Do biomedical researchers study topics related to their own ethnicity?: Evidence from the NIH

Wesley H. Greenblatt
Harvard Medical School and
Boston Children's Hospital
300 Longwood Avenue
Boston, MA 02115

Young Hou
Darden School of Business
University of Virginia
100 Darden Boulevard
Charlottesville, VA 22903

Sukhun Kang
Technology Management
UC Santa Barbara
1321 Phelps Hall
Santa Barbara, CA 93106

Aug 15, 2025

Abstract

The scientific interests of researchers are fundamental to determining the direction of innovation, yet may be shaped by an investigator's background, exposure to topics, and perceived problem importance. We study the role of ethnicity in shaping investigator research direction at the National Institutes of Health, analyzing over 100,000 R01 grants from more than 32,000 distinct investigators between 1980 and 2015. We combine image and name-based predictions with clerical review to identify the ethnicity of investigators. Across four measures of the ethnic orientation of research, including relative disease mortality, and keywords related to ethnicity, geographic location, and socioeconomic status and healthcare disparities, we find investigators are more likely to study topics related to their own ethnic background. We also examine if minority researchers have a disparity in the resources available to them for conducting research, which may contribute to distortions in the types of innovations produced. We find that while there is a disparity in the raw grant renewal rate of minority investigators, this disparity resolves after controlling for investigator, grant, and institutional characteristics. These findings suggest how frictions for minorities entering and remaining in research careers may impact the development of innovations that address the healthcare needs of minority populations.

Keywords: Ethnicity, disparities, biomedical innovation, government funding

JEL Classification: O32, I14, J71

Acknowledgements: Address correspondence to wesley.greenblatt@childrens.harvard.edu. We thank Damian Wen and Sienna Ro for their helpful research assistantship. We also thank Valentina Assenova and seminar participants of the West Coast Research Symposium on Technology Entrepreneurship, National Bureau of Economic Research Productivity Seminar, and Responsible Research in Business and Management for their feedback. The authors have no conflicts of interest.

1. Introduction

It seems a truism that what scientists choose to study—and which scientists receive funding to support their research—is fundamental to determining the rate and direction of innovation. While the choice of what to study is shaped by economic incentives (Budish et al. 2015), organizational and team environment (Lerner and Wulf 2007, Wu et al. 2019), institutions surrounding a researcher (Furman and Stern 2011), and perceived scientific and career risk and returns (Azoulay and Greenblatt 2025), researcher background and life experience, which can affect the perceived problem prominence, may also shape this choice. We study this in the context of investigator ethnicity, asking two related questions. First, are investigators more likely to study topics and diseases related to their own ethnicity? Second, do minority investigators have disparities in receiving resources to carry out projects, through which any differences in research topic choice could be amplified?

It has been a longstanding goal of leading scientists and policymakers to diversify the pipeline of biomedical researchers (National Research Council 2000, Sims 1992, Tilghman et al. 2021). These efforts are often motivated by equity considerations and a desire to fully tap the supply of scientific talent available (Oakes 1990), or in the case of physician-scientists, that patient-doctor homophily may lead to improved clinical care (LaVeist et al. 2003, Takeshita et al. 2020). Another potential reason to recruit diverse scientists is that, in addition to potentially improving the *quality* of science produced, they may also ask *different questions*, shaping in what direction the scientific frontier expands. This is especially salient in the context of biomedicine, as asking different questions may lead to different patient healthcare needs being addressed by biomedical innovation.

An emerging line of research has studied how investigator demographic background shapes innovation, focusing on the role of gender. For example, female inventors have been shown to be more likely to patent ideas and products related to women's health (Koning et al. 2020, Koning et al. 2021). Investigator demographic background may also shape the approach to research. In the clinical research context, female researchers may be more likely to include a gender analysis, recruit female patients, and focus on patient safety (Gupta 2022, Hemosilla 2023, Nielsen et al. 2017).

More generally, investigators may select their research topic based in part on exposure and perceived problem importance. The history of medicine has many anecdotes of physician-scientist research agendas that emerged after exposure to the clinical problems of their patients (Jain et al. 2019). For example, the work of Michael Brown and Joseph Goldstein on cholesterol metabolism—which earned them a Nobel Prize and contributed to the development of statins—was directly stimulated by caring for a patient with familial hypercholesterolemia (Goldstein and Brown 1997). More broadly, Bell et al. (2019) find a strong association between fathers and sons' propensity to patent in the exact same narrow patent class, a finding most easily explained by early socialization around the feasibility and desirability of a research career related to these topics. Innovators may also have intrinsic motivations for social impact related to innovator-user homophily. In the context of new consumer product ventures, founders are more likely to create products for the gender and socioeconomic status of the peers they have been exposed to, independent of financial incentives (Einio et al. 2023). Combined with same-ethnic homophily in social networks and between physicians and patients (Leszczensky and Pink 2019, Marrast et al. 2014), this suggests possible pathways that might lead investigators to pursue research topics related to their own ethnic background.

We explore this question at the National Institutes of Health (NIH), which plays a central role in funding biomedical research in the United States (Moses et al. 2015). The NIH is an attractive setting to study how ethnicity shapes the selection of research projects for two reasons. First, most academic biomedical research is published in peer-reviewed journals and funded NIH grant data is public. When combined with the rich bibliometrics in biomedicine, this provides a robust "paper trail" we can follow and characterize. Second, as NIH research is a source of significant improvement in health and positive economic spillovers, understanding how the NIH may more effectively utilize its funds to support innovation—

including in the relative supply of innovations addressing the healthcare needs of different ethnicities—is of considerable policy interest (Azoulay et al. 2019, Fleming et al. 2019).

We study the role of investigator ethnicity across 103,164 NIH-funded R01 and R01-equivalent grant cycles between 1980 and 2015. To study this, one would ideally have access to all grant proposals (regardless of funding status), reviewer scores, and grant outcomes in the form of publications and citations that can be traced to a particular funding stream. With access limited to publicly available data, we instead focus on funded grant applications. After controlling for a comprehensive set of covariates for investigator demographics, institution quality, and grant characteristics, we examine the relationship between ethnicity and the ethnic orientation of the research pursued as well as whether the grant was renewed after a competitive review for an additional cycle of funding.

A central challenge to our study is identifying investigator ethnicity (Borrell et al. 2021, Chaturvedi 2001, Mateos 2007). Unfortunately, the ethnicity of NIH-funded investigators is not publicly available and we need to infer ethnicity. Different methods of inferring ethnicity have distinct strengths and weaknesses (Marx 2024). For instance, name-based prediction in the United States tends to perform poorly at distinguishing Black vs. white investigators, but is more accurate at distinguishing Asian vs. white investigators. In contrast, artificial intelligence-based image analysis tend to perform well at classifying Asian vs. white but poorly at Hispanic vs. white.

To address this challenge, we build on the methodology developed by Cook et al. (2022). Specifically, for each investigator, we first obtain a picture. Our primary analysis assigns ethnicity using manual clerical review, integrating across both picture and full name, supplemented by biographical details from professional profiles such as languages spoken, location of training, or membership in ethnic-affiliation groups. We then perform extensive robustness checks using alternative methods of predicting ethnicity to support that our results are not driven by measurement error in assigning ethnicity.

The second empirical challenge to overcome is measuring the "ethnicness" of research. Given the lack of clarity in what constitutes ethnic research, it seems unlikely that a single measure can fully capture the phenomenon. Instead, we develop four distinct yet complementary measures of the ethnic orientation of research, using Medical Subject Headings (MeSH) to help us characterize the research. We first focus on what diseases are studied. We characterize each disease's ethnic orientation by calculating the ethnicity-specific mortality rate relative to the white mortality rate (e.g., sickle cell anemia has a much higher Black mortality rate than white mortality rate, but varicella has fairly similar mortality rates across ethnicities). Second, we look at research directly tagged with an ethnicity MeSH keyword (e.g., *Black or African-American*, *Hispanic or Latino*). Third, we examine keywords related to a country or geographic location (e.g. *China*, *Mexico*). Finally, we look at topics adjacent to ethnicity including minorities (e.g., *minority health*, *minority groups*), disparities (e.g., *healthcare disparities*, *socioeconomic disparities in health*), and socioeconomic status (e.g., *social determinants of health*, *poverty*).

Overall, we find that investigator ethnicity is associated with having research topics related to their own ethnicity. For instance, compared to white investigators, Asian investigators had a 0.13 (9.7%) increase in the maximum relative Asian disease mortality, calculated relative to white disease mortality, among grant cycle publications. Similarly, Hispanic investigators had a 1.28 (57.7%) higher maximum Hispanic relative disease mortality, and Black investigators a 12.91 (207.0%) higher maximum Black relative disease mortality. This effect was highly specific: ethnic investigators were not more likely to study topics related to other ethnicities besides their own. We also found that, while there are large disparities in the raw grant renewal rates by investigator ethnicity, no significant disparity is observed after controlling for observable investigator, grant, and institutional characteristics, including scientific productivity during the grant cycle.

Prior studies have explored how ethnicity relates to funding decisions at the NIH. For example, Ginther et al. (2011) found that both Asian and Black investigators were less likely to be funded than white investigators on initial grant applications. Other research has examined specific steps within the NIH peer review process at which disparities might arise (Erosheva et al. 2020, Hoppe et al. 2019), noting that Black

investigators were more likely to apply to NIH Institutes or Centers with lower funding rates (Lauer et al. 2021). Building on this work, we contribute to the literature in several important ways. First, we link investigator topic choice to a concept of potential public health significance, the ethnic orientation of the research. Second, we ask a novel research question: conditional on receiving a first grant, how likely is it to be renewed, after controlling for observable characteristics? This is important as separate mechanisms may be involved in grant renewal compared to initial grant application decisions, and there are distinct policy implications at each stage. Finally, our data spans 35 years, allowing us to trace how the role of ethnicity among NIH-funded investigators has evolved over time.

2. Results

Descriptive Statistics

Appendix table A1, table A2, and figure A1 present descriptives of investigators and grant cycles. Reflecting the very high level of competition for R01-equivalent grants, the sample consists of highly accomplished scientists. On average, investigators begin the grant cycle 22.6 years after completing their terminal degree, with 58.2 prior publications and 2,371 prior citations. Applications for grant renewal are funded at a much higher rate than applicants for a new R01-equivalent grant (Lauer 2016). Overall, 46.3% of grant cycles were renewed. Renewed grant cycles were associated with younger investigators (17.7 vs. 19.7 years of career experience), greater research productivity (8.3 vs 7.3 publications which received 691 vs. 450 long-run citations), and a greater percentage of the investigator's research activity funded by the focal grant (46.0% vs. 37.9% of publications during the grant cycle).

As a first step to identifying the ethnicity, we attempt to obtain a photo for each investigator in our sample. Overall, we found a picture for 32,011 (86.0%) of investigators. As would be expected, there is selection in which investigators have a photo available (Table A3). Investigators with a photo were older (24.1 vs. 21.7 years of career experience), more recently engaged in research (2004 vs. 1993 for year of most recent grant cycle), and more scientifically accomplished (60.9 vs. 39.5 publications). In particular, investigators that died, retired, moved abroad, or exited science, especially before the modern internet era, were less likely to have a picture. However, even among investigators whose last grant cycle started in the 1980s, 63% of investigators had a photo available (Figure A2). This selection is relevant to our research question only if the presence of a photo is associated with both ethnicity and our outcomes of interest. Reassuringly, this does not appear to be a major concern: predicted ethnicity based on last name is not correlated with photo availability after controlling for observable investigator characteristics (Table A4).

Our preferred method of assigning ethnicity is based on clerical review, including name, picture, and biographical details available online.[1] Using this method, our sample is 81% white, 15% Asian, 3% Hispanic, and 1% Black.[2]

[1] We also explore alternative methods of assigning ethnicity based solely on image analysis or name. Tables A5 and A6 present descriptive statistics, and figures A3-A6 show the distribution of assigned ethnicity by our preferred method of clerical review across different predicted ethnicity probabilities (i.e., by name or image). Table A7 shows the correlations across the different ethnicity measures. Two trends emerge from this. First, different methods vary in their effectiveness across ethnic groups. For instance, name-based identification performs relatively poorly for Black investigators, while image-based analysis is less accurate for identifying Hispanic investigators. This supports our approach of integrating multiple sources of information to infer potential ethnicity. Second, the overall distribution of ethnicities in the sample remains consistent across the different identification methods, suggesting robustness in our aggregate estimates.

[2] It is difficult to directly compare these numbers to publicly available information of NIH investigator ethnicity, as potential comparators are over a different time period and may include a different set of grant funding mechanisms. Additionally, in recent years over 10% of NIH research project grant applicants did not include a self-designated race (Lauer and Bernard 2022). Based on investigator self-reporting, in 2020, the National Cancer Institute R01-equivalent investigators were 1% black, 5% Hispanic, 29% Asian, and 64% white (this excludes 8% of investigators who did not report a race) (Ong 2021). Similarly, Hoppe et al. (2019), using data from 2011-2015 for awarded R01 NIH grants to only Black and white investigators found 1.4% of awarded grants were to Black investigators. Ginther et al. (2011),

Ethnicity and the direction of research

Table 1 reports the association between ethnicity and the ethnic orientation of research during the grant cycle, with descriptive statistics of outcome measures available in Table A8. Across these measures, there is a consistent pattern of investigators being more likely to conduct research on topics associated with their own ethnicity. For example, compared to white investigators, Asian investigators had a 0.13 (9.7%) increase in the maximum Asian disease relative mortality, calculated relative to white disease mortality, among their disease-oriented grant cycle publications, while Hispanic investigators had a 1.28 (57.7%) increase in mean Hispanic relative disease mortality and Black investigators had a 12.91 (207.0%) increase in mean Black relative disease mortality (Panel A).

A similar pattern is observed when measuring the ethnic orientation of research using ethnicity MeSH (Panel B) or country and geographic location MeSH (Panel C), although these are sparser outcomes. An additional 1.9% (a 93.0% increase) of Asian investigators had at least one publication tagged with an Asian ethnicity MeSH, and an additional 4.2% (a 121.1% increase) of Asian investigators had at least one publication tagged with an Asian country MeSH. Similarly, an additional 2.7% (a 103.5% increase) and 5.5% (a 203.0% increase) of Hispanic investigators have at least one publication with a Hispanic ethnicity MeSH or Hispanic country MeSH, respectively, while an additional 9.6% (a 177.2% increase) and 4.7% (a 236.5% increase) of Black investigators had at least one publication with Black ethnicity MeSH or Black country MeSH. These results are highly specific: across most measures, Asian investigators are more likely to study Asian research topics but not Black or Hispanic research topics, while Black investigators are more likely to study Black research topics but not Asian or Hispanic research topics and so forth

Finally, we examine ethnicity-adjacent topics, with MeSH focused on minority topics, healthcare disparities, or socioeconomic status (Panel D). Both Hispanic and Black investigators, but not Asian investigators, study these topics at higher rates relative to white investigators.

As a complementary approach to measuring the ethnic orientation of research, rather than focusing on an investigator's own publications, we instead look at how their publications are used by other researchers (Table 2). Specifically, we measure the ethnic orientation of articles citing an investigator's grant cycle publications. Using analogous measures, we examine the mean relative disease mortality among citations (Panel A) as well as the count of citations from papers with a particular ethnicity, country, or ethnic-adjacent topic MeSH adjusted with an offset for the total number of citations received (Panels B, C, and D). Overall, across these measures, we find similar results, where there is a strong and significant association between an investigator's ethnicity and the ethnic orientation of papers building on their work.

While the NIH does support clinical research, its focus is on basic science aimed at extending the scientific frontier. To ensure we are not capturing solely an ethnic orientation in clinical research, we limit our analysis to basic science grant cycle publications. We then examine the relative disease mortality of an investigator's basic science publications (Table A9) as well as look at the relative disease mortality and use of ethnicity and country topics among papers citing an investigator's basic science publications (Table A10).[3] While some estimates for Black and Hispanic investigators are imprecisely estimated, we continue to see associations between these measures of the ethnic orientation of basic science research and investigator ethnicity.

Ethnicity, ethnic research, and grant renewal

using awarded R01 grants from 2000-2006, had an investigator ethnicity distribution of 0.9% Black, 3.2% Hispanic, 11.7% Asian, 71.9% white, and 12.3% who did not self-report a race or ethnicity.

[3] As country and geographic topics are of limited relevance to basic science articles but are relevant for the translational and clinical research building on them, we include this measure for citations to basic science articles but not for the basic science articles themselves.

We next test for evidence of disparities in grant renewal rates by ethnicity through which any difference in topics explored could be compounded over time. Table 3 presents these results. When examining the "raw renewal rate," controlling only for grant year, both Asian and Black investigators have a lower renewal rate. The grants of Asian investigators have a 1.7% lower renewal rate (a 3.7% decline) and the grants of Black investigators have a 10.2% lower renewal rate (a 22.0% decline). There is no significant difference between Hispanic and white investigator renewal rates. However, after inclusion of controls for investigator, grant, and institutional characteristics, as well as scientific productivity during the grant cycle, there is no significant difference in the renewal rate, compared to white investigators, across each ethnicity.[4]

We also examine how the association between ethnicity and grant renewal rate may have evolved over time. Given the increasing competition for R01-equivalent grants, as well as evolving NIH policies, these results should be interpreted with caution. We find the magnitude of the association between ethnicity and grant renewal, controlling for observable factors, to remain stable (Figure A7).

In addition to focusing on the association between ethnicity and grant renewal, we also examine how conducting ethnic research topics can affect grant renewal rates, independent of investigator ethnicity (Table 4). When measuring the ethnic orientation of research by ethnicity or country topics, such research is renewed at a lower rate when controlling only for grant year, however, this difference resolves after including controls for investigator, grant, and institutional characteristics as well as productivity during the grant cycle. While there is a statistically significant difference for some ethnicities when measuring the ethnic orientation of research using disease relative mortality, the magnitude of this effect is small.[5]

Robustness analysis and ancillary results

We conduct a number of robustness checks to assess the sensitivity of our estimates to alternative assumptions. Given the challenges in measuring ethnicity, we test several alternative approaches of assigning ethnicity (Table A12). First, we repeat our analysis, but use the most likely ethnicity based solely on information from image analysis (Panel A) or name (Panel B). These have the advantage of avoiding any potential bias from manual review. Name-based prediction also allows us to use the full sample of investigators without restricting it to those with a profile picture identified. Second, rather than assigning each investigator to a single ethnicity, we allow each investigator to have a continuous predicted probability of being each ethnicity based on image analysis or name (Panels C and D). In addition to the advantages mentioned above, this may better capture uncertainty around individuals whose name or picture is more ethnically ambiguous. Finally, we focus on those investigators for which we have a high degree of certainty in the assigned race (Panel E). For this we limit our analysis to investigators where either image- and name-based methods assign an ethnicity with over a 95% probability that is also concordant with the manually assigned ethnicity. Our results using solely image or name analysis should to be interpreted with caution as these methods may be subject to systematic measurement error. For instance, name-only methods may perform less well in identifying Black investigators while image-only methods may perform less well for Hispanic investigators, or may interpret darker-skinned South Asians as Black. Overall, our results are robust to integrating across these different approaches.

