Race and Science

Gaia Dossi*

June 27, 2024

[Most recent version here]

Abstract

What are the consequences of the racial gap in science and innovation? I study this question by combining data on US patents, medical research articles, clinical trials, and research grants with the racial distribution of last names in the US population. Using last names as a proxy for race, I find that the racial composition of scientists affects the direction, as well as the rate, of medical research and innovation. First, Black scientists are three times as likely to design clinical trials with Black participants and twice as likely to publish articles focused on Black individuals. Second, Black scientists are more likely to research diseases frequent in the Black population, and white scientists in the white population. Third, I draw a link between race and the direction of research by focusing on diseases more common in Black individuals (e.g., sickle cell anemia) or white individuals (e.g., melanoma) due to evolutionary advantages in their ancestors' countries of origin. Fourth, I document the impact of relative disease incidence on the direction of research by studying an exogenous change in HIV-related mortality among Black compared to white Americans. I estimate a general equilibrium Roy model with racial frictions and endogenous choice of occupation. Using the data, I quantify the parameters and estimate that removing barriers would increase the overall number of inventors by 1 p.p., a 10% increase from the baseline.

JEL codes: J15, J24, I10, O31
Keywords: Innovation, Science, Racial gap

*Department of Economics and Centre for Economic Performance (CEP), London School of Economics and Political Science. Email: g.g.dossi@lse.ac.uk. I am grateful to John Van Reenen, Oriana Bandiera, and Xavier Jaravel for their invaluable mentoring, advice, and encouragement, and to Livia Alfonsi, Marcella Alsan, Mike Andrews, Nava Ashraf, Hadar Avivi, Marco Bellifemine, Enrico Berkes, Shai Bernstein, Fabio Bertolotti, Kaveh Danesh, Jeremiah Dittmar, Patrick Gaulé, Bob Gibbons, Paola Giuliano, Erik Hurst, Ethan Ilzetzki, Marlene Koffi, Jon Kolstad, Rem Koning, Peter Lambert, Camille Landais, Gabriel Leite-Mariante, Alan Manning, Will Matcham, Virginia Minni, Thomas Monk, Marta Morando, Marlene Koffi, Tom Nicholas, Will Parker, Torsten Persson, Steve Pischke, Paola Sapienza, Mark Schankerman, Johannes Spinnewijn, Fabian Waldinger, Bruce Weinberg, Alwyn Young, and participants to various seminars and conferences for feedback and helpful discussions. Joaquin Urria provided outstanding research assistance. I gratefully acknowledge financial support from POID and STICERD. All errors are my own.

1 Introduction

Black Americans are underrepresented in every scientific and innovation field. Compared to white Americans, they are one-fourth as likely to be scientists (NSF, 2015), and one-eighth as likely to hold a patent (Akcigit and Goldschlag, 2023). The underrepresentation of minorities in innovation is a salient issue in US policy. In 2022, Congress passed the "Unleashing American Innovators Act" encouraging the United States Patent and Trademark Office (USPTO) to implement outreach initiatives aimed at diversifying the inventor pool. However, to date, we know surprisingly little about the economic and distributional implications of this large gap.

The potential costs of minority gaps in scientists and inventors fall into two categories. First, private costs may emerge if the benefits of patent ownership are not shared equally among the population. Second, societal costs may arise if the lack of minority inventors not only leads to unequal opportunity, but also to economic and welfare losses for society at large. Recent research suggests that reducing barriers to occupational choice, and therefore minority gaps, anywhere in the economy will yield positive implications for growth (Hsieh, Hurst, Jones and Klenow, 2019; Bell, Chetty, Jaravel, Petkova and Van Reenen, 2019*a*,*b*).

In this paper, I provide novel evidence on the implications of the racial composition of scientists and inventors for the production of science. I document that it has implications for the direction of research and innovation, as well as its quantity and quality.

Addressing this question poses several challenges. The first is lack of data, as the demographic characteristics of inventors are not collected by patent offices, nor do inventors fall under any standard occupational category in labor survey data. The second challenge lies in identifying race separately from socio-economic status, as the two concepts are highly intertwined in the United States (e.g., Rose, 2023). The third challenge lies in studying whether innovation produced by Black or African Americans is the same as the one produced by white Americans. Previous research has found that female, wealthy, and older inventors tend to target their research towards their own demographic group. To document this pattern, these studies classify innovations by studying exclusively female diseases (Koning, Samila and Ferguson, 2020, 2021), consumption patterns by gender, age, and socio-economic status (Einiö, Feng and Jaravel, 2023). However, race is a social construct, not a biological one. For this reason, defining a measure of innovation specifically addressing Black or African Americans is challenging: classifying diseases based on race is not obvious, and consumption patterns may be driven by the social environment, rather than by race-specific factors.

In this paper, I measure race using information on the aggregate frequency of last names by race and ethnicity from the 2010 US Census (Comenetz, 2016). I merge this data with medical research articles from the PubMed database, ongoing clinical trials from the web portal `ClinicalTrials.gov`, research grants awarded by the National Institutes of Health (NIH), and applications and granted patents by the United States Patent and Trademark Office (USPTO). Because the focus of this paper is racial inequality in the United States and uses a the proxy for race relies on racial frequencies across US residents, I restrict the sample to researchers and Principal Investigators whose main affiliation is with a US institution and to inventors residing in the US.

I begin by documenting that having a Black-sounding name is correlated with a lower probability of being a scientist or inventor. An individual with a Black-sounding name is five times less likely to be a scientist or inventor compared to someone with a white-sounding name. How does this gap affect the production of science?

In the first part of the paper, I document that the racial composition of inventors shapes the direction of research and innovation. I focus on medical innovation, measured through research articles and the patents linked to them, to construct two measures of *Innovation directed towards Black or African Americans*. The first measure focuses on research and innovation that explicitly addresses Black or African Americans (I label this the "Demographics-based" approach). The second measure focuses on research on diseases with relatively higher incidence among Black or African Americans compared to white Americans (I label this the "Frequency-based" approach). These two measures provide complementary ways to classify the content of scientific production. The first is more direct, but more narrow, as only 1% of all articles in the dataset mention any demographic group. The second measure is less direct, but broader: over one third of all articles in the dataset can be classified using this approach.

To construct the first measure, I focus on research explicitly addressing Black or African Americans. I document that researchers with a Black-sounding last name are more than twice as likely compared to those with White-sounding last name to produce research directed towards Black or African Americans. This is largely driven by Black-sounding researchers publishing more articles reporting the results of clinical trials that include Black or African Americans. Turning to data on *ongoing* clinical trials, a similar pattern emerges: compared to scientists with a White-sounding last name, those with a Black-sounding one are almost four times as likely to include Black or African Americans in trials. I run placebo exercises and I find that Black-sounding researchers are not more likely than white-sounding ones to

conduct research addressing other demographic groups (such as Hispanic or Latinos, or Asian Americans), and they are significantly less likely to publish articles on demographic groups other than their own. This finding provides a novel explanation for the lack of inclusion of Black and African Americans in medical research and complements existing findings that point to the role of low enrolment in clinical trials (Alsan, Durvasula, Gupta, Schwartzstein and Williams, 2022), and low research funds dedicated to typically Black diseases (Farooq, Mogayzel, Lanzkron, Haywood and Strouse, 2020).

To construct the second measure, I build an index of relative mortality of Black or African Americans compared to white Americans using administrative data from the Center for Disease Prevention (CDC). A relative mortality index equal to 1 means that a Black individual and a white individual residing in the US are equally likely to die of that disease. I find that Black scientists are more likely to research and commercialize research on diseases with higher mortality among Black or African Americans. Symmetrically, white scientists are more likely to focus on diseases with higher mortality among white Americans. The magnitudes are large: compared to white researchers, Black or African American researchers focus on diseases that have a relative incidence that is 12% higher on average.

After documenting that Black scientists are more likely to conduct research directed towards Black or African Americans, I use two research designs to identify the drivers of these patterns. With the first design, I draw a causal link between race and the direction of research. This allows for disentangling other factors correlated with race such as the local environment or socio-economic status. With the second design, I exploit an exogenous shift in HIV-related mortality among Black Americans compared to white Americans, and I draw a causal link between race-specific mortality and the match between researchers and the content of their innovation.

With the first research design, I develop a novel approach building on findings in medicine and anthropology. I study conditions that gave an evolutionary advantage to ancestors and persist today, even if the advantage no longer holds. More specifically, I focus on genetic conditions that provided an advantage in the environment where ancestors were located. I focus on two case studies: sickle cell anemia and melanoma. Sickle cell anemia is more common among individuals of African ancestry because being a carrier of the condition protects against malaria. Melanoma is more common among white individuals as this condition is triggered by exposure to UV lights, from which Black Americans are more shielded thanks to the darker skin color. I find that scientists are more likely to focus on diseases more common in their own demographic group, while I find no correlation with research in similar diseases

with balanced incidence among demographic groups. This association is large: across all research articles, scientists with a Black-sounding last name are 50% more likely compared to those with a White-sounding last name to research sickle cell anemia. Within specific subfield, they are over twice as likely to research this disease compared to white scientists. Under the assumption that the incidence of these diseases varies with ancestry but not current socio-economic status, this approach allows to draw a causal link between race and the content of the research (not only via factors linked to socio-economic status, or the local environment).

With the second research design, I identify the impact of relative disease incidence on the direction of innovation by studying a shock to the relative mortality rate of Black Americans compared to white Americans. I build on evidence documenting that Human Immunodeficiency Viruses (HIV) became a higher contributor to the mortality of Black Americans (compared to white Americans) after the introduction of the Highly Active Antiretroviral Therapy (HAART) in 1996 (Levine, Briggs, Kilbourne, King, Fry-Johnson, Baltrus, Husaini and Rust, 2007). The evidence shown so far would predict that when HIV becomes a "Relatively More Black" disease, then the relative share of scientists with a Black-sounding last name researching the disease should go up. I test this hypothesis using data on the universe of research grants awarded by the National Institutes of Health between 1988 and 2012. Grant data is especially suited to study how researchers react to policy shocks because of its more immediate reflection of research activities. First, I find that HIV-related articles become five times more likely to mention Black Americans compared to the pre-period. I interpret this finding as a "first-stage" showing that HIV became a "relatively Black" disease. Turning to the direction of research, I find that Black researchers are 50% more likely to focus on HIV compared to whites, while there was no such pattern in the period before the introduction of HAART. A similar pattern holds when looking at research publications.

In the second part of the paper, I study how the racial composition of scientists and inventors affects the quantity and quality of scientific production. The large racial gaps suggest that there may be many "missing" scientists and inventors, thus affecting the quantity of the overall production of science and innovation. (Hsieh, Hurst, Jones and Klenow, 2019; Bell, Chetty, Jaravel, Petkova and Van Reenen, 2019*a*,*b*; Bloom, Van Reenen and Williams, 2019). Looking within the patenting process, I find evidence that racial gaps may also affect the the quality of science and innovation. I find that patent applications from Black inventors are granted 6% less often than patents from white inventors, both unconditionally and after controlling for detailed technology class fixed effects. Approximately half of this gap is explained by differential access to resources. After controlling for assignee fixed effects, state

of residence fixed effects, and proxies of lawyer quality, the residual gap is equal to 3%. Looking at granted patents, the gap is reversed. Patents from Black-sounding inventors have a higher impact (as measured by forward citations) compared to white-sounding ones. These findings suggest that minority gaps among scientists and inventors may result not only in lower quantity, but also in lower quality of innovation.

In the final part of the paper, I quantify how closing racial gaps would affect the quantity and quality of innovation through the lens of the general equilibrium Roy model developed by Hsieh, Hurst, Jones and Klenow (2019). Closing minority gaps may have general equilibrium effects for at least two reasons. First, if minorities anticipate discrimination, they may choose different careers. Second, as more individuals enter an occupation, the labor demand for that occupation will change, causing equilibrium wages to decrease. In the model, individuals are either white or Black and choose among a set of occupations based on heterogeneous preferences, heterogeneous talent, and group-specific preferences. The economy is composed of the home sector, standard market occupations, and the "science and innovation" sector. Black individuals face a "minority tax" in the acquisition of human capital and another tax in the labor market. I estimate the model using moments from the Current Population Survey (CPS), and externally calibrated moments from Hsieh, Hurst, Jones and Klenow (2019) to quantify the magnitude of this loss. I find that equalizing access to science and innovation will lead to an overall increase of 1 p.p., corresponding to 10% over baseline, in the number of scientists and inventors. This increase is driven by a higher number of Black inventors and scientists selecting into these occupations. The share of white inventors and scientists experiences a decrease, albeit quantitatively small (less than 1% over baseline).

Related Literature This paper is closely related to three strands of literature. First, it contributes to the literature showing that inventor demographics matter for the direction of innovation. Einiö, Feng and Jaravel (2023) document a match between gender, socio-economic status and age of inventors and entrepreneurs and their innovation. Koning, Samila and Ferguson (2020, 2021) documents that female researchers are more likely to carry out research on female diseases. A growing body of literature shows that interaction with diverse students peers (Truffa and Wong, 2022), representation in clinical trials (Michelman and Msall, 2023), access to data (Nagaraj, Shears and de Vaan, 2020), and geography (Moscona and Sastry, 2022; Fry, 2023) increase the representation in the focus of scientific research. Additionally, a growing body of evidence shows the impact of the location of inventors on the direction of their research. For example, Fry (2023) shows that the outbreak of a health crisis sparks research from local researchers on those diseases. I contribute to this literature

by documenting a link between the race of an inventor (scientist) and the content of their innovation (research).

Second, this paper contributes to the literature on the costs of talent misallocation. Hsieh, Hurst, Jones and Klenow (2019) show that lower barriers in accessing the labor market for minorities explain 40% of GDP growth between 1960 and 2010. Chetty, Dossi, Smith, Van Reenen, Zidar and Zwick (2023) uses a similar model to quantify the welfare gains from equalizing the access to entrepreneurship for women in the United States in the last two decades. Ashraf, Bandiera, Minni and Quintas-Martınez (2023) find that the misallocation of female talent across countries harms the productivity of firms. I contribute to this literature by showing that inequality in the direction of innovation and science is an additional consequence of misallocation.

Third, this paper contributes to the literature on the underprovision of medical research and innovation benefiting Black Americans. A set of studies has shown that low representation in clinical trials (Alsan, Durvasula, Gupta, Schwartzstein and Williams, 2022; Alsan, Campbell, Leister and Ojo, 2023) and low funding (Farooq, Mogayzel, Lanzkron, Haywood and Strouse, 2020) are drivers of this underprovision. I add to this literature by showing that the racial composition of scientists and inventors contributes to racial gap in medical research and innovation.

The rest of the paper is organized as follows. In section 2, I describe the data. In section 3, I study the link between the racial composition of scientists and inventors and the direction of research and innovation. In section 4, I study the mechanisms driving the match between researchers and the content of their research and innovation. In section 5, I study the link between the racial composition of scientists and inventors and the quantity and quality of research and innovation. In section 6, I estimate the labor market effects of equalizing access to innovation and science using a Roy model of occupational choice. In section 7, I discuss these findings. In section 8, I conclude.

2 Data

In this section, I describe the data sources and the construction of the variables used in the empirical analysis. In section 2.1, I discuss data on scientific publications and research scientists. In section 2.2, I discuss data on patents and inventors. In section 2.3, I discuss how I proxy the race of scientists and inventors. In section 2.4, I discuss the matching procedure of scientists and inventors with the measure of Black-sounding name. In section 2.5, I discuss

how I validate the measure of Black-sounding name as proxy of ethnicity. Sample statistics are reported in Table 1.

2.1 Measuring Scientific Production

To measure scientific production, I rely on three data sources. The first is research articles from the PubMed database. PubMed is a free search engine accessing the MEDLINE database of articles on life sciences and biomedical topics. It includes bibliographic information for articles from academic journals covering medicine, nursing, pharmacy, dentistry, veterinary medicine, and health care. It also covers much of the literature in biology and biochemistry, as well as fields such as molecular evolution. For each article, this dataset provides information on title, abstract, MeSH terms, journal, year of publication, publication type, institutional affiliation of first author. I complement this data with information on citations from the iCite database of the National Institutes of Health (NIH) (Hutchins, Yuan, Anderson and Santangelo, 2016). I link this data to the patents that cite these papers Marx and Fuegi (2020, 2022). Following Koning, Samila and Ferguson (2020, 2021), I restrict the main analysis to research articles in the top 1,000 journals in terms of their commercialization impact factor (JCIF) (Marx and Fuegi, 2020, 2022). For all datasets, I restrict the sample to years 2002 to 2018. Additionally, I restrict the data to researchers affiliated with an institution located in the United States.[1] Second, I use data on ongoing clinical trials from `ClinicalTrials.gov`, a registry run by the United States National Library of Medicine (NLM) at the National Institutes of Health. The third dataset is the universe of research grants awarded by the National Institutes of Health (NIH) between 1985 and 2012. Data on the universe of National Institutes of Health grants awarded between 1985 and 2012 comes from the NIH database.[2]

2.2 Measuring Patenting Activity

To measure patenting activity, I use data on the universe of patent applications filed in the United States between 2001 and 2018. The data were extracted from the Patent Examination Research dataset (PatEx) merged with the PatentsView dataset. The first dataset (PatEx) is

[1] Throughout, I restrict the sample to the first author of the article, grant, or clinical trial. For this reason, I impose that affiliation of the *first* author must be with a U.S. institution.

[2] The NIH is the largest funder of medical research in the United States after the U.S. Government. The National Science Foundation (NSF) documented in a 2015 report that, of the 86 billion spent on basic research in 2015, 44% came from federal agencies, 28% came from the National Institutes of Health, 29% came from pharmaceutical companies, 19% came from biotechnology companies.

maintained by the USPTO and contains detailed information on all patent applications filed since January 2001 (Graham, Marco and Miller, 2015). The second (PatentsView) is also maintained by the USPTO, but focuses on data linking inventors, their organizations, and locations. To construct the final sample, I include only applications filed between January 2001 and December 2018. Patents before 2001 are excluded because data on applications is not available before then.[3] I stop at December 2018 to allow enough time for grant decisions to be made, and I further exclude cases on which decisions have not been made, therefore excluding provisional, reissue, and Patent Cooperation Treaty applications. In this paper, I focus on the United States, and for this reason I restrict my sample to inventors who are resident in the US at time of filing.[4]

2.3 Measuring Race and Ethnicity

I proxy the race and ethnicity of scientists and inventors using their last names. Inferring unobserved characteristics through names is a procedure widely used in research on the ethnicity of inventors (Kerr, 2008*a*,*b*; Gaule and Piacentini, 2013) and in the economic history literature.[5] In this paper, I rely on last names, rather than first names, for two reasons. First, because they are more stable over time for the same individual, as well as more stable across cohorts. This is a crucial aspect as the age distribution of inventors is different from the one of the wider population. Second, for data availability: thanks to a recent initiative of the U.S. Census Bureau, we now have available data on the population frequency of each last name in the United States. For first names, much smaller and less representative datasets are instead available.[6] This approach follows best practices on how to infer ethnicity for the U.S. population (Kozlowski, Murray, Bell, Hulsey, Larivière, Monroe-White and Sugimoto, 2022).

I use aggregated data published by the Census Bureau (Comenetz, 2016) on the frequencies of last names by race from the 2010 U.S. population Census. These data report the probability

[3]The American Inventors Protection Act passed in 1999 made it mandatory to publish all non-provisional patent applications after December 2000.

[4]Information on the place of residence is listed on the patent, along with the name and last name of all inventors, the name of patent attorneys, and the name of the patent examiner who was assigned the patent application.

[5]Names have been used, among others, to measure race and ethnicity (Abramitzky, Boustan and Connor, 2020; Fouka, 2020), individualism (Bazzi, Fiszbein and Gebresilasse, 2020), socioeconomic background (Olivetti, Paserman, Salisbury and Weber, 2020), and religiosity (Berkes, Coluccia, Dossi and Squicciarini, 2023).

[6]To the best of my knowledge, the most comprehensive data on the racial and ethnic distribution of first names in the U.S. comes from Tzimioukis (2015) and is constructed on a sample of 8 million mortgage applicants.

that a given last name in the U.S. population is associated with a given race. It is provided for the entire population excluding those individuals whose last name appears in fewer than 100 records. This is the Bayes Optimal Classifier, that is the solution that minimizes the chance of misclassification of the race variable in the population. If you picked a random individual with last name *Washington* from the U.S. in 2010 and asked U.S. to guess this person's race (measured as crudely as by the Census), the best guess would be based on what is available from the aggregated Census file.

In the Census dataset, each last name is linked to six racial probabilities: white, Black, Hispanic, Asian or Pacific Islander, Native American, Two or more races. Empirically, these probabilities do not always add up to 1 as the U.S. Census Bureau censors all cells with fewer than 100 last names. For each last name, I re-standardize probabilities such that they add up to 1 and I split the "Two or more races" category equally across the other racial probabilities.

In the main analysis, I assign race or ethnicity based on the first researcher listed on the paper, or on the first inventor listed on the patent. In robustness checks (reported in the Appendix), I verify that results hold using alternative measures built leveraging data on all members of the team, or of both the first and the last listed authors.

2.4 Matching Procedure

To build the final dataset, I match data on scientists described in section 2.1 and on inventors described in section 2.2 with the distribution of last names by race described in section 2.3. To avoid false positives, I adopt a conservative approach in conducting the merge, keeping only *exact* matches. With this approach, I merge 78% of all research articles and 82% of all patents. The match is performed based on the last name of the first author. Sample statistics for the final sample are detailed in Table 1.

In Appendix Table B3 and B4, I report a test of balancedness for the match. Matched and unmatched observations are remarkably balanced across all demographics except geography: matched observations (both for scientists and for inventors) are significantly less likely to be located in California compared to the matched sample (25% compared to 31%). This is consistent with the hypothesis that California may have relatively more foreign-born inventors, whose last name may not be recorded among Census records.

In Figure 1, I plot the distribution of the US population over Pr(Black | Last name). Over 55% of individuals have a last name with a probability between 0 and 0.1 of being held by

a Black or African American individual. In contrast, when examining the distribution of scientists and inventors (shown in grey), this fraction increases to 70%. This large difference is more clearly illustrated in Figure 2. In this plot, I divide the population into quintiles based on Pr(Black | Last name) and normalize them by the first quintile. If scientists and inventors were represented proportionally to the overall population, all dots would lie on the red line. However, this figure tells a different story. Individuals in the top quintile of Pr(Black | Last name) are over five times less likely to be scientists or inventors.

2.5 Validation of Black-sounding Last Names as a Proxy for Race

To validate this data, I use a dataset of voter registration data matched to race of individuals who are registered to vote in the state of Florida from Dossi and Morando (2023). In Appendix Table B2, I report the individual-level correlation of Black-sounding name with a dummy = 1 if the individuals self-reported being Black or African American in the Voter Registration data. In colum (2), I show that the correlation is similar (and the R-squared as well) when including the vector or other racial probabilities.[7] In columns (3), I include a control for median income in zip code of residence. While the correlation decreases, it remains statistically significant and high. This result suggests that the vector of racial probabilities correlated with race even when conditioning for an individual-level measure of socio-economic status. In columns (4) and (5) I control for average median income by last name and its standard deviation. The association between Black-sounding name and self-reported race remains positive, significant, and close to 1.In Appendix Figure B1, I use the full Voter Florida data which consists of 9 million individuals to show that Pr(Black | Last name) in Florida has a 1:1 correlation to the one in the United States. I obtain a similar picture when building a last name-level vector of racial probabilities and correlate the two. This exercises lends confidence to the validity of Pr(Black | Last name) as a predictor of race even in subsamples of the United States (i.e., a single state such as Florida). Second, I test this correlation on the sample of scientists and inventors. I construct a vector of racial probabilities by last name in the subsample of, respectively, scientists and inventors. In Appendix Figure B2, I show the correlation of Pr(Black | Last name, scientist) and Pr(Black | Last name, inventor) with Pr(Black | Last name). The relationship is liner. However, the slope is lower than 1: this is consistent with the fact that the overall share of Black or African Americans is lower in these subsamples.

[7] White-sounding name is the omitted category.

3 Race and the Direction of Science and Innovation

In this section, I document that the race of scientists correlates with the content of their research. In section 3.1, I summarize the evidence on research addressing Black or African Americans from the existing literature. In section 3.2, I describe how I construct the indicators of innovation directed towards Black individuals. In section 3.3, I show results using the demographics-based approach, and in section 3.3, I show results using the frequency-based approach.

3.1 The Black-white Gap in Medical Research and Innovation

There is a large gap in health outcomes and life expectancy between Black and white Americans.[8] Social determinants of health, such as education, income, and access to the healthcare system, are documented drivers of these disparities (e.g., Cutler, Deaton and Lleras-Muney, 2006; Cutler, Lleras-Muney and Vogl, 2008; Chetty, Stepner, Abraham, Lin, Scuderi, Turner, Bergeron and Cutler, 2016; Schwandt, Currie, Bär, Banks, Bertoli, Bütikofer, Cattan, Chao, Costa, González et al., 2021). In addition to these factors, a growing body of evidence documents that research and innovations in healthcare may create or perpetuate these inequalities.

Do research and medical advances benefit Black or African Americans less often compared to white Americans? Answering this question is complicated because it is hard to determine what the optimal rate of innovation should be, and who in the population it should target (Bryan and Williams, 2021). In this section, I summarize some of the existing evidence suggesting that medical research and innovation benefits Black or African Americans less often compared to other groups in the population. As benchmark rates, I rely on the proportionality approach often used in public health, according to which it is optimal to target research efforts in a way proportional to the size of the population affected (unconditional of the demographic group).[9][10]

First, a large body of evidence documents that Black or African Americans are underrepresented in clinical trials compared to population size. The lack of diversity can significantly

[8] In 2018, the gap in life expectancy between Black and white Americans was 3.6 years (Schwandt, Currie, Bär, Banks, Bertoli, Bütikofer, Cattan, Chao, Costa, González et al., 2021). This difference has widened after the COVID-19 pandemic.

[9] Importantly, this approach may lead in itself to health inequality due to the smaller share of Black or African Americans compared to white Americans (Cutler, Meara and Richards-Shubik, 2012).

[10] For example, the NIH allocates research funds in accordance with disease burden (Farooq, Mogayzel, Lanzkron, Haywood and Strouse, 2020).

affect public health as the results of these studies may not accurately represent the broader population. This can result in drugs and treatments that are less effective or even harmful to certain groups of people. One example includes the use of certain cardiovascular medications, which have been found to produce clinical differences between white individuals of European ancestry and individuals of African ancestry (Johnson, 2008).[11]

Second, Black or African Americans are underrepresented in observational studies, both in research articles and in medical textbooks. This has real implications for health outcomes. For example, Louie and Wilkes (2018) finds that no textbook has images of six common skin cancers in skin of color. In turn, a recent study among medical students found a very high rate of underdiagnosis of skin conditions among Black individuals (Fenton, Elliott, Shahbandi, Ezenwa, Morris, McLawhorn, Jackson, Allen and Murina, 2020), plausibly due to lack of training and resources to learn how to diagnose these conditions.[12]

Third, diseases with higher incidence among Black or African Americans compared to whites are under-researched compared to similar diseases. Making this statement is complicated by the lack of systematic data on the incidence of diseases. However, suggesting (yet compelling) evidence on this comes from analyzing a genetic disease more frequent among Black individuals: Sickle Cell Anemia. The number of publications on Sickle Cell Anemia proportional to the affected population is almost five times smaller compared to cystic fibrosis, a genetic disease comparable to Sickle Cell Anemia but more frequent among whites (Farooq, Mogayzel, Lanzkron, Haywood and Strouse, 2020).

Fourth, recent evidence shows that artificial intelligence and healthcare algorithms are biased against Black patients (e.g., Obermeyer, Powers, Vogeli and Mullainathan, 2019). This is be both a *consequence* of lower research output on Black patients, hence a smaller training set compared to white ones and a *cause* of health inequality. However, it represents an additional way through which innovation in healthcare leads to inequality.

3.2 Measuring the Direction of Science and Innovation

In the rest of this section, I test the hypothesis that Black scientists and inventors located in the United States are more likely to produce medical research and innovation that benefits

[11] Several clinical trials have shown varying responses to the blood pressure-lowering effects of beta-blockers and ACE inhibitors.

[12] Another example is the utilization of the pulse oximeter. This is a medical device that measures blood oxygen saturation by passing light through the skin, typically on a fingertip. This device overestimates oxygen saturation in individuals with darker skin tone, leading to under-diagnosis (Sjoding, Dickson, Iwashyna, Gay and Valley, 2020).

Black or African Americans.

The first challenge in studying the correlation between race and the direction of innovation is defining which innovations benefit Black or African Americans. I propose two definitions of "Innovation directed towards Black or African Americans". The first is based on whether research explicitly addresses Black or African Americans as a demographic group. The second is based on the relative incidence of disease among Black or African Americans compared to white Americans. I focus on research articles and the patents linked to them as a measure of commercialization of the ideas introduced in these articles (Marx and Fuegi, 2020, 2022).

Linking race to the content of medical research is not straightforward because race is a social construct, rather than a biological one. However, it is widely documented in the medical literature that certain conditions have differential incidence conditional on race. This is driven both by environmental factors and by small genetic variations in our genetic heritage that link back to ancestors. To define diseases common among the Black or African American population, I propose two different strategies: a demographics-based approach, and a frequency-based approach. The first, which I label "demographics-based", defines research as directed to Black American if the research explicitly mentions Black or African Americans. The second, which I label "frequency-based", defines a disease as typically Black if its incidence is relatively higher among Black Americans than it is among white Americans. Symmetrically, I define typically-white diseases.

With the first definition, I define an indicator of innovation directed towards Black or African Americans coding a binary variable that takes value 1 if a research article is associated with the MeSH code corresponding to *Black or African American*, and takes value 0 otherwise. MeSH codes may vary over time, and are not updated retroactively. This should be a smaller concern in this context as I look at a relatively short time span. However, to address this I map articles to diseases using a dictionary approach based on the abstract of each publication. The results, shown in the Appendix, are qualitatively unchanged and quantitatively similar using the dictionary approach.[13]

This approach captures a direct link between the race of the researchers and the subject of their research. However, its main disadvantage is that it is fairly narrow: only approximately 1% of all articles are explicitly linked to a demographic group.

[13]In Appendix Figure B4, I report an example of an article with Black or African American among its MeSH codes, but which is not classified as 1 by the dictionary classification. In Appendix Figure B5, I report an example of an article which is classified as 1 by the dictionary classification, but does not have Black or African American among its MeSH codes. The two measures overlap for approximately 65% of observations.

With the second definition, I define a complementary measure that is less direct but more comprehensive. I use data on death rates by race from administrative records from the Center for Disease Control and Prevention (CDC). Because comprehensive data on the incidence of disease by race is not available (nor, for most diseases, systematically kept track of), I proxy disease incidence with disease-related mortality.[14] I use mortality rates between 1999 and 2015 as a proxy for disease incidence. I build a measure of *relative mortality* for each disease d reported in the official statistics as the cause of deaths at least 5,000 times (sum of non-Hispanic white and non-Hispanic Black individuals) over this period.[15] For each disease d, the measure is built as follows:

$$\text{Relative mortality}_d = \frac{(\text{N. deaths of Black or African Americans due to } d)/(\text{N. Black or African Americans})}{(\text{N. deaths of white Americans due to } d)/(\text{N. white Americans})} \tag{1}$$

This measure captures excess mortality of Black or African Americans compared to white Americans for a given disease d, and is similar to the one employed by Cutler, Meara and Richards-Shubik (2012) to study inequality in infant medical care. I classify diseases into three groups: *Typically White* if relative mortality is at least 1.5 times higher among non-Hispanic white Americans, *Typically Black* if relative mortality is at least 1.5 times higher among non-Hispanic Black or African Americans, *Similar Incidence* is defined as the residual category. I map articles to diseases using the set of associated Medical Subject Headings (MeSH), as standard in the literature.[16] Similarly to the Demographics-based approach, I consider an article as addressing a given disease if the MeSH code linked to that disease is listed as one of the MeSH codes associated with that publication. One third of all publications are matched to at least one ICD-10 Cause of Death. Whenever more than one MeSH code is linked to a cause of death, I assign to that article an average of the relative mortality of all diseases linked to the article. Approximately 10% of all articles of the sample are linked to more than one ICD-10 Causes of Deaths.

[14] I use data on Underlying Cause-of-Death by race and ethnicity. "Underlying Cause-of-Death" is defined by the World Health Organization as "the disease or injury which initiated the train of events leading directly to death, or the circumstances of the accident or violence which produced the fatal injury." It is selected from the conditions entered by the physician on the death certificate. When more than one cause or condition is listed, the underlying cause is determined by the sequence of conditions on the certificate.

[15] The relative ranking of diseases is fairly stable over time (and it is essentially identical for the period 1999 to 2001). While, magnitudes vary slightly over time, the results are robust to choosing different time spans. Mortality rates start in 1999 as this is the year when the CDC started using the International Classification of Diseases, Tenth Revision (ICD-10) to report causes of death.

[16] The MeSH thesaurus is a controlled and hierarchically-organized vocabulary produced by the National Library of Medicine (NLM). MeSH terms are assigned manually by indexers of the NLM.

3.3 Demographics-based Approach: Results

The first approach to measuring the direction of innovation is to define an indicator of innovation directed towards Black individuals coding a binary variable that takes value 1 if a research article is associated with the MeSH code corresponding to Black or African American, and takes value 0 otherwise. This variable captures a direct link between the race of the researchers and the subject of their research.

