NBER WORKING PAPER SERIES

NIH GRANT EXPANSION, ANCESTRAL DIVERSITY AND SCIENTIFIC DISCOVERY IN GENOMICS RESEARCH

Wei Fu
Shin-Yi Chou
Li-San Wang

Working Paper 30155
http://www.nber.org/papers/w30155

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
June 2022

We thank Feng Liu, Seth Richards-Shubik, Muzhe Yang and Fanny Leung for their valuable comments. All errors are our own. This study was supported by National Institute on Aging (U24AG041689, U54AG052427). The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

NIH Grant Expansion, Ancestral Diversity and Scientific Discovery in Genomics Research
Wei Fu, Shin-Yi Chou, and Li-San Wang
NBER Working Paper No. 30155
June 2022
JEL No. I14,J15,O31,O35,O38

ABSTRACT

In the approaching era of genomic medicine, the underrepresentation of minority populations in human genetics and genomics research has raised growing concerns regarding the distributive justice in the translation of biomedical innovations into human health across populations. Quantitative assessment of public funding policy in addressing the missing diversity is imperative yet lacking. In this paper, we fill this gap by empirically answering two central questions in the context of Genome-Wide Association Studies (GWAS): whether improved funding opportunity facilitates minority health research, and how significant is the scientific value of funding science on underrepresented populations. Our identification draws on an exogenous NIH grant expansion under the American Reinvestment and Recovery Act of 2009, and exploits variations in the share of people with medical conditions among minorities relative to whites. Our main findings are threefold. First, the ARRA-NIH grant expansion contributes to an increase in the inclusion of minority ancestries in GWAS. It also facilitates the engagement of minority scientists in academic activities and promotes their role in scientific collaborations. The grant expansion fosters the discoveries of disease-associated genetic variants within minority populations. This quantitative evidence speaks to the role that public funding policy can play in advancing science.

Wei Fu
Penn Neurodegeneration Genomics Center
Department of Pathology
and Laboratory Medicine
Perelman School of Medicine,
University of Pennsylvania
Leonard David Institute of Health Economics,
University of Pennsylvania,
3700 Hamilton Walk
Philadelphia, PA 19104
Fuwei@pennmedicine.upenn.edu

Li-San Wang
Penn Neurodegeneration Genomics Center
Department of Pathology and
Laboratory Medicine
Perelman School of Medicine,
University of Pennsylvania
3700 Hamilton Walk
Philadelphia, PA 19104
lswang@pennmedicine.upenn.edu

Shin-Yi Chou
Penn Neurodegeneration Genomics Center
Department of Pathology and
Laboratory Medicine
Perelman School of Medicine
University of Pennsylvania
3700 Hamilton Walk,
Philadelphia, PA, 19104
Department of Economics
Lehigh University
621 Taylor Street
Bethlehem, PA, 18015
and NBER
syc2@lehigh.edu

I. Introduction

With high-throughput genome sequencing technologies maturing and human genetic data accumulating at an ever-increasing pace, burgeoning advances in human genomics research enrich our weapons in the arsenal against diseases. One of the most revolutionary breakthroughs is genomic medicine (*i.e,* personalized medicine or precision medicine), which relies on a deep understanding of the genetic architecture of diseases to provide effective preventive and therapeutic strategies (Cohn, Henderson and Appelbaum, 2017).[1] Taking Ivacaftor (brand name Kalydeco)[2] as an example, it is estimated that overall and cystic fibrosis-related inpatient admissions dropped by 55% and 81%, respectively, among patients initiating Ivacaftor therapy (Feng et al., 2018). While applauding the health benefits that have resulted from the proliferation of genomic medicine and treatments, a growing chorus of voices is calling attention to the distributive justice in the translation of scientific advances to human health (Cohn, Henderson and Appelbaum, 2017; Beavis, Gravitt and Rositch, 2017; Singh, Kogan and Slifkin, 2017; Xu et al., 2018). Despite the well documented social-economic and environmental factors that affect health disparities in existing studies, the lack of ancestral diversity in biomedical science and clinical trials is gradually being recognized as a crucial hidden driver of the widening disparities, especially in the era of genomic medicine (Ramos, Callier and Rotimi, 2012; Bonham, Callier and Royal, 2016; Manrai et al., 2016; Landry et al., 2018; Martin et al., 2019).

Genomics research has witnessed persistent incongruities between disease burden and ancestral representation. For example, in the 230 clinical trials leading to FDA oncology drug approvals between 2008 and 2018, Blacks and Hispanics were largely underrepresented compared to their proportion of cancer incidence in the US, accounting for only 3.1% and 2.9% of the trial participants, respectively (Loree et al., 2019).[3] The lack of financial support for

[1]The Personalized Medicine Coalition defines personalized medicine as an evolving field in which physicians can provide targeted medical treatment and prevention plans to each patient, taking the diagnostic tests/genetic tests with the individual's medical history and family history into consideration. Personalized medicines are therapeutic products that are developed regarding specific biomarkers or genetic markers. Twenty-five of the 59 new molecular entities approved by the FDA in 2018 are personalized medicines, accounting for up to 42% of the FDA approvals, while only 5% in 2005 (Urban and Naylor, 2018; PMC, 2019).

[2]Ivacaftor was approved by the Food and Drug Administration (FDA) in January 2012. It is the first cystic fibrosis transmembrane conductance regulator (CFTR) modulator, targeting patients six years old and older with a G551D CFTR gene variant (De Boeck and Amaral, 2016; Feng et al., 2018).

[3]In addition, Chen Jr et al. (2014) finds that between 1993 and 2013, less than 2% of the cancer clinical trials funded by

projects on minority population could be a major contributor to the sluggish growth in diversity, apart from barriers such as cultural and linguistic obstacles, and lack of access to information on research for the minority community.[4] Against this background, this paper aims to examine the role of public funding policy in addressing the missing ancestral diversity in genomics research and its consequential effects on scientific advances.

In 1993, the Revitalization Act was signed into law to direct the National Institutes of Health (NIH) to require federally funded clinical research to prioritize the inclusion of women and minorities.[5] Decades after its implementation, there still exists a considerable funding disparity (Reardon, 2014; Hayden, 2015; Hoppe et al., 2019; Đoàn et al., 2019). For NIH R01 applications from 2011 to 2015, proposals from white Principal Investigators (PI) were 1.7 times more likely to be funded than those from African American PIs (Hoppe et al., 2019) and the gap persists even after matching a variety of characteristics of the proposals (Erosheva et al., 2020). In a survey by Jimenez et al. (2019), over 80% of the faculty respondents list insufficient funding as a barrier to engaging in diversity and inclusion activities. Inadequate funding support (Carnethon, Kershaw and Kandula, 2020; Taffe and Gilpin, 2021) and sizable recruitment cost of minority participants (Oh et al., 2015) hinder the minority health research.

There is consensus that certain fundamental changes, for example, in the funding mechanisms and peer review process, are required to promote the diversity and inclusion (Burchard, 2014; Popejoy and Fullerton, 2016; Carnethon, Kershaw and Kandula, 2020). However, most of the discussions on tackling the missing diversity to date have been descriptive. Our study expands their appeal by quantifying the effects of grant opportunities on promoting diversity in science. Quantitative analyses are critical to funding agencies when designing policies to cultivate research on the underrepresented populations.

Our empirical context is set in Genome-Wide Association Studies (GWAS), which seeks to identify disease-associated genetic variants. GWAS has been burgeoning since the first GWAS

the National Cancer Institute primarily focus on minority populations. Burchard et al. (2015) finds that among the NIH-funded pulmonary diseases-related studies published during 1993 and 2013, only 4.4% reported the inclusion of minority groups.

[4] George, Duran and Norris (2014) provides a systematic review of the barriers which impede the participation of minority communities.

[5] For more detail on the Revitalization Act of 1993, please see https://grants.nih.gov/policy/inclusion/women-and-minorities/guidelines.htm.

paper came out in 2005. As of September 2021, approximately 5,329 scientific articles are recorded in the GWAS Catalog.[6] A typical GWAS project is usually a case-control study and is designed to unearth the genetic architecture for diseases using genotype data from both normal individuals and patients.[7] GWAS projects require significant inter-institutional collaboration and resources as genome sequencing or genotyping of large-scale subjects is essentially needed to detect the very subtle but meaningful association between a single genetic variant and a disease. The identified gene-disease association is of importance in guiding the development of genomic medicine.

The grant expansion at the NIH due to the American Recovery and Reinvestment Act of 2009 (hereafter *ARRA-NIH grant expansion*) provides an ideal context for our study. To stimulate the economy through the advancement of science, NIH was appropriated a total of approximately $10 billion for two years on top of its regular funding budget, from February 2009 through September 2010. This expansion significantly raised NIH total funding level from $38.4 billion in 2008 to over $44.5 billion in both 2009 and 2010.

NIH being one of the largest funding agencies globally,[8] the ARRA-NIH grant expansion, by and large, greatly helped lessen two major barriers to minority health research – the shortage of minority scientists within the academic community,[9] and the inadequacy of financial support. Over 80% of the ARRA-NIH funding was allocated to support extramural scientific projects in all NIH-affiliated research institutes. It also financially supported the training of scientists from underrepresented populations, provided start-up packages, and set up new funding mechanisms, such as *Challenge Grants* and *Grand Opportunity Grants*, to support proposals with high impacts or scientific potential in a broad array of biomedical sciences. The unprecedented increase in grant opportunity, and a more transparent and expedited review process opened more access

[6]The GWAS Catalog is a platform collecting GWAS-related articles: https://www.ebi.ac.uk/gwas/.

[7]The Genome-Wide Association Studies fact sheet at the National Human Genome Research Institute provides plenty of background knowledge about GWAS: https://www.genome.gov/about-genomics/fact-sheets/Genome-Wide-Association-Studies-Fact-Sheet.

[8]Based on Mills and Rahal (2019) estimates, among published GWAS articles up to 2017, 85.11% of them made funding acknowledgments to US funding agencies, and primarily to NIH or Public Health Service.

[9]On the one hand, scientists from minority backgrounds have incentives to perform studies benefiting their communities (Bentley, Callier and Rotimi, 2017). On the other hand, they have cultural connections with the minority communities they are from, thus are able to mitigate mistrust and facilitate communications between the communities and the research teams (George, Duran and Norris, 2014; Hindorff et al., 2018).

to funding for junior or minority researchers to conduct GWAS projects focusing on minority ancestries, which are usually more costly in searching and recruiting participants than studies on European whites.

We construct a unique data set combining multiple data sources – the GWAS Catalog, PubMed portal, NIH RePORTER, and Medical Expenditure Panel Study (MEPS). Our final curated data consist of 2,706 GWAS studies from 27 broad disease categories between 2005 and 2016, and contain information on disease studied, study sample size, the ancestral background of subjects, the ancestral background of authors, discovered genetic variants (genes and SNPs), and grant information for each GWAS study. In the process of construction, we curate subjects' ancestral information from textual records in the GWAS Catalog, and assign subjects into seven broad ancestral categories.[10] We scrape a complete and ordered list of authors from PubMed Portal by the unique PubMed ID for each publication and predict the ancestral background associated with author's surname based on the 2000 US census. We also extract funding information from an article's acknowledgment and identify whether the study was supported by any ARRA-NIH grant by matching the grant series number in this article with an ARRA-NIH grant list from NIH RePORTER. We then group these curated studies into 27 broad disease categories based on the ICD-CM-10 code and Clinical Classification Software (CCS). By doing so, we are able to perform a disease-level analysis with an emphasis on the ancestral compositions of subjects recruited and the scientific engagement of minority scientists, as well as the disease-associated genetic variant discoveries.

It is challenging to causally quantify the impact of the ARRA-NIH grant expansion on ancestral diversity in genomics research. Simply regressing the diversity measure on the amount of funding received is not plausible for a causal interpretation. The source of endogeneity is twofold. First, reverse causality is a potential obstacle. As the NIH Revitalization Act of 1993 mandated the inclusion of minority and female subjects in clinical research, disease research involving a large pre-existing share of minority subjects may have a higher chance of obtaining an ARRA-NIH grant. In this situation, we might overestimate the impact of ARRA-NIH grant ex-

[10]For example, a typical description of the sample recorded in the GWAS Catalog before our curation is "560 African American cases, 360 African American controls, 420 European ancestry cases, 680 European ancestry controls in initial stage." For more details on our curation, please refer to data section and the appendix.

pansion. Besides, some unobserved confounding factors may also contaminate the estimate. One example is the grant reviewers' taste. Though the ARRA initiative did not explicitly prioritize some particular disease research, the NIH study session committees, which are responsible for the grant proposal review, may have particular preferences (Li and Agha, 2015; Gallo, Sullivan and Glisson, 2016; Li, 2017; Pier et al., 2018). For example, Burchard (2014) observes that NIH grant reviewers for genomics research tend to prefer proposals studying European white ancestry to proposals focusing on minority ancestries, as they believe that the former is more genetically homogeneous and the recruitment of the minority subjects are often more challenging so that the sample size for minority health research is too small to conclude convincing findings. It is difficult to fully control for unobserved confounders like the taste of the study session committee in the empirical exercise. This raises a concern that omitting those factors may lead to the impact of ARRA-NIH grant expansion being underestimated.

To address the concerns stated above, our empirical analysis employs a Difference-in-Differences model (DID) that exploits the across-disease variations, prior to the ARRA-NIH grant expansion, in the proportion of people with medical conditions among minorities relative to whites, which is referred to as relative referral base *(RRB)* in this paper. *RRB* is calculated as the ancestral difference (*minority vs white*) in the share of people with medical conditions in the US in 2005 based on the nationally-representative summary estimates by MEPS. We use the estimates in 2005 because that is the year when the first GWAS study came out, so *RRB* is less likely to be affected by the rapid development of GWAS. Ideally, *RRB* should well characterize how significantly each GWAS disease research benefits from the grant expansion. We provide the rationale from two perspectives below.

First, *RRB* captures, to some extent, variations in the feasibility of conducting minority health research across diseases. One of the significant sources of GWAS subjects is patients referred from hospitals or specialty care centers. A smaller *RRB* suggests that fewer minority patients, relative to whites patients, can be referred to and tracked by the research teams (Oh et al., 2015; Hindorff et al., 2018), thereby inducing significant threats to project feasibility, for example, by imposing a higher recruitment cost of minority subjects. Feasibility, always a critical review criterion for the study session committee, could significantly influence the odds

that a research proposal is approved by NIH. This is more true when it comes to minority health research proposals, as these projects often fall short of their recruitment goals (Stewart et al., 2020) and have difficulties collecting a sufficiently large number of subjects to obtain statistical power.[11]

Moreover, *RRB* also well describes the scientific value of GWAS projects on a minority population. A higher *RRB* suggests a relatively severe disease burden among the minority community, thereby underscoring an urgent need to understand the genetic architecture of such disease among the minority population.[12] Projects on diseases with a higher *RRB* are good candidates for some new types of funding mechanisms initiated by NIH during the ARRA-NIH grant expansion, for example, the *Grand Opportunity Grants*, which were designed to fund projects with significant short-term impacts.[13]

Our empirical analysis proceeds in four steps. First, we justify that *RRB* distinguishes which minority disease research would benefit from the grant expansion more significantly. Our exercise shows that the GWAS on diseases with a higher *RRB* tends to have a higher chance for being supported by an ARRA-NIH grant. The contribution of *RRB* to this likelihood quickly climbs to a peak during the ARRA period (2009-2010) and then diminishes gradually afterwards, which echoes the fact that the ARRA-NIH grant expansion is a two-year temporary stimulus. We do not observe any notable increase in funding support from other countries during the ARRA period.

Second, we investigate the impact of ARRA-NIH grant expansion on the diversity of GWAS subjects, taking the share of minority subjects in a particular disease research in a given year as the primary dependent variable. The descriptive analysis shows that the entire GWAS field has been developing rapidly since 2009, with the total number of subjects in 2016 surpassing 6 million. Minority subjects occupied approximately 8.4% of all subjects in 2016, which is better than its 3.4% in 2007, suggesting an improvement in diversity in the past decades. Nevertheless, mi-

[11] In GWAS research, in order to address the multiple testing issue, the criterion for significance is very stringent. After Bonferroni correction, the p value threshold for genome-wide significance is usually 1×10^{-8}.

[12] This idea is similar to the discussion on market size and pharmaceutical innovation, for example, as seen in Acemoglu and Linn (2004) and Dubois et al. (2015).

[13] Health disparity and genomics are two areas listed among the 15 grant challenge areas outlined by NIH in 2009 in the Omnibus of Broad Challenge Areas and Specific Topics. For more details, see https://grants.nih.gov/grants/funding/challenge_award/omnibus.doc, accessed on Oct 10, 2021.

nority ancestries are still underrepresented. Our empirical analysis reveals that the ARRA-NIH grant expansion contributes a sizable enrichment in the diversity of the ancestral composition of the GWAS subjects. It is noteworthy that the ARRA-NIH grant expansion leads to the most pronounced enhancement in ancestral diversity between 2009 and 2011, and the effect persists through 2016 after the ARRA-NIH initiative was terminated in 2010.

To explore what drives the long-lasting impact, we set our focus on the academic engagement of minority scientists. The inflow of minority researchers may raise the awareness of the lack of diversity in genomics research and prompt more research efforts on diseases prevalent among the minority population (Jimenez et al., 2019). Meanwhile, investigators of concordant ancestral background are helpful to facilitate communications, and build a mutual-trust partnership between the research team and the minority communities (Burchard, 2014; Bentley, Callier and Rotimi, 2017), thereby making the minority health research more feasible. In our data, we observe that the share of minority authors gradually grew from 8% in 2008 to 11.1% in 2016 with a peak of 14.7% in 2014. Our estimates demonstrate that the ARRA-NIH grant expansion substantially facilitates the academic engagements of minority scientists, as reflected by increasing share of minority authors. Moreover, the grant expansion also shapes the scientific collaboration pattern: it significantly increases the probability that a minority author serves as the first few author in a GWAS study, while having minimal impacts on minority researchers serving as principal investigators or fundraisers.

The inclusion of subjects from a variety of ancestries in GWAS studies contributes to more genetic information in detecting some novel ancestry-specific genetic variants for a common disease. It expands our understanding of a disease's genetic architecture by uncovering rare variants that are unique to some minority ancestries (Adeyemo and Rotimi, 2010) and confirming previous genetic findings from the European white ancestry in some minority ancestries (Ntzani et al., 2012). Having seen a considerable enrichment in diversity in subjects as a consequence of the ARRA-NIH grant expansion, our last exercise is to quantify the scientific returns to this expansion. We find that the improved funding support has yielded fruitful scientific discoveries – more disease-associated genetic variants from minority ancestries are discovered.

The remainder of the paper proceeds as follows. Section II briefly introduces background of

the ARRA legislation and GWAS. Section III describes the data and empirical approach in the analysis. Section IV presents the primary results of ancestral diversity in GWAS subjects. Sections V and VI further explore the diversity in the academic workforce and scientific discoveries. Section VII discusses our contributions, policy implications and concludes.

II. Background

A. Genome-Wide Association Studies

The completions of the Human Genome Project in 2003 and the International HapMap Project in 2005 initiated the genome era (Guttmacher and Collins, 2003; Roses, 2003).[14] In the new arsenal for geneticists, GWAS, the approach to compare genetic variation differences between affected and unaffected individuals using high-through genotyping technologies on large samples, is one of the most powerful weapons. Since the first GWAS article came out in 2005, this approach has been widely employed to explore the genetic variations associated with diseases like diabetes, Alzheimer's Disease, asthma, heart disease, and so on. For simplicity in this paper, we use the term "GWAS" to represent the human genomics research using GWAS approach, regardless of the underlying sequencing technology.

In a typical GWAS project, geneticists obtain DNA from two groups of participants, patients with the disease studied and healthy individuals with similar demographics. Selected positions on the chromosome, called single nucleotide polymorphisms (SNPs), are then scanned using high-throughput arrays that can genotype up to millions of SNPs for each individual.[15] SNPs are one of the common types of genetic variations in human genome. SNPs may potentially affect gene expressions and protein generations, especially when they are located within a gene or a regulatory region near a gene. The dysfunction of proteins may lead to variations in traits or onset of diseases. After a series of quality control processes, if one particular SNP or more

[14] In genomics, the genome era is also referred to the post-genomic era. Roses (2003) points out that this era would witness many new applications of genomic information to medicine development. For example, inexpensive high-throughput sequencing technologies enable us to identify multiple genetic variants associated with a given disease. The identification of disease-associated genetic variants can facilitate delivery of safe and effective medicines.

[15] The sequence of the four nucleotide bases Adenine(A), Guanine(G), Cytosine(C) and Thymine(T) codes the genetic information in DNA. Ninety-nine percent of the approximately 3 billion bases in human DNA are the same across populations, however, almost one in every 1,000 nucleotides on average differs from person to person. Those variations are called SNPs. For more information, please see https://ghr.nlm.nih.gov/primer/genomicresearch/snp.

has a significantly higher appearance among the patients relative to the healthy participants, then these SNPs or nearby genetic variants are biologically important for the disease and are likely to be associated with the risk of the disease studied. Scientists also employ gene-based association tests (GATES) to identify genes containing multiple risk SNPs that individually are weakly associated with a disease (Chung et al., 2019).[16] In our paper, we focus on SNPs and use disease-associated genetic variants and disease-associated SNPs interchangeably.

Over the past two decades, increasing genetic variants have been identified for many diseases using GWAS.[17] For example, the well-known Apolipoprotein E gene (known as $APOE$) on chromosome 19 is polymorphic, with three major alleles ($APOE$-$\epsilon 2$, $APOE$-$\epsilon 3$, and $APOE$-$\epsilon 4$). These three alleles can be discerned using two SNPs: rs7412 and rs429358.[18] The variant $APOE$-$\epsilon 4$ is found to be the strongest risk variant for Alzheimer's Disease (Koutsodendris et al., 2021). Other AD-associated SNPs include, for example, rs4236673 (closest gene: CLU) on chromosome 8 and rs10792832 (closest gene: $PICALM$) on chromosome 11.

The SNP-disease associations from GWAS are of great potential for the development of genomic medicine. Instead of the one-size-fits-all treatments, novel personalized medicine can be developed based on each patients' unique genetic profile, to provide targeted therapeutic treatments with high efficacy. One well-known example is the warfarin pharmacogenetics (Johnson and Cavallari, 2015). Warfarin is a widely-used medicine for the prophylaxis and treatment of venous thromboembolism and complications associated with atrial fibrillation and cardiac valve replacement. The cytochrome $P450$ (CYP) $2C9$ and vitamin K epoxide reductase complex 1 ($VKORC1$) genetic profiles have been incorporated into the dosing algorithms to design the optimal warfarin dosing to reduce the risk for over-anticoagulation and hemorrhage in the initial months of the treatment.[19]

[16] For example, $BRCA1$ and $BRCA2$ genes are strongly associated with breast cancer, and have been widely used as markers in genetic testing.

[17] DisGeNET is a large database for publicly available genes and variants associated to human diseases. For more information, see https://www.disgenet.org/.

[18] When hymine(T) appears at rs7412 and rs429358 (rs7412-T and rs429358-T), then the polymorphism is called $APOE$-$\epsilon 2$. In the case of rs7412-C and rs429358-T, it is called $APOE$-$\epsilon 3$. In the case of rs7412-C and rs429358-C, it is called $APOE$-$\epsilon 4$.

[19] Other examples are Trastuzumab for patients with $HER2$ positive breast cancers, and Vemurafenib in companion with $BRAF$ $V600E$ Mutation test for patients with late stage melanoma.For more details, please check: http://www.personalizedmedicinecoalition.org/Userfiles/PMC-Corporate/file/pmc_age_of_pmc_factsheet.pdf.

B. *Missing Diversity and Health Equity in the Era of Genomic Medicine*

Recent decades have witnessed noticeable progress in addressing the missing diversity in GWAS; however, it is barely adequate. Mills and Rahal (2019) finds that European white accounted for approximately 95.47% of the subjects in GWAS in 2007, and still as high as 87.96% in 2017.[20] In 2017, only 0.57% of the GWAS subjects were African, 1.2% were Hispanic or Latin American, and 6.33% were Asian. The missing diversity in genomics research raises the concern that whites may disproportionately benefit from the innovation in genomic medicine.

The "Asthma Inequality" is a well-known example showing the exacerbation in health disparity driven by missing minority subjects in research (Burchard, 2014; Cazzola et al., 2015).[21] Inhaled β2-agonists (such as Albuterol) are a widely used therapeutic drug for acute asthma.[22] The guidelines regarding the escalation of asthma therapy recommend an addition of a long-lasting-β2-agonist, superior to an increase in the dose of an inhaled glucocorticoid (*e.g.* Lemanske Jr et al., 2010; Peters et al., 2010). It is worth noting that this recommendation is mainly established by studies primarily conducted in the European white ancestry without accounting for variations in genetic architecture across ancestries (Wechsler et al., 2019).[23] Increasing evidence shows that long-lasting-β2-agonists medications do not work equally across populations: the risk of treatment failure among African American asthma patients is twice as large as that among their European white counterparts (Wechsler et al., 2011). The notable variations in the bronchodilator drug response reflect the rare and/or population-specific variants in the genetic architecture of asthma (Ortega et al., 2014; Mak et al., 2018).[24] The guidelines drawn from

[20] Popejoy and Fullerton (2016) also find the similar pattern. In their paper, GWAS subjects from European white ancestry account for about 96% of the 1.7 million samples in 373 studies in 2009, and about 81% of the 35 million samples in 2,511 studies in 2016.

[21] Oh et al. (2015) provides other disease examples in their table 1.

[22] Albuterol is often one of the commonly prescribed and the most affordable medication for low-income and minority populations. At major retail pharmacies, the price for Albuterol HFA is $15.79, compared to $203.91 for Asmanex, $96.44 for fluticasone-salmeterol (see `https://rxsaver.retailmenot.com/blog/asthma-medications-cost-without-insurance`).

[23] Another example is from the cardiovascular guidelines. Among all patients enrolled in the Randomized clinical trials (RCTs) cited in the American College of Cardiology/American Heart Association (ACC/AHA) guidelines for atrial fibrillation (AF), heart failure (HF), and unstable angina/nonST-segment elevation myocardial infarction (acute coronary syndromes [ACS]), 86% are white in RCTs for AF and ACS, 73% for HF RCTs (Sardar et al., 2014).

[24] Ortega et al. (2014) find that the rare ADRB2 variants are likely to cause a severe asthma exacerbation among LABA-treated white with the $Thr^{164}Ile$ variant, and among LABA-treated African American asthma patients with the -376 In-Del rare variant. The frequencies of the ADRB2 variants differ between ancestral populations. Mak et al. (2018) conduct a whole-genome sequence of pharmacogenetic drug response in Puerto Rican, Mexican, and African American children with asthma, and find novel genetic variants important for bronchodilator drug response near genes DNAH5, NFKB1, PLCB1, ADAMTS3, and COX18.

whites-based biomedical studies may cause inaccurate assessment of the disease risk and lead to improper treatment interventions among the underrepresented minority populations (Sirugo, Williams and Tishkoff, 2019), which underscores the calls for inclusion of diverse populations in genomics research (*e.g.* Rosenberg et al., 2010; Oh et al., 2015; Cohn, Henderson and Appelbaum, 2017; Hindorff et al., 2018; Wojcik et al., 2019).

C. ARRA-NIH Grant Expansion

In order to stimulate the economy after the Great Recession, the US carried out a prodigious stimulation package, the American Recovery and Reinvestment Act of 2009, in February 2009. One unprecedented initiative of the ARRA is to appropriate a total of approximately $10 billion directly to the National Institutes of Health to spark the economy through increasing spending on research. This financial package was available for two years, from February 2009 through September 2010.

The NIH was given a high flexibility in allocating the funding through various types of funding mechanisms to support projects that have potential to create new jobs and promote the advancement of science in two years.[25] The review process was more transparent and expedited: It asked for shorter research proposals without any requirements on preliminary data and analysis, and paid more emphasis on the impact of research and less on the technical details (Kaiser, 2009; Sorin and Hannum, 2014). Therefore, the volume of applications in 2009 and 2010 were boosted thanks to the NIH stimulus plan (Kaiser, 2009).

The NIH stimulus package is unprecedented, albeit temporary. The additional funding of $10 billion was put on top of the regular NIH budget. A report from the Congressional Research Service (Sekar, January 22, 2020) shows that: in the fiscal year 2008, the program-level funding at the NIH was about $38.4 billion (constant dollars in 2020 hereafter); it increased sharply to $44.8 billion in 2009 and $44.9 billion in 2010 (about $40 billion without ARRA-NIH grant in both years) and then dropped to $36.8 billion in 2011 as the stimulus package was terminated; it is worth mentioning that in the fiscal year 2020, the program-level fund was $41.6 billion,[26]

[25] See more details at the ARRA-NIH webpage: https://grants.nih.gov/recovery/.

[26] Estimate comes from FY 2020 By the Numbers: Extramural Investments in Research. For more details, see https://nexus.od.nih.gov/all/2021/04/21/fy-2020-by-the-numbers-extramural-investments-in-research/.

less than that either in 2009 or in 2010.

The ARRA-NIH funding were allocated through a variety of funding mechanisms. Of the total $10 billion to the NIH, about $1.8 billion was allocated for purposes such as extramural construction, repairs and capital equipment purchase, and the remaining $8.2 billion was designated to support extramural scientific research in all NIH-affiliated Institutes and Centers, Office of Director and Common Fund.[27] A broad set of research on diseases, such as diabetes, cancers, and heart and lung diseases, received substantial support from the ARRA-NIH grants through the existing funding mechanisms (such as R01) or new ARRA-NIH-wide programs. The ARRA-NIH-wide programs gave priority for projects with the potential to advance the frontiers of science through the *Challenge Grants* ($RC1$),[28] and also supported projects with significant short-term impacts through the *Grand Opportunity Grants* (short for $Go, RC2$).[29] The frontier biomedical research, including, for example, Alzheimer's Disease and large-scale sequencing, obtained exceptional financial support.

