The Importance of Race and Ethnic Background in Biomedical Research and Clinical Practice

Esteban González Burchard, M.D., Elad Ziv, M.D., Natasha Coyle, Ph.D., Scarlett Lin Gomez, Ph.D., Hua Tang, Ph.D., Andrew J. Karter, Ph.D., Joanna L. Mountain, Ph.D., Eliseo J. Pérez-Stable, M.D., Dean Sheppard, M.D., and Neil Risch, Ph.D.

A debate has recently arisen over the use of racial classification in medicine and biomedical research. In particular, with the completion of a rough draft of the human genome, some have suggested that racial classification may not be useful for biomedical studies, since it reflects “a fairly small number of genes that describe appearance” and “there is no basis in the genetic code for race.” In part on the basis of these conclusions, some have argued for the exclusion of racial and ethnic classification...
Definitions of race and ethnic background have often been applied inconsistently. The classification scheme used in the 2000 U.S. Census, which is often used in biomedical research, includes five major groups: black or African American, white, Asian, native Hawaiian or other Pacific Islander, and American Indian or Alaska native. In general, this classification scheme emphasizes the geographic region of origin of a person’s ancestry. Ethnic background is a broader construct that takes into consideration cultural tradition, common history, religion, and often a shared genetic heritage.

From the perspective of genetics, structure in the human population is determined by patterns of mating and reproduction. Historically, the greatest force influencing genetic differentiation among humans has been geography. Great physical distances and geographic barriers (e.g., high mountains, large deserts, and large bodies of water) have imposed impediments to human communication and interaction and have led to geographically determined endogamous (i.e., within-group) mating patterns resulting in a genetic substructure that largely follows geographic lines. The past two decades of research in population genetics has also shown that the greatest genetic differentiation in the human population occurs between continentally separated groups.

Endogamous mating within continents has given rise to further subdivisions, often corresponding to ethnic groups. This subdivision is again partially attributable to geography but is also associated with social factors, including religion, culture, language, and other sources of group identification. Thus, ethnic groups are genetically differentiated to varying degrees, depending on the extent of reproductive isolation and endogamy, but typically less so than are continentally defined groups.

Considerable debate has focused on whether race and ethnic identity are primarily social or biologic constructs. Unlike a biologic category such as sex, racial and ethnic categories arose primarily through geographic, social, and cultural forces and, as such, are not stagnant, but potentially fluid. Even though these forces are not biologic in nature, racial or ethnic groups do differ from each other genetically, which has biologic implications.

The racial or ethnic groups described above do not differ from each other solely in terms of genetic makeup, especially in a multiracial and multicultural society such as the United States. Socioeconomic status is strongly correlated with race and ethnic background and is a robust predictor of access to and quality of health care and education, which, in turn, may be associated with differences in the incidence of diseases and the outcomes of those diseases. For example, black Americans with end-stage renal disease are referred for renal transplantation at lower rates than white Americans. Black Americans are also referred for cardiac catheterization less frequently than white Americans. In some cases, these differences may be due to bias on the part of physicians and discriminatory practices in medicine. Nonetheless, racial or ethnic differences in the outcomes of disease sometimes persist even when discrepancies in the use of interventions known to be beneficial are considered. For example, the rate of complications from type 2 diabetes mellitus varies according to racial or ethnic category among members of the same health maintenance organization, despite uniform utilization of outpatient services and after adjustment for levels of education and income, health behavior, and clinical characteristics. The evaluation of whether genetic (as well as nongenetic) differences underlie racial disparities is appropriate in cases in which
important racial and ethnic differences persist after socioeconomic status and access to care are properly taken into account.

---

**EVIDENCE OF GENETIC DIFFERENTIATION AMONG RACES**

There are estimated to be at least 15 million genetic polymorphisms,\(^{1,2}\) and an as yet undefined subgroup of these polymorphisms underlie variation in normal and disease traits. The importance of such variation is underscored by the fact that a change of only a single base pair is required to cause many well-known inherited diseases, such as sickle cell disease, or to increase the risk of common disorders, such as Alzheimer’s disease. Studies in population genetics have revealed great genetic variation within racial or ethnic subpopulations, but also substantial variation among the five major racial groups, as defined above.\(^5\) This variation has been demonstrated in at least three ways.

