Racial and Ethnic Identification Practices in Public Health Data Systems in New England

M. BARTON LAWS, PHD[a]
RACHEL A. HECKSCHER, MPH[b]

SYNOPSIS

Objective. Efforts are underway to standardize "racial" and "ethnic" identification in public health data systems under the Revised Minimum Standards for the Classification of Federal Data on Race and Ethnicity issued in 1997. This study analyzed the racial and ethnic constructs and labels used in public health data systems maintained by the six New England states in light of these standards.

Methods. The authors surveyed public health officials responsible for ongoing individual-level data systems and reviewed relevant documents.

Results. Information was obtained on 169 of 170 identified data systems. Ninety-one systems (54%) conformed to the federal standard in having separate "race" and "ethnicity" fields, yet many of these did not conform to the standard in other respects. Fifty-five systems had only a race field; of these, 20 included no identifiers corresponding to Hispanic and/or Latino ethnicity. Three systems used only an ethnicity field. The systems used various lists of racial and/or ethnic categories, and overlapping but not fully comparable labels. Few systems allowed for identification of ancestry groups not included in the revised federal guidelines but with large populations in New England, such as Brazilians. Some definitions and coding instructions seemed inconsistent with social and geographic reality.

Conclusions. These public health data systems used inconsistent methods for classifying people by race and ethnicity. Standardization according to federal standards would improve comparability, but would limit options for defining and including some ethnic groups while forcing other groups to be aggregated in single race categories, perhaps inappropriately. Fundamental reconsideration of racial and ethnic categorization is called for.

---

[a] Latin American Health Institute, Boston, MA, and Tufts University School of Medicine, Department of Family Medicine and Community Health, Boston, MA

[b] Massachusetts Department of Public Health, Bureau of Communicable Disease Control, Boston, MA; at the time of this study, Ms. Heckscher was with the Latin American Health Institute.

Address correspondence to: M. Barton Laws, PhD, Senior Investigator, Latin American Health Institute, 95 Berkeley St., Boston, MA 02116; tel. 617-350-6900 x 142; fax 617-350-6901; e-mail <bart@lhi.org>.

Recent years have seen unprecedented attention to the problem of racial and ethnic classification of the U.S. population for purposes of public health surveillance and research. The Census has classified people by "race" from its inception in 1790.[1] Vital statistics in the U.S. have been analyzed by "race" since 1940.[2] However, practices for racial classification have varied over time and across federal programs. In 1977, the federal Office of Management and Budget (OMB) promulgated Statistical Policy Directive 15, which called for a minimum of four "race" categories, and two "ethnicity" categories ("Hispanic" and "not Hispanic").[3] Nevertheless, practices at the federal level continued to be inconsistent into the 1990s.[4,5]

In 1997, OMB issued Revised Minimum Standards for the Classification of Federal Data on Race and Ethnicity, which federal agencies are required to use for all surveys and administrative data by January 1, 2003.[3,6] Although these standards are not binding on the states, much of the data collection by state health departments is done under federally funded programs and conforms to the standards of these programs. Furthermore, states have a strong incentive to use categories consistent with those used in the Census—the source of population denominators for rate calculations—and in other federal and state systems, for comparability.[7]

The revised federal standards mandate that subjects first be asked whether they are "Hispanic or Latino" (or, optionally, "Spanish, Hispanic, or Latino") and then asked to identify themselves by "race."[6] In the order given in the document, the race categories are: white; black or African American; American Indian or Alaska Native; Asian; and Native Hawaiian or other Pacific Islander. More specific sub-categories are delineated in subsequent publications[8] and technical documentation for the 2000 Census[9] for the Hispanic/Latino ethnicity category, the Asian race category, and the Native Hawaiian or other Pacific Islander race category. Hispanic or Latino individuals can be "Mexican, Mexican American or Chicano," Puerto Rican, or Cuban, or they can choose "other Spanish, Hispanic, Latino" and write in a specification of their choice.

The standards permit the use of a single, combined race and ethnicity question when collecting data by observation only, but strongly encourage use of the two-question format and respondent self-identification. In later policy guidance, OMB has clarified that "[t]he 1997 standards emphasize self-reporting or self-identification as the preferred method for collecting data on race and ethnicity. The standards do not establish criteria or qualifications . . . to be used in determining a particular individual's racial or ethnic classification. . . . Self-identification for race and Hispanic or Latino origin means that the responses are based on self-perception. . . ."[8]

In the 2000 Census, people were not asked directly whether they were of Asian or Native Hawaiian or other Pacific Islander race. Rather, they were given as options for race a list of nine nationalities—including, for example, Asian Indian, Chinese, Filipino, and Japanese—and given the option of selecting "other Asian," "other Pacific Islander," or "some other race," and writing in a specification.

In most tabulations of Census results, these Asian and Native Hawaiian or other Pacific Islander nationalities are collapsed into a single Asian race and a single Pacific Islander race. The Census Bureau considers Filipinos part of the Asian race. People who wrote in Haitian, Dominican, Cape Verdean, or a sub-Saharan African nationality as their race on the 2000 Census were recoded as "black or African American."[9]

The U.S. Centers for Disease Control and Prevention (CDC) is encouraging consistency among public health data systems, and has developed a proposed standard coding scheme for race and ethnicity consistent with the revised standards.[10,11] This system allows for detailed specifications by assigning nationalities as subsets of races, e.g., Jamaicans are considered a subset of the black race, and Israelis are a subset of the white race. Nationalities associated with predominantly Spanish-speaking countries are represented as subsets of Hispanic ethnicity. Many nationalities such as Brazilians and South Africans were omitted, presumably because the residents of these countries do not fit the definition of Hispanic/Latino ethnicity and there did not appear to be any credible way to map them onto a race; these countries are notably multiracial. Indigenous American groups with original lands outside of the present-day United States (e.g., Maya, Quechua) were also omitted, although they could logically be included as subsets of the Native American "race."

