By Latrice G. Landry, Nadya Ali, David R. Williams, Heidi L. Rehm, and Vence L. Bonham

DOI: 10.1377/hlthaff.2017.1595
HEALTH AFFAIRS 37,
NO. 5 (2018): 780–785

Lack Of Diversity In Genomic Databases Is A Barrier To Translating Precision Medicine Research Into Practice

Latrice G. Landry (Latrice_Landry@hms.harvard.edu) is a fellow in the Laboratory for Molecular Medicine, Partners Personalized Medicine, in Cambridge, Massachusetts, and in the Office of Minority Health, Food and Drug Administration, in Silver Spring, Maryland.

Nadya Ali is an MD candidate in the Michigan State University College of Human Medicine, in East Lansing.

David R. Williams is a professor in the Department of Social and Behavioral Sciences, Harvard T. H. Chan School of Public Health, in Boston, Massachusetts, and in the Department of African and African American Studies, Harvard University, in Cambridge.

Heidi L. Rehm is the chief genomic officer in the Center for Genomic Medicine and Department of Medicine at Massachusetts General Hospital, in Boston, and medical director of the Broad Institute Clinical Research Sequencing Platform, in Cambridge.

Vence L. Bonham is an associate investigator in the Social and Behavioral Research Branch, Division of Intramural Research, and senior adviser to the director on genomics and health disparities at the National Human Genome Research Institute, National Institutes of Health, in Bethesda, Maryland.

ABSTRACT **Precision medicine is predicted to revolutionize the clinical practice of medicine, in part by using molecular biomarkers to assess patients' risk, prognosis, and therapeutic response more precisely. However, reliance on biomarkers could present challenges for diverse populations that are not equitably represented in precision medicine research. We examined the populations included in genomic studies whose data were available in the following two public databases: the Genome-Wide Association Study Catalog and the database of Genotypes and Phenotypes. We found significantly fewer studies of African, Latin American, and Asian ancestral populations in comparison to European populations. These patterns were consistent across both data types and disease areas. While the number of genomic research studies that include non-European populations is modestly improving, the overall numbers are still low, and decisive action is needed now to implement the changes necessary for realizing the promise of precision medicine for all.**

The success of the evolving field of precision medicine has been driven by the evolution of science and technology, which has enabled the sequencing of the human genetic code; the development of bioinformatic tools to process the vast amounts of genomic data generated; and the creation of databases that curate, organize, and share the information. Genomic databases are important and serve as a repository for shared knowledge related to the numerous research studies and clinical case reports that contribute to knowledge of the impact of genetics on human health, as well as a basic understanding of genetic differences between human beings. However, the majority of studies that contribute to this knowledge are based on populations of European ancestry, providing reasonable genetic representation of individuals of European ancestry in databases but poorer representation of other ethnic populations.[1–6]

The underrepresentation of non-European populations in genomic databases is problematic because it may miss gene-disease relationships for which the exposure or outcome is rare in European populations, it limits the generalizability of findings from genomic research, and it limits the evidence base for translating these findings into clinical care in diverse populations.[7,8] Understanding the extent of this underrepresentation is important so that we can address it as we solidify the foundation that will support precision medicine in the future and ensure the applicability of precision medicine for the global population.

In the evidence-based practice of precision medicine, sampling bias in research upstream has the potential to propagate bias in clinical translation downstream. Many researchers are concerned that the lack of inclusion of diverse populations in genomic research will cause problems for the translation of that research into the clinical practice of precision medicine.[1–6] In a comment in *Nature* in 2011,[1] Carlos Bustamante and colleagues drew the research community's attention to a 2009 study by Anna Need and David Goldstein,[6] who had reported that 96 percent of all genomewide association studies were of people of European descent. There have been improvements since then: In 2016 Alice Popejoy and Stephanie Fullerton reported that 81 percent of genomewide association studies were of people of European descent.[2] Of the 19 percent of the studies that focused on non-Europeans, 14 percent focused on populations of Asian descent, indicating clear improvements for those populations, but very little progress for other non-European groups. A 2016 study of data in the Cancer Genome Atlas repository found that of 5,729 tumor samples, 77 percent were from whites, 12 percent from blacks, and 3 percent from Asians.[9] In contrast, fewer than 16 percent of the world's population is of European descent, which suggests that new diversity inclusion strategies are needed before genomic data are representative of the global population.[10,11]