First, investigators studying the population genetics of indigenous groups from around the world have constructed ancestral-tree diagrams showing branching relationships among the various indigenous groups. Despite differences in the types of markers used, these studies have been consistent in showing that the human population has major branches corresponding to the major racial groups, with subbranches within each racial group associated with indigenous groups.\(^{13-15}\)

Second, analysis of genetic clusters has been applied to persons of diverse ancestry, with a focus on genotypes at multiple genetic loci. These analyses have also consistently resulted in the delineation of major genetic clusters that are associated with racial categories.\(^{16-19}\) The primary difference between the results of these studies and the categories used by the U.S. Census is that South, Central, and West Asians cluster with Europeans and are separate from East Asians.

Third, studies have examined the distribution of differences among racial groups in the frequency of alleles (genetic variants) at both microsatellite and single-nucleotide–polymorphism (SNP) markers, demonstrating a median difference in allele frequency of 15 to 20 percent, with 10 percent of markers showing a difference of 40 percent or more.\(^{5,20,21}\) Thus, for an allele with a frequency of 20 percent or greater in one racial group, the odds are in favor of seeing the same variant in another racial group. However, variants with a frequency below that level are more likely to be race-specific. This race-specificity of variants is particularly common among Africans, who display greater genetic variability than other racial groups and have a larger number of low-frequency alleles.\(^{17}\) These results indicate that the frequency of variant alleles underlying disease or normal phenotypes can vary substantially among racial groups, leading to differences in the frequency of the phenotypes themselves. Such differences in frequency are also found among ethnic groups, but these differences are typically not as great. Furthermore, self-defined ancestry is very highly correlated with genetically defined clusters.\(^5,19\)

---

**GENETIC DIFFERENCES IN DISEASE AMONG RACIAL AND ETHNIC GROUPS**

To what degree does genetic variability account for medically important differences in disease outcomes among racial and ethnic groups? The answer depends on the frequency of the genetic variants or alleles (mutations) underlying the susceptibility to the disease. For mendelian disorders, the relevance of race and ethnic background is readily apparent. Mutations that have frequencies of less than 2 percent are nearly always race-specific and, in fact, are often specific to single ethnic groups within a given race. For example, numerous mutations with frequencies in this range occur uniquely in Ashkenazi Jews, French Canadians, the Amish, or European gypsies. This is because such populations descend from a relatively small number of founders and have remained endogamous for a large part of their history. Mutant alleles with frequencies of more than 2 percent but less than 20 percent are typically prevalent within single racial groups but not in other racial groups. For example, hemochromatosis is associated with a mutant allele (C282Y) found in all European groups and at especially high frequency (8 to 10 percent) in northern Europeans, but is virtually absent in nonwhite groups.\(^22\)

“Complex” genetic disorders such as asthma, cancer, diabetes, and atherosclerosis are most likely due to multiple, potentially interacting, genes and environmental factors and are thus more challenging to study. The genetic determinants of the majority of these disorders are currently poorly understood, but the few examples that do exist demonstrate clinically important racial and ethnic differences in gene frequency. For example, factor V Leiden, a genetic variant that confers an increased
Risk of venous thromboembolic disease, is present in about 5 percent of white people. In contrast, this variant is rarely found in East Asians and Africans (prevalence, <1 percent).23, 24 Susceptibility to Crohn’s disease is associated with three polymorphic genetic variants in the CARD15 gene in whites25; none of these genetic variants were found in Japanese patients with Crohn’s disease.26 Another important gene that affects a complex trait is CCR5 — a receptor used by the human immunodeficiency virus (HIV) to enter cells. As many as 25 percent of white people (especially in northern Europe) are heterozygous for the CCR5–delta32 variant, which is protective against HIV infection and progression, whereas this variant is virtually absent in other groups, thus suggesting racial and ethnic differences in protection against HIV.27