As an alternative approach to measuring the ethnic orientation of research, we focus on disease that are less controversial as having a strong ethnic association, whether due to genetics, geographical exposure, or epidemiologic associations. We identify 19 such diseases, with examples including cystic fibrosis for whites, hepatitis B for Asians, Chagas disease for Hispanics, and sickle cell disease for Blacks (see Appendix B for details). While the association for Hispanics is imprecisely estimated due to those diseases with a strong Hispanic association tending to be very rare, our findings on white, Asian, and Black

[4] It is possible peer review committees may have distinct patterns of scrutiny for the first time a grant comes up for renewal compared to subsequent evaluations for renewal. We find overall similar dynamics for the first renewal cycle as subsequent renewal cycles (Table A11).

[5] We also do not find evidence ethnic investigators are penalized for studying topics related to their own ethnicity at grant renewal.

ethnicities are robust to this approach of measuring the ethnic orientation of research (Tables A13 and A14).

It might be possible that, due to ethnic homophily in coauthorship networks, ethnic coauthors rather than the NIH-funded investigator are driving work on ethnic-oriented research topics. To examine this, we repeat our analysis limiting it to papers where the investigator was the last author.[6] Our results are robust to this specification (Table A15).

We used Author-ity to disambiguate investigator publications. While it has been shown to be highly accurate, it is still likely that it is better at disambiguating the publications for rare as opposed to common names (Lerchenmueller and Sorenson 2016). To ensure our results are not being driven by investigators with poorly disambiguated publications, we split the sample at the median name frequency based on frequency within the entire corpus of PubMed. With the caveat that ethnicity is related to name frequency, we find similar results across both groups although some are imprecisely estimated (Table A16).

As another robustness check, we employ the "post-double-selection" lasso estimator from Belloni et al. (2014). This estimator uses lasso to select covariates to predict both ethnicity and the outcome variable, and then estimates the effect of ethnicity by the linear regression of the outcome on ethnicity and the union of the set of variables selected in the two variable selection steps. This approach can be especially helpful in selecting among a large number of potential controls, and can reduce the risk of overfitting the model and improve the precision of the estimates. The magnitudes of the coefficients by this method are similar to our earlier estimates (Table A17).

Finally, we use a bounding technique proposed by Oster (2019) to gauge the sensitivity of our results to omitted variable bias. The intuition behind this approach is that the stability of the coefficient on ethnicity when varying the set of control variables included in the model, scaled by the movement in R^2, provides information about the potential impact of unobserved covariates. Generating these bounds requires assuming proportionality between the covariances of the outcome with observed and unobserved covariates and positing a maximum value for R^2 if the regression could include all observed and unobserved covariates. Oster's technique generates δ, which can be interpreted as the degree of selection on unobservables relative to observables necessary to reduce the magnitude of the effect of the regressor of interest to zero. Across our measures of the ethnic orientation of research, δ is well above one, the threshold recommended by Oster to suggest robustness to the influence of unobservable covariates (Table A18).

3. Discussion

We find that, using a dataset of all awarded NIH grants over a thirty-five year period, investigator ethnicity is associated with studying topics and diseases related to their own ethnicity across a range of different measures of the "ethnicness" of their research. To examine how such differences in topic choice could be magnified over time, we also examine disparities in grant renewal rates. While there are large and significant disparities in the raw grant renewal rates, these resolve after controlling for investigator, grant, and institutional characteristics. Does who becomes a scientist matter for what science is undertaken? At least with regards to ethnicity and the ethnic orientation of their research, the answer seems to be yes.

A significant strength of our study is our ability to capitalize on recent advances in machine learning to combine artificial-intelligence analysis of investigator pictures with clerical review and name analysis to determine investigator ethnicity. Given the prominence of the NIH and the importance of R01 grant mechanisms in funding biomedical research in the United States, we believe our results are of substantial policy interest. While we focus on NIH-funded research, it seems likely our results can generalize to other biomedical research settings. More broadly, our study suggests that reducing ethnic disparities in the

[6] A norm in biomedical research assigns the last authorship position to the principal investigator (Dance 2012).

biomedical innovation process upstream of healthcare delivery may, over time, contribute to reducing disparities in healthcare outcomes.

Our result hinges on our ability to assign ethnicity to investigators. It is likely that there are some cases where our assigned ethnicity and an investigator's self-identified ethnicity are discordant. Each of the methods we employ to determine ethnicity—clerical review, image analysis, name-based prediction—has weaknesses as well as strengths. We are reassured against the concern that misidentification of ethnicity may be driving our results by the persistence of our findings across different methodologies.

The validity of our results further relies on the maintained assumption that there is no significant omitted variable bias: that is, there are no covariates associated with both ethnicity and the ethnic orientation of research or grant renewal not included in our model. We are reassured this is not a large concern given our rich controls as well as the stability of our findings across robustness checks. In particular, it strains credulity to suppose there is an unobserved factor associated with Asian investigator ethnicity and Asian-oriented research—yet not similarly be associated with Black- or Hispanic-oriented research—and so on for each ethnicity.

Our results may also understate the true association between investigator ethnicity and ethnic research topics. Our sample includes only those grant applications that were funded. To the extent ethnic investigators proposing ethnic research are funded at a lower rate in their initial grant application, it would make the observed association in our sample weaker than if we had access to the full grant applicant pool.

We emphasize that observing no disparity in grant renewal rate, conditional on observables, does not imply there are no ethnic disparities involved. Many of the controls in our regression have well-documented disparities. For instance, papers by minority researchers receive fewer citations and spend longer under review (Koffi et al. 2024, Liu et al. 2023) and scientists from lower socioeconomic status receive less recognition for their work (Airoldi et al. 2025, Abramitzky et al. 2024). Prior to entry into our sample, disparities may exist across several stages of the research career pipeline. These include differences in the likelihood of pursuing research careers, the quality of the research training or mentorship, job placement and access to resources, and the ability to secure an initial NIH grant (Bell et al. 2019, Ginther et al. 2011, Jackson et al. 2019, Low et al. 2019, Yehia et al. 2014)

The desired amount of ethnic diversity among researchers—or the ethnic orientation of supported research topics—is a decision for science policymakers and funders. It is striking that we find no differences in NIH grant renewal rates by ethnicity, conditional on observables, in contrast to earlier work finding ethnic disparities in the initial grant award rate at the NIH (Ginther et al. 2011). While both may be helpful, our results suggest that there may be greater returns to policy efforts to diversify the scientific workforce that focus on recruiting and retaining early-career minority scientists prior to winning their first R01-equivalent grant rather than supporting minority scientists who have already received R01 funding. Instead of focusing on investigator ethnicity, some policymakers might want to allocate resources to directly support ethnic-oriented research. The relative inelasticity of investigator research interests and penalty for investigators pivoting to new topics, devaluing their prior expertise, suggests it may be more effective to recruit and fund scientists interested in these topics rather than attempting to stimulate interest among scientists currently studying other topics (Myers 2020, Hill et al. 2024).

While we focus on ethnicity and research direction, similar dynamics may arise from other aspects of an investigator's background—such as socioeconomic class, challenges in their local community, or diseases experienced personally or among their social network. Our results suggest that scientific topic selection should not be viewed solely as a product of individual genius or calculated assessments of risk and return. The advances that emerge also reflect the broader system through which scientific talent is recruited, developed, and supported.

4. Materials and Methods

Data

We identify all awarded NIH grants from 1980 to 2015 using the NIH's Consolidated Grant Applicant File. We start our analysis in 1980 as we are not able to reliably link publications to the NIH grants which supported them prior to this date, and end in 2015 to allow sufficient time for follow up to observe research outcomes and renewal status of grants awarded in that year. We limit our analysis to R01-equivalent grants.[7] These investigator-initiated grants are a major funding mechanism to support the work of principal investigators who have achieved career independence and run their own laboratory. We also exclude grant awards from special emphasis study sections, which may not have an option for grant renewal, as well as grants without associated publications for which our measures of the ethnic orientation of research would be undefined. We focus our analysis on the grant cycle, which corresponds to a single competitive review of a funding proposal. A grant cycle begins in the year a proposal is funded and ends either at the conclusion of grant funding or when a proposal for grant renewal is competitively reviewed and funded. As grants may have multiple principal investigators and investigator characteristics are explicitly considered during renewal decisions, we analyze our sample at the investigator-grant-cycle level.[8] This gives a final sample of 103,164 investigator-grant-cycles from 63,101 grants and 37,222 investigators. All grants are adjusted to 2020 dollars using the Biomedical Research and Development Price Index.

Publication data is from *PubMed*. We limited our analysis to original research publications. This excludes other types of publications such as review articles, letters, and editorials. We use NIH RePORTER to link publications attributable to each NIH grant. This data relies on the grant having been listed in the publication's acknowledgement section as a funding source, which is required as a condition of accepting NIH funding. We consider publications attributable to a grant cycle if they are published up to one year after the last year of the grant cycle to allow for publication delays.[9] Citation data is from *Web of Science*.

We rely on Author-ity to identify the career publications of each investigator (Torvik and Smalheiser 2021, Torvik and Smalheiser 2009, Torvik et al. 2005). Author-ity disambiguates authors by leveraging the insight that individuals with the same name who share coauthors, affiliations, keywords, and other characteristics are likely to be the same person. Author-ity has been shown to be highly accurate (Lerchenmueller and Sorenson 2016).

We use Medical Subject Headings (MeSH) to help us characterize research topic and methodology. MeSH is a hierarchical controlled vocabulary created and maintained by the National Library of Medicine used for indexing and searching the biomedical literature. In the 2019 edition of MeSH, there are over 29,000 terms, with an average of about 12 MeSH keywords associated with each *PubMed* publication. Crucially, MeSH are assigned by professional indexers without involvement of publication authors.

Measuring ethnicity

Because the ethnicity of NIH-funded investigators is not publicly available, we infer investigator ethnicity from publicly available information. We locate pictures for investigators in our study, using sources such as faculty profile webpage, LinkedIn, obituaries, press releases, Facebook, and other internet resources identified via search. Only investigators for which a picture is identified are included in our primary sample for analysis. Our preferred approach to assigning ethnicity is based on clerical review, integrating across

[7] We consider R01-equivalent grants to be those with an NIH activity code of R01 or R37. We exclude several grant mechanisms considered by the NIH to be R01-equivalent, but which may have distinct dynamics are ethnicity. For instance, we exclude R56 and RL1 which may not be applied for by the investigator, U01 in which NIH staff support the research, and grants such as DP1, DP2 and DP5 which focus on high risk, high reward projects.

[8] 138 (0.13%) grant cycles appear in our sample more than once as they have multiple principal investigators.

[9] Among publications acknowledging a focal NIH grant in our sample, 20.8% are published more than one year after the end of grant funding. These publications and the results they are based on will typically not have been available to study sections when they review a grant for competitive renewal and so do not inform our grant renewal outcome.

profile picture, name, and any ancillary information available. This includes biographical information identified while locating an investigator's photo.[10]

We supplement this approach with robustness checks that use alternative methods for assigning ethnicity, relying solely on either last names or profile pictures.[11] Data for name-based ethnicity assignment stems from the U.S. Census Bureau (Word et al. 2007), the voter rolls from six U.S. Southern states (Rosenmann et al. 2023), and mortgage applications (Tzioumis 2018).[12] Clerical review of ethnicity was also informed by the ethnic distribution of first and middle names based on data from Rosenmann et al. (2023) and Tzioumis (2018). Image analysis was performed using Meta's DeepFace. Following the U.S. Census Bureau, we use ethnic categories of white (including individuals of Middle Eastern and North African descent), Asian or Pacific Islander, Hispanic or Latino, Black or African American, and other ethnicities (including American Indian, mixed-race individuals, or cases that remained unclear after clerical review).

Measuring ethnic-oriented research

Given the lack of clarity around what constitutes "ethnicity-oriented research," we employ five distinct measures, each capturing a different dimension of how research may relate to a particular ethnicity. While each measure has its own strengths and weaknesses, when considered in combination, they provide a more complete understanding of the ethnic orientation of research.

Our first measure focuses on the diseases being studied, using relative mortality to characterize each disease's ethnic orientation. This is calculated using age-adjusted mortality data for the ethnicity of interest relative to white mortality for the same disease with data from the National Center for Health Statistics Compressed Mortality File accessed from the U.S. Centers for Disease Control WONDER (Wide-ranging Online Data for Epidemiologic Research).[13] We used data from 1999-2016, as pre-1999 data did not include information on Hispanic ethnicity, and data from later time periods occurs after our study period and is contaminated by COVID-19. Mortality across 453 disease categories, based on ICD-10 codes, was mapped to MeSH terms to link to investigator publications.[14] Measures of the disease relative mortality examining an investigator's own research use the maximum of the disease relative mortality among grant cycle publications, while measures examining citations to an investigator's publications use the mean of this measure.[15] Note that this measure will not capture disease topics which might be tightly correlated with a particular ethnicity but do not result in death.

Our other measures of ethnicity-oriented research are based on MeSH keywords. These include those based on ethnicity keywords; country or geographic location keywords; and a set of three ethnicity-adjacent topics including keywords related to minorities, healthcare disparities, and socioeconomic status (see Appendix B for details). For each measure, we use an indicator variable set to one if there is at least one publication during the grant cycle that includes a MeSH term of the corresponding type.

---

[10] Information that can be used to infer ethnicity include membership in an ethnic college alumni society, location of educational institutions attended outside of the United States, languages spoken, self-reported parental ethnicity or birthplace, and information on forename and surname frequency by country from Forebears.com

[11] Because the prevalence of Hispanics and Blacks is much lower among NIH-funded investigators than the overall U.S. population, Bayes' theorem suggests this will lead to name-based prediction being less accurate in our population.

[12] Each of these data sources has different names which are subject to censoring to protect privacy, leading to slightly different sample sizes across analysis using solely named-based ethnicity prediction.

[13] We consider Hispanic as only those with white race and Hispanic ethnicity, with measures for white, Black, and Asian are calculated using only non-Hispanic mortality.

[14] Diseases categories were selected so that they were the most specific categorization possible such that mortality data was available and that it corresponded to a disease MeSH.

[15] This measure is defined only if there is at least one publication among grant cycle publications, or one publication among articles citing grant cycle publications, about a disease for which there is associated ethnicity-specific relative mortality data. As the CDC censors mortality data for cells with few individuals, the diseases for which mortality data is available differ slightly across each ethnicity.

Empirical strategy and econometric considerations

We observe the universe of NIH grants awarded, but do not observe grant applications that are not ultimately funded. Because of this, we empirically focus on NIH R01-equivalent grants that were awarded and measure if the grant was renewed. This approach lets us reliably observe all investigators who were at risk of having their grant renewed. A grant may not be renewed either because the investigator applied for but did not receive a competitive renewal, or because they never applied for a renewal. Given the career importance of grant funding in maintaining a lab, we expect most investigators will apply for renewal.[16]

The primary threat to unbiased estimation of the impact of ethnicity on the direction of research or grant renewal with our research design is omitted variable bias: that there are some unmeasured factors correlated with both ethnicity and the outcome of interest. To address this concern, we incorporate a rich set of controls for investigator, institution, and grant cycle characteristics. Given the limitations of this research design, we see our results as a correlation and do not claim to estimate a causal effect.

Our primary estimating equation relates the dependent variables of interest, $ethnicresearch_{ij}$ or $renewal_{ij}$, of investigator i and grant cycle j. Formally, we estimate variations of

$$E\left[renewal_{ij} = 1 \mid \delta_i, \gamma_j\right] = \beta_0 + \beta_1 ethnicity_i + \delta_i + \gamma_j$$

where $renewal_{ij}$ is an indicator variable equal to 1 if the grant cycle was renewed and 0 otherwise. We include high dimensional vectors of controls for investigator and grant cycle characteristics, δ_i and γ_j, respectively. This includes sets of indicator variables for investigator age, degree, and gender; log investigator career citations, log prior investigator NIH funding, an indicator variable for no prior investigator NIH funding, and log institutional NIH total funding as measures of investigator and institutional quality. Additionally, we include sets of indicator variables for year of evaluation, grant length, and the interaction between the NIH institutes and centers and study section. Select specifications also include controls based on measures of research quality and quantity during the grant cycle. This includes log of mean journal impact factor; sets of indicator variables for the number of publications during the grant cycle separately for publications acknowledging and not acknowledging funding from the grant; indicator variables for the research methodology employed including bench science, translational research, clinical studies, and clinical trials; and indicator variables for having publications in highly selective journals.[17] We use robust standard errors with two-way clustering at the investigator and grant levels throughout. The maintained assumption is that, after conditioning on observables, measures of ethnicity are uncorrelated with the error term.

Our primary specification is a linear probability model (LPM). Since LPM may generate fitted values outside of the interval [0,1], its use would not be appropriate for predictive analysis. However, because our objective is the estimation of the marginal effects from the conditional expectation function, LPM is appropriate (Angrist and Pischke 2008: pp. 94-106).[18]

[16] Applications for competitive grant renewal, using the NIH's more expansive definition of R01-equivalency (see footnote 6) were funded at a rate of 34.5% in 2015 (Lauer 2016), close to the 35.9% renewal rate in our sample in the same year, and consistent with the assumption that most grantees do apply for renewal.

[17] For this measure we include those publications which were published in *Nature*, *Cell*, *Science*, and the *New England Journal of Medicine*.

[18] While logit and probit models have the benefit of bounding estimates of the dependent variable between 0 and 1, the resulting estimates may be biased by the incidental parameter problem. This may be particularly problematic when including many fixed effects, as is necessary in our case.

References

Abramitzky, Ran, Lena Greska, Santiago Perez, Joseph Price, Carlo Schwarz, and Fabian Waldinger. 2024. "Climbing the Ivory Tower: How Socio-Economic Background Shapes Academia." CESifo Working Paper Series No. 11577.

Airoldi, Anna, and Petra Moser. 2024. "Inequality in Science: Who Becomes a Star?" National Bureau of Economic Research Working Paper 33063.

Angrist, Joshua D., and Jorn-Steffen Pischke. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press.

Azoulay, Pierre, Joshua S. Graff Zivin, Danielle Li, and Bhaven Sampat. 2019. "Public R&D and Private-Sector Patenting: Evidence from NIH Funding Rules." *The Review of Economic Studies,* **86**(1):117-152.