Underrepresentation in genomic databases and repositories is paralleled by underuse of genetic services in diverse populations. This underuse has been reported globally by the World Health Organization,[12,13] as well as domestically, in genetic services—including the ordering of genetic tests and genetic counseling.[14–17]

We conducted a comparison of the numbers of studies in the Genome-Wide Association Study Catalog, as well as the numbers of high-throughput sequencing studies in the database of Genotypes and Phenotypes, by ancestral population and disease area. By identifying disparities in genomic information by disease area, we sought to highlight which patient populations and disease areas were least represented.

Study Data And Methods

DATA We used data from two public sources of genomic information developed by the National Institutes of Health (NIH): the Genome-Wide Association Study Catalog, a curated catalogue of published genomewide association studies, and the database of Genotypes and Phenotypes, a data repository of genomic data sets (whether or not they were funded by the NIH) from which we obtained data capturing summary information on sequencing studies.[18,19]

METHODS We downloaded the Genome-Wide Association Study Catalog as of May 8, 2017, and the database of Genotypes and Phenotypes sequencing studies as of February 27, 2017, and categorized the studies by ancestral group and disease focus. The ancestral groups analyzed were European, Asian, and underrepresented minorities (defined as African or African American, Native American, and Hispanic or Latino—groups that were aggregated due to small sample sizes). European ancestry was the reference group for analysis purposes. We excluded studies whose ancestral groups were not defined. We also excluded multiethnic studies, meaning studies with participants from more than one ethnicity, as we had no clear or consistent way to capture the percentage of participants from specific ancestral populations within a multiethnic study. Additionally, we stratified disease categories by cancer or noncancer outcomes.

LIMITATIONS Our study had several limitations. First, the Genome-Wide Association Study Catalog and the database of Genotypes and Phenotypes might not capture all genomic studies. Second, although the catalog is professionally curated by a team whose members examine all published genomewide association studies, the database is a public repository of voluntarily submitted sequencing data and may be a less exhaustive list of studies. Third, our study does not report the number of participants in each study, only the number of studies. Lastly, these analyses represent the data available at the time of download. Newer entries are not reflected.

Study Results

GENOME-WIDE ASSOCIATION STUDY CATALOG We reviewed 2,817 genome-wide association studies from the catalog, of which 413 were cancer-related studies (exhibit 1). Of those studies, 67 percent were based on populations of European descent, 29 percent on populations of Asian descent, and 4 percent on populations from underrepresented minority groups. Of the 52 cancer studies that focused on hematologic or lymphatic cancer and the 47 that focused on cancer of the digestive tract, none were studies exclusively of underrepresented minority groups.

The remaining 2,404 genome-wide association studies were not related to cancer. Of those studies, 71 percent were of populations of European descent, 20 percent were of Asian populations, and 8 percent were of underrepresented minority groups. Fewer than 5 percent of the studies of gastrointestinal, reproductive system, or neurologic disease were of underrepresented minority groups.

DATABASE OF GENOTYPES AND PHENOTYPES The number of sequencing studies from the database of Genotypes and Phenotypes in our sample was limited. We excluded 394 studies because they contained no documentation of participants' ancestry, and we excluded 46 multiethnic studies for the purpose of this analysis, as explained above. Of the 113 sequencing studies analyzed (exhibit 2), twenty-three were focused

EXHIBIT 1

Genome-wide association studies, by disease area and study population demographic group, 2017