The impact of ARRA-NIH grant expansion is historical. From February 2009 through September 2010, more than 12,000 ARRA-NIH grants for 21,581 projects had been awarded. About 75% of the total funding was used to support the new science, and the rest for existing research.[30] Among the 13,000 Principal Investigator (PI) positions created by ARRA-NIH grants, about 9% were new PIs and they received one-eighth of the total ARRA-NIH funding. PIs who experienced a funding shortage before ARRA took one-sixth of the total funding (Sorin and Hannum, 2014).

The minority health research benefited largely from the expansion. For exmaple, owing to the Recovery Act in 2009, the National Institute on Minority Health and Health Disparities

[27] For more details on NIH's role in the American Recovery and Reinvestment Act, please see https://www.nih.gov/about-nih/who-we-are/nih-director/statements/nihs-role-american-recovery-reinvestment-act-arra.

[28] *Challenge Grants* was designated at least $200 million, to support novel research from domestic (United States) institutions/organizations. It supported scientific research up to $500 thousand total costs/year for up to two years, aiming to address specific knowledge gaps, scientific opportunities, new technologies, data generation, or research methods that would benefit from an influx of funds to advance the areas in significant ways quickly. For more details about the *Challenge Grants*, see https://grants.nih.gov/grants/funding/challenge_award/omnibus.pdf.

[29] The *Grand Opportunity Grants* with at least $200 million aimed to support large-scaled, well-defined high-impact, biomedical and bio-behavioral research endeavors. These projects should be new, not continuations of projects that have already been initiated previously. Moreover, the *Grand Opportunity Grants* were limited to up to 2 years. Supported projects should be without the expectation of continued NIH funding beyond two years. For more details about the *Grand Opportunity Grants*, see https://www.ninr.nih.gov/aboutninr/recovery/grand-opps.

[30] For more on ARRA-NIH grant summary, see https://report.nih.gov/recovery/NIH_ARRA_Funding.pdf.

(NIMHD) was awarded a two-year planning grant totaling $600,000 to study obesity-related disparities among Native Hawaiians and Pacific Islanders in 2009.[31] Moreover, the ARRA-NIH grants also provided remarkable financial support for researchers from underrepresented populations, including training junior or minority, or female scientists, providing start-up packages, and financing pilot research. These efforts contributed significantly to enhancing research capacity in academic institutions.

GWAS studies also benefited significantly from the ARRA-NIH grant expansion. Though we do not have a precise amount of the ARRA-NIH funding for GWAS, we can still intuit how tremendous the impact is from existing studies and reports. The National Human Genome Research Institute (NHGRI) reported that rewards valuing more than $113 million were granted to itself alone, on top of its regularly appropriated budget of $367 million.[32] Mills and Rahal (2019) estimates that up to 2017, about 85.11% of funding acknowledgments in GWAS publications were to US agencies. Most of those acknowledgments were primarily related to NIH grants (not limited to ARRA-NIH grants). The four NIH-affiliated institutes, National Heart, Lung & Blood Institute (NHLBI), National Cancer Institute (NCI), National Institute on Aging (NIA), and National Institute of Mental Health (NIMH), together with the UK Medical Research Council (MRC) are listed as the top five most frequently acknowledged agencies. As of 2017, approximately 57.6% of the cumulative acknowledgments in GWAS publications pointed to the top five agencies.

III. Data and Sample

A. Data Sources

Our primary data rely on four sources: (1) the GWAS Catalog, (2) the PubMed Portal, (3) NIH RePORTER, and (4) the Medical Expenditure Panel Survey (MEPS) summary tables.[33] Each data source serves different purposes, which are discussed below.

[31] https://recovery.nih.gov/stories/ViewStory.aspx?id=94.
[32] See https://recovery.nih.gov/Stories/ViewStory.aspx?id=35.
[33] The GWAS Catalog-EMBL-EBI: https://www.ebi.ac.uk/gwas/, accessed on Oct 14, 2019; The PubMed-NCBIL: https://www.ncbi.nlm.nih.gov/pubmed/; The NIH RePORTER: https://reporter.nih.gov/; The MEPS summary tables: https://meps.ahrq.gov/mepstrends/home/index.html.

The GWAS Catalog is a collection of GWAS articles, with the earliest article traced back to 2005. It is released and updated periodically by EMBL-EBI, an outstation of the European Molecular Biology Laboratory affiliated to the European Bioinformatics Institute. The Catalog keeps a comprehensive track of GWAS research development, comprising 5,329 publications and 292,054 associations as of September 15, 2020. The GWAS Catalog furnishes rich information on GWAS studies. It includes general journal-related information such as *PubMed ID, the first author, title, publication date, journal name, etc.* With the unique PubMed ID for each article, we are able to link the articles in the Catalog to the PubMed portal. The Catalog also provides research-related information including *a description of the research sample, disease/trait studied, Experimental Factor Ontology (EFO) assigned to each study,*[34] *disease-specific SNPs/Genes discovered, P-value for each SNP-association, location of each gene on the chromosome and their upstream and downstream genes, etc.* For details about the GWAS Catalog data, please refer to Appendix B.[35] The rich information allows us to calculate ancestral composition of the GWAS subjects to study the ancestral diversity, and to use the number of SNP-disease associations to measure the scientific output.

The second data source, the PubMed Portal, provides access to more than 30 million citations for literature from medicine and life science journals. With the unique PubMed ID for each article in the GWAS Catalog, we can query PubMed for more information on each article, including the abstract, a complete and ordered list of authors (with surname and first name), citations, and grant acknowledgments (if any). We predict the minority status of each author based on their surname using the 2000 US census, and calculate the share of minority authors for each disease research in a given year to capture the ancestral diversity in the scientific workforce.

NIH RePORTER, as the third data source, provides a list of ARRA-NIH grants, which were initiated or renewed by the ARRA initiative. We extract the grant series number from an article's grant acknowledgments. By matching the grant information for each study with this

[34]The EBI maps the curated traits/disease to a typical EFO term since the reported traits/disease in the GWAS research is highly diverse. The EFO system combines parts of several biological domain-specific ontologies, allows modeling of tests, diseases, anatomy, and anthropometry, and creates a hierarchical relationship between traits/diseases. It is also convenient to map the EFO term to other ontology systems such as ICD-10. For a detailed mapping rule, please refer to Ontology Xref Service https://www.ebi.ac.uk/spot/oxo/.

[35]Appendix B is available upon request.

list, we can identify which study was supported by ARRA-NIH grants.

The fourth data source is the summary tables of nationally-representative estimates provided by MEPS. MEPS is conducted by Agency for Health Care and Research Quality (AHRQ). It is a large-scale survey on U.S families and individuals, collecting information on their access to health care, medical conditions, health insurance, and health expenditure. Based on those individual data, AHRQ constructs various nationally-representative estimates summarized in the MEPS summary tables.[36] We construct the relative referral base for the minority subjects in GWAS based upon two national estimates from the MEPS summary table: the number of people with specific medical conditions and the size of the population, both by ancestry and disease.

B. Data Curation

We retrieved the data for analysis from the GWAS Catalog on October 14th, 2019. In our data, the earliest recorded GWAS publication came out on March 10th, 2005, and the latest one on September 9th, 2019. The raw data obtain 7,797 GWAS studies from 4,258 scientific publications.[37] A publication may involve multiple GWAS studies that use different samples of subjects. Given our interest in the diversity in GWAS studies, we organize the records by study and briefly describe the curations below. For more technical details and illustration examples, please refer to Appendix B.

Subjects' Ancestries – Of the 7,797 studies, 7,789 have reported ancestry information for their study sample. The ancestral information includes the *textual ancestry descriptions, sample size*, and *country of origin* for subjects in the initial sample and the replication sample, respectively.[38] Most importantly, GWAS subjects are classified into more than 180 broad ancestral categories in the GWAS Catalog. In order to facilitate our analysis, we aggregate these broad ancestral categories into seven ancestry categories following Mills and Rahal (2019). These

[36] MEPS summary tables: https://meps.ahrq.gov/mepstrends/home/index.html.

[37] A publication is a scientific article with a unique PubMed ID. Some publications may involve multiple genome-wide association studies, each of which with distinct diseases/traits, sample cohorts, or other unique characteristics is referred to as a study in the Catalog, and assigned with a unique accession number beginning with "GCST" in the data.

[38] We provide two examples. Example 1:"560 African American cases, 360 African American controls, 420 European white ancestry cases, 680 European white ancestry controls in initial stage." Example 2: "946 cases, 977 controls in the initial stage, 530 trios, 353 cases, 207 controls in the replication stage." Example 1 provides detailed information about subjects' ancestry, but Example 2 does not. In our curation, we assign subjects in Example 2 into the category "In Part not Reported."

seven broader ancestral categories in our analysis are: "African," "African American or Afro-Caribbean," "Asian," "Hispanic or Latin American," "European White," "Other or Mixed," and "In Part Not Reported." The last ancestral category is excluded from our analysis due to limited information. The first four ancestries make up the minority group in our analysis.

Authors' Ancestries – We scrape a complete and ordered list of authors from PubMed Portal by using the unique PubMed ID for each publication. Eventually, for the 7,797 studies, we identify 141,335 authors in total.[39] We predict the ancestral background associated with each surname based on the 2000 US census. The prediction leverages the python module *"ethnicolr"* and gives the probability that a name is from "Asian", "Black", "Hispanic" or "White" background.[40] We adopt a conservative rule: a surname is predicted to be associated with a minority background only if its probability of being either "Asian", "Black" or "Hispanic" is more than 0.95.[41] Of the 141,335 authors, 19,510 are predicted to be minority names based on this prediction criterion. We also use different probability thresholds in our robustness checks. For example, if setting the criterion to be 0.90, this will result in 24,661 minority names.

Funding – A grant list for the 7,797 studies is also queried from the PubMed Portal. Among them, 5,855 studies explicitly specify grant information, including grant series number and funding agencies, in their acknowledgments, resulting in a total of 55,142 grant acknowledgments.[42] After dropping the duplicated grant series numbers, we identify 10,926 unique grants issued by US funding agencies. In order to identify the studies supported by the ARRA-NIH grants, we implement a *fuzzy-wuzzy* method to match the grant series numbers between the list of 10,926 unique US-issued grants that we queried from PubMed portal and the list of all ARRA-NIH grants from the NIH RePORTER. After matching, 10,849 grant series numbers return with a match score. We set the match score threshold to be 100 and manually check some with a score over 97. Finally, 903 of the 10,849 matches are identified as ARRA-NIH grants, and the studies

[39] When we count authors, an author may be counted multiple times. For example, if he is listed in ten publications, his name will be counted ten times.

[40] The "ethnicolr" module can be found here: https://pypi.org/project/ethnicolr/.

[41] For example, the last name "Huang" is predicted to be an Asian name with a probability of over 0.99. However, we can not tell if the last name "William" is a White name or Black name given the probability for the former is 41.2% and 51.6% for the latter.

[42] Some studies do not explicitly specify grant acknowledgments. We assume that they do not receive any grant supports. Also some studies have multiple grant supports.

associated with them are considered to benefit from the ARRA-NIH grants expansion. We refer to these studies as ARRA-NIH-supported studies.

Harmonizing Disease Classifications – We complete our data for analysis by merging national estimates from MEPS to the curated data based on the GWAS Catalog via disease category. The fact that MEPS and the GWAS Catalog adopt different disease classification systems makes the linking challenging. Given that the MEPS uses Clinical Classification Software (CCS) based on ICD-10-CM codes, we convert the EFO ontology system adopted by the GWAS Catalog to the CCS via the ICD-10-CM codes. The conversion is not perfect. First, some GWAS studies focus on biomarkers rather than diseases.[43] It is common for a biomarker to be considered as a risk indicator for more than one disease. Therefore, classifying biomarkers to a particular disease is challenging. We have to remove these studies focusing on biomarkers. Second, the mapping from EFO to ICD-10-CM codes provided by EMBL-EBI Ontology Xref Service does not cover all EFO terms.[44] The attrition also arises from the conversion from ICD-10-CM codes to the CCS system, as not all ICD-10-CM codes can be found a match in the CCS system.[45]

We are aware of the potential issue due to the attrition in converting the EFO ontology to the CCS system though this is the best we can do. The sample attrition is summarized as follows: Of the 7,789 studies with complete ancestry information, 7,411 are successfully mapped to at least one EFO term, which results in 9,928 study-EFO pairs.[46] At last, 3,510 of the 9,928 study-EFO terms get merged with the CCS codes, which infers that 2,706 of the unique 7,411 studies can be matched to a CCS code. We finally have 4,802 study-CCS pairs, given that one EFO term may be matched to multiple CCS categories. By doing so, each observation in our sample is uniquely identified by study ID and CCS term together.

[43] A biomarker is defined as "a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention" in 1998 by the NIH Biomarkers Definitions Working Group. For example, the serum LDL is identified as a risk measure for coronary and vascular disease, C-reactive protein (CRP) for inflammation. See more details here: https://www.niehs.nih.gov/health/topics/science/biomarkers/index.cfm.

[44] EMBL-EBI Ontology Xref Service: https://www.ebi.ac.uk/spot/oxo/datasources/EFO.

[45] The *"hcuppy"* python module is used to finalize the dictionary for mapping the ICD-10-CM to CCS. We have to be cautious that in this step, the *"hcuppy"* python module does not fully convert a proportion of the ICD-10-CM codes, which have only the first two digits available. This is because the *"hcuppy"* requires a full format of the ICD-10-CM code for conversion. Therefore, we curate the dictionary manually by adding this missing part of the mapping with reference to the *"hcuppy"* technical document.

[46] The rest may be mapped to other ontology systems such as the Orphanet.

We are concerned that the attrition may cause our sample to be biased if the dropped studies are systematically different from the existing ones in terms of ancestral compositions of the subjects. To assess the influence of attrition, we compare the diversity measure in our data with that in Mills and Rahal (2019), which comprehensively consider all GWAS studies. The share of the minority subjects in our analysis is slightly higher than theirs. However, the aggregate time trend in ours is similar to theirs.

C. Sample Construction

Given our focus on the impact of ARRA, we restrict our analysis to studies with grants from any US funding agencies. We then aggregate the study-CCS pairs supported by any US funding agencies between 2005 and 2016 by year and broader disease categories to construct a disease-year sample for the empirical analysis. Note that the number of studies in the years before 2007 is minimal. Therefore, where not specifically mentioned, we recode 2007 to include 2007 and years prior to 2007.

To do so, we classify the study-CCS pairs into 52 broader disease categories based on the CCS-disease mapping provided by MEPS (see Table B7.2).[47] In order to keep enough observations in each disease category, we group some closely related disease categories into one category.[48] Then 38 broader diseases remain. We drop diseases named "residual codes" and "symptoms," which leads to 36 diseases remaining, as shown in Table 1. For nine diseases without an ID in Table 1, the number of people with medical conditions is not available in the MEPS, so we drop them.

Eventually, our analysis sample consists of 27 broader disease categories and a total of 183 disease-year pairs from 2007 to 2016. Figure A1.1 shows the number of studies for each disease in each year from 2007 to 2016. It is apparent that the intensity of research varies considerably from disease to disease and from year to year. The studies on diseases like cancer and mental

[47] The CCS-Disease mapping from MEPS is here: https://meps.ahrq.gov/data_stats/conditions04.shtml.

[48] We group disease categories as follows: "glaucoma" and "cataract" into "other eye disorders," "other GI" into "disorders of the upper GI," "CNS infection" into "other CNS disorders," "other endocrine, nutritional & immune disorder" into "thyroid disease," "non-malignant neoplasm" into "cancer," "epilepsy and convulsions" into "headache," "urinary tract infections" and "other urinary" into "kidney disease," "trauma-related disorders" into "mental disorders," "acute bronchitis and URI" into "pneumonia," and "back problems" into "osteoarthritis and other non-traumatic joint disorders".

disorders receive the most intensive attention each year. However, diseases like anemia and other deficiencies, and female genital disorders and contraception are comparatively less studied.

D. Variables

Relative Referral Base for Minority Subjects – The referral base for subjects in a disease research consists primarily of patients in hospitals or specialty care centers with a diagnosis of such disease. Physicians or hospitals can refer these patients to research teams. This referral base is a significant source for the GWAS subjects.

We construct the referral bases (hereafter RB) for minorities and whites, respectively, based on the MEPS summary table in 2005, the latest year before the earliest GWAS publication came out. We avoid using the MEPS data in the years after 2005 because of the concern that the development in GWAS may reversely affect the recruitment of minority patients.[49] Respondents in MEPS are classified into five ancestry groups: (1) Hispanic, (2) Black, (3) American Indian, Alaska Native, and multi-races, (4) Asian, Hawaiian and Pacific Islander, and (5) European White. For each ancestry group, we obtain two nationally-representative estimates in 2005 from the MEPS summary table: the number of people with medical conditions by disease and the population size. The minority population in our following analysis contains all the mentioned ancestries except European White.

We calculate the relative referral base (RRB_d^{2005}) for the minority subjects in 2005 using Equation (1),

$$RRB_d^{2005} = [\frac{P_{m,d}^{2005}}{Pop_m^{2005}} - \frac{P_{w,d}^{2005}}{Pop_w^{2005}}] \times 100. \tag{1}$$

$P_{m,d}^{2005}$ ($P_{w,d}^{2005}$) is the number of minorities (whites) with medical conditions associated with disease d in 2005, and Pop_m^{2005} (Pop_w^{2005}) is the population of minorities (whites) in 2005. The ratios, $\frac{P_{m,d}^{2005}}{Pop_m^{2005}}$ and $\frac{P_{w,d}^{2005}}{Pop_w^{2005}}$, are the RBs in 2005 for minorities and whites, respectively.

RRB_d^{2005} is indicative of the difference in the number of people with medical conditions

[49] For example, newly discovered genes from GWAS may boost medicine innovations, which possibly make the treatment cheaper and more accessible to disadvantage groups. However, it is well understood that the translations from scientific findings to the clinical trial and then to the market take a significantly long time.

associated with disease d per 100 individuals in 2005 between minorities and whites. We also prefer to view RRB_d^{2005} as a proxy reflecting variations across diseases in the recruitment cost of minority patients, which is standardized with respect to that of whites. A larger RRB_d^{2005} implies that for disease d, it is less challenging in subject recruitment for a GWAS study focusing on minorities because of the relatively lower cost in searching minority participants.

We cautiously prefer RRB to RB in our baseline analysis. The grant study session committee often faces a resource allocation problem involving cost-benefit analysis of financial investments into studies on minorities or whites, respectively. This reminds us that the relative rather than the absolute cost may matter more in funding allocation. In our decomposition analysis, we also examine the impact of each RB separately. We should also be cautious that this referral base is not equivalent to the disease prevalence rate used in epidemiology. The referral base based on MEPS data is not only determined by the disease prevalence rate but is also affected by factors such as access to care, insurance coverage, and income status.

Table 1 shows the RRB for minorities by disease. Diseases without an ID are not used in the empirical analysis due to the missing information on RRB in MEPS. There exist substantial variations in the referral base across diseases in either the European white ancestry or the minority ancestries. In Figure 1, we plot the referral bases for whites and minorities. Except for anemia and other deficiencies, the referral base for whites is considerably larger than the base for minorities, especially for mental disorders, hypertension, and asthma. A larger referral base for whites may be a consequence of their better access to medical services.

Ancestral Diversity in Subjects – We focus on the subjects in both the discovery and replication stages in GWAS studies. We define an ancestral diversity measure ($SubjectDiv_{d,t}$) as the share of minority subjects in GWAS studies on disease d published in year t, as shown in Equation (2):

$$SubjectDiv_{d,t} = \frac{n_{d,t}^s}{N_{d,t}^s}, \tag{2}$$

where $N_{d,t}^s$ and $n_{d,t}^s$ are the number of all subjects and the number of the minority subjects in the studies on disease d in year t. A higher value of $SubjectDiv_{d,t}$ is indicative of enrichment in

the ancestral diversity in the GWAS subjects.

In Figure A3.1, we draw a heat map illustrating the ancestral diversity by disease and year. The heat map shows several important points. First, some diseases are not thoroughly studied and even do not have any publications in some years. Second, in the majority of the disease research, the minority subjects occupy a share of less than 10% of the total subjects. Third, diseases like cancers and mental disorders involve relatively more minority subjects.

We describe the ancestral diversity from different angles in Figure 2. Figure (A) depicts the number and share of GWAS participants by ancestries and scientific stages over our study period. The scale of the GWAS studies increased rapidly from 2007 to 2016. In 2016, the total number of subjects in the discovery stage approached 6 million, and the number in the replication stage surpassed 2 million. Though the inclusion of minority subjects has been improving, it remains inadequate. As shown in figure (B), the share of minority subjects was approximately 3.0% in 2008 and 4.6% in 2009. It gradually increased and reached a peak of 25.5% in 2014, after which it experienced a quick drop.[50]

In figure (C), we break down the share of the minority subjects into four ancestries. The patterns across all ancestries are similar. Notably, Asian subjects is a sizable component of the minority subjects. For the break-downs in discovery and replications stages, see Figure A3.2.

We also separate the 27 broader disease categories into a "quasi-treated" and a "quasi-control" group based on their *RRB*. The "quasi-treated" includes 13 diseases with *RRB* larger than the mean of the 27 diseases, and the "quasi-control" includes the other 14 diseases. In Figure (D), the patterns across the two groups before 2009 were similar but diverged after the ARRA-NIH grant expansion. The share of minority subjects in the "quasi-treated" group experienced much faster growth after the expansion.

Ancestral Diversity in Scientific Workforce – Based on the ancestral pattern of the surname in the 2000 US census data, we predict the author's ancestral background, which falls into one of the four categories: Asian, Black, Hispanic, and White. The minority authors in our definition

[50]Compared to Mills and Rahal (2019), the share of the minority subjects in our analysis is higher than theirs. This is possible because their sample consists of both the disease-focused and biomarker-focused GWAS studies. Our analysis only includes the former. However, the aggregate time trend in ours is similar to theirs.

include names from the first three ancestry categories. The ancestral diversity in the scientific workforce in our analysis is defined as $AuthorDiv_{d,t}$ in Equation (3):

$$AuthorDiv_{d,t} = \frac{n^a_{d,t}}{N^a_{d,t}}, \tag{3}$$

where $AuthorDiv_{d,t}$ represents the share of minority authors in all authors in the GWAS studies on disease d in year t. $N^a_{d,t}$ and $n^a_{d,t}$ are the number of author's names and the number of minority author's names in studies on disease d in year t, respectively. A higher share of minority authors implies a more diverse academic community in the disease research, given that, thus far, the majority of the researchers are white. It should also be noted that $AuthorDiv_{d,t}$ captures the active research participation of minority researchers.[51] Figure A4.1 depicts the number and the share of minority authors by disease and year, respectively. There are considerable variations in the share across diseases and over the years. In the majority of the disease studies, less than 20% of the authors are minorities, which suggests a lack of involvement of researchers from underrepresented populations.

We illustrate the development of ancestral diversity of the scientific workforce in GWAS in Figure 3. As shown in Figure (A), the GWAS has expanded at a fast pace since 2007. In the publications in 2007, there were about merely 1000 active authors; however, it increased to more than ten thousand in 2015. Before 2009, the share of minority authors was small, taking less than 9% of the authors in 2007. In 2014, 14.7% of the authors were from the minority background.

In Figure (B), we again split the 27 diseases into the "quasi-treated" and the "quasi-control" groups as we did before. Before 2009, the two groups exhibited a similar trend in the share of minority authors. After the ARRA-NIH grant expansion, the share in the "quasi-treated" group began to increase promptly and exceeded the share in the "quasi-control" group.

Scientific Discoveries: SNP-disease Associations – We measure the scientific output by directly leveraging the number of SNP-disease associations from each study curated by the

[51] $N^a_{d,t}$ and $n^a_{d,t}$ count a name by its appearance. An author name is counted multiple times if it is listed in more than one publication. Thereby, $AuthorDiv_{d,t}$ is a weighted measure, by using the frequency of the author as a weight.

GWAS Catalog. It is worth mentioning that our interest lies in the associations from minority studies (*hereafter minority-based SNP-disease associations*). In our preferred definition, a SNP-disease association is recognized as a minority-based association as long as it is identified from a GWAS study in which all subjects in the discovery stage are minorities. We also adopt alternative criteria in the robustness checks.

In addition to the number of minority-based SNP-disease associations, we also use the share of minority-based SNP-disease associations in the total associations as calculated in Equation (4) to measure the genetic discoveries. $A_{d,t}$ and $a_{d,t}$ are the number of associations and the number of minority-based associations found for disease d in year t. The heat maps in Figure A5.1 illustrate the number and the share of minority-based associations by disease and year. We can see that in most of the disease studies, the minority-based SNP-disease associations are less than ten. However, the cancer study in 2016 is an extreme case, with more than 100 minority-based associations discovered. Those 100 minority-based associations account for over 40% of all the associations found for cancer in 2016.

$$SNPShare_{d,t} = \frac{a_{d,t}}{A_{d,t}}. \tag{4}$$

Figure 4 displays the time pattern of minority-based SNP-disease associations. In Figure (A), we see that there were no minority-based SNP-disease associations discovered before 2009. This echoes the lack of minority subjects in 2007 and 2008. Starting in 2009, both the number and the share began to proliferate. In 2016, 277 minority-based associations were identified, which occupied approximately 10% of the entire associations found in that year. Figure (B) compares the growth between the "quasi-treated" and the "quasi-control" groups. More minority-based genetic associations have been found for the diseases with a large RRB after 2009, before which the two groups had almost the same level of minority-based SNP-disease associations.

IV. Empirical Strategy

A. A Simple Conceptual Framework

We first illustrate the rationale of our identification strategy using a simple framework. We consider N diseases and two ancestral groups, for example, the European white and the minority ancestries. A funding agency, for example, the NIH, allocates funding in a total of F to its N affiliated disease research centers, each of which focuses on a particular disease. Each research center then allocates its grant funding among projects on each ancestral group. The goals of the funding agency and its affiliations are to maximize the scientific output. This two-stage funding allocation mimics the funding process between NIH and its affiliations.

First, NIH allocates research funding to genomics research on each disease and faces the following optimization problem:

$$\max_{\{f_1, f_2 \cdots f_N\}} \sum_{d=1}^{N} S_d(f_d, r_d), \quad s.t \sum_{d=1}^{N} f_d = F, \tag{5}$$

where $S_d(f_i, r_i)$ is the net scientific production function for genomics research on disease d depending on two parameters: the funding support f_d, and the referral base r_d (*i.e,* number of patients per 100 individuals). The r_d comes into the net scientific function in three ways. It is related to the difficulties in recruiting patients in genomics research. A large referral base r_d is associated with a lower recruiting cost. It also reflects the burden of the disease. A higher burden implies an emergency to performing research into this disease. A large share of patients suffering from such disease is always associated with a substantial medical cost. It could also be seen as a proxy of the size of a pharmaceutical market, reflecting the potential value of studying such disease. We assume some monotonic properties of the function $S_d(\bullet)$: $\frac{\partial S_d}{\partial f_d} > 0, \frac{\partial S_d}{\partial r_d} > 0, \frac{\partial^2 S_d}{\partial f_d^2} < 0, \frac{\partial^2 S_d}{\partial r_d^2} < 0, \frac{\partial^2 S_d}{\partial f_d \partial r_d} > 0$.

The optimal funding allocation for NIH is $(f_d^*, f_d^*, \ldots, f_N^*)$, satisfying:

$$\frac{\partial S_i(f_i^*, r_i)}{\partial f_i} = \frac{\partial S_j(f_j^*, r_j)}{\partial f_j} \text{ for } i \neq j. \tag{6}$$

Provided that the net scientific production function $S(\bullet)$ does not vary remarkably across disease research, then an important implication from Equation (6) is that the disease with a large referral base deserves more research funding support.

Within each research center, the funding from NIH is allocated between projects focusing on different ancestries. We denote that the referral base for ancestry 1 is $r_{d,1}$, and $r_{d,1} + \Delta_d$ for ancestry 2. Δ_d indicates the difference in the referral base for disease d across ancestries. We assume that research center d aims to maximize the total scientific production from studies on disease d in ancestry 1 and ancestry 2, under a fixed research budget f_d^*. For simplicity, we assume there is no spillover effect of scientific advances across ancestries. The research center d solves a problem similar to (5) , and follows the first order condition:

$$\frac{\partial S_d(f_{d,1}^*, r_{d,1})}{\partial f_{d,1}} = \frac{\partial S_d(f_{d.2}^*, r_{d,1} + \Delta_d)}{\partial f_{d,2}}, \tag{7}$$

where $f_{d,1}^*$ and $d_{d,2}^*$ are the optimal funding for studies on ancestry 1 and ancestry 2, respectively. Equation (7) states that keeping everything else constant, if Δ_d, namely the difference in the referral base, increases, then the research center will allocate more research funding towards projects focusing on ancestry 2. This is sensible in a way, that a relatively large referral base implies a smaller searching cost in recruiting study participants. In our study context, the conceptual framework suggests that GWAS projects on a minority ancestry with a larger *RRB*, namely, a larger Δ_d in the framework, would benefit more from the ARRA-NIH grant expansion.

