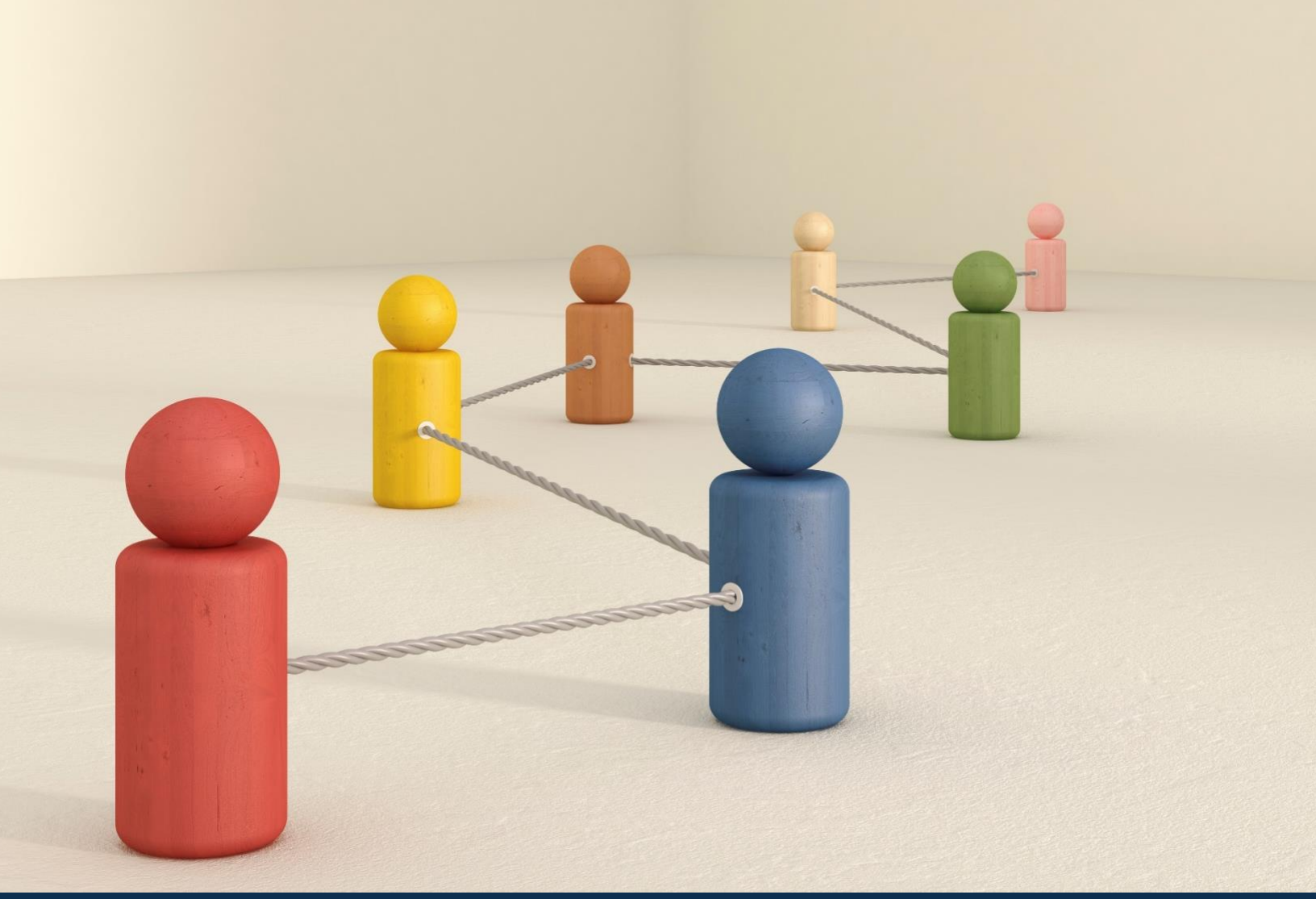

APPLYING CAUSAL ANALYSIS TO EQUITY ASSESSMENTS

RACIALLY DISPARATE NIH AWARD RATES

Authors: Saimun Habib, Mary Munro, Joshua Z. Stadlan, Lilly Boyer, Tamey M. Habtu

November 2023

MITRE | SOLVING PROBLEMS FOR A SAFER WORLD®

INTRODUCTION

Problem

Equity practitioners are pulled in different directions on the appropriate role of quantitative data in program equity assessments and intervention development. On the one hand, program data disaggregated by demographics can shed light on systemic disparities, allowing for the detection of patterns that individuals affected by the disparities may not realize were not just personal but pervasive and identity-based [1]. Quantitative disparity data also may be more convincing to skeptical decision makers who trust numbers over individuals voicing their experience [2].