The CDC scheme will not only apply to national disease surveillance systems and other systems overseen by the CDC, but has also been proposed for the standardized medical claims data system mandated by the Health Insurance Portability and Accountability Act of 1996[12] (Personal communication, Robert A. Davis, Statewide Planning and Research Cooperative System Coordinator, New York Department of Health, November 2001).

The separate Hispanic/Latino ethnicity category in federal data systems co-exists awkwardly with the race classifications. Hispanics are also asked to choose a race, but the plurality (42.2%) chose "some other race" in the 2000 Census.[13] It is difficult to see why a person may be of Chinese, Filipino, Native American, or Hawaiian race, but can only be of Hispanic ethnicity. Important, ethnically distinct immigrant groups such as Brazilians or Haitians are not distinguishable in this system, while individuals from certain countries are automatically coded to a race category, even when those countries are multiracial or their societies do not share the racial classification system of the U.S.

These decisions were made against a backdrop of vigorous controversy over the appropriate use of the concepts of race and ethnicity in public health. The races were once thought to be biological sub-species of humanity, in which the highly visible traits used to define them—principally complexion—were correlated with other supposedly genetically determined traits, such as intelligence and temperament. It is now understood that this division of humanity and various sub-grouping schemes were based on socially determined prejudices, not scientific data. Genetic variation within the "races" is far greater than the variation between them, and the selection of different traits as the determining factors would result in entirely different sets of "races."[14]

Many have noted that racial and ethnic labels are widely used without clear definitions, that there is frequently confusion between the concepts of race and ethnicity, and that the validity and reliability of racial/ethnic identification are often unclear.[15–17] Some have noted that the revised OMB categories are unsatisfactory in that they fail to accurately or completely capture the ethnic diversity of the population.[7,18,19] Some have argued for abolishing the collection of data on race and ethnicity entirely,[20] while others view it very skeptically.[21] However, most commentators have argued that race and ethnicity are social realities and that the social environment affects health in myriad ways; hence, failing to incorporate race and ethnicity as variables in health research would make real problems invisible.[22–26]

The American Association of Anthropologists[27]and the Institute of Medicine[28] have both proposed doing away with race in research and surveillance and instead using the concept of ethnicity exclusively. Their argument is that ethnicity—and the culturally patterned behaviors, social supports or stresses, and social statuses that come with it—can be linked with health outcomes through bio-psycho-social models of health and illness, whereas the race construct implies a false biological determinism. An objection to this proposal is that counting people by race, as long as we emphasize that we are using socially constructed categories, is essential to tracking the consequences of racism.[29,30] Fullilove argues for abandoning the race concept, substituting study of racism *per se*, and nuanced and sociologically valid concepts of ethnicity.[31]

This important discussion has taken place in something of an empirical vacuum. We have not had detailed, comprehensive information about how race and ethnicity have in fact been classified, and the relevant data collected, in public health data systems in recent years. Accordingly, we explored the uses of race and ethnicity in individual-level public health data collected by the six New England states.

METHODS

This study was undertaken on behalf of the New England Coalition for Health Equity, a regional organization concerned with health disparities. New England includes states that are predominantly rural and relatively ethnically homogeneous as well as states that are urbanized and more diverse. About 4.5% of the U.S. population lived in New England in 2000.[32]

The present study is a cross-sectional survey of public health data systems maintained by the public health agencies of the six New England states. The state bureaucracies are organized in various ways, and the public health agencies differ in designation and scope of responsibility. Nevertheless, in all of the states the surveyed agencies control core public health data systems: vital records, disease surveillance databases, and hospital discharge data.

The agencies included in this study were the Connecticut Department of Public Health, the Maine Bureau of Health, the Massachusetts Department of Public Health, the Public Health Services Division of the New Hampshire Department of Health and Human Services, the Rhode Island Department of Public Health, and the Vermont Department of Health. We also collected information on Youth Risk Behavior Surveillance Survey databases and/or immunization data systems in states where these are maintained by state Departments of Education, as these are core public health data systems.

We obtained what we believe to be complete listings of ongoing, individual-level, public health databases maintained by the public health agencies in the six New England states, plus the specific Department of Education databases mentioned above. By an individual-level database we mean any database in which records represent an individual or an event befalling an individual, and which can generate useful information about a definable population. These include program utilization data systems, which may not be generalizable beyond the beneficiaries. We did not include in this study systems in which the unit of analysis was an institution, such as a hospital or nursing home, or an individual professional, such as a physician. Although vital records registries maintain marriage and divorce records, they are not generally tabulated or analyzed for purposes related to public health; thus we excluded these systems from our analysis. Three states reported on linked infant birth-death files. Because these databases inherit their demographic information from birth certificates, we excluded them as redundant.

We obtained our listings in various ways. Massachusetts and Rhode Island had already prepared inventories of their systems. The New Hampshire Division of Health Statistics, Planning and Evaluation (a unit of the Public Health Services Division) assisted us by producing a listing of that state's systems and took responsibility for collecting much of the data. Officials in Maine and Vermont also assisted us in identifying systems. In Connecticut, the state Office of Multicultural Health conducted a survey of data systems at about the same time we did and provided us with raw data; we collected additional information from respondents to the Connecticut survey and identified additional data systems in the state, from which we collected data.

Data collection

We conducted this survey from January through July 2001. We mailed, faxed, or e-mailed questionnaires, depending on respondents' preferences, to officials responsible for all of the identified data systems, except for the 16 New Hampshire systems, where collaborating state officials implemented the survey for us. When contacts failed to respond, we often completed the questionnaire through telephone interviews. When a returned questionnaire contained incomplete, unclear, contradictory, or implausible information, we contacted the respondent to clarify the response. We also asked respondents to send copies of any data dictionaries or coding instructions. We received dictionaries for 38 databases.

