RAND CORPORATION

When Race/Ethnicity Data Are Lacking

Using Advanced Indirect Estimation Methods to Measure Disparities

Allen Fremont, Joel S. Weissman, Emily Hoch, and Marc N. Elliott

Key findings

- Most health plans and delivery systems lack complete race/ethnicity data, hampering their efforts to routinely monitor health utilization, track quality measures by racial/ethnic group, and effectively target community-based interventions.
- RAND's indirect estimation method, known as Bayesian Improved Surname Geocoding (BISG), can produce accurate estimates of racial/ethnic disparities within populations served when self-reported data are lacking.
- The BISG method is intended to estimate differences at the group or population level; greater caution should be used in classifying specific individuals' race/ethnicity.

A key aim of U.S. health care reforms is to ensure equitable care while improving quality for all Americans. Limited race/ethnicity data in health care records hamper efforts to meet this goal. Despite improvements in access and quality, gaps persist, particularly among persons belonging to racial/ethnic minority and low-income groups.

For example, most health plans lack race/ethnicity data on most of their enrollees, making monitoring of racial/ethnic differences in care impractical. While other entities, such as hospitals or state Medicaid programs, collect race/ethnicity data on more of their patients, the information is often missing or unreliable. Consequently, despite health plans and delivery systems routinely monitoring and reporting quality measures for their overall population, tracking quality scores by racial/ethnic group is uncommon.

Health plans and delivery systems are working to collect self-reported race/ethnicity information for all persons they serve, but it will take several more years, at least, to fully populate their files.

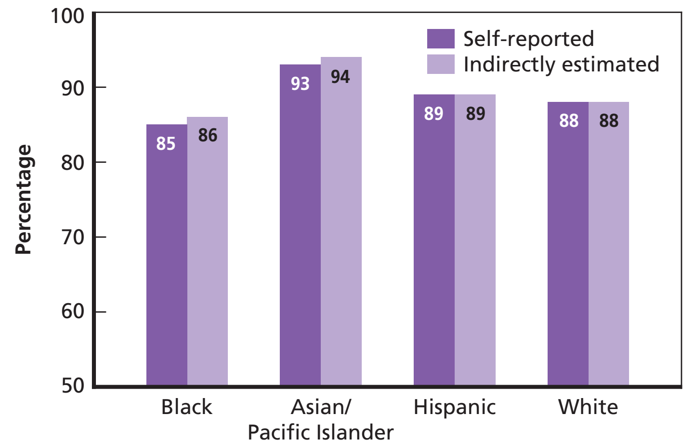
NEWER INDIRECT ESTIMATION METHODS

Recent advances in estimating race/ethnicity and improvements in accuracy have led the Institute of Medicine (IOM) and other entities to recommend using indirect estimation methods to produce probabilistic estimates when self-reported race/ethnicity data are not available for monitoring health care utilization.

RAND Corporation's method, known as Bayesian Improved Surname Geocoding (BISG), exemplifies newer estimation approaches recommended by the IOM. The method is optimized to produce accurate and reliable estimates at the group or population level rather than classifying each individual as belonging to a specific racial/ethnic group. It combines information about a person's likely race/ethnicity from the U.S. Census Surname list with information about the racial/ethnic composition of the neighborhood (i.e., Census Block Group) they live in. BISG can be used with any administrative data files that include a person's last name (i.e., surname) and residential address.

The method produces a set of probabilities that a given person belongs to each of six mutually exclusive racial/ethnic groups: Black, Asian/Pacific Islander, Hispanic, White, American Indian/Alaska Native, and Multiracial. The probabilities for each person can then be added or averaged across any subgroup of patients to produce counts and proportions, respectively, and thus can be used to compare performance rates on quality, cost, or access measures between different racial/ethnic groups. BISG probabilities can also be used with standard statistical methods, such as multivariate regression analyses, to examine the relationship between race/ethnicity and quality of care when such other factors as age, gender, and income are taken into account.

Figure 1. Differences in HbA1C Testing Among Diabetic Members, by Race/Ethinicity



RAND *RR1162-1*

ACCURACY OF NEWER ESTIMATES

Numerous validation studies have shown that BISG and related methods have an excellent ability to measure race/ethnicity. Concordance between self-reported race/ethnicity and BISG estimates is typically 90 to 96 percent for the four largest racial/ethnic groups—Blacks, Asians/Pacific Islanders, Hispanics, and Whites. Estimates are less accurate and not recommended, however, for American Indians/Alaskan Natives and Multiracial persons, due to small numbers.