I estimate a model where $y = 1$ if an article addresses a given disease, 0 otherwise:

$$y_{ip} = \alpha_t + \beta \,\text{Black-sounding name}_i + \delta X_i + \Gamma Z_p + \varepsilon_{ip} \tag{2}$$

Where i is the first author, p denotes the article, α_t are year of publication fixed effects, X_i is a vector including Hispanic-sounding name, Asian-sounding name, American Native-sounding name, with omitted category white-sounding name. Standard errors are clustered at last the name level. β is the propensity of Black-sounding scientists to research y compared to white-sounding ones.[17]

In Table 2, I report the results of the estimation of equation (2) on innovation directed towards Black individuals defined according to the demographics-based approach. The results in column (1) show that Black-sounding scientists are more than twice as likely compared to white-sounding ones to produce research that addresses African Americans.Next, I test whether this association is driven by observational studies, or by clinical trials. Disentangling the two is crucial because of mounting evidence that low participation of African Americans in clinical trials affects the "production" of clinical trials that include African Americans (Alsan, Durvasula, Gupta, Schwartzstein and Williams, 2022). In column (2), I find that Black-sounding researchers are 2.5 times more likely to mention African Americans in papers that report the results of clinical trials. In column (3), we see that a correlation similar to the one reported in column (1) holds for all other (observational) articles. While the result in column (2) could be the result of larger participation of Black individuals in clinical trials when the researcher is of their own race, consistently with findings by Alsan, Campbell, Leister and Ojo (2023), the association reported in column (3) shows that the willingness to participate in trials seems not to be the only driver of the match between the race and

[17] β will likely underestimate the Black-white gap because, as shown in Figure B1, Individuals with Black-sounding name are underrepresented among scientists. This can be visualized using Bayes rule: P(B | S, L) = P(I | B, L) × P(B | L) / P(S | L), where B is a Black individual, S is a scientist, L represents a last name. For this reason, whenever I refer to "Black-sounding name", I refer to the probability that that last name is held by a Black person in the US population.

ethnicity of scientists and the direction of their research.[18]

In columns (4) to (6), I re-estimate the results shown in columns (1) to (3) on the subsample of articles which focus on human subjects.[19] Results are similar to those in first three columns, and suggest that the positive and significant coefficients of Black-sounding last name are not simply driven by a higher propensity of Black-sounding researchers to research human subjects. To further test whether there is an association between researchers and direction of research based on their race or ethnicity, I re-run Table 2 on a dummy variable equal to one if the article has Hispanic or Latino among its MeSH terms, and equal to zero otherwise; and on a dummy variable equal to one if the article has Asian American among its MeSH terms, equal to zero otherwise.[20] The coefficients in column (1) are reported in Figure 4, while the full estimation results are reported in Appendix Table B14. These results reveal that scientists are disproportionately more likely to include individuals of their own demographic group in their research.[21]

In Table B15, I look at the sample of ongoing clinical trials registered on `ClinicalTrials.gov`. This sample consists of 12,366 trials registered in the portal between 2002 and 2018 and with Principal Investigator affiliated with a US institution. I define a set of indicators for Black or African American, Hispanic or Latino, and Asian using a dictionary approach on the project description.[22] To assign race to Trial Investigators, I take the average of each racial probability across all registered Principal Investigators (as researchers are not listed in a meaningful way, e.g. by contribution, like in the case of articles or patents). In column (1), I show the results of estimating equation (2) on the indicator for Black or African American. The association is positive and highly statistically significant: compared to a white-sounding one, a Black-sounding team of researchers is four times more likely to run trials with Black or African Americans. In columns (2) and (3), I test whether Black-sounding teams are more

[18] In Figure B6, I report coefficients across a more disaggregated set of publication types. "Publication type" is provided by PubMed for each research article. An article can be assigned to multiple publication types, but this is fairly rare. The split I report in this figure is virtually unchanged when excluding articles assigned to more than one publication type.

[19] More specifically, I keep only those articles whose assigned MeSH terms all fall under the Human category out of this article's MeSH terms that fall into the Human, Animal, or Molecular/Cellular Biology categories defined by Hutchins, Davis, Meseroll and Santangelo (2019).

[20] A dummy variable for research addressing white Americans cannot be built because the corresponding MeSH term is almost never used (similarly, white ethnicity is rarely explicitely mentioned in the abstract). This reflects the fact that including white Americans in trials and observational studies in the norm.

[21] It is worth noting that in this plot I do not exclude article addressing multiple demographic groups, so the variables on the LHS are not mutually exclusive. Results are even stronger when those observations are excluded.

[22] I assume that a project description that mentions a given demographic group is designed to include that group. I implement a dictionary-based classification as MeSH terms are not available in this data.

likely to research diverse populations, or whether this match between scientist and content of their innovation is linked to race. I find that these teams are not significantly more likely to include Hispanic or Latinos, or Asians, in their trial description. However, Hispanic-sounding teams are more likely to include Hispanic or Latinos in their project descriptions, and Asian-sounding teams are more likely to include Asians. This finding is line with the results on articles using the demographics-based apprach (Figure 4) and on white-sounding inventors found using the frequency-based approach: the match between race or ethnicity of researchers and the direction of their research is not specific to a given group, but rather an empirical regularity across demographic groups.[23]

This finding provides a novel explanation for the lack of inclusion of Black individuals in medical research and complements existing findings that point to the role of low enrolment in clinical trials (Alsan, Durvasula, Gupta, Schwartzstein and Williams, 2022), and lower funding allocated to typically Black diseases (Farooq, Mogayzel, Lanzkron, Haywood and Strouse, 2020).

3.4 Frequency-based Approach: Results

With the second approach, I define "Innovation benefiting Black or African Americans" as those research articles addressing diseases that have higher mortality among Black or African Americans compred to white Americans. In Table 3 I report the result of the estimation of equation (2) on relative mortality. Throughout, I control for the logarithm of total deaths among Black or African Americans and White Americans due to disease d, over the period 1999 to 2015. I estimate the following model:

$$y_{ipd} = \alpha_t + \beta \,\text{Black-sounding name}_i + \delta Z_i + \gamma \log(\text{Total n. deaths})_d + \epsilon_{ip} \tag{3}$$

In column (1), I estimate equation (3) on a dependent variable equal to the logarithm of the relative mortality of disease d among Black or African Americans compared to white Americans. In column (2), I estimate the same equation on a dependent variable that equals 1 if a disease is at least 1.5 more frequent among Black or African Americans, equals 0 otherwise. In column (3), I estimate the same equation on a dependent variable that equals 1 for diseases that have roughly balanced incidence among Black individuals and whites, that is those with relative mortality less than 1.5 among Black or African Americans

[23]It is worth noting that, while the magnitudes of the match between researcher and content of research reported in this table vary by demographic group, it is hard to infer real magnitudes of the match without keeping disease frequency and incidence constant.

compared to white Americans, and less than 1.5 among white Americans compared to Black or African Americans, and equals 0 otherwise. In column (4), I estimate the same equation on a dependent variable that equals 1 if a disease is at least 1.5 more frequent among white Americans, equals 0 otherwise.

Compared to white-sounding ones, Black-sounding researchers more likely to research diseases more frequent among Black or African Americans, while white-sounding researchers are more likely to research diseases frequent among white Americans. The coefficients shown in column (1) show that, compared to white-sounding researchers, Black-sounding researchers focus on diseases that have a relative mortality that is 12% higher on average among Black or African Americans. The results in column (2), looking at relatively more white diseases, reveal that white-sounding researchers are 15% more likely to research *typically white* diseases, such as Alzheimer's diseases or melanoma. The results in column (3) show that there is no significant difference for diseases with balanced incidence across whites and Black individuals. Finally, column (4) reveals that Black-sounding researchers are 20% more likely to research *typically Black* diseases. Results are qualitatively and quantitatively similar when adopting different thresholds.[24] In Appendix Table B23, I run specifications that mimic the ones in columns (1) to (4) of Table 3, but on the sample of articles linked to a patent (Marx and Fuegi, 2020, 2022).

These findings are in line with research by Koning, Samila and Ferguson (2020, 2021) for female innovation showing that women have higher probability to engage in research and innovation benefiting women, by Einiö, Feng and Jaravel (2023) showing that inventors are more likely to create new product benefiting individuals similar by gender, age, and socio-economic status.

3.5 Robustness Checks

I conduct a large set of robustness checks to gauge the validity of these findings.

3.5.1 Demographics-based Approach: Robustness

In Table B5, I define the dummy "Research directed toward Black or African Americans" with a dictionary classification based on the publication abstract. As the main outcome is binary, in Table B6, I re-estimate Table 2 using a logit model to make sure results are not driven by the choice of using a linear probability model. Results hold throughout.

[24] In Appendix Table B24 and Table B25, I report a version of this Table where I define the variables in column (2) through (5) setting thresholds at 1.3 and 1.7.

In Table B7, I control for the gender of the first author using data on gender inferred from first names from Koning, Samila and Ferguson (2020, 2021). In Table B8, I include, respectively, journal FE, US state where the institution of affiliation of the first author is located FE, and first author affiliation FE. Results hold similarly to those shown in Table 2.

In Table B9, I re-estimate equation 2 using different definitions of Black-sounding name (and symmetrically defined definitions for the vector of other racial probabilities). In Panel A, I include the vector of racial probabilities of the last author of the paper. Panel B, I construct an average of Pr(Black | Last name) across all researchers listed in the publication.[25] In Panel C, I recode Pr(Black | Last name) as binary variable taking value 1 if it is $\geq .5$, $= 0$ otherwise.[26] In Panel D, I restrict thee set of last names to those last names present in the 1930 US Census among US-born individuals, and with a frequency of at least 100. This exercise aims to test if results hold in the subset of US-born scientists. Results are robust to using different definitions of Black-sounding last name.[27]

In Table B10, I re-estimate Table 2 using data on all published articles by a first author affiliated with a US institution, regardless of the Commercialization Impact Factor of the journal. Results are consistent and larger in magnitude. Interestingly, the average Pr(Black | Last name) is higher (.072 compared to .064). In Table B11, I use the same sample used in the main Table and I further split it by Journal Commercialization Impact Factor (JCIF). In Panel A, I show results for articles published in journals with below-median JCIF. In panel B, in articles published in journals with above-median JCIF. The pattern is similar to the one observed in the previous table. While estimated coefficients are smaller, these are of similar magnitudes across the two tables if they are rescaled by the mean of the dependent variable, which is lower in the above-median sample. Finally, in Table B12 I show that similar patterns hold in the articles linked to a patent, providing an even more "applied" measure of research. The same patterns hold in this subsample.[28]

In Table B13, I re-estimate columns (1) to (3) of Table 2 on the sample of articles and MeSH terms linked to them. In columns (1), (3), and (5), I replicate the results int he

[25]Both in Panel A and in Panel B, I drop single-authored publications. For Panel A, the average is computed across all listed authors whose last name is successfully matched to a last name in the matrix of racial probabilities.

[26]I do this symmetrically for all racial probabilities, and construct an additional dummy "multiple race" taking value 1 if none of the other racial probabilities is above .5, = 0 otherwise. Only 2% of last names fall into this category.

[27]In Panel C, Column (2) and Column (5), coefficients are not significant due to lack of power: by using variation only in the set of last names with Pr(Black-sounding name $\geq .5$), variation in the RHS variable comes from very few observations in this smaller subsample of the data.

[28]In Column (2) and (5) significance of the coefficients is lost due to the very small sample size.

main Table. In columns (2), (4), and (6), I control for MeSH terms FE, and coefficients are essentially unchanged. This test suggests that even keeping the same medical conditions constant, scientists with a Black-sounding last name are more likely to include Black or African Americans in their research.

3.5.2 Frequency-based Approach: Robustness

In Tables B16 to B23 I implement the same robustness checks symmetric to those shown in Table B5 to B12 for the Frequency approach. Results remain consistent throughout, and are in line with the patterns observed for the demographics-based approach.

Additionally, in Table B24 and Tables B25 I show that results are robust to choosing different thresholds to compute the variables *Typically White* and *Typically Black*. Finally, in Table B26, I show that results are robust to using the absolute number of deaths, rather than the relative one as in the main analysis.

3.6 Impact

What is the scientific impact of research conducted by scientists and benefiting their own demographic group? To study this question, in Table B29 and Table B31 I test whether these publications are cited by other publications in a differential way compared to other papers. The results reveal no statistically significant evidence suggesting this is the case. This result is robust to using the Relative Citations Ratio, a metric that normalizes the total number of citations by the citations received by publications in the same area of research and year (Table B30 and Table B32).

Finally, I split the sample by below- and above- median Journal Impact Factor (JIF), and I re-estimate Table 2 and Table 3. The results, shown in Table B33 and B34, reveal that scientists with a Black-sounding last name are equally likely to conduct research benefiting their own demographic group across journal with lower or higher impact factors.

4 Mechanisms

In this section, I study the mechanisms behind the match between the race of researchers and the direction of research and innovation documented in the previous section.

4.1 Race or Socio-economic Status? Ancestry Variation

One residual question from the findings shown so far is whether the match between the race of researchers and the direction of innovation is linked to race, or to other factors that correlate with race such as enviornmental factors of socio-economics status. The ranking of diseases I introduced with the Frequency-base approach correlates both with relative incidence by race, but also mimics relative incidence by socio-economic status.

As a result, one argument could be that investment in preventive behaviors, as well as higher engagement with the health system, would act as a substitute for that research. To establish a link between the race of the researcher and the content of the innovation, I develop an approach that relies on variations in probability of having a certain disease due to different ancestral origins. Drawing on findings in medicine and anthropology, I focus on article that study diseases that are more common among Black individuals (white) for reasons linked to their ancestry, but not to environmental factors or to socioeconomic status. More specifically, I focus on a typically Black disease, Sickle Cell Anemia, and a typically white disease, Melanoma. The incidence of these diseases is linked to conditions that provided an evolutionary advantage to ancestors in their environment of origin, but through genetic inheritance, carry over to today. To disentangle race from socio-economic status through the ancestry-based approach, two assumptions must hold true.

The first assumption is that human mobility happens at a faster rate than human evolution. That is, even when humans migrate, their adaptation to the environment will be slow. The second assumption is that the incidence of the disease is not correlated with current socio-economic status (nor socio-economic status at birth), or with other factors linked to the individual's current environment (or environment at birth). The first assumption is likely to hold in this setting as evolution is a slow process, and much slower than human migrations to the United States. The second assumption is likely to hold, as detection of this disease is nearly perfect in the US because Sickle Cell Anemia is covered by Universal Newborn Screening. In addition, I provide falsification tests for each disease.

Sickle Cell Anemia is a multisystem disorder and the most common genetic disease in the United States, affecting 1 in 500 African Americans. It is caused by a mutation in the hemoglobin beta chain in which glutamic acid is substituted with valine at position six on chromosome 11. Sickle cell disease in more common in individuals with African ancestry because carriers of the gene are protected from severe forms of malaria.

In columns (1) and (2) of Table 4, I report the results of the estimation of equation (2) where y is an indicator for whether a research article lists Sickle Cell Anemia among its

MeSH codes. As a first falsification test, I study research on β-Thalassemia, the other major hereditary hemoglobinopathy. Both diseases are recessive autosomal disorders. Differently from Sickle Cell Anemia, the incidence of β Thalassemia is not only high among those of African ancestry, but also among other populations. Current studies have shown that there is no strong evidence in support of the fact that the carriers are protected against malaria (Introini, Marin-Menendez, Nettesheim, Lin, Kariuki, Smith, Jean, Brewin, Rees, Cicuta et al., 2022).

In columns (3) and (4) of Table 4, I report the results of the estimation of equation (2) where y is an indicator variable taking value 1 if the article addresses Thalassemia, 0 otherwise. The estimation results show that, across all articles, Black-sounding researchers are twice as likely to research Sickle Cell Anemia compared to white ones. In the subsample of hematology articles, they are almost five times more likely to research the disease compared to white-sounding researchers. As a comparison, they are not differentially likely to research Thalassemia. In terms of overall magnitudes, the number of articles on Sickle Cell Anemia over the period is roughly 2.3 times higher compared to the ones on Thalassemia, while the incidence of Sickle Cell Anemia in the United States is estimated to be 20 times higher compared to the one of beta Thalassemia.[29]

In order for Sickle Cell Anemia and Melanoma to deliver causal links between race and the direction of innovation, we need that they are uncorrelated with current socio-economic status. In the case of Sickle Cell Anemia, I provide three pieces of evidence in support of this hypothesis. First, a similar higher incidence among Black individuals compared to Caucasians is found in different contexts such as the United Kingdom. Because healthcare in the UK is provided universally by the government, socio-economic status should matter less here. While we lack systematic evidence on the incidence of Sickle Cell Anemia by socio-economic status, several studies have shown that both individuals from low socio-economic status and those from higher one may suffer from the disease.

4.1.1 Ancestry Variation: Melanoma

Melanoma, also known as malignant Melanoma, is a type of cancer that develops from the pigment-producing cells melanocytes. Melanoma typically occurs in the skin, and is caused both by genetic and environmental factors. Caucasian race, male sex, and older age are well-recognized factors associated with an increased risk of developing Melanoma (Azoury and

[29]The estimated number of individuals with Sickle Cell Anemia is 100,000 (CDC). The estimated number of individuals with beta-Thalassemia is 5,000 (estimates on the incidence of beta-Thalassemia are taken from NIH.gov.

Lange, 2014). A key environmental trigger of Melanoma is exposure to UV lights. Melanoma is less frequent among individuals with darker skin (such as those of African descent). This is because a darker skin tone is indication of more melanin, a molecule that protects against UV lights. Therefore, Black people are far less likely to develop Melanoma than non-Hispanic white people (at a rate of 1 per 100,000 compared to 30 per 100,000) due to the protection that melanin, the body's natural skin pigment, provides from damaging ultraviolet rays.[30]

In columns (1) and (2) of Table 5, I report the results of the estimation of equation (2) where y is a dummy equal to 1 if a research article addresses Melanoma, equal to 0 otherwise. Across all articles, scientists are 50% less likely than white-sounding ones to research Melanoma, and similarly among articles related to neoplasms.

Differently from the case of sickle cell anemic and Thalassemia, finding a comparable disease is more challenging. I adopt a data-driven approach and study research on thyroid neoplasma. Thyroid cancer had similar mortality among Black or African Americans and among White Americans, and for Black Americans the mortality rate is similar to the one due to Melanoma.[31].

In columns (3) and (4), I study the likelihood that a scientist with Black-sounding name engages in research on thyroid neoplasm. Both unconditionally, as shown in column (3), and within research articles on neoplasms, as shown in column (4), Black-sounding researchers are not differentially likely to research thyroid neoplasm compared to white-sounding ones.

4.2 A Shock to Relative Mortality: The Introduction of HAART

Today, Human Immunodeficiency Virus (HIV) is the disease with the highest relative mortality for Black Americans compared to white Americans: Black Americans are over 9 times more likely to die of HIV. However, this was not always the case. The medical literature has documented that HIV became a relatively higher contributor to the mortality of Black individuals (compared to whites) after the introduction of the Highly Active Antiretroviral Therapy (HAART) in 1996 (Levine, Briggs, Kilbourne, King, Fry-Johnson, Baltrus, Husaini and Rust, 2007; Levine, Rust, Pisu, Agboto, Baltrus, Briggs, Zoorob, Juarez, Hull,

[30]Interestingly, the evolutionary reason why individuals of African ancestry have darker skin tone (i.e., more melanin in their skin), is not linked to the incidence of Melanoma, but to the benefits and costs of UV lights for the functioning of the human body. The leading theory in anthropology is that the levels of melanin correlated with latitude to allow the human body to absorb enough vitamin D (triggered by exposure to UV lights), while preserving folate levels, which are depleted by UV light.

[31]Mortality rates by group due to each type of cancer are shown in Appendix Figure B10 and Appendix Figure B11.

Goldzweig et al., 2010). HAART is a medication regimen used to manage and treat HIV and it is composed of several drugs in the antiretroviral classes of medications. Approved by the U.S. Food and Drug Administration (FDA) in 1995, HAART became widely available in 1996. HAART has dramatically decreased morbidity and mortality among HIV-infected individuals. Antiretroviral therapy delays the progression of HIV-related disease and prolongs survival, but these benefits have not been equitably distributed by race.

The evidence shown so far would predict that, when HIV becomes a relatively more "Black" disease, the relative share of scientists with Black-sounding name researching the disease should go up. I test this prediction using data on the universe of research grants awarded by the National Institutes of Health between 1988 and 2012. Using data on awarded grants is more suited to study how researchers react to this policy shock due to their more "real-time" nature. I classify grants as being related to HIV using a dictionary-based approach on the description of the research grant. Additionally, I define an indicator for "Research addressing Black individuals" bsead on a dictionary classification on the abstract. Similarly to the rest of the analysis, I assign a vector of racial probabilities based on the last name of the first researcher listed on the article. First, I estimate a difference-in-differences equation around the introduction of HAART:

$$y_{pt} = \alpha_t + \beta_1 \text{Research about HIV}_p + \beta_2 \text{Research about HIV}_p \times \text{Post HAART} + \epsilon_{pt} \quad (4)$$

where p is a research grant, α_t are grant award year fixed effects, Research about HIV_i is a dummy $= 1$ if the research grant addresses HIV, $= 0$ otherwise, Post HAART is a dummy $= 1$ if the grant year is ≥ 2000, $= 0$ otherwise. Standard errors are clustered by last name.

Second, I estimate a similar differences-in-differences model where y is an indicator $= 1$ if a research grant addresses HIV, 0 otherwise

$$y_{ipt} = \delta_t + \gamma_1 \text{Black-s. name}_{it} + \gamma_2 \text{Black-s. name}_i \times \text{Post HAART} + \delta X_{it} + \theta X_{it} \times \text{Post HAART} + \epsilon_{ipt} \quad (5)$$

where i is the first researcher listed on the grant, p is a research grant, δ_t are grant award year fixed effects, Research about HIV_i is a dummy $= 1$ if the research grant addresses HIV, $= 0$ otherwise, Post HAART is a dummy $= 1$ if the grant year is > 1996, $= 0$ otherwise. X_i is a vector including Hispanic-sounding name, Asian-sounding name, American Native-sounding name, with omitted category white-sounding name. Standard errors are clustered by last name.

In column (1) of Table 6 I show that after the introduction of HAART, articles on research

projects addressing HIV become five times more likely to mention African Americans in their abstract. This result confirms that HIV becomes more associated with Black individuals.[32] In column (2), I test whether homophily between Black-sounding researchers changes after the introduction of HAART. After HAART, the difference in propensity to research HIV compared to white-sounding researchers increases four-fold compared to the pre-HAART period. The magnitude is similar to the increase in column (1), but the results in column (2) remain positive and significant even restricting the sample to those grants which do not mention explicitly African Americans.

I also estimate equations 4 and 5 in the sample of PubMed articles published between 1988 and 2017.[33] Results are consisted with the findings for the NIH grants, and are reported in Table XX. To investigate possible heterogeneity of treatment effects over time, I also estimate a more flexible model that, rather than interacting the vector of racial probabilities with the Post HAART indicator, interacts it with time dummies:[34]

$$y_{ipt} = \alpha_i + \alpha_{s(i)\times t} + \sum_{k=-3}^{+7} \beta^k \times I(k \leq t - 1997 \leq k+1) \times \text{HIV}_c + \boldsymbol{x}'_{ct}\boldsymbol{\delta} + +\varepsilon_{ipt}, \qquad (6)$$

where $I(k \leq t - 1997 \leq k+1)$ is an indicator variable that takes value one if t is in the three-year window indexed by k, and zero otherwise. $\boldsymbol{x}'_{ct}$ is a vector of year FE. y_{ipt} is an indicator taking value 1 if the paper has "Black or African American" among its MeSH code, 0 otherwise. The results of this estimation are reported in Figure 5, and confirm the findings of the DiD design. After the introduction of HAART, articles on HIV become relatively more likely to mention Black or African Americans.

$$y_{ipt} = \alpha_i + \alpha_{s(i)\times t} + \sum_{k=-3}^{+7} \beta^k \times I(k \leq t - 1997 \leq k+1) \times \text{Black-sounding name}_c + \boldsymbol{x}'_{ct}\boldsymbol{\delta} + \varepsilon_{ipt}, \quad (7)$$

where $I(k \leq t - 1997 \leq k+1)$ is an indicator variable that takes value one if t is in the three-year window indexed by k, and zero otherwise. $\boldsymbol{x}'_{ct}$ is a vector of racial probabilities interacted with time dummies, and it includes year FE. The results of this estimation on a dummy equal to 1 if the article has "HIV" among its MeSH codes are reported in Figure 6. The results of the dynamic differences-in-differences design confirm the results of Table 6

[32] The fact that the correlation was positive also before the introduction of HAART suggests that research mentioned African Americans also before HAART, but mostly jointly with mentioning white as well.

[33] I begin the analysis in 1988 as this is the first year the MeSH code corresponding to HIV was introduced.

[34] In the dynamic DiD specifications, I code periods over three-year windows to reduce noisy fluctuations in estimated treatment effects and to improve efficiency by pooling observations.

and Table XX. After the introduction of HAART, compared to White-sounding scientists, Black-sounding ones become relatively more likely to research HIV.

These results establish a more direct link between the relative incidence of a disease and homophily of researchers, and they suggest that homophily is not driven by specific diseases, but by the incidence of these diseases across demographic groups.

5 Race and the Quantity and Quality of Innovation

Having established a link between the racial composition of scientists inventors and the direction on innovation, in this section I study how this affects the quantity and quality of innovation. In section 5.1, I describe the research design. In section 5.2 and I describe how this correlates with access to resources within the patenting process, and with the impact of granted patents.

5.1 Research Design

I test the correlation between having a Black-sounding name and the likelihood of being granted a patent, conditional on applying. I estimate the following equation on the sample of patent applications submitted to the US Patent Office between 2001 and 2018:

$$y_{ip} = \alpha_t + \beta \,\text{Black-sounding name}_i + \delta X_i + \Gamma Z_p + \epsilon_{ip} \tag{8}$$

Where $y = 1$ if an application is granted, $= 0$ otherwise, i is an inventor, p is a patent application, α_t is filing year FE, X_i is a vector of racial probabilities including Hispanic-sounding, Asian-sounding, American Native-sounding. The omitted category is white-sounding, and Z_p is a vector of application-level controls. Standard errors are clustered at the last name level.[35] Results are shown in Table 8. In Table 9, I report the results of the estimation of equation 8 with the dependent variable y equal to the total number of forward citations received by the patent, restricting the sample to granted patents. In this specification, β is the average gap in patent granting (conditional on applying) for an inventor with a Black-sounding name, compared to an inventor with a white-sounding name.

[35] I cluster errors at the last name level because that is the level of variation of the independent variable of interest (Pr(Black | Last name). In robustness checks, I cluster by patent class, or by name and patent class and significance is virtually unchanged.

5.2 Results

The first specification (column 1) includes only year fixed effects. The coefficient of Black-sounding name is negative and significant: being Black is associated with 6% lower probability of being granted a patent conditional on applying. In column (2), I include USPC subclass FE. The coefficient decreases slightly but remains within the same 95% confidence interval. This suggests that the racial gap in patent-granting is not due to to differential sorting of Black-sounding inventors into patent classes with lower grant rates. In column (3), I add assignee FE. The assignee is the entity (such as a firm, or a university) that holds the property right on the patent. By law, a patent must report at least one inventor but can omit the assignee. In such a case, the inventor is implicitly assumed to be the holder of the property right. In this specification, I combine assignee fixed effect with an indicator for "small entity", granted by the USPTO to universities, single inventors, and firms of up to 500 employees. As a result, all entities without assignee will have assignee equal to "small entity". I adopt this strategy to keep the sample comparable to columns (1) and (2), but the coefficients in column (3) through (5) are similar when I drop observations without assignee. Assignee fixed effects explain approximately one third of the unconditional gap show in column (1). This suggests that Black-sounding inventors sort into entities with lower acceptance rate. However, even within the same organization (and patent subclass), they are still granted patents at lower rate conditional on applying for one. As assignees can be entities largely heterogeneous by location and size, I control for a set of additional variables to test whether, within organizations, Black-sounding inventors may be located in different areas compared to white-sounding ones, or they may have access to less support within the patenting process. In column (4), I include state of residence fixed effects. The magnitude of the gap remains almost identical compared to the one in column (3), which suggests that Black-sounding inventors do not sort into geographically different locations of the same firm (at least by state). In column (5), I include proxies for attorney quality. Patent attorneys play a key role in the patenting process as they help draft the patent document, and interact with patent examiners. I build two novel measures of attorney "quality" from the patent document. The list of attorneys for each patent is reported on the patent application, but the number of listed attorneys is typically very large (the average number across all applications is 15). For this reason, I build the average grant rate of all attorneys listed on the patent ("average grant rate of the legal team"). Looking at the universe of patent applications between 2001 and 018, I build for each attorney the average grant rate across all patents

they have worked on.[36] Then, I compute the average of the individual-specific grant rate across all attorneys listed on the patent.[37] The coefficient on the average grant rate of the legal team is positive and significant and explains one-fifth of the Black- vs white-sounding gap shown in column (4). These results suggest that differential access to resources indeed differs across white individuals and Black individuals, but approximately half of the racial gap in patent granting remains after accounting for the available metrics of differential access to resources.

I run several robustness checks to check the validity of these results to different assumptions and metrics. First, I run the specification assigning Pr(Black | Last name) (and the vector of racial probabilities) as the average across all members of the team. Second, I restrict the analysis to single-author patents (who do not have this potential measurement issue of multiple inventors). Results are comparable throughout.

One potential explanation for the results shown in Table 8 is that these patents are of lower quality. However, this hypothesis does not find support in the data. More specifically, patents granted to Black-sounding inventors have higher impact compared to the ones granted to white-sounding ones. Using a Poisson count model, I estimate a version of model 8 where I focus on the subset of granted patents and where y is the number of forward citations. These trace the acknowledged contributions of prior art, and are used in the literature as an ex-post measure of patent quality and economic value. The number of citations includes any citation received by the patent in the five years after its filing.[38] The results show that Black-sounding patents are cited more compared to white-sounding patents. The results are shown in Table 9, columns (1) to (5).

These results suggest that barriers in the patenting process may prevent patents from Black-sounding inventors from being granted. What would be the impact of equalizing access to science and innovation on the overall production of science? To answer this question, In the next section I estimate a model of occupational choice (Hsieh, Hurst, Jones and Klenow, 2019).

[36] I restrict the sample to those attorneys that have worked on at least 20 patents over the period, but results are consistent when I adopt different thresholds.

[37] Whenever a patent has no attorneys listed on the patent, I plug in a zero. I relax this assumption in robustness checks and similar results hold.

[38] Citations are constructed taking into account both citations from other patents, and citations from other applications.

6 Equalizing Access to Innovation and Science

In this section, I study the labor market consequences of equalizing access to innovation and science through a model of occupational choice. In section **??**, I describe the model. In section 6.1 I outline the setup of the model. In section 6.2, I discuss the model estimation. In section 6.3, I report policy counterfactuals.

6.1 Model Outline

To quantify the effect of equalizing access to innovation and science for Black Americans, I estimate a Roy model of occupational choice (Hsieh, Hurst, Jones and Klenow, 2019). There are at least two reasons why a general equilibrium model is useful to estimate the impact of removing barriers for Black individuals from the economy. First, if individuals anticipate discrimination, they may choose another career early on in life. Therefore, we would observe them select into different occupations where they may be less productive, or which they may enjoy less, but where barriers are lower. Second, as more individuals enter into an occupation, labor demand for that occupation will change, and equilibrium wages will go down.

The economy is composed by a continuum of individuals i who are either white or Black. Their group is indexed by $g = \{\text{white, Black}\}$. Each individual chooses an occupation j to maximize their lifetime utility. Individuals choose their lifetime occupation and decide how much time to dedicate to schooling before entering the labor market (the pre-period), and live for three periods. Their lifetime utility is equal to:

$$\log U_i = \alpha \sum_{t=1}^{3} \log c_{it} + \log (1 - s_{ij}) + \log z_{jg} + \log \mu_{ij} \tag{9}$$

α represents the tradeoff between utility in the pre-period and utility over the remaining of the lifetime. s is time spent in school to acquire human capital (so that $(1-s)$ is leisure), z is group-specific utility derived from working in occupation j, and μ is individual utility from working in occupation j. The parameter z_{jg} relaxes the assumption that, in the absence of barriers, all groups will select occupation j at the same rate. It can be interpreted as

preferences, beliefs, or experience.[39] Individual consumption c_{it} is equal to:

$$c_{it} = (1 - \tau^w_{jg}) w_{jt} \varepsilon_{ij} h_{ij} - (1 + \tau^h_{jg}) e_{ij} \tag{10}$$

Where τ_{jg} is the job- and group- specific tax to work in occupation j, w_{jt} is the efficiency wage, ε_{ij} is individual productivity of working in occupation j, (i.e., their "talent" for occupation j), h_{ij} is human capital acquired to work in occupation j. Over the lifetime, individuals repay the loan they got in the first period to acquire education. They repay it equally across all three periods. In every period they have to repay 1/3 $(1 + \tau^h_{jg}) e_{ij}$, where τ^h_{jg} is the tax on acquisition of human capital.[40] In this economy, output is produced by one firm that aggregates labor inputs from J occupations through the production function:

$$Y = \left[\sum_{j=1}^{J} (A_j \cdot H_j)^{\frac{\sigma-1}{\sigma}} \right]^{\frac{\sigma}{\sigma-1}} \tag{11}$$

Where H_j is total efficiency units of labor in each occupation, σ is the elasticity of substitution across occupations, A_j is the exogenously-given productivity of occupation j. In equilibrium, $w_j t$ clears the labor market in each sector so that $H^{supply}_{jt} = H^{demand}_{jt}$.

6.2 Estimation

To estimate the model, I impose a series of additional assumptions. First, in the main estimation, I assume that individuals only select occupations based on talent (not on individual preferences). That is, that $\mu = 1$. Secondly, I assume that talent is distributed equally across groups in a given sector. Third, I assume that white individuals do not face any barriers in the acquisition of human capital or in the labor market. That is, $\tau^h_{white} = 0$ and $\tau^w_{white} = 0$ in all occupations and all periods. Fourth, I normalize preference for the home sector to be equal to 1 for all groups. Fifth, I assume that the return to experience γ is the same for all sectors, groups, and cohorts. I estimate the model using internally calibrated parameters τ^h, τ^w, and z, for which I rely on data from the first two section of the paper and

[39] For example, Hsieh, Hurst, Jones and Klenow (2019) documents a decrease of z in the US between 1960 and 2010 a decrease women in the home sector, which they interpret as changes in social norms for women working in the market sector or changing preferences for fertility.