B. *Identification Strategy*

We construct a Difference-in-Differences (DID) model exploiting variations in RRB across diseases as shown in Equation (8):

$$y_{d,t} = \beta_0 + \zeta RRB_d^{2005} \times Post_t + \beta_1 X_{d,t-1} + \gamma_d + \theta_t + \epsilon_{d,t}, \tag{8}$$

where $y_{d,t}$ is the outcome of interest regarding disease d in year t. It is the ancestral diversity in subjects, the ancestral diversity in authors, or the minority-based SNP-disease association discoveries. RRB_d^{2005} is the relative referral base for the minority subjects for disease d in 2005. $Post_d$ is defined as one if the year is later than 2010 (≥ 2010); otherwise, zero. $X_{d,t-1}$ is a vector of control variables aiming to control for the disease-related research characteristics, including the total number of studies funded by non-US grants up to the year $t-1$ and the number of minority-based SNP-disease associations discovered in studies funded by only non-US grants up to the year $t-1$ to capture the scale and knowledge accumulation. We also control for the total number of grants from US funding agencies up to the year $t-1$ to capture the inertia in funding decisions by funding agencies in the US.[52] We superimpose the disease fixed effect γ_d to account for the disease-specific and time-invariant factors and year fixed effect θ_t for the general time trend in the scientific community.

The validity of a DID estimator rests with the pre-event parallel trend assumption. In order to test this, we implement a more parsimonious event study as shown in Equation (9):

$$y_{d,t} = \beta_0 + \sum_{j=2007; j\neq 2009}^{2016} \zeta_j RRB_d^{2005} \times 1\{\text{year} = j\} + \beta_1 X_{d,t-1} + \gamma_d + \theta_t + \epsilon_{d,t}, \tag{9}$$

where we particularly focus on the coefficients of the interactions of RRB and the indicator of each year, namely ζ_j. The year 2009 is the base year. Due to minimal publications in 2005 and

[52] Packalen and Bhattacharya (2020) note that the NIH may keep funding some scientific fields though those fields are less fruitful than expected. This inertia in funding decisions may keep those fields over-funded and lead to a less diversified funding portfolio. Hanna (2015) further finds that due to the inertia in funding decisions, diseases such as malaria and HIV/AIDS are over-funded by the NIH relative to the health burden on the American population. In contrast, diseases such as stroke and heart disease are still underfunded.

2006, we define $1\{year = 2007\}$ as 1 if j is 2007 or earlier. The coefficient ζ_j is indicative of the impact of the ARRA-NIH grant expansion in year j relative to year 2009. A necessary condition for the pre-event parallel trend across diseases requires that the estimates of ζ_{2007} and ζ_{2008} are both statistically insignificant.

We address the standard errors in two ways. We report the Huber-White robust standard errors in our primary estimation results. We also cluster the standard errors at the disease level to capture the correlations within disease. Given that there exist only 27 diseases in our analysis, we follow Roodman et al. (2019), and report wild-bootstrap cluster p-values and wild-bootstrap cluster 95 percent confidence intervals using the boottest Stata command (10000 replications). In the discussion below, we primarily interpret our estimates based on the Huber-White robust standard errors, as the results based on the two standard error approaches are almost consistent.

C. Validation of Identification Rationale: Impact on Funding

Recall that our identification is grounded on the rationale that the GWAS on diseases with a higher *RRB* would face a lower cost of recruiting minority subjects, and thus benefit more from an improved access to funding in the ARRA-NIH grant expansion. We first justify this intuition by examining whether the GWAS on diseases with a higher *RRB* tends to have a higher chance for obtaining an ARRA-NIH grant.

Our interest is the intensity of financial support from the ARRA-NIH grant expansion for GWAS projects on disease d in year t. As shown in Equation (10), this metric is defined as the share of studies supported by any ARRA-NIH grants in all studies that are supported by any US grants on disease d in year t. We focus on a sample of 183 disease-year pairs from 2007 to 2016, which is aggregated from studies acknowledging at least one grant from US funding agencies. In the placebo test, we extend the sample to include studies with any grant source.

$$\text{ARRA-NIH-Supported Study Share}_{d,t} = \frac{\#\text{ of ARRA-NIH-Supported Study}_{d,t}}{\#\text{ of US-Grant-Supported Study}_{d,t}}. \tag{10}$$

We estimate the DID model specified as Equation (8) with the dependent variable being the share of studies funded by any ARRA-NIH grants. Estimates in columns 1 and 2 in Table 2

suggest that a higher RRB is positively associated with a higher likelihood of being supported by an ARRA-NIH grant. With a one-percentage increase in RRB, the approval likelihood of a proposal submitted during the window of the ARRA-NIH grant expansion increases by 1.81 percentage points. Given that the magnitude of average RRB is 2.95 per 100 individuals in 2005, this estimate suggests that the ARRA-NIH grant expansion is associated with an increase in the approval likelihood, on average, of about 5.34 percentage points ($\simeq 1.81 \times 2.95$).

Figure 5 plots the coefficients from the event study as shown in Equation (9). Two findings emerge in the figure: First, it is apparent that there is a substantially large jump in the chance of being supported by ARRA-NIH grants from 2009 to 2010, which is in line with the timing of the ARRA-NIH grant expansion; second, the policy impact decreases over the years from 2010 through 2016, with a peak in 2010 and a bottom in 2014. This echoes the fact that NIH experienced a temporary grant stimulus during 2009 and 2010, and the stimulus faded in the subsequent years. It is worth highlighting that we do not observe any statistically significant impact in the years before the policy implementation. The coefficients in 2007 and 2008 are relatively smaller, close to zero, and insignificant.

We also perform a robustness check by extending our sample to studies with funding support from all sources, for example, funding agencies outside the US.[53] The dependent variable is the share of studies funded by any ARRA-NIH grants in the total number of studies on disease d in year t. Results summarized in columns 3 and 4 in Table 2 and in Figure (A) of Figure A2.1 are in line with the baseline conclusions.

In the last two columns, we perform a placebo test examining if the ARRA-NIH grant expansion leads to any changes in scientific investment outside the US. The dependent variable is the share of studies funded by any international grants in total number of studies on disease d in year t. The estimates are small in magnitude and statistically insignificant, suggesting minimal spillover effect or crowding out effect of the ARRA-NIH grant expansion. The event study is plotted in Figure (B) of Figure A2.1.

[53]The funding could come from the ARRA-NIH grant pool, regular NIH pool, other funding agencies rather than NIH in the US, funding agencies in other countries, or international organizations. Two examples of funding agencies outside the US are the Medical Research Council (MRC) and Wellcome Trust in the UK.

V. Impact on Ancestral Diversity in GWAS Subjects

Analysis in this section primarily focuses on a sample of 183 disease-year pairs from 2007 to 2016, which is aggregated from studies acknowledging at least one grant from any US funding agency. Our empirical exercise first quantifies the impact of the ARRA-NIH grant expansion in 2009 and 2010 on the ancestral diversity in GWAS subjects by leveraging the variations in *RRB* for minority subjects across diseases. We also perform a decomposition analysis distinguishing the impact of referral base for minorities from that of the base for whites. Lastly, we subject our findings to a battery of robustness checks, such as alternative treatment measures to the grant expansion and alternative estimation samples.

A. Ancestral Diversity in Subjects

Table 3 organizes the DID estimates from specifications with different controls and illuminates considerable contributions of the ARRA-NIH grant expansion to the growth of ancestral diversity in the GWAS subjects. The dependent variable is the share of minority subjects, a higher value of which demonstrates an improvement in diversity. The coefficient of the interaction between *post* and *RRB* represents the policy impact given a one-percentage point increase in *RRB*.

After controlling for the disease fixed effect and year fixed effect in column 1, the estimated coefficient of the interaction between *post* and *RRB* is 0.0281 and statistically significant. In our most preferred model specification in column 2, we additionally control for a set of disease research characteristics, as mentioned before, to capture the knowledge accumulation of the disease research and the inertia in funding decisions by US funding agencies. The coefficient of the key interaction term is 0.0341. Given that the average *RRB* is 2.95 in absolute value, this estimate suggests that the grant expansion leads the share of minority subjects to increase, on average, by almost two times ($\simeq 0.0341 \times 2.95/0.051$), equivalent to an increase of about ten percentage points. Against the background that minority ancestries are less underrepresented in GWAS (Mills and Rahal, 2019; Popejoy and Fullerton, 2016), our estimates imply an enrichment in ancestral diversity in GWAS subjects due to the grant expansion.

We also perform a more parsimonious event study following Equation (9), which allows us to examine the dynamic impact of the ARRA-NIH grant expansion over time. Figure 6 plots the coefficients of the interaction terms of RRB and the indicator of each year, and the associated 95% confidence intervals. The year 2009 serves as the reference year. As is readily apparent from the figure, the coefficients of the interactions in the pre-policy years (2007 and 2008) are 0.00237 and 0.00596, respectively. Their 95% confidence intervals cover zero, confirming minimal evidence of differential trends in the ancestral diversity prior to the grant expansion across disease research. It is also noteworthy that the ARRA-NIH grant expansion leads to the most pronounced jump in ancestral diversity in the window between 2009 and 2011. The coefficients are 0.0241 in 2010, double to 0.058 in 2011, then slightly fall to 0.0396 in 2012, and remains stable through 2016. The path of the coefficients shows a persistent and sizable effect in subsequent years even after the ARRA-NIH funding stimulus expired, in contrast to a diminishing pattern in the effect on the share of studies supported by ARRA-NIH grants in Figure 5. This comparison prompts us to explore the effect of the ARRA-NIH grant expansion on the scientific workforce in the next section.

B. Diversity at Different Scientific Stages

A GWAS project consists of a discovery sample (stage) and an independent replication sample (stage). Minority subjects could be involved in either sample or both. They help identify novel disease-associated genetic variants specific to the minority population itself at the discovery stage, and validate the previously identified associations, for example, from European white ancestry, at the replication stage. Exploring the influence of grants on the ancestral diversity in each scientific stage sheds light on how funding support contributes to knowledge creation.

The last two columns in Table 3 present estimates for each stage. The effects are broadly in line with our primary results. The ARRA-NIH grant expansion is estimated to lead the ancestral diversity to increase, on average, by 2.5 times ($\simeq 0.0225 \times 2.95/0.027$) in the discovery sample, and 1.8 times ($\simeq 0.0485 \times 2.95/0.078$) in the replication sample. These findings depict a possibly inspiring progress. Thanks to the grant expansion, more human genomics research with primary emphasis on the minority communities has been initiated; moreover, a growing number

of projects have directed their efforts to test whether their genetic findings are generalizable to the minority ancestries.

The event studies shown in Figure A3.3 reassure our findings in columns 3 and 4 in Table 3 and display a pattern similar to that in Figure 6. The overall patterns in both figures exhibit a significant jump in the impact immediately after the ARRA-NIH grant expansion was initiated. We also find no pre-existing trends for both stages.

C. Decomposition of the Impact of RRB

Recall that we assume that minority health research on diseases with a relatively lower cost of recruiting minority subjects are more feasible and thus more likely to be supported by ARRA-NIH grant. However, it remains unclear which part of *RRB* drives our findings thus far. To address this lingering conundrum, we distinguish the impact of the referral base for minorities from the impact of the base for whites using an event study following Equation (11):

$$
\begin{aligned}
y_{d,t} = \beta_0 + & \sum_{j=2007; j\neq 2009}^{2016} \zeta_{1,j} RBMinority_d^{2005} \times 1\{\text{year} = j\} \\
+ & \sum_{j=2007; j\neq 2009}^{2016} \zeta_{2,j} RBWhite_d^{2005} \times 1\{\text{year} = j\} + \beta_1 X_{d,t-1} + \gamma_d + \theta_t + \epsilon_{d,t},
\end{aligned} \tag{11}
$$

where $\zeta_{1,j}$ represents the expansion impact driven by the referral base for minorities in year j, while $\zeta_{2,j}$ reflects the impact driven by the referral base for whites in year j.

Figure 7 plots the estimated ζ_{1j} and ζ_{2j} from the event study. There are no discernible patterns showing differential trends in the ancestral diversity across diseases before the grant expansion. Following the funding expansion, with a one-percentage point increase in the referral base for minorities, the effect of funding expansion is positive, continuing through 2016. It experiences a sharp jump from 2009 to 2011, reaching a peak in 2011. The path of the funding impact stemming from a one-percentage point increase in the referral base for whites nearly mirrors that of the minorities but in the opposite direction.

The empirical patterns tied to both referral bases in the figure are consistent with the idea underlying our conceptual framework. A large pool of minority subject candidates, as evidenced

by a larger referral base for minorities, lowers the recruitment cost and increases the marginal scientific output per unit of funding invested in minority research. If the referral base for whites does not change, from Equation (7), we can infer that in equilibrium, the funding agencies are likely to raise funding support on minority studies. Thereby, it is not surprising to see positive ζ_{1j} in Figure 7.

D. Sensitivity Tests

Alternative Sample Specifications – In our primary sample, we include all available diseases with non-missing *RRB*. The heat map of the share of minority subjects by disease and over time in Figure A3.1 raises a concern that some diseases with a minimal number of publications in our study period may drive our results.[54] We assess the sensitivity of our findings to alternative sample specifications by removing disease research with the number of studies less than a given threshold in Panel A in Table 4. We present the results from the baseline sample in column 1 for comparison. Take column 2 as an example, we concentrate on a sample of diseases with a total number of studies no less than ten and perform analysis on the share of the minority subjects at each scientific stage, respectively. Results across each sample specification are effectively unchanged. Event studies corresponding to each cell are plotted in Figures A3.4, A3.5, A3.6, and A3.7.

Alternative Treatment Intensity Measure – Our identification thus far relies on the variations in the ancestral difference in the referral base in 2005. We caution that any sudden and temporary outbreak of some diseases around 2005 may significantly influence our *RRB* and possibly drive our results. To rule out this possibility, we re-estimate our models using an averaged *RRB* in a course from 2002 to 2005.[55] Results are summarized in Panel B in Table 4 and in Figures A3.4, A3.5, A3.6, and A3.7. The point estimates are all in line with our primary ones.

[54] For example, the "anemia and other deficiencies" has only one publication, and the "pneumonia" has four.

[55] We average the share of people with medical conditions across 2002 to 2005 in minority and white groups, respectively, and then calculate the difference between the two shares, which is called the average relative referral base.

VI. Impact on the Scientific Workforce

In the analysis of ancestral diversity among subjects, the impact of the ARRA-NIH grant expansion continues on a considerable scale through 2016. Considering that the ARRA initiative was temporary, lasting for only two years, a legitimate question to ask is what drives the long-lasting impact. We answer this question by examining how the ARRA-NIH grant expansion influences the scientific workforce in this section, with an particular emphasis on the academic activity of minority scientists and the collaboration patterns. We use the same sample as that used in the analysis of diversity in subjects.

A. Ancestral Diversity in Authors

The ancestral diversity in authors is measured by the share of minority authors among all authors in the studies on a given disease published each year. In our preferred specification, an author's surname is considered to be a minority name if its predicted probability of being a minority name is larger than 0.95. In robustness checks, we also consider thresholds of 0.90, 0.85, and 0.80 to define a minority name. Estimates from Equation (8) summarized in Table 5 show that minority author names appear increasingly in the GWAS publications as a consequence of the improved funding opportunity, a finding that is not sensitive to how we identify minority authors.

Specifically, in column 1, the coefficient of the interaction is 0.01, suggesting a statistically strong positive association between the ARRA-NIH grant expansion and the improved diversity in authors. After additionally controlling for disease research field characteristics in column 2, the effect remains, and, if anything, the point estimate becomes slightly larger in absolute value relative to column 1. With a calculation similar to that above, the grant expansion is associated, on average, with an approximately 57 percent ($\simeq 0.0145 \times 2.95/0.075$) increase in the share of minority authors, which equates to an increase of 4.28 percentage points ($\simeq 0.0145 \times 2.95$). Columns 3 to 5 report the results using alternative criteria to identify minority names, and the findings are unchanged.

It is not surprising to see the grant expansion facilitates the academic engagement of scien-

tists from minority ancestries, as evidenced by a large share of minority authors, as one of the missions of the ARRA-NIH initiative is to financially support the training of junior, minority, and female scientists, provide start-up packages, and fund pilot research. The entry of minority scientists could lead to a louder call for attention on the missing diversity in genomics research and spark the interest in minority studies. Their natural connection to the minority communities can also help alleviate mistrust and promote mutual-trust communications between the communities and scientific teams when they collaborate with senior white principal investigators (Burchard, 2014; Bentley, Callier and Rotimi, 2017).

Is the impact on scientific workforce transitory? Figure 8 displays the dynamics in the impact of the ARRA-NIH grant expansion by plotting the coefficients of the interactions between RRB with each year indicator from the event study. Prior to 2009, the point estimates slightly approach zero and are insignificant. However, since the beginning of the ARRA initiative, we have seen a sizable spike in the coefficients, with a peak in 2011. Thereafter, the effect gradually fades until 2014, but resurfaces in 2016. The path of coefficients yields several insights. The growing presence of minority authors in GWAS publications reveals that increased access to funding thanks to the ARRA-NIH grant expansion has facilitated the academic activities among minority scientists and diversified the scientific workforce. Furthermore, the impact, which continues to 2016, suggests that the academic positions for minority researchers created during the ARRA-NIH grant expansion are not temporary. The ARRA initiative has had a profound long-term impact on the academic community. One explanation for this long-term impact can be the career impact of the grant expansion on junior minority scientists. Studies such as Heggeness et al. (2016), Pickett (2019) and Conte et al. (2020) point out that fellowships and funding support have potential to promote retention of early-stage biomedical scientists and fuel their further funding.

We follow Equations (11) to distinguish the impact of the referral base for minorities on the diversity in authors from the impact of the base for whites. Figure 9 presents the event study estimates using the prediction cutoff of 0.95 to identify a minority author. The paths of coefficients resemble those in Figure 7. The effect of grant expansion driven by a one-percentage point increase in the referral base for minorities remains positive through 2016, with a rapid

jump from 2009 to 2011 and a peak in 2014; however, an equivalent change in the referral base for whites reverses the effect but exhibits similar magnitudes in absolute value. Figure A4.2 shows consistent findings using alternative criteria.

We also subject our analysis to a battery of robustness checks including alternative sample specifications, and alternative measures of treatment intensity. The results are organized in Tables A4.1 and A4.2. The conclusion remains unchanged.

B. Scientific Collaboration

Though we have seen an increasing academic activities of minority scientists thanks to the grant expansion, it is unclear what role they play in the scientific collaborations yet.

A GWAS project usually involves collaborations from multiple research centers in recruiting subjects, genetic testing, and genotype data analysis; therefore, it is not uncommon to see a GWAS article with more than ten authors. The highest credits are often given to the first few authors (hereafter *first authors*) and the last few authors (hereafter *last authors*). The former are those who make the major practical efforts leading the project and the latter are the principal investigators and fundraisers for the project.

Existing studies have suggested that research funding and the relevant funding mechanisms have a noticeable effect on scientific collaborations (Ubfal and Maffioli, 2011; Hoekman et al., 2013). Here, we focus on the first and last few authors, and examine the changes in the share of studies with a minority name listed in a given position before and after the ARRA-NIH grant expansion and across diseases.

The results are summarized in Table 6 with each column representing the dependent variable examined. We take the probability threshold of 0.95 to define if a name is from a minority ancestry. The estimates in the table shed light on the shift in scientific collaboration pattern in GWAS due to the grant expansion from at least two perspectives.

On the one hand, we see that the grant expansion contributes to an increasing presence of minority authors as the first authors. Estimate in column 1 shows that as a consequence of the ARRA-NIH grant expansion, the probability of a minority author as the first one author in a GWAS article, on average, increases by about 127 percent ($\simeq 0.0332 \times 2.95/0.077$), amounting

to 9.8 percentage points ($\simeq 0.0332 \times 2.95$). Furthermore, there is also consistent and discernible evidence in columns 2 and 3 showing a sizable increase in the presence of minority authors in the first two or first three author positions as a consequence of the improved funding opportunities.

On the other hand, we do not find noticeable evidence that the ARRA-NIH grant expansion has promoted the appearance of principal investigators or fundraisers from minority backgrounds, as shown in the last three columns in Table 6. Despite the fact that all point estimates are positive, they are statistically insignificant. Take column 4 as an example. The estimated effect is tantamount to increasing the possibility of the last author being a minority by 20.9 percent ($\simeq 0.0041 \times 2.95/0.058$), on average. Such effect size is far smaller compared to those in the first three columns. The event studies are presented in Figure A4.3.

With caution, our analysis suggests that the ARRA-NIH grant expansion seems to promote more collaborations between minority scientists with white PIs rather than more minority PIs leading the minority health research. As the grant expansion softens the budgetary constraints on minority health research, minority scientists can leverage their natural ties to their communities to a greater extent, giving them an advantage in participating in and leading these projects. Nevertheless, minority scientists remain underfunded, as reflected by fewer minority PIs (Hoppe et al., 2019; Jimenez et al., 2019; Erosheva et al., 2020).

VII. Impact on Scientific Discovery

To better capture the scientific advances contributed by improved inclusion of subjects from minority backgrounds in GWAS, we particularly concentrate on the SNP-disease associations identified from studies in which all subjects in the discovery stage are minorities, which we refer to as the minority-based SNP-Associations. We also subject our analysis to alternative definitions of the minority-based SNP-disease associations to reassure robustness. In the following empirical exercise, we construct two disease-level metrics regarding the minority-based SNP-disease associations: the absolute count each year and the share in total SNP-disease associations each year. We perform the same DID model and event study analysis following Equations (8) and (9) in the same sample as that in the analysis of diversity in subjects.

Table 7 presents our main results, with the first three columns focusing on the count of

minority-based SNPs and the last three columns on the share. Point estimates across all columns speak to an increase in genetic discoveries driven by the ARRA-NIH grant expansion. In columns 1 and 4, considering the absolute value of the average RRB is 2.95, the expansion in funding opportunities leads to 7.76 ($\simeq 2.63 \times 2.95$) more SNPs per year discovered from the minorities, which is tantamount to an increase of 9.71 percentage points ($\simeq 0.0329 \times 2.95$) in the share of minority-based SNPs. Figure 10 exhibits the event studies corresponding to columns 1 and 4. In both figures, we see a noticeable upward trend in coefficients following the ARRA initiative and a significant jump in the absolute effect size from 2009 to 2011. The effect of the funding expansion on the count of SNPs keeps increasing through 2016, while the effect on the share vacillates, reaching the peak in 2011.

In other columns in Table 7, we redefine the minority-based SNP-disease associations using alternative criteria. In columns 2 and 5, we slightly relax the criterion by including SNPs from studies in which at least 60 percent of the subjects in the discovery stage are minorities. In contrast, in columns 3 and 6, we impose a more stringent criterion, requiring that all subjects in both discovery and replication stages are minorities. The results in the four columns are consistent. Event studies are plotted in Figures A5.2.

Tables A5.1 and A5.2 organize the robustness checks based on alternative sample specifications and alternative measures of treatment intensity. The estimates are consistent.

VIII. Discussion

The NIH has initiated a variety of funding opportunities to support science on underrepresented populations in recent years. In 2015, the NIH initiated the "All of US" research program, aiming to build a more diverse health database for developing better-individualized health care treatments. In 2019, the NIH-affiliated National Institute on Minority Health and Health Disparities (NIMHD) awarded a total of $187 million in five years to three new awards and renewed awards from eight research centers in the Minority Institutions Specialized Center program.[56] Existing discussions on efficacy of these initiatives are descriptive, and rigorous quantitative assessments are imperative but remain inadequate still.

[56] https://www.nih.gov/news-events/news-releases/nih-funds-eleven-research-centers-minority-institutions.

Our paper timely fills this gap by empirically answering two central questions: whether improved funding opportunity helps facilitate minority health research, and how significant is the scientific value of funding science on underrepresented populations. The answers to these two questions are essential for policymakers when designating research grant allocation schemes and designing funding mechanisms.

We contribute to the literature on returns to public research investment in two aspects. First, our paper is among the few that investigate the role of public funding agencies in addressing the widening ancestral disparity in human genomics research. A steady stream of empirical studies has assessed the social benefit of publicly funded science from a wide array of angles, such as increasing productivity of the pharmaceutical industry (Cockburn and Henderson, 2000; Nayak, Avorn and Kesselheim, 2019; Cleary, Jackson and Ledley, 2020), knowledge production (Jacob and Lefgren, 2011; Li and Agha, 2015; Rosenbloom et al., 2015; Packalen and Bhattacharya, 2018; Azoulay, Fons-Rosen and Graff Zivin, 2019), and firm innovation (Bronzini and Iachini, 2014; Dechezleprêtre et al., 2016; Li, Azoulay and Sampat, 2017; Azoulay et al., 2019). Unlike them, our analysis concentrates on diversity and inclusion in genomics research. Increasing evidence on the lack of diversity in genomics research raises concerns regarding the distributive justice in the translation of biomedical innovations into human health across populations in the approaching era of genomic medicine (Hindorff et al., 2018; Wojcik et al., 2019). Advocates for targeting specific funding efforts towards minority health research have grown louder (Chen, 2019; Carnethon, Kershaw and Kandula, 2020; Zhang et al., 2020; Pichon, 2021). Against this background, evaluations on diversity, inclusion and equity policies deserve more attention.

Our paper also contributes to the literature tackling the lack of diversity in the scientific community. The disproportionate underrepresentation of female and minority faculties persists in various STEM fields (Puritty et al., 2017; AlShebli, Rahwan and Woon, 2018; Menon, 2021) and economics (Bayer and Rouse, 2016; Buckles, 2019; Boustan and Langan, 2019). Existing literature has proposed several approaches to address this disparity, such as revising the reward system in science (Hofstra et al., 2020; Davies et al., 2021), targeted mentoring (Guevara et al., 2013; Bayer and Rouse, 2016; Buckles, 2019), and improved support and opportunities (Xu, 2008; Khan et al., 2019). Evaluating the efficacy of these approaches in reality requires

more empirical evidence. Our analysis provides rigorous quantitative assessment on the role of funding support. We show that increasing funding opportunities helps facilitate the diversity in scientific workforce, as evidenced by increased scientific activities by minority researchers and more minority leaderships in the scientific collaborations.

Our study does not run counter to studies such as Li and Agha (2015), Li (2017), and Gallo, Sullivan and Glisson (2016). We are similar in spirit, sharing the same interests in how funding, in general, affects science, though we approach the core from different angles. Their studies carefully scrutinize the existing grant mechanisms and shed light on the impact of grant reviewers' preference, expertise proximity, and social relations in the grant review process on academic resource allocation given a fixed funding pool. For example, the NIH funding mechanisms are criticized for widening gender disparity (Boyle et al., 2015; Hechtman et al., 2018; Nikaj et al., 2018; Grogan, 2019; Witteman et al., 2019) and racial disparity (Ginther et al., 2011; Reardon, 2014; Hayden, 2015; Nikaj et al., 2018) in the grant recipients. Their findings awaken the awareness to refine the existing mechanisms to narrow funding disparity. Our study supplements their analysis with a new perspective. We attempt to answer the question: given the existing funding mechanisms, would a bigger "cake" help the underfed science grow and catch up steadily?

This paper provides evidence supporting that NIH's recent funding initiatives targeting minority health research have made a difference. Our findings show that apart from modifying existing funding mechanisms, scaled financial input may be potentially helpful in addressing the missing diversity. Although we focus on ancestral diversity, with careful considerations, the underlying takeaways studied here could be applied to future studies on diversity in gender or junior/senior scientists.

REFERENCES

Acemoglu, Daron, and Joshua Linn. 2004. "Market size in innovation: Theory and evidence from the pharmaceutical industry." *The Quarterly Journal of Economics*, 119(3): 1049–1090.

Adeyemo, Adebowale, and Charles Rotimi. 2010. "Genetic variants associated with complex human diseases show wide variation across multiple populations." *Public Health Genomics*, 13(2): 72–79.

AlShebli, Bedoor K, Talal Rahwan, and Wei Lee Woon. 2018. "The preeminence of ethnic diversity in scientific collaboration." *Nature Communications*, 9(1): 1–10.

Azoulay, Pierre, Christian Fons-Rosen, and Joshua S Graff Zivin. 2019. "Does science advance one funeral at a time?" *American Economic Review*, 109(8): 2889–2920.

Azoulay, Pierre, Joshua S Graff Zivin, Danielle Li, and Bhaven N Sampat. 2019. "Public R&D investments and private-sector patenting: Evidence from NIH funding rules." *The Review of Economic Studies*, 86(1): 117–152.

Bayer, Amanda, and Cecilia Elena Rouse. 2016. "Diversity in the economics profession: A new attack on an old problem." *Journal of Economic Perspectives*, 30(4): 221–42.

Beavis, Anna L, Patti E Gravitt, and Anne F Rositch. 2017. "Hysterectomy-corrected cervical cancer mortality rates reveal a larger racial disparity in the United States." *Cancer*, 123(6): 1044–1050.

Bentley, Amy R, Shawneequa Callier, and Charles N Rotimi. 2017. "Diversity and inclusion in genomic research: Why the uneven progress?" *Journal of Community Genetics*, 8(4): 255–266.

Bonàs-Guarch, Sílvia, Marta Guindo-Martínez, Irene Miguel-Escalada, Niels Grarup, David Sebastian, Elias Rodriguez-Fos, Friman Sánchez, Mercè Planas-Fèlix, Paula Cortes-Sánchez, Santi González, et al. 2018. "Re-analysis of public genetic data reveals a rare X-chromosomal variant associated with type 2 diabetes." *Nature Communications*, 9(1): 1–14.

Bonham, Vence L, Shawneequa L Callier, and Charmaine D Royal. 2016. "Will precision medicine move us beyond race?" *The New England Journal of Medicine*, 374(21): 2003.

Boustan, Leah, and Andrew Langan. 2019. "Variation in Women's Success across PhD Programs in Economics." *Journal of Economic Perspectives*, 33(1): 23–42.

Boyle, Paul J, Lucy K Smith, Nicola J Cooper, Kate S Williams, and Henrietta O'Connor. 2015. "Gender balance: Women are funded more fairly in social science." *Nature*, 525(7568): 181–183.