BISG estimates also compare well to self-reported race/ethnicity with respect to group-level results. For instance, the percentage of health plan members belonging to each of the four major racial/ethnic groups is similar whether self-reported race/ethnicity or BISG estimates are used to calculate percentage. BISG performs even better when estimating disparities, its primary intended purpose. Disparities calculated based on self-reported race/ethnicity and BISG estimates typically are within a few percentage points in size.

For example, Figure 1 compares rates at which diabetic patients belonging to the four major racial/ethnic groups received a recommended test in a commercial HMO plan. Regardless of whether self-reported race/ethnicity or indirect estimates for those patients are used, the calculated rates and gaps between different racial/ethnic groups are essentially the same.

The most accurate estimates of overall disparities or racial/ethnic composition are obtained by combining, rather than replacing, available race/ethnicity with indirect estimates. For instance, if an individual self-reports his or her race/ethnicity as Black (non-Hispanic), the estimates would be revised to show that this individual has a probability of "1" (i.e., 100-percent probability) of being Black, and a probability of "0" (i.e., no chance) of belonging to the other race/ethnicity groups, and so on.

USES OF INDIRECTLY ESTIMATED RACE/ETHNICITY

A growing number of health care organizations have begun routinely using methods like BISG to add or refine race/ethnicity information in their data files. The primary users thus far have been large national and regional health plans (e.g., Anthem, Cigna, Highmark, Kaiser) though some public agencies, including the Center for Medicare and Medicaid Services (CMS), and various research organizations also use BISG.

BISG estimates can avoid a potential bias when calculating performance measures based only on the subset of patients with self-reported race/ethnicity, since the subset may differ from other patients in ways that impact disparities. For instance, many plans obtain race/ethnicity information when members fill out optional health risk assessments on their online patient portal. Minority patients who use the portal and fill out risk assessments tend to be more affluent, educated, and likely to obtain good care.

Experience from these efforts has shown the estimates can generally be used in the same ways as race/ethnicity from self-reported or other direct sources for group or population level assessments of care (see Table 1).

RAND investigators and other experts discourage use of probabilistic estimates to classify a specific individual's race/ethnicity when calculating disparities or group comparisons,

Table 1. Example Uses of Self-Reported and BISG Race/Ethnicity Estimates

Example Use Cases	Self-Report or Other Direct Source	BISG Estimates
Racial/ethnic composition of a patient population	X	X
Racial/ethnic differences in care quality and outcomes	X	X
Community-level outreach and interventions	X	X
Comparative effectiveness of interventions	X	X
Classification of individuals' race/ethnicity in electronic medical records	X	–

since the overall accuracy of estimates decreases, both in terms of efficiency and bias. Under limited circumstances, it may be appropriate to target individuals with certain probabilities for follow-up, ideally after confirmation of their race/ethnicity.

Though newer indirect estimation methods have primarily been used thus far within individual health plans, the approach can also be used for assessments of racial/ethnic differences in care across plans at the national or regional level.

For example, as part of their oversight of Medicare beneficiaries' care, CMS maintains a file with member-level scores on the Healthcare Effectiveness Data and Information Set (HEDIS) quality of care measures from over 400 Medicare Advantage plans. Though Medicare administrative files include race/ethnicity information on most beneficiaries, that information is frequently incorrect for Hispanics and Asians/Pacific Islanders, making the data unsuitable for comparing HEDIS scores by race/ethnicity.

CMS worked with RAND investigators to improve race/ethnicity information in the administrative files using an extension of the BISG method that incorporates additional administrative information to estimate plan-level HEDIS performance by race/ethnicity.

Similarly, as part of their mission to monitor and ensure affordability, access, and quality care for all residents, the Massachusetts Center for Health Information and Analysis (CHIA) maintains an All Payer Claims Database (APCD) that includes health care data on 90 percent of the state's residents. Though the Massachusetts APCD is among the most comprehensive in the country, race/ethnicity information is missing for most residents, since that information is provided by health plans.