Because vital records in two states allowed for an open-ended write-in of "race," respondents from these states checked off all of the race and ethnicity categories specified on our questionnaire. However, we determined on follow-up that responses were coded into a limited set of categories used for analysis and reporting. We report here on the limited set of codes actually used.

In fact, many systems allowed a write-in option, but open-ended text fields cannot be analyzed without a good deal of processing to account for variant and wrong spellings, the multiple names used for some groups, and the not inconsiderable number of frivolous or tendentious responses that may appear in these fields. Generally, write-in responses were reported as "other" in these data systems.

We asked respondents to report the appropriate classification of each system as either Vital Records, Surveillance, Health Care Utilization, Other Program Utilization (e.g., Healthy Start), Health Status/Epidemiology, Surveillance, or Other. ("Other" systems include, for example, program activity reports for outreach programs, which include information on individual outreach contacts as well as other information such as locations of outreach efforts and materials distributed.) Health Status/Epidemiology and Surveillance were not always clearly distinguishable. The systems in the Health Status/Epidemiology category generally did not focus on a single diagnosis but on a spectrum of conditions or health status indicators, such as pediatric nutrition or chronic disease; they often tracked individuals over time and assessed factors such as severity or costs, rather than incidence or prevalence.

Respondents' choices of category were generally honored, but we did re-classify a few systems for which we considered the survey response inaccurate.

Data analysis

Data were entered into an SPSS database (Version 10.1; 2001). Dummy variables indicated whether each of the racial or ethnic labels we encountered more than four times was included in each of the systems. For uncommon labels, we created fields counting the number of labels that were aggregatable to the standard OMB "race" and "Hispanic" categories as well as to geographic constructs (e.g., the Caribbean or Latin America) that did not map onto OMB categories. When possible, data were checked against data systems' questionnaires or data dictionaries to confirm accuracy, and errors were corrected. Significance values reported here are based on *F* tests.

RESULTS

We obtained responses for 169 (99.4%) of 170 identified data systems. Massachusetts, the most populous state with the oldest and largest department of public health, had substantially more data systems than the other states. The Table shows the systems included in our data, by state and classification.

The number of Vital Records systems varied because not all states collected data on induced abortion.

Table. Number of data systems surveyed, by state and classification

	Classification						
State	*Vital records*	*Surveillance*	*Health care utilization*	*Other program utilization*	*Health status/ epidemiology*	*Other*	*Total*
Connecticut	4	13	5	7	—	—	29
Massachusetts	3	24	6	15	5	5	58
Maine	4	9	4	1	—	1	19
New Hampshire	3	8	4	1	—	—	16
Rhode Island	4	14	3	3	2	1	27
Vermont	4	9	4	1	1	1	20
Total	22	77	26	28	8	8	169

The Surveillance systems included communicable disease databases, Behavioral Risk Factor Surveillance System databases, lead poisoning databases, tumor registries, birth defect registries, and various systems monitoring injuries and violence. Health Care Utilization databases included hospital discharge system databases and immunization registries and survey databases.

Structure of racial and ethnic classification

Twenty systems (11.8%) did not collect any race or ethnicity data (see Figure 1). The total number of closed-ended race/ethnicity categories (including "other") used by each of the remaining 149 systems ranged from 4 to 44, with a mean of 13.4 (standard deviation [SD] = 10.2), a median of 8, and a mode of 6. (The mean number of categories is 11.8 when the 20 systems not collecting racial/ethnic data are included in the total.)

The systems used some fundamentally different classification strategies. The 89 systems (52.7%) that followed the revised OMB standards in separating race and ethnicity in principle permitted cross-classification by race and Hispanic ethnicity, which means that the possible number of reportable categories was the product of the two sets rather than their sum, although the resulting categories might not all be very meaningful (e.g., there are few Hispanic Vietnamese). In published reports, however, agencies often did not cross-classify. In some cases they reported on the race categories with Hispanics included in the totals for the various races and then reported a separate total for Hispanics. This double counting makes it impossible from published reports to determine the numbers of "black" and "white" non-Hispanics or to classify Hispanics by racial category. Alternatively, agencies sometimes subtracted Hispanics from the race categories. Twenty-one systems had fields labeled "black not Hispanic" and "white not Hispanic," although they did not explicitly exclude Hispanics from the other race categories, including "other race."

