CDC-ATSDR Workshop

A Case Study in the Use of Race and Ethnicity in Public Health Surveillance

RICHARD S. COOPER, MD

Dr. Cooper is Professor and Chairman of the Department of Preventive Medicine and Epidemiology, Loyola University Medical School. Joan Kennelly, RN, MPH, Director of the Midwest Maternal and Child Health Data Improvement Project, provided conceptual and practical assistance in the preparation of this manuscript.

Tearsheet requests to Richard S. Cooper, MD, Loyola University Medical School, 2160 S. First Ave., Maywood, IL 60153; tel. 708-327-9010, FAX 708-327-9009.

Synopsis

Increased attention has been focused on the use of racial and ethnic categories in public health research and surveillance in recent years. This attention, however, has done more to increase the collective awareness of the difficulties inherent in these classification schemes than to resolve either the theoretical or practical problems.

It has been recognized for many years that health outcomes for a particular ethnic population must be seen as an interaction between genetically determined factors and socially mediated exposures. The operational meaning of these concepts remains poorly defined, however.

Although the general biologic construct of race and ethnicity appears straightforward, appearance is deceiving and the technical requirements that allow formulation of a rigorous and objective working definition have never been fully developed. The social factors that influence ethnic health differentials are perhaps better understood in practical terms; measurement of many of the relevant variables remains difficult, however.

In this article an attempt is made to outline the implications of the new developments in molecular biology for the concept of race, and to provide an illustrative example of the continued evolution of the social determinants of ethnicity.

THE CONSTRUCTION OF RACIAL and ethnic categories is becoming an ever more contentious issue in public health and medicine (*1–3*). If it has any meaning, race in common usage is a biologic concept. The underlying idea is deceivingly simple—a racial group represents a single breeding population that varies in definable ways from other human subpopulations. An effective operational definition continues to elude us, however. Genetic traits are the basis for the classification scheme, yet, in the past at least, gene frequencies were rarely measured directly. In practice, of course, the common designation of race is based on socially defined phenotypic traits, usually skin color and facial features, as they are seen through the filter of individual as well as social prejudice. In the United States, self-definition is the only legal basis for racial classification.

Ethnicity will be used in this paper to denote a category of greater generality than race. A reasonable definition was proposed by Montagu in 1964 (*4*):

> An ethnic group represents one of a number of populations, comprising the single species Homo sapiens, which individually maintain their differences, physical and cultural, by means of isolating mechanisms such as geographic and social barriers. These differences will vary as the power of the geographic and social barriers acting upon the original genetic differences varies.

Attributes from a number of dimensions can thus determine ethnicity, and I will give first priority to three areas—genes, culture, and social

class (see box). In the end, it is the interaction between the elements of this three-legged stool that becomes crucial. The biologic component of race is fully absorbed into the category of ethnicity, both in the narrow sense in which genetic factors influence group differences and through the socially mediated effects of racial discrimination. For example, much of the meaning of ethnicity for black Americans is determined by social class factors, although the defining characteristic for the group may be genetic in origin (that is, skin pigmentation).

The development of ethnic subpopulations is a result of the ongoing process of social evolution. Like biologic adaptation, the form that ethnicity takes evolves in response to the challenges posed by stresses in the collective environment. Its advantage over race follows from the recognition of the primacy of social arrangements to health. In a species where cultural differentiation is the predominant force determining survival advantage, the relevant form of raciation should be seen as the process of ethnic differentiation. This view of human adaptation therefore contrasts to physical specialization among species that rely on lower order biologic systems. Ethnic differentiation, as complemented by biological processes, is therefore primarily the natural and inevitable response of our species to changing opportunities and challenges in the social environment. By definition, its particular meaning will change over time, and the expectation of creating a fixed and universal classification scheme is not justified.

How Can We Define Racial Identity?

A wide gap exists between the process one might use in the assignment of race in a logical and consistent manner and the reality of everyday practice. Ignoring for the moment the social implications of this phenomenon, it is reasonable to argue that only in the last several years has the technical capacity become available to test whether a scientific means of assigning race in fact exists. These techniques are derived from advances in molecular biology and prominently include the identification of the variable number tandem repeats (VNTRs) used in DNA fingerprinting (*5,6*). These markers are highly polymorphic and provide much more information than typical alleles; they are also much less expensive and time-consuming to run and can be carried out in larger numbers.