[40] For inventors, the barrier in human capital accumulation τ^h can be micro founded by Bell, Chetty, Jaravel, Petkova and Van Reenen (2019*b*,*a*), who show that exposure to more inventors in childhood leads to an increased propensity of becoming an inventor. The on-the-job barrier τ^w is motivated by the evidence shown in section 5. Similar evidence does not exist for scientists, but I assume it is plausible that similar frictions might be at play. In ongoing work, I re-estimate the model calibrating the barriers usineg these micro-founded estimated.

from the Current Population Survey (CPS) from 1990 to 2018. I use externally calibrated moments from Hsieh, Hurst, Jones and Klenow (2019). A summary of these moments, and of internally and externally calibrated parameters is provided in Appendix C.

6.3 Policy Counterfactuals

In this section, I estimate the impact of equalizing access to innovation and science, i.e. lifting barriers to human capital accumulation and in the labor market, on the total share of inventors and scientists in the economy. I define as "innovation and science" an occupation combining individuals who report working in "engineering" and "natural scientists". The results of this policy counterfactual are shown in Figure 7. Removing barriers increases the number of inventors and scientists by 10%. This increase is driven entirely by an increase in Black individuals, while the number of white individuals decreases by a negligible amount. Overall, in the economy, this corresponds to an increase in 1 p.p. in the number of scientists and inventors. In the plot, I report counterfactual estimations under a policy that lifts barriers only in the innovation and science sector (light grey bars), and a policy that lifts barriers everywhere in the economy. The broad result is consistent across the two different policies. The increase in the share of Black individuals is higher under the second policy. The intuition behind this result is that, at the margin, more Black individuals will sort into innovation if other occupations with high barriers elsewhere in the economy remain distorted (e.g., lawyers). Why does the overall number of inventors and scientists go up? This finding hinges on the assumption on the elasticity of substitution between occupations in aggregate production (σ). Following Hsieh, Hurst, Jones and Klenow (2019), I calibrate this parameter to have value 3, and experiment with different values in robustness checks. While the relative magnitudes of the increase in inventors vary with the value of σ, the core result on the increase in the number of inventors, and in the relative share of Black individuals among them, remains. Under any parameter value σ, the average productivity (or, quality) of scientitst and inventors increases after removing frictions, as talent is now optimally allocated across occupations. This result is in line with the empirical evidence on quality of granted patents discussed in section 5.

7 Discussion

The findings reported in section 6 suggest that equalizing access to innovation and science would lead to an increase in the quality and quantity of innovation, and to a larger share

of Black scientists and inventors. Would this lead to more research benefiting Black and African Americans?

Below, I summarize two arguments that suggest that the current underrepresentation of Black scientists and inventors leads to the underprovision of innovation benefiting Black or African Americans. This suggests that closing the Black-white gap in scientists and inventors will likely lead to more research benefiting Black or African Americans.

First, in section 3 I document that Black researchers are more likely to run clinical trials with Black participants. This is in line with recent evidence on the lack in participation in clinical trials of Black individuals (Alsan, Durvasula, Gupta, Schwartzstein and Williams, 2022; Alsan, Campbell, Leister and Ojo, 2023). Additionally, the fact that even in *registered* trials Black scientists tend to involve Black individuals more suggests that part of the effect is driven by researchers' decisions on which demographic groups to include in their studies. This hypothesis is supported by the fact that the coefficient in Table 2, Column (2) for published trials is larger compared to the one reported in Table B15, Column (2), which reports results on trial registration.

Second, when looking at white researchers (over-represented compared to the overall population), a similar pattern emerges. Compared to scientists with a Black-sounding last name, they are 15% more likely to research typically white diseases, and 30 to 40% more likely to research melanoma. Additionally, in articles and clinical trials I document that Hispanic-sounding researchers are more likely to design trials and publish articles with individuals of Hispanic or Latino origin, and Asian-sounding researchers with individuals of Asian origin. This suggests that, even when a group is over-represented in the pool of scientists and inventors compared to population size, the propensity to innovate for their own demographic group remains. The match between race or ethnicity and direction of research and innovation is not linked to being a minority group, but seems to be an empirical regularity across demographics.

Finally, an outstanding question is: *Why* do scientists and inventors produce research and innovation that benefits their own demographic group disproportionately more?

The match between researchers and HIV-related research suggests that this pattern is linked to demographic characteristics, rather than specific diseases. This is further supported by the finding that even conditioning on disease, Black-sounding scientists are more likely to include Black or African Americans in their research.

Differential returns are unlikely to explain per se why this match emerges. Research benefiting Black Americans does not have significantly different returns compared to other research

in the same area. Additionally, this match holds across bins of journal impact factors (high compared to low).

Plausible mechanisms through which the match between researcher and content of research may emerge are information asymmetries, differential expectations on market size, or preferences for innovating for one's own demographic group. One way to rationalize these findings through preferences is via intrinsic motivation: we know from seminal work by Stern (2004) that scientists are intrinsically motivated agents. The existence of a match between their race and ethnicity and the direction of their research and innovation would be consistent with intrinsic motivation being linked to a specific topic, rather than "general purpose" (i.e., scientists "paying" to be scientists *and* research a specific topic).

Understanding the relative contribution of information and preferences will have fundamental implications for the design of policies to ensure that the production of science and innovation benefits every group of society.

8 Conclusion

Previous literature has shown that minority gaps among inventors contribute to and perpetuate unequal opportunities within society (Bell, Chetty, Jaravel, Petkova and Van Reenen, 2019*a*,*b*). Such disparities can have large implications for the quality and quantity of innovation (Bloom, Van Reenen and Williams, 2019; Bell, Chetty, Jaravel, Petkova and Van Reenen, 2019*b*,*a*; Einiö, Feng and Jaravel, 2023). In this paper, I find evidence consistent with the racial gap in scientists and inventors leading to lower volume of research and innovation via a lower overall number of scientists and inventors. Inventors and scientists are a key input in the production function of science and innovation, and race-specific barriers in the economy lead to an inefficiently low number of them. I find that this is not only driven by differential sorting into these occupations, but also by Black inventors facing higher barriers within the patenting process.

What are the implications of too few Black inventors? In the second part of the paper, I test the hypothesis that Black scientists and inventors tend to innovate for Black individuals. I focus on medical innovation, a research field where innovation for Black individuals has been shown to be underprovided. Defining a set of novel metrics, I document that Black scientists are more likely to produce medical research and innovation benefiting Black individuals. This finding complements existing evidence on the importance of gender, socio-economic status, and age for the direction of innovation (Koning, Samila and Ferguson, 2020, 2021;

Einiö, Feng and Jaravel, 2023). This evidence suggests that equality among those producing science and research, as well as those commercializing these innovations, is fundamental to guarantee that all demographic groups in the population equally benefit from medical advancements.

Future research should focus on three issues of primary importance. First, on advancing our understanding of the reasons behind the racial gap in inventors and scientists, and on how to design policies and institutions aimed at closing this gap.[41] Second, on devising strategies to incentivize the production of innovation benefiting Black individuals in the transition to a more equal pool of researchers. This is likely to be a costly process as it takes a large amount of resources to get scientists to change the direction of their research (Myers, 2020). However, this represents an essential step to ensure that research and innovation benefit everyone in the population. Finally, an open question is whether, and to what extent, the racial and ethnic diversity of background and experience in the pool of innovators may foster, in itself, the production of novel ideas.

References

Abramitzky, Ran, Leah Platt Boustan, and Dylan Connor. 2020. "Leaving the enclave: Historical evidence on immigrant mobility from the industrial removal office." National Bureau of Economic Research.

Akcigit, Ufuk, and Nathan Goldschlag. 2023. "Measuring the characteristics and employment dynamics of US inventors." National Bureau of Economic Research.

Alsan, Marcella, Maya Durvasula, Harsh Gupta, Joshua Schwartzstein, and Heidi L Williams. 2022. "Representation and Extrapolation: Evidence from Clinical Trials." National Bureau of Economic Research.

Alsan, Marcella, Romaine A Campbell, Lukas Leister, and Ayotomiwa Ojo. 2023. "Investigator Racial Diversity and Clinical Trial Participation." National Bureau of Economic Research.

[41] A series of studies beginning with the seminal work by Bertrand and Mullainathan (2004), and most recently findings by Kline, Rose and Walters (2022), show that racial discrimination affects hiring decisions. Whether racial bias plays a role in patent granting decisions today is an open question. Recent findings (Jensen, Kovács and Sorenson, 2018; Avivi, 2023) show evidence consistent with gender bias at the patent office. Coluccia, Dossi and Ottinger (2023) finds that racial discrimination at the patent office harmed innovation at the beginning of the 20th century.

Ashraf, Nava, Oriana Bandiera, Virginia Minni, and Vıctor Quintas-Martınez. 2023. "Gender Roles and the Misallocation of Labour Across Countries." *Working paper*.

Avivi, Hadar. 2023. "Are Patent Examiners Gender Neutral?" *Working paper*.

Azoury, Saïd C., and Julie R. Lange. 2014. "Epidemiology, Risk Factors, Prevention, and Early Detection of Melanoma." *Surgical Clinics of North America*, 94(5): 945–962. Melanoma.

Bazzi, Samuel, Martin Fiszbein, and Mesay Gebresilasse. 2020. "Frontier Culture: The Roots and Persistence of "Rugged Individualism" in the United States." *Econometrica*, 88(6): 2329–2368.

Bell, Alex, Raj Chetty, Xavier Jaravel, Neviana Petkova, and John Van Reenen. 2019*a*. "Do tax cuts produce more Einsteins? The impacts of financial incentives versus exposure to innovation on the supply of inventors." *Journal of the European Economic Association*, 17(3): 651–677.

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

Berkes, Enrico, Davide Coluccia, Gaia Dossi, and Mara Squicciarini. 2023. "Dealing with Adversity: Religiosity and Science? Evidence from the Great Influenza Pandemic." *Working paper*.

Bertrand, Marianne, and Sendhil Mullainathan. 2004. "Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination." *American Economic Review*, 94(4): 991–1013.

Bloom, Nicholas, John Van Reenen, and Heidi Williams. 2019. "A Toolkit of Policies to Promote Innovation." *Journal of Economic Perspectives*, 33(3): 163–84.

Bryan, Kevin A, and Heidi L Williams. 2021. "Innovation: market failures and public policies." In *Handbook of industrial organization*. Vol. 5, 281–388. Elsevier.

Chetty, Raj, Gaia Dossi, Matthew Smith, John Van Reenen, Owen Zidar, and Eric Zwick. 2023. "America's Missing Entrepreneurs."

Chetty, Raj, Michael Stepner, Sarah Abraham, Shelby Lin, Benjamin Scuderi, Nicholas Turner, Augustin Bergeron, and David Cutler. 2016. "The association between income and life expectancy in the United States, 2001-2014." *Jama*, 315(16): 1750–1766.

Coluccia, Davide M, Gaia Dossi, and Sebastian Ottinger. 2023. "Racial Discrimination and Lost Innovation." *Working paper*.

Comenetz, J. 2016. "Frequently Occurring Surnames in the 2010 Census. United States Census Bureau."

Cutler, David, Angus Deaton, and Adriana Lleras-Muney. 2006. "The determinants of mortality." *Journal of economic perspectives*, 20(3): 97–120.

Cutler, David M, Adriana Lleras-Muney, and Tom Vogl. 2008. "Socioeconomic status and health: dimensions and mechanisms."

Cutler, David M, Ellen Meara, and Seth Richards-Shubik. 2012. "Induced innovation and social inequality evidence from infant medical care." *Journal of Human Resources*, 47(2): 456–492.

Dossi, Gaia, and Marta Morando. 2023. "Political Ideology and Innovation."

Einiö, Elias, Josh Feng, and Xavier Jaravel. 2023. "Social push and the direction of innovation." *Working Paper*.

Farooq, Faheem, Peter J Mogayzel, Sophie Lanzkron, Carlton Haywood, and John J Strouse. 2020. "Comparison of US federal and foundation funding of research for sickle cell disease and cystic fibrosis and factors associated with research productivity." *JAMA network open*, 3(3): e201737–e201737.

Fenton, Anne, Erika Elliott, Ashkan Shahbandi, Ekene Ezenwa, Chance Morris, Justin McLawhorn, James G Jackson, Pamela Allen, and Andrea Murina. 2020. "Medical students' ability to diagnose common dermatologic conditions in skin of color." *Journal of the American Academy of Dermatology*, 83(3): 957–958.

Fouka, Vasiliki. 2020. "Backlash: The Unintended Effects of Language Prohibition in US Schools After World War I." *The Review of Economic Studies*, 87(1): 204–239.

Fry, Caroline Viola. 2023. "Crisis and the Trajectory of Science: Evidence from the 2014 Ebola Outbreak." *The Review of Economics and Statistics*, 105(4): 1028–1038.

Gaule, Patrick, and Mario Piacentini. 2013. "Chinese graduate students and US scientific productivity." *Review of Economics and Statistics*, 95(2): 698–701.

Graham, Stuart JH, Alan C Marco, and Richard Miller. 2015. "The USPTO patent examination research dataset: A window on the process of patent examination." *Georgia Tech Scheller College of Business Research Paper No. WP*, 43.

Hsieh, Chang-Tai, Erik Hurst, Charles I Jones, and Peter J Klenow. 2019. "The Allocation of Talent and US Economic Growth." *Econometrica*, 87(5): 1439–1474.

Hutchins, B Ian, Matthew T Davis, Rebecca A Meseroll, and George M Santangelo. 2019. "Predicting translational progress in biomedical research." *PLoS biology*, 17(10): e3000416.

Hutchins, B Ian, Xin Yuan, James M Anderson, and George M Santangelo. 2016. "Relative citation ratio (RCR): a new metric that uses citation rates to measure influence at the article level." *PLoS biology*, 14(9): e1002541.

Introini, Viola, Alejandro Marin-Menendez, Guilherme Nettesheim, Yen-Chun Lin, Silvia N Kariuki, Adrian L Smith, Letitia Jean, John N Brewin, David C Rees, Pietro Cicuta, et al. 2022. "The erythrocyte membrane properties of beta thalassaemia heterozygotes and their consequences for Plasmodium falciparum invasion." *Scientific Reports*, 12(1): 8934.

Jensen, Kyle, Balázs Kovács, and Olav Sorenson. 2018. "Gender differences in obtaining and maintaining patent rights." *Nature biotechnology*, 36(4): 307–309.

Johnson, Julie A. 2008. "Ethnic differences in cardiovascular drug response: potential contribution of pharmacogenetics." *Circulation*, 118(13): 1383–1393.

Kerr, William. 2008*a*. "The ethnic composition of US inventors."

Kerr, William R. 2008*b*. "Ethnic scientific communities and international technology diffusion." *The Review of Economics and Statistics*, 90(3): 518–537.

Kline, Patrick, Evan K Rose, and Christopher R Walters. 2022. "Systemic discrimination among large US employers." *The Quarterly Journal of Economics*, 137(4): 1963–2036.

Koning, Rembrand, Sampsa Samila, and John-Paul Ferguson. 2020. "Inventor gender and the direction of invention." Vol. 110, 250–254, American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203.

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.

Kozlowski, Diego, Dakota S Murray, Alexis Bell, Will Hulsey, Vincent Larivière, Thema Monroe-White, and Cassidy R Sugimoto. 2022. "Avoiding bias when inferring race using name-based approaches." *Plos one*, 17(3): e0264270.

Levine, Robert S, George S Rust, Maria Pisu, Vincent Agboto, Peter A Baltrus, Nathaniel C Briggs, Roger Zoorob, Paul Juarez, Pamela C Hull, Irwin Goldzweig, et al. 2010. "Increased Black–White disparities in mortality after the introduction of lifesaving innovations: a possible consequence of US federal laws." *American Journal of Public Health*, 100(11): 2176–2184.

Levine, Robert S, Nathaniel C Briggs, Barbara S Kilbourne, William D King, Yvonne Fry-Johnson, Peter T Baltrus, Baqar A Husaini, and George S Rust. 2007. "Black–white mortality from HIV in the United States before and after introduction of highly active antiretroviral therapy in 1996." *American journal of public health*, 97(10): 1884–1892.

Louie, Patricia, and Rima Wilkes. 2018. "Representations of race and skin tone in medical textbook imagery." *Social Science & Medicine*, 202: 38–42.

Marx, Matt, and Aaron Fuegi. 2020. "Reliance on science: Worldwide front-page patent citations to scientific articles." *Strategic Management Journal*, 41(9): 1572–1594.

Marx, Matt, and Aaron Fuegi. 2022. "Reliance on science by inventors: Hybrid extraction of in-text patent-to-article citations." *Journal of Economics & Management Strategy*, 31(2): 369–392.

Michelman, Valerie, and Lucy Msall. 2023. "Sex, Drugs, and RD: Missing Innovation from Regulating Female Enrollment in Clinical Trials."

Moscona, Jacob, and Karthik Sastry. 2022. "Inappropriate technology: Evidence from global agriculture."

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

Nagaraj, Abhishek, Esther Shears, and Mathijs de Vaan. 2020. "Improving data access democratizes and diversifies science." *Proceedings of the National Academy of Sciences*, 117(38): 23490–23498.

NSF. 2015. "Women, Minorities, and Persons with Disabilities in Science and Engineering." *National Science Foundation Techinical Report*.

Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. "Dissecting racial bias in an algorithm used to manage the health of populations." *Science*, 366(6464): 447–453.

Olivetti, Claudia, M Daniele Paserman, Laura Salisbury, and E Anna Weber. 2020. "Who Married,(To) Whom, and Where? Trends in Marriage in the United States, 1850-1940." *NBER Working Paper*, , (No. 28033).

Rees, David C, Thomas N Williams, and Mark T Gladwin. 2010. "Sickle-cell disease." *The Lancet*, 376(9757): 2018–2031.

Rose, Evan K. 2023. "A Constructivist Perspective on Empirical Discrimination research." *Journal of Economic Literature*, 61(3): 906–923.

Schadendorf, Dirk, Alexander CJ van Akkooi, Carola Berking, Klaus G Griewank, Ralf Gutzmer, Axel Hauschild, Andreas Stang, Alexander Roesch, and Selma Ugurel. 2018. "Melanoma." *The Lancet*, 392(10151): 971–984.

Schwandt, Hannes, Janet Currie, Marlies Bär, James Banks, Paola Bertoli, Aline Bütikofer, Sarah Cattan, Beatrice Zong-Ying Chao, Claudia Costa, Libertad González, et al. 2021. "Inequality in mortality between Black and White Americans by age, place, and cause and in comparison to Europe, 1990 to 2018." *Proceedings of the National Academy of Sciences*, 118(40): e2104684118.

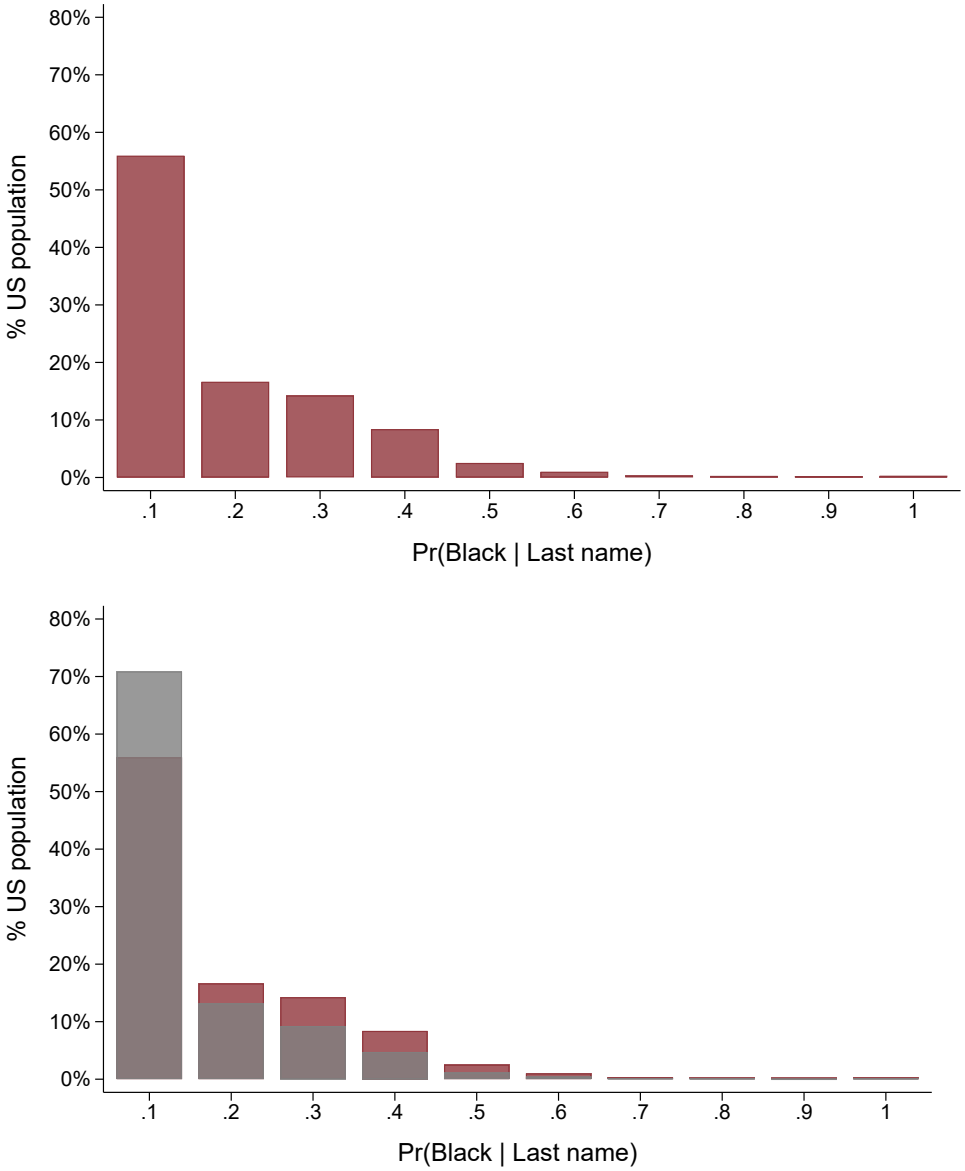
Sjoding, Michael W., Robert P. Dickson, Theodore J. Iwashyna, Steven E. Gay, and Thomas S. Valley. 2020. "Racial Bias in Pulse Oximetry Measurement." *New England Journal of Medicine*, 383(25): 2477–2478. PMID: 33326721.

Stern, Scott. 2004. "Do scientists pay to be scientists?" *Management science*, 50(6): 835–853.

Truffa, Francesca, and Ashley Wong. 2022. "Undergraduate gender diversity and direction of scientific research."

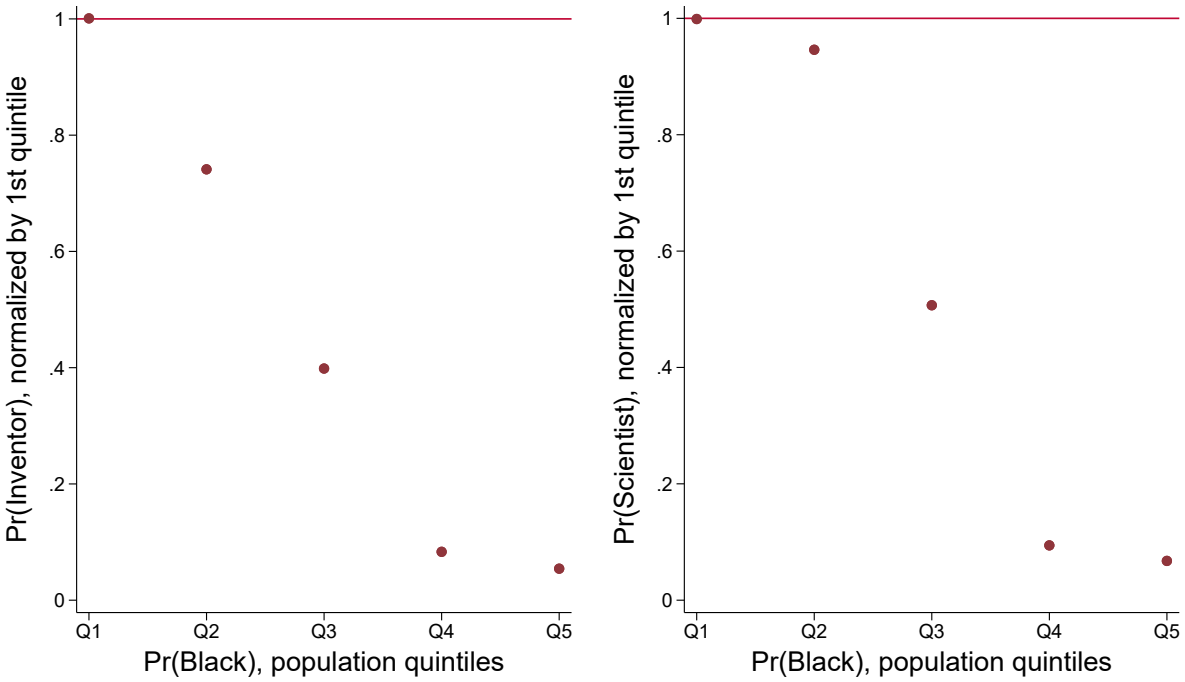
Figures

Figure 1: Pr(Black | Last name) in the US population, 2010 Census



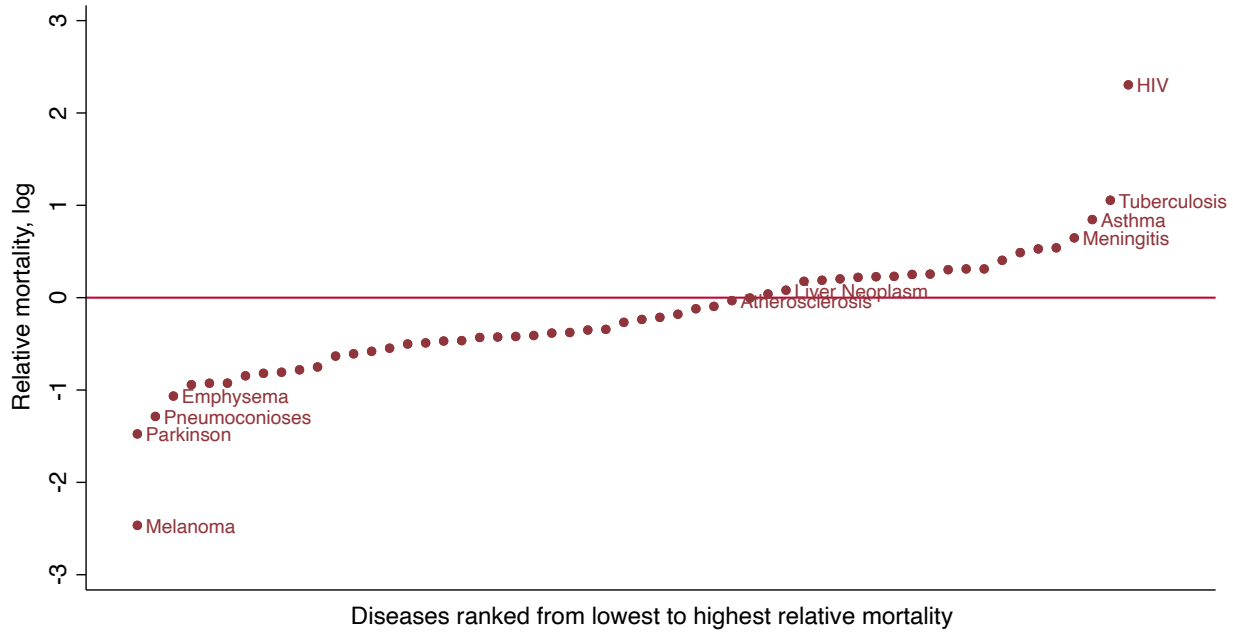
Notes. This figure plots the distribution of the US population across deciles of Pr(Black-sounding | Last name), using data from the 2010 US Census (Comenetz, 2016). One bar corresponds to the share of individuals in the US population who are in the given bin of probability of being Black or African American based on their last name.

Figure 2: Pr(Inventor) and Pr(Scientist) by quintiles of Pr(Black | Last name)



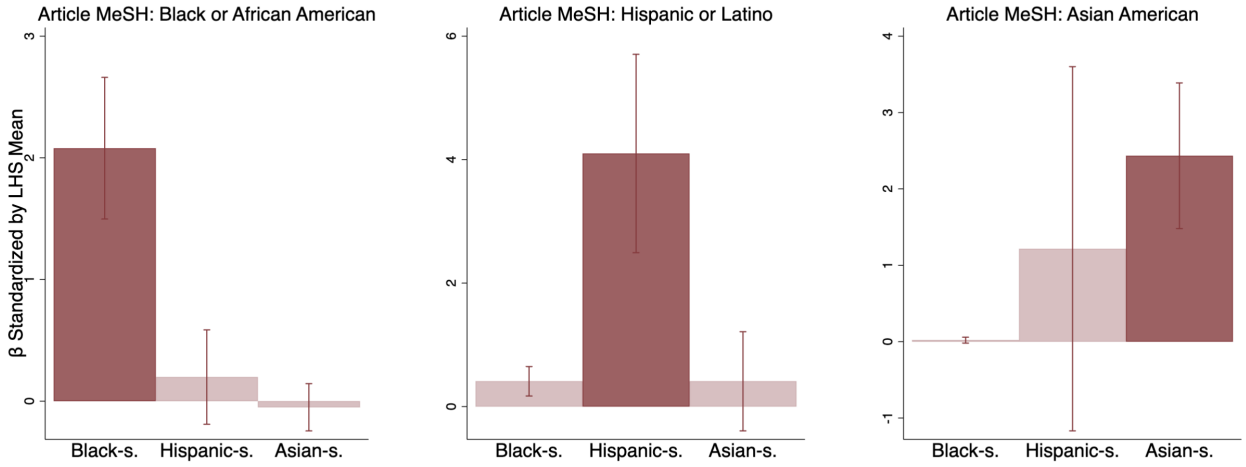
Notes. These plots report, respectively, the share of inventors (LHS) and the share of scientists (RHS), by quintile of Pr(Black | Last name). The shares are normalized by the first quintile. The share of inventors is computed on all US resident inventors listed as first author of a patent application submitted to the USPTO between January 2001 and December 2018. The share of scientists is computed on all US resident scientists listed as first author of a PubMed publication published between January 2002 and December 2018. On the x-axis, Pr(Black) refers to Pr(Black | Last name) from the 2010 US Census (Comenetz, 2016). The frequency of each last name in the US population is also taken from Comenetz (2016).

Figure 3: Causes of death ranked by relative mortality



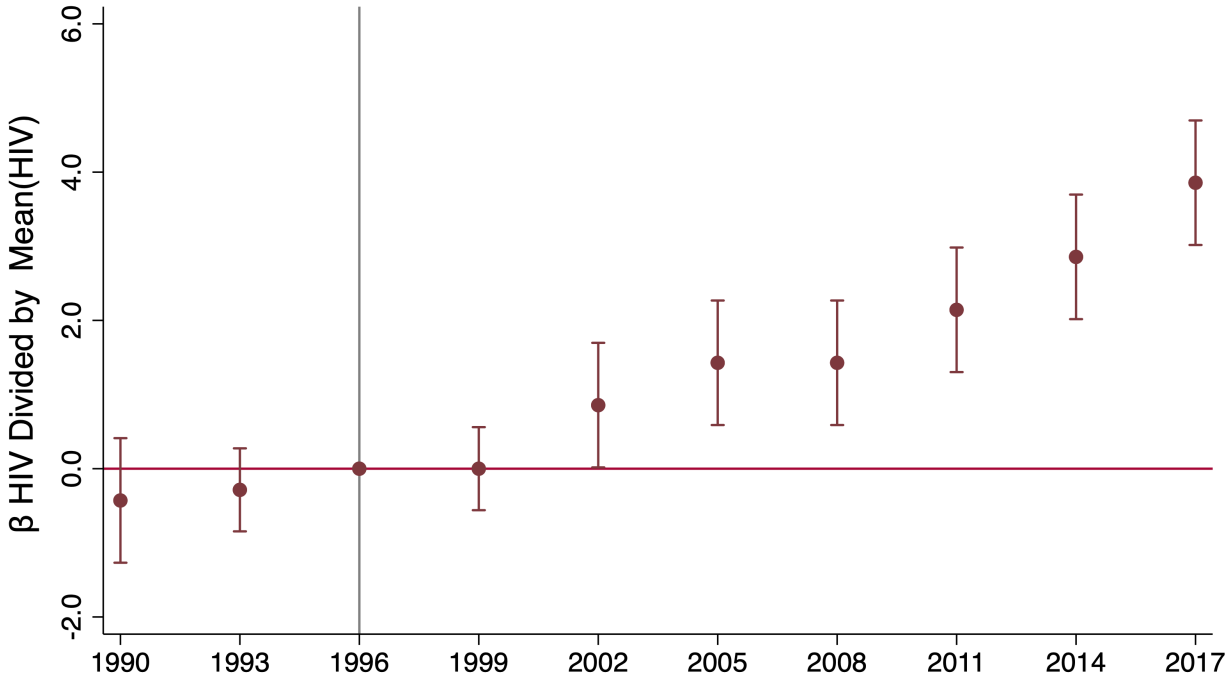
Notes. This figure displays relative mortality rate of Black individuals compared to white individuals in the United States in 2000 to 2016. One observation is one disease with at least 1,000 related deaths of white individuals and Black individuals over the full period. Data comes from the CDC. More details on the construction of this data are reported in section 2. On the x-axis, I rank disease by relative mortality rate of Black Americans vs white Americans. On the y-axis, I show the actual relative mortality rate of Black Americans compared to white Americans due to the given disease over the period. It is computed as total number of deaths of Black individuals reported due to the given disease divided but total number of Black individuals in the population, divided by total number of deaths of white individuals reported due to the given disease divided but total number of whites in the population, in log. The red line corresponds to equal mortality rate for Black individuals compared to white individuals (i.e. when y = 0). The sample is restricted to mortality reported between years 1999 and 2015, and includes diseases with at least 5,000 registered deaths (white Americans + Black Americans) over the period. The final sample includes 57 diseases. I report labels of the four diseases with highest relative mortality among whites, and of the four diseases with highest relative mortality among Black individuals.