Accordingly, meaningfully monitoring how some key aspects of health care vary between Massachusetts residents belonging to different racial/ethnic groups has been challenging. To help overcome this barrier, CHIA collaborated with RAND to assess the feasibility of applying BISG to the APCD. Preliminary results were promising, with estimates produced from APCD files corresponding well with reliable plan race/ethnicity that is available, Massachusetts death certificate data, and population data from the U.S. Census. More work is needed, however, to fully assess unique issues related to implementation, particularly ensuring appropriate use of the estimates in the APCD.

RELATED APPLICATIONS OF INDIRECT ESTIMATION

Although obtaining race/ethnicity information is an essential first step for identifying disparities, additional information and decision tools can help decisionmakers understand and more effectively target disparities in care. For example, some of the observed racial/ethnic disparities may relate to differences in socioeconomic status, health literacy, or language barriers. Because indirect estimation incorporates Census information about patients' neighborhoods, it is relatively easy to add information about these ecological (or geographic) characteristics (e.g., poverty rates) that—like race/ethnicity—may be missing from many health care databases. If implemented carefully, such information can be added without violating privacy concerns.

Knowing the neighborhood patients live in also allows mapping of how patterns of quality of care and distribution of patients from different racial/ethnic groups varies within and between different service areas. By identifying potential geographic "hotspots" and exploring patient characteristics of those within and outside of the area, decisionmakers can better understand contributing factors and how to effectively target interventions.

For instance, GIS mapping analyses RAND conducted with a large health plan revealed a community-level hotspot where members received quality of care scores significantly lower than in surrounding areas, regardless of members' race/ethnicity and even lower for Hispanics within the hotspot. Further analyses showed that patients within the area had lower income, education, and were more linguistically isolated on average than those in surrounding areas, where quality was better.

CONCLUSION

Advances in methods for estimating race/ethnicity are enabling health plans and other health care organizations to overcome a long-standing barrier to routine monitoring and actions to reduce disparities in care. Though these new estimation methods are promising, practical knowledge and guidance on how to most effectively apply newly available race/ethnicity data to address disparities can be greatly extended. Thus, in addition to continuing to make refinements to the method to improve accuracy and ease of use, RAND investigators are working on ways to estimate additional types of sociodemographic data when it's missing from health care data files. We are also exploring ways that health care decisionmakers can employ emerging data visualization and decision optimization tools to help them more effectively and efficiently target efforts to ensure the equitable, high-quality care for the diverse populations they serve.

REFERENCES

Adjaye-Gbewonyo, D., R. Bednarczyk, R. Davis, and S. Omer, "Using the Bayesian Improved Surname Geocoding Method (BISG) to Create a Working Classification of Race and Ethnicity in a Diverse Managed Care Population: A Validation Study," *Health Services Research*, Vol. 49, No. 1, 2014, pp. 268–283.

Derose, S., R. Contreras, K. J. Coleman, C. Koebnick, and S. Jacobsen, "Race and Ethnicity Data Quality and Imputation Using US Census Data in an Integrated Health System—The Kaiser Permanente Southern California Experience," *Medical Care Research and Review*, Vol. 70, No. 3, 2013, pp. 330–345.

Elliott, M., A. Fremont, P. Morrison, P. Pantoja, and N. Lurie, "A New Method for Estimating Race/Ethnicity and Associated Disparities Where Administrative Records Lack Self-Reported Race/Ethnicity," *Health Services Research*, Vol. 43, No. 5p1, 2008, pp. 1722–1736.

Elliott, M., P. Morrison, A. Fremont, D. McCaffrey, P. Pantoja, and N. Lurie, "Using the Census Bureau's Surname List to Improve Estimates of Race/Ethnicity and Associated Disparities," *Health Services and Outcomes Research Methodology*, Vol. 9, No. 2, 2009, pp. 69–83.

Elliott, M., K. Becker, M. Beckett, K. Hambarsoomian, P. Pantoja, and B. Karney, "Using Indirect Estimates Based on Name and Census Tract to Improve the Efficiency of Sampling Matched Ethnic Couples from Marriage License Data," *Public Opinion Quarterly*, Vol. 77, No. 1, 2013, pp. 375–384.