	Demographic group		
Disease area	**Europeans (%)**	**Asians (%)**	**Underrepresented minorities (%)**
Any type of cancer (*n* = 413)	67	29	4
Breast cancer (*n* = 59)	61	34	5
Gastrointestinal cancer (*n* = 47)	53	47	0
Lung cancer (*n* = 55)	44	49	7
Kidney cancer (*n* = 19)	84	11	5
Blood cancer (*n* = 52)	56	44	0
Reproductive cancer (*n* = 120)	81	13	6
Other cancers (*n* = 61)	82	16	2
Any noncancer disease or disorder (*n* = 2,404)	71	20	8
Cardiovascular disease (*n* = 219)	68	20	12
Neurologic disease (*n* = 418)	82	14	4
Respiratory disease (*n* = 111)	77	12	11
Gastrointestinal disease (*n* = 131)	78	21	1
Reproductive system disease (*n* = 93)	70	26	4
Kidney disease (*n* = 56)	70	23	7
Blood disorder (*n* = 155)	71	17	12
Other diseases (*n* = 1,221)	68	23	9

SOURCE Authors' analysis of data from the Genome-Wide Association Study Catalog. **NOTES** Demographic groups are defined in the text. Rows might not add to 100 percent because of rounding.

on cancer (62 percent Europeans, 15 percent Asians, and 23 percent underrepresented minorities). Of the noncancer studies, eleven were on cardiovascular disease (55 percent Europeans, 18 percent Asians, and 27 percent underrepresented minorities), thirteen on neurologic diseases (85 percent Europeans and 15 percent Asians), and fourteen on respiratory diseases (57 percent Europeans and 43 percent underrepresented minorities). Two of the disease areas

EXHIBIT 2

Genotypes and phenotypes studies, by disease area and study population demographic group, 2017

	Demographic group		
Disease area	**Europeans (%)**	**Asians (%)**	**Underrepresented minorities (%)**
Any type of cancer (*n* = 23)	74	9	17
Breast cancer (*n* = 2)	50	0	50
Gastrointestinal cancer (*n* = 3)	100	0	0
Lung cancer (*n* = 3)	33	67	0
Kidney cancer (*n* = 1)	100	0	0
Blood cancer (*n* = 1)	100	0	0
Reproductive cancer (*n* = 3)	33	0	67
Other cancers (*n* = 10)	90	0	10
Any noncancer disease (*n* = 90)	63	10	27
Cardiovascular disease (*n* = 11)	55	18	27
Neurologic disease (*n* = 13)	85	15	0
Respiratory disease (*n* = 14)	57	0	43
Gastrointestinal disease (*n* = 2)	100	0	0
Other diseases (*n* = 50)	60	10	30

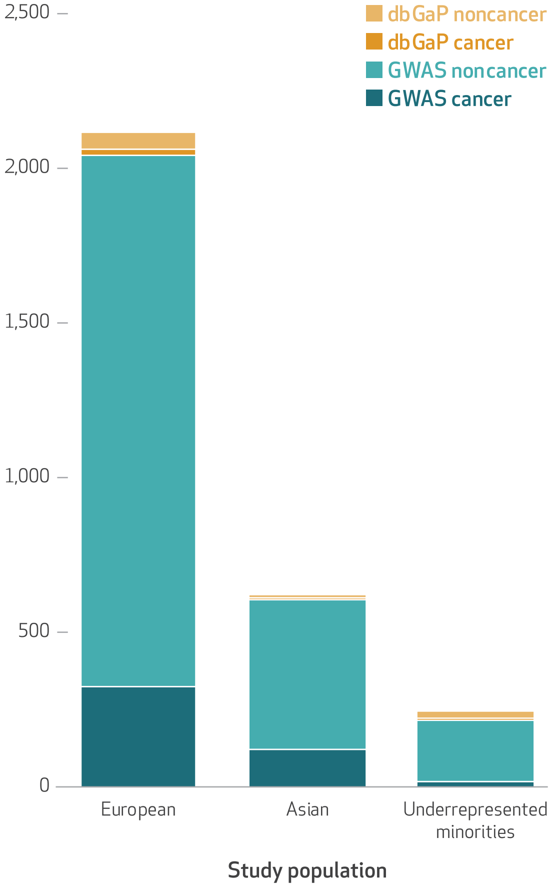
SOURCE Authors' analysis of data from the database of Genotypes and Phenotypes. **NOTE** Demographic groups are defined in the text.

had only one sequencing study. In each case in which there was only one ancestral population study in a disease area, the study population was of European ancestry. There were no sequencing studies in the database that focused on neurologic disease solely in underrepresented minority populations, respiratory diseases in Asian populations, or gasterointestinal disease in either Asians or underrepresented minorities.