Figure 4: Demographics-based approach, research on other demographic groups



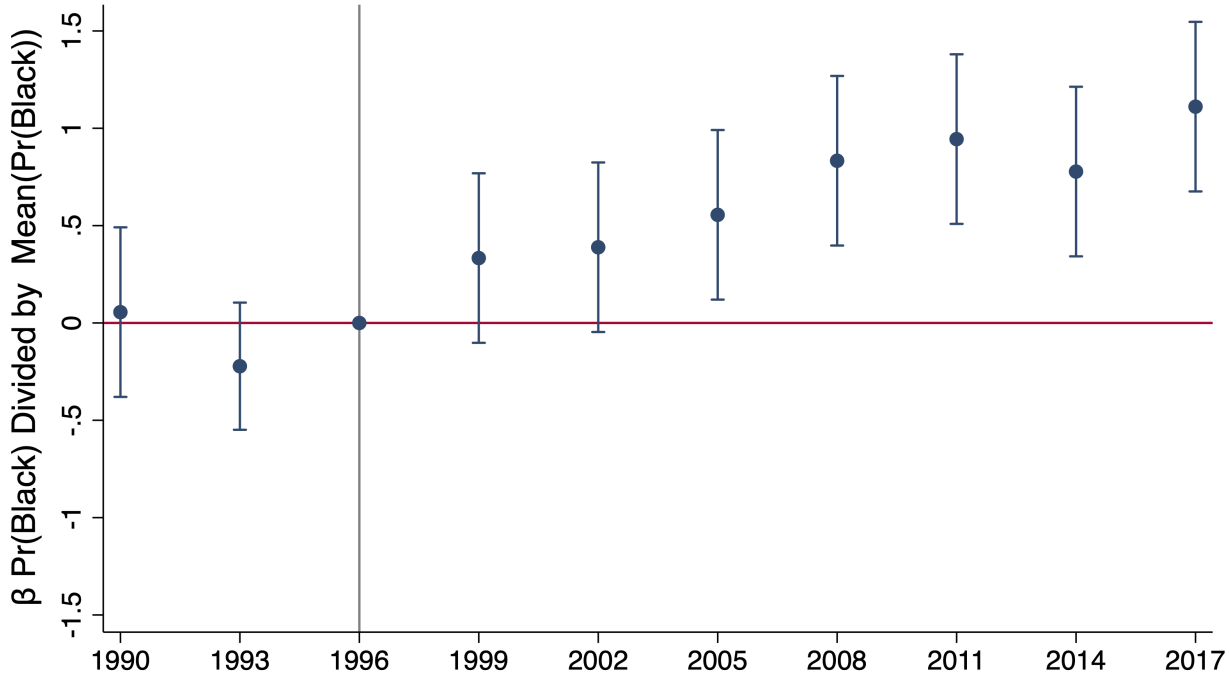
Notes. This plot reports the coefficients of column (1), Appendix Table B14, Panel A, Panel B, and Panel C. The bars plot 95% confidence intervals. S.e. are clustered by last name. The unit of observation is a research article in the PubMed database published between 2002 and 2018, with first author affiliated with a US institution, and published in a journal in the top 1,000 by Commercialization Impact Factor. Black-sounding name, Hispanic-sounding name, Asian-sounding name, and American Native-sounding name refer to the race or ethnicity of the first author listed on the publication. White-sounding name is the omitted category. All columns include year FE. The vector of racial frequencies by last name comes from the 2010 U.S. Census. The first plot from the left reports the result of estimating Equation 2 where the dependent variable is a dummy = 1 if the article has "Black or African American" among its MeSH codes, = 0 otherwise. The middle plot reports the result of estimating Equation 2 where the dependent variable is a dummy = 1 if the article has "Hispanic or Latino" among its MeSH codes, = 0 otherwise. The plot on the right reports the result of estimating Equation 2 where the dependent variable is a dummy = 1 if the article has "Asian American" among its MeSH codes, = 0 otherwise.

Figure 5: The impact of HAART on the likelihood that an HIV-related paper has "Black or African American" among its MeSH terms



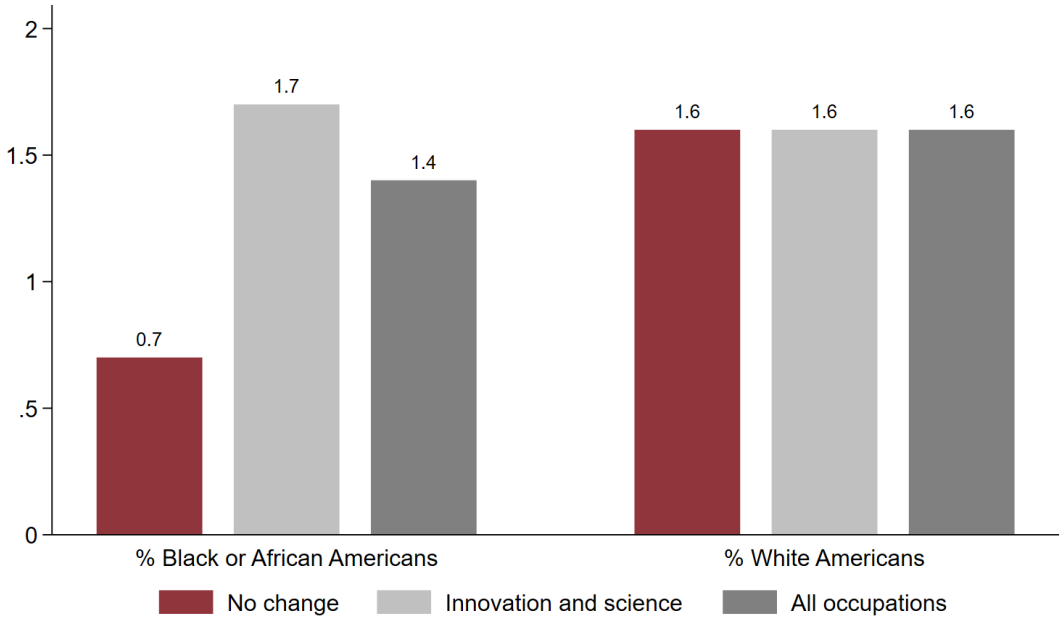
Notes. This plot reports the coefficients of an event study where the dependent variable id a dummy = 1 if the article has "Black or African American" among its MeSH terms, and = 0 otherwise. Time dummies (aggregate in three-year period) are interacted with a set of dummies = 1 if the article has "HIV" among its MeSH terms, = 0 otherwise. The sample includes all articles published by a first author residing in the US between 1988 and 2017. Standard errors are clustered at the last name level.

Figure 6: The impact of HAART on the match between a Black-sounding researcher and HIV research



Notes. This plot reports the coefficients of an event study where the dependent variable id a dummy = 1 if the article has "HIV" among its MeSH terms, and = 0 otherwise. Time dummies (aggregated in three-year periods) are interacted with Pr(Black | Last name) of the first author of the article. The vector of controls includes the other racial probabilities associate with the first author's last name, i.e. Hispanic-sounding last name, Asian-sounding last name, Am. Native-sounding last name, and this vector of racial probabilities interacted with time dummies. White-sounding name is the omitted category. The vector of racial frequencies by last name comes from the 2010 U.S. Census. The sample includes all articles published by a first author residing in the US between 1988 and 2017. Standard errors are clustered at the last name level.

Figure 7: Equalizing access to innovation and science, % of inventors and scientists



Notes. This figure reports the result of estimating the model outlined in section 6. The red bar (first and fourth bar from the left) reports the share of life scientists and engineers in an economy with race-specific frictions, estimated using Current Population Survey (CPS) data for 1995 to 2018. The first bar reports the share among whites (i.e., the number of white scientists and inventors normalized by the total number of whites). The fourth bar reports the share among Black individuals (i.e., the number of Black scientists and inventors normalized by the total number of Black individuals). The light gray bars (second and fifth bars) report, respectively, the share of white scientists and inventors, and the share of Black scientists and inventors in the economy after removing race-specific barriers to science an innovation. In this policy counterfactual, race-specific barriers to human capital accumulation and within the labor market are only removed in the innovation and science sector, and remain equal to the baseline in all other occupations. The dark gray bars (third and sixth bar), show respectively the results of a policy counterfactual where race-specific barriers to human capital accumulation and within the labor market are removed everywhere in the economy. The third bar from the left show the share of inventors and scientists among whites. The sixth bar from the left shows the share of inventor and scientists among Black individuals.

Tables

Table 1: Sample statistics, research articles and patents

Panel A: PubMed Articles

Variable	Mean	Std. dev.	Min.	Max.	N
MeSH: Black or African American	0.007	0.086	0.000	1.000	651,253
Relative mortality, log	-0.039	0.899	-2.465	2.304	138,657
Typically White	0.331	0.470	0.000	1.000	138,657
Similar Incidence	0.488	0.500	0.000	1.000	138,657
Typically Black	0.181	0.385	0.000	1.000	138,657
MeSH: Sickle Cell Anemia	0.001	0.036	0.000	1.000	651,253
MeSH: Thalassemia	0.000	0.019	0.000	1.000	651,253
MeSH: Melanoma	0.006	0.079	0.000	1.000	651,253
MeSH: Thyroid Neoplasm	0.002	0.042	0.000	1.000	651,253
Black-sounding last name	0.063	0.115	0.000	1.000	651,253
Hispanic-sounding last name	0.046	0.151	0.000	1.000	651,253
Asian-sounding last name	0.316	0.432	0.000	1.000	651,253
American Native-sounding last name	0.005	0.013	0.000	0.981	651,253
White-sounding last name	0.571	0.407	0.000	1.000	651,253

Panel B: USPTO Patent Applications

Variable	Mean	Std. dev.	Min.	Max.	N
Granted patent	0.717	0.450	0	1.000	2,009,485
Number of total citations	23.299	263.724	0	27522.000	1,483,345
Black-sounding name	0.080	0.120	0	1.000	2,009,485
Hispanic-sounding name	0.038	0.128	0	0.996	2,009,485
Asian-sounding name	0.193	0.368	0	0.997	2,009,485
American native-sounding name	0.005	0.012	0	0.978	2,009,485
White-sounding name	0.684	0.358	0	1.000	2,009,485

Notes. In Panel A, the unit of observation is a PubMed researched article published between 2002 and 2018. In Panel B, the unit of observation is a patent application with first inventor resident in the US filed at the the USPTO between 2001 and 2018. Black-sounding name, Hispanic-sounding name, Asian-sounding name, American-native sounding name, and white-sounding name refer to race or ethnicity of the first author listed on the patent application. Black-sounding name, Hispanic-sounding name, and Asian-sounding name, American-native sounding name, and white-sounding name refer to race or ethnicity of the first author listed on the publication.

Table 2: Demographics-based approach

	All	Linked to trial	Not linked to trial	All	Linked to trial	Not linked to trial
	(1)	(2)	(3)	(4)	(5)	(6)
	MeSH: Black or African American					
Black-sounding name	0.010***	0.025***	0.009***	0.028***	0.029***	0.028***
	(0.002)	(0.009)	(0.002)	(0.005)	(0.009)	(0.005)
Hispanic-sounding name	0.000	0.001	0.000	0.003	0.001	0.003
	(0.001)	(0.003)	(0.001)	(0.003)	(0.004)	(0.003)
Asian-sounding name	-0.003***	-0.002	-0.003***	0.000	-0.002	-0.000
	(0.000)	(0.001)	(0.000)	(0.001)	(0.002)	(0.001)
Am. Native-sounding name	-0.006	-0.011	-0.005	-0.015	-0.011	-0.018
	(0.007)	(0.029)	(0.007)	(0.016)	(0.031)	(0.016)
Article on human subjects				X	X	X
Observations	651,253	41,944	609,309	199,455	33,793	165,662
R-squared	0.001	0.001	0.001	0.001	0.002	0.001
LHS (mean)	0.007	0.010	0.007	0.019	0.010	0.021
RHS (mean)	0.063	0.072	0.063	0.072	0.072	0.071

Notes. Robust s.e. in parenthesis clustered by last name. The unit of observation is a research article in the PubMed database published between 2002 and 2018, with first author affiliated with a US institution, and published in a journal in the top 1,000 by Commercialization Impact Factor. Black-sounding name, Hispanic-sounding name, Asian-sounding name, and American Native-sounding name refer to the race or ethnicity of the first author listed on the publication. White-sounding name is the omitted category. All columns include year FE. The vector of racial frequencies by last name comes from the 2010 U.S. Census. Column (1) reports the result of estimating Equation 2 where the dependent variable is a dummy = 1 if the article has "Black or African American" among its MeSH codes, = 0 otherwise. Column (2) reports the results of estimating the same equation as in column (1), but on the subsample of articles linked to a clinical trial. Column (3) reports the results of estimating the same equation as in column (1), but on the subsample of articles not linked to a clinical trial. In columns (4) through (6), I report the results of columns (1) to (3) estimated on the subsample of articles focusing on human subjects.

Table 3: Frequency-based approach

	(1) Relative mortality, log	(2) Typically White	(3) Similar incidence	(4) Typically Black
Black-sounding name	0.122***	-0.048***	-0.002	0.050***
	(0.034)	(0.018)	(0.017)	(0.014)
Hispanic-sounding name	0.006	-0.005	0.000	0.005
	(0.024)	(0.014)	(0.012)	(0.009)
Asian-sounding name	-0.006	-0.022***	0.044***	-0.022***
	(0.009)	(0.005)	(0.005)	(0.004)
Am. Native-sounding name	-0.227	0.053	-0.021	-0.032
	(0.177)	(0.112)	(0.124)	(0.084)
Total mortality, log	-0.108***	-0.039***	0.139***	-0.100***
	(0.002)	(0.001)	(0.001)	(0.001)
Observations	138,657	138,657	138,657	138,657
R-squared	0.032	0.015	0.164	0.144
LHS (mean)	-0.039	0.331	0.488	0.181
RHS (mean)	0.064	0.064	0.064	0.064

Notes. Robust s.e. in parenthesis clustered by last name. The unit of observation is a research article in the PubMed database published between 2002 and 2018, with first author affiliated with a US institution, and published in a journal in the top 1,000 by Commercialization Impact Factor. Black-sounding name, Hispanic-sounding name, Asian-sounding name, and American Native-sounding name refer to race or ethnicity of the first author listed on the publication. White-sounding name is the omitted category. The vector of racial frequencies by last name comes from the 2010 U.S. Census. All columns control for the log of total mortality among white individuals and Black individuals over the period 1999 to 2016. All columns include year FE. Column (1) reports the result of estimating equation (3) where the dependent variable is a continuous variable equal to the log of relative mortality among Black Americans compared to white Americans, calculated according to equation 1. Column (2) reports the results of estimating equation 3 where the dependent variable is a dummy = 1 if relative mortality is at least 1.5 times higher among white Americans compared to Black Americans. Column (3) reports the results of estimating equation (3) where the dependent variable is a dummy = 1 if relative mortality is less than 1.5 among Black Americans compared to white Americans, and less than 1.5 among white Americans compared to Black Americans. Column (4) reports the results of estimating equation (3) where the dependent variable is a dummy = 1 if relative mortality is at least 1.5 times higher among Black individuals compared to white individuals.

Table 4: Ancestry Variation: Sickle cell anemia

	(1)	(2)	(3)	(4)
	MeSH: Sickle Cell Anemia		MeSH: Thalassemia	
Black-sounding name	0.001*	0.091***	-0.000	-0.011
	(0.001)	(0.034)	(0.000)	(0.010)
Hispanic-sounding name	-0.000	-0.015	-0.000	-0.011*
	(0.000)	(0.017)	(0.000)	(0.007)
Asian-sounding name	-0.000**	-0.004	-0.000	0.002
	(0.000)	(0.008)	(0.000)	(0.004)
Am. Native-sounding name	-0.006***	-0.232**	0.000	0.021
	(0.002)	(0.100)	(0.002)	(0.089)
Hemoglobin-related articles		X		X
Observations	651,253	13,297	651,253	13,297
R-squared	0.000	0.004	0.000	0.004
LHS (mean)	0.001	0.062	0.000	0.017
RHS (mean)	0.063	0.063	0.063	0.063

Notes. Robust s.e. in parenthesis clustered by last name. The unit of observation is a research article in the PubMed database published between 2002 and 2018, with first author affiliated with a US institution, and published in a journal in the top 1,000 by Commercialization Impact Factor. Black-sounding name, Hispanic-sounding name, Asian-sounding name, and American Native-sounding name refer to race or ethnicity of the first author listed on the publication. White-sounding name is the omitted category. All columns include year FE. The vector of racial frequencies by last name comes from the 2010 U.S. Census. Column (1) reports the result of estimating equation 2 where the dependent variable is a dummy = 1 if the article has sickle cell anemia among its MeSH codes, = 0 otherwise. Column (2) reports the results of estimating the same equation as in column (1), but on the subsample of hemoglobin-related diseases. "Hemoglobin-related" diseases are defined as those articles with hemoglobin among their MeSH codes. Column (3) reports the result of estimating equation 2 where the dependent variable is a dummy = 1 if the article has thalassemia among its MeSH codes, = 0 otherwise. Column (4) reports the results of estimating the same equation as in column (3), but on the subsample of hemoglobin-related diseases. Column (5) reports the result of estimating equation 2 where the dependent variable is a dummy = 1 if the article has hemophilia among its MeSH codes, = 0 otherwise. Column (6) reports the results of estimating the same equation as in column (5), but on the subsample of hemoglobin-related diseases.

Table 5: Ancestry Variation: Melanoma

	(1)	(2)	(3)	(4)
	MeSH: Melanoma		MeSH: Thyroid neoplasm	
Black-sounding name	-0.002*	-0.023**	0.000	0.001
	(0.001)	(0.011)	(0.001)	(0.005)
Hispanic-sounding name	0.001	-0.004	0.000	-0.000
	(0.001)	(0.009)	(0.000)	(0.003)
Asian-sounding name	0.000	-0.005	-0.000	-0.003**
	(0.000)	(0.003)	(0.000)	(0.001)
Am. Native-sounding name	0.000	0.040	-0.005*	-0.023
	(0.006)	(0.070)	(0.003)	(0.028)
Neoplasm-related articles		X		X
Observations	651,253	67,804	651,253	67,804
R-squared	0.000	0.002	0.000	0.001
LHS (mean)	0.006	0.060	0.002	0.011
RHS (mean)	0.063	0.061	0.063	0.061

Notes. Robust s.e. in parenthesis clustered by last name. The unit of observation is a research article in the PubMed database published between 2002 and 2018, with first author affiliated with a US institution, and published in a journal in the top 1,000 by Commercialization Impact Factor. Black-sounding name, Hispanic-sounding name, Asian-sounding name, and American Native-sounding name refer to race or ethnicity of the first author listed on the publication. White-sounding name is the omitted category. All columns include year FE. The vector of racial frequencies by last name comes from the 2010 U.S. Census. Column (1) reports the result of estimating equation (2) where the dependent variable is a dummy = 1 if the article has "Melanoma" among its MeSH codes, = 0 otherwise. Column (2) reports the results of estimating the same equation as in column (1), but on the subsample of neoplasm-related articles. "Neoplasm-related" articles are defined as those articles with neoplasm among their MeSH codes. Column (3) reports the result of estimating equation (2) where the dependent variable is a dummy = 1 if the article has "Thyroid neoplasm" among its MeSH codes, = 0 otherwise. Column (4) reports the results of estimating the same equation as in column (3), but on the subsample of neoplasm-related articles.

Table 6: DD around the introduction of HAART, NIH grants

	(1) Research on African Americans	(2) Research on HIV
HIV	0.004***	
	(0.001)	
HIV X Post HAART	0.020***	
	(0.003)	
Black-sounding name		0.004
		(0.007)
Black-sounding name X Post HAART		0.021**
		(0.008)
Vector of racial prob.		X
Vector of racial prob. X Post HAART		X
Observations	589,277	589,277
R-squared	0.006	0.005
LHS (mean)	0.042	0.042
RHS (mean)	0.083	0.083

Notes. Robust s.e. in parenthesis clustered by last name. The unit of observation is a research grant awarded by the National Institutes of Health between 1988 and 2017 by a first author affiliated with a US institution. Black-sounding name, Hispanic-sounding name, Asian-sounding name, and and American Native-sounding name refer to the last names of the first author listed on the publication. White-sounding name is the omitted category. All columns include FE for the year in which the grant was awarded. The vector of racial frequencies by last name comes from the 2010 U.S. Census. Column (1) reports the result of estimating equation (4). The dependent variable is a dummy = 1 if the research grant mentions Black or African Americans, = 0 otherwise. Post HAART is an indicator = 1 if the award grant year is > 1996, = 0 otherwise. Column (2) reports the result of estimating equation (5). The dependent variable is a dummy = 1 if the research grant mentions HIV, = 0 otherwise. The results in column (2) include a vector of controls including Hispanic-sounding name, Asian-sounding name, and American-Native-sounding name. It also includes these controls interacted with the Post HAART indicator. All columns include year FE.

Table 7: DD around the introduction of HAART, PubMed articles

	(1) Research on African Americans	(2) Research on HIV
HIV	0.004***	
	(0.001)	
HIV X Post HAART	0.020***	
	(0.003)	
Black-sounding name		0.004
		(0.007)
Black-sounding name X Post HAART		0.021**
		(0.008)
Vector of racial prob.		X
Vector of racial prob. X Post HAART		X
Observations	589,277	589,277
R-squared	0.006	0.005
LHS (mean)	0.042	0.042
RHS (mean)	0.083	0.083

Notes. Robust s.e. in parenthesis clustered by last name. The unit of observation is a research article in the PubMed database published between 1988 and 2017, with first author affiliated with a US institution, and published in a journal. Black-sounding name, Hispanic-sounding name, Asian-sounding name, and American Native-sounding name refer to the last names of the first author listed on the publication. White-sounding name is the omitted category. All columns include FE for the year in which the grant was awarded. The vector of racial frequencies by last name comes from the 2010 U.S. Census. Column (1) reports the result of estimating equation (4). The dependent variable is a dummy = 1 if the article has "Black or African Americans" among its MeSH terms, = 0 otherwise. Post HAART is an indicator = 1 if the award grant year is > 1996, = 0 otherwise. Column (2) reports the result of estimating equation (5). The dependent variable is a dummy = 1 if the research has "HIV" among its MeSH terms, = 0 otherwise. The results in column (2) include a vector of controls including Hispanic-sounding name, Asian-sounding name, and American-Native-sounding name. It also includes these controls interacted with the Post HAART indicator. All columns include year FE.

Table 8: Inventors with a Black-sounding name are less likely to be granted a patent

	(1)	(2)	(3)	(4)	(5)
			Pr(Granted)		
Black-sounding name	-0.060***	-0.064***	-0.037***	-0.037***	-0.029***
	(0.008)	(0.007)	(0.006)	(0.006)	(0.006)
Hispanic-sounding name	-0.052***	-0.049***	-0.033***	-0.034***	-0.031***
	(0.006)	(0.005)	(0.005)	(0.005)	(0.005)
Asian-sounding name	0.052***	0.043***	0.010***	0.008***	0.007***
	(0.003)	(0.003)	(0.002)	(0.002)	(0.002)
Am. Native-sounding name	-0.118**	-0.097*	-0.112**	-0.128**	-0.124***
	(0.055)	(0.051)	(0.050)	(0.050)	(0.048)
Average grant rate of legal team					0.312***
					(0.003)
Observations	2,009,485	2,009,485	2,009,485	2,009,283	2,009,283
R-squared	0.015	0.057	0.162	0.164	0.178
LHS (mean)	1.000	1.000	1.000	1.000	1.000
RHS (mean)	0.080	0.080	0.080	0.080	0.080

Notes. Robust s.e. clustered by last name in parenthesis. The unit of observation is a patent with first inventor resident in the US granted by the the USPTO between 2001 and 2018. Black-sounding name, Hispanic-sounding name, Asian-sounding name, and and American Native-sounding name refer to race or ethnicity of the first author listed on the publication. White-sounding name is the omitted category. The vector of racial frequencies by last name comes from the 2010 U.S. Census. In Column (1), I report the results of the estimation of equation (8) where y is a dummy = 1 if a patent is granted by the USPTO, = 0 otherwise. y is standardized to have mean 1. Its mean prior to standardization is equal to 0.718. In this specification, I control for year FE. In Column (2), I run the same specification as Column (1), but adding patent subclass FE. In Column (3), I add assignee FE. In column (4), I add state of residence FE. In column (5), I add controls for the "quality" of the legal team that handled the patent application. I control for the total number of lawyers listed on the patent, and on the average grant rate of the legal team, computed as the average of the individual average grant rate of each lawyer listed on the patent.

Table 9: Patents granted to inventors with a Black-sounding name have higher impact

	(1)	(2)	(3)	(4)	(5)
	N. forward citations				
Black-sounding name	0.968*	0.785**	0.812**	0.812**	0.816**
	(0.542)	(0.397)	(0.374)	(0.374)	(0.378)
Hispanic-sounding name	-0.449***	-0.336***	-0.281***	-0.281***	-0.279***
	(0.154)	(0.100)	(0.097)	(0.097)	(0.099)
Asian-sounding name	-0.371***	-0.212***	-0.265***	-0.265***	-0.272***
	(0.098)	(0.048)	(0.047)	(0.047)	(0.047)
Am. Native-sounding name	-0.961	-0.145	0.056	0.056	-0.067
	(1.725)	(1.309)	(1.287)	(1.287)	(1.298)
Average grant rate of legal team					0.317***
					(0.117)
Observations	1,483,345	1,483,345	1,445,735	1,445,735	1,445,735
LHS (mean)	23.676	23.718	23.986	23.986	23.986
RHS (mean)	0.078	0.078	0.078	0.078	0.078

Notes. Robust s.e. clustered by last name in parenthesis. The unit of observation is a patent with first inventor resident in the US granted by the the USPTO between 2001 and 2018. Black-sounding name, Hispanic-sounding name, Asian-sounding name, and American Native-sounding name refer to race or ethnicity of the first author listed on the publication. White-sounding name is the omitted category. The vector of racial frequencies by last name comes from the 2010 U.S. Census. In Column (1), I report the results of the estimation of a equation (8) using a Poisson count model where y is the total number of forward citations received by the patent. In this specification, I control for year FE. In Column (2), I run the same specification as Column (1), but adding patent subclass FE. In Column (3), I add assignee FE. In column (4), I add state of residence FE. In column (5), I add controls for the "quality" of the legal team that handled the patent application. I control for the total number of lawyers listed on the patent, and on the average grant rate of the legal team, computed as the average of the individual average grant rate of each lawyer listed on the patent.

ONLINE APPENDIX for "Race and Science"

June 2024

A Appendix Figures

Figure B1: Pr(Black | Last name), US Census vs Florida

Figure B2: Pr(Black) | Last name), US Census vs Florida scientists and inventors

Figure B3: Correlation between Pr(Black | Last name) of first and last author

Figure B4: Example of a PubMed article with Black or African American among its MeSH codes, but would not be recognized with a dictionary approach

Figure B5: Example of a PubMed article classified as addressing Black or African Americans with a dictionary classification, but would not be recognized using MeSH codes

Figure B6: Demographics-based approach, additional split by publication type

Figure B7: Relative mortality, weighted by total number of deaths

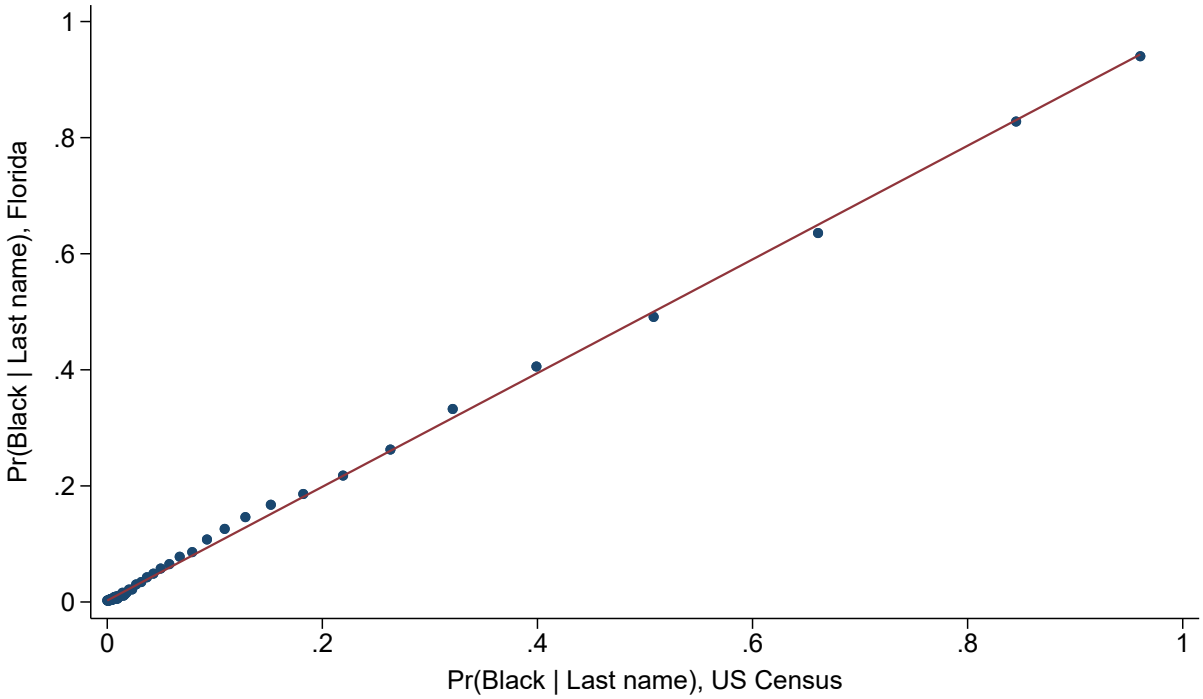
Figure B8: Incidence of sickle cell anemia by region of the world

Figure B9: Incidence of melanoma by country of the world

Figure B10: Mortality rates of White and Black/African Americans, Melanoma

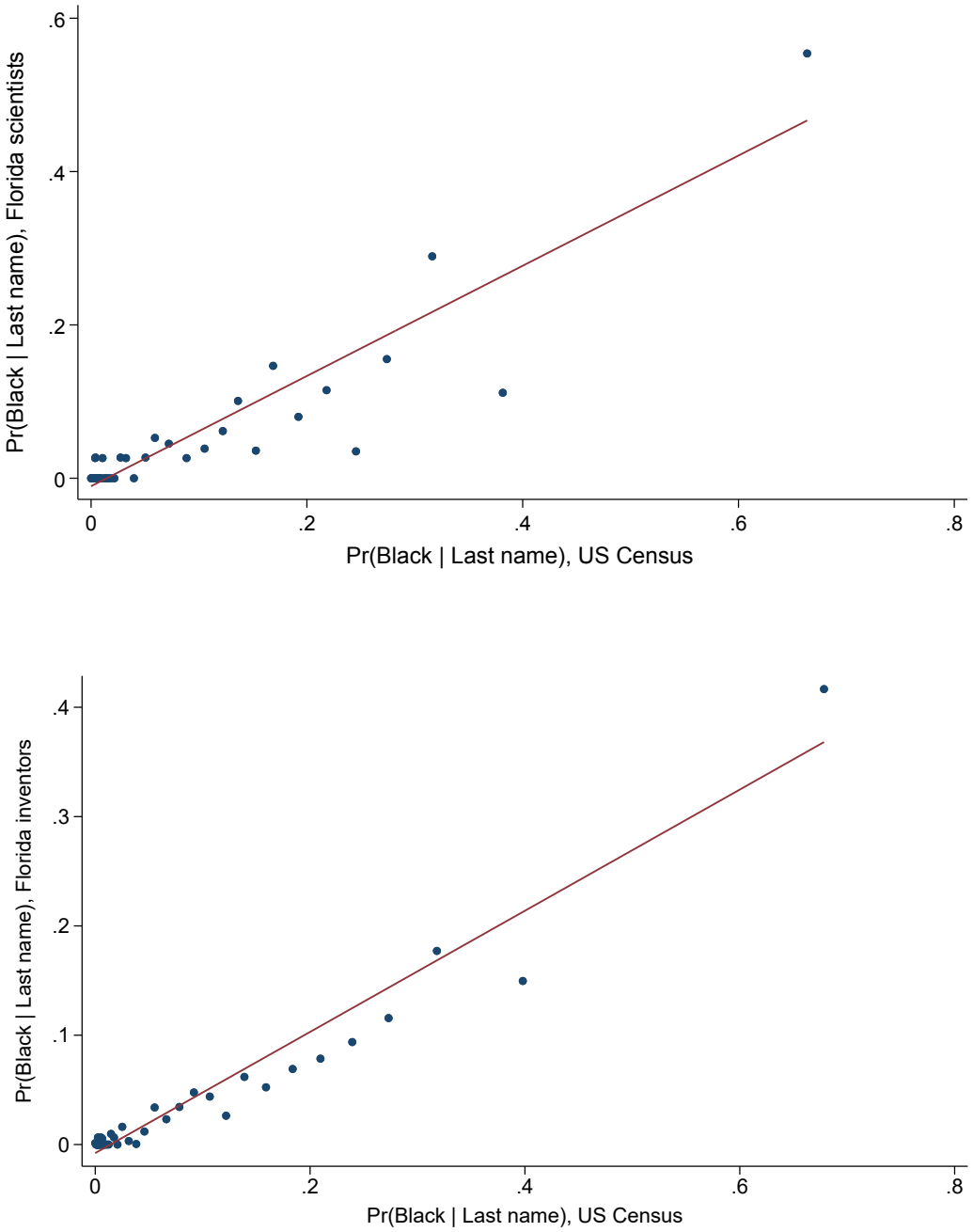
Figure B11: Mortality rates of White and Black/African Americans, Thyroid Neoplasm

Figure B1: Pr(Black | Last name), US Census vs Florida



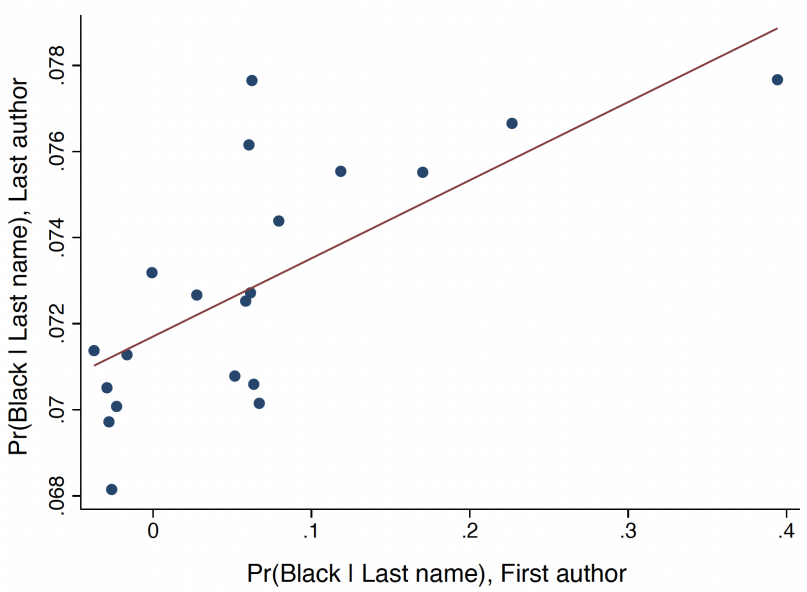
Notes. One observation is one last name. The underlying data is the sample of registered voters in Florida (Dossi and Morando, 2023) merged with information on racial frequency by last name from Comenetz (2016). This Figure reports a binned scatter plot with Pr(Black | Last name) on the x-axis, and the share of individuals who reported race = Black, by last name, on the y-axis. $\hat{\beta} = 0.979$, robust s.e.= 0.004.