Azoulay, Pierre, and Wesley H. Greenblatt. 2025. "Does Peer Review Penalize Scientific Risk Taking? Evidence from NIH Grant Renewals." National Bureau of Economic Research Working Paper 33495.

Bell, Alexander M., Raj Chetty, Xavier Jaravel, Neviana Petkova, and John Van Reenen. 2019. "Who Becomes an Inventor in America? The Importance of Exposure to Innovation." *Quarterly Journal of Economics* **134**(2):647-713.

Belloni, Alexandre, Victor Chernozhukov, Christian Hansen. 2014. "Inference on treatment effects after selection among high-dimensional controls." *Review of Economics Studies*, **81**(2):608-650.

Borrell, Luisa N., Jennifer R. Elhawary, Elena Fuentes-Afflick, Jonathan Witonsky, Nirav Bhakta, Alan H.B. Wu, Kirsten Bibbons-Domingo, Jose R. Rodriguez-Santana, Michael A. Lenoir, James R. Gavin III, Rick A. Kittles, Noah A. Zaitlen, David S. Wilkes, Neil R. Powe, Elad Ziv, and Esteban G. Burchard. 2021. "Race and Genetic Ancestory in Medicine—A Time for Reckoning with Racism." *New England Journal of Medicine*, **384**(5): 474-480.

Budish, Eric, Benjamin N. Roin, and Heidi Williams. 2015. "Do Firms Underinvest in Long-Term Research? Evidence from Cancer Clinical Trials." *American Economic Review*, **105**(7):2044-2085.

Chaturvedi, Nish. 2001. "Ethnicity as an epidemiological determinant—crudely racist or crucially important?" *Internal Journal of Epidemiology*, **30**(5):925-927.

Cook, Lisa D., Matt Marx, and Emmanuel Yimfor. 2022. "Funding Black High-Growth Startups." National Bureau of Economic Research Working Paper 30682.

Dance, Amber. 2012. "Who's on First?" *Nature*, 489(7417):591-593.

Einio, Elias, Josh Feng, and Xavier Jaravel. 2023. "Social Push and the Direction of Innovation." SSRN Working Paper 3383703.

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

Fleming, Lee, Hillary Greene, Guan-Cheng Li, Matt Marx, and Dennis Yao. 2019. "Government-funded research increasingly fuels innovation." *Science*, **364**(6446):1139-1141.

Furman, Jeffrey L., Scott Stern. 2011. "Climbing atop the Shoulders of Giants: The Impact of Institutions on Cumulative Research." *American Economic Review*, **101**(5):1933-1963.

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.

Goldstein, Joseph L., and Michael S. Brown. 1997. "The Clinical Investigator: Bewitched, Bothered, and Bewildered—But Still Beloved." *Journal of Clinical Investigation* **99**(12): 2803-2812.

Gupta, Harsh. 2022. "Do Female Researchers Increase Female Enrollment in Clinical Trials?" Working paper.

Hermosilla, Manuel. 2023. "Are Female-led Trials Safer?" SSRN Working Paper 4638767.

Hill, Ryan, Yian Yin, Carolyn Stein, Xizhao Wang, Dashun Wang, and Benjamin F. Jones. 2024. "Adaptability and the Pivot Penalty in Science and Technology." Working paper arXiv:2107.06476.

Hoppe, Travis A., Avia 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, George M Santangelo. 2019. "Topic choice contributes to the lower rate of NIH awards to African-American/black scientists." *Science Advances*, **5**(10):eaaw7238.

Jackson, Joanna R., Ann M. Holmes, Elizabeth Golembiewski, Brittany L. Brown-Podgorski, and Nir Menachemi. 2019. "Graduation and Academic Placement of Underrepresented Racial/Ethnic Minority Doctoral Recipients in Public Health Disciplines, Unites States, 2003-2015.

Jain, Mukesh K., Tadataka Yamada, and Robert Lefkowitz. 2019. "We Need More Doctors Who Are Scientists." *New York Times*. https://www.nytimes.com/2019/09/23/opinion/doctor-scientist-medical-research.html

Koffi, Marlene, Roland Pongou, and Leonard Wantchekon. 2024. "The Color of Ideas: Racial Dynamics and Citations in Economics." National Bureau of Economic Research Working Paper 33150.

Koning, Rembrand, Sampsa Samila, and John-Paul Ferguson. 2020. "Inventor Gender and the Direction of Invention." *AEA Papers and Proceedings*, **110**:250-254.

Koning, Rembrand, Sampsa Samila, and John-Paul Ferguson. 2021. "Who do we invent for? Patents by women focus more on women's health, but few women get to invent." *Science*, **372**(6548):1345-1348.

Lauer, Mike and Marie A. Bernard. 2022. "Research Project Grant Funding Rates and Principal Investigator Race and Ethnicity." National Institues of Health Office of Extramural Research. https://nexus.od.nih.gov/all/2022/06/14/research-project-grant-funding-rates-and-principal-investigator-race-and-ethnicity/.

Lauer, Michael S., Jamie Doyle, Joy Wang, and Deepshikha Roychowdhury. 2021. "Associations of topic-specific peer review outcomes and institute and center award rates with funding disparities at the National Institutes of Health." *eLife*, **10**:e67173.

Lauer, Mike. 2016. "Are Attempts at Renewal Successful?." National Institutes of Health Office of Extramural Research.

LaVeist, Thomas A., Amani Nuru-Jeter, and Keisha E. Jones. 2003. "The Association of Doctor-Patient Race Concordance with Health Services Utilization." *Journal of Public Health Policy,* **24**(3/4):312-323.

Lercehmueller, Marc J. and Olav Sorenson. 2016. "Author Disambiguation in PubMed: Evidence on the Precision and Recall of Author-ity among NIH-Funded Scientists." *PLoS One,* 11(7):e0158731.

Lerner, Josh and Julie Wulf. 2007. "Innovation and Incentives: Evidence from Corporate R&D." *Review of Economics and Statistics*, **89**(4):634-644.

Leszczensky, Lars and Sebastian Pink. 2019. "What Drived Ethnic Homophily? A Relational Approach on How Ethnic Identification Moderates Preferences for Same-Ethnic Friends." *American Sociological Review*, **84**(3):394-419.

Liu, Fengyun, Talal Rahwan, and Bedoor Alshebli. 2023. "Non-White scientists appear on fewer editorial boards, spend more time under review, and receive fewer citations." *Proceedings of the National Academy of Sciences,* **120**(13):e2215324120.

Low, Daniel, Samantha W. Pollack, Zachary C. Liao, Ramoncita Maestas, Larry E. Kirven, Anne M. Eacker, and Leo S. Morales. 2019. "Racial/Ethnic Disparities in Clinical Grading in Medical School." *Teaching and Learning in Medicine,* **31**(5):487-496.

Marrast, Lyndonna M., Leah Zallman, Steffie Woolhandler, David H. Bor, and Danny McCormick. 2014. "Minority Physicians' Role in the Care of Underserved Patients." *JAMA Internal Medicine*, **174**(2):289-291.

Marx, Matt. 2024. "Can we win the 'names game'? Tactics for large-scale demographic research in entrepreneurship and innovation." *Industry and Innovation*, **31**(3):271-279.

Mateos, Pablo. 2007. "A Review of Name-based Ethnicity Classification Methods and their Potential in Population Studies." *Population Space and Place*, **13**(4):243-263.

Moses 3rd, Hamilton, David H. M. Matheson, Sarah Cairns-Smith, Benjamin P. George, Chase Palisch, and E. Ray Dorsey. 2015. "The anatomy of medical research: US and international comparisons." *Journal of the American Medical Association*, **313**(2):174-89.

Myers, Kyle. 2020. "The Elasticity of Science." *American Economic Journal: Applied Economics*, **12**(4): 103-134.

National Research Council Committee on National Needs for Biomedical and Behavioral Scientists. 2000. "Developing a Research Workforce that Reflects the Nation and Addresses its Health Needs." In *Addressing the Nation's Changing Needs for Biomedical and Behavioral Scientists*. Washington, D.C.: National Academy Press.

Nielsen, Mathias W., Jens P. Andersen, Londa Schiebinger, and Jesper W. Schneider. 2017. "One and a half million medical papers reveal a link between author gender and attention to gender and sex analysis." *Nature Human Behavior*, **1**(11):791-796.

Oakes, Jeannie. 1990. *Lost Talent: The Underparticipation of Women, Minorities, and Disabled Persons in Science.* Santa Monica, CA: Rand Corporation.

Ong, Matthew B.H. 2021. "Single digits: Black, Hispanic scientists strikingly underrepresented at NCI among senior workforce and grantees." *Cancer Letter*, **47**(26). https://cancerletter.com/health-equity/20210702_1/.

Oster, Emily. 2019. "Unobservable Selection and Coefficient Stability: Theory and Evidence." *Journal of Business and Economic Statistics,* **37**(2):187-204.

Rosenman, Evan T.R., Santiago Olivella, and Kosuke Imai. 2023. "Race and ethnicity data for first, middle and surnames." *Scientific Data*, **10**:299.

Sims, Calvin. 1992. "What Went Wrong: Why Programs Failed." *Science*, **258**(5085):1185-1187.

Takeshita, Junko, Shiyu Wang, Alison W. Loren, Nandita Mitra, Justine Shults, Daniel B. Shin, Deirdre L. Sawinski. 2020. "Association of Racial/Ethnic and Gender Concordance Between Patients and Physicians With Patient Experience Ratings." *JAMA Network Open*, **3**(11):e2024583.

Tilghman, Shirley, Bruce Alberts, Daniel Colon-Ramos, Kafui Dzirasa, Judith Kimble and Harold Varmus. 2021. "Concrete steps to diversity the scientific workforce." *Science*, **372**(6538):133-135.

Torvik, Vetle I. and Neil R. Smalheiser. 2009. "Author name disambiguation in MEDLINE." *ACM Transactions on Knowledge Discovery from Data*, 3(3):11.

Torvik, Vetle I. and Neil R. Smalheiser. 2021. Author-ity 2018 - PubMed author name disambiguated dataset. University of Illinois at Urbana-Champaign.

Torvik, Vetle I, Marc Weeber, Don R. Swanson, and Neil R. Smalheiser. 2005. "A Probabilistic Similarity Metric for Medline Records: A Model for Author Name Disambiguation." *Journal of the American Society for Information Science and Technology*, 56(2):140-158.

Tzioumis, Konstantinos. 2018. "Data Descriptor: Demographic aspects of first names." *Scientific Data*, 5:180025.

Word, David L., Charles D. Coleman, Robert Nunziata, and Robert Kominski. 2007. "Demographic Aspects of Surnames from Census 2000." U.S. Census Bureau. https://www2.census.gov/topics/genealogy/2010surnames/surnames.pdf

Wu, Lingfei, Dashun Wang, and James A. Evans. 2019. "Large teams develop and small teams disrupt science and technology." *Nature*, **566**(7744):378-382.

Yehia, Baligh R., Peter F. Cronholm, Nicholas Wilson, Steven C. Palmer, Stephen D. Sisson, Conair E. Guilliames, Norma I. Poll-Hunter, and John-Paul Sanchez. 2014. "Mentorship and pursuit of academic medical careers: a mixed methods study of residents from diverse backgrounds." *BMC Medical Education*, **14**:26.

Table 1. Effect of ethnicity on research topic

A. Disease relative mortality

	Disease relative mortality		
	Asian	Hispanic	Black
Asian ethnicity	0.1281**	-0.0055	0.1994
	(0.0287)	(0.1158)	(0.4512)
Hispanic ethnicity	0.0597	1.2834**	1.1612
	(0.0587)	(0.4330)	(0.8670)
Black ethnicity	0.1476	0.7676	12.9128*
	(0.1055)	(0.4747)	(5.3748)
Mean of Dependent Variable	1.316	2.223	6.239
Adjusted R2	0.1710	0.0922	0.0607
Nb. of Investigators	25,631	25,849	25,916
Nb. of Grant cycles	64,577	65,441	65,701

B. Ethnicity

	Ethnicity MeSH topics			
	White	Asian	Hispanic	Black
White ethnicity	-0.0002			
	(0.0024)			
Asian ethnicity		0.0186**	-0.0018	-0.0047†
		(0.0024)	(0.0017)	(0.0024)
Hispanic ethnicity		0.0014	0.0269**	0.0059
		(0.0036)	(0.0052)	(0.0060)
Black ethnicity		0.0075	0.0273*	0.0957**
		(0.0104)	(0.0124)	(0.0180)
Mean of Dependent Variable	0.053	0.020	0.026	0.054
Adjusted R2	0.2675	0.1210	0.2202	0.2814
Nb. of Investigators	31,777	31,777	31,777	31,777
Nb. of Grant cycles	91,467	91,467	91,467	91,467

C. Country

	Country MeSH topics			
	White	Asian	Hispanic	Black
White ethnicity	0.0205**			
	(0.0036)			
Asian ethnicity		0.0424**	-0.0051**	-0.0033*
		(0.0034)	(0.0017)	(0.0016)
Hispanic ethnicity		0.0005	0.0548**	0.0007
		(0.0046)	(0.0075)	(0.0035)
Black ethnicity		-0.0046	0.0175	0.0473**
		(0.0115)	(0.0125)	(0.0147)
Mean of Dependent Variable	0.180	0.035	0.027	0.020
Adjusted R2	0.4135	0.1170	0.1237	0.1457
Nb. of Investigators	31,777	31,777	31,777	31,777
Nb. of Grant cycles	91,467	91,467	91,467	91,467

D. Ethnicity-adjacent topics

	Ethnicity-adjacent MeSH topics		
	Minorities	Disparities	Socioeconomic status
Asian ethnicity	-0.0008	-0.0003	-0.0052*
	(0.0007)	(0.0007)	(0.0020)
Hispanic ethnicity	0.0097**	0.0099**	0.0105*
	(0.0028)	(0.0034)	(0.0051)
Black ethnicity	0.0290*	0.0218*	0.0419**
	(0.0116)	(0.0087)	(0.0150)
Mean of Dependent Variable	0.007	0.006	0.067
Adjusted R2	0.1121	0.1598	0.4552
Nb. of Investigators	31,777	31,777	31,777
Nb. of Grant cycles	91,467	91,467	91,467

Note: Estimates stem from an OLS regression (panel A) or linear probability model (panels B, C, and D) where the unit of analysis is the investigator-grant-cycle. The independent variables are indicator variables for white, Asian, Hispanic and Black ethnicity. If white ethnicity is the independent variable of interest, all other ethnicities are the omitted category, while if Asian, Hispanic and Black ethnicities are the independent variables of interest, white ethnicity and other ethnicity are the omitted category. The outcome variable of interest is the maximum disease relative mortality calculated relative to white mortality (panel A), any ethnicity topic (panel B), any country topic (panel C), and any ethnicity adjacent topics (panel D) among grant cycle publications. All estimates include controls for grant year, gender, investigator degree, investigator career age, length of the grant cycle, a set of 2,911 indicator variables corresponding to each NIH institute or center and study section, log NIH funding to the investigator's institution, log investigator prior career citations, log investigator prior NIH grant funding, no prior NIH grant funding, a set of four indicator variable for research methodology used, and number of publications during the grant cycle. Robust standard errors were used, clustered at the investigator and grant. $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$.

Table 2. Effect of ethnicity on the research direction of forward citations

A. Disease relative mortality

	Disease relative mortality		
	Asian	Hispanic	Black
Asian ethnicity	0.0290**	-0.0083	0.0194
	(0.0107)	(0.0473)	(0.1239)
Hispanic ethnicity	0.0225	0.3112*	0.1459
	(0.0251)	(0.1382)	(0.3188)
Black ethnicity	0.0257	0.7890	4.6202*
	(0.0364)	(0.5122)	(2.3038)
Mean of Dependent Variable	0.845	1.491	2.798
Adjusted R2	0.2778	0.2153	0.0978
Nb. of Investigators	31,314	31,345	31,355
Nb. of Grant cycles	89,699	89,791	89,827

B. Ethnicity

		Ethnicity MeSH topics		
	White	Asian	Hispanic	Black
White ethnicity	-0.0247			
	(0.0559)			
Asian ethnicity		0.2684**	-0.3210**	-0.2608**
		(0.0580)	(0.0856)	(0.0693)
Hispanic ethnicity		0.0907	1.0587**	0.2923*
		(0.0816)	(0.1317)	(0.1477)
Black ethnicity		0.5321**	0.4803**	1.0530**
		(0.1888)	(0.1516)	(0.1594)
Pseudo-R2	0.7018	0.6733	0.7037	0.7085
Log-likelihood	-286,455	-244,359	-157,694	-313,221
Nb. of Investigators	31,968	31,994	31,861	31,959
Nb. of Grant cycles	91,928	92,007	91,534	91,905

C. Country

		Country MeSH topics		
	White	Asian	Hispanic	Black
White ethnicity	0.1205**			
	(0.0347)			
Asian ethnicity		0.3890**	0.0178	-0.0606
		(0.0739)	(0.1063)	(0.1118)
Hispanic ethnicity		-0.0922	0.3796**	-0.2956**
		(0.0603)	(0.0938)	(0.1123)
Black ethnicity		0.3123	0.2308	0.3747*
		(0.1975)	(0.1473)	(0.1796)
Pseudo-R2	0.8502	0.7422	0.6748	0.6469
Log-likelihood	-1,051,333	-432,531	-293,229	-296,012
Nb. of Investigators	32,001	31,997	31,993	31,989
Nb. of Grant cycles	92,045	92,026	92,004	91,984

D. Ethnicity-adjacent topics

	Ethnicity-adjacent MeSH topics		
	Minorities	Disparities	Socioeconomic status
Asian ethnicity	-0.3269**	-0.3034**	-0.2099**
	(0.0807)	(0.0799)	(0.0719)
Hispanic ethnicity	0.4969**	0.6967*	0.3616
	(0.1858)	(0.2787)	(0.2321)
Black ethnicity	1.0651**	0.8409**	0.3659*
	(0.2323)	(0.1970)	(0.1594)
Pseudo-R2	0.6574	0.7019	0.8293
Log-likelihood	-37,141	-55,971	-368,103
Nb. of Investigators	30,830	30,995	31,987
Nb. of Grant cycles	86,604	87,982	91,953

Note: Estimates stem from an OLS regression (panel A) or Poisson regression (panels B, C, and D) where the unit of analysis is the investigator-grant-cycle. The independent variables are indicator variables for white, Asian, Hispanic and Black ethnicity. If white ethnicity is the independent variable of interest, all other ethnicities are the omitted category, while if Asian, Hispanic and Black ethnicity are the independent variables of interest, white ethnicity and other ethnicity are the omitted category. The outcome variable of interest is the ethnicity-specific mean disease relative mortality relative to white mortality (panel A), and the count of papers with the corresponding ethnicity topic (panel B), country topic (panel C), and ethnicity-adjacent topics (panel D), calculated using citations to an investigator's grant cycle publications. Panels B, C, and D include an offset for the total number of citations. All estimates include controls for grant year, gender, investigator degree, investigator career age, length of the grant cycle, a set of 2,911 indicator variables corresponding to each NIH institute or center and study section, log NIH funding to the investigator's institution, log investigator prior career citations, log investigator prior NIH grant funding, no prior NIH grant funding, a set of indicator variables for research methodology, and number of publications during the grant cycle. Robust standard errors were used, clustered at the investigator and grant. $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$.