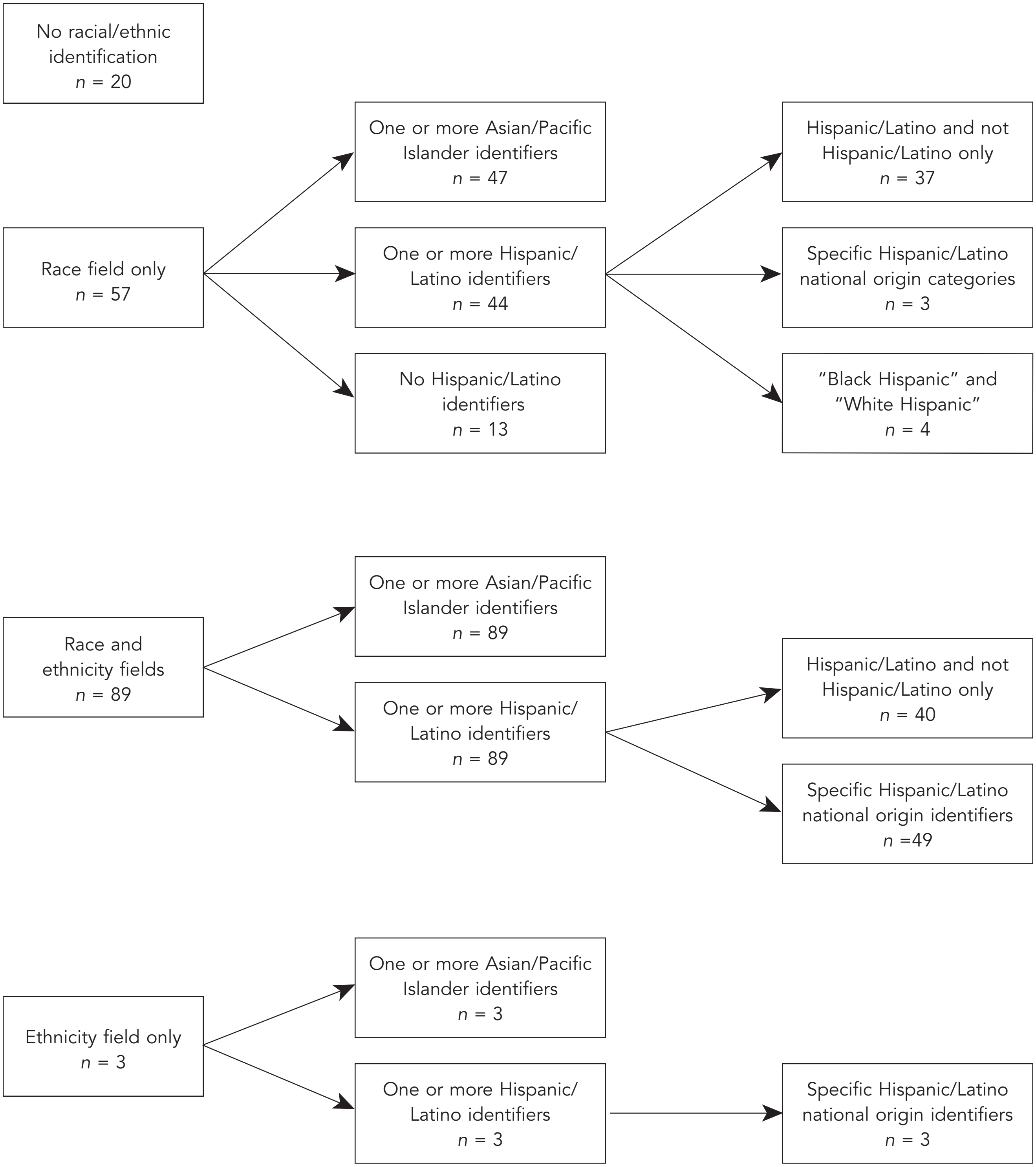
Fifty-seven systems used only a single race field. Two of these, the immunization registries for Maine and New Hampshire, which are jointly administered, presented an anomalous situation. In fact they each had a field labeled "race," which was never used. They also had a field labeled "ethnicity," which offered standard "race" labels (e.g., American Indian, black, Chinese, Japanese, white) and did not include any indication of Hispanic ethnicity. For tabulation purposes, we classified this as a race field. Three systems (1.8%) used a single ethnicity field, as discussed below.

In addition to black, white, and so on, the response categories in the race field in 44 of the 57 systems using only a single race field included some indication of Hispanic or Latino identification. Thirty-seven of these systems had only a single Hispanic category (which may have been labeled Hispanic/Latino). In three of these systems, Hispanics/Latinos were divided in some way by national origin. One offered a choice of Puerto Rican or "other Hispanic"; one offered Puerto Rican, Dominican, Central American, and "other"; and one offered Puerto Rican, Dominican, Salvadoran, Mexican, and Cuban. In four systems, the options were "white Hispanic" and "black Hispanic," with no other Hispanic/Latino option. In 13 systems, data were collected on race but no Hispanic, Latino, or related response option was offered.

Race labels

These systems used various sets of race or ethnicity labels. Some alternative wordings are considered synonymous by some observers, but not others (e.g., Native American and American Indian), while some pairs represent overlapping but not identical groups (e.g., African American and black).

Figure 1. Racial and ethnic identifiers used in 169 public health data systems in the six New England states: overall structure, use of Asian/Pacific Islander identifiers, and use of Hispanic/Latino identifiers



Of the 146 systems that had a race field, 115 (78.8%) used the label "black." Twenty-one used the label "black, not Hispanic," and one "African American, not Hispanic." Seventeen used the label "African American," although 14 of these allowed a choice of "black" or "African American," while three used "African-American" but not "black" as a race, and two used the U.S. Census label "Black, African American, Negro." Ten systems also allowed a choice of "Cape Verdean" as a race, and one system classified "African immigrant" as a separate race.

For the Asian and Pacific Islander group, the picture was even more complex. Eighty-one systems used the combined label "Asian/Pacific Islander." Twenty-five systems used "Asian," of which two also used the simple label "Pacific Islander" and one also had "Asian/Pacific Islander." Only 11 of these used "Native Hawaiian or other Pacific Islander," or some variant, as called for by the revised OMB standards. Two states' vital records systems specified "Native Hawaiian, including part Hawaiian," consistent with National Center for Health Statistics (NCHS) coding instructions for vital records systems.[33] This was the only case in which assignment to a single race classification was explicitly required for people who were only partially of a given heritage. The remaining 10 systems with an "Asian" specification had no Pacific Islander designation.

Of the remaining 40 systems that had a race field, 10 offered no Asian or Pacific Islander designations at all. The remainder used a list of Asian or Pacific/Islander nationalities plus an "other" category, rather than a generic "Asian or Pacific Islander" label or a choice between "Asian" and "Pacific Islander." These categories were highly variable. Thirty systems included Chinese, 28 included Japanese, 27 included Filipino, 13 included Vietnamese, and seven included Hmong (a Southeast Asian ethnic group). The most common Pacific Islander nationality specified was Hawaiian, used in 35 systems. Sometimes an "other" category was augmented by a list of examples, e.g., "Tongan, Tahitian, Fiji Islander, Chamorran, PI NOS." ("NOS" appeared frequently as an abbreviation for "not otherwise specified.")