Bronzini, Raffaello, and Eleonora Iachini. 2014. "Are incentives for R&D effective? Evidence from a regression discontinuity approach." *American Economic Journal: Economic Policy*, 6(4): 100–134.

Buckles, Kasey. 2019. "Fixing the leaky pipeline: Strategies for making economics work for women at every stage." *Journal of Economic Perspectives*, 33(1): 43–60.

Burchard, Esteban G. 2014. "Medical research: Missing patients." *Nature News*, 513(7518): 301.

Burchard, Esteban G, Sam S Oh, Marilyn G Foreman, and Juan C Celedón. 2015. "Moving toward true inclusion of racial/ethnic minorities in federally funded studies. A key step for achieving respiratory health equality in the United States." *American Journal of Respiratory and Critical Care Medicine*, 191(5): 514–521.

Carnethon, Mercedes R, Kiarri N Kershaw, and Namratha R Kandula. 2020. "Disparities research, disparities researchers, and health equity." *JAMA*, 323(3): 211–212.

Cazzola, Mario, Paola Rogliani, Alessandro Sanduzzi, and Maria Gabriella Matera. 2015. "Influence of ethnicity on response to Asthma drugs." *Expert Opinion on Drug Metabolism & Toxicology*, 11(7): 1089–1097.

Chen Jr, Moon S, Primo N Lara, Julie HT Dang, Debora A Paterniti, and Karen Kelly. 2014. "Twenty years post-NIH Revitalization Act: Enhancing minority participation in clinical trials (EMPaCT): Laying the groundwork for improving minority clinical trial accrual: Renewing the case for enhancing minority participation in cancer clinical trials." *Cancer*, 120: 1091–1096.

Chen, Moon S. 2019. "Rectifying disparities in funding of Asian American, Native Hawaiian, and Pacific Islander research by the US National Institutes of Health." *JAMA Network Open*, 2(7): e197561–e197561.

Chung, Jaeyoon, Gyungah R Jun, Josée Dupuis, and Lindsay A Farrer. 2019. "Comparison of methods for multivariate gene-based association tests for complex diseases using common variants." *European Journal of Human Genetics*, 27(5): 811–823.

Cleary, Ekaterina, Matthew J Jackson, and Fred Ledley. 2020. "Government as the first investor in biopharmaceutical innovation: Evidence from new drug approvals 2010–2019." *https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3731819*.

Cockburn, Iain M, and Rebecca M Henderson. 2000. "Publicly funded science and the productivity of the pharmaceutical industry." *Innovation Policy and the Economy*, 1: 1–34.

Cohn, Elizabeth G, Gail E Henderson, and Paul S Appelbaum. 2017. "Distributive justice, diversity, and inclusion in precision medicine: What will success look like?" *Genetics in Medicine*, 19(2): 157–159.

Conte, Marisa, Santiago Schnell, Adrienne Ettinger, and Bishr Omary. 2020. "Trends in NIH-supported career development funding: implications for institutions, trainees, and the future research workforce." *JCI Insight*, 5(17).

Davies, Sarah W, Hollie M Putnam, Tracy Ainsworth, Julia K Baum, Colleen B Bove, Sarah C Crosby, Isabelle M Côté, Anne Duplouy, Robinson W Fulweiler, Alyssa J Griffin, et al. 2021. "Promoting inclusive metrics of success and impact to dismantle a discriminatory reward system in science." *PLoS Biology*, 19(6): e3001282.

De Boeck, Kris, and Margarida D Amaral. 2016. "Progress in therapies for cystic fibrosis." *The Lancet Respiratory Medicine*, 4(8): 662–674.

Dechezleprêtre, Antoine, Elias Einiö, Ralf Martin, Kieu-Trang Nguyen, and John Van Reenen. 2016. "Do tax incentives for research increase firm innovation? An RD design for R&D." National Bureau of Economic Research.

Dubois, Pierre, Olivier De Mouzon, Fiona Scott-Morton, and Paul Seabright. 2015. "Market size and pharmaceutical innovation." *The RAND Journal of Economics*, 46(4): 844–871.

Erosheva, Elena A, Sheridan Grant, Mei-Ching Chen, Mark D Lindner, Richard K Nakamura, and Carole J Lee. 2020. "NIH peer review: Criterion scores completely account for racial disparities in overall impact scores." *Science Advances*, 6(23): eaaz4868.

Feng, Lisa B, Scott D Grosse, Ridgely Fisk Green, Aliza K Fink, and Gregory S Sawicki. 2018. "Precision medicine in action: The impact of Ivacaftor on cystic fibrosis–related hospitalizations." *Health Affairs*, 37(5): 773–779.

Gallo, Stephen A, Joanne H Sullivan, and Scott R Glisson. 2016. "The influence of peer reviewer expertise on the evaluation of research funding applications." *PLoS One*, 11(10): e0165147.

George, Sheba, Nelida Duran, and Keith Norris. 2014. "A systematic review of barriers and facilitators to minority research participation among African Americans, Latinos, Asian Americans, and Pacific Islanders." *American Journal of Public Health*, 104(2): e16–e31.

Ginther, Donna K, Walter T Schaffer, Joshua Schnell, Beth Masimore, Faye Liu, Laurel L Haak, and Raynard Kington. 2011. "Race, ethnicity, and NIH research awards." *Science*, 333(6045): 1015–1019.

Grogan, Kathleen E. 2019. "How the entire scientific community can confront gender bias in the workplace." *Nature Ecology & Evolution*, 3(1): 3.

Guevara, James P, Emem Adanga, Elorm Avakame, and Margo Brooks Carthon. 2013. "Minority faculty development programs and underrepresented minority faculty representation at US medical schools." *JAMA*, 310(21): 2297–2304.

Guttmacher, Alan E, and Francis S Collins. 2003. "Welcome to the genomic era."

Hanna, Michael. 2015. "Matching taxpayer funding to population health needs." *Circulation Research*, 116(8): 1296–1300.

Hayden, Erika. 2015. "Racial bias continues to haunt NIH grants." *Nature News*, 527(7578): 286.

Hechtman, Lisa A, Nathan P Moore, Claire E Schulkey, Andrew C Miklos, Anna Maria Calcagno, Richard Aragon, and Judith H Greenberg. 2018. "NIH funding longevity by gender." *Proceedings of the National Academy of Sciences*, 115(31): 7943–7948.

Heggeness, Misty L, Lisa Evans, Jennifer Reineke Pohlhaus, and Sherry L Mills. 2016. "Measuring diversity of the National Institutes of Health-funded workforce." *Academic Medicine: Journal of the Association of American Medical Colleges*, 91(8): 1164.

Hindorff, Lucia A, Vence L Bonham, Lawrence C Brody, Margaret EC Ginoza, Carolyn M Hutter, Teri A Manolio, and Eric D Green. 2018. "Prioritizing diversity in human genomics research." *Nature Reviews Genetics*, 19(3): 175.

Hoekman, Jarno, Thomas Scherngell, Koen Frenken, and Robert Tijssen. 2013. "Acquisition of European research funds and its effect on international scientific collaboration." *Journal of Economic Geography*, 13(1): 23–52.

Hofstra, Bas, Vivek V Kulkarni, Sebastian Munoz-Najar Galvez, Bryan He, Dan Jurafsky, and Daniel A McFarland. 2020. "The diversity–innovation paradox in science." *Proceedings of the National Academy of Sciences*, 117(17): 9284–9291.

Hoppe, Travis A, Aviva Litovitz, Kristine A Willis, Rebecca A Meseroll, Matthew J Perkins, B Ian Hutchins, Alison F Davis, Michael S Lauer, Hannah A Valantine, James M Anderson, et al. 2019. "Topic choice contributes to the lower rate of NIH awards to African-American/Black scientists." *Science Advances*, 5(10): eaaw7238.

Huang, Jie, Maria Sabater-Lleal, Folkert W Asselbergs, David Tregouet, So-Youn Shin, Jingzhong Ding, Jens Baumert, Tiphaine Oudot-Mellakh, Lasse Folkersen, Andrew D Johnson, et al. 2012. "Genome-wide association study for circulating levels of PAI-1 provides novel insights into its regulation." *Blood, The Journal of the American Society of Hematology*, 120(24): 4873–4881.

Jacob, Brian A, and Lars Lefgren. 2011. "The impact of research grant funding on scientific productivity." *Journal of Public Economics*, 95(9-10): 1168–1177.

Jimenez, Miguel F, Theresa M Laverty, Sara P Bombaci, Kate Wilkins, Drew E Bennett, and Liba Pejchar. 2019. "Underrepresented faculty play a disproportionate role in advancing diversity and inclusion." *Nature Ecology & Evolution*, 3(7): 1030–1033.

Johnson, Julie, and Larisa Cavallari. 2015. "Warfarin pharmacogenetics." *Trends in Cardiovascular Medicine*, 25(1): 33–41.

Kaiser, Jocelyn. 2009. "NIH stimulus plan triggers flood of applications–and Anxiety."

Khan, Mishal Sameer, Fatim Lakha, Melisa Mei Jin Tan, Shweta Rajkumar Singh, Rina Yu Chin Quek, Emeline Han, See Mieng Tan, Victoria Haldane, Montserrat Gea-Sánchez, and Helena Legido-Quigley. 2019. "More talk than action: Gender and ethnic diversity in leading public health universities." *The Lancet*, 393(10171): 594–600.

Klein, Robert J, Caroline Zeiss, Emily Y Chew, Jen-Yue Tsai, Richard S Sackler, Chad Haynes, Alice K Henning, John Paul SanGiovanni, Shrikant M Mane, Susan T Mayne, et al. 2005. "Complement factor H polymorphism in age-related macular degeneration." *Science*, 308(5720): 385–389.

Koutsodendris, Nicole, Maxine R Nelson, Antara Rao, and Yadong Huang. 2021. "Apolipoprotein E and Alzheimer's Disease: Findings, hypotheses, and potential mechanisms." *Annual Review of Pathology: Mechanisms of Disease*, 17: 73–99.

Landry, Latrice G, Nadya Ali, David R Williams, Heidi L Rehm, and Vence L Bonham. 2018. "Lack of diversity in genomic databases is a barrier to translating precision medicine research into practice." *Health Affairs*, 37(5): 780–785.

Lemanske Jr, Robert F, David T Mauger, Christine A Sorkness, Daniel J Jackson, Susan J Boehmer, Fernando D Martinez, Robert C Strunk, Stanley J Szefler, Robert S Zeiger, Leonard B Bacharier, et al. 2010. "Step-up therapy for children with uncontrolled Asthma receiving inhaled Corticosteroids." *New England Journal of Medicine*, 362(11): 975–985.

Li, Danielle. 2017. "Expertise versus Bias in Evaluation: Evidence from the NIH." *American Economic Journal: Applied Economics*, 9(2): 60–92.

Li, Danielle, and Leila Agha. 2015. "Big names or big ideas: Do peer-review panels select the best science proposals?" *Science*, 348(6233): 434–438.

Li, Danielle, Pierre Azoulay, and Bhaven N Sampat. 2017. "The applied value of public investments in biomedical research." *Science*, 356(6333): 78–81.

Loree, Jonathan M, Seerat Anand, Arvind Dasari, Joseph M Unger, Anirudh Gothwal, Lee M Ellis, Gauri Varadhachary, Scott Kopetz, Michael J Overman, and

Kanwal Raghav. 2019. "Disparity of race reporting and representation in clinical trials leading to cancer drug approvals from 2008 to 2018." *JAMA Oncology*, 5(10): e191870–e191870.

Mak, Angel CY, Marquitta J White, Walter L Eckalbar, Zachary A Szpiech, Sam S Oh, Maria Pino-Yanes, Donglei Hu, Pagé Goddard, Scott Huntsman, Joshua Galanter, et al. 2018. "Whole-genome sequencing Of pharmacogenetic drug response in racially diverse children with Asthma." *American Journal of Respiratory and Critical Care Medicine*, 197(12): 1552–1564.

Manrai, Arjun K, Birgit H Funke, Heidi L Rehm, Morten S Olesen, Bradley A Maron, Peter Szolovits, David M Margulies, Joseph Loscalzo, and Isaac S Kohane. 2016. "Genetic misdiagnoses and the potential for health disparities." *New England Journal of Medicine*, 375(7): 655–665.

Martin, Alicia R, Masahiro Kanai, Yoichiro Kamatani, Yukinori Okada, Benjamin M Neale, and Mark J Daly. 2019. "Clinical use of current polygenic risk scores may exacerbate health disparities." *Nature Genetics*, 51(4): 584.

Menon, Binuraj RK. 2021. "The missing colours of chemistry." *Nature Chemistry*, 13(2): 101–106.

Mills, Melinda C, and Charles Rahal. 2019. "A scientometric review of genome-wide association studies." *Communications Biology*, 2(1): 1–11.

Nayak, Rahul K, Jerry Avorn, and Aaron S Kesselheim. 2019. "Public sector financial support for late stage discovery of new drugs in the United States: Cohort study." *The BMJ*, 367.

Nikaj, Silda, Deepshikha Roychowdhury, P Kay Lund, Marguerite Matthews, and Katrina Pearson. 2018. "Examining trends in the diversity of the US National Institutes of Health participating and funded workforce." *The FASEB Journal*, 32(12): 6410–6422.

Ntzani, Evangelia E, George Liberopoulos, Teri A Manolio, and John PA Ioannidis. 2012. "Consistency of genome-wide associations across major ancestral groups." *Human Genetics*, 131(7): 1057–1071.

Đoàn, Lan N, Yumie Takata, Kari-Lyn K Sakuma, and Veronica L Irvin. 2019. "Trends in clinical research including Asian American, Native Hawaiian, and Pacific Islander participants funded by the US National Institutes of Health, 1992 to 2018." *JAMA Network Open*, 2(7): e197432–e197432.

Oh, Sam S, Joshua Galanter, Neeta Thakur, Maria Pino-Yanes, Nicolas E Barcelo, Marquitta J White, Danielle M de Bruin, Ruth M Greenblatt, Kirsten Bibbins-Domingo, Alan HB Wu, et al. 2015. "Diversity in clinical and biomedical research: A promise yet to be fulfilled." *PLoS Medicine*, 12(12): e1001918.

Ortega, Victor E, Gregory A Hawkins, Wendy C Moore, Annette T Hastie, Elizabeth J Ampleford, William W Busse, Mario Castro, Domingo Chardon, Serpil C Erzurum, Elliot Israel, et al. 2014. "Effect of rare variants in *ADRB*2 on risk of severe exacerbations and symptom control during longacting β agonist treatment in a multiethnic Asthma population: A genetic study." *The Lancet Respiratory Medicine*, 2(3): 204–213.

Packalen, Mikko, and Jay Bhattacharya. 2018. "Does the NIH Fund Edge Science?" National Bureau of Economic Research.

Packalen, Mikko, and Jay Bhattacharya. 2020. "NIH funding and the pursuit of edge science." *Proceedings of the National Academy of Sciences*, 117(22): 12011–12016.

Peters, Stephen P, Susan J Kunselman, Nikolina Icitovic, Wendy C Moore, Rodolfo Pascual, Bill T Ameredes, Homer A Boushey, William J Calhoun, Mario Castro, Reuben M Cherniack, et al. 2010. "Tiotropium bromide step-up therapy for adults with uncontrolled Asthma." *New England Journal of Medicine*, 363(18): 1715–1726.

Pichon, Anne. 2021. "Funding a more equitable research community." *Nature Chemistry*, 13(5): 387–389.

Pickett, Christopher. 2019. "The increasing importance of fellowships and career development awards in the careers of early-stage biomedical academic researchers." *PloS One*, 14(10): e0223876.

Pier, Elizabeth L, Markus Brauer, Amarette Filut, Anna Kaatz, Joshua Raclaw, Mitchell J Nathan, Cecilia E Ford, and Molly Carnes. 2018. "Low agreement among reviewers evaluating the same NIH grant applications." *Proceedings of the National Academy of Sciences*, 115(12): 2952–2957.

PMC. 2019. "Personalized medicine at FDA: A progress and outlook report 2018." *http://www.personalizedmedicinecoalition.org/Userfiles/PMC-Corporate/file/PM_at_FDA_A_Progress_and_Outlook_Report.pdf*.

Popejoy, Alice B, and Stephanie M Fullerton. 2016. "Genomics is failing on diversity." *Nature News*, 538(7624): 161.

Puritty, Chandler, Lynette R Strickland, Eanas Alia, Benjamin Blonder, Emily Klein, Michel T Kohl, Earyn McGee, Maclovia Quintana, Robyn E Ridley, Beth Tellman, et al. 2017. "Without inclusion, diversity initiatives may not be enough." *Science*, 357(6356): 1101–1102.

Ramos, Edward, Shawneequa L Callier, and Charles N Rotimi. 2012. "Why personalized medicine will fail if we stay the course." *Personalized Medicine*, 9(8): 839–847.

Reardon, Sara. 2014. "NIH to probe racial disparity in grant awards." *Nature*, 512(7514): 243.

Roodman, David, Morten Ørregaard Nielsen, James G MacKinnon, and Matthew D Webb. 2019. "Fast and wild: Bootstrap inference in Stata using boottest." *The Stata Journal*, 19(1): 4–60.

Rosenberg, Noah A, Lucy Huang, Ethan M Jewett, Zachary A Szpiech, Ivana Jankovic, and Michael Boehnke. 2010. "Genome-wide association studies in diverse populations." *Nature Reviews Genetics*, 11(5): 356–366.

Rosenbloom, Joshua L, Donna K Ginther, Ted Juhl, and Joseph A Heppert. 2015. "The effects of research & development funding on scientific productivity: Academic chemistry, 1990-2009." *PloS One*, 10(9): e0138176.

Roses, Allen D. 2003. "The genome era begins..." *Nature Genetics*, 33(3): 217–217.

Sardar, Muhammad Rizwan, Marwan Badri, Catherine T Prince, Jonathan Seltzer, and Peter R Kowey. 2014. "Underrepresentation of women, elderly patients, and racial minorities in the randomized trials used for cardiovascular guidelines." *JAMA Internal Medicine*, 174(11): 1868–1870.

Sekar, Kavya, Congressional Research Service. January 22, 2020. "National Institutes of Health (NIH) funding: FY1994-FY2020." *https://pesquisa.busalud.org/global-literature-on-novel-coronavirus-2019-ncov/resource/pt/grc-740681*.

Singh, Gopal K, Michael D Kogan, and Rebecca T Slifkin. 2017. "Widening disparities in infant mortality and life expectancy between Appalachia and the rest of the United States, 1990–2013." *Health Affairs*, 36(8): 1423–1432.

Sirugo, Giorgio, Scott M Williams, and Sarah A Tishkoff. 2019. "The missing diversity in human genetic studies." *Cell*, 177(1): 26–31.

Sorin, Martin D, and Randall Hannum. 2014. "Which extramural scientists were funded by NIH from its ARRA Funds?" *https://academicworks.cuny.edu/ny_pubs/196/*.

Stewart, Anita L, Anna Maria Nápoles, Sorbarikor Piawah, Jasmine Santoyo-Olsson, and Jeanne A Teresi. 2020. "Guidelines for evaluating the feasibility of recruitment in pilot studies of diverse populations: an overlooked but important component." *Ethnicity & Disease*, 30(Suppl 2): 745.

Taffe, Michael A, and Nicholas W Gilpin. 2021. "Equity, Diversity and Inclusion: Racial inequity in grant funding from the US National Institutes of Health." *elife*, 10: e65697.

Ubfal, Diego, and Alessandro Maffioli. 2011. "The impact of funding on research collaboration: Evidence from a developing country." *Research Policy*, 40(9): 1269–1279.

Urban, Kiernan, and Stephen Naylor. 2018. "Precision medicine drugs." *Drug Discovery World*, 9.

Wechsler, Michael E, Mario Castro, Erik Lehman, Vernon M Chinchilli, E Rand Sutherland, Loren Denlinger, Stephen C Lazarus, Stephen P Peters, and Elliot Israel. 2011. "Impact of race on Asthma treatment failures in the Asthma clinical research network." *American Journal of Respiratory and Critical Care Medicine*, 184(11): 1247–1253.

Wechsler, Michael E, Stanley J Szefler, Victor E Ortega, Jacqueline A Pongracic, Vernon Chinchilli, John J Lima, Jerry A Krishnan, Susan J Kunselman, David Mauger, Eugene R Bleecker, et al. 2019. "Step-up therapy in black children and adults with poorly controlled Asthma." *New England Journal of Medicine*, 381(13): 1227–1239.

Welter, Danielle, Jacqueline MacArthur, Joannella Morales, Tony Burdett, Peggy Hall, Heather Junkins, Alan Klemm, Paul Flicek, Teri Manolio, Lucia Hindorff, et al. 2014. "The NHGRI GWAS Catalog, a curated resource of SNP-trait associations." *Nucleic Acids Research*, 42(D1): D1001–D1006.

Witteman, Holly O, Michael Hendricks, Sharon Straus, and Cara Tannenbaum. 2019. "Are gender gaps due to evaluations of the applicant or the science? A natural experiment at a national funding agency." *The Lancet*, 393(10171): 531–540.

Wojcik, Genevieve L, Mariaelisa Graff, Katherine K Nishimura, Ran Tao, Jeffrey Haessler, Christopher R Gignoux, Heather M Highland, Yesha M Patel, Elena P Sorokin, Christy L Avery, et al. 2019. "Genetic analyses of diverse populations improves discovery for complex traits." *Nature*, 570(7762): 514–518.

Xu, Jiaquan, Sherry L Murphy, Kenneth D Kochanek, Brigham Bastian, and Elizabeth Arias. 2018. "Deaths: Final data for 2016." *https://stacks.cdc.gov/view/cdc/*

57989.

Xu, Yonghong Jade. 2008. "Gender disparity in STEM disciplines: A study of faculty attrition and turnover intentions." *Research in Higher Education*, 49(7): 607–624.

Zhang, Mingyu, Brooke A Jarrett, Keri N Althoff, Frances S Burman, Laura Camarata, Sally B Coburn, Aisha S Dickerson, Kathryn Foti, Maneet Kaur, Kathryn M Leifheit, et al. 2020. "Recommendations to the Society for Epidemiologic Research for further promoting diversity and inclusion at the annual meeting and beyond." *American Journal of Epidemiology*, 189(10): 1037–1041.

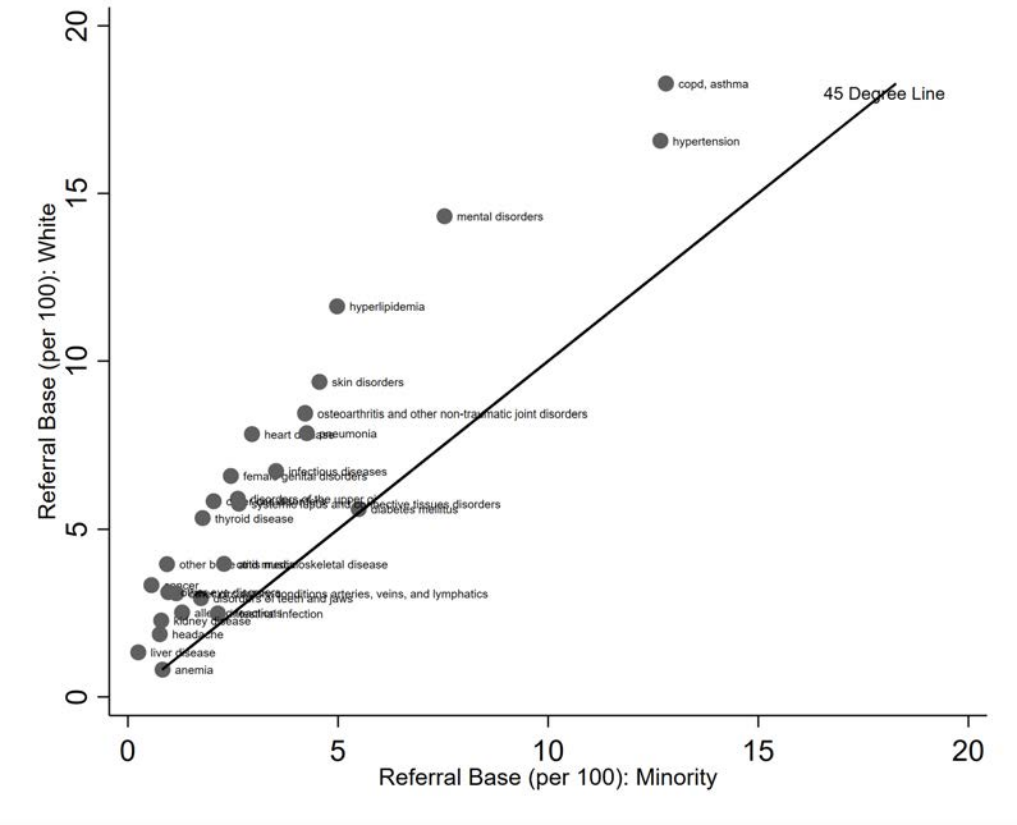
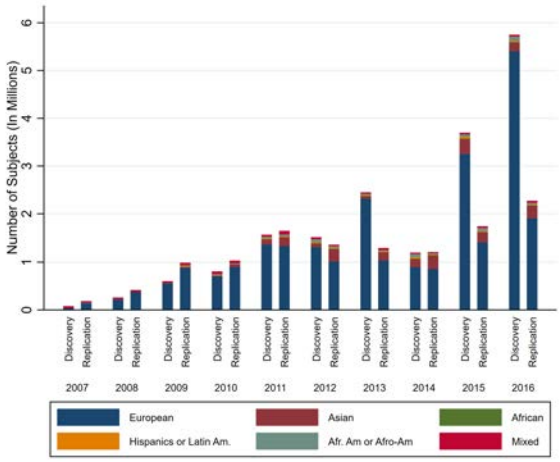
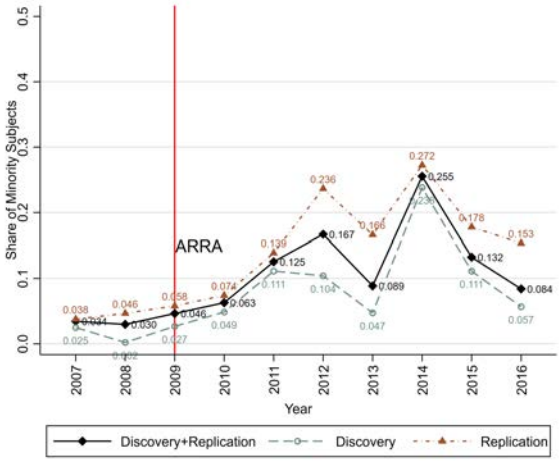


Figure 1. Referral Bases for Whites and Minorities

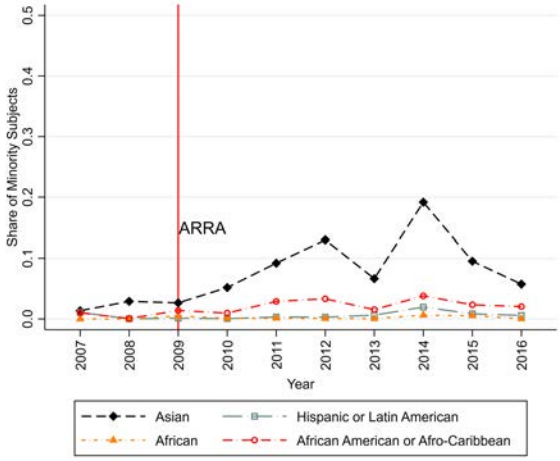
Notes: The figure is created by the author based on the MEPS summary tables. The 27 diseases in our final sample with non-missing number of people with medical conditions are plotted in the figure. Respondents in MEPS are classified into five ancestry groups: (1) Hispanic, (2) Black, (3) American Indian, Alaska Native, and multi-races, (4) Asian, Hawaiian and Pacific Islander, and (5) European white. The minority population in our following analysis contains all the mentioned ancestries except the European white.



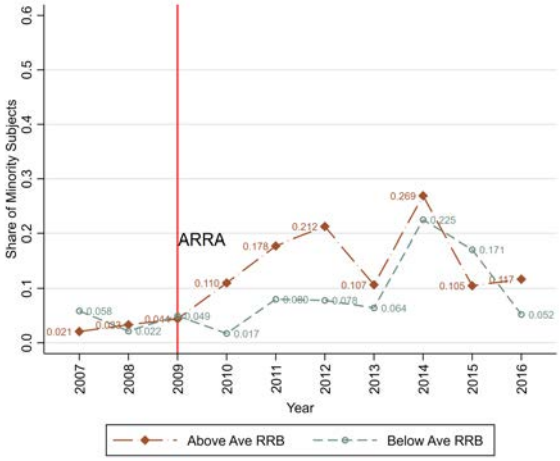
(A) Number of GWAS Subjects over Years



(B) Share of Minority Subjects over Years



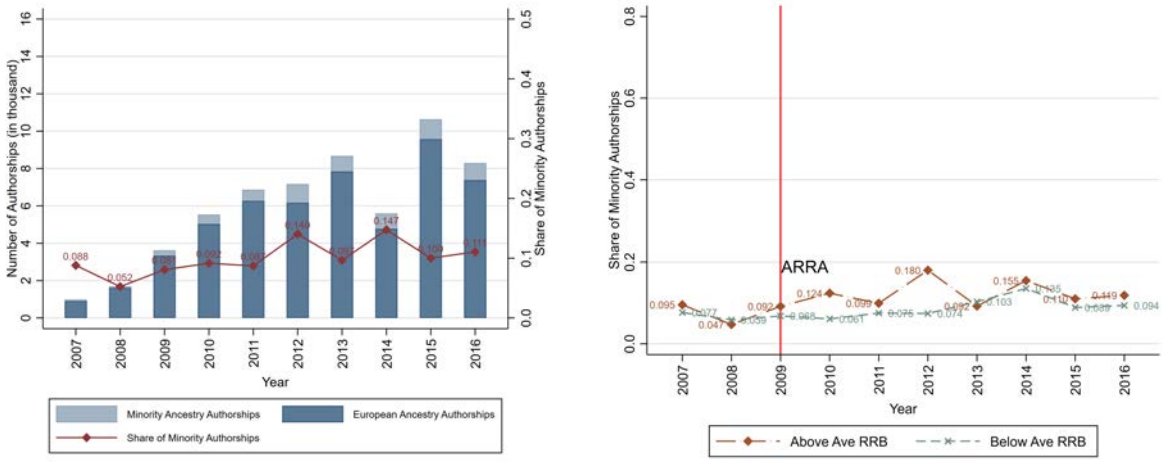
(C) Share of Minority Subjects By Ancestries



(D) Quasi-Treated vs Quasi-Control

FIGURE 2. DIVERSITY IN GWAS SUBJECTS

Notes: The sample comprises of 27 broader diseases from 2007 to 2016, in total 183 disease-year pairs. Given minimal GWAS publications in 2005 and 2006, the data in 2007 presented in the figure includes subjects from GWAS publications in 2005, 2006 and 2007. We separate the 27 broader disease categories into a "quasi-treated" and a "quasi-control" group based on their *RRB*. The "quasi-treated" includes 13 diseases with *RRB* larger than the mean of the 27 diseases, and the "quasi-control" includes the other 14 diseases.