Table 3. Effect of ethnicity on grant renewal

	(1)	(2)	(3)	(4)
Asian ethnicity	-0.017**	-0.002	0.005	0.003
	(0.006)	(0.005)	(0.005)	(0.004)
Hispanic ethnicity	-0.001	0.008	0.016†	0.015
	(0.013)	(0.010)	(0.009)	(0.009)
Black ethnicity	-0.102**	-0.049*	-0.010	-0.009
	(0.025)	(0.021)	(0.019)	(0.019)
Grant year	Yes	Yes	Yes	Yes
Investigator demographics	No	Yes	Yes	Yes
IC x study section	No	Yes	Yes	Yes
Institutional quality	No	No	Yes	Yes
Investigator quality	No	No	Yes	Yes
Pub. quantity, quality, methodology during cycle	No	No	No	Yes
Mean of dependent variable	0.463	0.465	0.465	0.465
Std. Dev. of dependent variable	0.499	0.499	0.499	0.499
Effect Asian ethnicity, in s.d. units	-0.035	-0.004	0.010	0.005
Effect Hispanic ethnicity, in s.d. units	-0.002	0.017	0.033	0.029
Effect Black ethnicity, in s.d. units	-0.204	-0.099	-0.019	-0.017
Adjusted R^2	0.0344	0.2403	0.3105	0.3221
Nb. of investigators	32,012	31,793	31,790	31,790
Nb. of investigator-grant-cycles	92,126	91,554	91,505	91,505

Note: Estimates stem from a linear probability model where the unit of analysis is the investigator-grant-cycle and the dependent variable is an indicator variable for if the grant cycle was renewed. The independent variables are indicator variables for Asian, Hispanic and Black ethnicity, with white ethnicity and other ethnicity as the omitted category. Column 1 includes only an indicator variable for grant year, while column 2 adds indicator variables for gender, investigator degree, investigator career age, length of the grant cycle, and a set of 2,911 indicator variables corresponding to each NIH institute or center and study section. Column 3 includes all of the previous controls, but also includes log NIH funding to the investigator's institution, log investigator prior career citations, log investigator prior NIH grant funding, and an indicator for no prior NIH grant funding. Finally, column 4 includes all previous controls, but also add log mean journal impact factor, indictor variables for grant and non-grant publications in one of four highly selective journals, a set of 4 indictors variables for the research methodology used, and a set of 21 indicator variables separately for the number of publications during the grant cycle acknowledging and not acknowledging the grant. Robust standard errors were used, clustering at the investigator and grant. $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$.

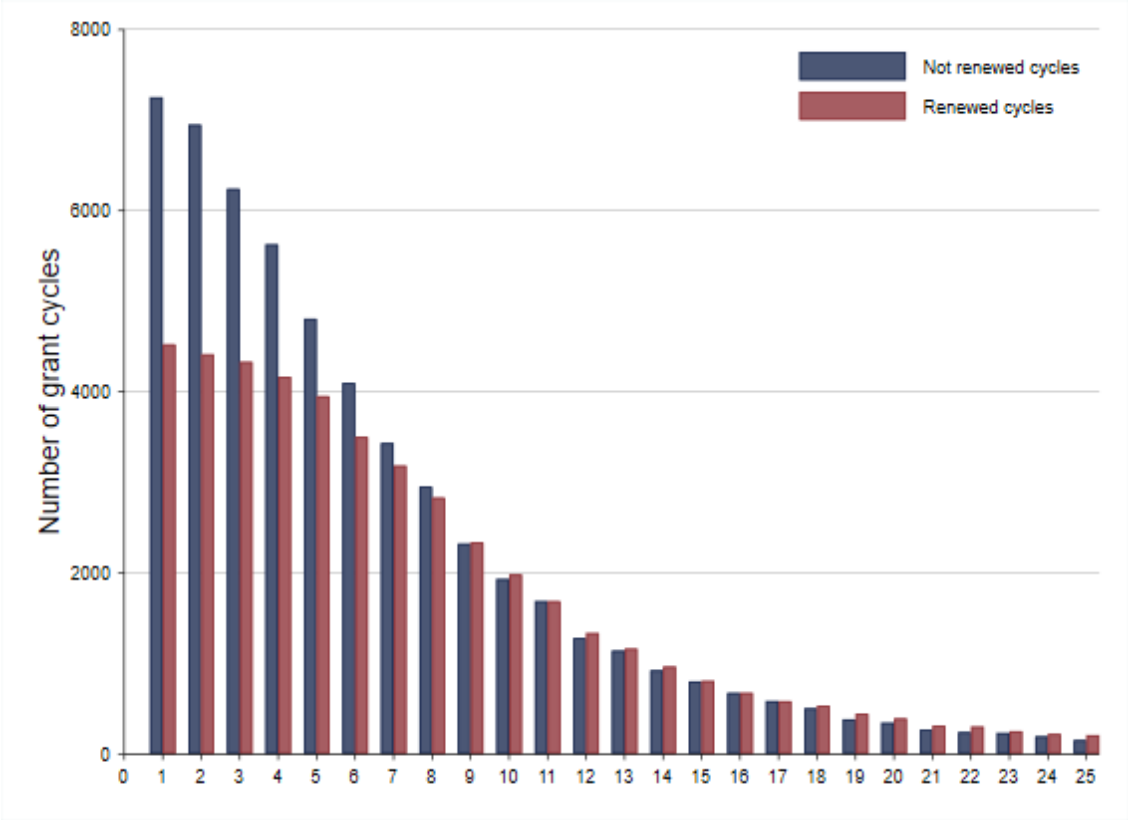
Table 4. Effect of ethnic research topic on grant renewal

	Year controls			Full controls		
	(1)	(2)	(3)	(4)	(5)	(6)
Asian relative disease mortality	-0.003			-0.003*		
	(0.013)			(0.001)		
Hispanic relative disease mortality	0.004			-0.000		
	(0.003)			(0.000)		
Black relative disease mortality	-0.002**			0.000		
	(0.001)			(0.000)		
Asian ethnicity topic		-0.019			-0.013	
		(0.013)			(0.011)	
Hispanic ethnicity topic		-0.067**			-0.008	
		(0.012)			(0.010)	
Black ethnicity topic		-0.101**			-0.008	
		(0.009)			(0.008)	
Asian country topic			-0.063**			-0.004
			(0.010)			(0.008)
Hispanic country topic			-0.074**			-0.008
			(0.011)			(0.009)
Black country topic			-0.040**			-0.012
			(0.013)			(0.011)
Mean of dependent variable	0.430	0.463	0.463	0.432	0.466	0.466
Adjusted R^2	0.0325	0.0378	0.0357	0.302	0.322	0.322
Nb. of investigators	25,831	31,998	31,998	25,572	31,776	31,776
Nb. of investigator-grant-cycles	65,003	92,087	92,087	64,373	91,465	91,465

Note: Estimates stem from a linear probability model where the unit of analysis is the investigator-grant-cycle and the dependent variable is an indicator variable for if the grant cycle was renewed. The independent variables are continuous variables for Asian, Hispanic and Black relative disease mortality; indicator variables for Asian, Hispanic, and Black ethnicity MeSH topics; and indicator variables for Asian Hispanic, and Black country MeSH topics all based on grant cycle publications. Columns 1-3 include only a set of indicator variables for grant year as controls, while columns 4-6 includes identical controls as Table 3, column 4. Robust standard errors were used, clustering at the investigator and grant. $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$.

Appendix A
Ancillary Results and Robustness Checks

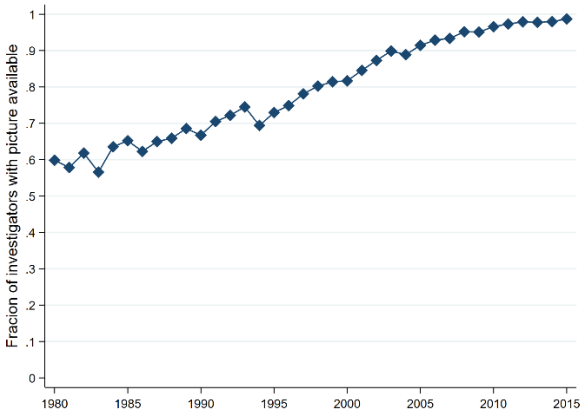
Figure A1. Distribution of grant cycle publications by renewal status



Note: Histogram of the number of publications acknowledging the grant within the grant cycle. Unit of analysis is at the investigator-grant-cycle level. Two thousand nine hundred fifteen (3.2% of the sample) outliers with over 25 grant cycle publications are not shown.

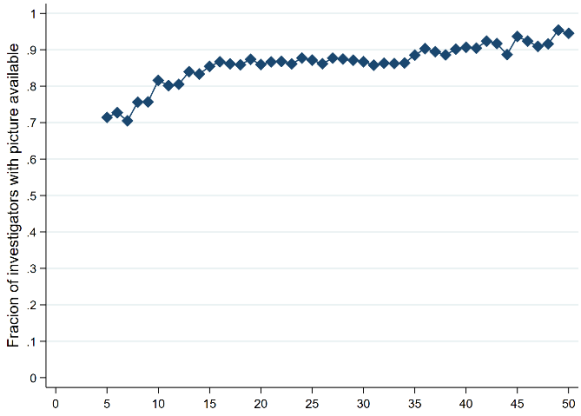
Figure A2. Investigators with missing picture

A. Fraction of investigators by most recent grant cycle year



Note: Plot of the fraction of investigators with a photo identified online, by the year the investigator's most recent grant cycle began. Unit of analysis is at the investigator level.

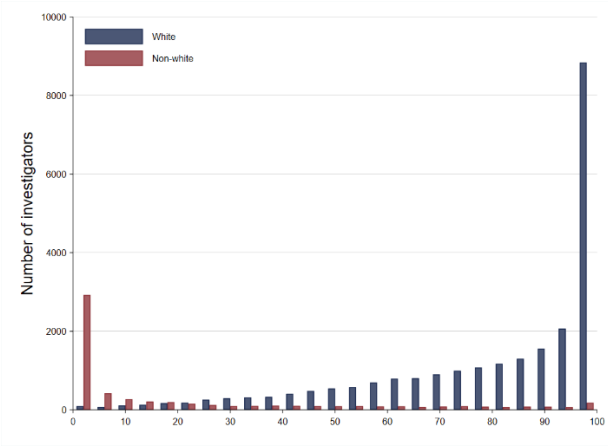
B. Fraction of investigators by career age



Note: Plot of the fraction of investigators with a photo identified online, by career age. Career age is defined as the number of years from terminal degree to the start of the grant cycle, here calculated using an investigator's most recent grant cycle. Ninety-nine investigators with career age under 5 years and 259 investigators with career age over 50 are omitted. Unit of analysis is at the investigator level.

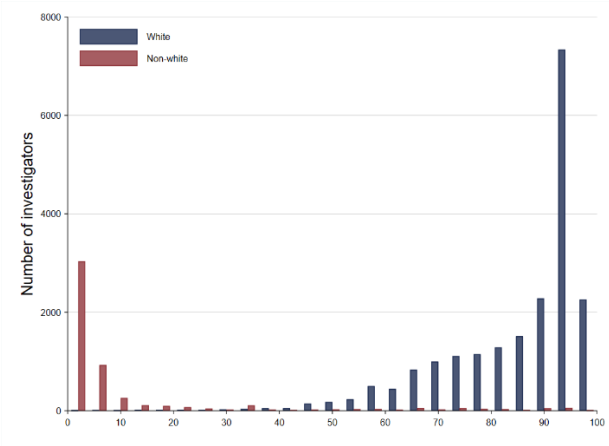
Figure A3. Distribution of measures of ethnicity: White

A. Deepface score



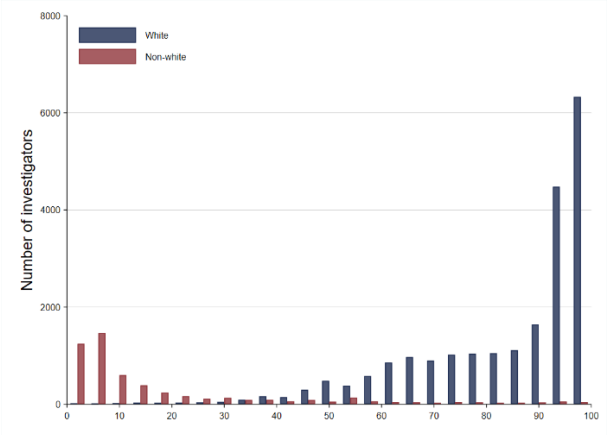
Note: Histogram of the percent likelihood of white non-Hispanic ethnicity based on analysis of investigator picture using Deepface, by assigned ethnicity on manual review. Unit of analysis is at the investigator level.

B. Census last name score



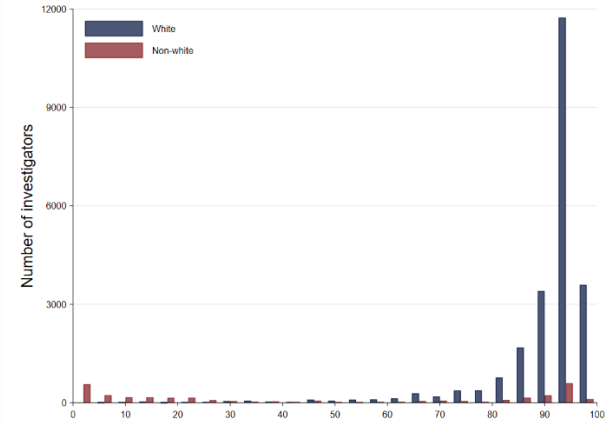
Note: Histogram of the percent likelihood of white non-Hispanic ethnicity based on last name using data from the U.S. Census Bureau, by assigned ethnicity on manual review. Unit of analysis is at the investigator level.

C. Rosenman et al. (2023) last name score



Note: Histogram of the percent likelihood of white non-Hispanic ethnicity based on last name using data from Rosenman et al. (2023), by assigned ethnicity on manual review. Unit of analysis is at the investigator level.

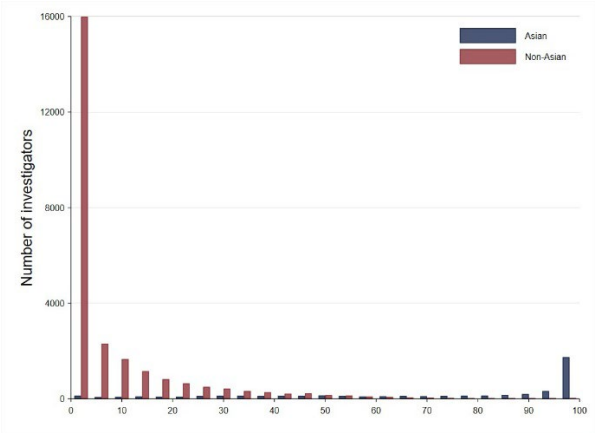
D. Tzioumis (2018) first name score



Note: Histogram of the percent likelihood of white non-Hispanic ethnicity based on first name using data from Tzioumis (2018), by assigned ethnicity on manual review. Unit of analysis is at the investigator level.

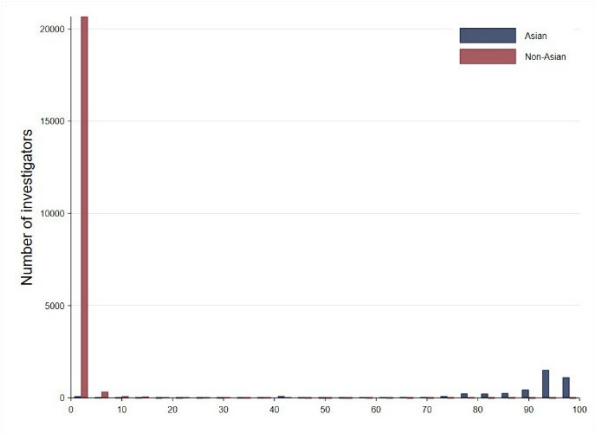
Figure A4. Distribution of measures of ethnicity: Asian

A. Deepface score



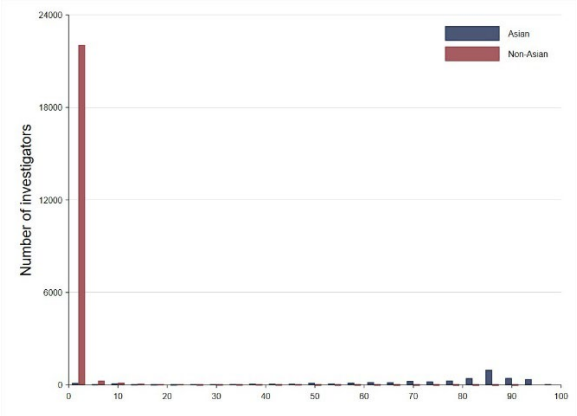
Note: Histogram of the percent likelihood of Asian ethnicity based on analysis of investigator picture using Deepface, by assigned ethnicity on manual review. Unit of analysis is at the investigator level.

B. Census last name score



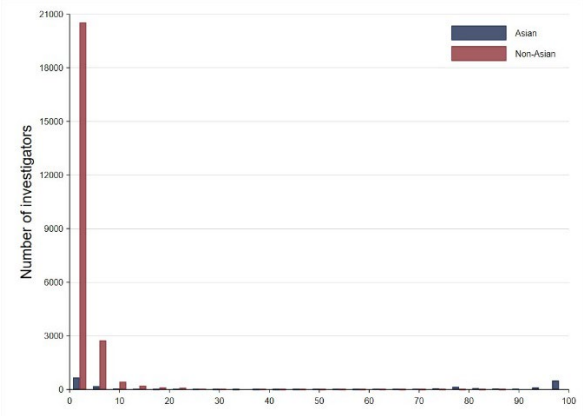
Note: Histogram of the percent likelihood of Asian ethnicity based on last name using data from the U.S. Census Bureau, by assigned ethnicity on manual review. Unit of analysis is at the investigator level.

C. Rosenman et al. (2023) last name score



Note: Histogram of the percent likelihood of Asian ethnicity based on last name using data from Rosenman et al. (2023), by assigned ethnicity on manual review. Unit of analysis is at the investigator level.