The systems used a variety of categories for indigenous Americans. A bare majority (75/146) used the encompassing label "American Indian/Eskimo/Aleut." Fourteen used "Native American," and 35 used "American Indian," while 19 used some variation on "American Indian or Alaska Native," with rare twists such as "American Native/Alaska Native." "American Indian/Alaskan/Eskimo" also appears, as does "American Indian/Alaska Native/Eskimo/Aleut." Five systems—the Rhode Island vital records and work-related fatality surveillance systems—used the label "Indian (N, C, So. Am & Eskimo & Aleut)," specifically allowing for origin or ancestry from outside the U.S. No system permitted specific tribal identification, or an indication as to whether an individual was a member of a federally or state-recognized group.

The Connecticut death certificate had a free write-in field, but written responses were coded to a fixed set of categories, which were largely, but not entirely, consistent with the NCHS instructions for the race field code structure.[33] The NCHS race categories are similar to the U.S. Census categories, with one exception. While the Census maintains that Hispanics can be of any race, in the NCHS instructions "white" is defined as "includ[ing] Mexican, Puerto Rican, and other Caucasian." The Connecticut death registry followed NCHS instructions by including a general Asian or Pacific Islander category, with a list of specified subcategories.

The Hispanic origin categories, according to NCHS instructions, are Mexican, Puerto Rican, Cuban, Central or South American, and "other and unknown Hispanic." Written responses that are to be coded as white, according to NCHS include, among others, Bolivian, Colombian, Cuban, Brazilian, and Mexican, which represent Latin American countries with relatively small numbers of people of exclusively European heritage. "White" also includes South American, Latino, Latina, Chicano, and Latin American as well as various religious labels such as "Mohammedan," Muslim, Moslem, Islamic, Sunni, Zoroastrian, and "Jew." Consistent with Census practice, the Connecticut death registry coded people from Middle Eastern and North African countries as white. The registry's list of white race categories also included Brava (the name of a Somalian town and ethnic group), Crucian (people from the island of Saint Croix), and Ebian (which is the name of a region in Szechuan province, China). It also includes some terms we could not identify: "Bravo," and "Marshenese."

The white category in the Connecticut death registry also included Creole, which in Mexico means a person of Spanish descent, although elsewhere in Latin America and in the Southern U.S., it means a person of mixed African and European descent. "Mulatto" was to be coded as black. Also included in the list of black categories were some African nationalities, some Caribbean nationalities, and miscellaneous terms such as "quadroon," "octaroon," and "Hamitic."

The Connecticut death registry grouped together some of the terms in NCHS's "other entries" category, including Caribbean, Belizian, Guatemalan, Honduran, Nicaraguan, Panamanian, and Salvadoran, suggesting an attempt to separate out Caribbean Basin countries. However, also included in the same category in the death registry were the archaic term "Moor" as well as *Mestizo* (a Spanish term for mixed race), "mixed," and "Trigueno," apparently intended to be *Trigueño*, a Spanish-language term for a person with dark skin.

Finally, the system incorporated the remaining NCHS "other entries," which include Bahamian, British Honduran, Guyanese, and various unidentifiable terms that had apparently been encountered in the past, such as "Alocona," "Begri," and "Colestran," plus Mosotho, a South African ethnic group. In spite of the fairly lengthy list of "races" and Hispanic origin subgroups in the death registry database, Connecticut vital statistics reports used only white, black, and other "race" and Hispanic or not Hispanic "ethnicity."

Ethnicity labels

Of the 92 systems that had any ethnicity field (54.4% of all systems), 34 offered only Hispanic (and/or Latino) and not Hispanic (and/or Latino) as ethnic categories. Three systems used a Hispanic and/or Latino category and an "unknown ethnicity" category and hence were counted as having two ethnicity categories. Three additional systems allowed for coding a single national origin category along with a general Hispanic and/or Latino category—one used Mexican, one used Puerto Rican, and one, oddly, used "Spanish, Mexican, or Cuban descent." Others had longer lists of Latin American nationalities. The modal number was five nationalities (22 systems), and the maximum was 10 (one system).

Forty-eight systems had specific ethnicity categories other than "not Hispanic" (or an equivalent such as "not Hispanic/Latino"). Thirteen systems had African American as an ethnicity category. Six had "African immigrant," two used the label "African," and small numbers had specific African nationality or regional categories (two Nigerian, two Somalian, two North African). The most common of all non-Hispanic ethnicity categories was Cape Verdean, found in 21 systems, 20 of them in Massachusetts and one in Rhode Island. Seventeen systems—one in Rhode Island, the remaining 16 in Massachusetts—included Brazilian as an ethnicity category.

Various European or European-American national origin groups were used as ethnicity categories in small numbers of systems, although there was no discernible pattern to the different sets. For example, one system offered British, Irish, German, French, Polish, and Italian, while another offered French, Greek, Italian, and Irish, and yet another Canadian, Dutch, English, and French. Three subdivided Eastern European "ethnicity." Five systems in Rhode Island offered "other Western European," "other North European," "other Eastern European," and "other South European" along with some specific nationalities. New Hampshire vital records systems used 15 European and European American (e.g., French Canadian) "ancestry" categories, which we treated as ethnicity categories.