Other alleles occur in all ethnic groups but with highly variable frequency. Increasingly, researchers and clinicians are focusing on identifying and studying the genetic variants that influence responses to drugs and the metabolism of drugs (an area of study termed pharmacogenetics). One example is N-acetyltransferase 2 (NAT2), an enzyme involved in the detoxification of many carcinogens and the metabolism of many commonly used drugs. Genetic variants of NAT2 result in two phenotypes, slow and rapid acetylators. Population-based studies of NAT2 and its metabolites have shown that the slow-acetylator phenotype ranges in frequency from approximately 14 percent among East Asians to 34 percent among black Americans to 54 percent among whites.28 Genetic variants of NAT2 are important because they may predict toxic effects of drugs and because they may contribute to racial and ethnic variation in the incidence of environmentally induced cancers.

**Racial and Ethnic Differences as Clues to Interactions**

Even when all racial and ethnic groups share a genetic variant that causes a disease, studies of different groups may offer important insights. One of the best-known examples of a gene that affects a complex disease is APOE. A patient harboring a variant of this gene, APOE e4, has a substantially increased risk of Alzheimer’s disease. APOE e4 is relatively common and is seen in all racial and ethnic groups, albeit at different frequencies, ranging from 9 percent in Japanese populations to 14 percent in white populations to 19 percent in black American populations.29 However, a recent meta-analysis has demonstrated that the effect of APOE e4 on the risk of Alzheimer’s disease varies according to race.29 Homozygosity for the e4 allele increases risk by a factor of 33 in Japanese populations and by a factor of 15 in white populations, but only by a factor of 6 in black American populations; similarly, homozygosity for the e4 allele increases the risk by a factor of 5.6 in Japanese populations, by a factor of 3.0 in white populations, and by a factor of 1.1 in black American populations. Although the reason for this variation in risk remains unknown, it suggests that there may be genetic or environmental modifiers of this gene. Thus, even when a genetic determinant of a complex disease is present in all racial and ethnic groups, racial and ethnic classification may offer additional important insights.

**Racially Admixed Populations**

Although studies of population genetics have clustered persons into a small number of groups corresponding roughly to five major racial categories, such classification is not completely discontinuous, because there has been intermixing among groups both over the course of history and in recent times. In particular, genetic admixture, or the presence in a population of persons with multiple races or ethnic backgrounds, is well documented in the border regions of continents and may represent genetic gradations (clines) — for example, among East Africans (e.g., Ethiopians)18 and some central Asian groups.19 In the United States, mixture among different racial groups has occurred recently, although in the 2000 U.S. Census, the majority of respondents still identified themselves as members of a single racial group. Genetic studies of black Americans have documented a range of 7 to 20 percent white admixture, depending on the geographic location of the population studied.30 Despite the admixture, black Americans, as a group, are still genetically similar to Africans. Hispanics, the largest and fastest growing minority population in the United States, are an admixed group that includes white and Native American ancestry, as well as African ancestry.31 The proportions of admixture in this group also vary according to geographic region.

Although the categorization of admixed groups poses special challenges, groups containing persons with varying levels of admixture can also be particularly useful for genetic–epidemiologic studies. For example, Williams et al. studied the associ-
ation between the degree of white admixture and the incidence of type 2 diabetes mellitus among Pima Indians. They found that the self-reported degree of white admixture (reported as a percentage) was strongly correlated with protection from diabetes in this population. Furthermore, as noted above, information on race or ethnic background can provide important clues to effects of culture, access to care, and bias on the part of caregivers, even in genetically admixed populations. It is also important to recognize that many groups (e.g., most Asian groups) are highly underrepresented both in the population of the United States and in typical surveys of population genetics, relative to their global numbers. Thus, primary categories that are relevant for the current U.S. population might not be optimal for a globally derived sample.