Figure B2: Pr(Black) | Last name), US Census vs Florida scientists and inventors



Notes. One observation is one last name. The underlying data is the sample of registered voters in Florida (Dossi and Morando, 2023) who are inventors, merged with information on racial frequency by last name from Comenetz (2016). Details on the match between voter register data and inventors is reported in Dossi and Morando (2023). This Figure reports a binned scatter plot with Pr(Black | Last name) on the x-axis, and the share of inventors who reported race = Black, by last name, on the y-axis.

Figure B3: Correlation between Pr(Black | Last name) of first and last author



Notes. This figure displays the correlation between Pr(Black | Last name) of first authors and of last authors in the sample of PubMed publications described in Table 1. The plot includes controls for Pr(Hispanic | Last name), Pr(Asian | Last name), Pr(Am. Native | Last name) of both first and last author. The sample excludes single-authored publications.

Figure B4: Example of a PubMed article with Black or African American among its MeSH codes, but which would not be recognized with a dictionary approach



Clinical Trial > Anesthesiology. 2005 Apr;102(4):715-9.
doi: 10.1097/00000542-200504000-00004.

Effects of skin pigmentation on pulse oximeter accuracy at low saturation

Philip E Bickler [1], John R Feiner, John W Severinghaus

Affiliations + expand
PMID: 15791098 DOI: 10.1097/00000542-200504000-00004
Free article

Abstract

Background: It is uncertain whether skin pigmentation affects pulse oximeter accuracy at low HbO2 saturation.

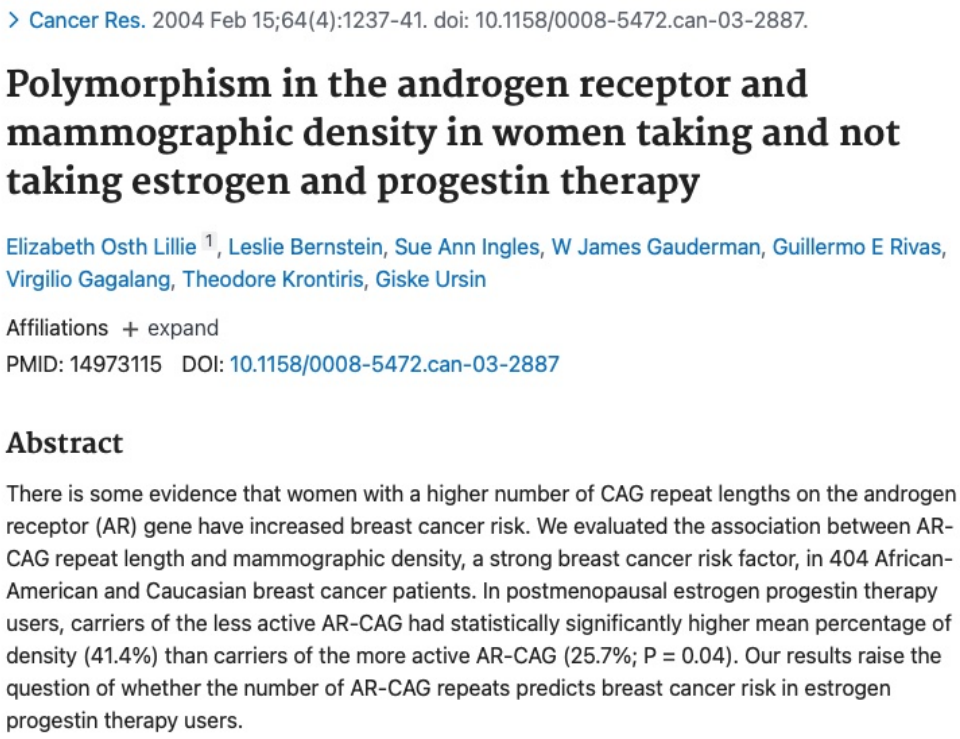
Methods: The accuracy of finger pulse oximeters during stable, plateau levels of arterial oxygen saturation (Sao2) between 60 and 100% were evaluated in 11 subjects with darkly pigmented skin and in 10 with light skin pigmentation. Oximeters tested were the Nellcor N-595 with the OxiMax-A probe (Nellcor Inc., Pleasanton, CA), the Novametrix 513 (Novametrix Inc., Wallingford, CT), and the Nonin Onyx (Nonin Inc., Plymouth, MN). Semisupine subjects breathed air-nitrogen-carbon dioxide mixtures through a mouthpiece. A computer used end-tidal oxygen and carbon dioxide concentrations determined by mass spectrometry to estimate breath-by-breath Sao2, from which an operator adjusted inspired gas to rapidly achieve 2- to 3-min stable plateaus of desaturation. Comparisons of oxygen saturation measured by pulse oximetry (Spo2) with Sao2 (by Radiometer OSM3) were used in a multivariate model to determine the interrelation between saturation, skin pigmentation, and oximeter bias (Spo2 - Sao2).

Results: At 60-70% Sao2, Spo2 (mean of three oximeters) overestimated Sao2 (bias +/- SD) by 3.56 +/- 2.45% (n = 29) in darkly pigmented subjects, compared with 0.37 +/- 3.20% (n = 58) in lightly pigmented subjects (P < 0.0001). The SD of bias was not greater with dark than light skin. The dark-light skin differences at 60-70% Sao2 were 2.35% (Nonin), 3.38% (Novametrix), and 4.30% (Nellcor). Skin pigment-related differences were significant with Nonin below 70% Sao2, with Novametrix below 90%, and with Nellcor at all ranges. Pigment-related bias increased approximately in proportion to desaturation.

Conclusions: The three tested pulse oximeters overestimated arterial oxygen saturation during hypoxia in dark-skinned individuals.

Notes. This is an example of a PubMed article of the dataset which has "Black or African American" among its MeSH codes, but would *not* be classified as such using a dictionary approach. This is because its abstract does not explicitly mention "Black American", "African American", or even "Black".

Figure B5: Example of a PubMed article classified as addressing Black or African Americans using a dictionary classification, but which would not be recognized using MeSH codes



> Cancer Res. 2004 Feb 15;64(4):1237-41. doi: 10.1158/0008-5472.can-03-2887.

Polymorphism in the androgen receptor and mammographic density in women taking and not taking estrogen and progestin therapy

Elizabeth Osth Lillie [1], Leslie Bernstein, Sue Ann Ingles, W James Gauderman, Guillermo E Rivas, Virgilio Gagalang, Theodore Krontiris, Giske Ursin

Affiliations + expand

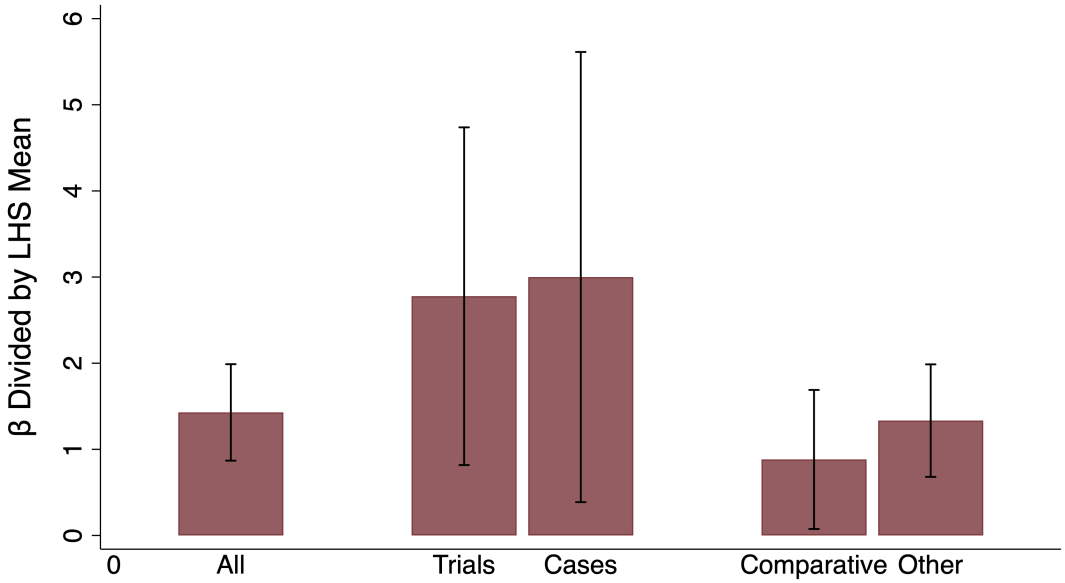
PMID: 14973115 DOI: 10.1158/0008-5472.can-03-2887

Abstract

There is some evidence that women with a higher number of CAG repeat lengths on the androgen receptor (AR) gene have increased breast cancer risk. We evaluated the association between AR-CAG repeat length and mammographic density, a strong breast cancer risk factor, in 404 African-American and Caucasian breast cancer patients. In postmenopausal estrogen progestin therapy users, carriers of the less active AR-CAG had statistically significantly higher mean percentage of density (41.4%) than carriers of the more active AR-CAG (25.7%; P = 0.04). Our results raise the question of whether the number of AR-CAG repeats predicts breast cancer risk in estrogen progestin therapy users.

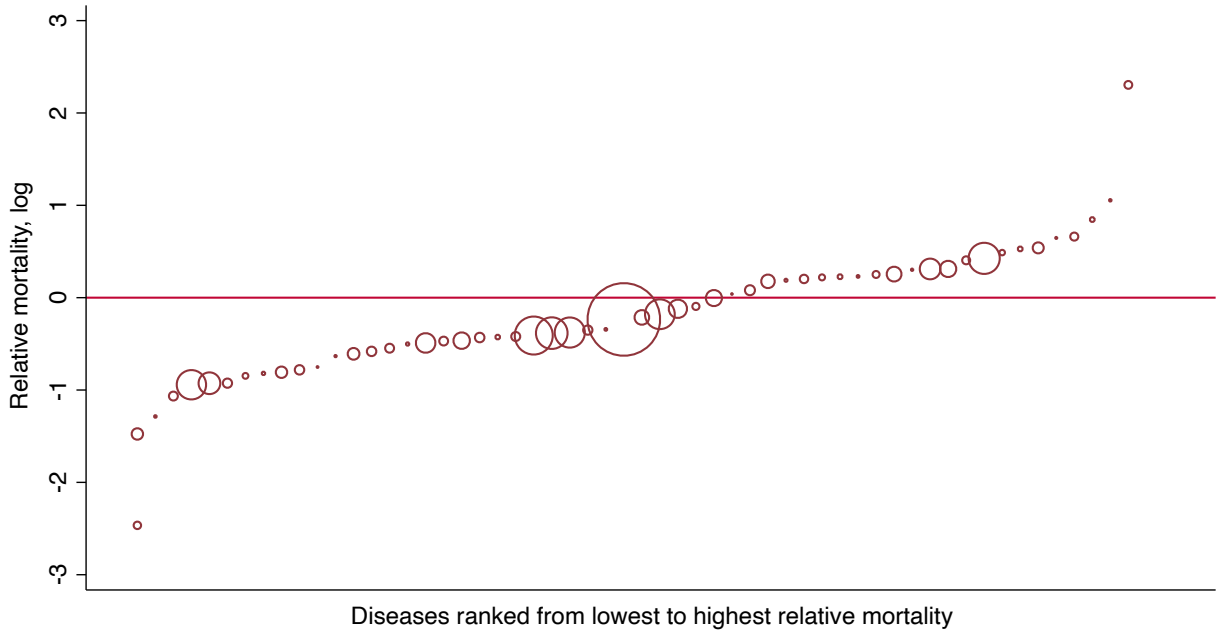
Notes. This is an example of a PubMed article of the dataset which does not have "Black or African American" among its MeSH codes. This article is classified as benefit Black or African Americans because the term "African American" appears in its abstract.

Figure B6: Demographics-based approach, additional split by publication type



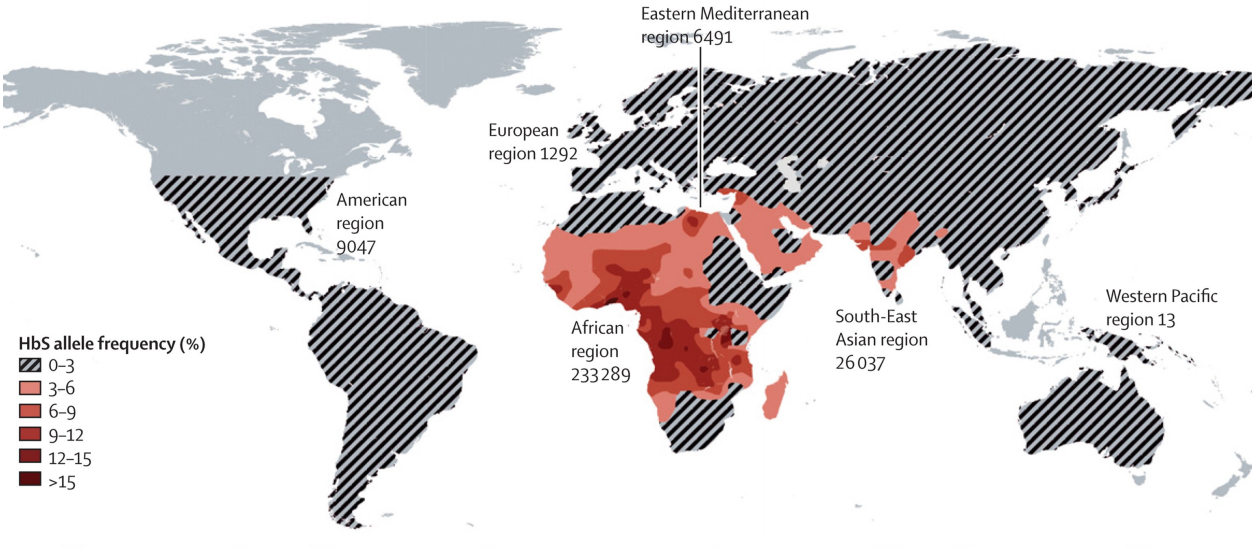
Notes. This figure reports the coefficients of Black-sounding last name for the estimation of equation 2 on the sample shown in column (1), Table 2, and split by publication type. Publication type is provided by PubMed. In this plot, the first bar on the left reports the coefficient of Black-sounding last name in column (1), Table 2, divided by the mean of the dependent variable. The second bar ("Trials") reports the coefficient of Black-sounding last name for equation 2 estimated on the subsample of articles linked to a clinical trial. The third bar ("Cases") reports the coefficient of Black-sounding last name for equation 2 estimated on the subsample of case reports. The fourth bar ("Comparative") reports the coefficient of Black-sounding last name for equation 2 estimated on the subsample of comparative studies. The fifth bar ("Other") reports the coefficient of Black-sounding last name for equation 2 estimated on all other articles..)

Figure B7: Relative mortality, weighted by total number of deaths



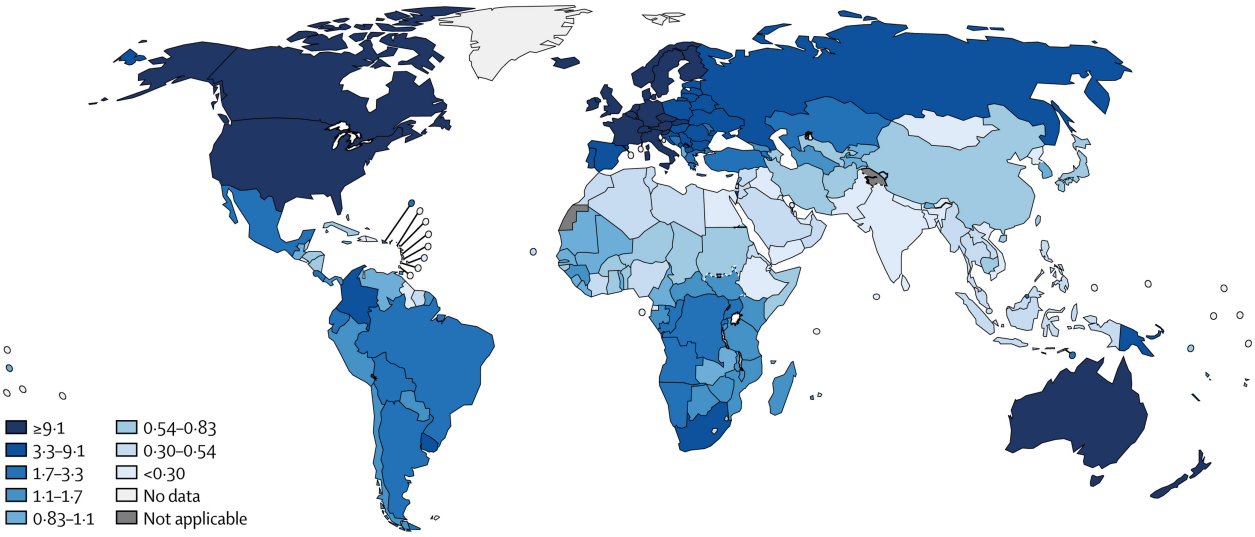
Notes. This figure displays relative mortality rate of Black individuals compared to white individuals in the United States in 2000 to 2015. It is equivalent to Figure 3 reported in the main text with the difference that observations are weighted by the total number of deaths by the given disease over the period. One observation is one disease with at least 1,000 related deaths of white individuals and Black individuals over the full period. Data comes from the CDC. More details on the construction of this data in section ??. On the x-axis, I rank disease by relative mortality rate of Black individuals vs white individuals. On the y-axis is the relative mortality rate of Black individuals compared to white individuals due to the given disease over the period. It is computed as total number of deaths of Black individuals reported due to the given disease divided but total number of Black individuals in the population, divided by total number of deaths of white individuals reported due to the given disease divided but total number of white individuals in the population, in log. The red line corresponds to equal mortality rate for Black individuals compared to white individuals (i.e. when y = 0).

Figure B8: Incidence of sickle cell anemia by region of the world



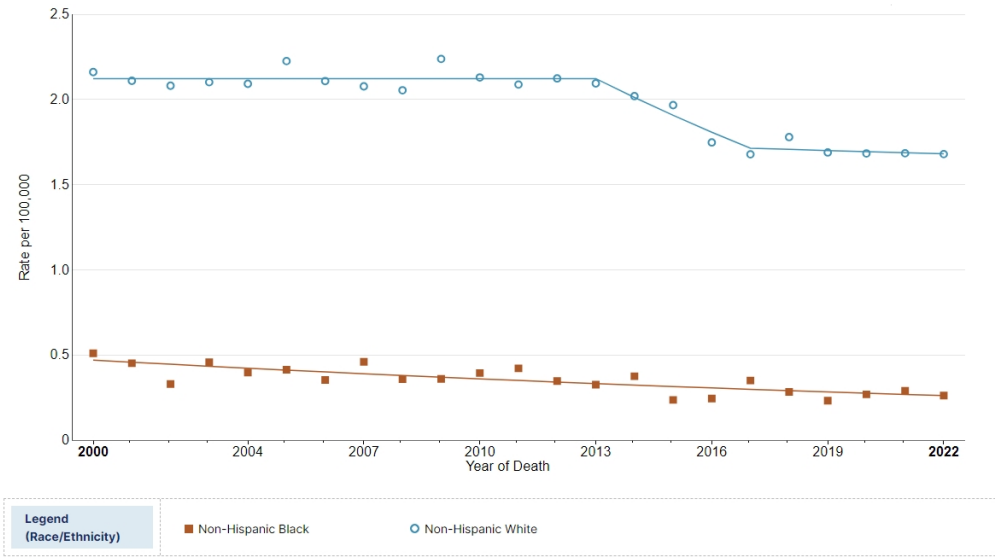
Notes. Incidence of sickle cell anemia by world region (Rees, Williams and Gladwin, 2010).

Figure B9: Incidence of melanoma by country of the world



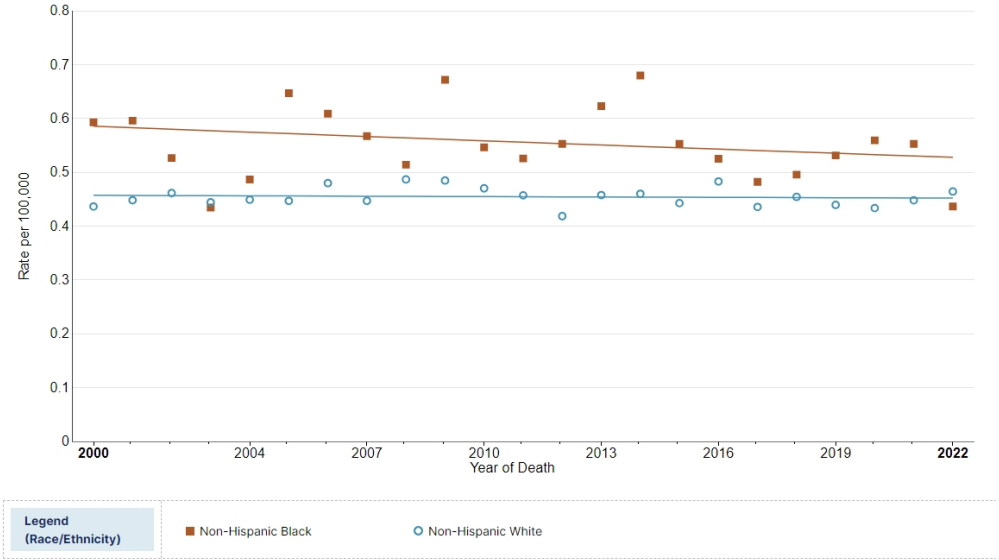
Notes. Incidence of melanoma by world region (Schadendorf, van Akkooi, Berking, Griewank, Gutzmer, Hauschild, Stang, Roesch and Ugurel, 2018).

Figure B10: Mortality rates of White and Black/African Americans, Melanoma



Notes. Source: U.S. Mortality Data (1969-2022), National Center for Health Statistics, CDC.

Figure B11: Mortality rates of White and Black/African Americans, Thyroid Neoplasm



Notes. Source: U.S. Mortality Data (1969-2022), National Center for Health Statistics, CDC..

B Appendix Tables

B.I Validation of Last Names as a Proxy for Race

Table B1: Extract of racial probabilities by last name, 2010 Census

Table B2: Validation of Black-sounding names as a predictor of race, Florida

Table B3: Differences in observables between matched and unmatched scientists

Table B4: Differences in observables between matched and unmatched inventors

B.II Demographics-based Approach, Additional Tables

Table B5: Demographics-based approach, dictionary classification

Table B6: Demographics-based approach, logit model

Table B7: Demographics-based approach, controlling for gender of first author

Table B8: Demographics-based approach, controlling for Journal FE, State FE, first author affiliation FE

Table B9: Demographics-based approach, alternative definitions of Black-sounding name

Table B10: Demographics-based approach, including all articles (not only top 1,000 journals by Journal Commercialization Impact Factor (JCIF)

Table B11: Demographics-based approach, split by below/above Journal Commercialization Impact Factor (JCIF)

Table B12: Demographics-based approach, articles linked to a patent

Table B13: Demographics-based approach, controlling for MeSH codes linked to article

Table B14: Demographics-based approach, research on other demographic groups

Table B15: Demographics-based approach: Evidence from ongoing clinical trials

B.III Frequency-based approach, Additional Tables

Table B16: Frequency-based approach, dictionary classification

Table B17: Frequency-based approach, logit model

Table B18: Frequency-based approach, controlling for gender of first author

Table B19: Frequency-based approach, controlling for Journal FE, State FE, first author affiliation FE

Table B20: Frequency-based approach, alternative definitions of Black-sounding name

Table B21: Frequency-based approach, including all articles (not only top 1,000 journals by Journal Commercialization Impact Factor (JCIF)

Table B22: Frequency-based approach, split by below/above Journal Commercialization Impact Factor (JCIF)

Table B23: Frequency-based approach, articles linked to a patent

Table B24: Frequency-based approach: Threshold set at 1.3 higher mortality

Table B25: Frequency-based approach: Threshold set at 1.7 higher mortality

Table B26: Frequency-based approach: Alternative definitions of frequency

B.IV Comparison with Gender

Table B27: Demographics-based Approach, match by gender

Table B28: Frequency-based approach, match by gender

B.V Impact

Table B29: Demographics-based approach, Total Citations

Table B30: Demographics-based approach, Citations (Relative Citation Ratio)

Table B31: Demographics-based approach, Total Citations

Table B32: Frequency-based approach, Citations (Relative Citation Ratio)

Table B33: Demographics-based approach, split by below/above median Journal Impact Factor (JIF)

Table B34: Frequency-based approach, split by below/above median Journal Impact Factor (JIF)

Table B1: Extract of racial probabilities by last name, 2010 Census

Last name	White	Black	Hispanic	Asian	Am. Native
Top 10 last names by Pr(Black | Last name)					
WASHINGTON	5%	91%	3%	0%	1%
JEFFERSON	18%	77%	3%	0%	2%
BOOKER	29%	68%	2%	0%	0%
BANKS	40%	56%	3%	0%	0%
JOSEPH	30%	56%	3%	10%	1%
MOSLEY	42%	55%	2%	0%	1%
JACKSON	41%	55%	3%	0%	1%
CHARLES	35%	54%	8%	1%	2%
DORSEY	43%	54%	2%	0%	0%
RIVERS	42%	52%	3%	1%	2%
Top 10 last names by Pr(White | Last name)					
YODER	98%	0%	1%	0%	0%
FRIEDMAN	97%	0%	2%	1%	0%
KRUEGER	97%	0%	2%	1%	0%
SCHWARTZ	97%	0%	2%	1%	0%
SCHMITT	97%	0%	2%	1%	0%
MUELLER	97%	0%	2%	1%	0%
WEISS	96%	0%	2%	1%	0%
NOVAK	96%	0%	2%	1%	0%
OCONNELL	97%	0%	2%	1%	0%
KLEIN	97%	0%	2%	1%	0%

Notes. This table reports the vector of racial probabilities from the US Census (2010) for the 10 last names with highest probability of being Black, and the 10 last names with highest probability of being white among the 1,000 most frequent last names. *white* refers to % non-Hispanic or Latino white alone; *Black* refers to % non-Hispanic or Latino Black or African American alone; *Hispanic* refers to percent Hispanic or Latino origin; *Asian* refers to percent non-Hispanic or Latino Asian and Native Hawaiian and other Pacific Islander alone; *Am. Native* refers to percent non-Hispanic or Latino American Indian and Alaska Native alone. Probabilities do not always add up to one due to censoring of those cells with less than 100 observation in the population, and due to the residual category *two or more races*. They are rescaled to add up to 1. Data source: Comenetz (2016).

Table B2: Validation of Black-sounding names as a predictor of race, Florida

	(1)	(2)	(3)	(4)	(5)
			Race = Black		
Black-sounding name	1.021***	1.019***	0.956***	0.937***	0.919***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Asian-sounding name		0.006***	0.013***	0.015***	0.015***
		(0.001)	(0.001)	(0.001)	(0.001)
Hispanic-sounding name		-0.000	-0.027***	-0.035***	-0.046***
		(0.000)	(0.000)	(0.000)	(0.000)
Am. Native-sounding name		0.325***	0.256***	0.234***	0.223***
		(0.007)	(0.007)	(0.007)	(0.007)
Median zip code income (log)			-0.197***	-0.194***	-0.194***
			(0.000)	(0.000)	(0.000)
Average zip code income by last name (log)				-0.066***	-0.150***
				(0.002)	(0.002)
Average zip code income by last name (sd)					0.000***
					(0.000)
Observations	10,167,678	10,167,678	10,133,338	10,133,338	10,126,273
R-squared	0.245	0.246	0.276	0.276	0.276
LHS (mean)	0.141	0.141	0.141	0.141	0.141
RHS (mean)	0.132	0.132	0.132	0.132	0.132

Notes. One observation is a registered voter in the state of Florida in 2017 (Dossi and Morando, 2023). Black-sounding name, Hispanic-sounding name, and Asian-sounding name, and American Native-sounding name are the probability that the individual is Black, Hispanic, Asian, or American Native based on their last name, and is assigned based on the 2010 U.S. Census (Comenetz, 2016). White-sounding name is the omitted category. The vector of racial frequencies by last name comes from the 2010 U.S. Census. Median zip code income (log) is the average median income in the zip code of residence of the voter. Average zip code income by last name (log) and Average zip code income by last name (sd) are, respectively, is the average median zip code income by last name, computed on the population of registered voters in Florida in 2017, and its standard deviation. The dependent variable is a dummy = 1 if the voter's reported race or ethnicity is "Black or African American", = 0 otherwise.

Table B3: Differences in observables between matched and unmatched scientists

	Matched		Unmatched		Matched-Unmatched	
	Mean	Standard Deviation	Mean	Standard Deviation	Standardized Difference	P-value Equivalence Test
	(1)	(2)	(3)	(4)	(5)	(6)
Year	2011.000	4.952	2011.000	4.897	0.000	0.000
Team size	2.723	2.018	2.781	2.022	-0.029	0.000
Granted patent	0.716	0.451	0.719	0.450	-0.007	0.000
Forward citations	23.510	262.600	20.240	141.000	0.013	0.000
Small entity	0.341	0.474	0.320	0.467	0.044	0.000
Assignee	0.293	0.455	0.297	0.457	-0.011	0.000
Female	0.067	0.250	0.081	0.273	-0.055	0.000
Male	0.811	0.391	0.792	0.406	0.049	0.000
Unassigned gender	0.059	0.235	0.074	0.261	-0.063	0.000
Unknown gender	0.063	0.243	0.054	0.225	0.040	0.000
California	0.249	0.433	0.305	0.460	-0.127	1.000
Texas	0.070	0.256	0.059	0.236	0.045	0.000
New York	0.060	0.237	0.072	0.258	-0.049	0.000
Massachusetts	0.048	0.213	0.058	0.233	-0.047	0.000
Washington state	0.042	0.202	0.043	0.203	-0.003	0.000
Michigan	0.039	0.193	0.041	0.198	-0.012	0.000
Illinois	0.036	0.186	0.037	0.189	-0.006	0.000
New Jersey	0.032	0.176	0.042	0.201	-0.056	0.000
Florida	0.031	0.174	0.030	0.171	0.007	0.000
Pennsylvania	0.031	0.173	0.030	0.172	0.004	0.000
Ohio	0.030	0.172	0.026	0.158	0.028	0.000
Minnesota	0.031	0.172	0.025	0.155	0.036	0.000
North Carolina	0.025	0.156	0.019	0.138	0.036	0.000
Colorado	0.022	0.146	0.017	0.130	0.032	0.000
Connecticut	0.018	0.134	0.019	0.138	-0.008	0.000
Georgia	0.019	0.137	0.014	0.119	0.036	0.000
Other states	0.216	0.412	0.163	0.369	0.133	1.000

Notes. One observation is a patent application filed at the USPTO between 2001 and 2018 with first inventors resident resident in the United States. "Matched" refers to patents whose first inventors has a match in the Census distribution of last names by race. "Unmatched" refers to patents that are not matched. Descriptive statistics (mean and standard deviation) of inventors matched to a last name from Comenetz (2016) (Columns 1 & 2) and unmatched (Columns 3 & 4). Column 5 shows the standardized difference between matched and unmatched in the full sample of inventors. Column 6 reports the largest p-value for the equivalence test of means using a two one-sided t-tests approach. The null hypothesis is that the difference is larger than 10% of a sd, or smaller than 10% of a sd. The share of matched patents is equal to 82%.