In addition to the total number of studies, we observed the biggest difference in numbers of studies across ancestral populations in noncancer studies in the Genome-Wide Association Study Catalog (exhibit 3). These results do not report numbers of participants in each study. However, the European population studies were larger, and more European studies had the statistical power to conduct robust genomic research.

EXHIBIT 3

Numbers of genomewide association studies and genotype and phenotype studies, by disease area and study population demographic group, 2017



SOURCE Authors' analysis of data from the Genome-Wide Association Study Catalog and the database of Genotypes and Phenotypes (dbGaP). **NOTES** Underrepresented minorities are explained in the text. GWAS is genome-wide association study.

EXCLUDED STUDIES Though our analysis excluded multiethnic studies, studies with populations not included in the demographic groups we focused on, and studies with populations of unknown descent, we provide summaries of the numbers and disease focuses of these studies in online appendix exhibits 1A and 1B.[20]

Discussion

Sequencing—specifically high-throughput sequencing—is rapidly becoming the dominant technology for genomics research and in molecular diagnostics. In this study we found a lack of diversity not only in genomewide association studies, but also in sequencing studies in the database of Genotypes and Phenotypes. In both databases, the number of studies for individual underrepresented groups was so small that we aggregated them into a single group, which focuses the discussion on the aggregate as apposed to the distinct ancestral groups. Additionally, our analyses showed differences across ancestral groups in the numbers of studies by disease area. One of the most striking findings is the lack of ancestral information included in these data sets, specifically in the database of Genotypes and Phenotypes. Ancestral information is important to the clinical use of genetic information, and thus its inclusion in these databases is of the utmost importance.

We found that several disease areas had no studies of underrepresented minorities or Asian populations. Because the database of Genotypes and Phenotypes is a repository of data sets available for secondary analysis, improving minority representation in that database may decrease bias in future studies that rely on secondary data analysis.

These databases allow precision medicine leaders to track the progress of diversity inclusion in research and can potentially serve as a tool in developing national research priorities. However, when determining which disease areas and populations to focus on in precision medicine research, data on underrepresentation in research (which we provide here) ideally would be accompanied by information on disease prevalence; disparities in disease morbidity or mortality; and pertinent genetic factors such as penetrance (the probability that a variant will lead to phenotypic expression), expressivity (variation in phenotypic expression when a variant is penetrant), and number of variants across ancestral groups.

We identified several clinical and research priorities to address underrepresentation in precision medicine research. Clearly, there is a persistent need to include diverse populations in research, which would reduce sampling bias, enhance knowledge about human genetic variation, and increase the generalizability of genetic research findings. The need applies to all study designs, including case series, observational studies, case control studies, and clinical trials- Increasing diversity in studies will require multiple strategies. Policies and guidance to promote inclusion in research have already begun to emerge, such as targeted requests for applications, an NIH policy on including women and members of minority groups in research,[21] and guidance from the Food and Drug Administration on collecting and reporting race and ethnicity information in clinical trials.[22] We could envision editorial boards specifying certain inclusion standards, or justification for the lack of diverse sample populations, as a requirement for publication. However, one major barrier to enhancing diversity in trial participation is that the established system for and approaches to recruiting participants might not be generalizable to all. Thus, innovative and culturally respectful strategies for the recruitment of underrepresented groups in research are required.[23,24]

It has been suggested that access to care, access to health insurance, and socioeconomic status may affect participation in genomic research and the use of genomic services.[13,16] However, in a study of direct-to-consumer genetic testing, in which consumers in populations that received testing were similar in educational attainment, socioeconomic status, and insurance, the number of non-European consumers participating in research was significantly lower than the number of European consumers.[25] As the field seeks to increase participation in research, as well as to increase uptake of precision medicine, a multilayered approach with patient, provider, and researcher education at the center is needed. This must entail educating the public about the value of research participation, including the importance of contributing biological samples to biobanks.[26]

The routine collection of information on genetic ancestry (both self-described and biological), as well as on socioeconomic status is important[27] and could also occur at the patient bedside. For instance, genetic data could be collected in patients' electronic health records and incorporated into decision-making tools within those records.