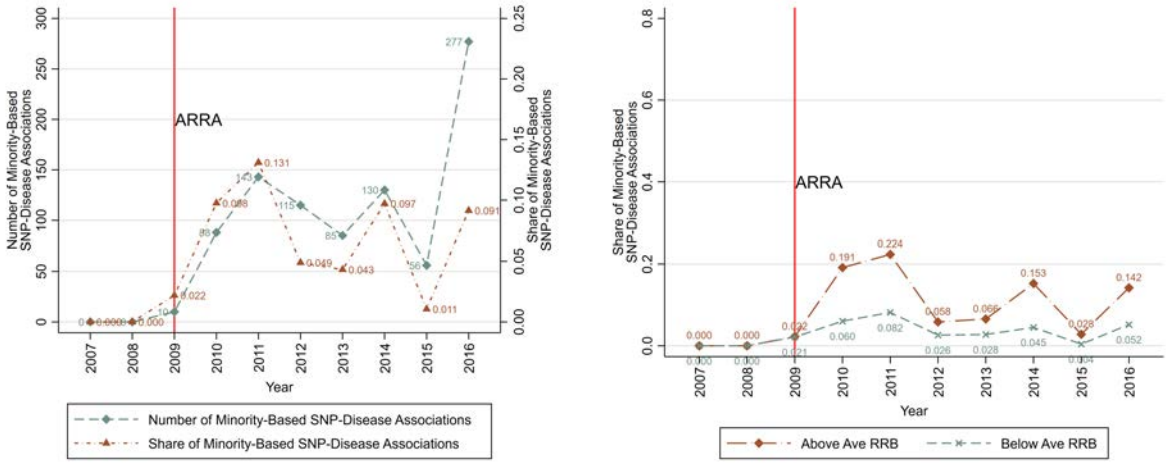


(A) GWAS Authors over Years

(B) Quasi-Treated vs Quasi-Control

FIGURE 3. DIVERSITY IN GWAS AUTHORS

Notes: The sample comprises of 27 broader diseases from 2007 to 2016, in total 183 disease-year pairs. Given minimal GWAS publications in 2005 and 2006, the data in 2007 presented in the figure includes authors from GWAS publications in 2005, 2006 and 2007. We predict the ancestral background associated with each surname based on the 2000 US census. We adopt a conservative rule in prediction of minority authors: a last name is predicted to be associated with a minority background only if its probability of being either "Asian", "Black" or "Hispanic" is more than 0.95. We separate the 27 broader disease categories into a "quasi-treated" and a "quasi-control" group based on their *RRB*. The "quasi-treated" includes 13 diseases with *RRB* larger than the mean of the 27 diseases, and the "quasi-control" includes the other 14 diseases.



(A) Minority-Based SNP-Disease Associations over Years

(B) Quasi-Treated vs Quasi-Control

FIGURE 4. MINORITY-BASED SNP-DISEASE ASSOCIATIONS

Notes: The sample comprises of 27 broader diseases from 2007 to 2016, in total 183 disease-year pairs. Given minimal GWAS publications in 2005 and 2006, the data in 2007 presented in the figure includes minority-based SNP-Disease associations from GWAS publications in 2005, 2006 and 2007. A SNP-disease association is referred to as a minority-based association as long as it is identified from a GWAS study in which all subjects in the discovery stage are minorities.

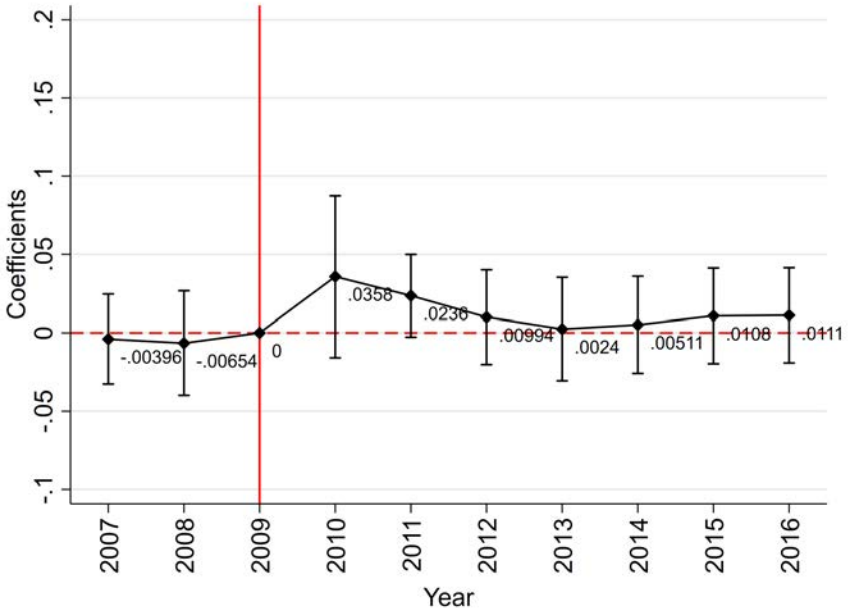


Figure 5. Impact on ARRA-NIH Grant Expansion on Grant Approval Likelihood

Notes: The sample comprises of 27 broader diseases from 2007 to 2016, in total 183 disease-year pairs. The sample is constructed based on studies supported by any US funding agencies between 2005 and 2016. The dependent variable is the share of studies supported by ARRA-NIH grants over all studies supported by the grants from any US funding agencies for each disease in each year. The relative referral base is the difference in the share of people with medical conditions between minorities and whites. The disease research field characteristics includes the total number of studies funded by only non-US grants up to the previous year, the number of minority-based SNPs discovered in studies funded by only non-US grants up to the previous year, and the total number of grants from the US funding agencies up to the previous year. Coefficients of the interactions of RRB and each year indicator are plotted. The 95% confidence intervals based on robust standard errors are plotted.

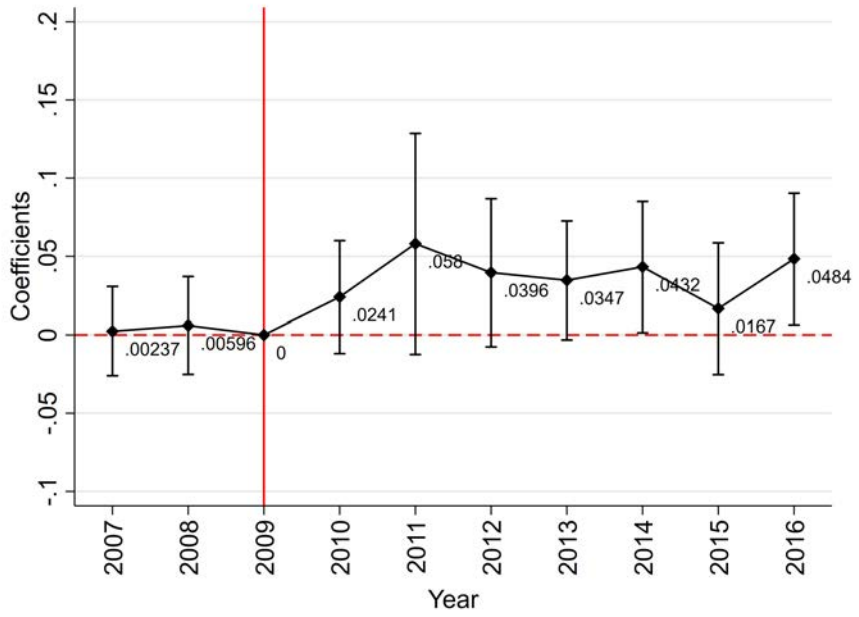


Figure 6. Impact of ARRA-NIH Grant Expansion on Ancestral Diversity in Subjects

Notes: The sample comprises of 27 broader diseases from 2007 to 2016, in total 183 disease-year pairs. The sample is constructed based on studies supported by any US funding agencies between 2005 and 2016. The dependent variable is the share of minority subjects for each disease in each year. The relative referral base (RRB) is the difference in the share of people with medical conditions between minorities and whites. The disease research field characteristics includes the total number of studies funded by only non-US grants up to the previous year, the number of minority-based SNPs discovered in studies funded by only non-US grants up to the previous year, and the total number of grants from the US funding agencies up to the previous year. Coefficients of the interactions of RRB and each year indicator are plotted. The 95% confidence intervals based on robust standard errors are plotted.

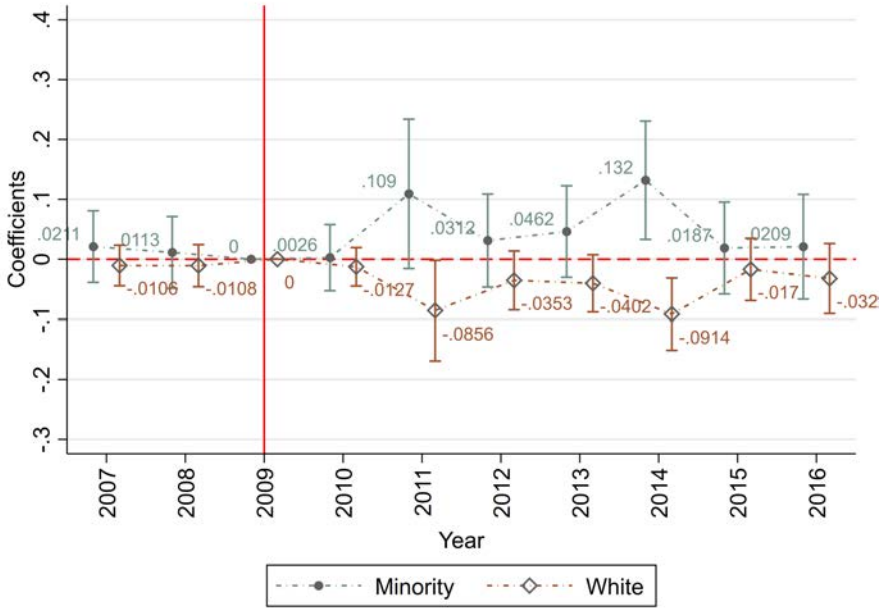


FIGURE 7. IMPACT OF ARRA-NIH GRANT EXPANSION ON ANCESTRAL DIVERSITY IN SUBJECTS – RRB DECOMPOSITION

Notes: The sample comprises of 27 broader diseases from 2007 to 2016, in total 183 disease-year pairs. The sample is constructed based on studies supported by any US funding agencies between 2005 and 2016. The dependent variable is the share of minority subjects for each disease in each year. The referral base (RB) for minorities (whites) is the share of people with medical conditions among minorities (whites). The disease research field characteristics includes the total number of studies funded by only non-US grants up to the previous year, the number of minority-based SNPs discovered in studies funded by only non-US grants up to the previous year, and the total number of grants from the US funding agencies up to the previous year. Coefficients of the interactions of RB and each year indicator are plotted. The 95% confidence intervals based on robust standard errors are plotted.

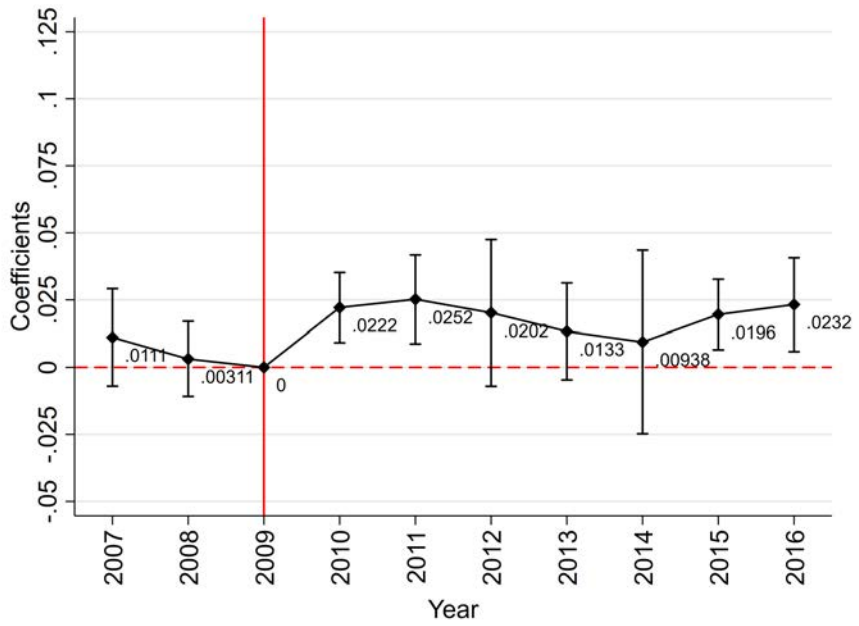


Figure 8. Impact of ARRA-NIH Grant Expansion on Ancestral Diversity in Authors

Notes: The sample comprises of 27 broader diseases from 2007 to 2016, in total 183 disease-year pairs. The sample is constructed based on studies supported by any US funding agencies between 2005 and 2016. The dependent variable is the share of minority authors for each disease in each year. We adopt a conservative rule in prediction of minority authors: a surname name is predicted to be associated with a minority background only if its probability of being either "Asian", "Black" or "Hispanic" is more than 0.95. The relative referral base (RRB) is the difference in the share of people with medical conditions between minorities and whites. The disease research field characteristics includes the total number of studies funded by only non-US grants up to the previous year, the number of minority-based SNPs discovered in studies funded by only non-US grants up to the previous year, and the total number of grants from the US funding agencies up to the previous year. Coefficients of the interactions of RRB and each year indicator are plotted. The 95% confidence intervals based on robust standard errors are plotted.

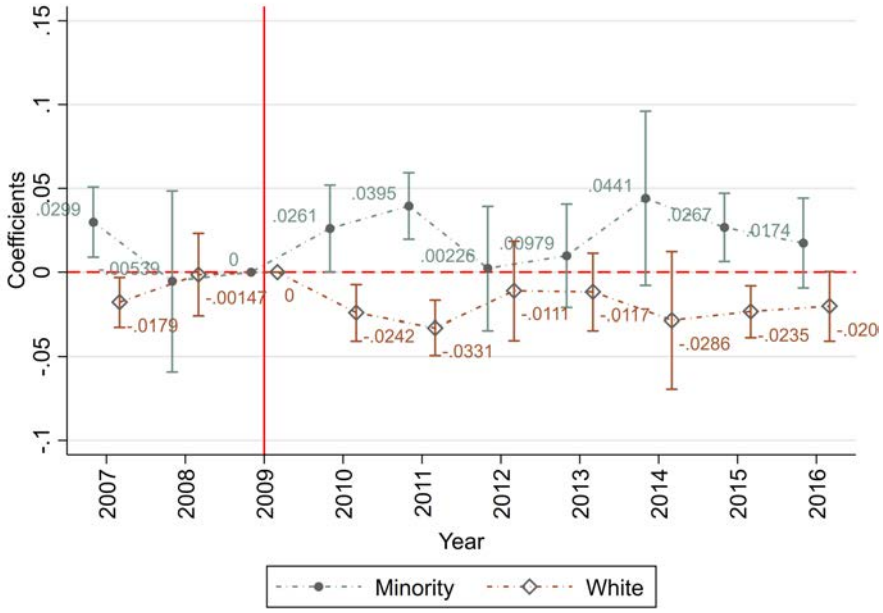
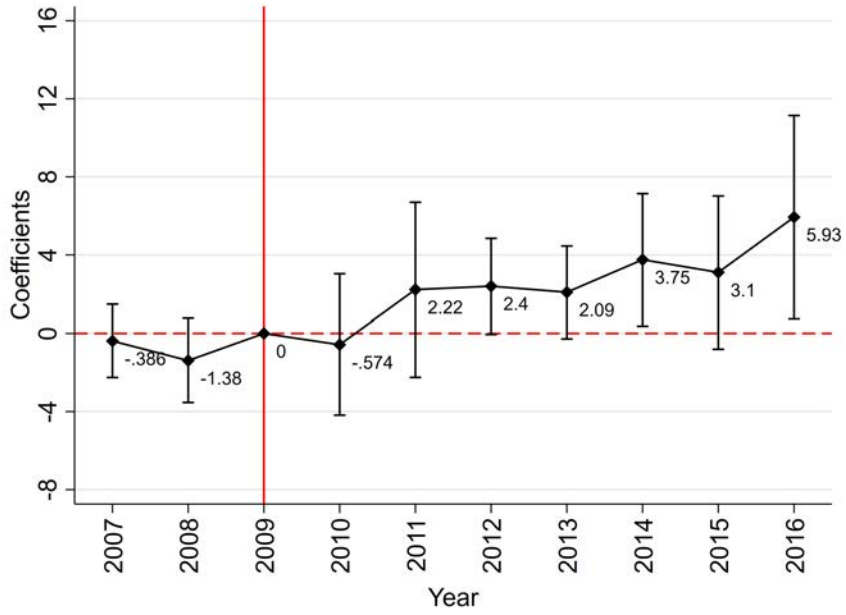
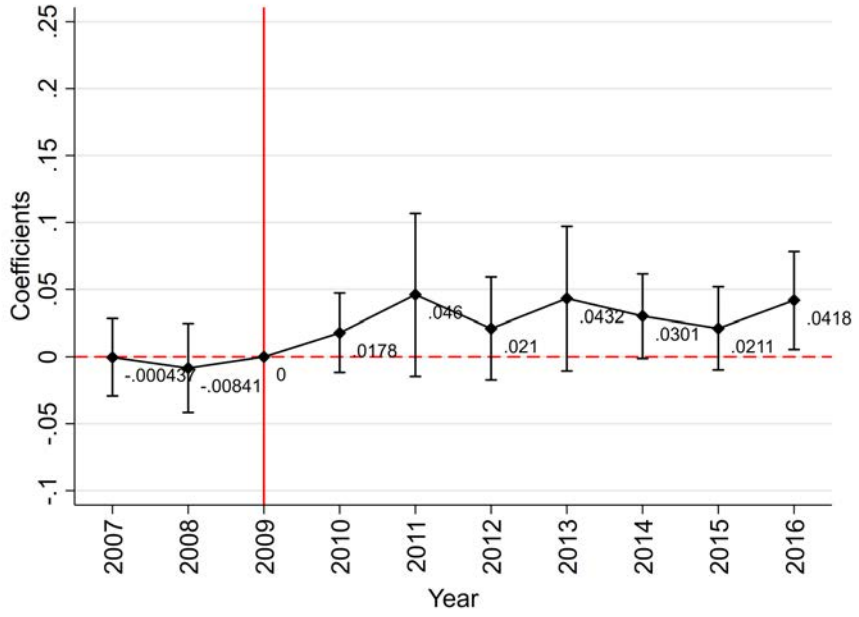


FIGURE 9. IMPACT OF ARRA-NIH GRANT EXPANSION ON ANCESTRAL DIVERSITY IN AUTHORS – RRB DECOMPOSITION

Notes: The sample comprises of 27 broader diseases from 2007 to 2016, in total 183 disease-year pairs. The sample is constructed based on studies supported by any US funding agencies between 2005 and 2016. The dependent variable is the share of minority authors for each disease in each year. We adopt a conservative rule in prediction of minority authors: a last name is predicted to be associated with a minority background only if its probability of being either "Asian", "Black" or "Hispanic" is more than 0.95. The referral base (RB) for minorities (whites) is the share of people with medical conditions among minorities (whites). The disease research field characteristics includes the total number of studies funded by only non-US grants up to the previous year, the number of minority-based SNPs discovered in studies funded by only non-US grants up to the previous year, and the total number of grants from the US funding agencies up to the previous year. Coefficients of the interactions of RB and each year indicator are plotted. The 95% confidence intervals based on robust standard errors are plotted.



(A) Number of Minority-Based SNP-Disease Associations



(B) Share of Minority-Based SNP-Disease Associations

Figure 10. Impact of ARRA-NIH Grant Expansion on Minority-Based SNP-Disease Association Discoveries

Notes: The sample comprises of 27 broader diseases from 2007 to 2016, in total 183 disease-year pairs. The sample is constructed based on studies supported by any US funding agencies between 2005 and 2016. The dependent variable is specified under each figure. A SNP-disease association is referred to as a minority-based association as long as it is identified from a GWAS study in which all subjects in the discovery stage are the minorities. The relative referral base (RRB) is the difference in the share of people with medical conditions between minorities and whites. The disease research field characteristics includes the total number of studies funded by only non-US grants up to the previous year, the number of minority-based SNPs discovered in studies funded by only non-US grants up to the previous year, and the total number of grants from the US funding agencies up to the previous year. Coefficients of the interactions of RRB and each year indicator are plotted. The 95% confidence intervals based on robust standard errors are plotted.

Table 1—Referral Base for GWAS Participants by Diseases (per 100 Individuals)

ID	Disease	European White	Minority
1	Allergic Reactions	2.52	1.29
2	Anemia and Other Deficiencies	0.81	0.82
	Appendicitis	na	na
3	Cancer	3.33	0.56
	Cerebrovascular Disease	0.85	na
	Complications of Pregnancy and Birth	na	na
	Congenital Anomalies	0.63	na
4	Copd, Asthma	18.28	12.80
5	Diabetes Mellitus	5.60	5.49
	Disorders of Mouth and Esophagus	0.46	na
6	Disorders of Teeth and Jaws	2.94	1.74
7	Disorders of the Upper GI	5.91	2.62
8	Female Genital Disorders, and Contraception	6.59	2.45
9	Gallbladder, Pancreatic, and Liver Disease	1.33	0.25
10	Headache	1.87	0.76
11	Heart Disease	7.84	2.95
	Hemorrhagic, Coagulation, and Disorders of White Blood Cells	na	na
	Hereditary, Degenerative, Other Nervous System Disorders	0.96	na
12	Hyperlipidemia	11.63	4.98
13	Hypertension	16.57	12.67
14	Infectious Diseases	6.74	3.52
15	Intestinal Infection	2.49	2.14
16	Kidney Disease	2.28	0.79
	Male Genital Disorders	1.73	na
17	Mental Disorders	14.32	7.53
18	Osteoarthritis and Other Non-Traumatic Joint Disorders	8.45	4.22
19	Other Bone and Musculoskeletal Disease	3.96	0.93
20	Other Circulatory Conditions Arteries, Veins, and Lymphatics	3.09	1.15
21	Other CNS Disorders	5.84	2.04
22	Other Eye Disorders	3.12	0.96
	Other Stomach and Intestinal Disorders	0.88	na
23	Otitis Media	3.97	2.29
24	Pneumonia	7.87	4.26
25	Skin Disorders	9.38	4.56
26	Systemic Lupus and Connective Tissues Disorders	5.76	2.64
27	Thyroid Disease	5.33	1.78

Notes: The table is created by the author based on the MEPS summary tables. Diseases with IDs are used in our final empirical analysis. Respondents in MEPS are classified into five ancestry groups: (1) Hispanic, (2) Black, (3) the American Indian, Alaska Native, and multi-races, (4) Asian, Hawaiian and Pacific Islander, and (5) European whites. The minority population in our following analysis contains all the mentioned ancestries except the European whites. The average referral bases for whites and minorities in 2005 are 6.22 and 3.27, respectively. The mean RRB is -2.95.

Table 2—Impact on ARRA-NIH Grant Expansion on Grant Approval Likelihood

Dependent Var	Share of Studies Supported by ARRA-NIH Grants over Those Supported by US Grants		Share of Studies Supported by ARRA-NIH Grants over Those Supported by All Grants		Share of Studies Supported by International Grants	
	(1)	(2)	(3)	(4)	(5)	(6)
Relative Referral Base (2005) * Post	0.0150**	0.0181**	0.0130**	0.0165**	0.0006	-0.0088
	(0.008)	(0.008)	(0.007)	(0.007)	(0.016)	(0.018)
	[0.022]	[0.005]	[0.032]	[0.006]	[0.983]	[0.747]
	[0.0019, 0.0310]	[0.0056, 0.0336]	[0.0010, 0.0299]	[0.0045, 0.0330]	[-0.0485, 0.0546]	[-0.0598, 0.0453]
Disease Research Field Characteristics	No	Yes	No	Yes	No	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Disease FE	Yes	Yes	Yes	Yes	Yes	Yes
Outcome Mean (Pre-Policy)	0.041	0.041	0.034	0.034	0.210	0.210
Observations	183	183	183	183	183	183

Notes: The sample in column 1 comprises of 27 broader diseases from 2007 to 2016, in total 183 disease-year pairs. The sample is constructed based on studies supported by any US funding agencies between 2005 and 2016. In column 2 and 3, we do not impose any restrictions on the source of funding. The dependent variable is specified in the column title. The relative referral base is the difference in the share of people with medical conditions between minorities and European whites. The disease research field characteristics includes the total number of studies funded by only non-US grants up to the previous year, the number of minority-based SNPs discovered in studies funded by only non-US grants up to the previous year, and the total number of grants from the US funding agencies up to the previous year. We report the Huber-White robust standard errors in the parentheses. We also report wild bootstrap cluster p-values in square brackets and wild bootstrap cluster 95% confidence intervals in square brackets, generated using boottest command (10000 replications) in Stata 14 (Roodman et al. 2019) for standard errors clustered at disease level with 27 disease clusters. Significance stars are labeled with respect to the robust standard errors. $*p < 0.1, **p < 0.05, ***p < 0.01$.

Table 3—Impact of ARRA-NIH Grant Expansion on Ancestral Diversity in Subjects

Dependent Var	Share of Minority Subjects in			
	All Subjects		Discovery Sample	Replication Sample
	(1)	(2)	(3)	(4)
Relative Referral Base (2005) * Post	0.0281***	0.0341***	0.0225**	0.0485**
	(0.011)	(0.013)	(0.010)	(0.024)
	[0.009]	[0.006]	[0.040]	[0.015]
	[0.0101, 0.0537]	[0.0110, 0.0580]	[0.0011, 0.0440]	[0.0125, 0.0779]
Disease Research Characteristics	No	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
Disease FE	Yes	Yes	Yes	Yes
Outcome Mean (Pre-Policy)	0.051	0.051	0.027	0.078
Observations	183	183	182	170

Notes: The sample comprises of 27 broader diseases from 2007 to 2016, in total 183 disease-year pairs. The sample is constructed based on studies supported by any US funding agencies between 2005 and 2016. The dependent variable is the share of minority subjects for each disease in each year. The relative referral base is the difference in the share of people with medical conditions between minorities and whites. The disease research field characteristics includes the total number of studies funded by only non-US grants up to the previous year, the number of minority-based SNPs discovered in studies funded by only non-US grants up to the previous year, and the total number of grants from the US funding agencies up to the previous year. We report the Huber-White robust standard errors in the parentheses. We also report wild bootstrap cluster p-values in square brackets and wild bootstrap cluster 95% confidence intervals in square brackets, generated using boottest command (10000 replications) in Stata 14 (Roodman et al. 2019) for standard errors clustered at disease level with 27 disease clusters. Significance stars are labeled with respect to the robust standard errors. $*p < 0.1, **p < 0.05, ***p < 0.01$.