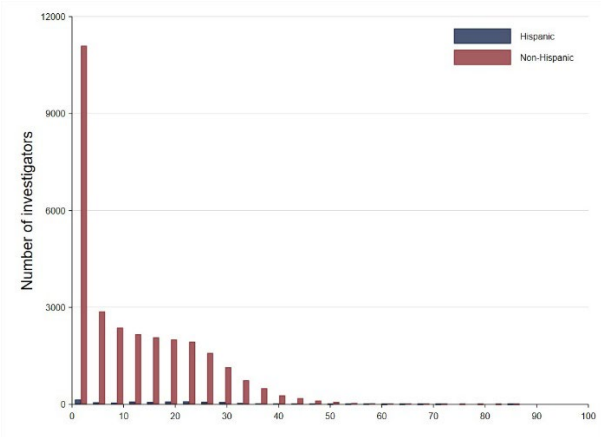
D. Tzioumis (2018) first name score



Note: Histogram of the percent likelihood of Asian ethnicity based on first name using data from Tzioumis (2018), by assigned ethnicity on manual review. Unit of analysis is at the investigator level.

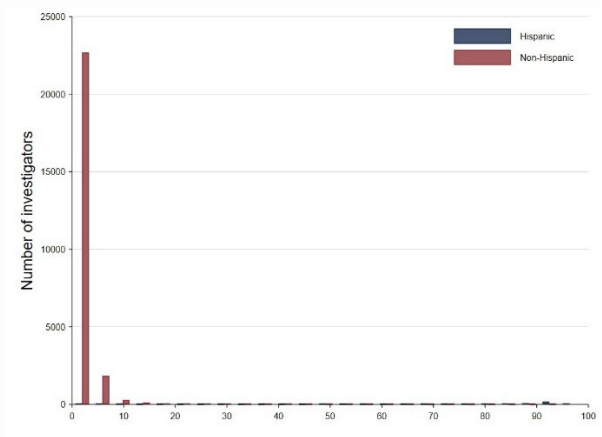
Figure A5. Distribution of measures of ethnicity: Hispanic

A. Deepface score



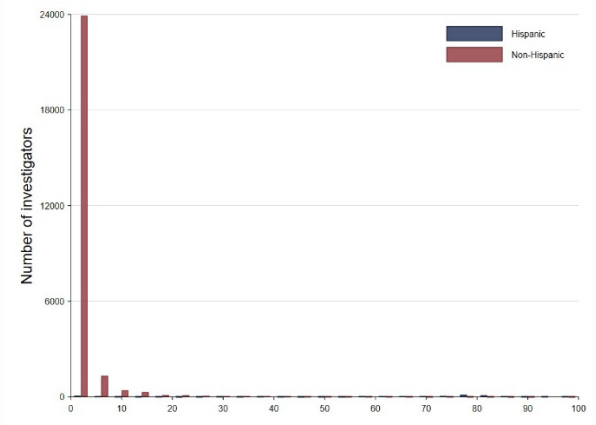
Note: Histogram of the percent likelihood of Hispanic ethnicity based on analysis of investigator picture using Deepface, by assigned ethnicity on manual review. Unit of analysis is at the investigator level.

B. Census last name score



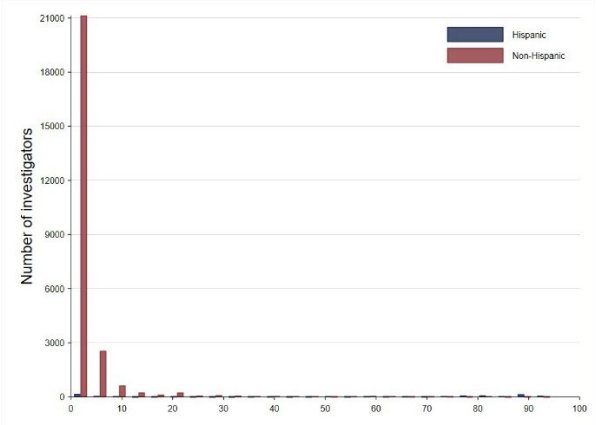
Note: Histogram of the percent likelihood of Hispanic ethnicity based on last name using data from the U.S. Census Bureau, by assigned ethnicity on manual review. Unit of analysis is at the investigator level.

C. Rosenman et al. (2023) last name score



Note: Histogram of the percent likelihood of Hispanic ethnicity based on last name using data from Rosenman et al. (2023), by assigned ethnicity on manual review. Unit of analysis is at the investigator level.

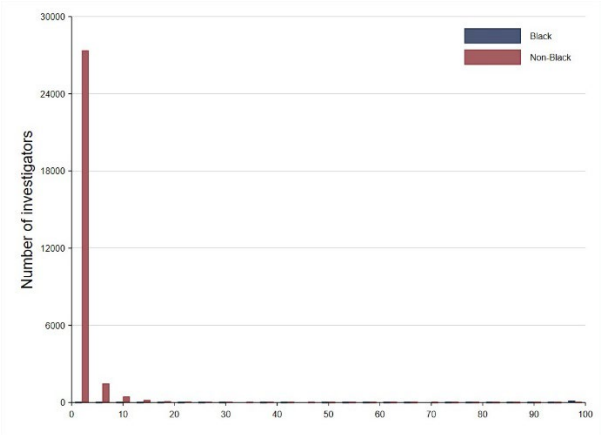
D. Tzioumis (2018) first name score



Note: Histogram of the percent likelihood of Hispanic ethnicity based on first name using data from Tzioumis (2018), by assigned ethnicity on manual review. Unit of analysis is at the investigator level.

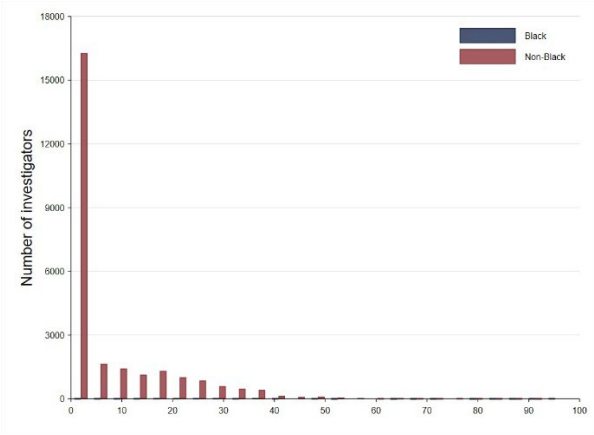
Figure A6. Distribution of measures of ethnicity: Black

A. Deepface score



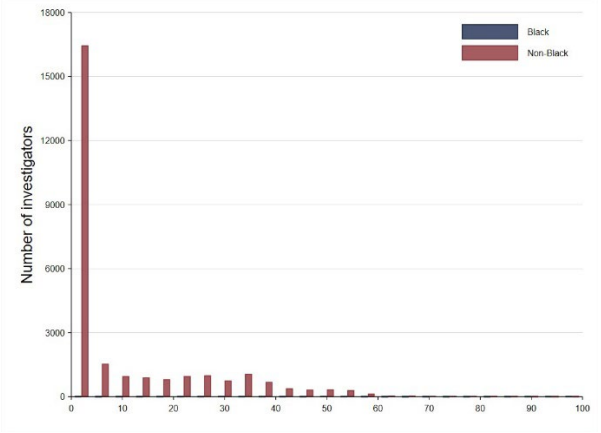
Note: Histogram of the percent likelihood of Black non-Hispanic ethnicity based on analysis of investigator picture using Deepface, by assigned ethnicity on manual review. Unit of analysis is at the investigator level.

B. Census last name score



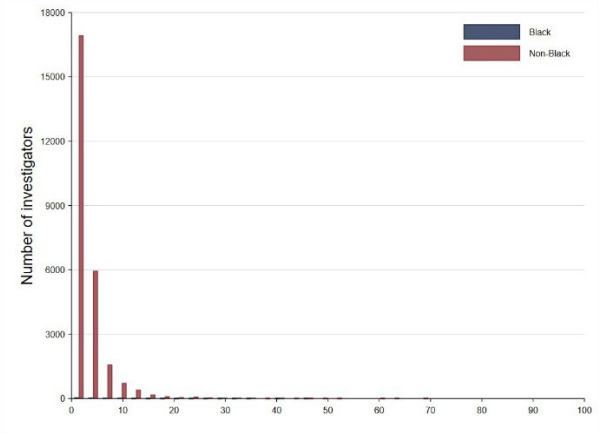
Note: Histogram of the percent likelihood of Black non-Hispanic ethnicity based on last name using data from the U.S. Census Bureau, by assigned ethnicity on manual review. Unit of analysis is at the investigator level.

C. Rosenman et al. (2023) last name score



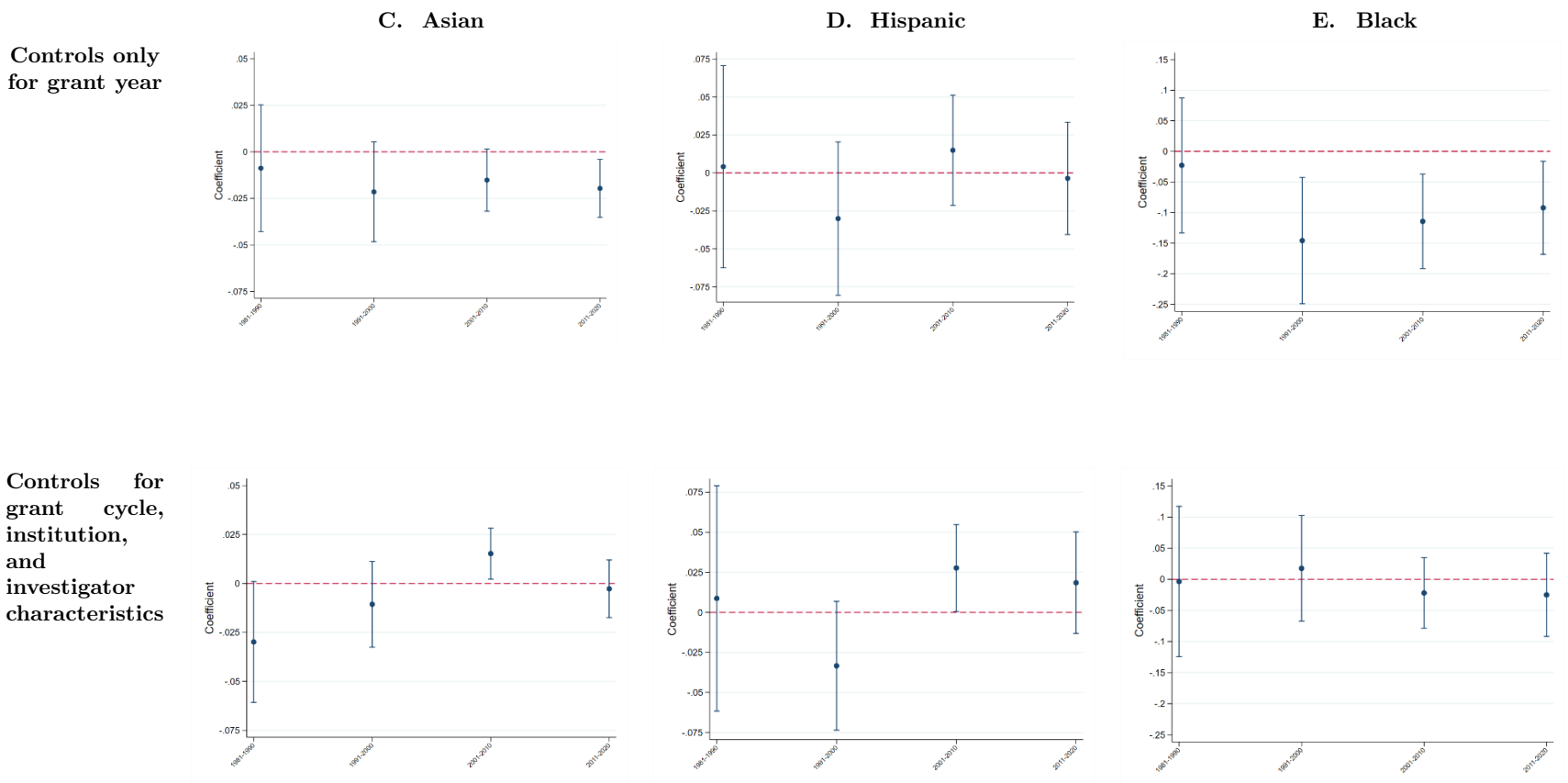
Note: Histogram of the percent likelihood of Black non-Hispanic ethnicity based on last name using data from Rosenman et al. (2023), by assigned ethnicity on manual review. Unit of analysis is at the investigator level.

D. Tzioumis (2018) first name score



Note: Histogram of the percent likelihood of Black non-Hispanic ethnicity based on first name using data from Tzioumis (2018), by assigned ethnicity on manual review. Unit of analysis is at the investigator level.

Figure A7. Time trends in grant renewal



Note: The plots above present the coefficient estimates (and associated 95% confidence intervals) from a linear probability model where the dependent variable is an indicator variable for if the grant cycle was renewed and independent variable of interest is an indicator for Asian (panel A), Hispanic (panel B), or Black (panel c) ethnicity. Each row-time period represents the results of a separate regression. Row 1 includes controls only for the grant cycle evaluation year modeled on Table ***, column ***, while row 2 was modeled on Table ***, column ***. The unit of analysis is the investigator-grant-cycle. Robust standard errors were used, clustered at the investigator and grant.

Table A1. Descriptive statistics: Investigator characteristics

	Mean	Median	Std. Dev.	Min.	Max.
Female	0.24	0	0.43	0	1
Ethnicity: White	0.81	1	0.39	0	1
Ethnicity: Asian	0.15	0	0.36	0	1
Ethnicity: Hispanic	0.03	0	0.17	0	1
Ethnicity: Black	0.01	0	0.09	0	1
Ethnicity: Other	0.00	0	0.05	0	1
Degree year	1983.54	1984	12.05	1932	2012
Degree: PhD	0.67	1	0.47	0	1
Degree: MD	0.21	0	0.40	0	1
Degree: MD/PhD	0.11	0	0.32	0	1
Degree: Other	0.02	0	0.13	0	1
Nb of R01-equivalent grants	1.81	1	1.22	1	14
Nb of R01-equivalent grant cycles	3.29	2	2.92	1	37
Career NIH R01 funding ($2020)	11,266,722	7,941,796	10,967,157	133,632	134,399,904
Career NIH funding ($2020)	18,463,929	10,160,652	28,647,272	133,632	1,743,186,688

Note: Number of grants and cycles refers to only those grants that started 1980-2015. N=32,011 investigators with a photo available.

Table A2. Descriptive statistics: Grant cycle

	Renewed grant cycles					Not renewed grant cycles				
	Mean	Median	Std. Dev.	Min.	Max.	Mean	Median	Std. Dev.	Min.	Max.
Grant cycle number	1.96	1	1.30	1	10	1.71	1	1.18	1	11
First cycle of grant	0.50	1	0.50	0	1	0.61	1	0.49	0	1
Year started	1998.52	1999	9.00	1980	2015	2001.49	2004	9.46	1980	2015
Cycle length (years)	4.41	4	1.60	1	19	3.21	3	1.11	1	14
Time, degree to cycle start (years)	17.70	16	8.50	-22	61	19.74	18	9.46	-18	66
Investigator pubs. prior to cycle	54.76	37	58.91	0	1,148	61.23	41	66.11	0	1,196
Investigator citations prior to cycle	2,222.74	941	4,090.39	0	89,857	2,499.34	1,055	4,606.50	0	113,601
Publications acknowledging grant										
Nb of publications	8.27	6	8.02	1	184	7.34	5	7.57	1	164
Journal impact factor	6.39	5	4.03	0	72	5.49	5	3.56	0	43
Citations	691.40	376	1,183.29	1	64,421	449.58	219	837.41	1	67,451
Highly selective journal	0.13	0	0.33	0	1	0.07	0	0.26	0	1
Publications not acknowledging grant										
Nb of publications	9.70	6	13.60	0	328	12.03	7	15.89	0	258
Journal impact factor	6.18	5	4.28	0	43	5.28	4	3.42	0	43
Citations	864.50	311	1,824.23	0	50,927	837.17	318	1,743.62	0	65,447
Highly selective journal	0.17	0	0.38	0	1	0.13	0	0.34	0	1
Fraction with any NIH funding	0.58	1	0.33	0	1	0.61	1	0.32	0	1

Note: Unit of analysis is at the investigator-grant-cycle level. N=92,125 (42,686 renewed and 49,439 not renewed investigator-grant-cycles). Grant cycle number minus one is the number of times the grant has previously been competitively renewed. Time from degree to cycle is measured from the last terminal degree earned by the investigator; 36 individuals (60 investigator-grant-cycles) earned an R01-equivalent grant prior to their last degree. This largely reflects investigators with multiple doctorates, frequently with large time gaps between them. Publications acknowledging a grant reflect those publications citing funding from the focal grant which were published up to one year after the conclusion of the grant cycle; publications not acknowledging the focal grant represents the investigator's other publications during the same time period. Citations are measured through 2020.

Table A3. Descriptive Statistics: Comparison of investigators with and without photo available

	Photo available	Photo unavailable	T-Stat
Grant cycle starting year	2004.200	1993.463	77.148
	(0.052)	(0.124)	
Career age	24.106	21.735	15.979
	(0.056)	(0.134)	
Career publications	60.923	39.512	22.801
	(0.368)	(0.546)	
Career citations	2527.315	1121.023	21.203
	(25.832)	(42.887)	
Number of R01-equivalent grants	6.340	4.623	20.093
	(0.033)	(0.058)	
Amount of R01-equivalent grants ($2020)	6,777,889	4,138,936	20.998
	(49,112)	(77,365)	
Last name frequency among NIH grantees	78.287	99.077	7.093
	(1.070)	(3.094)	
Female	0.244	0.222	3.301
	(0.002)	(0.006)	
Degree: MD	0.206	0.155	8.502
	(0.002)	(0.005)	
Degree: PhD	0.666	0.752	12.442
	(0.003)	(0.006)	
Degree: MD/PhD	0.112	0.070	9.104
	(0.002)	(0.004)	
Degree: Other	0.016	0.022	3.013
	(0.001)	(0.002)	
Census score: White	69.999	74.821	9.175
	(0.207)	(0.421)	
Census score: Asian	15.662	10.549	9.767
	(0.207)	(0.414)	
Census score: Hispanic	4.216	3.173	5.675
	(0.074)	(0.113)	
Census score: Black	7.445	8.783	7.077
	(0.073)	(0.177)	
Rosenman et al. score: White	70.802	74.982	8.441
	(0.194)	(0.401)	
Rosenman et al. score: Asian	12.257	8.031	9.848
	(0.169)	(0.332)	
Rosenman et al. score: Hispanic	3.513	2.263	7.326
	(0.069)	(0.106)	
Rosenman et al. score: Black	10.557	12.492	7.584
	(0.098)	(0.240)	

Note: Means, standard errors, and t-statistics are reported. NIH grants are reported in 2020 dollars after adjustment with the Biomedical R&D Producer Price Index. Grant cycle starting year, career age, career publications, career citations, and number and amount of R01-equivalent grants are evaluated at the start of the investigators last grant cycle during the study period (1980-2015). Last name frequency is calculated among all unique investigators who received any NIH funding. Census score and Rosenman et al. (2023) score report the continuous probability of each assigned ethnicity based on their respective data. Analysis is at the level of the investigator. N = 37,222 (32,011 with photo available and 5,211 without photo available).

