetical situation is created where two blood group systems, A and B, are considered. Furthermore, each of the two

The preparation of these materials was financially assisted through a federal grant from the Law Enforcement Assistance Administration and the California Office of Criminal Justice Planning and under Title I of the Crime Control Act of 1973. The opinions, findings, and conclusions in this publication are those of the authors and not necessarily those of OCJP or LEAA, which reserve a royalty-free, nonexclusive, and irrevocable license to reproduce, publish, and use these materials, and to authorize others to do so. A copy of these materials may be obtained from OCJP or LEAA upon payment of the cost for reproducing the materials. Received for publication 25 May 1978; revised manuscript received 17 Aug. 1978; accepted for publication 23 Aug. 1978.

[1] Associate professor of biostatistics, School of Public Health; research assistant, White Mountain Research Station; research biochemist, White Mountain Research Station; and professor of physiology, emeritus, respectively; University of California, Berkeley.

TABLE 1—*The eleven blood protein systems employed to discriminate among four ethnic groups.*

Blood Protein Systems	Abbreviation
Autosomal	
ABO system	ABO
Rhesus system	Rh
Phosphoglucomutase	PGM
Adenylate kinase	AK
Adenosine deaminase	ADA
Erythrocyte acid phosphatase	EAP
Esterase D	EsD
Hemoglobin	Hb
Blood specific component	Gc
6-Phosphogluconate dehydrogenase	PGD
Sex-linked	
Glucose-6-phosphate dehydrogenase	G-6-PD

systems is assumed to have two phenotypes (1 and 2) with the frequencies given in Table 2. With respect to these two blood groups, every individual in this population has total phenotype of either A1B1, A1B2, A2B1, or A2B2.

Matching of Two Randomly Chosen Individuals

The probability that individuals match with respect to a series of independent characteristics was first discussed by Simpson [2] in a general context, later by Fisher [3] for blood groups, and more recently by Jones [4]. The probability that two randomly chosen individuals match with respect to a series of blood group phenotypes will be represented by P(match). The quantity [1 − P(match)] is the probability that two randomly chosen individuals are not identical for a set of phenotypes and is sometimes called the probability of discrimination [3].

In general the probability P(match) is calculated by multiplying together a series of probabilities represented by q_j. The quantity q_j is

$$\sum_{i=1}^{s} p_i^2$$

with p_i representing the phenotypic frequency of the ith phenotype from the jth blood group system and s is the number of phenotypes within that system [3]. The probability P(match) is

$$P\,(\text{match}) = q_1 q_2 \ldots q_k$$

with k representing the total number of blood groups considered. As an example, for the hypothetical population

$$q_1 = (0.1)^2 + (0.9)^2 = 0.82$$

$$q_2 = (0.5)^2 + (0.5)^2 = 0.50$$

and

$$P(\text{match}) = q_1 q_2 = (0.82)\,(0.50) = 0.41$$

TABLE 2—*The phenotypic frequencies associated with two genetic systems, A and B, in a hypothetical population.*

	Phenotypes	
Blood Group System	1	2
A	0.1	0.9
B	0.5	0.5

Thus, in the hypothetical population, the probability that two randomly selected individuals match with respect to two blood groups is 0.41 and P(discrimination) $= 1 - 0.41 = 0.59$.

The probability of the random match of two individuals is useful in summarizing the efficacy of blood group protein systems for identifying individuals. The smaller this probability is, the less chance that two randomly chosen individuals match for all genetic systems considered. The probability is minimized when the possible phenotypes within a genetic system are equally likely to occur [*4*]. Therefore, genetic systems in which one phenotype occurs with a very high probability (for example, greater than 0.9) do not contribute much to the reduction of P(match). It should be emphasized that P(match) or P(discrimination) is only a summary indication of how well a set of blood groups will discriminate.

Probability of Matching a Particular Phenotype

The probability that a randomly chosen individual matches a specific predetermined set of phenotypes termed P(coincidence) is of greater direct use. For example, if blood found at the scene of a crime is analyzed for a particular set of phenotypes, it may be of interest to know the probability of finding a match to that blood sample in the general population. The quantity represented by P(coincidence) can be calculated if the frequencies of the individual phenotypes within each blood group are known. Again assuming statistical independence among all systems considered, the probability of coincidence is the product of a series of specific phenotypic frequencies or

$$P(\text{coincidence}) = p_1 p_2 \ldots p_k$$

where p_i is the frequency of the ith given phenotypic genetic variant and k is the number of systems considered. For example, in the hypothetical population defined above, the probability of finding a random individual with a phenotype of A1B1 is simply P(A1) × P(B1) = (0.10) (0.5) = 0.05. The corresponding probabilities for the other three possible phenotypes A1B2, A2B1, and A2B2 are 0.05, 0.45, and 0.45, respectively.