Table 4—Impact of ARRA-NIH Grant Expansion on Ancestral Diversity in Subjects – Robustness Checks

Sample	Disease with Studies ≥ 0 (1)	Disease with Studies ≥ 10 (2)	Disease with Studies ≥ 20 (3)	Disease with Studies ≥ 30 (4)
Panel A – Key Independent Variable: Relative Referral Base (2005)				
Dependent Variable: Share of Minority Subject				
Relative Referral Base (2005) * Post	0.0341***	0.0345***	0.0409***	0.0434***
	(0.013)	(0.013)	(0.014)	(0.014)
	[0.006]	[0.005]	[0.005]	[0.005]
	[0.0110, 0.0580]	[0.0106, 0.0591]	[0.0189, 0.0628]	[0.0207, 0.0655]
Outcome Mean (Pre-Policy)/Observations	0.051/183	0.051/174	0.051/157	0.051/134
Dependent Variable: Share of Minority Subject in Discovery Sample				
Relative Referral Base (2005) * Post	0.0225**	0.0232**	0.0305**	0.0325***
	(0.010)	(0.010)	(0.012)	(0.012)
	[0.040]	[0.039]	[0.003]	[0.003]
	[0.0011, 0.0440]	[0.0012, 0.0453]	[0.0135, 0.0515]	[0.0153, 0.0541]
Outcome Mean (Pre-Policy)/Observations	0.027/182	0.027/173	0.027/156	0.027/133
Dependent Variable: Share of Minority Subject in Replication Sample				
Relative Referral Base (2005) * Post	0.0485**	0.0490**	0.0465*	0.0493**
	(0.024)	(0.024)	(0.024)	(0.024)
	[0.015]	[0.021]	[0.069]	[0.064]
	[0.0125, 0.0779]	[0.0107, 0.0792]	[-0.0059, 0.0811]	[-0.0041, 0.0853]
Outcome Mean (Pre-Policy)/Observations	0.078/170	0.078/163	0.078/149	0.078/127
Panel B – Key Independent Variable: Average Relative Referral Base (2002-2005)				
Dependent Variable: Share of Minority Subject				
Ave. Relative Referral Base (2002 -2005) * Post	0.0335***	0.0339***	0.0410***	0.0438***
	(0.013)	(0.013)	(0.014)	(0.014)
	[0.008]	[0.006]	[0.006]	[0.004]
	[0.0101, 0.0590]	[0.0096, 0.0601]	[0.0199, 0.0626]	[0.0222, 0.0658]
Outcome Mean (Pre-Policy)/Observations	0.051/183	0.051/174	0.051/157	0.051/134
Dependent Variable: Share of Minority Subject in Discovery Sample				
Ave. Relative Referral Base (2002 -2005) * Post	0.0217**	0.0223**	0.0303**	0.0328***
	(0.010)	(0.010)	(0.012)	(0.012)
	[0.048]	[0.051]	[0.002]	[0.002]
	[0.0002, 0.0437]	[-0.0001, 0.0450]	[0.0138, 0.0493]	[0.0161, 0.0525]
Outcome Mean (Pre-Policy)/Observations	0.027/182	0.027/173	0.027/156	0.027/133
Dependent Variable: Share of Minority Subject in Replication Sample				
Ave. Relative Referral Base (2002 -2005) * Post	0.0496**	0.0500**	0.0478*	0.0510**
	(0.025)	(0.025)	(0.025)	(0.025)
	[0.016]	[0.021]	[0.068]	[0.063]
	[0.0136, 0.0821]	[0.0112, 0.0830]	[-0.0065, 0.0862]	[-0.0056, 0.0910]
Outcome Mean (Pre-Policy)/Observations	0.078/170	0.078/163	0.078/149	0.078/127
Disease Research Field Characteristics	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
Disease FE	Yes	Yes	Yes	Yes

Notes: The sample comprises of 27 broader diseases from 2007 to 2016, in total 183 disease-year pairs. The sample is constructed based on studies supported by any US funding agencies between 2005 and 2016. The relative referral base is the difference in the share of people with medical conditions between minorities and whites. The disease research field characteristics includes the total number of studies funded by only non-US grants up to the previous year, the number of minority-based SNPs discovered in studies funded by only non-US grants up to the previous year, and the total number of grants from the US funding agencies up to the previous year. We report the Huber-White robust standard errors in the parentheses. We also report wild bootstrap cluster p-values in square brackets and wild bootstrap cluster 95% confidence intervals in square brackets, generated using boottest command (10000 replications) in Stata 14 (Roodman et al. 2019) for standard errors clustered at disease level with 27 disease clusters. Significance stars are labeled with respect to the robust standard errors. $*p < 0.1, **p < 0.05, ***p < 0.01$.

Table 5—Impact of ARRA-NIH Grant Expansion on Ancestral Diversity in Authors

Dependent Var	Share of Minority Authors				
	Minority Prediction ≥ 0.95		Minority Prediction ≥ 0.90	Minority Prediction ≥ 0.85	Minority Prediction ≥ 0.80
	(1)	(2)	(3)	(4)	(5)
Relative Referral Base (2005) * Post	0.0101**	0.0145**	0.0168**	0.0162**	0.0177**
	(0.005)	(0.006)	(0.007)	(0.007)	(0.007)
	[0.017]	[0.010]	[0.010]	[0.014]	[0.009]
	[0.0034, 0.0175]	[0.0052, 0.0230]	[0.0060, 0.0257]	[0.0048, 0.0255]	[0.0062, 0.0270]
Disease Research Field Characteristics	No	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
Disease FE	Yes	Yes	Yes	Yes	Yes
Outcome Mean (Pre-Policy)	0.075	0.075	0.094	0.105	0.114
Observations	183	183	183	183	183

Notes: The sample comprises of 27 broader diseases from 2007 to 2016, in total 183 disease-year pairs. The sample is constructed based on studies supported by any US funding agencies between 2005 and 2016. The dependent variable is the share of minority authors for each disease in each year. We adopt a conservative rule in prediction of minority authors in columns 1 and 2: a last name is predicted to be associated with a minority background only if its probability of being either "Asian", "Black" or "Hispanic" is more than 0.95. In columns 3 to 5, we use alternative prediction probability thresholds. The relative referral base is the difference in the share of people with medical conditions between minorities and whites. The disease research field characteristics includes the total number of studies funded by only non-US grants up to the previous year, the number of minority-based SNPs discovered in studies funded by only non-US grants up to the previous year, and the total number of grants from the US funding agencies up to the previous year. We report the Huber-White robust standard errors in the parentheses. We also report wild bootstrap cluster p-values in square brackets and wild bootstrap cluster 95% confidence intervals in square brackets, generated using boottest command (10000 replications) in Stata 14 (Roodman et al. 2019) for standard errors clustered at disease level with 27 disease clusters. Significance stars are labeled with respect to the robust standard errors. $*p < 0.1, **p < 0.05, ***p < 0.01$.

Table 6—Impact of ARRA-NIH Grant Expansion on Collaboration Patterns

Dependent Var	Share of Studies with Minority Authors in					
	First 1 Author	First 2 Authors	First 3 Authors	Last 1 Author	Last 2 Authors	Last 3 Authors
	(1)	(2)	(3)	(4)	(5)	(6)
Relative Referral Base (2005) * Post	0.0332*	0.0408*	0.0472**	0.0041	0.0131	0.0300
	(0.018)	(0.021)	(0.023)	(0.021)	(0.022)	(0.028)
	[0.078]	[0.075]	[0.067]	[0.689]	[0.384]	[0.199]
	[-0.0050, 0.0633]	[-0.0045, 0.0771]	[-0.0042, 0.0923]	[-0.0139, 0.0267]	[-0.0145, 0.0446]	[-0.0170, 0.0757]
Disease Research Field Characteristics	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Disease FE	Yes	Yes	Yes	Yes	Yes	Yes
Outcome Mean (Pre-Policy)	0.077	0.130	0.169	0.058	0.080	0.195
Observations	183	183	183	183	183	183

Notes: The sample comprises of 27 broader diseases from 2007 to 2016, in total 183 disease-year pairs. The sample is constructed based on studies supported by any US funding agencies between 2005 and 2016. The dependent variable is the share of minority authors at different positions in the author rank, as specified in each column title. We adopt a conservative rule in prediction of minority authors: a last name is predicted to be associated with a minority background only if its probability of being either "Asian", "Black" or "Hispanic" is more than 0.95. The relative referral base is the difference in the share of people with medical conditions between minorities and whites. The disease research field characteristics includes the total number of studies funded by only non-US grants up to the previous year, the number of minority-based SNPs discovered in studies funded by only non-US grants up to the previous year, and the total number of grants from the US funding agencies up to the previous year. We report the Huber-White robust standard errors in the parentheses. We also report wild bootstrap cluster p-values in square brackets and wild bootstrap cluster 95% confidence intervals in square brackets, generated using boottest command (10000 replications) in Stata 14 (Roodman et al. 2019) for standard errors clustered at disease level with 27 disease clusters. Significance stars are labeled with respect to the robust standard errors. $*p < 0.1, **p < 0.05, ***p < 0.01$.

Table 7—Impact of ARRA-NIH Grant Expansion on Minority-Based SNP-Disease Association Discoveries

Dependent Var	Number of Minority-Based SNP-Associations			Share of Minority-Based SNP-Associations		
	100% Minority in Discovery (1)	60% Minority in Discovery (2)	100% Minority in All (3)	100% Minority in Discovery (4)	60% Minority in Discovery (5)	100% Minority in All (6)
Relative Referral Base (2005) * Post	2.6296**	2.4891*	1.7868	0.0329***	0.0331**	0.0162*
	(1.185)	(1.279)	(1.176)	(0.013)	(0.013)	(0.009)
	[0.128]	[0.171]	[0.120]	[0.085]	[0.114]	[0.067]
	[-0.7523, 5.645]	[-1.0585, 5.761]	[-0.4755, 4.177]	[-0.0041, 0.0702]	[-0.0072, 0.0728]	[-0.0010, 0.0328]
Disease Research Field Characteristics	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Disease FE	Yes	Yes	Yes	Yes	Yes	Yes
Outcome Mean (Pre-Policy)	0.250	0.250	0.200	0.010	0.010	0.009
Observations	183	183	183	183	183	183

Notes: The sample comprises of 27 broader diseases from 2007 to 2016, in total 183 disease-year pairs. The sample is constructed based on studies supported by any US funding agencies between 2005 and 2016. The dependent variable is specified in the column title. A SNP-disease association is referred to as a minority-based association as long as it is identified from a GWAS study which should satisfy a certain criterion as specified in the column title. The relative referral base (RRB) is the difference in the share of people with medical conditions between minorities and whites. The disease research field characteristics includes the total number of studies funded by only non-US grants up to the previous year, the number of minority-based SNPs discovered in studies funded by only non-US grants up to the previous year, and the total number of grants from the US funding agencies up to the previous year. Coefficients of the interactions of RRB and each year indicator are plotted. The 95% confidence intervals based on robust standard errors are plotted.

For Online Publication, Appendix A: Additional Figures and Tables

A1. Sample

Table A1.1—Referral Base for GWAS Participants by Diseases and Ancestries (per 100 Individuals)

Disease	White	Hispanics	Black	Asian	Native American
Allergic Reactions	2.52	1.22	2.10	na	na
Anemia and Other Deficiencies	0.81	1.00	1.08	na	na
Appendicitis	na	na	na	na	na
Cancer	3.33	0.49	0.96	na	na
Cerebrovascular Disease	0.85	na	na	na	na
Complications of Pregnancy and Birth	na	na	na	na	na
Congenital Anomalies	0.63	na	na	na	na
Copd, Asthma	18.28	11.49	13.89	10.66	19.98
Diabetes Mellitus	5.60	4.99	7.31	4.98	na
Disorders of Mouth and Esophagus	0.46	na	na	na	na
Disorders of Teeth and Jaws	2.94	1.78	2.67	na	na
Disorders of the Upper GI	5.91	2.98	3.64	na	na
Female Genital Disorders, and Contraception	6.59	2.80	3.40	na	na
Gallbladder, Pancreatic, and Liver Disease	1.33	0.56	na	na	na
Headache	1.87	1.01	0.88	na	na
Heart Disease	7.84	2.86	4.73	na	na
Hemorrhagic, Coagulation, and Disorders of White Blood Cells	na	na	na	na	na
Hereditary, Degenerative and Other Nervous System Disorders	0.96	na	na	na	na
Hyperlipidemia	11.63	4.21	6.00	7.15	na
Hypertension	16.57	8.57	17.96	11.03	14.35
Infectious Diseases	6.74	3.71	3.89	na	7.57
Intestinal Infection	2.49	2.89	2.43	na	na
Kidney Disease	2.28	1.00	0.98	na	na
Male Genital Disorders	1.73	na	na	na	na
Mental Disorders	14.32	7.28	8.18	3.15	14.77
Osteoarthritis and Other Non-Traumatic Joint Disorders	8.45	3.56	5.76	na	8.89
Other Bone and Musculoskeletal Disease	3.96	1.19	1.14	na	na
Other Circulatory Conditions and Arteries, Veins, and Lymphatics	3.09	1.15	1.81	na	na
Other CNS Disorders	5.84	2.11	3.10	na	na
Other Eye Disorders	3.12	0.99	1.47	na	na
Other Stomach and Intestinal Disorders	0.88	na	na	na	na
Otitis Media	3.97	3.43	2.18	na	na
Pneumonia	7.87	4.35	3.84	3.87	6.66
Skin Disorders	9.38	4.41	5.10	5.82	na
Systemic Lupus and Connective Tissues Disorders	5.76	2.81	3.93	na	na
Thyroid Disease	5.33	1.90	2.63	na	na

Notes: The table is created by the author based on the MEPS summary tables. Respondents in MEPS are classified into five ancestry groups: (1) Hispanic, (2) Black, (3) American Indian, Alaska Native, and multi-races, (4) Asian, Hawaiian and Pacific Islander, and (5) European whites. The minority population in our following analysis contains all the mentioned ancestries except the European whites.

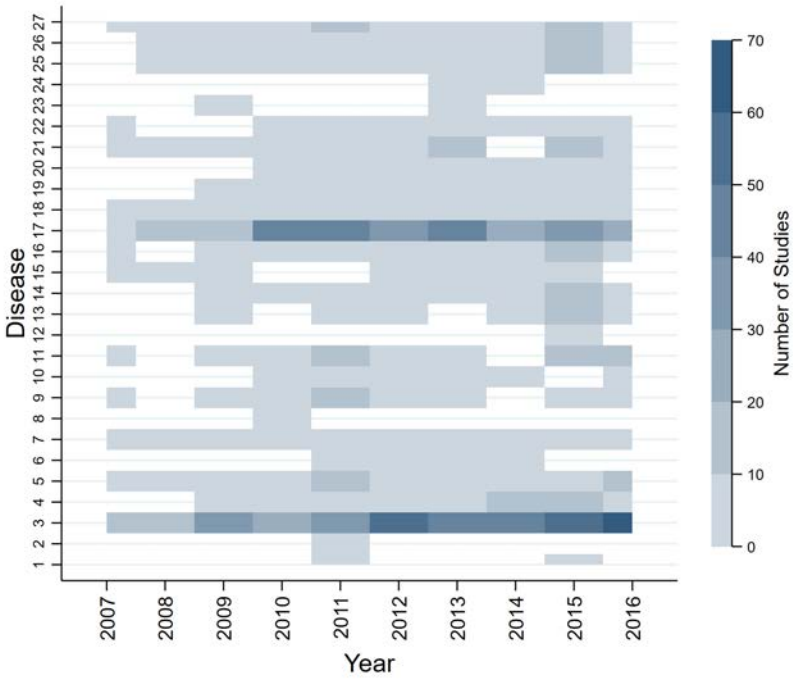
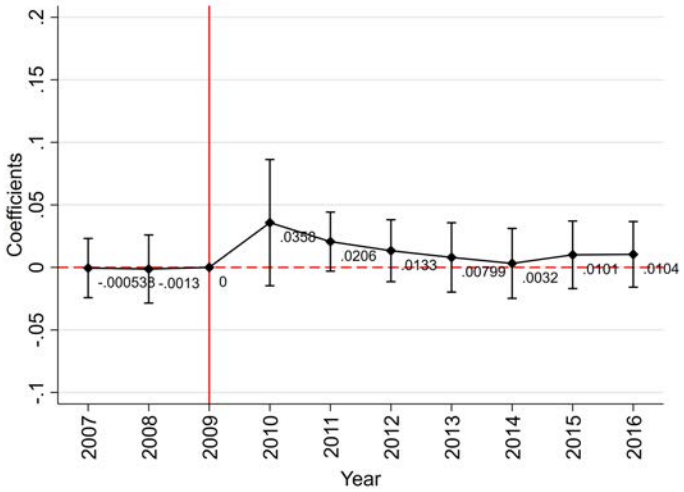


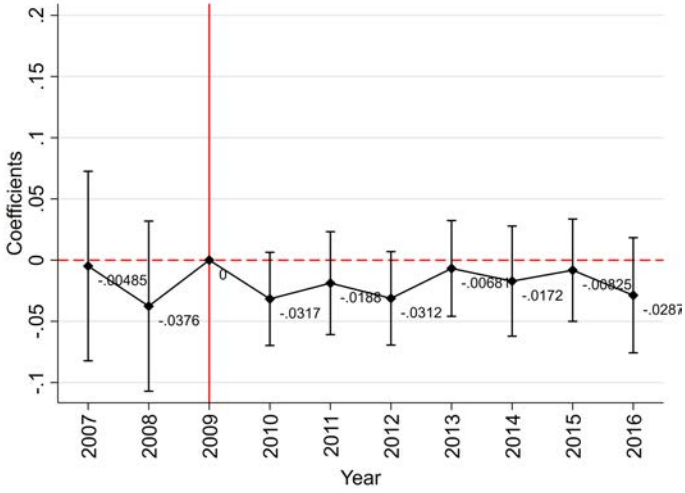
Figure A1.1. Heat Map for Disease Studies

Notes: The map sample comprises of 27 broader diseases from 2007 to 2016, in total 183 disease-year pairs. Given minimal GWAS publications in 2005 and 2006, the data in 2007 presented in the figure includes subjects from GWAS publications in 2005, 2006 and 2007. The IDs corresponds to the IDs in Table 1.

A2. ARRA-NIH Funding Approval Likelihood



(A) Share of Studies Supported by ARRA-NIH Grants over Studies Supported by Grants from Any Sources



(B) Share of Studies Supported by International Grants over Studies Supported by Grants from Any Sources

FIGURE A2.1. IMPACT ON GRANT SUPPORT

Notes: The sample comprises of 27 broader diseases from 2007 to 2016, in total 183 disease-year pairs. The sample is constructed based on studies supported by any U.S funding agencies between 2005 and 2016. The relative referral base is the difference in the share of people with medical conditions between minorities and whites. The disease research field characteristics includes the total number of studies funded by only non-US grants up to the previous year, the number of minority-based SNPs discovered in studies funded by only non-US grants up to the previous year, and the total number of grants from the US funding agencies up to the previous year. Coefficients of the interactions of *RRB* and each year indicator are plotted. The 95% confidence intervals based on robust standard errors are plotted.

A3. Ancestral Diversity in GWAS Subjects

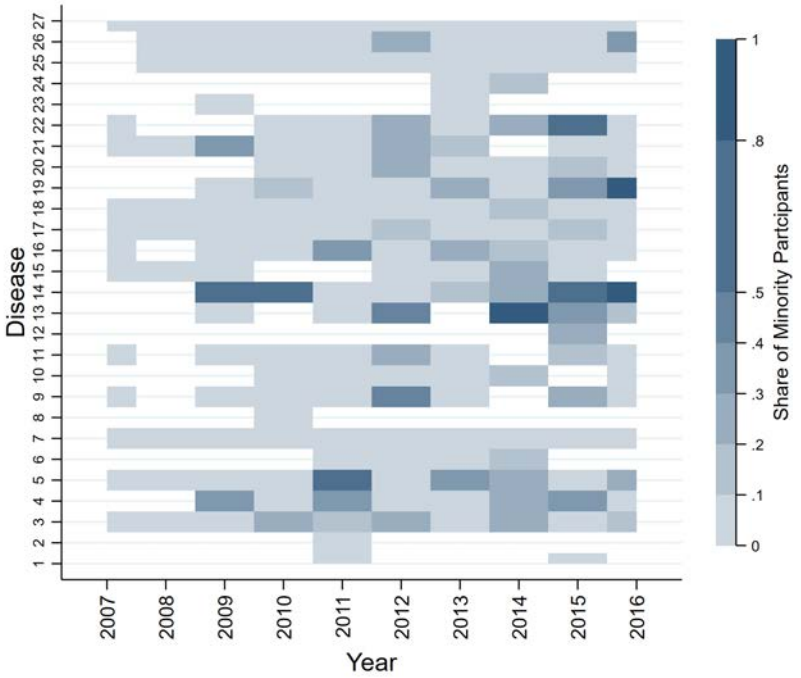


FIGURE A3.1. HEAT MAP FOR ANCESTRAL DIVERSITY IN GWAS SUBJECTS

Notes: The sample comprises of 27 broader diseases from 2007 to 2016, in total 183 disease-year pairs. Given minimal GWAS publications in 2005 and 2006, the data in 2007 presented in the figure includes subjects from GWAS publications in 2005, 2006 and 2007. The IDs corresponds to the IDs in Table 1.

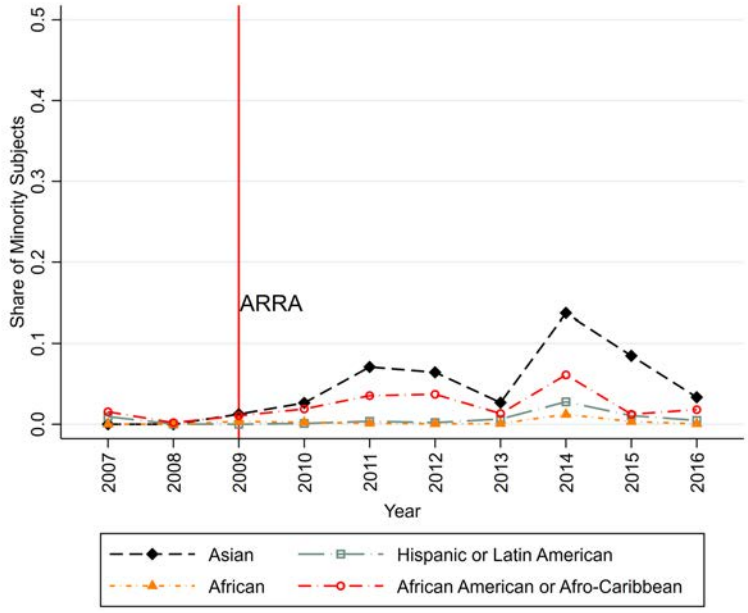


Figure A: Share of Minority Subjects in Discovery Sample by Races over Years

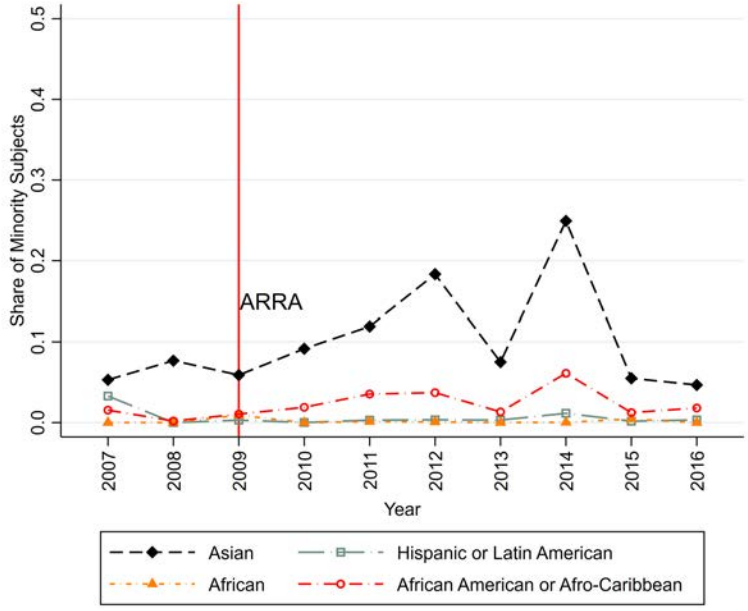
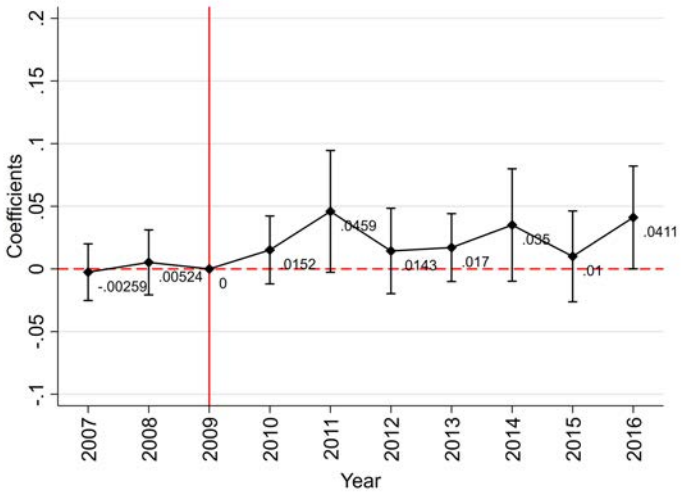


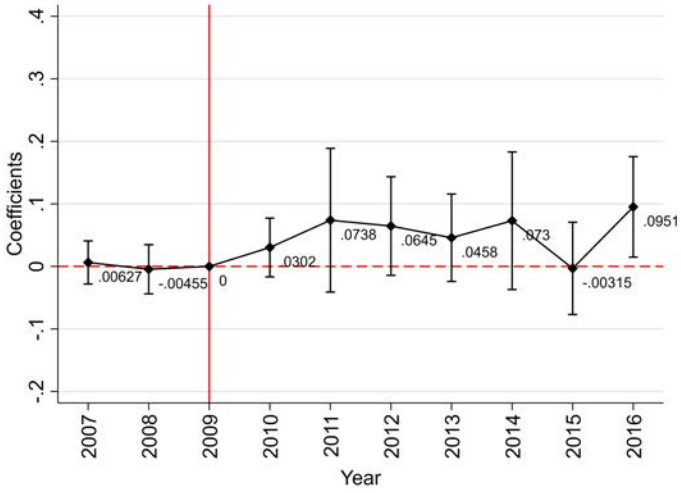
Figure B: Share of Minority Subjects in Replication Sample by Races over Years

FIGURE A3.2. SHARE OF MINORITY SUBJECTS BY RACES OVER YEARS

Notes: The sample comprises of 27 broader diseases from 2007 to 2016, in total 182 disease-year pairs for discovery stage analysis and 170 disease-year pairs for replication stage analysis. Given minimal GWAS publications in 2005 and 2006, the data in 2007 presented in the figure includes subjects from GWAS publications in 2005, 2006 and 2007.



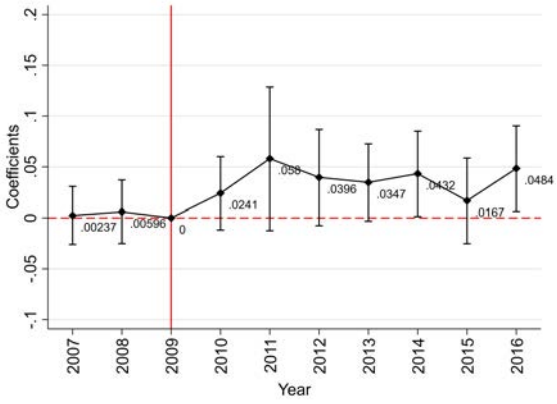
(A) Share of Minority Subjects in the Discovery Stage



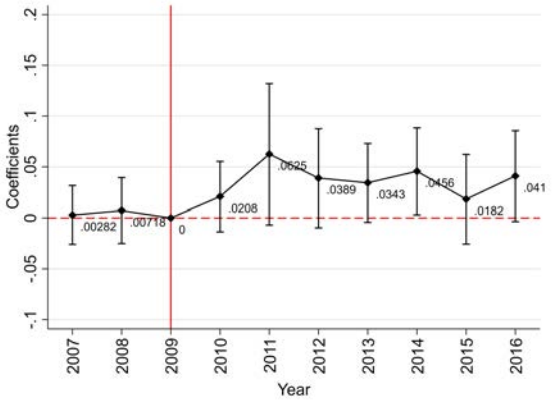
(B) Share of Minority Subjects in the Replication Stage

FIGURE A3.3. IMPACT OF ARRA-NIH GRANT EXPANSION ON ANCESTRAL DIVERSITY IN SUBJECTS – BY STAGE

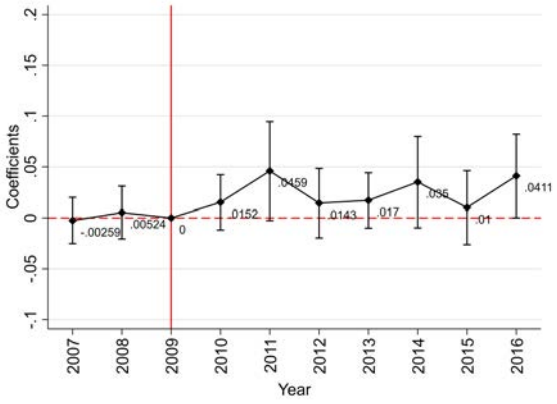
Notes: The sample comprises of 27 broader diseases from 2007 to 2016, in total 182 disease-year pairs for discovery stage analysis and 170 disease-year pairs for replication stage analysis. The sample is constructed based on studies supported by any US funding agencies between 2005 and 2016. The dependent variable is the share of minority subjects for each disease in each year. The relative referral base (RRB) is the difference in the share of people with medical conditions between minorities and whites. The disease research field characteristics includes the total number of studies funded by only non-US grants up to the previous year, the number of minority-based SNPs discovered in studies funded by only non-US grants up to the previous year, and the total number of grants from the US funding agencies up to the previous year. Coefficients of the interactions of RRB and each year indicator are plotted. The 95% confidence intervals based on robust standard errors are plotted.