On the other hand, the pervasiveness of disparities by race/ethnicity, gender, and other protected classes in data across measures of life resources, access, and life outcomes [3]—in health, industry, education, housing, the criminal justice system—may obscure a program's potential to reduce inequity in their own program and may not lead to interventions [4]. Overreliance on data collected by particular researchers, organizations, or methodologies may embed social assumptions and biases [5], especially as academia has historically prioritized methods and researchers from certain backgrounds over others [6]. Worse, the all-encompassing disparities in data on life outcomes without context may be misinterpreted to reinforce stereotypes and evidence-less essentialism [7].

We argue that a causal analysis framing to equity assessment and inequity intervention can access the benefits of quantitative evidence while limiting the drawbacks of quantitative-only analyses, as a mixed-method approach [8]. Causal framing requires producing a fleshed-out model of a process and interrogating the relationships between observed variables and unobserved variables. The mapping of variables and the outcome in question can reveal sources of inequity and pathways in which to intervene. This mapping is called a Directed Acyclic Graph (DAG) and represents formal causal relationships. Having this mapping in place guides appropriate selection of research questions, models, and even variables in an analysis. Furthermore, it prevents misinterpretation of the data. For instance, humans can intuitively understand that when there is rain, people will use umbrellas, and in very modest terms, the presence of rain causes umbrellas. There is an association between rain and umbrellas but only a one-way causation. A statistical model is naïve about the direction of causation and will just as confidently suggest that umbrellas cause rain as vice versa based purely on estimates of correlation. It is up to the researcher to put in place a causal framework beforehand to prevent such misinterpretations of the data.

Consider a research grant program that awards grants to White principal investigators (PIs) at higher rates than Black PIs. By causal framing to equity assessment, we mean focusing on an outcome of a program (receiving the grant) that is disparate for a protected class: race. A causal framing entails hypothesizing the network of cause-and-effect that brings about that outcome: the causal models. To create a valid model, the participants must use their experience and expertise of how the grant program works, the candidate variables that are considered, and how they are related to receiving the grant. With these relationships put into place, empirical analysis can be done with confidence that it is accurately capturing sources of inequity. In the equity assessment context, the hypothesized cause of focus is the protected class—race—to evaluate to what extent

the program is causing disparate outcomes: Had applicant X been of a different race, would the outcome have been different?

This framing helps isolate how a program *is* perpetuating inequity and provides a basis for impactful interventions. A causal approach can also distinguish between causes: The data is unlikely to equally support the causal pathway hypotheses of overt discrimination, indirect discrimination, and differences in population characteristics. It also integrates crucial qualitative data of lived experience and professional expertise: Those impacted by the program and the designers of the program alike share their mental models based on their own experience of how the program is working and interacting with the outcome disparities. The output from the causal framing lends itself to these groups co-designing interventions based on the causal mechanisms identified and estimating their relative impact using the causal model. The DAG itself is a qualitative heuristic-based tool that integrates different sources of input, and it provides the foundation and boundaries for further quantitative analysis. Statistical modeling and analysis do not draw their validity from the data itself but instead are externally validated by the researcher's understanding of the problem. Statistical models ultimately are not, themselves, the scientific models or theory but rather tools for verifying models and theories. Causal frameworks, such as DAGs, put forth a model or theory for how a system is inequitable, and statistical modeling allows for evaluation of those effects.

This framing also helps shift the burden of evidence in addressing disparities from the those raising the issue to those representing the inequitable status quo. The following sequence of events may sound familiar: Imagine that a group of Black researchers approach the program manager, sharing their sense that the grant selection process is stacked against them, and even providing data evidence of the disparity in grant awards (the program outcome). A defensive program manager may point to the program's name-blind and therefore ostensibly "color-blind" selection process as a rejection of any power over the disparate outcome. A more sympathetic program manager may assume the same blamelessness of the program itself but set up an ad-hoc diversity committee to conduct outreach to Black researchers.

A causal framework can advance equity advocates' cause as follows: The petitioners can share their causal view—perhaps they put forward the hypothesis of explicit discrimination, that grant selection is determined by race (which we can represent as "race → grant selection"). In other words, they are presenting a straightforward DAG, proposing that the Total Effect is entirely a Direct Effect. Given a disparate outcome, and no competing hypotheses offered by the program designers, that will by default be the best hypothesis under consideration.