Assuming the availability of appropriate data, is there the theoretical basis for a reasonable method for assigning membership in a racial category? At the outset, I would argue that the problem of race cannot be solved within a discrete classification scheme. It would be important to consider, however, whether the advent of new techniques have altered the utility of this exercise. Rather than relying on morphologic and skeletal features, it would appear more useful to attempt to accommodate the data emerging in the field of molecular genetics. This approach is also more consistent with the language of public health.

The following proposal is a naive outline of such a scheme. First, one would have to make an assumption about the existence of subspecies among humans. For the purposes of this paper I will take the majority point of view and assume that true sub-species do not exist (*3,4,7*). The underlying distribution of allele frequencies is therefore continuous across the species and subpopulations will be defined as sets drawn from this single distribution; assignment of a person will not be deterministic but estimated as a probability. Gene markers can provide information on discrete categories that could be used then to assign people to one or another group.

Theoretically, one could create a genetic reference data base and calculate the probability that any given person was likely to fall into one or another of these groups. Assuming the data consist of randomly selected, adaptively neutral alleles, the resulting categories could be arbitrarily defined. Thus, the category "black American" would be defined to capture the largest amount of discriminatory information in the data base, and distinguish blacks from whites or other groups in the United States. A polychotomous logistic model or some form of discriminant function analysis could

Race versus Ethnicity

Race:
- biologic construct
- single breeding population
- in practice based on superficial phenotypic traits such as skin color, facial features

Ethnicity:
- results from the interaction of genes, culture, and social class
- produced by social evolution
- the relevant form of raciation among a species where cultural differentiation predominates.

'In practice, of course, the common designation of race is based on socially defined phenotypic traits, usually skin color and facial features, as they are seen through the filter of individual as well as social prejudice. In the United States, self-definition is the only legal basis for racial classification.'

be used to calculate the probability of belonging to one or the other of these categories so defined.

To my knowledge, attempts to assign racial identity by this means have not been published. Classification based on genetic information has been applied to several other related problems, however. Extensive experience has been accumulated in estimating genetic distance among specific population groups (*8–10*). In this instance, however, one is dealing with mean differences between populations and the statistical analyses are more straightforward. DNA fingerprinting is used primarily as a tool to identify specific persons in criminal or paternity cases (*6*). Current debate focuses on the level of certainty with which a person can be identified on the basis of genotype. Although the level of certainty is high, where conviction of a serious crime is at stake, even low error rates may be unacceptable. Finally, methods currently exist to estimate genetic admixture, both for populations and individuals (*11*). These methods are again more effective at the population level, however, than for individual people (*12*).

Short of a direct test of this method, it is difficult to make precise statements about the nature of the data required and the usefulness of the final exercise. A number of contingencies are likely to apply. First, it would be necessary to identify appropriately informative alleles. Given the availability of VNTR technology, and the current explosion in genetic information, this requirement may be satisfied in the coming period. Second, it would be necessary to collect a sufficiently large data base to permit accurate definition of the population sub-group. This requirement surfaced as a major criticism of DNA fingerprinting and undermined confidence in its use in court. With sufficient resources, however, this obstacle likewise could be overcome.

Finally, the statistical model would need to have the ability to classify an individual person with sufficient precision. Thus, if the probability of being assigned to Group A was only 5 percent greater than the probability of being assigned to Group B, not much has been gained. This outcome would be determined by the actual distribution of the data. In a sense, with new techniques one would be re-investigating the question posed by Lewontin in the 1970s, "to what degree are we more different than alike?" (*13*). The ultimate test of race as a scientific concept, therefore, would be determined by the proportion of individuals who could be assigned with an acceptable degree of certainty to a genetically defined population subgroup.

Clearly nothing like this applies in practice. In the United States today race is used for groups that a casual observer can accurately classify (that is, black, white, Asian), whereas ethnicity is reserved for segments of the population in which subcategories are given greater importance and the migration history is recent (that is, Hispanic).