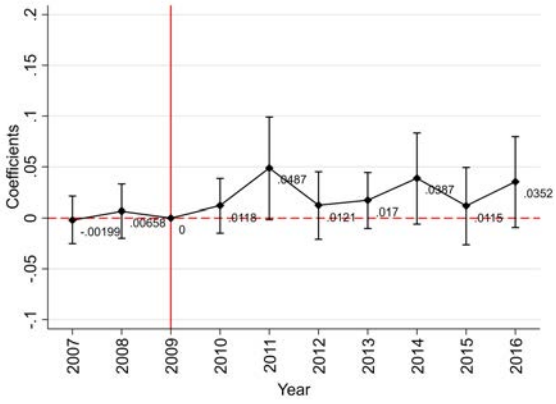
(A) RRB (2005), Discovery + Replication Stages



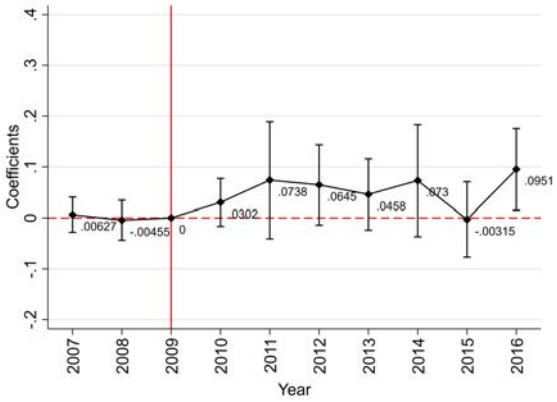
(B) Ave RRB (2002-2005), Discovery + Replication Stages



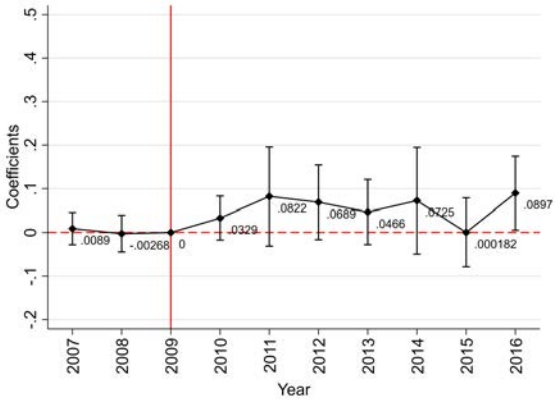
(C) RRB (2005), Discovery Stage



(D) Ave RRB (2002-2005), Discovery Stage



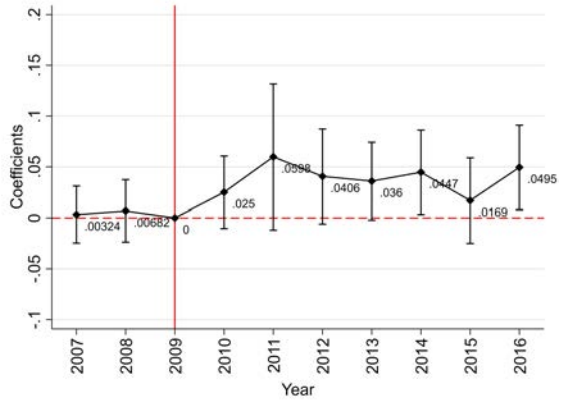
(E) RRB (2005), Replication Stage



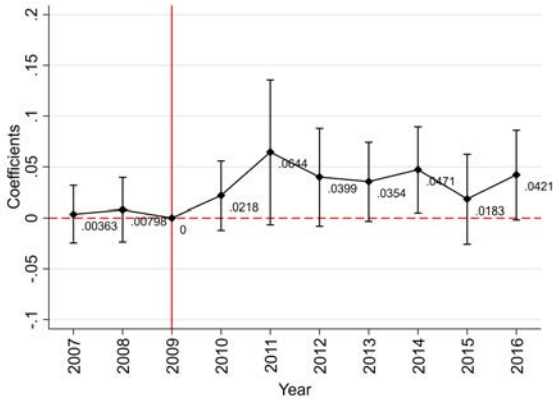
(F) Ave RRB (2002-2005), Replication Stage

FIGURE A3.4. IMPACT OF ARRA-NIH GRANT EXPANSION ON ANCESTRAL DIVERSITY IN SUBJECTS: DISEASES WITH A NUMBER OF STUDIES ≥ 0

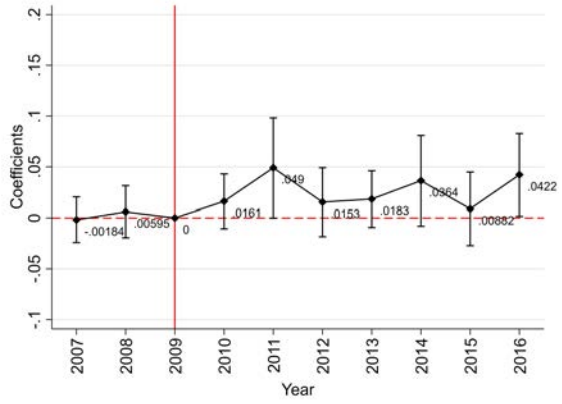
Notes: The sample comprises of 27 broader diseases from 2007 to 2016. There are in total 183 disease-year pairs for analysis combining discovery and replication stages, 182 disease-year pairs for discovery stage analysis, and 170 disease-year pairs for replication stage analysis. The sample is constructed based on studies supported by any US funding agencies between 2005 and 2016. The dependent variable is the share of minority subjects for each disease in each year. The relative referral base (RRB) is the difference in the share of people with medical conditions between minorities and whites. The disease research field characteristics includes the total number of studies funded by only non-US grants up to the previous year, the number of minority-based SNPs discovered in studies funded by only non-US grants up to the previous year, and the total number of grants from the US funding agencies up to the previous year. Coefficients of the interactions of RRB and each year indicator are plotted. The 95% confidence intervals based on robust standard errors are plotted.



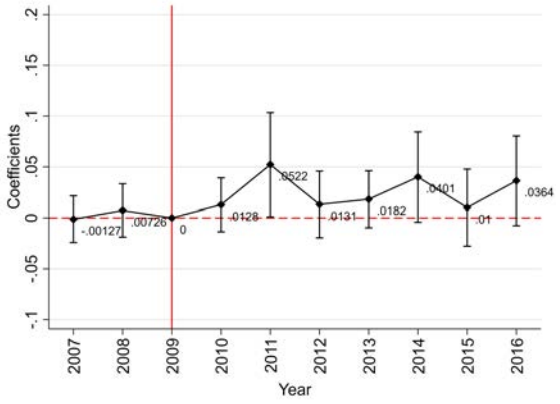
(A) RRB (2005), Discovery + Replication Stages



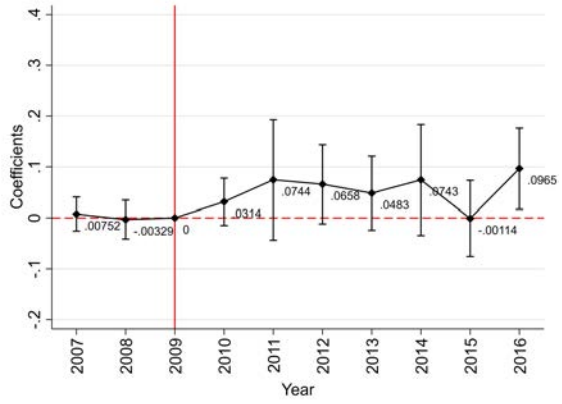
(B) Ave RRB (2002-2005), Discovery + Replication Stages



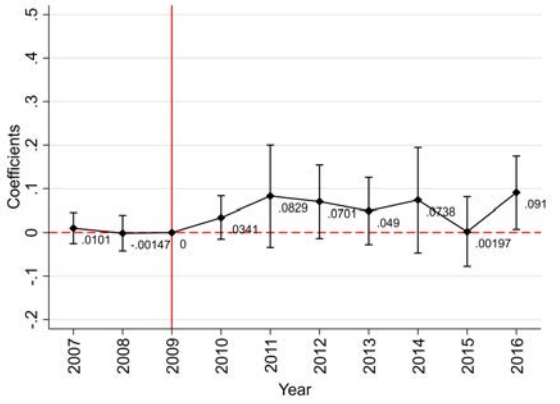
(C) RRB (2005), Discovery Stage



(D) Ave RRB (2002-2005), Discovery Stage



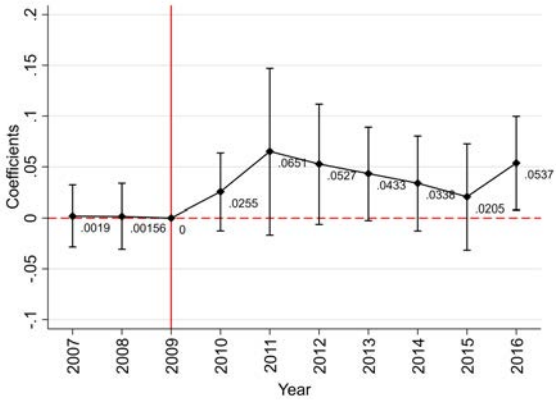
(E) RRB (2005), Replication Stage



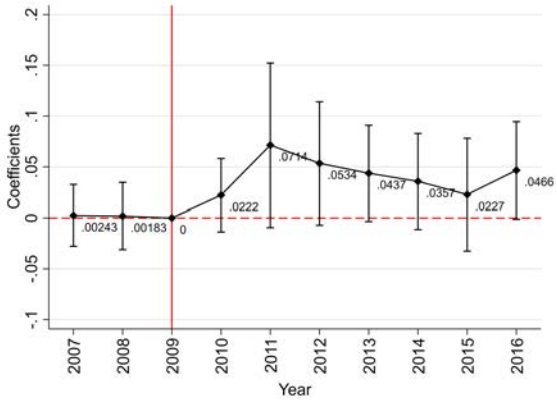
(F) Ave RRB (2002-2005), Replication Stage

FIGURE A3.5. IMPACT OF ARRA-NIH GRANT EXPANSION ON ANCESTRAL DIVERSITY IN SUBJECTS: DISEASES WITH A NUMBER OF STUDIES ≥ 10

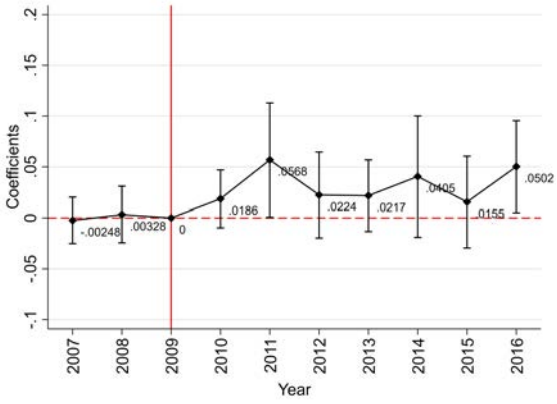
Notes: The sample comprises of 27 broader diseases from 2007 to 2016. There are in total 174 disease-year pairs for analysis combining discovery and replication stages, 173 disease-year pairs for discovery stage analysis, and 163 disease-year pairs for replication stage analysis. The sample is constructed based on studies supported by any U.S funding agencies between 2005 and 2016. The dependent variable is the share of minority subjects for each disease in each year. The relative referral base (RRB) is the difference in the share of people with medical conditions between minorities and whites. The disease research field characteristics includes the total number of studies funded by only non-US grants up to the previous year, the number of minority-based SNPs discovered in studies funded by only non-US grants up to the previous year, and the total number of grants from the US funding agencies up to the previous year. Coefficients of the interactions of RRB and each year indicator are plotted. The 95% confidence intervals based on robust standard errors are plotted.



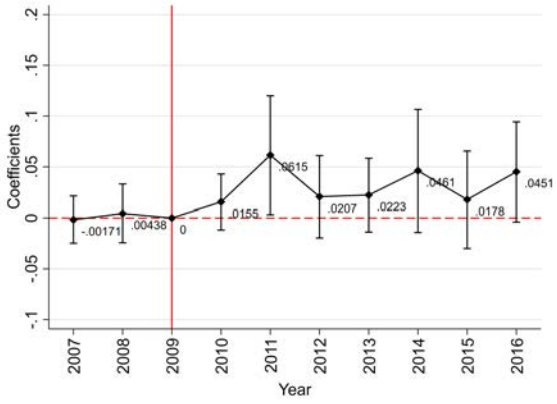
(A) RRB (2005), Discovery + Replication Stages



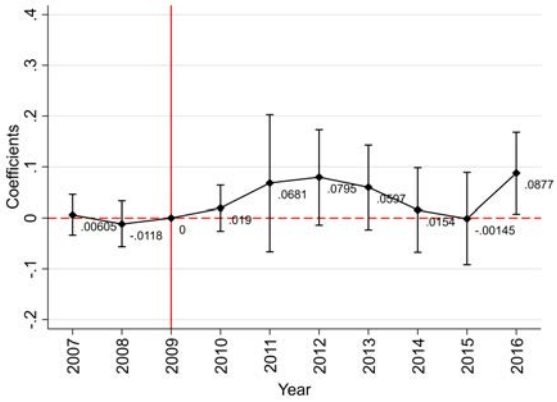
(B) Ave RRB (2002-2005), Discovery + Replication Stages



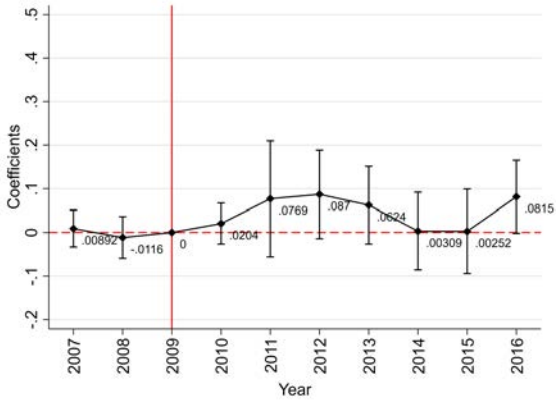
(C) RRB (2005), Discovery Stage



(D) Ave RRB (2002-2005), Discovery Stage



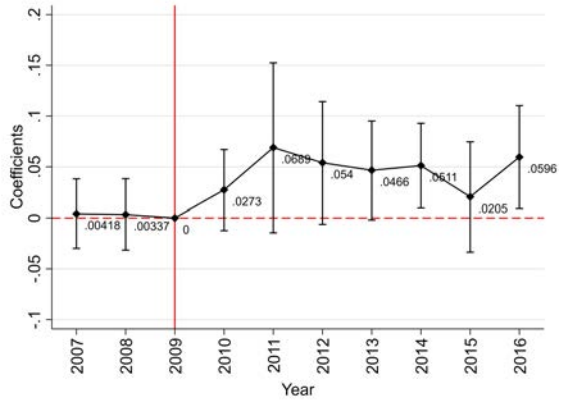
(E) RRB (2005), Replication Stage



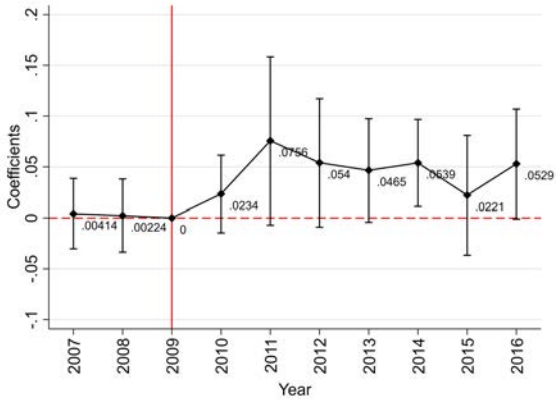
(F) Ave RRB (2002-2005), Replication Stage

FIGURE A3.6. IMPACT OF ARRA-NIH GRANT EXPANSION ON ANCESTRAL DIVERSITY IN SUBJECTS: DISEASES WITH A NUMBER OF STUDIES ≥ 20

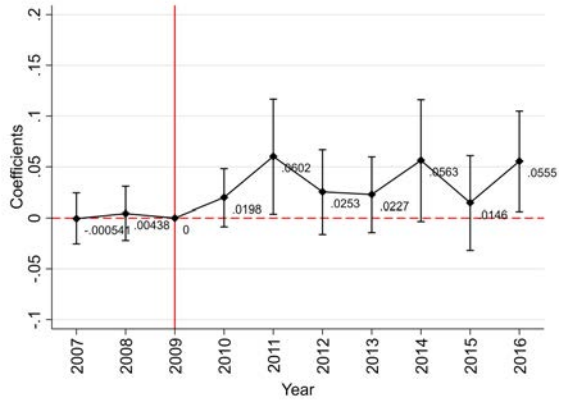
Notes: The sample comprises of 27 broader diseases from 2007 to 2016. There are in total 157 disease-year pairs for analysis combining discovery and replication stages, 156 disease-year pairs for discovery stage analysis, and 149 disease-year pairs for replication stage analysis. The sample is constructed based on studies supported by any US funding agencies between 2005 and 2016. The dependent variable is the share of minority subjects for each disease in each year. The relative referral base (RRB) is the difference in the share of people with medical conditions between minorities and whites. The disease research field characteristics includes the total number of studies funded by only non-US grants up to the previous year, the number of minority-based SNPs discovered in studies funded by only non-US grants up to the previous year, and the total number of grants from the US funding agencies up to the previous year. Coefficients of the interactions of RRB and each year indicator are plotted. The 95% confidence intervals based on robust standard errors are plotted.



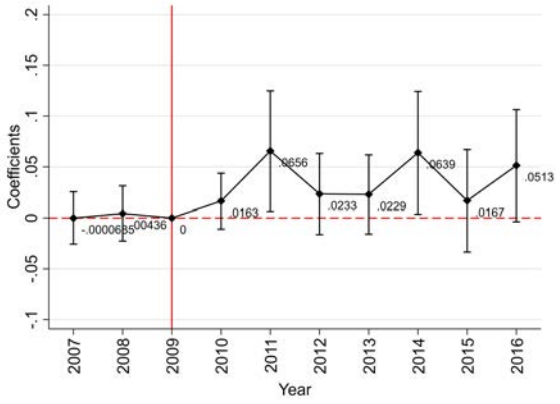
(A) RRB (2005), Discovery + Replication Stages



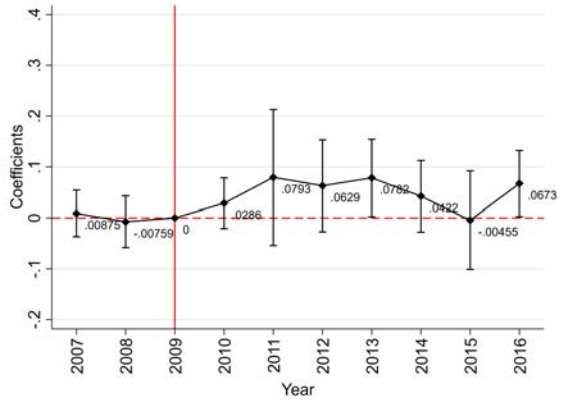
(B) Ave RRB (2002-2005), Discovery + Replication Stages



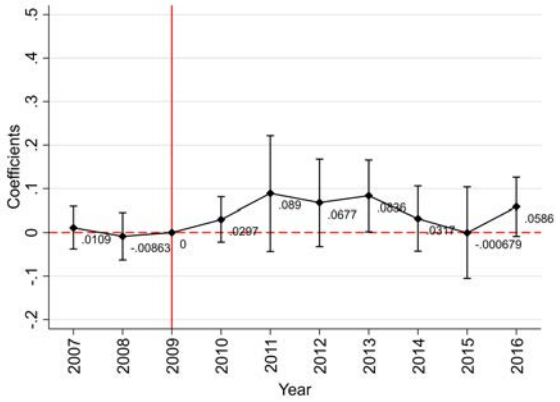
(C) RRB (2005), Discovery Stage



(D) Ave RRB (2002-2005), Discovery Stage



(E) RRB (2005), Replication Stage

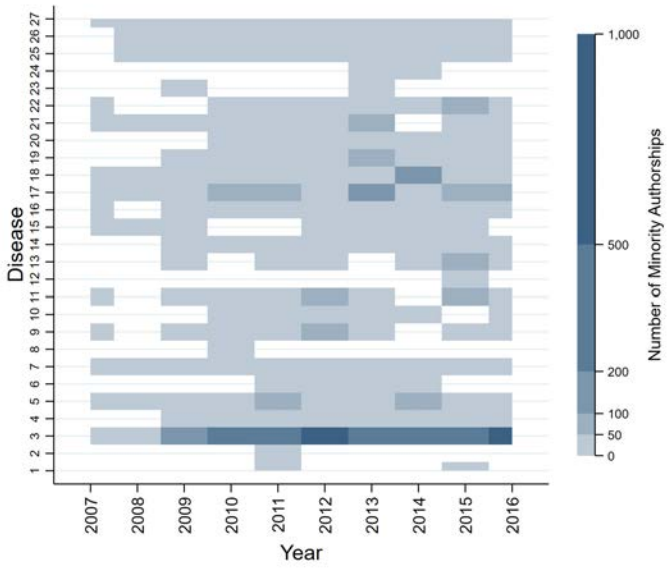


(F) Ave RRB (2002-2005), Replication Stage

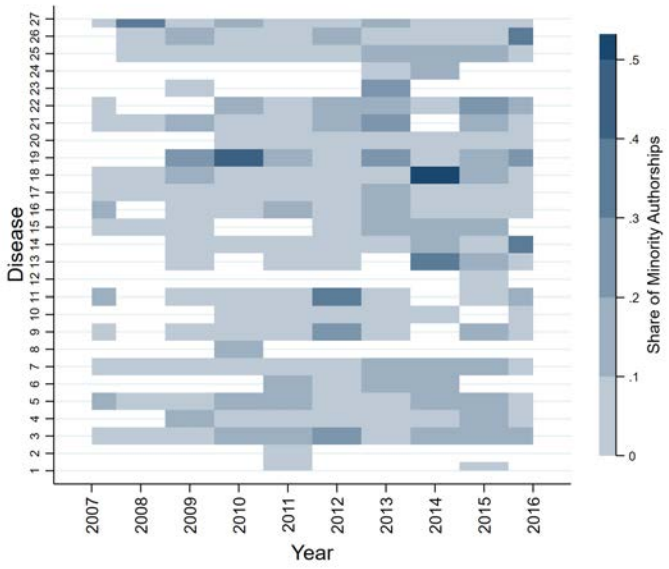
FIGURE A3.7. IMPACT OF ARRA-NIH GRANT EXPANSION ON ANCESTRAL DIVERSITY IN SUBJECTS: DISEASES WITH A NUMBER OF STUDIES $\geq$ 30

Notes: The sample comprises of 27 broader diseases from 2007 to 2016. There are in total 134 disease-year pairs for analysis combining discovery and replication stages, 133 disease-year pairs for discovery stage analysis, and 127 disease-year pairs for replication stage analysis. The sample is constructed based on studies supported by any US funding agencies between 2005 and 2016. The dependent variable is the share of minority subjects for each disease in each year. The relative referral base (RRB) is the difference in the share of people with medical conditions between minorities and whites. The disease research field characteristics includes the total number of studies funded by only non-US grants up to the previous year, the number of minority-based SNPs discovered in studies funded by only non-US grants up to the previous year, and the total number of grants from the US funding agencies up to the previous year. Coefficients of the interactions of RRB and each year indicator are plotted. The 95% confidence intervals based on robust standard errors are plotted.

A4. Ancestral Diversity in GWAS Authors



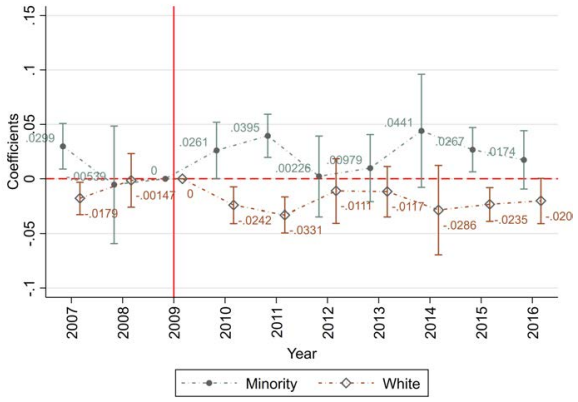
(A) Heat Map for the Number of Minority Authors



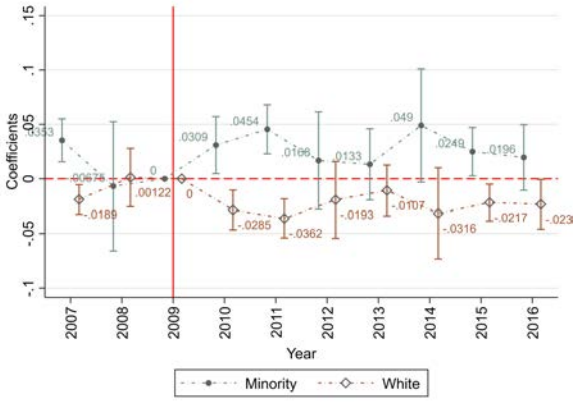
(B) Heat Map for the Share of Minority Authors

FIGURE A4.1. HEATMAP FOR ANCESTRAL DIVERSITY IN AUTHORS

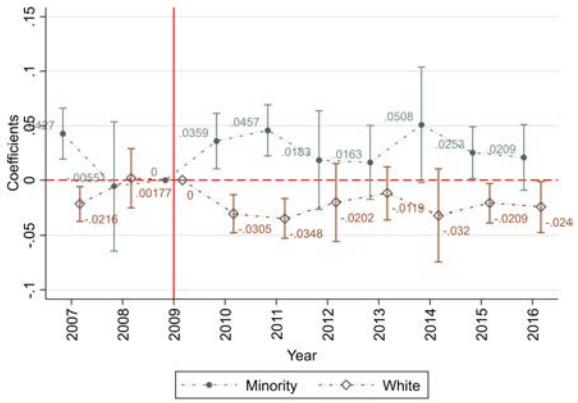
Notes: The sample comprises of 27 broader diseases from 2007 to 2016, in total 183 disease-year pairs. Given minimal GWAS publications in 2005 and 2006, the data in 2007 presented in the figure includes authors from GWAS publications in 2005, 2006 and 2007. We predict the ancestral background associated with each surname based on the 2000 US census. We adopt a conservative rule in prediction of minority authors: a last name is predicted to be associated with a minority background only if its probability of being either "Asian", "Black" or "Hispanic" is more than 0.95. The IDs corresponds to the IDs in Table 1.



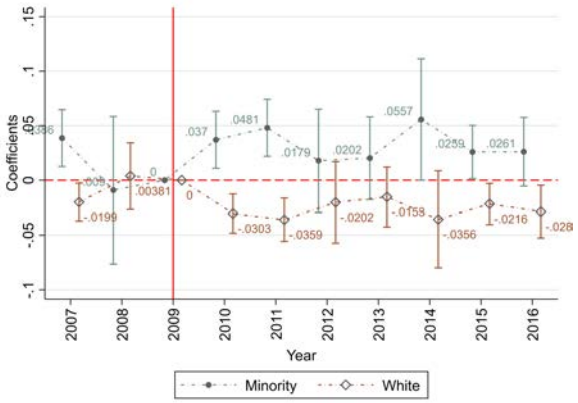
(A) RRB (2005), Minority Prediction $\geq$ 0.95



(B) RRB (2005), Minority Prediction $\geq$ 0.90



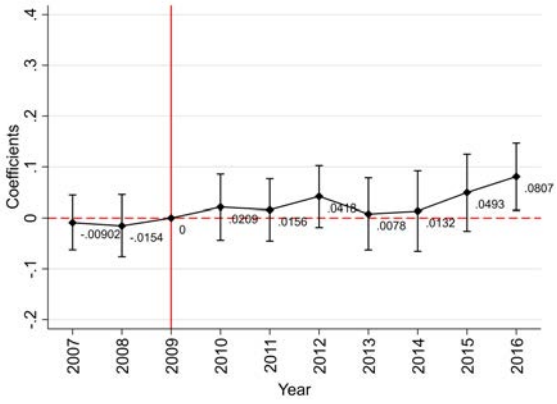
(C) RRB (2005), Minority Prediction $\geq$ 0.85



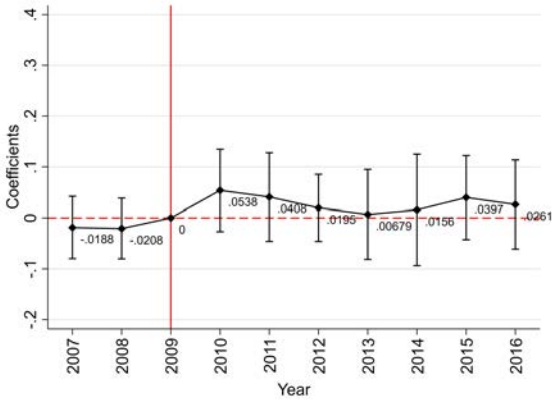
(C) RRB (2005), Minority Prediction $\geq$ 0.8

FIGURE A4.2. IMPACT OF ARRA-NIH GRANT EXPANSION ON ANCESTRAL DIVERSITY IN AUTHORS – RRB DECOMPOSITION, ALTERNATIVE CRITERIA

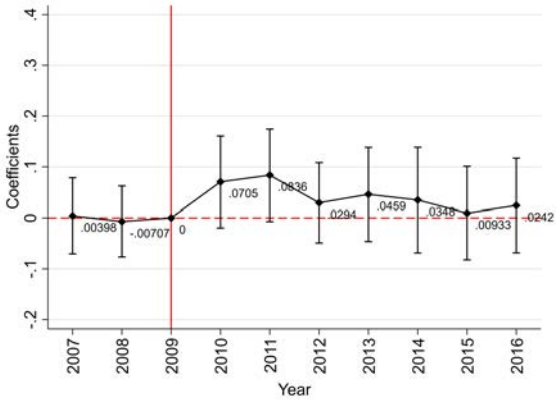
Notes: The sample comprises of 27 broader diseases from 2007 to 2016, in total 183 disease-year pairs. The sample is constructed based on studies supported by any US funding agencies between 2005 and 2016. The dependent variable is the share of minority authors for each disease in each year. We adopt a different rules in prediction of minority authors: a surname is predicted to be associated with a minority background only if its probability of being either "Asian", "Black" or "Hispanic" is more than 0.95 in Figure (A), 0.90 in Figure (B), 0.85 in Figure (C), and 0.80 in Figure (D). The relative referral base is the difference in the share of people with medical conditions between minorities and whites. The disease research field characteristics includes the total number of studies funded by only non-US grants up to the previous year, the number of minority-based SNPs discovered in studies funded by only non-US grants up to the previous year, and the total number of grants from the US funding agencies up to the previous year. Coefficients of the interactions of *RRB* and each year indicator are plotted. The 95% confidence intervals based on robust standard errors are plotted.