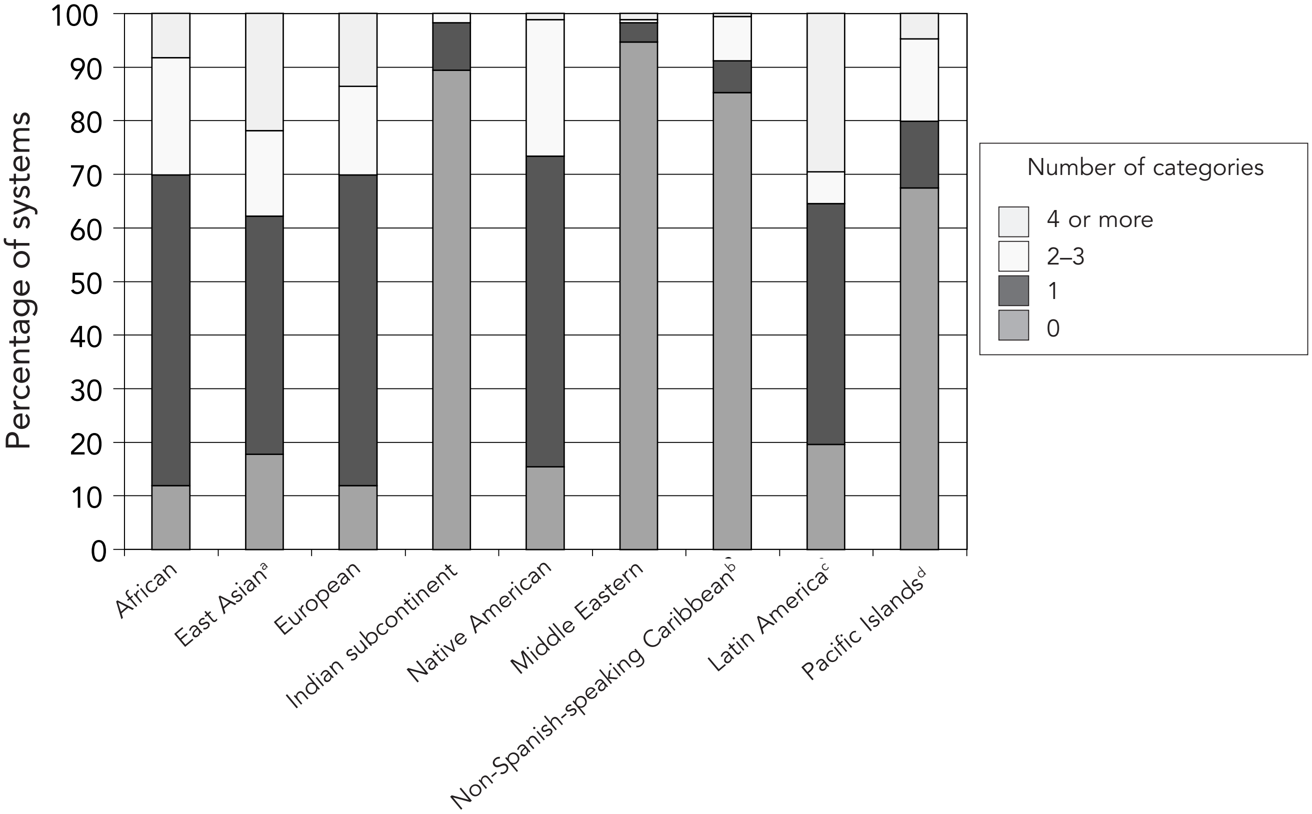
The three systems that had only an ethnicity field and no race field were all WIC program databases in Massachusetts and Rhode Island. The three systems used essentially the same set of 20 ethnicity categories, although the dictionary supplied by the Massachusetts system included more specific annotations. For example, all three of the systems had the labels Puerto Rican, Dominican, Central American, and "other Hispanic." The Massachusetts dictionary specified that Central American meant "Guatemala, El Salvador, Nicaragua, Costa Rica, Panama, Honduras, and Ecuador" and that "other Hispanic" meant "South America, Mexico, Cuba, Spain, etc." (Note that the assignment of Ecuador to Central America is erroneous.) The Massachusetts dictionary had apparently been updated more recently; it used the label "Russian," with the notation "includes all former Soviet Republics," while the Rhode Island system used the label "Soviet."

These ethnicity-only WIC systems handled various categories normally subsumed under race with no basic difficulty. Under the black heading, Massachusetts listed Haitian, West Indian (annotated as "Jamaica, Trinidad, Antigua, Barbados"), black American, and "other black" (annotated "Ethiopia, Somalia, Guyana, Surinam, Nigeria, South Africa, etc."). Under "white," the systems included "other East European" and "other white" in addition to "Russian" (or "Soviet"). "Other white" was annotated by Massachusetts as "North America, Europe, Middle East, Morocco, etc."

Figure 2 shows the distribution of racial and ethnic labels that appear to correspond to geographic regions of the world. For this analysis, we separated the Indian subcontinent (India, Pakistan, Bangladesh, Sri Lanka) from the rest of Asia. We included the generic Asian category, and the generic Asian/Pacific Islander categories with the East Asian countries, on the assumption that most of the population encompassed by these categories originates in East Asia. We assigned black or African American to our African descent category, which also includes Cape Verde, located off the coast of Africa, although many Cape Verdeans may not report their race as black or African American.

We grouped the non-Spanish-speaking countries of the Caribbean; Haiti, Barbados, and "other West Indian (incl. Jamaican)," "West Indian," and "other Caribbean/West Indian" were the labels actually used. We

Figure 2. Number of racial and ethnic labels corresponding to geographic regions of the world used in 169 public health data systems in the six New England states



[a]Includes "Asian/Pacific Islander" and "Asian" labels

[b]Includes Haiti, Barbados, and general "West Indian" and similar designations

[c]Includes Spanish-speaking countries of the Americas, and Brazil

[d]Does not include "Asian/Pacific Islander"; includes "Pacific Islander" and specific Pacific Islands nationalities, including Filipino

did not subsume this group into the African descent category; although many people from these countries trace their ancestry principally to Africa, not all do so.

Latin America includes Brazil, but does not include Spain, which we classified with the other European countries even though the OMB Standard definition of Hispanic includes Spain, its European possessions, and its non-Castilian-speaking ethnic groups..

Only the 20 systems that had no racial or ethnic identifiers failed to include at least black and white. Native Americans were completely unrepresented in 26 systems, and Asians (or Asians/Pacific Islanders) were unrepresented in 32. Latin America was unrepresented in 33 systems. Middle Eastern, non-Spanish-speaking Caribbean, and Indian subcontinent categories were seldom represented. Not counting the generic Asian/Pacific Islander label, Pacific Islanders were also unrepresented in most systems.