The frequency of each phenotype plays a major role in the magnitude of the probability of coincidence. For example, if each of the ten phenotypes had a probability of occurrence of 0.5, then the probability of coincidence of the set is $(0.5)^{10} = 0.00098$, or about 1 in 1000. However, if ten blood group systems are used and a particular sample is found to have the most common phenotype occurring with a probability of 0.9 in each system, then the total probability of coincidence is $(0.9)^{10} = 0.349$. This probability indicates that this particular set of phenotypes would be of little use for identification of an individual since about 35% of the population match with respect to these ten phenotypes. Intermediate to these two possibilities is the situation where a blood group system contains one or two rare phenotypes. For example, instead of all ten phenotypes occurring with frequency 0.9, suppose two phenotypes were found with frequency 0.1. In this case the probability of coincidence is $(0.9)^8\ (0.1)^2 = 0.004$. Clearly, rare phenotypes are the key to producing small probabilities of coincidence, which implies that when a set of phenotypes containing rare genetic variants occurs it will be useful in the process of identification.

Predicting Ethnic Source from a Blood Sample (Bayes' Theorem)

Calculations for both the P(match) and P(coincidence) are dependent on having population estimates of the phenotypic frequencies. However, if the population under consideration is made up of subgroups that differ with respect to phenotypic frequencies, the population estimates are more complex. For example, phenotypic frequencies vary among ethnic groups and therefore the phenotypic frequencies for the entire population must be estimated by employing a weighted average of the phenotypic frequencies using as weights the proportion of each ethnic group in the population. However, there are subjective elements involved in deciding what proportions to use for each subgroup. Suppose, for example, a blood sample to be matched is found in San Francisco. Should the ethnic proportions of San Francisco, of California, of the United States, or of North America be used to calculate the probability of coincidence? In many cases, the choice of base population will lead to very different results.

However, the probability of coincidence, which can reflect the differences in phenotypic frequencies among ethnic groups, can be directly employed to gain information about ethnic origin of a blood sample. Using Bayes' theorem of conditional probability, the probability that a given blood sample comes from any ethnic group can be estimated. Bayes' theorem states that for two events, called A and B,

$$P(A\,|\,B) = P(B\,|\,A)\,[P(A)/(B)]$$

where $P(A|B)$ represents the probability of Event A given Event B has occurred and $P(B|A)$ is the probability of Event B given Event A. Belonging to a particular ethnic group (for example, white) could be Event A and the possession of a particular set of phenotypes (called phenotype X) could be Event B; then,

$$P(\text{white}\,|\,\text{phenotype}\,X) = P(\text{phenotype}\,X\,|\,\text{white})\;[P(\text{white})/P(\text{phenotype}\,X)]$$

The three probabilities P(phenotype X|white), P(white), and P(phenotype X) can be estimated from a set of data making it possible to calculate P(white|phenotype X) from Bayes' theorem. The term P(phenotype X|white) is P(coincidence) calculated as previously described by using the phenotypic frequencies of whites. The probability that an individual is white can be estimated by the proportion of whites in the total population. The value of P(phenotype X) is the probability an individual possesses phenotype X regardless of racial group and is, as mentioned, a weighted average expressed as P(phenotype X|white)· P(white) + P(phenotype X|black)P(black) . . ., summed for all relevant ethnic subpopulations. Therefore, once the individual phenotypic frequencies in each group are estimated and the proportion of each group in the total population is known, it is possible to calculate P(any ethnic source|phenotype X).

The probability of belonging to a specific racial category given the condition that a particular set of phenotypes has been observed serves as the basis for a system to classify individuals by race. The classification rule is simply to place an individual into the most probable ethnic category. For example, if the probability an individual is white given a set of phenotypes X is 0.8 (that is, P[white|phenotype X] = 0.8) and the probability a person is black given the same set of phenotypes X is 0.9 (that is, P[black|phenotype X] = 0.9), then the individual is classified as black. The process can be applied to individuals who possibly belong to any number of ethnic categories by applying Bayes' theorem to each ethnic group under consideration. The individual is then classified into the most probable racial category. Note that for comparing a particular blood sample for a series of ethnic groups, the denominator P(phenotype X) is the same for every probability and therefore does not need to be calculated for the classification scheme.