Although rigorous tests of this system have not been carried out, the DNA data are informative about several aspects of the current practice, and the conclusions are not particularly encouraging. Creating a rigorous model for racial designation based on the definitions currently in use is extremely problematic, however. A list of racial categories implies the concept of "unit race," that somehow the population subgroups being defined can be appropriately compared. In a statistical analogy, one hopes to find equal variance between groups. The genetic data clearly demonstrate, however, that current schemes do not subtend equivalent portions of the distribution; this is most likely a result of the disjunction between the current geographic distribution of populations and their long-term history.

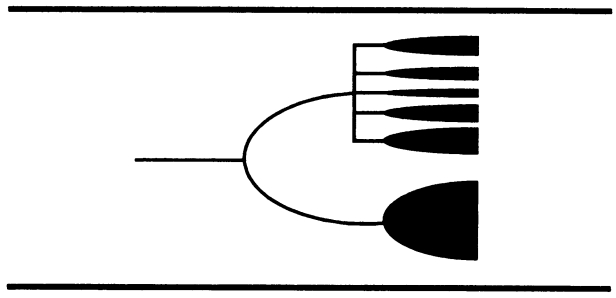
Evolution is often illustrated through the image of a phylogenetic tree. In that system, the African population represents the trunk, while those of Europe and Asia are the branches. The internal composition—or extent of diversity—within the branches and the trunk, if you will, may be very different. American blacks, as derived from populations of African origin, may be more heterogeneous as a population than are whites or Hispanics (*5,14*). The degree of internal heterogeneity is most likely correlated with the age of the population. Our species originated on the African continent, and native populations may therefore have had the greatest opportunity for further differentiation to

occur; genetic analyses, at least of single gene systems, repeatedly demonstrate the long evolutionary history of several African ethnic groups (*8*). Cavali-Sforza estimates that there is approximately twice as much diversity in Africa as elsewhere in the world (*9*). It must be acknowledged that the mitochondrial DNA findings are less supportive of greater internal genetic diversity in Africa, but relatively little data are available on this question (*15*). Even at the population level, therefore, the construct of race may not conform to actual distributions of populations. The figure shows a re-configuration of the traditional dendrogram, where the width of the branch now represents the heterogeneity within a group; the lowest branch is taken to be African in origin. Viewed from this dimension, it is possible that a person classified as black could be more dissimilar from another black than from a member of the white population.

In addition, the assignment of a person to a particular racial group on the basis of gene markers does not imply that the phylogenetic history of any particular gene will follow the same pattern. In medical research, since one is generally concerned about the health risk associated with particular genes, it would be necessary to specify the role that gene complex played in conditioning the risk of a given disease for the racial group of interest. The collective history of a racial group will therefore occasionally be misleading in relation to a single gene, and often for a single individual. Finally, there is growing evidence that for complex polygenic disorders, like hypertension and diabetes, different genes may condition risk in different subpopulations. Thus, specific alleles for the glucokinase gene are associated with risk of noninsulin-dependent diabetes mellitus in blacks in the United States, but not whites (*16,17*), and the angiotensinogen alleles that co-segregate with risk of hypertension in whites have not yet been associated with that disease among blacks (*18,19*). The implications of this phenomenon for assessing disease burden across racial groups are obvious. Only after the associations have been quantified in each of the groups under study can inferences be made about cross-group differences.

Is Ethnicity an Appropriate Replacement?

Within the schema applied in this presentation, ethnicity is proposed as the appropriate classification for public health research and practice. A potential objection to the use of ethnicity is the absence of a permanent reference system. Thus,



The width of the branches represents the relative heterogeneity within populations.

Japanese migrants to the United States become Japanese-Americans, and Native Americans can be disaggregated into a wide array of subgroups. By the same token, ongoing change occurs in all populations and racial groups in the absence of migration or cultural assimilation.