Quantitative analyses

Given such a bewildering variety of labels and approaches, quantitative analysis of these data may offer only limited insight. We found that Vital Records systems (n = 22) tended to have the largest number of race/ethnicity categories, a mean of 22.5 (SD = 10.1), followed by Health Status/Epidemiology systems (n = 8) with 17.75 (SD = 10.7). Health Care Utilization systems (mean = 9.2; SD = 8.5)—principally Hospital Discharge systems—and Surveillance systems (mean = 9.1; SD = 9.5) had the fewest. Vital Records systems used significantly more categories than Surveillance systems (p = 0.001), Health Care Utilization (p = 0.001), Other Utilization (p = 0.015), and Other systems (p = 0.014). Vital Records systems were not significantly different from Health Status/Epidemiology systems, nor were there any other significant differences.

Other findings

Data on U.S. vs. foreign birth were included in 39 (23.1%) of all systems. Only 35 systems (20.7%) reported having written protocols as to how race/ethnicity information was to be collected. Just over 60% (103) of systems reportedly included at least some racial and/or ethnic analysis in published reports.

DISCUSSION

These observations confirm the difficulties many commentators have noted with inconsistent and frequently unclear systems of classification. In general, data from different systems were not comparable. Categories could not be matched or "rolled up" to match. The attempts to map nationalities and ethnic groups onto "races" were often quite inappropriate, while the reliance on racial classification in most systems made it impossible to distinguish important immigrant communities, such as Brazilians, Cape Verdeans, Russians, Haitians, and Asian Indians, all of whom have substantial populations in New England, in particular in the southern New England states. Some systems in Massachusetts included such groups, but this was not generally the case in the other states.

Furthermore, it appears that many classification policies had been developed without sufficient study of the underlying social and geographic reality. Such practices as placing Ecuador in Central America, automatically coding Bolivians as white, assigning religions such as Moslem to the white category, or assuming that Hispanics must either be black or white suggest that many of these systems had developed with little serious reflection or study and without participation by the various groups who are affected.

A movement underway to standardize systems in accordance with the revised OMB standards may ameliorate some of the problems of comparability among systems. However, it is clear that the states have a massive challenge ahead of them to bring their systems into conformity. Standardization will also complicate and discourage efforts, such as those in Massachusetts, to include groups that cannot readily be forced into the OMB scheme, such as Brazilian immigrants and their descendants. It will also force groups such as Haitian and Cape Verdean immigrants and their descendants to be included in a single "black" category, whether or not that is appropriate or in any way meaningful for public health.

Although we applaud the move toward uniformity, we do not believe that the OMB's Revised Minimum Standards provide a rational basis for uniformity. While OMB makes it clear that the "race" concept is not scientifically valid and is to be understood as a social construction, its continued use as the basic structural framework for classifying the population reifies the concept and tends to legitimize it. It does not even provide a valid basis for understanding the effects of racism in society, because there is no evidence that the conflation of various nationalities with "races" corresponds to people's perceptions of race or to the likelihood that an individual will experience racist treatment. Are Arabs generally perceived as "white"? Is the impact of racism the same on fifth generation African Americans, on recent immigrants from Haiti or Cape Verde, and on people of mixed African and European heritage, all of whom are to be automatically coded as "black"? Are people of Pakistani, Tamil, Japanese, Vietnamese, and Filipino birth or descent (of whatever earlier heritage) perceived by the U.S. public as belonging to a single Asian "race"? There is no evidence for such conclusions. And there is a major practical problem in that some important immigrant groups, such as Brazilians, are simply impossible to classify as belonging to a single "race."

Making the construct of ethnicity the basis for classification avoids these problems. It is important to note that this would serve, among other benefits, to affirm and clarify African American ethnic identity, which at present is inappropriately conflated with black race. However, the current standards fail to make proper use of the concept. There is no evident reason why only people identified as Hispanic or Latino should be allowed to claim an ethnicity, while everyone else must somehow fit into a race. Furthermore, the very definition of "Hispanic or Latino" used in the revised standards is sociologically invalid. Latin American countries share important commonalities such as the predominance of Iberian languages, shared history of European conquest and colonialism, and the syncretism of European and American or African religions. But Spain is not part of Latin America, and there is no evident reason why Spaniards should be grouped with Mayans from Guatemala or Afro-Cubans, or why Brazilians should have no place in the scheme at all.

There is an urgent need to re-examine this question. The national commitment to eliminate health disparities cannot be honored unless disparities can be properly understood in terms of the sociological realities of "racial," "ethnic," and cultural diversity. Our present data systems, for the most part, do not do this, nor will the revised OMB standards represent progress in this regard.

We believe that systems based on an ethnic concept not limited to Hispanics/Latinos, such as the one used by the three Massachusetts and Rhode Island WIC databases, have the greatest underlying face validity. While these particular systems might be improved in specific ways in terms of the list of categories and their definitions, this approach does not pose any structural barriers to accurately reflecting people's self-identity and sociologically meaningful differences among communities.

For purposes of historical comparison, people could still be asked to pick a race, provided we are not afraid to accept "other" or "none" as answers. But we see no plausible argument for restricting the ethnicity concept to "Hispanics" or "Latinos." Rather, we propose creating a full menu of national origin/descent ethnicity categories, not mapped onto race. If people of any ethnicity are able to choose any race from a range of choices, as Hispanics now may do, the goals of understanding and eliminating health disparities will be better served.