Table B4: Differences in observables between matched and unmatched inventors

	Matched		Unmatched		Matched-Unmatched	
	Mean (1)	Standard Deviation (2)	Mean (3)	Standard Deviation (4)	Standardized Difference (5)	P-value Equivalence Test (6)
Year	2011.000	4.952	2011.000	4.897	0.000	0.000
Team size	2.723	2.018	2.781	2.022	-0.029	0.000
Granted patent	0.716	0.451	0.719	0.450	-0.007	0.000
Forward citations	23.510	262.600	20.240	141.000	0.013	0.000
Small entity	0.341	0.474	0.320	0.467	0.044	0.000
Assignee	0.293	0.455	0.297	0.457	-0.011	0.000
Female	0.067	0.250	0.081	0.273	-0.055	0.000
Male	0.811	0.391	0.792	0.406	0.049	0.000
Unassigned gender	0.059	0.235	0.074	0.261	-0.063	0.000
Unknown gender	0.063	0.243	0.054	0.225	0.040	0.000
California	0.249	0.433	0.305	0.460	-0.127	1.000
Texas	0.070	0.256	0.059	0.236	0.045	0.000
New York	0.060	0.237	0.072	0.258	-0.049	0.000
Massachusetts	0.048	0.213	0.058	0.233	-0.047	0.000
Washington state	0.042	0.202	0.043	0.203	-0.003	0.000
Michigan	0.039	0.193	0.041	0.198	-0.012	0.000
Illinois	0.036	0.186	0.037	0.189	-0.006	0.000
New Jersey	0.032	0.176	0.042	0.201	-0.056	0.000
Florida	0.031	0.174	0.030	0.171	0.007	0.000
Pennsylvania	0.031	0.173	0.030	0.172	0.004	0.000
Ohio	0.030	0.172	0.026	0.158	0.028	0.000
Minnesota	0.031	0.172	0.025	0.155	0.036	0.000
North Carolina	0.025	0.156	0.019	0.138	0.036	0.000
Colorado	0.022	0.146	0.017	0.130	0.032	0.000
Connecticut	0.018	0.134	0.019	0.138	-0.008	0.000
Georgia	0.019	0.137	0.014	0.119	0.036	0.000
Other states	0.216	0.412	0.163	0.369	0.133	1.000

Notes. One observation is a patent application filed at the USPTO between 2001 and 2018 with first inventors resident resident in the United States. "Matched" refers to patents whose first inventors has a match in the Census distribution of last names by race. "Unmatched" refers to patents that are not matched. Descriptive statistics (mean and standard deviation) of inventors matched to a last name from Comenetz (2016) (Columns 1 & 2) and unmatched (Columns 3 & 4). Column 5 shows the standardized difference between matched and unmatched in the full sample of inventors. Column 6 reports the largest p-value for the equivalence test of means using a two one-sided t-tests approach. The null hypothesis is that the difference is larger than 10% of a sd, or smaller than 10% of a sd. The share of matched patents is equal to 82%.

Table B5: Demographics-based approach, dictionary classification

	All	Linked to trial	Not linked to trial	All	Linked to trial	Not linked to trial
	(1)	(2)	(3)	(4)	(5)	(6)
	Black or African American (dictionary classification on abstract)					
Black-sounding name	0.008***	0.023***	0.007***	0.022***	0.027***	0.021***
	(0.002)	(0.009)	(0.001)	(0.004)	(0.009)	(0.004)
Hispanic-sounding name	-0.000	-0.003	0.000	0.003	-0.002	0.003
	(0.001)	(0.003)	(0.001)	(0.002)	(0.003)	(0.003)
Asian-sounding name	-0.002***	0.000	-0.002***	0.002**	0.001	0.002
	(0.000)	(0.002)	(0.000)	(0.001)	(0.002)	(0.001)
Am. Native-sounding name	-0.002	-0.046**	0.001	-0.004	-0.036*	-0.002
	(0.006)	(0.020)	(0.007)	(0.015)	(0.019)	(0.018)
Article on human subjects				X	X	X
Observations	651,253	41,944	609,309	199,455	33,793	165,662
R-squared	0.000	0.001	0.000	0.001	0.001	0.001
LHS (mean)	0.007	0.010	0.006	0.017	0.010	0.018
RHS (mean)	0.063	0.072	0.063	0.072	0.072	0.071

Notes. Robust s.e. in parenthesis clustered by last name. The unit of observation is a research article in the PubMed database published between 2002 and 2018, with first author affiliated with a US institution, and published in a journal in the top 1,000 by Commercialization Impact Factor. Black-sounding name, Hispanic-sounding name, Asian-sounding name, and American Native-sounding name refer to the race or ethnicity of the first author listed on the publication. White-sounding name is the omitted category. All columns include year FE. The vector of racial frequencies by last name comes from the 2010 U.S. Census. Column (1) reports the result of estimating Equation 2 where the dependent variable is a dummy = 1 if the article abstract contains either the word "African American", or the word "Black" jointly with "race", "racial", or "ethnic", = 0 otherwise. Column (2) reports the results of estimating the same equation as in column (1), but on the subsample of articles linked to a clinical trial. Column (3) reports the results of estimating the same equation as in column (1), but on the subsample of articles not linked to a clinical trial. In columns (4) through (6), I report the results of columns (1) to (3) estimated on the subsample of articles focusing on human subjects.

Table B6: Demographics-based approach, logit model

	All	Linked to trial	Not linked to trial	All	Linked to trial	Not linked to trial
	(1)	(2)	(3)	(4)	(5)	(6)
	MeSH: Black or African American					
Black-sounding name	1.001***	1.875***	0.916***	1.212***	2.002***	1.094***
	(0.155)	(0.473)	(0.151)	(0.162)	(0.452)	(0.161)
Hispanic-sounding name	0.001	0.108	-0.005	0.148	0.111	0.131
	(0.119)	(0.348)	(0.123)	(0.131)	(0.388)	(0.136)
Asian-sounding name	-0.527***	-0.234	-0.535***	-0.014	-0.306	-0.038
	(0.057)	(0.178)	(0.058)	(0.059)	(0.209)	(0.061)
Am. Native-sounding name	-0.558	-0.474	-0.556	-0.739	-0.289	-0.905
	(1.095)	(3.509)	(1.076)	(1.166)	(3.439)	(1.162)
Article on human subjects				X	X	X
Observations	651,253	41,944	609,309	199,455	33,793	165,662
LHS (mean)	0.007	0.010	0.007	0.019	0.010	0.021
RHS (mean)	0.063	0.072	0.063	0.072	0.072	0.071

Notes. Robust s.e. in parenthesis clustered by last name. The unit of observation is a research article in the PubMed database published between 2002 and 2018, with first author affiliated with a US institution, and published in a journal in the top 1,000 by Commercialization Impact Factor. Black-sounding name, Hispanic-sounding name, Asian-sounding name, and American Native-sounding name refer to the race or ethnicity of the first author listed on the publication. White-sounding name is the omitted category. All columns include year FE. The vector of racial frequencies by last name comes from the 2010 U.S. Census. Column (1) reports the result of estimating Equation 2 where the dependent variable is a dummy = 1 if the article has "Black or African American" among its MeSH codes, = 0 otherwise. Column (2) reports the results of estimating the same equation as in column (1), but on the subsample of articles linked to a clinical trial. Column (3) reports the results of estimating the same equation as in column (1), but on the subsample of articles not linked to a clinical trial. In columns (4) through (6), I report the results of columns (1) to (3) estimated on the subsample of articles focusing on human subjects. I estimate logit models in columns (1) through (6). The table reports exp(coefficients).

Table B7: Demographics-based approach, controlling for gender of first author

	All	Linked to trial	Not linked to trial	All	Linked to trial	Not linked to trial
	(1)	(2)	(3)	(4)	(5)	(6)
	MeSH: Black or African American					
Black-sounding name	0.011***	0.026***	0.010***	0.028***	0.029***	0.027***
	(0.002)	(0.009)	(0.002)	(0.005)	(0.010)	(0.005)
Hispanic-sounding name	0.000	0.003	0.000	0.005*	0.003	0.005
	(0.001)	(0.004)	(0.001)	(0.003)	(0.004)	(0.003)
Asian-sounding name	-0.003***	-0.002	-0.003***	-0.001	-0.003	-0.001
	(0.000)	(0.002)	(0.000)	(0.001)	(0.002)	(0.001)
Am. Native-sounding name	-0.007	-0.006	-0.007	-0.018	-0.008	-0.022
	(0.008)	(0.034)	(0.007)	(0.018)	(0.035)	(0.018)
Female dummy	0.004***	0.007***	0.004***	0.014***	0.008***	0.015***
	(0.000)	(0.001)	(0.000)	(0.001)	(0.002)	(0.001)
Article on human subjects				X	X	X
Observations	539,707	38,032	501,675	177,170	30,777	146,393
R-squared	0.001	0.003	0.001	0.003	0.003	0.003
LHS (mean)	0.008	0.010	0.008	0.019	0.010	0.021
RHS (mean)	0.069	0.074	0.069	0.074	0.074	0.074

Notes. Robust s.e. in parenthesis clustered by last name. The unit of observation is a research article in the PubMed database published between 2002 and 2018, with first author affiliated with a US institution, and published in a journal in the top 1,000 by Commercialization Impact Factor. Black-sounding name, Hispanic-sounding name, Asian-sounding name, and American Native-sounding name refer to the race or ethnicity of the first author listed on the publication. White-sounding name is the omitted category. All columns include year FE. The vector of racial frequencies by last name comes from the 2010 U.S. Census. Female dummy is a variable = 1 if the gender of the first author estimated from their first name is female, = 0 otherwise. The omitted category is male first author. Column (1) reports the result of estimating Equation 2 where the dependent variable is a dummy = 1 if the article has "Black or African American" among its MeSH codes, = 0 otherwise. Column (2) reports the results of estimating the same equation as in column (1), but on the subsample of articles linked to a clinical trial. Column (3) reports the results of estimating the same equation as in column (1), but on the subsample of articles not linked to a clinical trial. In columns (4) through (6), I report the results of columns (1) to (3) estimated on the subsample of articles focusing on human subjects.

Table B8: Demographics-based approach, controlling for Journal FE, State FE, Affiliation FE

	All	Linked to trial	Not linked to trial	All	Linked to trial	Not linked to trial
	(1)	(2)	(3)	(4)	(5)	(6)
			MeSH: Black or African American			
Panel A: Journal FE						
Black-sounding name	0.011***	0.024***	0.010***	0.029***	0.027***	0.029***
	(0.002)	(0.009)	(0.002)	(0.004)	(0.009)	(0.005)
Observations	630,029	41,398	588,556	194,839	33,331	161,439
R-squared	0.031	0.031	0.035	0.036	0.033	0.042
LHS (mean)	0.008	0.010	0.007	0.020	0.010	0.022
RHS (mean)	0.063	0.072	0.063	0.071	0.072	0.071
Panel B: State FE						
Black-sounding name	0.009***	0.025***	0.008***	0.008***	0.029***	0.025***
	(0.002)	(0.009)	(0.002)	(0.002)	(0.009)	(0.005)
Observations	649,249	41,748	607,501	607,501	33,629	164,912
R-squared	0.002	0.005	0.002	0.002	0.006	0.006
LHS (mean)	0.007	0.010	0.007	0.007	0.010	0.021
RHS (mean)	0.063	0.072	0.063	0.063	0.072	0.071
Panel C: Affiliation of first author FE						
Black-sounding name	0.008***	0.047**	0.006***	0.031***	0.052***	0.027***
	(0.002)	(0.018)	(0.002)	(0.007)	(0.019)	(0.007)
Observations	342,843	17,353	325,379	89,198	13,659	75,425
R-squared	0.008	0.033	0.008	0.020	0.047	0.021
LHS (mean)	0.007	0.011	0.007	0.021	0.012	0.023
RHS (mean)	0.062	0.072	0.062	0.071	0.073	0.071

Notes. Robust s.e. in parenthesis clustered by last name. The unit of observation is a research article in the PubMed database published between 2002 and 2018, with first author affiliated with a US institution, and published in a journal in the top 1,000 by Commercialization Impact Factor. All columns include year FE and a vector of racial frequencies. Black-sounding name, Hispanic-sounding name, Asian-sounding name, and American Native-sounding name refer to the race or ethnicity of the first author listed on the publication. White-sounding name is the omitted category. The vector of racial frequencies by last name comes from the 2010 U.S. Census. Column (1) reports the result of estimating Equation 2 where the dependent variable is a dummy = 1 if the article has "Black or African American" among its MeSH codes, = 0 otherwise. Column (2) reports the results of estimating the same equation as in column (1), but on the subsample of articles linked to a clinical trial. Column (3) reports the results of estimating the same equation as in column (1), but on the subsample of articles not linked to a clinical trial. In columns (4) through (6), I report the results of columns (1) to (3) estimated on the subsample of articles focusing on human subjects. In Panel A, all columns include journal FE. In Panel B, all columns include State FE, where state refers to the US state of the affiliation of the first author. In Panel C, all columns include affiliation of first author FE.

Table B9: Demographics-based approach, alternative definitions of Black-sounding last name

	All	Linked to trial	Not linked to trial	All	Linked to trial	Not linked to trial
	(1)	(2)	(3)	(4)	(5)	(6)
			MeSH: Black or African American			
Panel A: Black-sounding last name of first and last author						
Black-sounding name	0.009***	0.026**	0.008***	0.025***	0.030***	0.023***
	(0.002)	(0.011)	(0.002)	(0.005)	(0.011)	(0.005)
Black-sounding name (Last author)	0.008***	0.017***	0.008***	0.025***	0.019**	0.025***
	(0.002)	(0.006)	(0.002)	(0.004)	(0.007)	(0.005)
Article on human subjects				X	X	X
Observations	501,771	29,063	472,708	149,926	23,309	126,617
R-squared	0.001	0.002	0.001	0.001	0.003	0.001
LHS (mean)	0.007	0.009	0.007	0.019	0.010	0.021
RHS (mean)	0.064	0.072	0.064	0.073	0.073	0.073
Panel B: Average of Pr(Black | Last name) across all authors						
Black-sounding name	0.010***	0.024***	0.009***	0.029***	0.027***	0.028***
	(0.002)	(0.009)	(0.002)	(0.005)	(0.009)	(0.005)
Article on human subjects				X	X	X
Observations	621,326	40,962	580,364	189,148	32,987	156,161
R-squared	0.001	0.001	0.001	0.001	0.001	0.001
LHS (mean)	0.008	0.010	0.007	0.020	0.010	0.022
RHS (mean)	0.063	0.072	0.063	0.072	0.073	0.071
Panel C: Dummy for Pr(Black | Last name) ≥ 0.5						
Black-sounding name (Binary)	0.011***	0.012	0.011***	0.026***	0.014	0.028***
	(0.002)	(0.012)	(0.002)	(0.006)	(0.012)	(0.007)
Article on human subjects				X	X	X
Observations	638,409	41,348	597,061	195,895	33,322	162,573
R-squared	0.001	0.001	0.001	0.001	0.001	0.001
LHS (mean)	0.007	0.010	0.007	0.019	0.010	0.021
RHS (mean)	0.009	0.009	0.009	0.010	0.009	0.010
Panel D: Restricting sample to last names present in the 1930 Census						
Black-sounding name	0.009***	0.022***	0.008***	0.026***	0.027***	0.026***
	(0.002)	(0.007)	(0.002)	(0.005)	(0.008)	(0.005)
Article on human subjects				X	X	X
Observations	507,339	35,656	471,683	164,658	28,763	135,895
R-squared	0.000	0.001	0.000	0.001	0.001	0.001
LHS (mean)	0.008	0.010	0.008	0.020	0.011	0.022
RHS (mean)	0.076	0.080	0.075	0.081	0.080	0.081

Notes. Robust s.e. in parenthesis clustered by last name. The unit of observation is a research article in the PubMed database published between 2002 and 2018, with first author affiliated with a US institution, and published in a journal in the top 1,000 by Commercialization Impact Factor. All columns include year FE and a vector of racial frequencies. Black-sounding name, Hispanic-sounding name, Asian-sounding name, and American Native-sounding name refer to the race or ethnicity of the first author listed on the publication. White-sounding name is the omitted category. The vector of racial frequencies by last name comes from the 2010 U.S. Census. Column (1) reports the result of estimating Equation 2 where the dependent variable is a dummy = 1 if the article has "Black or African American" among its MeSH codes, = 0 otherwise. Column (2) reports the results of estimating the same equation as in column (1), but on the subsample of articles linked to a clinical trial. Column (3) reports the results of estimating the same equation as in column (1), but on the subsample of articles not linked to a clinical trial. In columns (4) through (6), I report the results of columns (1) to (3) estimated on the subsample of articles focusing on human subjects. In Panel A, all columns include journal FE. In Panel B, all columns include State FE, where state refers to the US state of the affiliation of the first author. In Panel C, all columns include affiliation of first author FE.

Table B10: Demographics-based approach, including all articles (not only top 1,000 journals by Journal Commercialization Impact Factor (JCIF)

	All	Linked to trial	Not linked to trial	All	Linked to trial	Not linked to trial
	(1)	(2)	(3)	(4)	(5)	(6)
	MeSH: Black or African American					
Black-sounding name	0.034***	0.051***	0.033***	0.045***	0.053***	0.045***
	(0.001)	(0.006)	(0.001)	(0.002)	(0.007)	(0.002)
Hispanic-sounding name	0.004***	0.002	0.004***	0.006***	0.003	0.006***
	(0.001)	(0.003)	(0.001)	(0.001)	(0.003)	(0.001)
Asian-sounding name	-0.003***	-0.001	-0.002***	-0.002***	-0.002	-0.002***
	(0.000)	(0.001)	(0.000)	(0.000)	(0.001)	(0.000)
Am. Native-sounding name	-0.006	-0.021	-0.005	-0.008	-0.022	-0.008
	(0.007)	(0.017)	(0.008)	(0.009)	(0.017)	(0.010)
Article on human subjects				X	X	X
Observations	2,274,170	99,314	2,174,856	1,624,717	95,828	1,528,889
R-squared	0.002	0.004	0.002	0.003	0.004	0.003
LHS (mean)	0.009	0.014	0.009	0.013	0.015	0.013
RHS (mean)	0.072	0.077	0.071	0.074	0.077	0.074

Notes. Robust s.e. in parenthesis clustered by last name. The unit of observation is a research article in the PubMed database published between 2002 and 2018, with first author affiliated with a US institution. Black-sounding name, Hispanic-sounding name, Asian-sounding name, and American Native-sounding name refer to the race or ethnicity of the first author listed on the publication. White-sounding name is the omitted category. All columns include year FE. The vector of racial frequencies by last name comes from the 2010 U.S. Census. Column (1) reports the result of estimating Equation 2 where the dependent variable is a dummy = 1 if the article has "Black or African American" among its MeSH codes, = 0 otherwise. Column (2) reports the results of estimating the same equation as in column (1), but on the subsample of articles linked to a clinical trial. Column (3) reports the results of estimating the same equation as in column (1), but on the subsample of articles not linked to a clinical trial. In columns (4) through (6), I report the results of columns (1) to (3) estimated on the subsample of articles focusing on human subjects.

Table B11: Demographics-based approach, split by below/above median Journal Commercialization Impact Factor

	All	Linked to trial	Not Linked to trial	All	Linked to trial	Not Linked to trial
	(1)	(2)	(3)	(4)	(5)	(6)
	MeSH: Black or African American					
Panel A: Journal Commercialization Impact Factor (JCIF) below median in the sample						
Black-sounding name	0.015***	0.037***	0.012***	0.026***	0.040***	0.023***
	(0.003)	(0.012)	(0.003)	(0.005)	(0.011)	(0.005)
Observations	253,951	22,606	231,345	122,497	18,983	103,514
R-squared	0.001	0.003	0.001	0.001	0.003	0.001
LHS (mean)	0.010	0.010	0.010	0.018	0.010	0.019
RHS (mean)	0.067	0.072	0.067	0.072	0.072	0.072
Panel B: Journal Commercialization Impact Factor (JCIF) above median in the sample						
Black-sounding name	0.007***	0.011	0.007***	0.033***	0.015	0.037***
	(0.002)	(0.008)	(0.002)	(0.007)	(0.010)	(0.008)
Observations	397,302	19,338	377,964	76,958	14,810	62,148
R-squared	0.001	0.002	0.001	0.001	0.002	0.001
LHS (mean)	0.006	0.009	0.005	0.022	0.010	0.025
RHS (mean)	0.061	0.071	0.060	0.070	0.072	0.070

Notes. Robust s.e. in parenthesis clustered by last name. The unit of observation is a research article in the PubMed database published between 2002 and 2018, with first author affiliated with a US institution, and published in a journal in the top 1,000 by Commercialization Impact Factor. The sample is further restricted to those articles linked to a patent (Marx and Fuegi, 2020, 2022). Black-sounding name, Hispanic-sounding name, Asian-sounding name, and American Native-sounding name refer to the race or ethnicity of the first author listed on the publication. White-sounding name is the omitted category. All columns include year FE. The vector of racial frequencies by last name comes from the 2010 U.S. Census. Column (1) reports the result of estimating Equation 2 where the dependent variable is a dummy = 1 if the article has "Black or African American" among its MeSH codes, = 0 otherwise. Column (2) reports the results of estimating the same equation as in column (1), but on the subsample of articles linked to a clinical trial. Column (3) reports the results of estimating the same equation as in column (1), but on the subsample of articles not linked to a clinical trial. In columns (4) through (6), I report the results of columns (1) to (3) estimated on the subsample of articles focusing on human subjects.

Table B12: Demographics-based approach, articles linked to a patent

	All	Linked to trial	Not linked to trial	All	Linked to trial	Not linked to trial
	(1)	(2)	(3)	(4)	(5)	(6)
	MeSH: Black or African American					
Black-sounding name	0.004**	0.008	0.004*	0.025***	0.008	0.030***
	(0.002)	(0.007)	(0.002)	(0.009)	(0.008)	(0.012)
Hispanic-sounding name	-0.001	-0.004***	-0.001	-0.005	-0.004***	-0.006
	(0.001)	(0.001)	(0.001)	(0.004)	(0.001)	(0.006)
Asian-sounding name	-0.002***	-0.000	-0.002***	-0.001	-0.003**	-0.003
	(0.000)	(0.002)	(0.000)	(0.002)	(0.001)	(0.002)
Am. Native-sounding name	0.000	0.000	-0.000	0.004	-0.016	-0.000
	(0.007)	(0.024)	(0.007)	(0.032)	(0.018)	(0.038)
Article on human subjects				X	X	X
Observations	184,146	10,057	174,089	29,140	7,348	21,792
R-squared	0.000	0.001	0.000	0.002	0.002	0.002
LHS (mean)	0.003	0.004	0.003	0.014	0.003	0.017
RHS (mean)	0.058	0.070	0.058	0.067	0.071	0.066

Notes. Robust s.e. in parenthesis clustered by last name. The unit of observation is a research article in the PubMed database published between 2002 and 2018, with first author affiliated with a US institution, and published in a journal in the top 1,000 by Commercialization Impact Factor. The sample is further restricted to those articles linked to a patent (Marx and Fuegi, 2020, 2022). Black-sounding name, Hispanic-sounding name, Asian-sounding name, and American Native-sounding name refer to the race or ethnicity of the first author listed on the publication. White-sounding name is the omitted category. All columns include year FE. The vector of racial frequencies by last name comes from the 2010 U.S. Census. Column (1) reports the result of estimating Equation 2 where the dependent variable is a dummy = 1 if the article has "Black or African American" among its MeSH codes, = 0 otherwise. Column (2) reports the results of estimating the same equation as in column (1), but on the subsample of articles linked to a clinical trial. Column (3) reports the results of estimating the same equation as in column (1), but on the subsample of articles not linked to a clinical trial. In columns (4) through (6), I report the results of columns (1) to (3) estimated on the subsample of articles focusing on human subjects.

Table B13: Demographics-based approach, controlling for MeSH codes linked to article

	All articles		Linked to trial		Not linked to trial	
	(1)	(2)	(3)	(4)	(5)	(6)
	Black or African American					
Black-sounding name	0.011***	0.010***	0.029***	0.026**	0.010***	0.008***
	(0.002)	(0.002)	(0.011)	(0.010)	(0.002)	(0.002)
Hispanic-sounding name	-0.000	0.000	0.002	0.002	-0.000	0.000
	(0.001)	(0.001)	(0.004)	(0.003)	(0.001)	(0.001)
Asian-sounding name	-0.004***	-0.001**	-0.001	-0.001	-0.004***	-0.001***
	(0.000)	(0.000)	(0.002)	(0.001)	(0.000)	(0.000)
Am. Native-sounding name	-0.005	-0.009	-0.015	-0.016	-0.004	-0.009
	(0.009)	(0.008)	(0.031)	(0.028)	(0.008)	(0.008)
MeSH code FE		X		X		X
Observations	8,668,848	8,667,886	616,278	613,189	8,052,570	8,051,509
R-squared	0.001	0.136	0.002	0.125	0.001	0.139
LHS (mean)	0.009	0.009	0.011	0.011	0.009	0.009
RHS (mean)	0.062	0.062	0.071	0.071	0.062	0.062

Notes. Robust s.e. in parenthesis clustered by last name and publication id. The unit of observation is a research article in the PubMed database published between 2002 and 2018, with first author affiliated with a US institution, and published in a journal in the top 1,000 by Commercialization Impact Factor, with associated MeSH code. Black-sounding name, Hispanic-sounding name, Asian-sounding name, and American Native-sounding name refer to race or ethnicity of the first author listed on the publication. White-sounding name is the omitted category. All columns include year FE. The vector of racial frequencies by last name comes from the 2010 U.S. Census. Columns (1) and (2) report the result of estimating Equation 2 where the dependent variable is a dummy = 1 if the article has Black or African American among its MeSH codes, = 0 otherwise. Columns (3) and (4) report the results of estimating the same equation as in column (1), but on the subsample of articles linked to a clinical trial. Columns (5) and (6) report the results of estimating the same equation as in column (1), but on the subsample of articles not linked to a clinical trial. Columns (2), (4), and (6) include MeSH FE.

Table B14: Demographics-based approach, research on other demographic groups

Panel A: Research on Black or African American Individuals

	All	Linked to trial	Not linked to trial	All	Linked to trial	Not linked to trial
	(1)	(2)	(3)	(4)	(5)	(6)
	MeSH: Black or African American (no other race/ethnicity)					
Black-sounding name	0.007***	0.021**	0.006***	0.020***	0.024***	0.019***
	(0.002)	(0.009)	(0.001)	(0.004)	(0.009)	(0.004)
Hispanic-sounding name	-0.001	0.001	-0.001	-0.001	0.000	-0.001
	(0.001)	(0.003)	(0.001)	(0.002)	(0.003)	(0.002)
Asian-sounding name	-0.002***	-0.002	-0.002***	-0.001	-0.002*	-0.001
	(0.000)	(0.001)	(0.000)	(0.001)	(0.001)	(0.001)
Article on human subjects				X	X	X
Observations	651,253	41,944	609,309	199,455	33,793	165,662
LHS (mean)	0.005	0.007	0.005	0.013	0.007	0.014

Panel B: Research on Hispanic or Latino Individuals

	All	Linked to trial	Not linked to trial	All	Linked to trial	Not linked to trial
	MeSH: Hispanic or Latino (no other race/ethnicity)					
Black-sounding name	0.001	0.004	0.001	0.004*	0.008	0.003
	(0.001)	(0.004)	(0.001)	(0.002)	(0.005)	(0.002)
Hispanic-sounding name	0.004***	0.005	0.004***	0.014***	0.008*	0.016***
	(0.001)	(0.004)	(0.001)	(0.003)	(0.004)	(0.003)
Asian-sounding name	-0.001***	-0.000	-0.001***	0.001	0.001	0.000
	(0.000)	(0.001)	(0.000)	(0.001)	(0.001)	(0.001)
Article on human subjects				X	X	X
Observations	651,253	41,944	609,309	199,455	33,793	165,662
LHS (mean)	0.003	0.006	0.003	0.008	0.006	0.009

Panel C: Research on Asian American

	All	Linked to trial	Not linked to trial	All	Linked to trial	Not linked to trial
	MeSH: Asian American (no other race/ethnicity)					
Black-sounding name	0.000	-0.002	0.001	-0.001	-0.001	-0.001
	(0.001)	(0.003)	(0.001)	(0.002)	(0.003)	(0.002)
Hispanic-sounding name	-0.000	-0.002	-0.000	0.001	-0.002	0.001
	(0.001)	(0.002)	(0.001)	(0.002)	(0.002)	(0.002)
Asian-sounding name	0.000*	0.009***	0.000	0.006***	0.010***	0.006***
	(0.000)	(0.001)	(0.000)	(0.001)	(0.002)	(0.001)
Article on human subjects				X	X	X
Observations	651,253	41,944	609,309	199,455	33,793	165,662
LHS (mean)	0.004	0.006	0.004	0.009	0.006	0.009

Notes. Robust s.e. in parenthesis clustered by last name. The unit of observation is a research article in the PubMed database published between 2002 and 2018, with first author affiliated with a US institution, and published in a journal in the top 1,000 by Commercialization Impact Factor. All columns include controls for Black-sounding name, Hispanic-sounding name, Asian-sounding name, and American Native-sounding name. Black-sounding name, Hispanic-sounding name, Asian-sounding name, and American Native-sounding name refer to the race or ethnicity of the first author listed on the publication. White-sounding name is the omitted category. All columns include year FE. The vector of racial frequencies by last name comes from the 2010 U.S. Census. In Panel A, Column (1) reports the result of estimating Equation 2 where the dependent variable is a dummy = 1 if the article if the article has "Black or African American" among its MeSH codes, = 0 otherwise. The dummy is = 0 if the article has either "Hispanic or Latino" or "Asian American" among its MeSh codes. Column (2) reports the results of estimating the same equation as in column (1), but on the subsample of articles linked to a clinical trial. Column (3) reports the results of estimating the same equation as in column (1), but on the subsample of articles not linked to a clinical trial. In columns (4) through (6), I report the results of columns (1) to (3) estimated on the subsample of articles focusing on human subjects. Panel B is identical to Panel A, but the dependent variable is a dummy = 1 if the article if the article has "Hispanic or Latino" among its MeSH codes, = 0 otherwise. The dummy is = 0 if the article has either "Black or African American" or "Asian American" among its MeSh codes. Panel C is identical to Panel A, but the dependent variable is a dummy = 1 if the article if the article has "Asian American" among its MeSH codes, = 0 otherwise. The dummy is = 0 if the article has either "Black or African American" or "Hispanic or Latino" among its MeSh codes.

Table B15: Demographics-based approach: Evidence from ongoing clinical trials

	(1) Black or African American	(2) Hispanic or Latino	(3) Asian
Black-sounding name	0.041***	0.007	0.005
	(0.015)	(0.006)	(0.004)
Hispanic-sounding name	0.003	0.024***	0.004
	(0.007)	(0.008)	(0.004)
Asian-sounding name	-0.004	-0.001	0.006***
	(0.003)	(0.002)	(0.002)
Observations	12,366	12,366	12,366
LHS (mean)	0.012	0.007	0.004
Black-sounding name (mean)	0.074	0.074	0.074
Hispanic-sounding name (mean)	0.065	0.065	0.065
Asian-sounding name (mean)	0.195	0.195	0.195

Notes. Robust s.e. in parenthesis clustered by Pr(Black | Black-sounding last name). The unit of observation is an ongoing clinical trial registered between 2001 and 2018 with a US institution. Black-sounding name, Hispanic-sounding name, Asian-sounding name, and and American Native-sounding name refer to the average race and ethnicity across the names of all investigators listed on the trial. Column (1) reports the result of estimating equation (2) where the dependent variable is a dummy = 1 if the description of the clinical trial mentions Black, = 0 otherwise. Column (2) reports the result of estimating equation (2) where the dependent variable is a dummy = 1 if the description of the clinical trial mentions Hispanics or Latinos, = otherwise. Column (3) reports the result of estimating equation (2) where the dependent variable is a dummy = 1 if the description of the clinical trial mentions Asians, = 0 otherwise. This data comes from the web portal `Clinicaltrials.gov`. All columns include trial registration year FE.

Table B16: Frequency-based approach, dictionary classification

	(1) Relative mortality, log	(2) Typically White	(3) Similar incidence	(4) Typically Black
Black-sounding name	0.181***	-0.039***	-0.014	0.052**
	(0.050)	(0.011)	(0.022)	(0.021)
Hispanic-sounding name	-0.002	0.004	-0.001	-0.003
	(0.034)	(0.009)	(0.014)	(0.013)
Asian-sounding name	-0.004	-0.001	0.016**	-0.016***
	(0.015)	(0.004)	(0.006)	(0.006)
Am. Native sounding name	-0.118	-0.073	0.058	0.015
	(0.282)	(0.054)	(0.162)	(0.146)
Observations	90,929	90,929	90,929	90,929
R-squared	0.003	0.001	0.001	0.002
LHS (mean)	0.126	0.085	0.603	0.312
RHS (mean)	0.066	0.066	0.066	0.066

Notes. Robust s.e. in parenthesis clustered by last name. The unit of observation is a research article in the PubMed database published between 2002 and 2018, with first author affiliated with a US institution, and published in a journal in the top 1,000 by Commercialization Impact Factor. Black-sounding name, Hispanic-sounding name, Asian-sounding name, and American Native-sounding name refer to race or ethnicity of the first author listed on the publication. White-sounding name is the omitted category. The vector of racial frequencies by last name comes from the 2010 U.S. Census. All columns include year FE. Column (1) reports the result of estimating equation (3) where the dependent variable is a continuous variable equal to the log of relative mortality among Black Americans compared to white Americans, calculated according to equation 1. Column (2) reports the results of estimating equation 3 where the dependent variable is a dummy = 1 if relative mortality is at least 1.5 times higher among white Americans compared to Black Americans. Column (3) reports the results of estimating equation (3) where the dependent variable is a dummy = 1 if relative mortality is less than 1.5 among Black Americans compared to white Americans, and less than 1.5 among white Americans compared to Black Americans. Column (4) reports the results of estimating equation (3) where the dependent variable is a dummy = 1 if relative mortality is at least 1.5 times higher among Black individuals compared to white individuals. Differently from Table 3 in the main text, here articles are linked to diseases through a dictionary classification on the article abstract.