**Risks Entailed by Ignoring Race in Biomedical Research and Clinical Practice**

Given its controversial social and political history, it may be tempting to abandon the notion of race altogether, particularly if we believe that continued attention to differences among races may perpetuate discrepancies in health status and well-being. Indeed, some have advocated discontinuing the collection of information about race and ethnic background, presumably as a way of protecting minority groups. In California, advocates of this move are pushing for a state law — through the Racial Privacy Initiative — that would prohibit racial classification by the state or other public entities. Although this initiative formally excludes a ban on classification for the purposes of medical research, the abolition of the collection of data on race or ethnic group for all other purposes would eliminate these data from many public data bases on which clinicians and scientists rely in order to make meaningful inferences about the effects of race and ethnic background on health and disease in persons and populations.

We believe that ignoring race and ethnic background would be detrimental to the very populations and persons that this approach allegedly seeks to protect. Information about patients’ ethnic or racial group is imperative for the identification, tracking, and investigation of the reasons for racial and ethnic differences in the prevalence and severity of disease and in responses to treatment. This information is also crucial for identifying different risk-factor profiles even when a disease does not occur with dramatically different frequencies in different racial or ethnic groups. Furthermore, knowledge of a person’s ancestry may facilitate testing, diagnosis, and treatment when genetic factors are involved. For example, there are already tests to screen for disease-causing mutations that are tailored to specific racial or ethnic groups. Currently, racial and ethnic minorities in the United States are underrepresented in many clinical studies. If investigators ignored race and ethnic background in research studies and persons were sampled randomly, the overwhelming majority of participants in clinical studies in the United States would be white, and minority populations would never be adequately sampled. In cases in which there are important racial and ethnic differences in the causes of disease or other outcomes or in which there are interactions between race or ethnic background and other factors contributing to these outcomes, such patterns would never be discovered, their causes could not be identified, and the appropriate interventions would never be applied in the groups in which they were needed. Despite the fact that the National Institutes of Health requires reporting of all racial or ethnic groups participating in biomedical research, limited progress has been made in the inclusion of minority groups.

**Conclusions**

There are racial and ethnic differences in the causes, expression, and prevalence of various diseases. The relative importance of bias, culture, socioeconomic status, access to care, and environmental and genetic influences on the development of disease is an empirical question that, in most cases, remains unanswered. Although there are potential social costs associated with linking race or ethnic background with genetics, we believe that these potential costs are outweighed by the benefits in terms of diagnosis and research. Ignoring racial and ethnic differences in medicine and biomedical research will not make them disappear. Rather than ignoring these differences, scientists should continue to use them as starting points for further research. Only by focusing attention on these issues can we hope to understand better the variations among racial and ethnic groups in the prevalence and severity of diseases and in responses to treat-
ment. Such understanding provides the opportunity to develop strategies for the improvement of health outcomes for everyone.

From the Lung Biology Center (E.G.B., N.C., D.S.), the Division of General Internal Medicine (E.Z., E.J.P.-S.), the Department of Medicine (E.G.B., E.Z., N.C., E.J.P.-S., D.S.), and the Medical Effectiveness Research Center for Diverse Populations (E.G.B., E.Z., E.J.P.-S.), University of California, San Francisco; and San Francisco General Hospital (E.G.B., N.C., D.S.) — both in San Francisco; the Division of Epidemiology, Department of Health Research and Policy (S.L.G.), the Department of Genetics (N.R.), Stanford University School of Medicine, and the Department of Statistics (H.T.) and the Department of Anthropological Sciences (J.L.M.), Stanford University — both in Stanford, Calif.; the Northern California Cancer Center, Union City, Calif. (S.L.G.); and the Department of Epidemiology and Health Services Research, Division of Research, Kaiser Permanente, Oakland, Calif. (A.J.K., N.R.). Address reprint requests to Dr. González Burchard at the University of California, San Francisco, San Francisco, CA 94143-0833, or at eburch@itsa.ucsf.edu.

Drs. González Burchard and Ziv contributed equally to the article.

34. Gifford AL, Cunningham WE, Heslin KC, et al. Participation in a racialprivacy.org.)