Results

The procedure was tested on a sample of 5043 male individuals who had complete sets of phenotypes determined for eleven genetic systems (Table 1). For purposes of the study, the population was considered to be made up of four ethnic groups (white, black, Chicano/Amerindian, Asian) and only those males belonging to one of these four groups were considered. Additionally, the population of California was employed as the base population with the proportion of the four ethnic groups as 75.9% white, 7.0% black, 15.6% Chicano/Amerindian, and 1.5% Asian. Phenotypic frequencies of Grunbaum et al [1] were used. The results of this approach are given in Table 3. The overall rate of misclassification of the 5043 males into four ethnic categories was 21.7%.

Predicting Ethnic Source from a Blood Sample (Discriminant Function)

An alternative method of classification is the linear discriminant function. The linear discriminant function was originated by Fisher [5] and applied to the problem of classifying three species of iris. Fisher shows that the linear discriminant function is optimum in the sense that under certain conditions, the misclassification rate is minimized for a given set of data. Subsequent to the work of Fisher, several measures of classification based on genetic data have been proposed to differentiate between various groupings, each method dealing in a different manner with the problem (for example, see Refs *6* and *7*).

Recently, Spielman and Smouse [8], employing a linear discriminant function, classified Brazilian Indians into their respective villages based on genetic and anthropomorphic measurements. A logical extension of the technique is to attempt to classify individuals into racial categories based on the variation in frequencies among genetic phenotypes found in blood.

In order to use a linear discriminant function, the traits under investigation must be metric. Following the method of Spielman and Smouse and others each phenotype was given an arbitrary score. For example, consider the phosphoglucomutase (PGM) phenotypes where the 1-1 variant was assigned the value 1, the 1-2 variant was assigned the value 2, and the 2-2 variant was assigned the value 3. This coding introduces a certain subjectivity into the analysis that should be recognized.

TABLE 3—*The results of employing eleven blood group protein systems to classify individuals into four ethnic categories by using an application of Bayes' theorem.*[a]

		Predicted							
Observed		White		Black		Chicano/ Amerindian		Asian	
Category	*n*	*n*	%	*n*	%	*n*	%	*n*	%
White	2351	2344	99.7	7	0.3	0	0.0	0	0.0
Black	533	337	63.2	196	36.8	0	0.0	0	0.0
Chicano/Amerindian	758	742	97.9	14	1.8	1	0.1	1	0.1
Asian	1401	1382	98.6	3	0.2	4	0.3	12	0.9

[a] Average correctly classified, 34.38%; probability of misclassification, 0.217.

Results

Table 4 deals with the performance of the linear discriminant function given as

$$\begin{aligned} D_i = {} & (-0.093)\mathrm{ABO}_i + (0.034)\mathrm{Rh}_i + (0.010)\mathrm{PGM}_i + (0.256)\mathrm{AK}_i \\ & + (0.187)\mathrm{ADA}_i + (-0.038)\mathrm{EAP}_i + (0.670)\mathrm{EsD}_i + (-3.844)\mathrm{Hb}_i \\ & + (0.385)\mathrm{Gc}_i + (0.158)\mathrm{PGD}_i + (-2.380)\mathrm{G\text{-}6\text{-}PD}_i + 4.349 \end{aligned}$$

in classifying each of the 5043 blood samples into one of four ethnic groups. The mean discriminant scores for each of the four ethnic groups are whites = 0.127, blacks = −1.729, Chicano/Amerindian = 0.010, and Asians = 0.438. The discriminant scores are statistically adjusted to have a standard deviation equal to 1.0. The only group that shows any appreciable genetic distance as measured by the linear discriminant function is blacks. The difference is primarily influenced by the hemoglobin and G-6-PD systems, as seen by comparing the magnitude of the discriminant function coefficients. The overall population of individuals misclassified by the linear discriminant function is 21.9%.

Discussion

The evaluation of the racial classification employed here rests on the race declared when blood was donated at a series of California and Hawaii blood banks and is certainly subject to possibly severe bias. The data deal with "declared" race and not true race. Therefore, the classification of individuals by race is applied in this relative sense and not as a way of determining a person's true racial identity (if such a state exists).

The classification methods based on Bayes' theorem and the linear discriminant function method produce practically the same results by entirely different routes. The method based on Bayes' theorem is essentially nonparametric (that is, it depends on no assumption about the sampled population). The calculation of the misclassification probabilities associated with the discriminant function depends on the assumption that the discriminant function produces values with a normal distribution. The discriminant scores will be normally distributed when the variables under study have at least approximately normal distributions. This is not the case with phenotypic frequency data. In fact, phenotypic scores are discrete variables with rather skewed distributions in most cases. For the data employed here the scores are certainly not normally distributed, especially when the Hb and the G-6-PD systems are included. The lack of normality introduces an unknown degree of error in the calculation of the misclassification probabilities based on the linear discriminant function.