Table A4. Regression analysis of photo availability

	(1)	(2)
Ln(Last name frequency among NIH grantees)	1.061**	1.063**
	(0.012)	(0.011)
Ln(Career publications)	0.692**	0.681**
	(0.029)	(0.029)
Ln(Career citations)	1.171**	1.183**
	(0.026)	(0.026)
Number of R01-equivalent grants	0.953**	0.952**
	(0.006)	(0.006)
Ln(Amount of R01-equivalent grants)	0.947*	0.943*
	(0.022)	(0.022)
Ln(Amount of R01-equivalent grants to institution)	1.000	1.000
	(0.003)	(0.003)
Degree year	0.954**	0.954**
	(0.003)	(0.003)
Grant cycle starting year	0.938**	0.937**
	(0.003)	(0.003)
Female	1.125**	1.114*
	(0.050)	(0.049)
Degree: MD	0.475**	0.480**
	(0.024)	(0.024)
Degree: MD/PhD	0.715**	0.729**
	(0.049)	(0.048)
Degree: Other	1.165	1.155
	(0.155)	(0.151)
Census last name score: White	1.071	
	(0.111)	
Census last name score: Asian	1.062	
	(0.113)	
Census last name score: Hispanic	0.983	
	(0.103)	
Census last name score: Black	1.073	
	(0.119)	
Rosenman et al. last name score: White		1.089†
		(0.056)
Rosenman et al. last name score: Asian		1.087
		(0.063)
Rosenman et al. last name score: Hispanic		0.977
		(0.056)
Rosenman et al. last name score: Black		1.082
		(0.058)
Number of investigators	30,326	31,805
Pseudo-R^2	0.184	0.184
Log likelihood	-10,412	-10,867

Note: Estimates stem from a logistic regression model where the unit of analysis is at the investigator level and the dependent variable is an indicator variable for if the investigator's picture was unavailable. Each column reflects the results of a separate regression, with exponentiated coefficients presented. Grant cycle starting year, career publications, career citations, and number and amount of R01-equivalent grants are evaluated at the start of the investigators last grant cycle during the study period (1980-2015). Last name frequency is calculated among all unique investigators who received any NIH funding. Census score and Rosenman et al. (2023) score report the continuous probability of each assigned ethnicity based on their data, scaled so a change of 1 corresponds to a 10% change the frequency of individuals with the corresponding ethnicity. Only investigators with available ethnicity predictions using last name from Census data (Column 1) and Rosenman et al. data (Column 2) are included. Robust standard errors were used. $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$.

Table A5. Descriptive statistics: Assigned ethnicities

	Mean	Median	Std. Dev.	Min.	Max.
Deepface image					
White	0.789	1	0.408	0	1
Asian	0.174	0	0.379	0	1
Hispanic	0.026	0	0.158	0	1
Black	0.011	0	0.106	0	1
Last name from U.S. Census Bureau					
White	0.801	1	0.399	0	1
Asian	0.166	0	0.372	0	1
Hispanic	0.021	0	0.143	0	1
Black	0.011	0	0.106	0	1
Last name from Rosenman et al. (2023)					
White	0.797	1	0.402	0	1
Asian	0.146	0	0.354	0	1
Hispanic	0.022	0	0.146	0	1
Black	0.034	0	0.180	0	1
First name from Tzioumis et al. (2018)					
White	0.925	1	0.263	0	1
Asian	0.045	0	0.206	0	1
Hispanic	0.029	0	0.169	0	1
Black	0.001	0	0.032	0	1
Integrated full name					
White	0.798	1	0.401	0	1
Asian	0.147	0	0.354	0	1
Hispanic	0.031	0	0.172	0	1
Black	0.023	0	0.151	0	1

Note: Analysis limited to those investigators with a picture available (n = 32,011). Assigned ethnicity is the ethnicity with highest predicted probability based on image analysis or name. Integrated name combines data from the U.S. Census Bureau, Rosenman et al. (2023) and Tzioumis (2018).

Table A6. Descriptive statistics: Ethnicity scores

	Mean	Median	Std. Dev.	Min.	Max.
Deepface image					
White	68.21	81	33.85	0	100
Asian	17.95	3	29.32	0	100
Hispanic	11.81	8	12.08	0	87
Black	2.03	0	9.18	0	100
Last name from U.S. Census Bureau					
White	70.00	85	33.28	0	100
Asian	15.66	1	33.20	0	100
Hispanic	4.22	2	11.93	0	98
Black	7.45	1	11.69	0	97
Last name from Rosenman et al. (2023)					
White	70.80	85	31.96	0	100
Asian	12.26	0	27.85	0	100
Hispanic	3.51	1	11.29	0	100
Black	10.56	1	16.10	0	100
First name from Rosenman et al. (2023)					
White	71.21	82	25.91	0	100
Asian	6.67	1	20.65	0	100
Hispanic	5.90	2	13.86	0	100
Black	13.73	10	12.07	0	100
Middle name from Rosenman et al. (2023)					
White	73.13	85	28.12	0	100
Asian	7.21	0	21.62	0	100
Hispanic	5.42	1	14.93	0	89
Black	11.57	8	11.91	0	99
First name from Tzioumis et al. (2018)					
White	84.96	93	21.85	0	100
Asian	6.49	2	17.81	0	100
Hispanic	5.07	2	13.28	0	95
Black	3.19	2	4.18	0	70

Note: Analysis limited to those investigators with a picture available (n = 32,011). Ethnicity score reflects the percent likelihood of having the corresponding ethnicity based on name or photo. Data is from U.S. Census, Rosenman et al. (2023) and Tzioumis (2018).

Table A7. Correlation across measures of race and ethnicity

A. Assigned ethnicity: White

Variables	(1)	(2)	(3)	(4)
(1) White (Manual)	1.000			
(2) White (Deepface image)	0.694	1.000		
(3) White (Census last name)	0.896	0.637	1.000	
(4) White (Rosenman et al. last name)	0.810	0.577	0.880	1.000

B. Assigned ethnicity: Asian

Variables	(1)	(2)	(3)	(4)
(1) Asian (Manual)	1.000			
(2) Asian (Deepface image)	0.737	1.000		
(3) Asian (Census last name)	0.957	0.712	1.000	
(4) Asian (Rosenman et al. last name)	0.926	0.689	0.953	1.000

C. Assigned ethnicity: Hispanic

Variables	(1)	(2)	(3)	(4)
(1) Hispanic (Manual)	1.000			
(2) Hispanic (Deepface image)	0.064	1.000		
(3) Hispanic (Census last name)	0.821	0.061	1.000	
(4) Hispanic (Rosenman et al. last name)	0.795	0.065	0.900	1.000

D. Assigned ethnicity: Black

Variables	(1)	(2)	(3)	(4)
(1) Black (Manual)	1.000			
(2) Black (Deepface image)	0.632	1.000		
(3) Black (Census last name)	0.178	0.128	1.000	
(4) Black (Rosenman et al. last name)	0.140	0.103	0.515	1.000

E. Predicted ethnicity

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
(1) Deepface image score: white	1.000											
(2) Census last name score: white	0.654	1.000										
(3) Rosenman et al. last name score: white	0.616	0.969	1.000									
(4) Deepface image score: Asian	-0.897	-0.705	-0.661	1.000								
(5) Census last name score: Asian	-0.702	-0.880	-0.831	0.785	1.000							
(6) Rosenman et al. last name score: Asian	-0.689	-0.858	-0.827	0.778	0.983	1.000						
(7) Deepface image score: Hispanic	-0.367	0.002	0.005	0.002	-0.030	-0.043	1.000					
(8) Census last name score: Hispanic	0.007	-0.225	-0.184	-0.050	-0.103	-0.108	0.105	1.000				
(9) Rosenman et al. last name score: Hispanic	-0.032	-0.243	-0.225	-0.004	-0.045	-0.061	0.100	0.935	1.000			
(10) Deepface image score: Black	-0.314	-0.121	-0.125	0.041	0.067	0.057	0.075	0.006	0.002	1.000		
(11) Census last name score: Black	0.148	-0.076	-0.178	-0.194	-0.258	-0.260	-0.031	-0.072	-0.123	0.137	1.000	
(12) Rosenman et al. last name score: Black	0.151	-0.062	-0.195	-0.190	-0.258	-0.266	-0.032	-0.067	-0.124	0.112	0.951	1.000

Note: The tables above present the correlation across measures of ethnicity. Panels A, B, C, and . include indicator variables for the most likely race based on manual review, Deepface picture analysis, last name using data from the U.S. Census Bureau and last name using data from Rosenman et al. (2023). Panel E includes measures of the continuous probability of each assigned ethnicity. Analysis is at the level of the investigator (N = 32,011 with photo available).

Table A8. Descriptive statistics: Measures of ethnic research orientation

A. Measures using investigator publications

	Mean	Median	Std. Dev.	Min.	Max.
Relative disease mortality					
Asian	1.313	1	1.690	0	8
Hispanic	2.221	1	7.755	0	92
Black	6.250	2	30.350	0	301
Ethnicity					
White	0.054	0	0.225	0	1
Asian	0.020	0	0.141	0	1
Hispanic	0.027	0	0.161	0	1
Black	0.055	0	0.227	0	1
Country					
White	0.181	0	0.385	0	1
Asian	0.035	0	0.183	0	1
Hispanic	0.027	0	0.163	0	1
Black	0.020	0	0.141	0	1
Ethnicity-adjacent topics					
Minorities	0.007	0	0.081	0	1
Health Disparities	0.006	0	0.080	0	1
Socioeconomic status	0.068	0	0.251	0	1
Strongly ethnic diseases					
White	0.016	0	0.125	0	1
Asian	0.027	0	0.162	0	1
Hispanic	0.004	0	0.066	0	1
Black	0.011	0	0.106	0	1

B. Measures using citations to investigator publications

	Mean	Median	Std. Dev.	Min.	Max.
Relative disease mortality					
Asian	0.844	1	0.764	0	8
Hispanic	1.490	1	3.690	0	92
Black	2.801	2	8.845	0	301
Ethnicity					
White	3.929	0	22.363	0	1,083
Asian	3.354	0	17.550	0	873
Hispanic	1.809	0	12.857	0	551
Black	4.121	0	25.486	0	1,217
Country					
White	36.531	3	182.053	0	11,795
Asian	8.780	1	44.640	0	3,797
Hispanic	4.197	0	21.453	0	1,555
Black	3.154	0	23.205	0	1,457
Ethnicity-adjacent topics					
Minorities	0.316	0	2.695	0	216
Health Disparities	0.554	0	5.436	0	584
Socioeconomic status	7.864	0	57.051	0	4,915
Strongly ethnic diseases					
White	1.967	0	21.782	0	1,795
Asian	6.152	0	58.867	0	6,492
Hispanic	0.589	0	9.541	0	986
Black	1.171	0	16.433	0	1,776

Note: Outcomes are measured at the investigator-grant-cycle level (N= 92,125 investigator-grant-cycles). Measures are only defined for those cycles with qualifying associated publications (see the Materials and Methods section for details).

Table A9. Effect of ethnicity on research topic: Basic science publications

	Disease relative mortality		
	Asian	Hispanic	Black
Asian ethnicity	0.1134**	0.0788	0.2067
	(0.0283)	(0.1235)	(0.4061)
Hispanic ethnicity	0.1050	0.9094*	1.0906
	(0.0735)	(0.4365)	(0.9605)
Black ethnicity	-0.0222	0.5591	9.6432*
	(0.1050)	(0.3529)	(4.8332)
Mean of Dependent Variable	1.126	1.891	4.475
Adjusted R2	0.1529	0.1012	0.0663
Nb. of Investigators	18,247	18,490	18,564
Nb. of Grant cycles	43,566	44,364	44,634

Note: Estimates stem from an OLS regression where the unit of analysis is the investigator-grant-cycle. The independent variables are indicator variables for Asian, Hispanic and Black ethnicity with white ethnicity and other ethnicity as the omitted category. The outcome variable of interest is the maximum disease relative mortality calculated relative to white mortality, calculated using only basic science publications. All estimates include controls for grant year, gender, investigator degree, investigator career age, length of the grant cycle, a set of 2,911 indicator variables corresponding to each NIH institute or center and study section, log NIH funding to the investigator's institution, log investigator prior career citations, log investigator prior NIH grant funding, no prior NIH grant funding, research methodology, and number of publications during the grant cycle. Robust standard errors were used, clustered at the investigator and grant. $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$.

Table A10. Effect of ethnicity on the research direction of forward citations: Basic science publications

A. Disease relative mortality

	Disease relative mortality		
	Asian	Hispanic	Black
Asian ethnicity	0.0279*	-0.0067	-0.0672
	(0.0115)	(0.0563)	(0.1442)
Hispanic ethnicity	0.0210	0.2323	0.2330
	(0.0302)	(0.1676)	(0.4088)
Black ethnicity	0.0466	0.1905	3.4545
	(0.0509)	(0.2139)	(2.1089)
Mean of Dependent Variable	0.887	1.568	2.768
Adjusted R2	0.2271	0.1855	0.0814
Nb. of Investigators	26,060	26,094	26,110
Nb. of Grant cycles	74,787	74,948	74,998

B. Ethnicity

	Ethnicity MeSH topics		
	Asian	Hispanic	Black
Asian ethnicity	0.1282*	-0.1687	-0.1967†
	(0.0624)	(0.1544)	(0.1100)
Hispanic ethnicity	0.0002	0.5852†	0.1995
	(0.1348)	(0.3341)	(0.2592)
Black ethnicity	0.6376*	0.0231	1.1830**
	(0.2608)	(0.2238)	(0.3406)
Pseudo-R2	0.5810	0.5073	0.5494
Log-likelihood	-109,288	-35,086	-87,634
Nb. of Investigators	27,240	26,801	27,085
Nb. of Grant cycles	78,737	77,304	78,184

C. Country

	Country MeSH topics		
	Asian	Hispanic	Black
Asian ethnicity	0.4707**	0.2648	0.1330
	(0.1285)	(0.1854)	(0.1522)
Hispanic ethnicity	-0.1131	0.2539†	-0.2323†
	(0.1064)	(0.1368)	(0.1307)
Black ethnicity	0.6576*	0.5146	0.2753
	(0.3192)	(0.3326)	(0.4981)
Pseudo-R2	0.6193	0.5401	0.5602
Log-likelihood	-192,251	-123,971	-102,257
Nb. of Investigators	27,300	27,258	27,171
Nb. of Grant cycles	78,870	78,757	78,571

D. Ethnicity-adjacent topics

	Ethnicity-adjacent MeSH topics		
	Minorities	Disparities	Socioeconomic status
Asian ethnicity	-0.3821†	-0.6119**	-0.3055**
	(0.2160)	(0.1425)	(0.0767)
Hispanic ethnicity	0.3423	0.2810	-0.0343
	(0.4412)	(0.1984)	(0.1442)
Black ethnicity	0.9189	1.2732*	0.6607**
	(0.5888)	(0.6296)	(0.2141)
Pseudo-R2	0.3242	0.3680	0.4641
Log-likelihood	-2,969	-7,541	-42,362
Nb. of Investigators	17,929	23,011	27,226
Nb. of Grant cycles	44,585	63,420	78,626

Note: Estimates stem from an OLS regression (panel A) or linear probability model (panels B, C, and D) where the unit of analysis is the investigator-grant-cycle. The independent variables are indicator variables for white, Asian, Hispanic and Black ethnicity. If white ethnicity is the independent variable of interest, all other ethnicities are the omitted category, while if Asian, Hispanic and Black ethnicity is the independent variable of interest, white ethnicity and other ethnicity are the omitted category. The outcome variable of interest is the average maximum disease relative mortality calculated relative to white mortality (panel A), and the count of papers with the corresponding ethnicity topic (panel B), country topic (panel C), and ethnicity-adjacent topics (panel D), calculated using citations to an investigator's basic science grant cycle publications. Panels B, C, and D include an offset for the total number of citations. All estimates include controls for grant year, gender, investigator degree, investigator career age, length of the grant cycle, a set of 2,911 indicator variables corresponding to each NIH institute or center and study section, log NIH funding to the investigator's institution, log investigator prior career citations, log investigator prior NIH grant funding, no prior NIH grant funding, research methodology, and number of publications during the grant cycle. Robust standard errors were used, clustered at the investigator and grant. $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$.