The authors would like to thank Maria Elena Orejuela, MPH, of the Bureau of Health Statistics, Research and Evaluation of the Massachusetts Department of Public Health; Ratan Singh, PhD, MPH, Connecticut Department of Public Health; and William Walker, New Hampshire Department of Health and Human Services, for their assistance in facilitating data collection. The authors also thank research assistants Rachel Suazo and Fouad Pervez. They extend particular thanks to Bruce Cohen, PhD, of the Bureau of Health Statistics, Research and Evaluation of the Massachusetts Department of Public Health and Elliot Stone of the Massachusetts Health Data Consortium for comments on drafts of the manuscript and to Janet Scott Harris, Office of Minority Health, U.S. Department of Health and Human Services, for important support throughout the study process. This study was supported in part by a grant from the Jessie B. Cox Charitable Trust. A preliminary study was supported by the U.S. Office of Minority Health.

REFERENCES

1. Nobles M. History counts: a comparative analysis of racial/color categorization in U.S. and Brazilian censuses. Am J Public Health 2000;90:1738-45.
2. Schulman KA, Rubenstein LE, Chesley FD, Eisenberg JM. The roles of race and socioeconomic factors in health services research. Health Serv Res 1995;30:179-95.
3. Wallman KK, Evinger S, Schechter S. Measuring our nation's diversity: developing a common language for data on race/ethnicity. Am J Public Health 2000;90:1704-8.
4. Warren RC, Hahn RA, Bristow L, Yu ES. The use of race and ethnicity in public health surveillance. Public Health Rep 1994;109:4-6.
5. Hahn RA. The state of federal health statistics on racial and ethnic groups. JAMA 1992;267:268-71.
6. Office of Management and Budget (US). Revisions to the Standards for Classification of Federal Data on Race and Ethnicity. Fed Regist 1997;62:58781-90.
7. Friedman DJ, Cohen BB, Averbach AR, Norton JM. Race/ethnicity and OMB Directive 15: implications for state public health practice. Am J Public Health 2000; 90:1714-9.
8. Interagency Committee for the Review of Standards for Data on Race and Ethnicity, Tabulation Working Group. Provisional guidance on the implementation of the 1997 standards for federal data on race and ethnicity. Washington: Office of Management and Budget (US), Office of Information and Regulatory Affairs; 2000 Dec 3.
9. Bureau of the Census (US). Census 2000 redistricting data (Public Law 94-171) summary file: 2000 Census of Population and Housing: technical documentation. Washington: Census Bureau; 2001.
10. Department of Health and Human Services (US). Integrating public health information systems: a report and recommendations from the CDC/ATSDR Steering Committee on Public Health Information and Surveillance System Development. Atlanta (GA): DHHS; 1995.
11. Health Information and Surveillance Systems Board, Standards and Liaison Committee. Common data elements implementation guide version 2.4. Atlanta (GA): Centers for Disease Control and Prevention (US); 2000 Jan 3.
12. Iezzoni L. What health services researchers need to know about the Health Insurance Portability and Accountability Act. Med Care 1997;35:993-5.
13. Bureau of the Census (US). Overview of race and Hispanic origin: Census 2000 brief. March 2001 [cited 2002 Mar 8]. Available from: URL: http://www.census.gov/prod/2001pubs/c2kbr01-1.pdf
14. Goodman AH. Why genes don't count (for racial differences in health). Am J Public Health 2000;90:1699-702.
15. Weissman A. "Race-ethnicity": a dubious scientific concept [editorial]. Public Health Rep 1990;105:102-3.
16. Warren RC, Hahn RA, Bristow L, Yu ES. The use of race and ethnicity in public health surveillance. Public Health Rep 1994;109:4-6.
17. Schulman KA, Rubenstein LE, Chesley FD, Eisenberg JM. The roles of race and socioeconomic factors in health services research. Health Serv Res 1995;30:179-95.
18. Willis DP. Them and us: the politics of population taxonomy [editorial]. Am J Public Health 2001;91:1048-9.
19. Srinivasan S, Guillermo T. Toward improving health: disaggregating Asian American and native Hawaiian/Pacific Islander data. Am J Public Health 2000;90:1731-4.
20. Azuonye IO. Guidelines will encourage the thinking that underpins racism in medicine [letter]. BMJ 1996; 313:426.
21. Bhopal R. Is research into ethnicity and health racist, unsound, or important science? BMJ 1997;314:1751-6.
22. Rabin R. The use of race as a variable in public health research [editorial]. Am J Public Health 1999;89:783.
23. Buehler JW. Abandoning race as a variable in public health research [editorial]. Am J Public Health 1999; 89:783.

24. Szabo CP. Race as a research parameter [editorial]. S Afr Med J 1996;86:1559.
25. Krieger N. Counting accountably: implications of the new approaches to classifying race/ethnicity in the 2000 Census [editorial]. Am J Public Health 2000;90:1687-9.
26. Laws M. Race and ethnicity in biomedical and health services research [letter]. Arch Pediatr Adolesc Med 2001;155:972-3.
27. American Anthropological Association. American Anthropological Association response to OMB Directive 15 [cited 2002 Mar 8]. Available from: URL: http://www.aaanet.org/gvt/ombdraft.htm
28. Haynes MA, Smedley BD, editors. The unequal burden of cancer: an assessment of NIH research and programs for ethnic minorities and the medically underserved. Washington: National Academy Press; 1999.
29. Oppenheimer GM. Paradigm lost; race, ethnicity and the search for a new population taxonomy. Am J Public Health 2001;91:1049-55.
30. Thomas SB. The color line: race matters in the elimination of health disparities. Am J Public Health 2001; 91:1046-8.
31. Fullilove MT. Comment: abandoning "race" as a variable in public health research: an idea whose time has come. Am J Public Health 1998;88:1297-8.
32. Bureau of the Census (US). Census 2000 summary file 1: 100 percent data. Washington: Census Bureau; 2000.
33. National Center for Health Statistics (US), Division of Vital Statistics. Instruction manual. Part 4: classification and coding instructions for death records, 1999. Research Triangle Park (NC): NCHS; 1999.