Table B17: Frequency-based approach (logit model)

	(1) Typically White	(2) Similar incidence	(3) Typically Black
Black-sounding name	0.801***	0.986	1.428***
	(0.066)	(0.077)	(0.145)
Hispanic-sounding name	0.905***	1.229***	0.859***
	(0.020)	(0.027)	(0.027)
Asian-sounding name	0.975	1.002	1.031
	(0.061)	(0.059)	(0.073)
Am. Native-sounding name	1.290	0.861	0.860
	(0.639)	(0.512)	(0.602)
Total mortality, log	0.839***	2.024***	0.477***
	(0.005)	(0.017)	(0.004)
Observations	138,657	138,657	138,657
LHS (mean)	0.331	0.488	0.181
RHS (mean)	0.064	0.064	0.064

Notes. Robust s.e. in parenthesis clustered by last name. The unit of observation is a research article in the PubMed database published between 2002 and 2018, with first author affiliated with a US institution, and published in a journal in the top 1,000 by Commercialization Impact Factor. Black-sounding name, Hispanic-sounding name, and Asian-sounding name refer to race or ethnicity of the first author listed on the publication. White-sounding name is the omitted category. The vector of racial frequencies by last name comes from the 2010 U.S. Census. All columns control for the log of total mortality among white individuals and Black individuals over the period 1999 to 2015. All columns include year FE. Column (1) reports the results of estimating Equation 3 using a logit model where the dependent variable is a dummy = 1 if relative mortality is at least 1.5 times higher among white individuals compared to Black individuals. Column (2) reports the results of estimating equation (3) using a logit model where the dependent variable is a dummy = 1 if relative mortality is less than 1.5 among Black individuals compared to white individuals, and less than 1.5 among white individuals compared to Black individuals. Column (3) reports the results of estimating equation (3) using a logit model where the dependent variable is a dummy = 1 if relative mortality is at least 1.5 times higher among Black individuals compared to white individuals. The table reports exp(coefficients).

Table B18: Frequency-based approach, controlling for gender of first author

	(1) Relative mortality, log	(2) Typically White	(3) Similar Incidence	(4) Typically Black
Black-sounding name	0.136***	-0.056***	0.006	0.050***
	(0.037)	(0.020)	(0.020)	(0.014)
Hispanic-sounding name	-0.006	-0.003	0.009	-0.005
	(0.028)	(0.016)	(0.015)	(0.009)
Asian-sounding name	-0.003	-0.019***	0.039***	-0.019***
	(0.011)	(0.006)	(0.006)	(0.004)
Am. Native-sounding name	-0.269	0.098	-0.025	-0.073
	(0.191)	(0.125)	(0.123)	(0.079)
Total mortality, log	-0.076***	-0.012***	0.087***	-0.075***
	(0.004)	(0.002)	(0.002)	(0.001)
Female dummy	0.053***	-0.062***	0.055***	0.007**
	(0.009)	(0.005)	(0.005)	(0.003)
Observations	118,008	118,008	118,008	118,008
R-squared	0.011	0.004	0.040	0.063
LHS (mean)	-0.126	0.411	0.460	0.129
RHS (mean)	0.063	0.063	0.063	0.063

Notes. Robust s.e. in parenthesis clustered by last name. The unit of observation is a research article in the PubMed database published between 2002 and 2018, with first author affiliated with a US institution, and published in a journal in the top 1,000 by Commercialization Impact Factor. Black-sounding name, Hispanic-sounding name, Asian-sounding name, and and American Native-sounding name refer to race or ethnicity of the first author listed on the publication. White-sounding name is the omitted category. The vector of racial frequencies by last name comes from the 2010 U.S. Census. Female dummy is a variable equal to 1 if the gender of the first author assigned based on their last name if female, = 0 otherwise. All columns include year FE. Column (1) reports the result of estimating equation (3) where the dependent variable is a continuous variable equal to the log of relative mortality among Black Americans compared to white Americans, calculated according to equation 1. Column (2) reports the results of estimating equation 3 where the dependent variable is a dummy = 1 if relative mortality is at least 1.5 times higher among white Americans compared to Black Americans. Column (3) reports the results of estimating equation (3) where the dependent variable is a dummy = 1 if relative mortality is less than 1.5 among Black Americans compared to white Americans, and less than 1.5 among white Americans compared to Black Americans. Column (4) reports the results of estimating equation (3) where the dependent variable is a dummy = 1 if relative mortality is at least 1.5 times higher among Black individuals compared to white individuals.

Table B19: Frequency-based approach, controlling for Journal FE, State FE, first author affiliation FE

	(1) Relative mortality, log	(2) Typically White	(3) Similar Incidence	(4) Typically Black
Panel A: Journal FE				
Black-sounding name	0.083***	-0.053***	0.030**	0.023**
	(0.026)	(0.014)	(0.012)	(0.011)
Observations	138,601	138,601	138,601	138,601
LHS (mean)	-0.039	0.331	0.488	0.181
Panel B: State FE				
Black-sounding name	0.116***	-0.047***	0.002	0.045***
	(0.034)	(0.017)	(0.016)	(0.014)
Observations	138,171	138,171	138,171	138,171
LHS (mean)	-0.039	0.331	0.488	0.181
Panel C: Affiliation of first author FE				
Black-sounding name	0.095**	-0.056**	0.018	0.039**
	(0.043)	(0.023)	(0.023)	(0.017)
Observations	70,072	70,072	70,072	70,072
LHS (mean)	-0.027	0.332	0.483	0.185

Notes. Robust s.e. in parenthesis clustered by last name. The unit of observation is a research article in the PubMed database published between 2002 and 2018, with first author affiliated with a US institution, and published in a journal in the top 1,000 by Commercialization Impact Factor. All columns control for Black-sounding name, Hispanic-sounding name, Asian-sounding name, and and American Native-sounding name, which refer to race or ethnicity of the first author listed on the publication. White-sounding name is the omitted category. All columns also control for the total of total number of deaths (Black + white Americans) due to the given disease. The vector of racial frequencies by last name comes from the 2010 U.S. Census. All columns include year FE. Column (1) reports the result of estimating equation (3) where the dependent variable is a continuous variable equal to the log of relative mortality among Black Americans compared to white Americans, calculated according to equation 1. Column (2) reports the results of estimating equation 3 where the dependent variable is a dummy = 1 if relative mortality is at least 1.5 times higher among white Americans compared to Black Americans. Column (3) reports the results of estimating equation (3) where the dependent variable is a dummy = 1 if relative mortality is less than 1.5 among Black Americans compared to white Americans, and less than 1.5 among white Americans compared to Black Americans. Column (4) reports the results of estimating equation (3) where the dependent variable is a dummy = 1 if relative mortality is at least 1.5 times higher among Black individuals compared to white individuals. In Panel A, all columns include journal FE. In Panel B, all columns include State FE, where state refers to the US state of the affiliation of the first author. In Panel C, all columns include affiliation of first author FE.

Table B20: Frequency-based approach, alternative definitions of Black-sounding last name

	(1) Relative mortality, log	(2) Typically White	(3) Similar Incidence	(4) Typically Black
Panel A: Black-sounding last name of first and last author				
Black-sounding name	0.103***	-0.039**	0.001	0.038**
	(0.037)	(0.019)	(0.018)	(0.015)
Black-sounding name (Last author)	0.102***	-0.035**	-0.034**	0.069***
	(0.034)	(0.016)	(0.016)	(0.013)
Observations	100,029	100,029	100,029	100,029
R-squared	0.030	0.015	0.161	0.142
LHS (mean)	-0.046	0.331	0.490	0.178
RHS (mean)	0.064	0.064	0.064	0.064
Panel B: Average of Pr(Black | Last name) across all authors				
Black-sounding name	0.115***	-0.045**	-0.002	0.047***
	(0.035)	(0.018)	(0.017)	(0.014)
Observations	133,465	133,465	133,465	133,465
R-squared	0.032	0.015	0.163	0.144
LHS (mean)	-0.040	0.331	0.489	0.180
RHS (mean)	0.064	0.064	0.064	0.064
Panel C: Dummy for Pr(Black | Last name) ≥ 0.5				
Black sounding name (Dummy)	0.105***	-0.050***	0.021	0.029**
	(0.032)	(0.018)	(0.020)	(0.012)
Observations	138,657	138,657	138,657	138,657
R-squared	0.032	0.015	0.164	0.144
LHS (mean)	-0.039	0.331	0.488	0.181
RHS (mean)	0.009	0.009	0.009	0.009
Panel D: Restricting sample to last names present in 1930 Census				
Black-sounding name	0.093**	-0.031	-0.013	0.044***
	(0.039)	(0.020)	(0.019)	(0.016)
Observations	107,551	107,551	107,551	107,551
R-squared	0.031	0.016	0.167	0.144
LHS (mean)	-0.037	0.333	0.481	0.185
RHS (mean)	0.076	0.076	0.076	0.076

Notes. Robust s.e. in parenthesis clustered by last name. The unit of observation is a research article in the PubMed database published between 2002 and 2018, with first author affiliated with a US institution, and published in a journal in the top 1,000 by Commercialization Impact Factor. All columns control for Black-sounding name, Hispanic-sounding name, Asian-sounding name, and and American Native-sounding name, which refer to race or ethnicity of the first author listed on the publication. White-sounding name is the omitted category. All columns also control for the total of total number of deaths (Black + white Americans) due to the given disease. The vector of racial frequencies by last name comes from the 2010 U.S. Census. All columns include year FE. Column (1) reports the result of estimating equation (3) where the dependent variable is a continuous variable equal to the log of relative mortality among Black Americans compared to white Americans, calculated according to equation 1. Column (2) reports the results of estimating equation 3 where the dependent variable is a dummy = 1 if relative mortality is at least 1.5 times higher among white Americans compared to Black Americans. Column (3) reports the results of estimating equation (3) where the dependent variable is a dummy = 1 if relative mortality is less than 1.5 among Black Americans compared to white Americans, and less than 1.5 among white Americans compared to Black Americans. Column (4) reports the results of estimating equation (3) where the dependent variable is a dummy = 1 if relative mortality is at least 1.5 times higher among Black individuals compared to white individuals. In Panel A, all columns include journal FE. In Panel B, all columns include State FE, where state refers to the US state of the affiliation of the first author. In Panel C, all columns include affiliation of first author FE.

Table B21: Frequency-based approach, including all articles (not only top 1,000 journals by Journal Commercialization Impact Factor (JCIF)

	(1) Relative mortality, log	(2) Typically White	(3) Similar Incidence	(4) Typically Black
Black-sounding name	0.201***	-0.074***	-0.007	0.081***
	(0.029)	(0.012)	(0.012)	(0.012)
Hispanic-sounding name	-0.011	-0.003	0.007	-0.004
	(0.020)	(0.010)	(0.009)	(0.008)
Asian-sounding name	-0.093***	-0.004	0.062***	-0.058***
	(0.008)	(0.004)	(0.003)	(0.003)
Am. Native-sounding name	-0.131	0.014	0.085	-0.099
	(0.171)	(0.078)	(0.073)	(0.073)
Total mortality, log	-0.125***	-0.039***	0.144***	-0.105***
	(0.002)	(0.001)	(0.001)	(0.001)
Observations	403,985	403,985	403,985	403,985
R-squared	0.044	0.017	0.191	0.152
LHS (mean)	0.045	0.301	0.482	0.217
RHS (mean)	0.070	0.070	0.070	0.070

Notes. Robust s.e. in parenthesis clustered by last name. The unit of observation is a research article in the PubMed database published between 2002 and 2018, with first author affiliated with a US institution. Black-sounding name, Hispanic-sounding name, Asian-sounding name, and and American Native-sounding name refer to race or ethnicity of the first author listed on the publication. White-sounding name is the omitted category. The vector of racial frequencies by last name comes from the 2010 U.S. Census. All columns include year FE. Column (1) reports the result of estimating equation (3) where the dependent variable is a continuous variable equal to the log of relative mortality among Black Americans compared to white Americans, calculated according to equation 1. Column (2) reports the results of estimating equation 3 where the dependent variable is a dummy = 1 if relative mortality is at least 1.5 times higher among white Americans compared to Black Americans. Column (3) reports the results of estimating equation (3) where the dependent variable is a dummy = 1 if relative mortality is less than 1.5 among Black Americans compared to white Americans, and less than 1.5 among white Americans compared to Black Americans. Column (4) reports the results of estimating equation (3) where the dependent variable is a dummy = 1 if relative mortality is at least 1.5 times higher among Black individuals compared to white individuals.

Table B22: Frequency-based approach, split by below/above Journal Commercialization Impact Factor (JCIF)

	(1) Relative mortality, log	(2) Typically White	(3) Similar incidence	(4) Typically Black
Panel A: Journal Commercialization Impact Factor (JCIF) below median in the sample				
Black-sounding name	0.091**	-0.059***	0.034*	0.024
	(0.038)	(0.020)	(0.020)	(0.016)
Observations	60,958	60,958	60,958	60,958
R-squared	0.019	0.023	0.141	0.102
LHS (mean)	-0.091	0.341	0.499	0.161
RHS (mean)	0.067	0.067	0.067	0.067
Panel B: Journal Commercialization Impact Factor (JCIF) above median in the sample				
Black-sounding name	0.148***	-0.037	-0.036*	0.073***
	(0.048)	(0.023)	(0.020)	(0.019)
Observations	77,699	77,699	77,699	77,699
R-squared	0.042	0.011	0.184	0.178
LHS (mean)	0.002	0.323	0.479	0.198
RHS (mean)	0.061	0.061	0.061	0.061

Notes. Robust s.e. in parenthesis clustered by last name. The unit of observation is a research article in the PubMed database published between 2002 and 2018, with first author affiliated with a US institution, and published in a journal in the top 1,000 by Commercialization Impact Factor. The sample is restricted to those articles linked to a patent (Marx and Fuegi, 2020, 2022). Black-sounding name, Hispanic-sounding name, and Asian-sounding name refer to race or ethnicity of the first author listed on the publication. White-sounding name is the omitted category. The vector of racial frequencies by last name is taken from the 2010 U.S. Census. All columns control for the log of total mortality among white individuals and Black individuals over the period 1999 to 2015. All columns include year FE. Column (1) reports the result of estimating equation (3) where the dependent variable is a continuous variable equal to the log of relative mortality among Black individuals compared to white individuals, calculated according to equation 1. Column (2) reports the results of estimating Equation 3 where the dependent variable is a dummy = 1 if relative mortality is at least 1.5 times higher among white individuals compared to Black individuals. Column (3) reports the results of estimating equation (3) where the dependent variable is a dummy = 1 if relative mortality is less than 1.5 among Black individuals compared to white individuals, and less than 1.5 among white individuals compared to Black individuals. Column (4) reports the results of estimating equation (3) where the dependent variable is a dummy = 1 if relative mortality is at least 1.5 times higher among Black individuals compared to white individuals.

Table B23: Frequency-based approach: Articles linked to a patent

	(1) Relative mortality, log	(2) Typically White	(3) Similar incidence	(4) Typically Black
Black-sounding	0.106*	-0.043	-0.016	0.059***
	(0.059)	(0.030)	(0.027)	(0.022)
Hispanic-sounding	-0.046	0.003	0.010	-0.013
	(0.043)	(0.023)	(0.021)	(0.015)
Asian-sounding	0.004	-0.031***	0.055***	-0.024***
	(0.015)	(0.007)	(0.007)	(0.006)
Am. Native-sounding	-0.563	0.446*	-0.214	-0.233
	(0.403)	(0.261)	(0.232)	(0.170)
Total mortality, log	-0.118***	-0.040***	0.154***	-0.114***
	(0.004)	(0.002)	(0.002)	(0.002)
Observations	41,357	41,357	41,357	41,357
LHS (mean)	-0.070	0.350	0.481	0.169
RHS (mean)	0.058	0.058	0.058	0.058

Notes. Robust s.e. in parenthesis clustered by last name. The unit of observation is a research article in the PubMed database published between 2002 and 2018, with first author affiliated with a US institution, and published in a journal in the top 1,000 by Commercialization Impact Factor. The sample is restricted to those articles linked to a patent (Marx and Fuegi, 2020, 2022). Black-sounding name, Hispanic-sounding name, and Asian-sounding name refer to race or ethnicity of the first author listed on the publication. White-sounding name is the omitted category. The vector of racial frequencies by last name is taken from the 2010 U.S. Census. All columns control for the log of total mortality among white individuals and Black individuals over the period 1999 to 2015. All columns include year FE. Column (1) reports the result of estimating equation (3) where the dependent variable is a continuous variable equal to the log of relative mortality among Black individuals compared to white individuals, calculated according to equation 1. Column (2) reports the results of estimating Equation 3 where the dependent variable is a dummy = 1 if relative mortality is at least 1.5 times higher among white individuals compared to Black individuals. Column (3) reports the results of estimating equation (3) where the dependent variable is a dummy = 1 if relative mortality is less than 1.5 among Black individuals compared to white individuals, and less than 1.5 among white individuals compared to Black individuals. Column (4) reports the results of estimating equation (3) where the dependent variable is a dummy = 1 if relative mortality is at least 1.5 times higher among Black individuals compared to white individuals.

Table B24: Frequency-based approach: Threshold set at 1.3 higher mortality

	(1) Relative mortality, log	(2) Typically White	(3) Similar incidence	(4) Typically Black
Black-sounding name	0.122***	-0.047**	0.010	0.036**
	(0.034)	(0.019)	(0.017)	(0.017)
Hispanic-sounding name	0.006	-0.016	0.005	0.011
	(0.024)	(0.014)	(0.012)	(0.012)
Asian-sounding name	-0.006	-0.009*	0.040***	-0.031***
	(0.009)	(0.005)	(0.005)	(0.004)
Am. Native-sounding name	-0.227	0.082	-0.051	-0.031
	(0.177)	(0.118)	(0.129)	(0.118)
Total mortality, log	-0.108***	0.023***	0.059***	-0.082***
	(0.002)	(0.001)	(0.001)	(0.001)
Observations	138,657	138,657	138,657	138,657
R-squared	0.032	0.005	0.035	0.076
LHS (mean)	-0.039	0.416	0.321	0.263
RHS (mean)	0.064	0.064	0.064	0.064

Notes. Robust s.e. in parenthesis clustered by last name. The unit of observation is a research article in the PubMed database published between 2002 and 2018, with first author affiliated with a US institution, and published in a journal in the top 1,000 by Commercialization Impact Factor. Black-sounding name, Hispanic-sounding name, and Asian-sounding name refer to race or ethnicity of the first author listed on the publication. White-sounding name is the omitted category. The vector of racial frequencies by last name comes from the 2010 U.S. Census. All columns control for the log of total mortality among white individuals and Black individuals over the period 1999 to 2015. All columns include year FE. Column (1) reports the result of estimating equation (3) where the dependent variable is a continuous variable equal to the log of relative mortality among Black individuals compared to white individuals, calculated according to equation 1. Column (2) reports the results of estimating Equation 3 where the dependent variable is a dummy = 1 if relative mortality is at least 1.3 times higher among white individuals compared to Black individuals. Column (3) reports the results of estimating equation (3) where the dependent variable is a dummy = 1 if relative mortality is less than 1.3 among Black individuals compared to white individuals, and less than 1.3 among white individuals compared to Black individuals. Column (4) reports the results of estimating equation (3) where the dependent variable is a dummy = 1 if relative mortality is at least 1.3 times higher among Black individuals compared to white individuals.

Table B25: Frequency-based approach: Threshold set at 1.7 higher mortality

	(1) Relative mortality, log	(2) Typically White	(3) Similar incidence	(4) Typically Black
Black-sounding name	0.122***	-0.032**	-0.015	0.047***
	(0.034)	(0.016)	(0.017)	(0.013)
Hispanic-sounding name	0.006	-0.003	-0.003	0.006
	(0.024)	(0.013)	(0.013)	(0.009)
Asian-sounding name	-0.006	-0.011**	0.023***	-0.013***
	(0.009)	(0.004)	(0.005)	(0.004)
Am. Native-sounding name	-0.227	0.126	-0.104	-0.022
	(0.177)	(0.113)	(0.124)	(0.085)
Total mortality, log	-0.108***	-0.067***	0.147***	-0.080***
	(0.002)	(0.001)	(0.001)	(0.001)
Observations	138,657	138,657	138,657	138,657
R-squared	0.032	0.052	0.189	0.105
LHS (mean)	-0.039	0.239	0.607	0.153
RHS (mean)	0.064	0.064	0.064	0.064

Notes. Robust s.e. in parenthesis clustered by last name. The unit of observation is a research article in the PubMed database published between 2002 and 2018, with first author affiliated with a US institution, and published in a journal in the top 1,000 by Commercialization Impact Factor. Black-sounding name, Hispanic-sounding name, and Asian-sounding name refer to race or ethnicity of the first author listed on the publication. White-sounding name is the omitted category. The vector of racial frequencies by last name comes from the 2010 U.S. Census. All columns control for the log of total mortality among white individuals and Black individuals over the period 1999 to 2015. All columns include year FE. Column (1) reports the result of estimating equation (3) where the dependent variable is a continuous variable equal to the log of relative mortality among Black individuals compared to white individuals, calculated according to equation 1. Column (2) reports the results of estimating Equation 3 where the dependent variable is a dummy = 1 if relative mortality is at least 1.7 times higher among white individuals compared to Black individuals. Column (3) reports the results of estimating equation (3) where the dependent variable is a dummy = 1 if relative mortality is less than 1.7 among Black individuals compared to white individuals, and less than 1.7 among white individuals compared to Black individuals. Column (4) reports the results of estimating equation (3) where the dependent variable is a dummy = 1 if relative mortality is at least 1.7 times higher among Black individuals compared to white individuals.

Table B26: Frequency-based approach: Alternative definitions of frequency

	(1)	(2)	(3)	(4)	(5)	(6)
	Relative mortality (log)			Mortality, Black population (log)		
Black-sounding name	0.122***	0.116***	0.083***	0.106***	0.100***	0.077***
	(0.034)	(0.034)	(0.026)	(0.033)	(0.033)	(0.026)
Hispanic-sounding name	0.005	0.017	-0.004	-0.002	0.009	-0.006
	(0.025)	(0.024)	(0.019)	(0.024)	(0.024)	(0.018)
Asian-sounding name	-0.006	-0.004	0.000	-0.002	-0.000	0.003
	(0.009)	(0.010)	(0.008)	(0.009)	(0.009)	(0.008)
Am. Native-sounding name	-0.229	-0.135	-0.098	-0.195	-0.109	-0.063
	(0.178)	(0.181)	(0.154)	(0.174)	(0.178)	(0.153)
Mortality, (log)	-0.108***	-0.106***	-0.057***			
	(0.002)	(0.002)	(0.002)			
Mortality, White population (log)				0.818***	0.821***	0.861***
				(0.002)	(0.002)	(0.002)
Year FE	X	X	X	X	X	X
State FE		X			X	
Journal FE			X			X
Observations	138,657	138,171	138,601	138,657	138,171	138,601
R-squared	0.032	0.044	0.293	0.675	0.679	0.755
LHS (mean)	-0.039	-0.039	-0.039	10.961	10.960	10.961
RHS (mean)	0.064	0.064	0.064	0.064	0.064	0.064

Notes. Robust s.e. in parenthesis clustered by last name. The unit of observation is a research article in the PubMed database published between 2002 and 2018, with first author affiliated with a US institution, and published in a journal in the top 1,000 by Commercialization Impact Factor. Black-sounding name, Hispanic-sounding name, and Asian-sounding name refer to race or ethnicity of the first author listed on the publication. White-sounding name is the omitted category. The vector of racial frequencies by last name comes from the 2010 U.S. Census. Columns (1), (2), and (3) control for the log of total mortality among white individuals + Black individuals over the period 1999 to 2015. All columns include year FE. Columns (1), (2), and (3) report the results of estimating equation (3) where the dependent variable is a continuous variable equal to the log of relative mortality among Black individuals compared to white individuals, calculated according to equation 1. Column (1) is the baseline. In Column (2), I add FE for the state where the institution of the first author is affiliated is located. In column (3), I add journal FE. Columns (1), (2), and (3) control for the log of total mortality among white individuals + Black individuals over the period 1999 to 2015. Columns (4), (5), and (6) report the results of estimating equation (3) where the dependent variable is a continuous variable equal to the log total mortality among Black individuals in 1999 to 2015. Column (4) is the baseline. In Column (5), I add FE for the state where the institution of the first author is affiliated is located. In column (6), I add journal FE. Columns (4), (5), and (6) control for the log of total mortality among white individuals over the period 1999 to 2015.

Table B27: Demographics-based Approach, by Gender

	(1)	(2)	(3)	(4)	(5)	(6)
	Article mentions Black or African Am.			Article mentions female		
Black-sounding	0.008***	0.022***	0.008***	0.002	0.007	0.001
	(0.002)	(0.009)	(0.001)	(0.002)	(0.007)	(0.002)
Female	0.013***	0.003**	0.014***	0.030***	0.018***	0.033***
	(0.000)	(0.001)	(0.000)	(0.000)	(0.001)	(0.001)
Asian-sounding	-0.001**	0.001	-0.001***	-0.002***	0.003	-0.002***
	(0.000)	(0.002)	(0.000)	(0.000)	(0.002)	(0.000)
Hispanic-sounding	0.000	-0.003	0.000	-0.000	0.002	-0.001
	(0.001)	(0.003)	(0.001)	(0.001)	(0.005)	(0.001)
Observations	651,253	41,944	609,309	651,253	41,944	609,309
R-squared	0.006	0.001	0.007	0.018	0.003	0.020
LHS (mean)	0.007	0.010	0.006	0.012	0.019	0.012
RHS (mean)	0.063	0.072	0.063	0.325	0.845	0.295

Notes. Robust s.e. in parenthesis clustered by last name. The unit of observation is a research article in the PubMed database published between 2002 and 2018, with first author affiliated with a US institution, and published in a journal in the top 1,000 by Commercialization Impact Factor. Black-sounding name refers to the race or ethnicity of the first author listed on the publication. All columns include controls for , Hispanic-sounding name, Asian-sounding name, and American Native-sounding name, and controls for these variables interacted with Sickle cell anemia (Column (1) and (2)), with Thalassemia (Column (2) and (3)), and with Hemophilia (Column (4) and (5)). Throughout, white-sounding name is the omitted category. All columns include year FE. In all columns, the dependent variable (Relative Citation Ratio) is calculated as the citations of a paper normalized to the citations received publications in the same area of research and year. Sickle cell anemia is a dummy = 1 if the article has sickle cell anemia among its MeSH codes, = 0 otherwise. Column (2) reports the results of estimating the same equation as in column (1), but on the subsample of hemoglobin-related diseases. "Hemoglobin-related" diseases are defined as those articles with hemoglobin among their MeSH codes. Thalassemia is a dummy = 1 if the article has thalassemia among its MeSH codes, = 0 otherwise. Column (4) reports the results of estimating the same equation as in column (3), but on the subsample of hemoglobin-related diseases. Hemophilia is a dummy = 1 if the article has hemophilia among its MeSH codes, = 0 otherwise. Column (6) reports the results of estimating the same equation as in column (5), but on the subsample of hemoglobin-related diseases. All columns report coefficients from the estimation of a Poisson count model.

Table B28: Frequency-based approach, Frequency by Gender

	(1) Relative mortality, log	(2) Typically male	(3) Similar	(4) Typically female
Female	0.020***	-0.033***	0.013***	0.020***
	(0.003)	(0.003)	(0.003)	(0.003)
Total deaths, log	0.087***	-0.143***	0.153***	-0.009***
	(0.001)	(0.001)	(0.001)	(0.001)
Observations	112,333	112,333	112,333	112,333
LHS (mean)	-0.185	0.356	0.515	0.130
RHS (mean)	0.485	0.485	0.485	0.485

Notes. Robust s.e. in parenthesis clustered by last name. The unit of observation is a research article in the PubMed database published between 2002 and 2018, with first author affiliated with a US institution, and published in a journal in the top 1,000 by Commercialization Impact Factor. Black-sounding name refers to the race or ethnicity of the first author listed on the publication. All columns include controls for , Hispanic-sounding name, Asian-sounding name, and American Native-sounding name, and controls for these variables interacted with Sickle cell anemia (Column (1) and (2)), with Thalassemia (Column (2) and (3)), and with Hemophilia (Column (4) and (5)). Throughout, white-sounding name is the omitted category. All columns include year FE. In all columns, the dependent variable (Relative Citation Ratio) is calculated as the citations of a paper normalized to the citations received publications in the same area of research and year. Sickle cell anemia is a dummy = 1 if the article has sickle cell anemia among its MeSH codes, = 0 otherwise. Column (2) reports the results of estimating the same equation as in column (1), but on the subsample of hemoglobin-related diseases. "Hemoglobin-related" diseases are defined as those articles with hemoglobin among their MeSH codes. Thalassemia is a dummy = 1 if the article has thalassemia among its MeSH codes, = 0 otherwise. Column (4) reports the results of estimating the same equation as in column (3), but on the subsample of hemoglobin-related diseases. Hemophilia is a dummy = 1 if the article has hemophilia among its MeSH codes, = 0 otherwise. Column (6) reports the results of estimating the same equation as in column (5), but on the subsample of hemoglobin-related diseases. All columns report coefficients from the estimation of a Poisson count model.

Table B29: Demographics-based approach, Total Citations

	All	Linked to trial	Not linked to trial	All	Linked to trial	Not linked to trial
	(1)	(2)	(3)	(4)	(5)	(6)
				Total Citations		
Black-sounding name	-0.275***	-0.130	-0.277***	-0.199**	-0.123	-0.188**
	(0.040)	(0.136)	(0.042)	(0.079)	(0.156)	(0.091)
MeSH: Black or African American	0.068	-0.363***	0.114**	0.112**	-0.323***	0.230***
	(0.046)	(0.104)	(0.050)	(0.054)	(0.113)	(0.058)
MeSH: Black or African Am. X Black-s. name	0.281	0.069	0.301	0.281	0.145	0.290
	(0.254)	(0.531)	(0.278)	(0.281)	(0.557)	(0.310)
Article on human subjects				X	X	X
Observations	651,253	41,944	609,309	199,455	33,793	165,662
LHS (mean)	71.340	98.402	69.477	66.904	98.406	60.477
RHS (mean)	0.063	0.072	0.063	0.072	0.072	0.071

Notes. Robust s.e. in parenthesis clustered by last name. The unit of observation is a research article in the PubMed database published between 2002 and 2018, with first author affiliated with a US institution, and published in a journal in the top 1,000 by Commercialization Impact Factor. Black-sounding name refers to the race or ethnicity of the first author listed on the publication. White-sounding name is the omitted category. All columns include year FE, and controls for Hispanic-sounding, Asian-sounding, and American Native-sounding last name of the first author, and controls for these variables interacted with "MeSH: Black or African American". The vector of racial frequencies by last name comes from the 2010 U.S. Census. In all columns, the dependent variable is the total number of citations. "MeSH: Black or African American" is a dummy = 1 if the article has Black or African American among its MeSH codes, = 0 otherwise. Column (2) reports the results of estimating the same equation as in Column (1), but on the subsample of articles linked to a clinical trial. Column (2) reports the results of estimating the same equation as in column (1), but on the subsample of articles not linked to a clinical trial. Columns (4) to (6) are symmetric to columns (1) to (3), but the model is estimated on the subsample of articles focusing on human subjects. All columns report coefficients from the estimation of a Poisson count model.

Table B30: Demographics-based approach, Citations (Relative Citation Ratio)

	All	Linked to trial	Not linked to trial	All	Linked to trial	Not linked to trial
	(1)	(2)	(3)	(4)	(5)	(6)
	Relative Citation Ratio					
Black-sounding name	-0.245***	-0.130	-0.240***	-0.167**	-0.121	-0.149*
	(0.040)	(0.126)	(0.042)	(0.073)	(0.144)	(0.086)
MeSH: Black or African American	0.074*	-0.342***	0.125***	-0.006	-0.302**	0.089*
	(0.042)	(0.123)	(0.045)	(0.048)	(0.136)	(0.051)
MeSH: Black or A.A. × Black-s. name	0.250	-0.085	0.275	0.217	-0.076	0.235
	(0.246)	(0.513)	(0.268)	(0.274)	(0.549)	(0.301)
Article on human subjects				X	X	X
Observations	651,247	41,944	609,303	199,453	33,793	165,660
LHS (mean)	2.100	3.339	2.015	2.380	3.397	2.173
RHS (mean)	0.063	0.072	0.063	0.072	0.072	0.071

Notes. Robust s.e. in parenthesis clustered by last name. The unit of observation is a research article in the PubMed database published between 2002 and 2018, with first author affiliated with a US institution, and published in a journal in the top 1,000 by Commercialization Impact Factor. Black-sounding name refers to the race or ethnicity of the first author listed on the publication. White-sounding name is the omitted category. All columns include year FE, and controls for Hispanic-sounding, Asian-sounding, and American Native-sounding last name of the first author, and controls for these variables interacted with "MeSH: Black or African American". The vector of racial frequencies by last name comes from the 2010 U.S. Census. In all columns, the dependent variable (Relative Citation Ratio) is calculated as the citations of a paper normalized to the citations received by publications in the same area of research and year (Hutchins, Yuan, Anderson and Santangelo, 2016). "MeSH: Black or African American" is a dummy = 1 if the article has Black or African American among its MeSH codes, = 0 otherwise. Column (2) reports the results of estimating the same equation as in Column (1), but on the subsample of articles linked to a clinical trial. Column (2) reports the results of estimating the same equation as in column (1), but on the subsample of articles not linked to a clinical trial. Columns (4) to (6) are symmetric to columns (1) to (3), but the model is estimated on the subsample of articles focusing on human subjects. All columns report coefficients from the estimation of a Poisson count model.