Table A11. Effect of ethnicity on grant renewal by renewal cycle

	First renewal				Second and future renewals			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Asian ethnicity	0.012†	0.007	0.010†	0.005	-0.034**	-0.007	0.004	0.004
	(0.006)	(0.006)	(0.006)	(0.006)	(0.008)	(0.008)	(0.007)	(0.007)
Hispanic ethnicity	0.006	0.010	0.011	0.008	0.006	0.007	0.025†	0.026†
	(0.014)	(0.012)	(0.012)	(0.012)	(0.019)	(0.016)	(0.014)	(0.014)
Black ethnicity	-0.096**	-0.070**	-0.035	-0.034	-0.072*	-0.010	0.036	0.037
	(0.027)	(0.024)	(0.023)	(0.023)	(0.037)	(0.034)	(0.030)	(0.031)
Grant year	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Investigator demographics	No	Yes	Yes	Yes	No	Yes	Yes	Yes
IC x study section	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Institutional quality	No	No	Yes	Yes	No	No	Yes	Yes
Investigator quality	No	No	Yes	Yes	No	No	Yes	Yes
Pub. quantity, quality during cycle	No	No	No	Yes	No	No	No	Yes
Mean of dependent variable	0.415	0.415	0.415	0.415	0.526	0.530	0.530	0.530
Std. Dev. of dependent variable	0.493	0.493	0.493	0.493	0.499	0.499	0.499	0.499
Effect Asian ethnicity, in s.d. units	0.024	0.015	0.020	0.010	-0.069	-0.014	0.007	0.009
Effect Hispanic ethnicity, in s.d. units	0.012	0.020	0.022	0.017	0.011	0.014	0.051	0.052
Effect Black ethnicity, in s.d. units	-0.195	-0.143	-0.071	-0.068	-0.145	-0.021	0.072	0.074
Adjusted R^2	0.0413	0.2279	0.2901	0.3016	0.0367	0.2494	0.3254	0.3360
Nb. of investigators	30,468	30,160	30,160	30,160	16,374	16,078	16,065	16,065
Nb. of investigator-grant-cycles	51,805	50,953	50,952	50,952	40,320	39,853	39,803	39,803

Note: Estimates stem from a linear probability model where the unit of analysis is the investigator-grant-cycle and the dependent variable is an indicator variable for if the grant cycle was renewed. Columns 1-4 limit the sample to the first cycle for each investigator-grant, while columns 5-8 limit it to cycles 2 and above for each investigator-grant. The independent variables are indicator variables for Asian, Hispanic and Black ethnicity, with white ethnicity and other ethnicity as the omitted category. Column 1 and 5 include only an indicator variable for grant year, while columns 2 and 6 add indicator variables for gender, investigator degree, investigator career age, length of the grant cycle, and a set of 2,911 indicator variables corresponding to each NIH institute or center and study section. Columns 3 and 7 include all of the previous controls, but also include log NIH funding to the investigator's institution, log investigator prior career citations, log investigator prior NIH grant funding, and an indicator for no prior NIH grant funding. Finally, columns 4 and 8 include all previous controls, but also add log mean journal impact factor, indictor variables for grant and non-grant publications in one of four highly selective journals, a set of 4 indictors variables for the research methodology used, and a set of 21 indicator variables separately for the number of publications during the grant cycle acknowledging and not acknowledging the grant. Robust standard errors were used, clustering at the investigator and grant. $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$.

Table A12. Robustness: Alternative measures of ethnicity

A. Image-based predicted ethnicity

	Relative disease mortality			Ethnicity MeSH topics			Country MeSH topics			Grant renewal	
	Asian	Hispanic	Black	Asian	Hispanic	Black	Asian	Hispanic	Black	Year controls	Full controls
Asian ethnicity	0.0514†	-0.0462	0.2878	0.0138**	-0.0006	-0.0023	0.0334**	-0.0016	-0.0020	-0.0149**	0.0002
	(0.0267)	(0.1252)	(0.4653)	(0.0021)	(0.0017)	(0.0023)	(0.0031)	(0.0019)	(0.0017)	(0.0056)	(0.0042)
Hispanic ethnicity	0.0374	-0.3897*	0.9087	0.0044	0.0022	0.0013	-0.0018	0.0045	0.0011	-0.0119	0.0170†
	(0.0709)	(0.1635)	(1.1829)	(0.0049)	(0.0050)	(0.0074)	(0.0054)	(0.0058)	(0.0044)	(0.0132)	(0.0098)
Black ethnicity	0.1700†	0.1009	1.8791	0.0064	0.0061	0.0259**	0.0008	0.0069	0.0218**	-0.0575**	-0.0223
	(0.1008)	(0.2866)	(2.0181)	(0.0061)	(0.0072)	(0.0098)	(0.0081)	(0.0077)	(0.0080)	(0.0211)	(0.0149)
Mean of Dependent Variable	1.319	2.223	6.282	0.021	0.027	0.055	0.035	0.027	0.020	0.464	0.466
Adjusted R2	0.1696	0.0954	0.0607	0.1193	0.2213	0.2824	0.1174	0.1203	0.1467	0.0351	0.3228
Nb. of Investigators	24,281	24,483	24,545	29,941	29,941	29,941	29,941	29,941	29,941	30,179	29,953
Nb. of Grant cycles	61,202	62,021	62,259	86,145	86,145	86,145	86,145	86,145	86,145	86,812	86,178

Note: The estimates for relative disease mortality, ethnicity topics, country topics, and grant renewal are very similar to Table 2 panel A, Table 2 panel B, Table 2 panel C, and columns 1 and 4 of Table 3, respectively. The sample is limited to those investigators with a profile picture available for which DeepFace could analyze the faces; it excludes 7,043 investigators (16,532 investigator-grant-cycles) which had a profile picture but for which DeepFace could not make an ethnicity prediction. The independent variable is a set of indicator variables for assigned ethnicity from DeepFace image analysis corresponding to the ethnicity with the highest assigned probability. White ethnicity is the omitted category. Robust standard errors were used, clustered at the investigator and grant. $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$.

B. Name-based predicted ethnicity

	Relative disease mortality			Ethnicity MeSH topics			Country MeSH topics			Grant renewal	
	Asian	Hispanic	Black	Asian	Hispanic	Black	Asian	Hispanic	Black	Year controls	Full controls
Asian ethnicity	0.1300**	-0.0400	-0.0460	0.0178**	-0.0026	-0.0062*	0.0424**	-0.0063**	-0.0030†	-0.0139*	0.0032
	(0.0292)	(0.1098)	(0.4450)	(0.0024)	(0.0017)	(0.0024)	(0.0034)	(0.0017)	(0.0016)	(0.0059)	(0.0046)
Hispanic ethnicity	0.0698	0.2657	2.0348	-0.0033	0.0192**	-0.0042	-0.0061	0.0504**	-0.0021	0.0134	0.0119
	(0.0670)	(0.3609)	(1.2960)	(0.0034)	(0.0062)	(0.0063)	(0.0060)	(0.0087)	(0.0035)	(0.0166)	(0.0119)
Black ethnicity	0.0658	0.2512	4.5437	0.0084	0.0085	0.0270*	0.0077	0.0030	0.0149†	0.0061	0.0132
	(0.0973)	(0.3710)	(3.6338)	(0.0081)	(0.0081)	(0.0110)	(0.0099)	(0.0081)	(0.0085)	(0.0231)	(0.0147)
Mean of Dependent Variable	1.299	2.207	6.174	0.019	0.025	0.052	0.034	0.025	0.019	0.453	0.455
Adjusted R2	0.1672	0.1060	0.0597	0.1161	0.2160	0.2752	0.1208	0.1272	0.1448	0.0307	0.3139
Nb. of Investigators	23,665	23,912	23,980	30,106	30,106	30,106	30,106	30,106	30,106	30,329	30,118
Nb. of Grant cycles	58,038	58,937	59,172	83,812	83,812	83,812	83,812	83,812	83,812	84,455	83,849

Note: The estimates for relative disease mortality, ethnicity topics, country topics, and grant renewal are very similar to Table 2 panel A, Table 2 panel B, Table 2 panel C, and columns 1 and 4 of Table 3, respectively. The sample is limited to those investigators whose last name the U.S. Census Bureau has public information on the distribution of ethnicities for the name; this excludes 11,816 investigators (18,709 investigator-grant-cycles) for which the U.S. Census Bureau with rare names for which the ethnicity distribution is withheld to protect confidentiality. The independent variable is a set of indicator variables for assigned ethnicity based on last name using U.S. Census data corresponding to the ethnicity with the highest assigned probability. White ethnicity is the omitted category. Robust standard errors were used, clustered at the investigator and grant. $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$.

C. Ethnicity as continuous variable from image-based prediction

	Relative disease mortality			Ethnicity MeSH topics			Country MeSH topics			Grant renewal	
	Asian	Hispanic	Black	Asian	Hispanic	Black	Asian	Hispanic	Black	Year controls	Full controls
Asian ethnicity	0.0968**	-0.2356	0.2223	0.0211**	-0.0030	-0.0063*	0.0521**	-0.0036	-0.0054**	-0.0187**	-0.0011
	(0.0349)	(0.1469)	(0.5591)	(0.0030)	(0.0020)	(0.0029)	(0.0043)	(0.0023)	(0.0020)	(0.0073)	(0.0054)
Hispanic ethnicity	-0.0629	-0.5743†	1.5088	-0.0139**	-0.0002	0.0000	-0.0318**	0.0072	0.0009	0.0085	0.0060
	(0.0830)	(0.3481)	(1.4457)	(0.0053)	(0.0054)	(0.0076)	(0.0070)	(0.0061)	(0.0054)	(0.0171)	(0.0124)
Black ethnicity	0.2788*	0.5258	4.0861	0.0033	0.0201*	0.0472**	-0.0069	0.0181†	0.0338**	-0.0725**	-0.0152
	(0.1293)	(0.4560)	(2.7138)	(0.0081)	(0.0101)	(0.0124)	(0.0110)	(0.0106)	(0.0107)	(0.0244)	(0.0177)
Mean of Dependent Variable	1.319	2.223	6.282	0.021	0.027	0.055	0.035	0.027	0.020	0.464	0.466
Adjusted R2	0.1698	0.0955	0.0608	0.1198	0.2214	0.2825	0.1193	0.1204	0.1469	0.0352	0.3228
Nb. of Investigators	24,281	24,483	24,545	29,941	29,941	29,941	29,941	29,941	29,941	30,179	29,953
Nb. of Grant cycles	61,202	62,021	62,259	86,145	86,145	86,145	86,145	86,145	86,145	86,812	86,178

Note: The estimates for relative disease mortality, ethnicity topics, country topics, and grant renewal are very similar to Table 2 panel A, Table 2 panel B, Table 2 panel C, and columns 1 and 4 of Table 3, respectively. The sample is limited to those investigators with a profile picture available for which DeepFace could analyze the faces; it excludes 7,043 investigators (16,532 investigator-grant-cycles) which had a profile picture but for which DeepFace could not make an ethnicity prediction. The independent variable is a set of variables for the predicted probability of each ethnicity from DeepFace image analysis, scaled from zero to one hundred. White ethnicity is the omitted category. Robust standard errors were used, clustered at the investigator and grant. $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$.

D. Ethnicity as continuous variable from name-based prediction

	Relative disease mortality			Ethnicity MeSH topics			Country MeSH topics			Grant renewal	
	Asian	Hispanic	Black	Asian	Hispanic	Black	Asian	Hispanic	Black	Year controls	Full controls
Asian ethnicity	0.1463**	-0.0433	-0.0230	0.0208**	-0.0028	-0.0071*	0.0494**	-0.0060**	-0.0026	-0.0162*	0.0044
	(0.0339)	(0.1308)	(0.5229)	(0.0028)	(0.0020)	(0.0029)	(0.0039)	(0.0021)	(0.0019)	(0.0068)	(0.0053)
Hispanic ethnicity	0.0727	0.3479	1.8331	-0.0056	0.0253**	-0.0030	-0.0111†	0.0685**	-0.0033	0.0077	0.0082
	(0.0780)	(0.4605)	(1.5576)	(0.0044)	(0.0077)	(0.0080)	(0.0058)	(0.0107)	(0.0045)	(0.0204)	(0.0145)
Black ethnicity	0.0006	-0.0877	0.4174	0.0010	-0.0027	0.0045	0.0034	0.0018	0.0116†	-0.0264	0.0150
	(0.0884)	(0.3638)	(2.0672)	(0.0055)	(0.0057)	(0.0084)	(0.0071)	(0.0067)	(0.0063)	(0.0190)	(0.0131)
Mean of Dependent Variable	1.299	2.207	6.174	0.019	0.025	0.052	0.034	0.025	0.019	0.453	0.455
Adjusted R2	0.1672	0.1060	0.0595	0.1162	0.2160	0.2750	0.1212	0.1276	0.1447	0.0307	0.3139
Nb. of Investigators	23,665	23,912	23,980	30,106	30,106	30,106	30,106	30,106	30,106	30,329	30,118
Nb. of Grant cycles	58,038	58,937	59,172	83,812	83,812	83,812	83,812	83,812	83,812	84,455	83,849

Note: The estimates for relative disease mortality, ethnicity topics, country topics, and grant renewal are very similar to Table 2 panel A, Table 2 panel B, Table 2 panel C, and columns 1 and 4 of Table 3, respectively. The sample is limited to those investigators whose last name the U.S. Census Bureau has public information on the distribution of ethnicities for the name; this excludes 11,816 investigators (18,709 investigator-grant-cycles) for which the U.S. Census Bureau with rare names for which the ethnicity distribution is withheld to protect confidentiality. The independent variable is a set of variables for the predicted probability of each ethnicity based on last name using U.S. Census data, scaled from zero to one hundred. White ethnicity is the omitted category. Robust standard errors were used, clustered at the investigator and grant. $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$.

E. High-confidence predicted ethnicity

	Relative disease mortality			Ethnicity MeSH topics			Country MeSH topics			Grant renewal	
	Asian	Hispanic	Black	Asian	Hispanic	Black	Asian	Hispanic	Black	Year controls	Full controls
Asian ethnicity	0.0973*	-0.2669†	-0.5293	0.0222**	-0.0008	-0.0076*	0.0606**	-0.0030	-0.0052**	-0.0120	-0.0028
	(0.0379)	(0.1480)	(0.5497)	(0.0034)	(0.0021)	(0.0031)	(0.0049)	(0.0025)	(0.0020)	(0.0079)	(0.0061)
Hispanic ethnicity	-0.0755	-1.2395	-2.4807*	0.0155	0.0290	-0.0273*	0.0050	0.0494	-0.0185*	-0.0106	-0.0295
	(0.2533)	(0.7993)	(1.0455)	(0.0221)	(0.0272)	(0.0123)	(0.0191)	(0.0331)	(0.0082)	(0.0610)	(0.0318)
Black ethnicity	0.2142	0.5474	5.7487	0.0037	0.0299†	0.1017**	0.0180	0.0301	0.0825**	-0.1006**	-0.0058
	(0.1326)	(0.5133)	(5.1760)	(0.0131)	(0.0176)	(0.0235)	(0.0209)	(0.0186)	(0.0221)	(0.0339)	(0.0262)
Mean of Dependent Variable	1.313	2.170	6.188	0.021	0.027	0.054	0.037	0.026	0.020	0.465	0.469
Adjusted R2	0.1713	0.0834	0.0608	0.1326	0.2372	0.3045	0.1394	0.1272	0.1544	0.0374	0.3258
Nb. of Investigators	14,429	14,555	14,596	17,939	17,939	17,939	17,939	17,939	17,939	18,161	17,944
Nb. of Grant cycles	36,558	37,070	37,206	52,019	52,019	52,019	52,019	52,019	52,019	52,632	52,034

Note: The estimates for relative disease mortality, ethnicity topics, country topics, and grant renewal are very similar to Table 2 panel A, Table 2 panel B, Table 2 panel C, and columns 1 and 4 of Table 3, respectively. The sample is limited to the 18,161 investigators (52,632 investigator-grant-cycles) for which a high-confidence ethnicity prediction can be made. An ethnicity assignment is considered high confidence if either the image-based prediction or name-based prediction has an assigned probability of over 95% and this agrees with the assigned ethnicity on clerical review. The independent variable is a set of indicator variables for assigned ethnicity, with white ethnicity as the omitted category. Robust standard errors were used, clustered at the investigator and grant. $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$.

Table A13. Effect of ethnicity on research topic: Strongly ethnic diseases

	Strongly ethnic diseases			
	White	Asian	Hispanic	Black
White ethnicity	0.0038*			
	(0.0015)			
Asian ethnicity		0.0147**	-0.0015†	-0.0000
		(0.0028)	(0.0009)	(0.0014)
Hispanic ethnicity		-0.0014	-0.0015	0.0056
		(0.0048)	(0.0014)	(0.0046)
Black ethnicity		0.0018	0.0025	0.0312*
		(0.0104)	(0.0040)	(0.0143)
Mean of Dependent Variable	0.016	0.027	0.004	0.011
Adjusted R2	0.0642	0.1014	0.0404	0.0475
Nb. of Investigators	31,777	31,777	31,777	31,777
Nb. of Grant cycles	91,467	91,467	91,467	91,467

Note: Estimates stem from a linear probability model where the unit of analysis is the investigator-grant-cycle. The independent variables are indicator variables for white, Asian, Hispanic and Black ethnicity. In column 1, the omitted category is all other ethnicities, while in columns 2-4, white ethnicity and other ethnicity are the omitted category. The outcome variable of interest an indicator variable for having any research with strong ethnic association during the grant cycle. All estimates include controls for grant year, gender, investigator degree, investigator career age, length of the grant cycle, a set of 2,911 indicator variables corresponding to each NIH institute or center and study section, log NIH funding to the investigator's institution, log investigator prior career citations, log investigator prior NIH grant funding, no prior NIH grant funding, research methodology, and number of publications during the grant cycle. Robust standard errors were used, clustered at the investigator and grant. $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$.

Table A14. Effect of ethnicity on the research direction of forward citations: strongly ethnic diseases

	Strongly ethnic diseases			
	White	Asian	Hispanic	Black
White ethnicity	0.3767**			
	(0.1415)			
Asian ethnicity		0.3712**	-0.1169	-0.1641
		(0.1131)	(0.1760)	(0.1104)
Hispanic ethnicity		-0.0578	-0.1146	0.0070
		(0.2200)	(0.3150)	(0.2911)
Black ethnicity		0.0987	0.4878	1.9669**
		(0.2403)	(0.4909)	(0.4619)
Mean of Dependent Variable	0.4404	0.6222	0.5352	0.4275
Adjusted R2	-360,931	-621,036	-105,425	-236,645
Nb. of Investigators	31,881	31,950	30,127	31,792
Nb. of Grant cycles	91,646	91,875	87,249	91,457

Note: Estimates stem from a linear probability model where the unit of analysis is the investigator-grant-cycle. The independent variables are indicator variables for white, Asian, Hispanic and Black ethnicity. In column 1, the omitted category is all other ethnicities, while in columns 2-4, white ethnicity and other ethnicity are the omitted category. The outcome variable of interest an indicator variable for having any research with strong ethnic association during the grant cycle. All estimates include controls for grant year, gender, investigator degree, investigator career age, length of the grant cycle, a set of 2,911 indicator variables corresponding to each NIH institute or center and study section, log NIH funding to the investigator's institution, log investigator prior career citations, log investigator prior NIH grant funding, no prior NIH grant funding, research methodology, and number of publications during the grant cycle. Robust standard errors were used, clustered at the investigator and grant. $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$.