Elliott, M., A. Haviland, J. Adams, D. McCaffrey, A. Fremont, and N. Lurie, "Using Indirect Estimation to Improve CMS Administrative Information on Race/Ethnicity and Estimate Plan-Level HEDIS Performance by Race/Ethnicity," briefing slides, June 28, 2010. As of June 25, 2015: http://www.academyhealth.org/files/2010/monday/elliott.pdf

Fiscella, K., and A. Fremont, "Use of Geocoding and Surname Analysis to Estimate Race and Ethnicity," *Health Services Research*, Vol. 41, No. 4p1, 2006, pp. 1482–1500.

Grundmeier, R., L. Song, M. Ramos, A. Fiks, M. Elliott, A. Fremont, W. Pace, R. Wasserman, and R. Localio, "Imputing Missing Race/Ethnicity in Pediatric Electronic Health Records: Reducing Bias with Use of US Census Location and Surname Data," *Health Services Research*, in press, 2015.

High-Value Health Care Project, *Moving Toward Racial and Ethnic Equity in Health Care*, March 2011.

Institute of Medicine, *Race, Ethnicity, and Language Data: Standardization for Health Care Quality Improvement*, Washington, D.C.: National Academies Press, 2009.

Lurie, N. and A. Fremont, "Building Bridges Between Medical Care and Public Health," *JAMA*, Vol. 302, No. 1, 2009, pp. 84–86.

Lurie, N., A. Fremont, S. Somers, K. Coltin, A. Gelzer, R. Johnson, W. Rawlins, G. Ting, W. Wong, and D. Zimmerman, "The National Health Plan Collaborative to Reduce Disparities and Improve Quality," *Joint Commission Journal on Quality and Patient Safety*, Vol. 34, No. 5, 2008, pp. 256–265.

Martino, S., R. Weinick, D. Kanouse, J. Brown, A. Haviland, E. Goldstein, J. L. Adams, K. Hambarsoomian, D. Klein, and M. Elliott, "Reporting CAHPS and HEDIS Data by Race/Ethnicity for Medicare Beneficiaries," *Health Services Research*, Vol. 48, No. 2p1, 2013, pp. 417–434.

RAND Health, "Q-DART: Innovative Solutions to Target Gaps in Health Care Quality & Health Outcomes," website, undated. As of January 27, 2016: http://www.rand.org/health/projects/qdart.html

Weissman, J. S., and R. Hasnain-Wynia, "Advancing Health Care Equity Through Improved Data Collection," *New England Journal of Medicine*, Vol. 364, No. 24, 2011, pp. 2276–2277.

About This Report

A key aim of U.S. health care reforms is to ensure equitable care while improving quality for all Americans. Limited race/ethnicity data in health care records hamper efforts to meet this goal. Despite improvements in access and quality, gaps persist, particularly among persons belonging to racial/ethnic minority and low-income groups.

This report describes the use of indirect estimation methods to produce probabilistic estimates of racial/ethnic populations to monitor health care utilization and improvement. One method described, called Bayesian Indirect Surname Geocoding, uses a person's Census surname and the racial/ethnic composition of their neighborhood to produce a set of probabilities that a given person belongs to one of a set of mutually exclusive racial/ethnic groups.

Advances in methods for estimating race/ethnicity are enabling health plans and other health care organizations to overcome a long-standing barrier to routine monitoring and actions to reduce disparities in care. Though these new estimation methods are promising, practical knowledge and guidance on how to most effectively apply newly available race/ethnicity data to address disparities can be greatly extended.

This report is part of Q-DART. The Q-DART project applies emerging analytic tools to better target gaps in the quality of care and health outcomes in diverse populations, helping decisionmakers more wisely allocate scarce resources. Q-DART uses an array of tools to generate data on disparities and report the data in a variety of visually compelling ways for use by health plans, public health organizations, and others concerned about improving the care that people receive. Q-DART is part of RAND Health, a division of the RAND Corporation. For more information about Q-DART, visit www.rand.org/health/projects/qdart.html.

For more information on this publication, visit www.rand.org/t/rr1162.

www.rand.org

RAND CORPORATION

The RAND Corporation is a research organization that develops solutions to public policy challenges to help make communities throughout the world safer and more secure, healthier and more prosperous. RAND is nonprofit, nonpartisan, and committed to the public interest.

RAND's publications do not necessarily reflect the opinions of its research clients and sponsors. **RAND**® is a registered trademark.

RR-1162-COMMASS