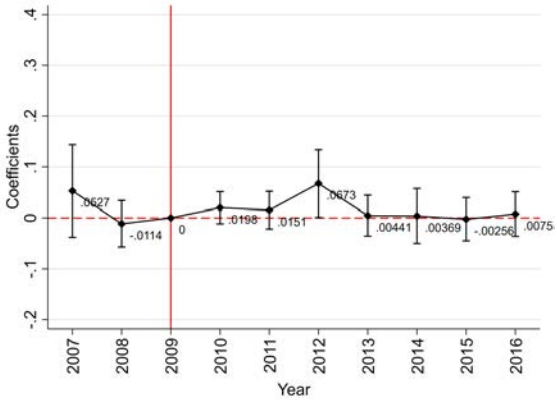
(A) Share of Studies w/ First 1 Author Being Minority



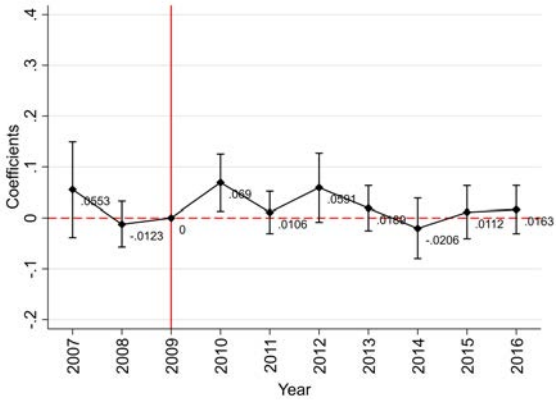
(B) Share of Studies w/ Minority Author in First 2 Authors



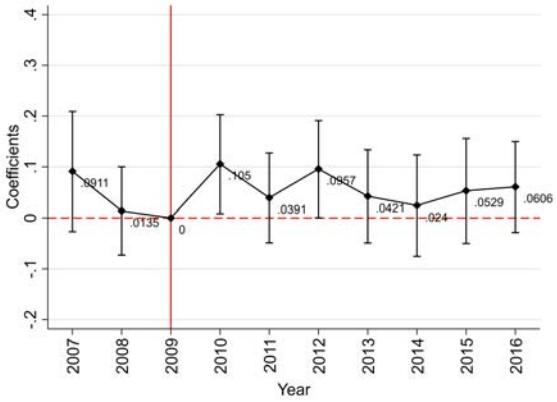
(C) Share of Studies w/ Minority Author in First 3 Authors



(D) Share of Studies w/ Last 1 Author Being Minority



(E) Share of Studies w/ Minority Author in Last 2 Authors



(F) Share of Studies w/ Minority Author in Last 3 Authors

FIGURE A4.3. IMPACT OF ARRA-NIH GRANT EXPANSION ON COLLABORATION PATTERNS

Notes: The sample comprises of 27 broader diseases from 2007 to 2016, in total 183 disease-year pairs. The sample is constructed based on studies supported by any US funding agencies between 2005 and 2016. The dependent variable is specified under each figure. We adopt a conservative rule in prediction of minority authors: a last name is predicted to be associated with a minority background only if its probability of being either "Asian", "Black" or "Hispanic" is more than 0.95. The relative referral base (RRB) is the difference in the share of people with medical conditions between minorities and whites. The disease research field characteristics includes the total number of studies funded by only non-US grants up to the previous year, the number of minority-based SNPs discovered in studies funded by only non-US grants up to the previous year, and the total number of grants from the US funding agencies up to the previous year. Coefficients of the interactions of RRB and each year indicator are plotted. The 95% confidence intervals based on robust standard errors are plotted.

Table A4.1—Impact of ARRA-NIH Grant Expansion on Ancestral Diversity in Authors – Robustness Checks using RRB

Sample	Disease with Studies ≥ 0 (1)	Disease with Studies ≥ 10 (2)	Disease with Studies ≥ 20 (3)	Disease with Studies ≥ 30 (4)
Key Independent Variable: Relative Referral Base (2005)				
Dependent Variable: Share of Minority Authors – Prob. Cutoff of 0.95				
Relative Referral Base (2005) * Post	0.0145**	0.0130**	0.0119*	0.0118*
	(0.006)	(0.006)	(0.007)	(0.007)
	(0.010)	(0.021)	(0.036)	(0.041)
	[0.0052, 0.0230]	[0.0030, 0.0210]	[0.0012, 0.0198]	[0.0011, 0.0190]
Outcome Mean (Pre-Policy)/Observations	0.075/183	0.075/174	0.075/157	0.075/134
Dependent Variable: Share of Minority Authors – Prob. Cutoff of 0.9				
Relative Referral Base (2005) * Post	0.0168**	0.0155**	0.0145**	0.0145**
	(0.007)	(0.007)	(0.007)	(0.007)
	(0.010)	(0.017)	(0.031)	(0.034)
	[0.0060, 0.0257]	[0.0039, 0.0241]	[0.0022, 0.0230]	[0.0019, 0.0223]
Outcome Mean (Pre-Policy)/Observations	0.094/183	0.094/174	0.094/157	0.094/134
Dependent Variable: Share of Minority Authors – Prob. Cutoff of 0.85				
Relative Referral Base (2005) * Post	0.0162**	0.0150**	0.0141*	0.0143*
	(0.007)	(0.007)	(0.008)	(0.008)
	(0.014)	(0.023)	(0.040)	(0.041)
	[0.0048, 0.0255]	[0.0029, 0.0239]	[0.0011, 0.0234]	[0.0011, 0.0228]
Outcome Mean (Pre-Policy)/Observations	0.105/183	0.105/174	0.105/157	0.105/134
Dependent Variable: Share of Minority Authors – Prob. Cutoff of 0.8				
Relative Referral Base (2005) * Post	0.0177**	0.0166**	0.0149*	0.0152*
	(0.007)	(0.007)	(0.008)	(0.008)
	(0.009)	(0.018)	(0.032)	(0.035)
	[0.0062, 0.0270]	[0.0044, 0.0260]	[0.0018, 0.0239]	[0.0017, 0.0236]
Outcome Mean (Pre-Policy)/Observations	0.114/183	0.114/174	0.114/157	0.114/134
Disease Research Field Characteristics	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
Disease FE	Yes	Yes	Yes	Yes

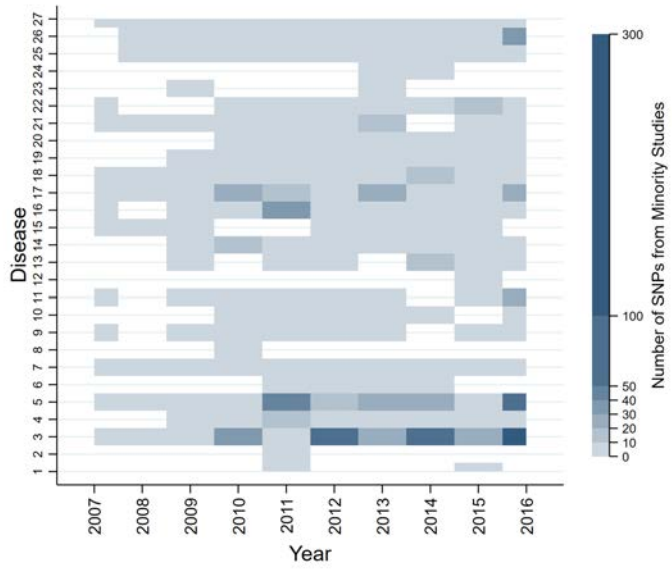
Notes: The most complete sample comprises of 27 broader diseases from 2007 to 2016, with a total 183 disease-year pairs. The sample is constructed based on studies supported by any US funding agencies between 2005 and 2016. We also require the disease research domain to have a minimum number of studies as specified in each column. The dependent variable is the share of minority authors for each disease in each year. We adopt different rules in prediction of minority authors: a surname is predicted to be associated with a minority background only if its probability of being either "Asian", "Black" or "Hispanic" is more than 0.95, 0.90, 0.85 or 0.80. The relative referral base is the difference in the share of people with medical conditions between minorities and whites. The disease research field characteristics includes the total number of studies funded by only non-US grants up to the previous year, the number of minority-based SNPs discovered in studies funded by only non-US grants up to the previous year, and the total number of grants from the US funding agencies up to the previous year. We report the Huber-White robust standard errors in the parentheses. We also report wild bootstrap cluster p-values in square brackets and wild bootstrap cluster 95% confidence intervals in square brackets, generated using boottest command (10000 replications) in Stata 14 (Roodman et al. 2019) for standard errors clustered at disease level with 27 disease clusters. Significance stars are labeled with respect to the robust standard errors. $*p < 0.1, **p < 0.05, ***p < 0.01$.

Table A4.2—Impact of ARRA-NIH Grant Expansion on Ancestral Diversity in Authors – Robustness Checks using 2002-2005 Ave. RRB

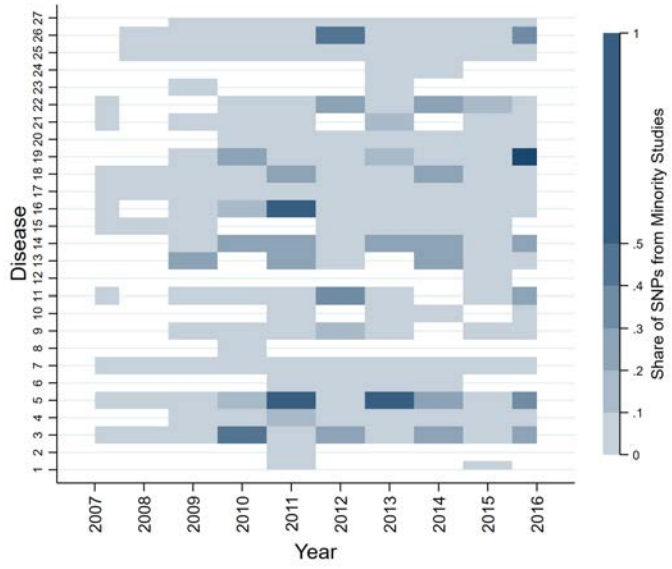
Sample	Disease with Studies ≥ 0 (1)	Disease with Studies ≥ 10 (2)	Disease with Studies ≥ 20 (3)	Disease with Studies ≥ 30 (4)
Key Independent Variable: Average Relative Referral Base (2002-2005)				
Dependent Variable: Share of Minority Authors – Prob. Cutoff of 0.95				
Ave. Relative Referral Base (2002 -2005) * Post	0.0138**	0.0123*	0.0109	0.0109
	(0.006)	(0.006)	(0.007)	(0.007)
	(0.016)	(0.034)	(0.067)	(0.068)
	[0.0038, 0.0220]	[0.0014, 0.0202]	[-0.0012, 0.0182]	[-0.0014, 0.0175]
Outcome Mean (Pre-Policy)/Observations	0.075/183	0.075/174	0.075/157	0.075/134
Dependent Variable: Share of Minority Authors – Prob. Cutoff of 0.9				
Ave. Relative Referral Base (2002 -2005) * Post	0.0161**	0.0149**	0.0137*	0.0136*
	(0.007)	(0.007)	(0.008)	(0.008)
	(0.014)	(0.026)	(0.048)	(0.052)
	[0.0046, 0.0246]	[0.0025, 0.0231]	[0.0002, 0.0212]	[-0.0002, 0.0205]
Outcome Mean (Pre-Policy)/Observations	0.094/183	0.094/174	0.094/157	0.094/134
Dependent Variable: Share of Minority Authors – Prob. Cutoff of 0.85				
Ave. Relative Referral Base (2002 -2005) * Post	0.0157**	0.0145**	0.0134	0.0134
	(0.007)	(0.007)	(0.008)	(0.008)
	(0.021)	(0.034)	(0.061)	(0.063)
	[0.0035, 0.0245]	[0.0015, 0.0230]	[-0.0009, 0.0219]	[-0.0012, 0.0212]
Outcome Mean (Pre-Policy)/Observations	0.105/183	0.105/174	0.105/157	0.105/134
Dependent Variable: Share of Minority Authors – Prob. Cutoff of 0.8				
Ave. Relative Referral Base (2002 -2005) * Post	0.0170**	0.0159**	0.0139	0.0138
	(0.008)	(0.008)	(0.008)	(0.009)
	(0.016)	(0.029)	(0.058)	(0.072)
	[0.0045, 0.0259]	[0.0024, 0.0249]	[-0.0008, 0.0221]	[-0.0019, 0.0217]
Outcome Mean (Pre-Policy)/Observations	0.114/183	0.114/174	0.114/157	0.114/134
Disease Research Field Characteristics	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
Disease FE	Yes	Yes	Yes	Yes

Notes: The most complete sample comprises of 27 broader diseases from 2007 to 2016, with a total 183 disease-year pairs. The sample is constructed based on studies supported by any US funding agencies between 2005 and 2016. We also require the disease research domain to have a minimum number of studies as specified in each column. The dependent variable is the share of minority authors for each disease in each year. We adopt different rules in prediction of minority authors: a surname is predicted to be associated with a minority background only if its probability of being either "Asian", "Black" or "Hispanic" is more than 0.95, 0.90, 0.85 or 0.80. The 2002-2005 average relative referral base is the average difference in the share of people with medical conditions between minorities and whites between 2002 and 2005. The disease research field characteristics includes the total number of studies funded by only non-US grants up to the previous year, the number of minority-based SNPs discovered in studies funded by only non-US grants up to the previous year, and the total number of grants from the US funding agencies up to the previous year. We report the Huber-White robust standard errors in the parentheses. We also report wild bootstrap cluster p-values in square brackets and wild bootstrap cluster 95% confidence intervals in square brackets, generated using boottest command (10000 replications) in Stata 14 (Roodman et al. 2019) for standard errors clustered at disease level with 27 disease clusters. Significance stars are labeled with respect to the robust standard errors. $*p < 0.1, **p < 0.05, ***p < 0.01$.

A5. SNP-Disease Associations



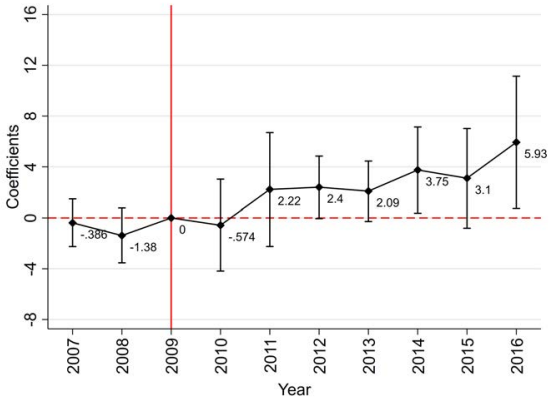
(A) Heat Map for the Number of Minority-Based SNP-Disease Associations



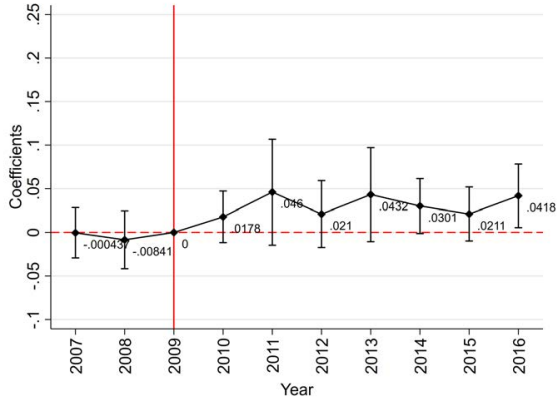
(B) Heat Map for the Share of Minority-Based SNP-Disease Associations

FIGURE A5.1. HEATMAP FOR MINORITY-BASED SNP-DISEASE ASSOCIATION DISCOVERIES

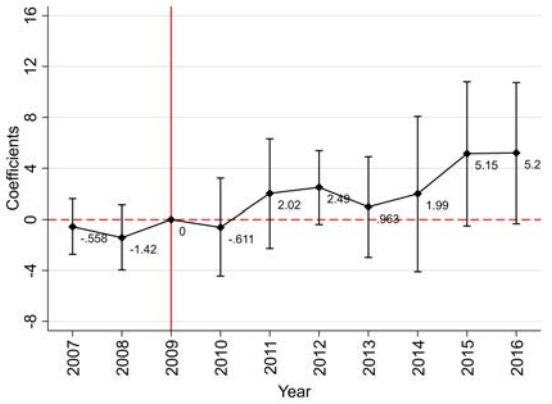
Notes: The sample comprises of 27 broader diseases from 2007 to 2016, in total 183 disease-year pairs. Given minimal GWAS publications in 2005 and 2006, the data in 2007 presented in the figure includes SNP-Disease associations from GWAS publications in 2005, 2006 and 2007. A SNP-disease association is referred to as a minority-based association as long as it is identified from a GWAS study in which all subjects in the discovery stage are the minorities. The IDs corresponds to the IDs in Table 1.



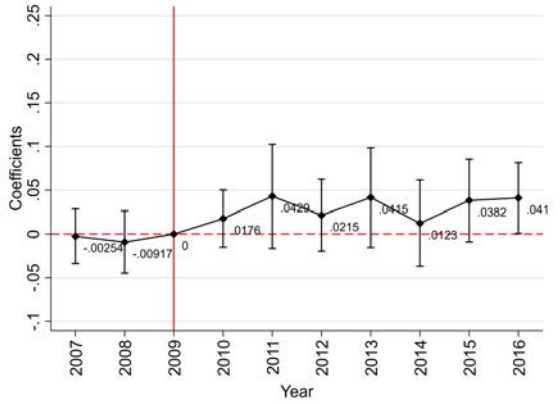
(A) Number of Minority-Based SNPs – 100% minority in discovery



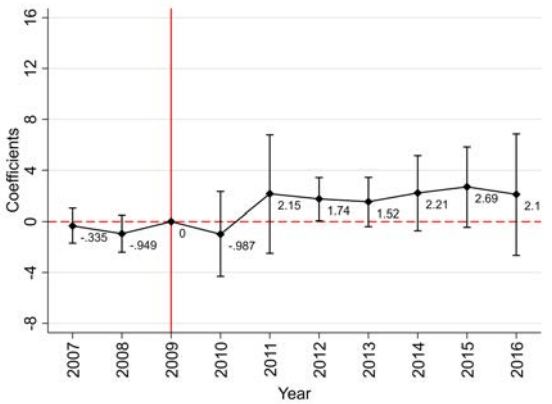
(B) Share of Minority-Based SNPs – 100% minority in discovery



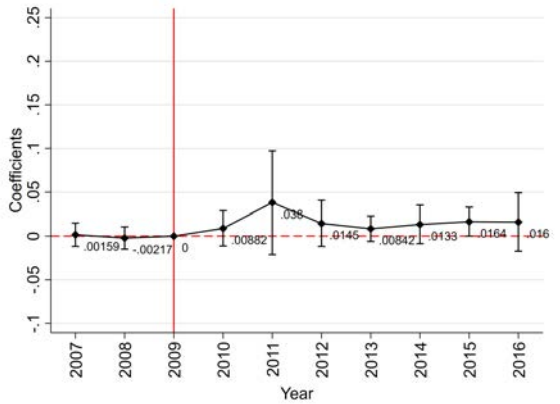
(C) Number of Minority-Based SNPs – 60% minority in discovery



(C) Share of Minority-Based SNPs – 60% minority in discovery



(E) Number of Minority-Based SNPs – 100% minority in discovery + replication



(E) Share of Minority-Based SNPs – 100% minority in discovery + replication

FIGURE A5.2. IMPACT OF ARRA-NIH GRANT EXPANSION ON MINORITY-BASED SNP-DISEASE ASSOCIATION DISCOVERIES – ALTERNATIVE CRITERIA

Notes: The sample comprises of 27 broader diseases from 2007 to 2016, in total 183 disease-year pairs. The sample is constructed based on studies supported by any US funding agencies between 2005 and 2016. The dependent variable is specified under each figure. A SNP-disease association is referred to as a minority-based association as long as it is identified from a GWAS study in which all subjects in the discovery stage are minorities. The relative referral base (RRB) is the difference in the share of people with medical conditions between minorities and whites. The disease research field characteristics includes the total number of studies funded by only non-US grants up to the previous year, the number of minority-based SNPs discovered in studies funded by only non-US grants up to the previous year, and the total number of grants from the US funding agencies up to the previous year. Coefficients of the interactions of RRB and each year indicator are plotted. The 95% confidence intervals based on robust standard errors are plotted.

Table A5.1—Impact of ARRA-NIH Grant Expansion on the Number of Minority-Based SNP-Disease Association Discoveries: Robustness Checks

	(1) Disease with Studies ≥ 0	(2) Disease with Studies ≥ 10	(3) Disease with Studies ≥ 20	(4) Disease with Studies ≥ 30
Panel A – Key Independent Variable: Relative Referral Base (2005)				
Dependent Var: Num of Minority-Based SNP-Disease Associations: 100% Minority at Discovery Stage				
Relative Referral Base (2005) * Post	2.6296**	2.6653**	2.8495**	2.8438**
	(1.185)	(1.174)	(1.312)	(1.310)
	(0.128)	(0.133)	(0.180)	(0.184)
	[-0.7523, 5.645]	[-0.7792, 5.641]	[-1.0577, 6.403]	[-1.0634, 6.430]
Outcome Mean (Pre-Policy)/Observations	0.250/183	0.250/174	0.250/157	0.250/134
Dependent Var: Num of Minority-Based SNP-Disease Associations: 60% Minority at Discovery Stage				
Relative Referral Base (2005) * Post	2.4891*	2.5321**	2.8426*	2.8409*
	(1.279)	(1.265)	(1.454)	(1.458)
	(0.171)	(0.174)	(0.223)	(0.222)
	[-1.0585, 5.761]	[-1.0648, 5.801]	[-1.4170, 6.935]	[-1.4366, 7.052]
Outcome Mean (Pre-Policy)/Observations	0.250/183	0.250/174	0.250/157	0.250/134
Dependent Var: Num of Minority-Based SNP-Disease Associations: 100% Minority in Discovery and Replication				
Relative Referral Base (2005) * Post	1.7868	1.8001	1.7157	1.7047
	(1.176)	(1.162)	(1.222)	(1.213)
	(0.120)	(0.126)	(0.159)	(0.159)
	[-0.4755, 4.177]	[-0.5499, 4.211]	[-0.6464, 4.509]	[-0.6931, 4.490]
Outcome Mean (Pre-Policy)/Observations	0.200/183	0.200/174	0.200/157	0.200/134
Panel B – Key Independent Variable: Average Relative Referral Base (2002-2005)				
Dependent Var: Num of Minority-Based SNP-Disease Associations: 100% Minority at Discovery Stage				
Ave. Relative Referral Base (2002 -2005) * Post	2.6252**	2.6592**	2.8724**	2.8767**
	(1.215)	(1.202)	(1.357)	(1.349)
	(0.123)	(0.128)	(0.175)	(0.176)
	[-0.7073, 5.751]	[-0.7282, 5.779]	[-0.9825, 6.335]	[-0.9615, 6.425]
Outcome Mean (Pre-Policy)/Observations	0.250/183	0.250/174	0.250/157	0.250/134
Dependent Var: Num of Minority-Based SNP-Disease Associations: 60% Minority at Discovery Stage				
Ave. Relative Referral Base (2002 -2005) * Post	2.5305*	2.5720**	2.9313*	2.9504*
	(1.296)	(1.281)	(1.493)	(1.495)
	(0.156)	(0.160)	(0.211)	(0.208)
	[-0.9352, 5.906]	[-0.9306, 5.921]	[-1.2794, 6.839]	[-1.2927, 7.020]
Outcome Mean (Pre-Policy)/Observations	0.250/183	0.250/174	0.250/157	0.250/134
Dependent Var: Num of Minority-Based SNP-Disease Associations: 100% Minority in Discovery and Replication				
Ave. Relative Referral Base (2002 -2005) * Post	1.7928	1.8057	1.7236	1.7083
	(0.127)	(0.134)	(0.164)	(0.163)
	[-0.4983, 4.201]	[-0.5594, 4.243]	[-0.6294, 4.508]	[-0.6504, 4.518]
Outcome Mean (Pre-Policy)/Observations	0.200/183	0.200/174	0.200/157	0.200/134
Disease Research Field Characteristics	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
Disease FE	Yes	Yes	Yes	Yes

Notes: The most complete sample comprises of 27 broader diseases from 2007 to 2016, with a total 183 disease-year pairs. The sample is constructed based on studies supported by any US funding agencies between 2005 and 2016. We also require the disease research domain to have a minimum number of studies as specified in each column. The dependent variable is specified in the table. A SNP-disease association is referred to as a minority-based association as long as it is identified from a GWAS study which should satisfy a certain criterion as specified in the column title. The relative referral base (RRB) is the difference in the share of people with medical conditions between minorities and whites. The disease research field characteristics includes the total number of studies funded by only non-US grants up to the previous year, the number of minority-based SNPs discovered in studies funded by only non-US grants up to the previous year, and the total number of grants from the US funding agencies up to the previous year. Coefficients of the interactions of RRB and each year indicator are plotted. The 95% confidence intervals based on robust standard errors are plotted.

Table A5.2—Impact of ARRA-NIH Grant Expansion on the Share of Minority-Based SNP-Disease Association Discoveries: Robustness Checks

	(1) Disease with Studies ≥ 0	(2) Disease with Studies ≥ 10	(3) Disease with Studies ≥ 20	(4) Disease with Studies ≥ 30
Panel A – Key Independent Variable: Relative Referral Base (2005)				
Dependent Var: Share of Minority-Based SNP-Disease Associations: 100% Minority at Discovery Stage				
Relative Referral Base (2005) * Post	0.0329***	0.0341***	0.0442***	0.0425***
	(0.013)	(0.013)	(0.015)	(0.015)
	(0.085)	(0.082)	(0.054)	(0.065)
	[-0.0041, 0.0702]	[-0.0040, 0.0716]	[-0.0007, 0.0769]	[-0.0037, 0.0720]
Outcome Mean (Pre-Policy)/Observations	0.010/183	0.010/174	0.010/157	0.010/134
Dependent Var: Share of Minority-Based SNP-Disease Associations: 60% Minority at Discovery Stage				
Relative Referral Base (2005) * Post	0.0331**	0.0344***	0.0460***	0.0443***
	(0.013)	(0.013)	(0.016)	(0.016)
	(0.114)	(0.112)	(0.062)	(0.078)
	[-0.0072, 0.0728]	[-0.0068, 0.0741]	[-0.0028, 0.0821]	[-0.0058, 0.0779]
Outcome Mean (Pre-Policy)/Observations	0.010/183	0.010/174	0.010/157	0.010/134
Dependent Var: Share of Minority-Based SNP-Disease Associations: 100% Minority in Discovery and Replication				
Relative Referral Base (2005) * Post	0.0162*	0.0168*	0.0207*	0.0189*
	(0.009)	(0.009)	(0.011)	(0.011)
	(0.067)	(0.058)	(0.049)	(0.074)
	[-0.0010, 0.0328]	[-0.0006, 0.0343]	[0.0002, 0.0431]	[-0.0028, 0.0421]
Outcome Mean (Pre-Policy)/Observations	0.009/183	0.009/174	0.009/157	0.009/134
Panel B – Key Independent Variable: Average Relative Referral Base (2002-2005)				
Dependent Var: Share of Minority-Based SNP-Disease Associations: 100% Minority at Discovery Stage				
Ave. Relative Referral Base (2002 -2005) * Post	0.0307**	0.0319**	0.0425***	0.0417**
	(0.013)	(0.013)	(0.016)	(0.016)
	(0.120)	(0.119)	(0.081)	(0.095)
	[-0.0067, 0.0717]	[-0.0067, 0.0735]	[-0.0054, 0.0794]	[-0.0079, 0.0752]
Outcome Mean (Pre-Policy)/Observations	0.010/183	0.010/174	0.010/157	0.010/134
Dependent Var: Share of Minority-Based SNP-Disease Associations: 60% Minority at Discovery Stage				
Ave. Relative Referral Base (2002 -2005) * Post	0.0313**	0.0325**	0.0449***	0.0442***
	(0.013)	(0.013)	(0.017)	(0.017)
	(0.150)	(0.148)	(0.088)	(0.100)
	[-0.0094, 0.0742]	[-0.0090, 0.0760]	[-0.0063, 0.0833]	[-0.0093, 0.0800]
Outcome Mean (Pre-Policy)/Observations	0.010/183	0.010/174	0.010/157	0.010/134
Dependent Var: Share of Minority-Based SNP-Disease Associations: 100% Minority in Discovery and Replication				
Ave. Relative Referral Base (2002 -2005) * Post	0.0148	0.0154	0.0194*	0.0182
	(0.009)	(0.009)	(0.011)	(0.012)
	(0.109)	(0.099)	(0.079)	(0.111)
	[-0.0030, 0.0324]	[-0.0027, 0.0335]	[-0.0023, 0.0417]	[-0.0052, 0.0418]
Outcome Mean (Pre-Policy)/Observations	0.009/183	0.009/174	0.009/157	0.009/134
Disease Research Field Characteristics	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
Disease FE	Yes	Yes	Yes	Yes

Notes: The most complete sample comprises of 27 broader diseases from 2007 to 2016, with a total 183 disease-year pairs. The sample is constructed based on studies supported by any US funding agencies between 2005 and 2016. We also require the disease research domain to have a minimum number of studies as specified in each column. The dependent variable is specified in the table. A SNP-disease association is referred to as a minority-based association as long as it is identified from a GWAS study which should satisfy a certain criterion as specified in the column title. The relative referral base (RRB) is the difference in the share of people with medical conditions between minorities and whites. The disease research field characteristics includes the total number of studies funded by only non-US grants up to the previous year, the number of minority-based SNPs discovered in studies funded by only non-US grants up to the previous year, and the total number of grants from the US funding agencies up to the previous year. Coefficients of the interactions of RRB and each year indicator are plotted. The 95% confidence intervals based on robust standard errors are plotted.

Appendix B is available upon request.