In addition, information on gene-disease associations and individual genetic variants must be shared with patients in an accessible way. This information and subsequent interpretation are presented to patients and providers through the genomic report (the report of the genetic test results). Therefore, it is important that the report be sufficiently approachable and useful for diverse patient populations.

Americans must make a commitment to the equitable diffusion of precision medicine into clinical care.

Conclusion

Americans must make a commitment to the equitable diffusion of precision medicine into clinical care. The factors that create inequality in the uptake of precision medicine are multifaceted and related to the individual, the provider, insurance coverage, health care system and care infrastructure, and genomic data infrastructure (including the populations included in research, documentation of ancestral populations in genomic databases, as well as access to and use of genomic data in health care). The translation of genomic information into clinical tests for diagnostic, prognostic, therapeutic, and disease-monitoring purposes is an evidence-based process. Ensuring the safety and efficacy of precision medicines creates the need for a sizable amount of evidence in the form of genomic research and clinical case reports for genetic test development and use. The lack of diversity in genomic research can affect the understanding of the relationships between genes and disease in unstudied populations including, erroneous rare variant-disease associations in poorly studied populations, and insufficient evidence regarding the effect of variants on disease in diverse populations. Given the importance of genomic databases for both genomic knowledge dissemination and clinical translation, it is important that prioritization be given to the inclusion of diverse ancestral populations in the data supplying these databases and the documentation of ancestral information in them. ■

The authors thank the Food and Drug Administration (FDA) and its staff members Martin Mendoza, Jonca Bull, and David Litwack. This research was supported in part by a fellowship appointment for Latrice Landry through the FDA Office of Minority Health, administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the Department of Energy and the FDA. Additional support was provided in part by the Intramural Research Program of the National Human Genome Research Institute (NHGRI), National Institutes of Health (NIH). The opinions expressed are the authors' own and do not reflect the policies or positions of the FDA, NHGRI, NIH, or the Department of Health and Human Services.

NOTES

1 Bustamante CD, Burchard EG, De la Vega FM. Genomics for the world. Nature. 2011;475(7355):163–5.

2 Popejoy AB, Fullerton SM. Genomics is failing on diversity. Nature. 2016; 538(7624):161–4.

3 Ramos E, Callier SL, Rotimi CN. Why personalized medicine will fail if we stay the course. Per Med. 2012;9(8): 839–47.

4 Hindorff LA, Bonham VL, Brody LC, Ginoza MEC, Hutter CM, Manolio TA, et al. Prioritizing diversity in human genomics research. Nat Rev Genet. 2018;19(3):175–85.

5 Bentley AR, Callier S, Rotimi CN. Diversity and inclusion in genomic research: why the uneven progress? J Community Genet. 2017;8(4): 255–66.

6 Need AC, Goldstein DB. Next generation disparities in human genomics: concerns and remedies. Trends Genet. 2009;25(11):489–94.

7 Manrai AK, Funke BH, Rehm HL, Olesen MS, Maron BA, Szolovits P, et al. Genetic misdiagnoses and the potential for health disparities. N Engl J Med. 2016;375(7):655–65.

8 Landry LG, Rehm HL. Association of racial/ethnic categories with the ability of genetic tests to detect a cause of cardiomyopathy. JAMA Cardiol. 2018 Feb 28. [Epub ahead of print].

9 Spratt DE, Chan T, Waldron L, Speers C, Feng FY, Ogunwobi OO, et al. Racial/ethnic disparities in genomic sequencing. JAMA Oncol. 2016;2(8):1070–4.