TABLE 4—*The results of employing eleven blood group protein systems to classify individuals into four ethnic categories by using an application of the linear discriminant function.*[a]

Observed		Predicted							
		White		Black		Chicano/ Amerindian		Asian	
Category	*n*	*n*	%	*n*	%	*n*	%	*n*	%
White	2351	2342	99.7	9	0.4	0	0.0	0	0.0
Black	533	344	64.5	189	35.5	0	0.0	0	0.0
Chicano/Amerindian	758	746	98.4	12	1.6	0	0.0	0	0.0
Asian	1401	1399	99.9	2	0.1	0	0.0	0	0.0

[a] Average correctly classified, 33.8%; probability of misclassification, 0.219.

Both methods of classification applied to the eleven blood group protein systems produce a classification error probability of about 22%. The seemingly low error rate must be considered moderately successful. It should be noted that the California population is 75.9% white. This fact implies that the strategy of declaring all individuals as white without reference to blood phenotypes yields a misclassification probability of 0.241. From this point of view, it is clear that the amount of racial heterögeneity among the four major California ethnic groups is not sufficient for either method to classify individuals into racial categories with a high degree of accuracy based on the eleven blood protein systems considered.

There are, however, certain cases where the use of Bayes' theorem for ethnic classification is more successful. For example, if a phenotype occurs primarily in one ethnic group, a blood sample containing that phenotype might well have a high probability of coming from that particular ethnic group. Sickle-cell trait, although sometimes found in Southern European whites, occurs more commonly in blacks and thus a blood sample containing sickle-cell trait is far more likely to have come from a black individual than from any other racial group. There are other rare phenotypes that are also confined primarily to one ethnic group and can be used in a similar way. When such rare phenotypes do occur, they will greatly reduce the probability of coincidence and, therefore, the error of classification.

An issue that should be explicitly stated with regard to P(match) and P(coincidence) involves the role of phenotypic frequencies for identification rather than racial classification. The probability of coincidence is most useful for identification when rare blood groups are among the set of phenotypes found. These sets of phenotypes occur infrequently. On the other hand, when the phenotypic frequencies for a specific blood group protein system are equal, the probability of discrimination is minimum or P(match) is maximum [3,4]. This fact implies that the most useful systems for identification for all individuals in a population of interest are those with several evenly distributed phenotypes. The more such systems that are examined in a given blood sample, the lower the probability of coincidence, which necessarily increases the likelihood a specific individual will be correctly identified.

Summary

Two methods employing frequencies of blood group phenotypes are assessed as methods of accurately classifying individuals into racial categories. The data used consist of eleven blood group protein systems from 5043 males distributed into four ethnic categories (white, black, Chicano/Amerindian, and Asian). Both methods work equally well and yield a rate of misclassification of about 22%. Also included is a discussion of two probabilistic calculations relevant to employing blood group protein systems in the context of an identification tool.

References

[1] Grunbaum, B. W., Selvin, S., Pace, N., and Black, D. M., "Frequency Distribution and Discrimination Probability of Twelve Protein Genetic Variants in Human Blood as a Function of Race, Sex, and Age," *Journal of Forensic Sciences*, Vol. 23, No. 3, July 1978, pp. 577-587.

[2] Simpson, E. H., "Measure of Diversity," *Nature* (London), Vol. 163, April 1949, p. 688.

[3] Fisher, R. A., "Standard Calculation for Evaluating a Blood-Group System," *Heredity* (London), Vol. 5, 1951, pp. 95-102.

[4] Jones, D. A., "Blood Samples: Probability of Discrimination," *Journal of the Forensic Science Society*, Vol. 12, 1972, pp. 355-359.

[5] Fisher, R. A., "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Eugenics*, Vol. 7, April 1936, pp. 179-188.

[*6*] Cavalli-Sforza, L. L. and Bodmer, W. F., *The Genetics of Human Populations*, W. H. Freeman and Co., San Francisco, 1971.

[*7*] Felsenstein, J., "Maximum-Likelihood Estimation of Evaluating Trees from Continuous Characters," *American Journal of Human Genetics*, Vol. 27, July 1973, pp. 491-492.

[*8*] Spielman, R. S. and Smouse, P. E., "Multivariate Classification of Human Populations. I. Allocation of Yanomama Indians to Villages," *American Journal of Human Genetics*, Vol. 28, July 1976, pp. 317-331.

Address requests for reprints or additional information to
Steve Selvin, Ph.D.
School of Public Health
140 Earl Warren Hall
University of California
Berkeley, Calif. 94720