From another perspective, this mutability is precisely the strength of the concept of ethnicity. If we forsake the notion that human population groups are arranged in a fixed array, then accurate designation of current circumstances provides maximum information. A cursory listing of the relationship between ethnicity and health status will confirm this proposition. Thus, black Americans are most closely related to West Africans genetically, yet the latter do not share high rates of the major disease syndromes common in the United States, including coronary heart disease, hypertension, diabetes, and lung cancer. Ethnicity, as a working designation for population subgroups, can subsume the minor contribution to risk accounted for by known genes, such as sickle cell, as well as the secondary effects, like the health consequences of discrimination that result from darker skin color.

How Will Ethnicity Function in Surveillance?

The composition of the population of the United States is changing rapidly as new ethnic groups rise in prominence and surveillance systems must respond to this challenge. Rates of infant mortality and its associated risk factors are important data items in current monitoring systems and provide a clear example of the usefulness of ethnicity. The State of Massachusetts has recently provided data on low birth weight among African-origin ethnic groups (see table) (*20*). American blacks experienced substantially higher rates than did Haitians. Variation in the incidence of low birth weight and infant mortality among Hispanic groups is now well recognized (*21*). A complex set of social and historical factors most likely explains the stepwise

Crude and adjusted odds ratios of low birth weight by race-ethnic group, Massachusetts, 1987–89

Race-ethnic group	Crude ratio	Adjusted ratio
Black:		
American	3.3	1.5
Haitian	1.6	1.2
Cape Verdean	1.7	1.3
West Indian	1.9	1.4
Hispanic	2.0	1.4
White:		
Non-Hispanic	1.0	1.0

SOURCE: Reference 20.

increase in rates of low birth weight among infants born to Mexican Americans, Cubans, and Puerto Ricans. It is particularly interesting that, in the face of straightforward data, skepticism regarding the low infant mortality among Mexican immigrants lasted at least 20 years.

In the last several years, considerable attention has been given to the heterogeneity of groups classified as Asian. Approximately 850,000 Southeast Asian refugees have come to the United States since 1980, and surveillance systems have been required to respond to a new set of problems (*22–24*). In the national data, infant mortality has remained low among persons classified as Asian, despite lower average birth weights. Rapid change has occurred during this recent surveillance period. In 1978, for example, the prevalence of low birth weight among Asians was 15 percent, identical to the rate found among blacks (*23*). During the next decade, however, the rate of low birth weight fell to 9 percent, comparable to the rate among whites and Hispanics. Use of a flexible construct for ethnicity has made it very efficient to disaggregate the rates for Asians and identify significant within-group heterogeneity, particularly in relation to immigrants from Southeast Asia (*24*). In San Diego County, CA, for example, infant mortality rates for the Hmong are 50 percent higher than those for all Indochinese immigrants combined (*23*). At the same time, the post-neonatal mortality rates are higher than among U.S. whites for both the Hmong and Lao.

The immediate relevance of a data collection system built on ethnicity is apparent in the findings of the Midwest Maternal and Child Health Data Improvement Project (MMDIP), a collaborative effort involving the 10 States of Public Health Service Regions V and VII and the Department of Community Health Sciences at the University of Illinois School of Public Health. Funded through the Public Health Service's Maternal and Child Health Bureau, the project is designed to enhance the ability of participating States to access in a timely fashion data on maternal and child health services and outcomes, according to MMDIP Director Joan Kennelly, RN.

Based on linked birth-death files provided by State health departments, a full analysis of the birth cohorts for the years 1986 and 1987 was available by race in January 1992 (*25*). Infant mortality rates for Asians were 8.5 per 1,000 live births, versus 8.6 for whites. They were considerably lower than the 20.6 recorded for blacks and 15.5 for Native Americans (*25*). In the context of the data from San Diego, however, it was clear that analysis of all Asians together might overlook important differences for subgroups.

Census data were used to supplement the information currently available on vital records and to provide estimates of the size of the Asian subpopulations in individual States. A more precise surveillance measure of at-risk populations could therefore be obtained. For example, Hmong comprise 42 percent of the Asian population of Wisconsin and represent the largest Asian subgroup in that State, although they account for less than 10 percent in the region as a whole (*25*).