10 Worldometers. Countries in the world by population (2018) [Internet]. Dover (DE): Worldometers; [cited 2018 Apr 16]. Available from: http://www.worldometers.info/world-population/

11 United Nations. World population prospects: the 2017 revision: key findings and advance tables [Internet]. New York (NY): UN; 2017 [cited 2018 Mar 28]. (Working Paper No. ESA/P/WP/248). Available from: https://esa.un.org/unpd/wpp/publications/Files/WPP2017_KeyFindings.pdf

12 Ballantyne A, Goold I, Pearn A, WHO Human Genetics Programme. Medical genetic services in developing countries: the ethical, legal, and social implications of genetic testing and screening [Internet]. Geneva: World Health Organization; 2006 [cited 2018 Mar 28]. Available from: http://apps.who.int/iris/bitstream/handle/10665/43288/924159344X_eng.pdf?sequence=1&isAllowed=y

13 World Health Organization. Community genetics services: report of a WHO consultation on community genetics in low- and middle-income countries [Internet]. Geneva: WHO; 2010 [cited 2018 Mar 28]. Available from: http://apps.who.int/iris/bitstream/handle/10665/44532/9789241501149_eng.pdf?sequence=1&isAllowed=y

14 Shields AE, Burke W, Levy DE. Differential use of available genetic tests among primary care physicians in the United States: results of a national survey. Genet Med. 2008; 10(6):404–14.

15 Wideroff L, Vadaparampil ST, Breen N, Croyle RT, Freedman AN. Awareness of genetic testing for increased cancer risk in the year 2000 National Health Interview Survey. Community Genet. 2003;6(3): 147–56.

16 McCarthy AM, Bristol M, Domchek SM, Groeneveld PW, Kim Y, Motanya UN, et al. Health care segregation, physician recommendation, and racial disparities in *BRCA1/2* testing among women with breast cancer. J Clin Oncol. 2016;34(22): 2610–8.

17 Cragun D, Weidner A, Lewis C, Bonner D, Kim J, Vadaparampil ST, et al. Racial disparities in *BRCA* testing and cancer risk management across a population-based sample of young breast cancer survivors. Cancer. 2017;123(13):2497–505.

18 National Human Genome Research Institute. The NHGRI-EBI catalog of published genome-wide association studies [Internet]. Bethesda (MD): NHGRI; c 2017 Nov [cited 2018 Mar 28]. Available from: http://www.ebi.ac.uk/gwas/

19 National Center for Biotechnology Information. dbGaP [Internet]. Bethesda (MD): NCBI; [cited 2018 Mar 28]. Available from: https://www.ncbi.nlm.nih.gov/gap

20 To access the appendix, click on the Details tab of the article online.

21 National Institutes of Health. Inclusion across the lifespan—policy implementation [Internet]. Bethesda (MD): NIH; [last updated 2018 Jan 26; cited 2018 Mar 28]. Available from: https://grants.nih.gov/grants/funding/lifespan/lifespan.htm

22 Food and Drug Administraton. Collection of race and ethnicity data in clinical trials: guidance for industry and Food and Drug Administration staff [Internet]. Silver Spring (MD): FDA; 2016 Oct 26 [cited 2018 Mar 28]. Available from: https://www.fda.gov/downloads/RegulatoryInformation/Guidances/ucm126396.pdf

23 Otado J, Kwagyan J, Edwards D, Ukaegbu A, Rockcliffe F, Osafo N. Culturally competent strategies for recruitment and retention of African American populations into clinical trials. Clin Transl Sci. 2015;8(5): 460–6.

24 Hamel LM, Penner LA, Albrecht TL, Heath E, Gwede CK, Eggly S. Barriers to clinical trial enrollment in racial and ethnic minority patients with cancer. Cancer Control. 2016; 23(4):327–37.

25 Landry L, Nielsen DE, Carere DA, Roberts JS, Green RC. Racial minority group interest in direct-to-consumer genetic testing: findings from the PGen study. J Community Genet. 2017;8(4):293–301.

26 Sanderson SC, Brothers KB, Mercaldo ND, Clayton EW, Antommaria AHM, Aufox SA, et al. Public attitudes toward consent and data sharing in biobank research: a large multi-site experimental survey in the US. Am J Hum Genet. 2017; 100(3):414–27.

27 Williams DR, Mohammed SA, Leavell J, Collins C. Race, socioeconomic status, and health: complexities, ongoing challenges, and research opportunities. Ann N Y Acad Sci. 2010;1186:69–101.