Table A15. Effect of ethnicity on research topic: Last author publications

A. Disease relative mortality

	Disease relative mortality		
	Asian	Hispanic	Black
Asian ethnicity	0.0985**	0.0607	0.2510
	(0.0278)	(0.0958)	(0.3754)
Hispanic ethnicity	0.0433	1.1292*	0.8823
	(0.0513)	(0.4568)	(0.6897)
Black ethnicity	0.0809	0.5265	11.6189*
	(0.0762)	(0.3678)	(5.3764)
Mean of Dependent Variable	1.052	1.717	4.507
Adjusted R2	0.1787	0.1514	0.0804
Nb. of Investigators	20,892	21,192	21,274
Nb. of Grant cycles	49,448	50,419	50,700

B. Ethnicity

	Ethnicity MeSH topics		
	Asian	Hispanic	Black
Asian ethnicity	0.0084**	-0.0007	-0.0002
	(0.0015)	(0.0011)	(0.0018)
Hispanic ethnicity	0.0017	0.0215**	0.0103*
	(0.0024)	(0.0041)	(0.0050)
Black ethnicity	0.0040	0.0062	0.0783**
	(0.0075)	(0.0088)	(0.0157)
Mean of Dependent Variable	0.008	0.012	0.026
Adjusted R2	0.0727	0.1532	0.1905
Nb. of Investigators	31,789	31,789	31,789
Nb. of Grant cycles	91,503	91,503	91,503

C. Country

	Country MeSH topics		
	Asian	Hispanic	Black
Asian ethnicity	0.0202**	-0.0021†	-0.0010
	(0.0023)	(0.0011)	(0.0011)
Hispanic ethnicity	-0.0010	0.0183**	-0.0025
	(0.0030)	(0.0042)	(0.0021)
Black ethnicity	0.0078	0.0214*	0.0224*
	(0.0111)	(0.0104)	(0.0108)
Mean of Dependent Variable	0.015	0.011	0.010
Adjusted R2	0.0727	0.0796	0.1161
Nb. of Investigators	31,789	31,789	31,789
Nb. of Grant cycles	91,503	91,503	91,503

D. Ethnicity-adjacent topics

	Ethnicity-adjacent MeSH topics		
	Minorities	Disparities	Socioeconomic status
Asian ethnicity	-0.0001	0.0002	-0.0026†
	(0.0004)	(0.0005)	(0.0015)
Hispanic ethnicity	0.0038*	0.0041†	0.0054
	(0.0017)	(0.0021)	(0.0042)
Black ethnicity	0.0052	0.0112*	0.0311*
	(0.0060)	(0.0056)	(0.0125)
Mean of Dependent Variable	0.002	0.003	0.035
Adjusted R2	0.0691	0.0978	0.3084
Nb. of Investigators	31,789	31,789	31,789
Nb. of Grant cycles	91,503	91,503	91,503

E. Strongly ethnic diseases

	Strongly ethnic diseases		
	Asian	Hispanic	Black
Asian ethnicity	0.0078**	-0.0007	0.0006
	(0.0019)	(0.0005)	(0.0009)
Hispanic ethnicity	-0.0029	-0.0008	0.0027
	(0.0028)	(0.0009)	(0.0025)
Black ethnicity	-0.0011	0.0028	0.0223*
	(0.0053)	(0.0032)	(0.0113)
Mean of Dependent Variable	0.012	0.002	0.005
Adjusted R2	0.0798	0.0450	0.0417
Nb. of Investigators	31,789	31,789	31,789
Nb. of Grant cycles	91,503	91,503	91,503

Note: Estimates stem from an OLS regression (panel A) or linear probability model (panels B, C, D, and E) where the unit of analysis is the investigator-grant-cycle. The independent variables are indicator variables forAsian, Hispanic and Black ethnicity, with white ethnicity and other ethnicity as the omitted category. The outcome variable of interest is the maximum disease relative mortality calculated relative to white mortality (panel A), any ethnicity topic (panel B), any country topic (panel C), any ethnicity adjacent topics (panel D), and any research with strong ethnic association (panel E) during the grant cycle, calculated using only publications for which the investigator was the last author. All estimates include controls for grant year, gender, investigator degree, investigator career age, length of the grant cycle, a set of 2,911 indicator variables corresponding to each NIH institute or center and study section, log NIH funding to the investigator's institution, log investigator prior career citations, log investigator prior NIH grant funding, no prior NIH grant funding, research methodology, and number of publications during the grant cycle. Robust standard errors were used, clustered at the investigator and grant. $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$.

Table A16. Robustness: Investigator name frequency

A. Ethnic research orientation

	Full sample	Below median name frequency	Above median name frequency
Relative disease mortality			
Asian ethnicity	0.1281**	0.0695	0.1343**
	(0.0287)	(0.0577)	(0.0353)
Hispanic ethnicity	1.2834**	1.2301*	1.0935†
	(0.4330)	(0.5734)	(0.6508)
Black ethnicity	12.9128*	13.9111†	11.8484
	(5.3748)	(7.2923)	(7.2270)
Ethnicity			
Asian ethnicity	0.0186**	0.0049	0.0208**
	(0.0024)	(0.0036)	(0.0029)
Hispanic ethnicity	0.0269**	0.0193**	0.0347**
	(0.0052)	(0.0058)	(0.0097)
Black ethnicity	0.0957**	0.1247**	0.0689**
	(0.0180)	(0.0255)	(0.0233)
Country			
Asian ethnicity	0.0424**	0.0106†	0.0480**
	(0.0034)	(0.0056)	(0.0041)
Hispanic ethnicity	0.0548**	0.0361**	0.0836**
	(0.0075)	(0.0079)	(0.0144)
Black ethnicity	0.0473**	0.0680**	0.0367*
	(0.0147)	(0.0240)	(0.0166)

B. Grant cycle renewal

	Full sample		Below median name frequency		Above median name frequency	
	Year controls	Full controls	Year controls	Full controls	Year controls	Full controls
Asian ethnicity	-0.017**	0.003	-0.043**	-0.002	-0.013†	0.001
	(0.006)	(0.004)	(0.012)	(0.009)	(0.007)	(0.006)
Hispanic ethnicity	-0.001	0.015	0.008	0.009	-0.016	0.020
	(0.013)	(0.009)	(0.016)	(0.011)	(0.022)	(0.016)
Black ethnicity	-0.102**	-0.009	-0.137**	-0.009	-0.065†	0.007
	(0.025)	(0.019)	(0.034)	(0.026)	(0.036)	(0.028)
Mean of dependent variable	0.463	0.465	0.463	0.466	0.464	0.468
Adjusted R^2	0.0344	0.3221	0.0338	0.3279	0.0357	0.3136
Nb. of investigators	32,012	31,790	16,182	15,980	15,829	15,619
Nb. of investigator-grant-cycles	92,126	91,505	46,065	45,540	46,060	45,491

Note: Each column and row representing a distinct outcome variable corresponds to the results of a separate regression. For panel A, relative disease mortality, ethnicity, and country are regressions are very similar to panels A, B, and C of Table 1, respectively. For panel B, columns 1, 3, and 5 are very similar to Table 3 column 1 while columns 2, 4, and 6 are very similar to Table 3, column 4. Columns 1 of Panel A and 1 and 2 of Panel B reproduce the results using the full sample. Columns 2 of Panel A and columns 3 and 4 of Panel B limit the sample to those investigator-grant-cycles with below median name frequency, calculated relative to the corpus of PubMed, while columns 3 of Panel A and columns 5 and 6 of Panel B limit the sample to above median name frequency. Robust standard errors were used, clustered at the investigator and grant. $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$.

Table A17. Robustness: PDS lasso

A. Ethnic research orientation

	Base estimates	PDS Lasso
Relative disease mortality		
Asian	0.1281**	0.1297**
	(0.0287)	(0.0182)
Hispanic	1.2834**	1.2277**
	(0.4330)	(0.2950)
Black	12.9128*	12.4853**
	(5.3748)	(3.3904)
Ethnicity		
Asian	0.0186**	0.0180**
	(0.0024)	(0.0016)
Hispanic	0.0269**	0.0277**
	(0.0052)	(0.0042)
Black	0.0957**	0.1023**
	(0.0180)	(0.0141)
Country		
Asian	0.0424**	0.0414**
	(0.0034)	(0.0022)
Hispanic	0.0548**	0.0555**
	(0.0075)	(0.0054)
Black	0.0473**	0.0529**
	(0.0147)	(0.0121)

B. Grant cycle renewal

	Base estimates	PDS Lasso
Grant cycle renewal		
Asian	0.003	-0.0024
	(0.004)	(0.0043)
Hispanic	0.015	0.0113
	(0.009)	(0.0090)
Black	-0.009	0.0031
	(0.019)	(0.0195)
Mean of Dependent Variable	0.465	0.463
Adjusted R2	0.3221	0.2803
Nb. of Investigators	31,790	32,007
Nb. of Grant cycles	91,505	92,068

Note: Each cell (Panel A) and column (panel B) contains an estimate from a separate regression. Column 1 reproduces estimates from Table 1 where the outcome variable is research oriented at the same ethnicity (Panel A) and Table 3, Column 4 where the outcome variable is grant cycle renewal (Panel B). Column 2 repeats this analysis, but uses the "post-double-selection" lasso estimator due to Belloni et al. (2014). Robust errors in parentheses. $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$.

Table A18. Robustness: Oster's Delta

	Oster's Delta
Relative disease mortality	
Asian	43.848
Hispanic	145.451
Black	28.241
Ethnicity	
Asian	47.093
Hispanic	10.955
Black	5.922
Country	
Asian	90.951
Hispanic	15.482
Black	10.276

Note: The score reported corresponds to the δ parameter from Oster (2019), which can be interpreted as the degree of selection on unobservables relative to observables necessary to explain away the effect of ethnicity on research topic. The δ parameter is calculated from regressions very similar to Table 1, where the outcome is the relative disease mortality, ethnic topics, and country topics of the same ethnicity. We follow Oster's recommendation of setting $R^{max} = 1.3 \times R^2$. $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$.

Appendix B
Measurement and Methods

We take advantage of MeSH terms to characterize both the direction of research and the research methodology being employed by NIH-funded investigators. MeSH terms are a hierarchical controlled vocabulary maintained by the National Library of Medicine used primarily for searching the biomedical literature. MeSH terms are assigned by professional indexers employed by the National Library of Medicine without involvement of the authors. Below we outline the MeSH terms used to characterize the diseases studied, research methodology employed, and the ethnic orientation of research. The measures of disease and research methodology follow Azoulay et al. (2021) and Greenblatt (2022).

Disease

We identifying disease-oriented articles as those under the MeSH tree disease category (C01-C26). We exclude C22, which primarily measures veterinary diseases, and include F03, mental disorders. All together, this measure includes 4,895 unique MeSH terms. Examples include *hematuria*, *aortic valve stenosis* and *Klippel-Feil Syndrome*.

Research methods

We use a set of indicator variables to identify the research methodology being employed by an investigator during the grant cycle. This includes indicators for bench science research, translational research, clinical trial research, and other clinical studies.

Our measure of bench science research is focused on identify the tools of bench biomedical research by using a molecular biology method or model organism. An article is considered bench science if it has at least one MeSH term in either of these categories. An article is considered translational research if it is uses the tools of bench science but is also tagged with at least one disease MeSH term.

Articles using molecular biology methods were identified primarily based on the MeSH category for investigative techniques (F05). These were manually reviewed to eliminate any terms that may have a potential direct clinical application outside of a laboratory, such as *angioplasty* or *Glasgow coma scale*. The final list of MeSH codes includes: E05.017, E05.091, E05.118, E05.181, E05.196 (but excluding E05.196.353), E05.197, E05.198, E05.242.223, E05.242.335, E05.242.363.342, E05.242.373, E05.242.383.910, E05.242.551, E05.242.800, E05.295, E05.301, E05.313, E05.318.416, E05.393, E05.478, E05.481, E05.484, E05.490, E05.522, E05.588, E05.591, E05.595, E05.598, E05.601, E05.624, E05.650, E05.657, E05.830, E05.916.680, and A11.251 (excluding A11.251.476). Together, these codes identify 510 unique MeSH terms with examples including *immunoelectrophoresis*, *nucleic acid hybridization*, and *real-time polymerase chain reaction*.

To capture use of model organisms, we compiled a list of 53 different model organisms used in biomedical research: *tobacco mosaic virus, bacteriophage lambda, bacteriophage phiX174, SV40, T4 phage, Escherichia coli, Bacillus subtilis, Caulobacter crescentus, Aliivibrio fischeri, Synechocystis, Pseudomonas fluorescens, Azotobacter vinelandii, Streptomyces coelicolor, Chlamydomonas reinhardtii, Dictyostelium discoideum, Tetrahymena thermophila, Eremothecium, Aspergillus nidulans, Coprinus cinereus, Cryptococcus neoformans, Ceurospora crassa, Saccharomyces cerevisiae, Schizophyllum, Schizosaccharomyces pombe, Ustilago maydis, Arbacia punctulata, Aplysia, Caenorhabditis elegans, Ciona intestinalis, Drosophila, Loligo pealei, Trichoplax adhaerens, Ambystoma mexicanum, cat, chicken, dog, Mesocricetus auratus, guinea pigs, rabbits, Oryzias latipes, mice, genetically modified animals* (B01.050.050.136 and B01.050.050.199.520), *mole-rat, pigeon, Poecilia reticulata, rat, rhesus macaque, Petromyzon marinus, Takifugu, Xenopus laevis,* and *zebrafish*. The associated MeSH codes for these organisms resulted in 125 unique MeSH terms.

To measure clinical trial research, I combine three different measures, with an emphasis on capturing phase 1-4 trials. I start with using the following MeSH codes: V03.175.250.100, E05.318.372.250.250.200, V03.175.250.200, E05.318.372.250.250.210, V03.175.250.300, E05.318.372.250.250.220, V03.175.250.400, E05.318.372.250.250.230, and E05.337.800.200. This identifies 8 distinct MeSH terms, with examples including *phase 2 clinical trial* and *phase 4 clinical trial*. I supplement this approaching using the four article publication types corresponding to phase 1-4 clinical trials (e.g., *clinical trial, phase 1*). Finally, articles with titles including the phrases "phase 1," "phase 2," "phase 3," or "phase 4" were also included. These

were reviewed to exclude topics that might be picked up by these search terms but were unrelated to clinical trials, such as those related to bundle branch blocks, blood pressure, and cell depolarization.

We also identify other clinical studies, to capture non-clinical trial clinical research. This includes case reports, epidemiologic studies, and observational studies.

Case reports were identified by publications tagged with the MeSH code V03.100 corresponding to the single MeSH term *case reports [publication type]*. Articles were also considered case reports if their title included the phrase "case report" or if they were the publication type *case reports*.

Epidemiologic studies were characterized by the following MeSH codes as well as all codes under them in the hierarchical tree: E05.318.372.500, N05.715.360.330.500, and N06.850.520.450.500. In total, this identifies 12 MeSH terms. Examples include *cohort studies*, *case-control studies*, and *seroepidemiologic studies*. Care was taken to exclude terms reflecting a specific study or patient population rather than a general epidemiologic method, such as excluding the MeSH for *National Longitudinal Study of Adolescent Health* and *HIV Seroprevalence*. This measure was expanded to also include any papers with one the following title key words: "case-control," "retrospective," and "cohort."

Observational studies includes those with the MeSH codes E05.318.372.250, E05.318.372.658, N05.715.360.330.250, N05.715.360.330.250.500, N05.715.360.330.643, N06.850.520.450.250, N06.850.520.450.250.500, N06.850.520.450.643, V03.175 (but not V03.175.750), and V03.175.500. Examples include *clinical study, observational study*. We also included articles with a publication type of *clinical study* or *observational study*. We exclude from this measure any clinical trial, epidemiologic study, or case report as identified above.

Ethnic-orientation of research

We first identify MeSH related to ethnicity. The MeSH codes for white ethnicity are M01.686.877, M01.686.445, M01.686.461 (but not M01.686.461.500), and M01.686.477.625.594.812. Examples include *white*, *white people*, and *European people*. The MeSH codes for Asian ethnicity include M01.686.330, M01.270.906, M01.686.477.625.594.563, M01.686.477.625.594.625, and M01.686.493.500.800. Examples include *Asian*, *Asian people*, and *Asian American Native Hawaiian and Pacific Islander*. The MeSH codes for Black ethnicity include M01.270.500, M01.686.254, M01.686.372, M01.686.461.500, and M01.686.477.625.594.594. Examples include *Black people*, *Black or African American*, *African people*. The MeSH codes for Hispanic ethnicity include M01.270.953, M01.686.429, M01.686.477.625.188.441, and M01.686.685. Examples include *Hispanic or Latino*, *Central American people*, and *South American people*.

We also identify MeSH related to country and geographic location. We consider countries related to white ethnicity to include Z01.107.567.176, Z01.107.567.403, Z01.107.567.875, Z01.252.100, Z01.252.122.500, Z01.252.245.500, Z01.542, Z01.542.248.775, Z01.542.816.124.500, Z01.639.100, Z01.639.400, Z01.639.760.747, Z01.678.100.373, and Z01.678.100.747 but not Z01.542.727 or Z01.542.846. Examples include *Australia, Belgium, Europe,* and *Canada*. We consider countries related to Asian ethnicity to include Z01.252.145, Z01.252.245.782, Z01.252.474, Z01.639.760, Z01.678.100.873 but not Z01.639.760.747 or Z01.678.100.747. Examples include *China*, *Eastern Asia*, *India*, and *Thailand*. We consider countries related to Hispanic ethnicity to include Z01.107.084, Z01.107.169, Z01.107.424, Z01.107.567.589, Z01.542.727, Z01.107.757, and Z01.542.846. Examples include *Central America*, *Dominican Republic*, *Guatemala*, and *Mexico*. We consider countries related to Black ethnicity to include Z01.058. Examples include *Africa*, *Ghana*, *Kenya*, and *Zimbabwe*.

We also look at a set of measures examining ethnicity-adjacent topics. We consider minority topics to include I01.880.853.125, I01.880.853.300.437, and N01.400.512. Examples include *ethnic and racial minorities*, *minority health*, and *minority groups*. We consider topics related to disparities to include I01.240.425.513, M01.270 (but not M01.270.976), N01.224.425.394, and N06.850.505.400.425.675. Examples include *health inequities*, *health status disparities*, *socioeconomic disparities in health*. We consider topics related to socioeconomic status to include I01.240.425.756, I01.880.735.634, I01.880.853.996, M01.189, M01.920, and M01.987. Examples include *poverty*, *housing instability*, and *low socioeconomic status*, and *social determinants of health*.

Finally, we review the medical literature to identify nineteen diseases which have a strong ethnic association based on genetics, geographical exposure, or epidemiologic associations. This includes Tay-Sachs disease, cystic fibrosis, and hereditary hemochromatosis for white ethnicity; hepatitis B, hepatitis C, gastric cancer, tuberculosis and lactose intolerance for Asian ethnicity; dengue, leishmaniasis, Chagas disease,

Hermansky-Pudlak syndrome for Hispanic ethnicity; and sickle cell disease, glucose-6-phosphate-dehydrogenase deficiency, sarcoidosis, trypanosomiasis, Lassa fever, Ebola, malaria for Black ethnicity.

References

Greenblatt, Wesley. 2022. "Anchoring Exploration: Knowledge Certification and the Dynamics of Innovation." SSRN working paper 3919866.