Up to this point, I have avoided direct comments on current national practice. With the previous examples as a background, however, it seems appropriate to open this discussion. The Public Health Service is, of course, actively attempting to reformulate and improve data collection systems related to race and ethnicity. In my view, however, the basic construct used in this process is limiting.

One of the more interesting examples relates to the birth certificate. In 1988, the National Center for Health Statistics document on Classification and Coding Instructions for Live Birth Records provided eight race codes; other than "white" and "black," all codes were subgroups for Native Americans or Asians (*26*). A comprehensive appendix defined how population groups of particular geographic origin were to be coded to one of the eight race codes. In this system, as always, race for Hispanics was coded as "white." An additional data item was provided for "Origin or Descent," however (*26*). A total of 24 options were permitted, including a number of Hispanic subgroups as well as some geographic codes (for example, "North African, Other African"). In 1991, the race code remained unchanged, but the "Origin or Descent" variable was gone and the designation of ethnicity was permitted only for Hispanic subgroups (*27*).

What ideas appear to guide this process? A

hierarchical view emerges, in which race is more fundamental than ethnicity and takes precedence as the basic organizing element. The motive of the scheme is to force all population groups into a category designating biological race, with little apparent concern for the outcome. The dissonant examples continue. Persons from India and Korea appear in the same category, and Hispanics are white, unless they are black, but Native Americans in the United States belong to a different race. Once the option of using a code for ancestry was raised, however, it filled a latent need.

Additional instructions in the "Hospitals' and Physicians' Handbook on Birth Registration and Fetal Death Reporting" continued the evolutionary process, in the following way: "Some States have a very small Hispanic population and may wish to obtain data on other groups. Therefore, they may opt to include a general ancestry item on their report instead of a specific Hispanic origin item" (*28*). Ancestry in this context was defined essentially as country of origin; religious affiliation—Jewish, Moslem, and so on—was not to be accepted.

The State of Michigan adopted these recommendations in 1988 (*29*), permitting self-declaration of ancestry as an ethnicity code. I have no information about the outcome of the Michigan experience, but this procedure should provide more extensive data than previously available and potentially obviates the need for relying on a system that dichotomizes race and ethnicity into separate categories. Race can be the definition of ethnicity for some groups, and the traditional hierarchy is thereby reversed. Rather than inventing some mal-fitting system that is then imposed on the world's populations, the classification scheme attempts to describe things as they are.

Conclusions

The purpose of this discussion has been to reexamine the difficulties inherent in using the category of race as a tool for public health surveillance and to suggest its universal replacement with ethnicity. A mounting series of theoretical and practical objections have been raised against the category of race. While the category of race will continue to exist, given the momentum of social habit, public health practice could be modified within data systems that have been constructed to meet pre-defined goals for health improvement. The potential impact of population differences in gene frequencies could be subsumed under the category of ethnicity. Although the judgment regarding the specific categories of ethnicity that are to be applied could become problematic if contested, in everyday practice this may not be a frequent occurrence. In any event, the outcome is likely to be more satisfactory than continued reliance on a construct defined in strictly biological terms that cannot be made to fit biological reality. Abandoning the concept of race for surveillance purposes would thus communicate an important message and constitute a step forward in the struggle to reject the misuse of race in public health.

'While the category of race will continue to exist, given the momentum of social habit, public health practice could be modified within data systems that have been constructed to meet pre-defined goals for health improvement. The potential impact of population differences in gene frequencies could be subsumed under the category of ethnicity.'

References

1. Cooper, R. S.: Celebrate diversity - or should we? Ethnicity Dis 1: 3-7 (1991).
2. Osborne, N. G., and Feit, M. D.: The use of race in medical research. JAMA 267: 275-279, Jan. 8 1992.
3. Hahn, P. A., Mulinare, J., and Teutsch, S. M.: Inconsistencies in coding of race and ethnicity between birth and death in U.S. infants. JAMA 267: 259-263, Jan. 8, 1992.
4. Montagu, A. M. F.: Discussion and criticism on the race concept. Curr Anthropology 5: 317 (1964).
5. Balazs, I., Baird, M., Clyne. M., and Meade, E.: Human population genetic studies of five hypervariable DNA loci. Am J Hum Genet 44: 182-190 (1989).
6. Cohen, J. E., Lynch, M., and Taylor, C. E.: Forensic DNA tests and Hardy-Weinberg equilibrium. Science 253: 1037-1038 (1991).
7. Cooper, R. S.: A note on the biological concept of race and its application in epidemiological research. Am Heart J 108: 715-773 (1984).
8. Nei, M., and Roychoudhury, A. K.: Genetic relationship and evolution of human races. Evol Biology 14: 1-59 (1982).
9. Cavali-Sforza, L. L.: Genes, peoples and languages. Sci Am 265: 104-111, November 1991.
10. Hill, A. V. S., et al.: Extensive genetic diversity in the HLA class II region of Africans, with a focal predominant