Table B31: Frequency-based approach, Total Citations

	(1)	(2)	(3)	(4)
	Total Citations			
Black-sounding name	-0.114	-0.070	-0.148*	-0.161*
	(0.071)	(0.087)	(0.089)	(0.082)
Relative mortality (log)	0.424***			
	(0.107)			
Relative mortality (log) X Black-sounding name	0.101			
	(0.072)			
Typically White		0.664***		
		(0.129)		
Typically White X Black-sounding name		-0.177		
		(0.141)		
Similar incidence			-0.625***	
			(0.135)	
Similar incidence X Black-sounding name			0.038	
			(0.139)	
Typically Black				0.825***
				(0.158)
Typically Black X Black-sounding name				0.231
				(0.160)
Observations	138,657	138,657	138,657	138,657
LHS (mean)	80.731	80.731	80.731	80.731
RHS (mean)	0.064	0.064	0.064	0.064

Notes. Robust s.e. in parenthesis clustered by last name. The unit of observation is a research article in the PubMed database published between 2002 and 2018, with first author affiliated with a US institution, and published in a journal in the top 1,000 by Commercialization Impact Factor. Black-sounding name refers to race or ethnicity of the first author listed on the publication. All columns include controls for Hispanic-sounding, Asian-sounding, and American Indian-sounding, and for those variables interacted with Relative mortality (log) (Column (1)), with Typically white (Column (2)), with Similar incidence (Column (4)), with Typically Black individuals (Column (4)). In all columns, white-sounding is the omitted category. All columns control for the log of total mortality (white + Black Americans) over the period 1999 to 2015. All columns include year FE. In all columns, the dependent variable it the total number of citations. In Column (1), relative mortality (log) is the log of rate among Black individuals compared to white individuals. In Column (2), Typically white is a dummy = 1 if relative mortality is at least 1.5 times higher among white individuals compared to Black individuals. In Column (3), similar incidence is a dummy = 1 if relative mortality is less than 1.5 among Black individuals compared to white individuals, and less than 1.5 among white individuals compared to Black individuals. In Column (4), Typically Black is a dummy = 1 if relative mortality is at least 1.5 times higher among Black individuals compared to white individuals. All columns report coefficients from the estimation of a Poisson count model.

Table B32: Frequency-based approach, Citations (Relative Citation Ratio)

	(1)	(2)	(3)	(4)
	Relative citations ratio (RCR)			
Black-sounding name	-0.111	-0.064	-0.143	-0.157*
	(0.072)	(0.088)	(0.090)	(0.082)
Relative mortality (log)	0.366***			
	(0.109)			
Relative mortality (log) × Black-sounding name	0.096			
	(0.069)			
Typically White		0.635***		
		(0.131)		
Typically White × Black-sounding name		-0.181		
		(0.137)		
Similar incidence			-0.624***	
			(0.141)	
Similar incidence × Black-sounding name			0.038	
			(0.139)	
Typically Black				0.579***
				(0.159)
Typically Black × Black-sounding name				0.220
				(0.168)
Observations	138,656	138,656	138,656	138,656
LHS (mean)	2.366	2.366	2.366	2.366
RHS (mean)	0.064	0.064	0.064	0.064

Notes. Robust s.e. in parenthesis clustered by last name. The unit of observation is a research article in the PubMed database published between 2002 and 2018, with first author affiliated with a US institution, and published in a journal in the top 1,000 by Commercialization Impact Factor. Black-sounding name refers to race or ethnicity of the first author listed on the publication. All columns include controls for Hispanic-sounding, Asian-sounding, and American Indian-sounding, and for those variables interacted with Relative mortality (log) (Column (1)), with Typically white (Column (2)), with Similar incidence (Column (4)), with Typically Black individuals (Column (4)). In all columns, white-sounding is the omitted category. All columns control for the log of total mortality among white individuals and Black individuals over the period 1999 to 2015 interacted with Black-sounding name. All columns include year FE. In all columns, the dependent variable (Relative Citation Ratio) is calculated as the citations of a paper normalized to the citations received by publications in the same area of research and year (Hutchins, Yuan, Anderson and Santangelo, 2016). In Column (1), relative mortality (log) is the log of rate among Black individuals compared to white individuals. In Column (2), Typically white is a dummy = 1 if relative mortality is at least 1.5 times higher among white individuals compared to Black individuals. In Column (3), similar incidence is a dummy = 1 if relative mortality is less than 1.5 among Black individuals compared to white individuals, and less than 1.5 among white individuals compared to Black individuals. In Column (4), Typically Black is a dummy = 1 if relative mortality is at least 1.5 times higher among Black individuals compared to white individuals. All columns report coefficients from the estimation of a Poisson count model.

Table B33: Demographics-based approach, split by below/above median Journal Impact Factor (JIF)

	All	Linked to trial	Not Linked to trial	All	Linked to trial	Not Linked to trial
	(1)	(2)	(3)	(4)	(5)	(6)
	MeSH: Black or African American					
Panel A: Journal Commercialization Impact Factor (JCIF) below median in the sample						
Black-sounding name	0.015***	0.037***	0.012***	0.026***	0.040***	0.023***
	(0.003)	(0.012)	(0.003)	(0.005)	(0.011)	(0.005)
Observations	253,951	22,606	231,345	122,497	18,983	103,514
R-squared	0.001	0.003	0.001	0.001	0.003	0.001
LHS (mean)	0.010	0.010	0.010	0.018	0.010	0.019
RHS (mean)	0.067	0.072	0.067	0.072	0.072	0.072
Panel B: Journal Commercialization Impact Factor (JCIF) above median in the sample						
Black-sounding name	0.007***	0.011	0.007***	0.033***	0.015	0.037***
	(0.002)	(0.008)	(0.002)	(0.007)	(0.010)	(0.008)
Observations	397,302	19,338	377,964	76,958	14,810	62,148
R-squared	0.001	0.002	0.001	0.001	0.002	0.001
LHS (mean)	0.006	0.009	0.005	0.022	0.010	0.025
RHS (mean)	0.061	0.071	0.060	0.070	0.072	0.070

Notes. Robust s.e. in parenthesis clustered by last name. The unit of observation is a research article in the PubMed database published between 2002 and 2018, with first author affiliated with a US institution, and published in a journal in the top 1,000 by Commercialization Impact Factor. Black-sounding name, Hispanic-sounding name, Asian-sounding name, and American Native-sounding name refer to the race or ethnicity of the first author listed on the publication. White-sounding name is the omitted category. All columns include year FE. The vector of racial frequencies by last name comes from the 2010 U.S. Census. Column (1) reports the result of estimating Equation 2 where the dependent variable is a dummy = 1 if the article has "Black or African American" among its MeSH codes, = 0 otherwise. Column (2) reports the results of estimating the same equation as in column (1), but on the subsample of articles linked to a clinical trial. Column (3) reports the results of estimating the same equation as in column (1), but on the subsample of articles not linked to a clinical trial. In columns (4) through (6), I report the results of columns (1) to (3) estimated on the subsample of articles focusing on human subjects.

Table B34: Frequency-based approach, split by below/above median Journal Impact Factor (JIF)

	(1) Relative mortality, log	(2) Typically White	(3) Similar incidence	(4) Typically Black
Panel A: Journal Impact Factor (JIF) below median in the sample				
Black-sounding name	0.122***	-0.088***	0.048**	0.040**
	(0.040)	(0.023)	(0.023)	(0.017)
Observations	49,872	49,872	49,872	49,872
R-squared	0.023	0.020	0.145	0.114
LHS (mean)	-0.098	0.363	0.476	0.161
RHS (mean)	0.066	0.066	0.066	0.066
Panel B: Journal Impact Factor (JIF) above median in the sample				
Black-sounding name	0.118***	-0.022	-0.027	0.049***
	(0.045)	(0.022)	(0.020)	(0.018)
Observations	84,170	84,170	84,170	84,170
R-squared	0.040	0.014	0.190	0.178
LHS (mean)	-0.003	0.313	0.496	0.191
RHS (mean)	0.063	0.063	0.063	0.063

Notes. Robust s.e. in parenthesis clustered by last name. The unit of observation is a research article in the PubMed database published between 2002 and 2018, with first author affiliated with a US institution, and published in a journal in the top 1,000 by Commercialization Impact Factor. Black-sounding name refers to race or ethnicity of the first author listed on the publication. All columns include controls for Hispanic-sounding, Asian-sounding, and American Indian-sounding, and for those variables interacted with Relative mortality (log) (Column (1)), with Typically white (Column (2)), with Similar incidence (Column (4)), with Typically Black individuals (Column (4)). In all columns, white-sounding is the omitted category. All columns control for the log of total mortality among white individuals and Black individuals over the period 1999 to 2015 interacted with Black-sounding name. All columns include year FE. In all columns, the dependent variable (Relative Citation Ratio) is calculated as the citations of a paper normalized to the citations received by publications in the same area of research and year (Hutchins, Yuan, Anderson and Santangelo, 2016). In Column (1), relative mortality (log) is the log of rate among Black individuals compared to white individuals. In Column (2), Typically white is a dummy = 1 if relative mortality is at least 1.5 times higher among white individuals compared to Black individuals. In Column (3), similar incidence is a dummy = 1 if relative mortality is less than 1.5 among Black individuals compared to white individuals, and less than 1.5 among white individuals compared to Black individuals. In Column (4), Typically Black is a dummy = 1 if relative mortality is at least 1.5 times higher among Black individuals compared to white individuals. All columns report coefficients from the estimation of a Poisson count model.

Table B35: Sickle cell anemia, by MeSH code

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
	Anatomy		Organisms		Diseases		Chemicals & Drugs		Diagnostic equip.		Biological sciences		Healthcare	
Black-s.	0.000	0.030	0.002*	0.092***	0.003*	0.091***	0.001	0.085**	0.001	0.078**	0.001	0.095**	0.002*	0.119***
	(0.001)	(0.037)	(0.001)	(0.034)	(0.001)	(0.034)	(0.001)	(0.040)	(0.001)	(0.037)	(0.001)	(0.041)	(0.001)	(0.046)
Hispanic-s.	-0.000	-0.023	-0.000	-0.014	-0.001	-0.015	-0.000	-0.004	-0.000	-0.009	-0.000	-0.014	-0.001	-0.029
	(0.001)	(0.022)	(0.000)	(0.017)	(0.001)	(0.017)	(0.001)	(0.020)	(0.000)	(0.016)	(0.000)	(0.018)	(0.001)	(0.024)
Asian-s.	-0.000**	-0.011	-0.000**	-0.004	-0.000	-0.004	-0.001***	-0.009	-0.000**	-0.005	-0.000***	-0.010	0.000	0.012
	(0.000)	(0.009)	(0.000)	(0.008)	(0.000)	(0.008)	(0.000)	(0.008)	(0.000)	(0.008)	(0.000)	(0.008)	(0.000)	(0.013)
Am. Native-s.	-0.007***	-0.316*	-0.006***	-0.231**	-0.011***	-0.232**	-0.008***	-0.394**	-0.007***	-0.242**	-0.008***	-0.525**	-0.006**	-0.170**
	(0.002)	(0.183)	(0.002)	(0.099)	(0.003)	(0.100)	(0.002)	(0.162)	(0.002)	(0.097)	(0.002)	(0.262)	(0.003)	(0.072)
Observations	353,238	6,875	599,099	13,268	316,191	13,297	486,374	10,257	435,345	9,536	489,509	9,572	243,224	6,265
R-squared	0.000	0.006	0.000	0.004	0.000	0.004	0.000	0.006	0.000	0.006	0.000	0.005	0.000	0.007
LHS (mean)	0.001	0.058	0.001	0.062	0.003	0.062	0.001	0.061	0.001	0.057	0.001	0.060	0.002	0.075
RHS (mean)	0.060	0.059	0.064	0.063	0.066	0.063	0.061	0.061	0.063	0.063	0.061	0.061	0.069	0.067

Notes. Robust s.e. in parenthesis clustered by last name. The unit of observation is a research article in the PubMed database published between 2002 and 2018, with first author affiliated with a US institution. Black-sounding name, Hispanic-sounding name, Asian-sounding name, and American Native-sounding name refer to race or ethnicity of the first author listed on the publication. White-sounding name is the omitted category. All columns include year FE. The vector of racial frequencies by last name comes from the 2010 U.S. Census. Columns (1), (3), (5), (7), (9), (11), (13) report the result of estimating Equation 2 where the dependent variable is a dummy = 1 if the article has sickle cell anemia among its MeSH codes, = 0 otherwise. Columns (2), (4), (6), (8), (10), (12), and (14) report the results of estimating the same equation as in column (1), but on the subsample of hemoglobin-related diseases. "Hemoglobin-related" diseases are defined as those articles with neoplasm among their MeSH codes.

C Model Appendix

In section 6 of the main text, I estimate the model of occupational choice built by Hsieh, Hurst, Jones and Klenow (2019). In this model, each individual selects the occupation where they obtain the highest utility given their talents and preferences. In this model of the labor market, there are three forces which cause individuals to choose occupations where they do not have a comparative advantage: i) discrimination in the labor market; ii) barriers to acquiring human capital; iii) group-specific preferences or social norms. I use the data to calibrate the magnitude of these three forces for the period 2001 to 2018. Then I compute a policy counterfactual where barriers to human capital acquisition and in the labor market are lifted for Black individuals.

C.1 Workers

The economy is composed by a continuum of individuals j who are either white or Black. Their group is indexed by $g = \{\text{white, Black}\}$. Each individual chooses an occupation j to maximize their lifetime utility. Individuals choose their lifetime occupation and decide how much time to dedicate to schooling before entering the labor market (the pre-period), and live for three periods. Their lifetime utility is equal to:

$$\log U_i = \alpha \sum_{t=1}^{3} \log \mathrm{c}_{it} + \log\left(1 - s_{ij}\right) + \log z_{jg} + \log \mu_{ij} \tag{12}$$

α represents the tradeoff between utility in the pre-period and utility over the remaining of the lifetime. s is time spent in school to acquire human capital (so that $(1-s)$ is leisure), z is group-specific utility derived from working in occupation j, and μ is individual utility from working in occupation j. The parameter z_{jg} relaxes the assumption that, in the absence of barriers, all groups will select occupation j at the same rate. It can be interpreted as preferences, beliefs, or experience.[42] Individual consumption c_{it} is equal to:

$$c_{it} = (1 - \tau_{jg}^{w}) w_{jt} \varepsilon_{ij} h_{ij} - (1 + \tau_{j,g}^{h}) e_{ij} \tag{13}$$

Where τ_{jg} is the job- and group- specific tax to work in occupation j, w_{jt} is the efficiency wage, ε_{ij} is individual productivity of working in occupation j, (i.e., their "talent" for occupation j), h_{ij} is human capital acquired to work in occupation j. Over the lifetime, individuals

[42] For example, Hsieh, Hurst, Jones and Klenow (2019) documents a decrease of z in the US between 1960 and 2010 a decrease women in the home sector, which they interpret as changes in social norms for women working in the market sector or changing preferences for fertility.

repay the loan they got in the first period to acquire education. They repay it equally across all three periods. In every period they have to repay $1/3\ (1+\tau_{jg}^{h})e_{ij}$, where τ_{jg}^{h} is the tax on acquisition of human capital.

Occupation-specific human capital acquired in the pre-period equals:

$$h_{ij} = \bar{h}_{jg}\gamma s_{ij}^{\phi_j} e_{ij}^{\eta} \tag{14}$$

Where $\bar{h}_{jg}$ are differences in human capital endowment that are specific to a group and a given occupation, γ is the return to experience, ϕ_j is the occupation-specific return to time investments in human capital, and η is the elasticity of human capital with respect to human capital expenditures.

Individuals draw a vector of idiosyncratic talent ϵ_j or preferences μ_j across occupations. When they draw idiosyncratic talent, then preference μ_j is assumed to be the same in all occupations, and $= 1$ in each occupation j.

Talent in occupation j, ϵ_j, is assumed to be distributed according to the multivariate Fréchet distribution:

$$F_g\left(\epsilon_{i1}, \ldots, \epsilon_{iJ}\right) = \exp\left[-\sum_{j=1}^{J} \epsilon_{ij}^{-\theta}\right] \tag{15}$$

Where J is the total number of occupations j, and θ is the shape parameter that governs the dispersion of talent across occupations, with higher θ corresponding to smaller dispersion. The mean of the distribution is normalized to one in each occupation and each group (white individuals and Black individuals).

When individuals draw a vector of idiosyncratic preferences, μ_j is assumed to be distributed according to same Fréchet distribution, but with shape parameter $\frac{\theta(1-\eta)}{3\beta}$. In this way, the labor supply elasticity of a given occupation of individuals with heterogeneous preferences is equal to the one with heterogeneous talent. In this case, talent ϵ_i is assumed to be the same in all occupations, and equals $\Gamma^{1-\eta}$, where $\Gamma \equiv \Gamma\left(1-\frac{1}{\theta(1-\eta)}\right)$. In this way, the average talent is the same in the case of heterogeneous preferences, and in the case of heterogeneous talent.

C.2 Occupational choice

Solving the worker's problem for a given young cohort c at time t, indirect utility of an individual in occupation j and group g (omitting from now on the individual subscript i) is

equal to:

$$U^*_{j,g} = \mu_j \left[\bar{\gamma}\tilde{w}_{j,g}\epsilon_j\right]^{\frac{3\beta}{1-\eta}} \tag{16}$$

where:

$$\tilde{w}_{j,g} \equiv w_j s_j^{\phi_j} \left(1 - s_j\right)^{\frac{1-\eta}{3\beta}} \cdot \frac{\bar{h}_{j,g}\tilde{z}_{j,g}}{\tau_{j,g}} \tag{17}$$

$$\tau_{j,g} \equiv \frac{\left(1 + \tau^h_{j,g}\right)^{\eta}}{1 - \tau^w_{j,g}} \tag{18}$$

$$\tilde{z}_{j,g} \equiv z_{j,g}\frac{1-\eta}{3\beta} \tag{19}$$

Utility of working in occupation j increases with the occupation-specific preference (μ_j) and occupation-specific talent (ϵ_j). A higher value of $\tau_{j,g}$ is associated with lower individual utility. A higher value of $\tilde{z}_{j,g}$, which represents the utility of group g from working in occupation j is also associated with higher utility.

Each individual chooses the occupation with the highest U^*. Because the heterogeneity is drawn from an extreme value distribution, the highest utility can also be characterized by an extreme value distribution (McFadden, 1974). The overall occupation share is obtained by aggregating the optimal choice across people. The occupation choice problem is equivalent to selecting occupation j with the highest value of $U^*_{j,g}$.

C.3 Workers' equilibrium

The equilibrium is described by the following five propositions.

Proposition 1: Occupational choice

The fraction of individuals from group g (Black, or white) who select into occupation j (i.e., $p_{j,g}$), is equal to:

$$p_{j,g} = \frac{\tilde{w}^{\theta}_{j,g}}{\sum_{s=1}^{J} \tilde{w}^{\theta}_{s,g}} \tag{20}$$

where

$$\widetilde{w}_{j,g} \equiv w_j s_j^{\phi_j} \left[1 - s_j\right]^{\frac{1-\eta}{3\beta}} \cdot \frac{\bar{h}_{j,g}\tilde{z}_{j,g}}{\tau_{j,g}} \tag{21}$$

Equation (20) does not depend on the cohort because the choice is made once, when individuals are young, and remains the same throughout their lifetime.

C.3.1 Proposition 2: Average quality of workers in occupation j

The geometric average of worker quality in each occupation is equal to:

$$\exp\left(\mathbb{E}\log\left[h_{j,g,c,t}\,\epsilon_{j,g,c}\right]\right) = \bar{\Gamma} s_{j,c}^{\phi_{j,t}} \gamma(t-c) \left(\frac{\eta s_{j,c}^{\phi_{j,c}} \bar{\gamma} \bar{h}_{j,g} w_{j,c}\left[1-\tau_{j,g,c}^{w}\right]}{1+\tau_{j,g,c}^{h}}\right)^{\frac{\eta}{1-\eta}} \left(\frac{1}{p_{j,g,c}}\right)^{\frac{1-\delta}{\theta(1-\eta)}} \tag{22}$$

The average quality of workers in occupation j varies by c, group g, and time t.

C.3.2 Proposition 3: Average wages

The geometric average of earnings in occupation j by cohort c in period t of group g equals:

$$\begin{aligned}
&\overline{\text{wage}}_{j,g,c,t} \equiv \left(1-\tau_{j,g,t}^{w}\right) w_{j,t} e^{\mathbb{E}\log[h_{j,g,c,t}\epsilon_{j,g}]} \\
&= \bar{\Gamma}\bar{\eta}\left[p_{j,g,c}^{\delta} m_{g,c}\right]^{\frac{1}{\sigma(1-\eta)}} \tilde{z}_{j,g,c}^{-\frac{1}{1-\eta}} \left[1-s_{j,c}\right]^{-\frac{1}{3\beta}} \times \frac{1-\tau_{j,g,t}^{w}}{1-\tau_{j,g,c}^{w}} \frac{w_{j,t}}{w_{j,c}} \frac{\gamma(t-c)}{\bar{\gamma}} \frac{s_{j,c}^{\phi_{j,t}}}{s_{j,c}^{\phi_{j,c}}}
\end{aligned} \tag{23}$$

C.3.3 Proposition 4: Relative propensities

The proportion of Black individuals of group g in cohort c employed in occupation j relative to white individuals equals:

$$\frac{p_{j,Black}}{p_{j,white}} = \left(\frac{\tau_{j,Black}}{\tau_{j,white}}\right)^{-\frac{\theta}{1-\delta}} \left(\frac{\bar{h}_{j,Black}}{\bar{h}_{j,white}}\right)^{\frac{\theta}{1-\delta}} \left(\frac{\overline{\text{wage}}_{j,Black}}{\overline{\text{wage}}_{j,white}}\right)^{-\frac{\theta(1-\eta)}{1-\delta}} \tag{24}$$

C.3.4 Proposition 5: Relative labor force participation

The proportion of Black individuals in the home occupation relative to white individuals for cohort c is equal to:

$$\frac{1-\text{LFP}_{Black}}{1-\text{LFP}_{white}} = \left(\frac{\overline{\text{wage}}_{j,Black}}{\overline{\text{wage}}_{j,white}}\right)^{-\theta(1-\eta)} \left(\frac{\tilde{z}_{j,Black}}{\tilde{z}_{j,white}}\right)^{-\theta} \left(\frac{p_{j,Black}}{p_{j,white}}\right)^{\delta} \tag{25}$$

C.4 Firms

In this economy, output is produced by one firm that aggregates labor inputs from J occupations through the production function:

$$Y = \left[\sum_{j=1}^{J} \left(A_j \cdot H_j \right)^{\frac{\sigma-1}{\sigma}} \right]^{\frac{\sigma}{\sigma-1}} \tag{26}$$

Where H_j is total efficiency units of labor in each occupation, σ is the elasticity of substitution across occupations, A_j is the exogenously-given productivity of occupation j.

C.5 Equilibrium

$H_{j,t}^{demand}$ that satisfies the profit maximization equals:

$$H_{j,t}^{demand} = \left(\frac{A_{j,t}^{\frac{\sigma-1}{\sigma}}}{w_{j,t}} \right) Y_t \tag{27}$$

And $w_{j,t}$ clears the labor market in each occupation such that $H_{j,t}^{supply} = H_{j,t}^{demand}$.

C.6 Estimation

To estimate the model, I follow the steps of Hsieh, Hurst, Jones and Klenow (2019). I use both internally calibrated parameters (on occupational shares and average wages by group and occupation) from the CPS and externally calibrated moments from Hsieh, Hurst, Jones and Klenow (2019). I use data from the March Annual Demographic Survey files of the Current Population Survey (CPS), and build an occupation that I call "scientists and inventors" combining individuals who report working in "engineering" and "natural scientists".[43]

I start from proposition 4, which governs the propensity of Black individuals compared to white individuals to be employed in occupation j. This equation tells us that the propensity of Black individuals to work in occupation j compared to white individuals depends on relative frictions, on relative talent, and on the average wage Black-white gap.

To estimate the model, I impose a series of additional assumptions following Hsieh, Hurst, Jones and Klenow (2019). First, in the main estimation, I assume that individuals only select

[43] More details coming soon.

occupations based on talent (not on individual preferences). That is, that $\mu_j = 1$ for each occupation j. Secondly, I assume that talent is distributed equally across groups in a given sector. Third, I assume that white individuals do not face any barriers in the acquisition of human capital or in the labor market. That is, $\tau^h_{white} = 0$ and $\tau^w_{white} = 0$ in all occupations and all periods. Fourth, I normalize preference for the home sector to be equal to 1 for all groups. Fifth, I assume that the return to experience γ is the same for all sectors, groups, and cohorts.

I begin by estimating ϕ_j (return to schooling for occupation j), $z_{j,white}$ (preference of white individuals to work in occupation j with respect to staying at home), and w_j (wage per efficiency unit for occupation j).

The return to schooling for occupation j (ϕ_j) is pinned down by the first order condition for schooling:

$$\phi_j = \frac{1-\eta}{3\beta} \cdot \frac{s_j}{1-s_j} \tag{28}$$

To recover s_j, I assume that the pre-market period is 28 years long (although the estimation results are virtually unchanged if I use a lower year threshold).

$$s_j = \frac{\text{Years of education}}{28} \tag{29}$$

To estimate the preference of white individuals to work in occupation j with respect to staying at home (z_j) for J occupations, I use data on the earnings of white individuals in the young cohort. Defining $m_{white} = \sum_{j=1}^{J} \widetilde{w}^{\theta}_{j,white}$ and rearranging the equation for average wage:

$$m_{white} = \left[\frac{\overline{\text{wage}}_{j,white} \tilde{z}_{j,white} \left(1-s_j\right)^{\frac{1}{3\beta}}}{\bar{\Gamma}\bar{\eta}} \right]^{\theta(1-\eta)} \tag{30}$$

To recover m_{white} (which does not vary by occupation), I plug in values for j = home. By assumption, $\tilde{z}_{home,white}$ is normalized to 1. Since there is no wage data for the home occupation, I assume that earnings in home occupation are equal to the average earnings in another occupation. I follow Hsieh, Hurst, Jones and Klenow (2019) and use the average wage in occupation "Secretaries". To compute $\tilde{z}_{j,white}$ for all occupations, I plug m_{white} into equation (30).

The wage per efficiency unit (w_j) is obtained by rearranging the occupational share equation

for white individuals:

$$w_j = \frac{\left[p_{j,white} \cdot m_{white}\right]^{\frac{1}{\theta}}}{\bar{\gamma} \cdot s_j^{\phi_j} \left[\left(1 - s_j\right) z_{j,white}\right]^{\frac{1-\eta}{3\beta}}} \tag{31}$$

After making an initial guess about the return to experience γ, all other variables are known, which allows to pin down the value of w_j (which does not vary by group or cohort).

To estimate the return to experience (γ), I use the change in the average wage of cohort c and occupation j over time. The ratio of average wage in occupation j at time t with respect to time c (when that cohort is young) is equal to:

$$\frac{\overline{\text{wage}}_{j,white,c,t}}{\overline{\text{wage}}_{j,white,c,c}} = \frac{w_{j,t} \cdot \gamma(t-c) \cdot s_j^{\phi_{j,t}}}{w_{j,c} \cdot s_j^{\phi_c}} \tag{32}$$

Since I have $\overline{\text{wage}}$ from the empirical moments, w_j, s_j, and ϕ_j from the previous steps, I can recover γ. By assumption, it is the same across all occupations j and cohorts c. $\bar{\gamma}$ is then computed as the average across all occupations j and cohorts c. The final step to recover $z_{j,white}$ and w_j is to iterate over equations (30) through (32).

Now, I turn to estimating τ^h and τ^w. After applying assumptions 1. to 3. and rearranging terms, I obtain the following expression for $\tau_{j,Black,c}$:

$$\tau_{j,Black,c} = \left(\frac{p_{j,Black,c}}{p_{j,white,c}}\right)^{-\frac{1}{\theta}} \left(\frac{\overline{\text{wage}}_{j,Black,c}}{\overline{\text{wage}}_{j,white,c}}\right)^{-(1-\eta)} \tag{33}$$

The first term on the right comes from the data: it is the ratio of Black to white individuals in occupation j and cohort c. The second term also comes from the data: it is the ratio between the geometric average of the wage of Black individuals in occupation j and cohort c to white individuals same occupation-cohort. By assumption, $\theta = 2$ and $\eta = 0.103$.

As $\tau_{j,Black,c}$ is estimated from the data only for the young cohort in each period, τ varies at the period- and occupation- level.
Finally, I use the cohort structure to recover the components of composite τ: τ^w and τ^h. For each occupation j and period t, $\tau^w_{j,Black,t}$ represents *labor market* barriers faced by all Black individuals in the labor market at time t. For each occupation j and cohort c, $\tau^h_{j,Black,c}$ represents *human capital accumulation* barriers, faced by all Black individuals who entered the labor market in period $t = c$. By definition, $\tau_{j,Black,c,t}$ is equal to:

$$\tau_{j,Black,c,t} \equiv \frac{\left(1 + \tau^h_{j,Black,c}\right)^{\eta}}{1 - \tau^w_{j,Black,t}} \tag{34}$$

To recover the τ's, I start from equation 34 for the young cohort in each period t. In what follows, I will omit subscripts j (for occupation) and Black individuals, but all τs are always specific to an occupation, and specific to Black individuals (relative to white individuals).

1. Recovering $\tau^h_{Cohort=3}$ and $\tau^w_{t=1}$

   (a) Define α as the Cobb-Douglas split of composite τ. Specifically:

   $$\tau^{\alpha}_{c,t} = \frac{1}{1-\tau^w_t} \tag{35}$$

   and

   $$\tau^{1-\alpha}_{c,t} = (1+\tau^h_{c-t})^{\eta} \tag{36}$$

   (b) Setting the initial value of α to 0.5 for the young cohort in period 1 (i.e. cohort 3), and plugging in $\tau_{Cohort\ 3,Period\ 1}$ from the data, I compute τ^w for period 1, and τ^h for cohort 3 (i.e. the young in period 1).

2. Recovering $\tau^w_{t=2}$

   (a) When $\delta = 0$, individuals are heterogeneous on talent (and not on preferences), average talent (and, by Proposition 2, also average wages) are inversely related to the share of the group working in occupation j. Rearranging Proposition 3 combined with Proposition 2, wage growth for cohort c in a given occupation by group (white or Black) is equal to:

   $$\frac{\overline{\text{wage}}_{j,g,c,t+1}}{\overline{\text{wage}}_{j,g,c,t}} = \frac{1-\tau^w_{j,g,t+1}}{1-\tau^w_{j,g,t}} \cdot \frac{w_{j,t+1}}{w_{j,t}} \cdot \frac{(s_{j,c})^{\phi_{j,t+1}}}{(s_{j,c})^{\phi_{j,t}}} \tag{37}$$

   (b) By assumption, both τ^h and τ^w equal zero for white individuals. Therefore, equation 37 for white individuals is equal to:

   $$\frac{\overline{\text{wage}}_{j,white,c,t+1}}{\overline{\text{wage}}_{j,white,c,t}} = \frac{w_{j,t+1}}{w_{j,t}} \cdot \frac{(s_{j,c})^{\phi_{j,t+1}}}{(s_{j,c})^{\phi_{j,t}}} \tag{38}$$

   (c) Dividing equation 37 by equation 38 and rearranging terms:

   $$\frac{\overline{\text{wage}}_{j,Black,c,t+1}}{\overline{\text{wage}}_{j,Black,c,t}} = \frac{1-\tau^w_{j,Black,t+1}}{1-\tau^w_{j,Black,t}} \cdot \frac{\overline{\text{wage}}_{j,white,c,t+1}}{\overline{\text{wage}}_{j,white,c,t}} \tag{39}$$

   For a given occupation-cohort, the wage growth for Black individuals is equal

to the wage growth for white individuals, times the growth rate of labor market barriers faced by Black individuals.

(d) Wage growth comes from the data, and $\tau^w_{j,Black,t}$ is known from above. Therefore, this equation allows to compute $\tau^w_{j,Black,t=2}$.

3. Recovering $\tau^h_{j,Cohort=2}$

   To compute the barriers to human capital accumulation faced by the cohort who enters the market in period 2, I plug $\tau^w_{j,t=2}$ into the definition of composite τ for the young cohort in period 2. In the benchmark estimation, the minimum value of τ^h is set to -0.80.

4. Recovering $\tau^w_{j,t=2}$

   To recover labor market barriers for period 2, I follow the steps outlined in point 2. plugging in $t = 2$, $t + 1 = 3$, and $c = 2$.

5. Recovering $\tau^h_{j,Cohort=1}$

   To compute human capital barriers faced by the cohort who enters the labor market in period 2, I plug $\tau^w_{j,t=2}$ into the definition of composite τ (equation 34) for the young cohort in period 2.

Finally, since in the model occupations are chosen when young, labor force participation remains the same when middle-aged and old. The wage moments need to be corrected for the fact that the composition of workers changes as individuals move in and out of the labor force. Hsieh, Hurst, Jones and Klenow (2019) apply a common adjustment across all occupations to obtain the wage growth estimate. As the elasticity of labor force participation with respect to wage growth is $\frac{\theta(1-\eta)}{1-\delta}$, the common adjustment is:

$$\left(\frac{\text{wagegrowth}_{\ j,Black}}{\text{wagegrowth}_{\ j,white}}\right)^{\text{Estimation}} = \left(\frac{\text{wagegrowth}_{\ j,Black}}{\text{wagegrowth}_{\ j,white}}\right)^{\text{Data}} \left(\frac{\text{LFPgrowth}_{\ j,Black}}{\text{LFPgrowth}_{\ j,white}}\right)^{\frac{1-\delta}{(1-\eta)}} \tag{40}$$

D Description of Data Sources

Updated Description Coming soon