allele, DRB*1304. Proc Natl Acad Sci U S A 89: 2277–2281 (1992).
11. Chakraborty, R., Kamboh, M. I., and Ferrell, R.: 'Unique' alleles in admixed populations: a strategy for determining 'hereditary' population differences in disease frequencies. Ethnicity Dis 1: 245–256 (1991).
12. Weiss, K. M.: Natural and unnatural crosses. Ethnicity Dis 1: 231–235 (1991).
13. Lewontin, R. C.: Apportionment of human diversity. Evol Biology 6: 381–398 (1972).
14. Uterman, G.: The mysteries of lipoprotein (a). Science 246: 904–910, Nov. 17, 1989.
15. Templeton, A. R.: The "Eve" hypothesis: a genetic critique and reanalysis. Am Anthroplogist 95: 51–72 (1993).
16. Chiu, K. C., Province, M. A., and Permutt, M. A.: Glucokinase gene is genetic marker for NIDDM in American blacks. Diabetes 41: 843–849 (1992).
17. Permutt, A., Chiu, K. C., and Tanizawa Y.: Glucokinase and NIDDM. A candidate gene that paid off. Diabetes 41: 1367–1372 (1992).
18. Jeunemaitre, X., et al.: Molecular basis of human hypertension: role of angiotensinogen. Cell 71: 169–180 (1992).
19. Rotimi, C. N., Cooper, R. S., Ward, R. H., and Morrison, L.: The role of the angiotensinogen gene in human hypertension: absence of an association among African Americans. Genet Epidemiol. In press.
20. Friedman, D. J., et al.: Ethnic variation and maternal risk characteristics among blacks: Massachusetts, 1987 and 1988. MMWR Morb Mortal Wkly Rep 40: 403–411, July 21, 1991.
21. Ventura, S. J.: Births of Hispanic parentage, 1982. Monthly Vital Statistics Report, vol. 34, No. 4, Supp. DHHS Publication No. (PHS) 85–1120, National Center for Health Statistics, Hyattsville, MD, July 23, 1985.
22. Weeks J. P., and Rumbaut, R. G.: Infant health and mortality among Indochinese refugees in San Diego County. International Population Center, San Diego State University, San Diego, CA, August 1988.
23. Shino, P. H., et al.: Birth weight among women of different ethnic groups. JAMA 255: 48–52, Jan. 3, 1986.
24. Helsel, D., Petitti, D. B., and Kunstadter, T.: Pregnancy among the Hmong. Birthweight, age, and parity. Am J Public Health 82: 13612–1364 (1992).
25. Kennelly, J., Rhodes, C., Harders, R., and Morris, N.: Perinatal and infant data book addendum; Native American and Asian data. The Midwest Maternal and Child Health Data Improvement Project, Chicago, IL, January 1992.
26. Instruction manual, Pt. 3a. Classification and coding instructions for live birth records, 1987. National Center for Health Statistics, Hyattsville, MD, 1988.
27. Instruction manual, Pt. 3a. Classification and coding instructions for live birth records, 1991. National Center for Health Statistics, Hyattsville, MD, 1991.
28. Hospitals' and physicians' handbook on birth registration and fetal death reporting. National Center for Health Statistics, Hyattsville, MD, 1989.
29. Instruction letter number 11. Office of the State Registrar and Center for Health Statistics, Michigan Department of Public Health, Lansing, Jan. 1, 1989.