We figure that the program designers will be interested in defending against the hypothesis that they are engaged in explicit discrimination, whether or not they are personally invested in reducing inequity. Therefore, they are incentivized to support their team in investigating and improving the accuracy of the DAG—uncovering confounders and mediators that they think are justified and reducing the estimated Direct Effect of race on the grant award rate outcome. Meanwhile, the confounders and mediators also provide the means to reduce the disparate impact of the program; program co-designers can try to propose factors for the decision-making process

that are less discriminatory than the current mediators, and launch initiatives that reduce the influence of the confounders.

Benefit to Anti-Discrimination Legal Cases

In certain cases, implicit discrimination is even proscribed by law. The U.S. Supreme Court ruled that the Civil Rights Act Title VII prohibition of employment discrimination based on "race, color, religion, sex or national origin" extends to "not only overt discrimination but also practices that are fair in form, but discriminatory in operation" [9]. The Sixth Circuit U.S. Court of Appeals adopted this understanding as well for the Fair Housing Act, as noted in the Reinstatement of the U.S. Department of Housing and Urban Development's (HUD) Discriminatory Effects Standard [10].

HUD's Discriminatory Effects Standard in CFR 100.500 establishes liability for a housing action that "actually or predictably results in a disparate impact on a group of persons or creates, increases, reinforces, or perpetuates segregated housing patterns because of race, color, religion, sex, handicap, familial status, or national origin," independent of "discriminatory intent." A valid defense of the action is that the action "is necessary to achieve one or more substantial, legitimate, nondiscriminatory interests." A successful counter to this defense is that an alternative practice can accomplish the same ends with less of a discriminatory effect.

We suggest that a causal inference framing could provide a process for these evidentiary standards. The first claim amounts to demonstrating a treatment effect from a protected class to a housing action outcome. A defense of the action is to propose a data-validated causal model that explains this treatment effect exclusively through mediators that serve legitimate interests. A valid counterargument would be a causal model with these legitimate interests as the outcome variable and demonstrating that this new model would reduce the disparity of outcome.

Case Study: NIH Grants

To demonstrate the technical process of applying a causal inference framework to equity assessment, we focus on our example of racial disparities in research grant awards. Specifically, we look at the U.S. National Institutes of Health's (NIH) R01 award. An R01 grant is the most widely used investigator-initiated research project grant for hypothesis-driven research projects with strong preliminary data [11].

Ginther, Schaffer, Schnell, et al. explored the association between an applicant's race or ethnicity and the probability of receiving an R01 award in a 2011 report, "Race, ethnicity, and NIH Research Awards" [12]. They found that, compared to White applicants, Asian applicants were 4 percentage points less likely to receive NIH research funding, and Black/African-American applicants were 13 percentage points less likely to receive NIH research funding. In addition, after controlling for a number of factors, they found that Black/African-American applicants remained 10 percentage points less likely to be awarded NIH research funding than White applicants. The report cautioned against understanding the variable correlations found as causal impact and called for further study.

NIH responded to these findings with investigations and deliberations by the NIH Advisory Committee to the Director (ACD) and Working Group on Diversity in the Biomedical Research Workforce (WGDBRW) [13]. NIH has since worked on implementing the ACD's recommendations in data collection, mentoring, institutional support, and testing bias interventions; the approval rate percentage point gap grew in 2011–2019, and narrowed in 2020–2021 [14].

However, the working group was not able to directly connect their data analysis and investigations to their recommendations:

The WGDBRW was unable to precisely distinguish among funding disparities caused by the potential presence of bias (unintended or otherwise) during the peer review process (see Section V for a discussion of bias) and application quality, which in turn may be affected by a wide range of factors including mentorship, resource availability, release time from teaching/administrative responsibilities, all of which could potentially be influenced by institutional bias (unintended or otherwise). Thus, because the WGDBRW's analyses and discussions did not point to a single, definitive cause for NIH-funding disparities—and the group recognizes fully that causes are unlikely to be mutually exclusive—the WGDBRW has proposed a set of complementary interventions…

(Working Group on Diversity in the Biomedical Research Workforce and the Advisory Committee to the Director, 2012) [15]

Without explicitly presenting the hypothesized root causes of inequity, it is challenging to develop interventions and then to understand whether they are working once implemented.

Because of the open question around the root causes in the racially disparate NIH R01 approval rates, we returned to the data analyzed in Ginther et al. with a causal inference framework. In the original study, the authors used data from the NIH grant application and award database (IMPAC II), the Thomson Reuters Web of Science, and other sources. For this study, we used the de-identified dataset created for replication purposes[1]. The dataset includes key variables used in the original study, such as race, type of grant, employer characteristics, and previous NIH grants. The dataset, however, is not identical, with several variables having been omitted or recoded for privacy purposes. Our analysis builds on the associations found in the original study by identifying causal pathways and the strength of the causal pathways from race to receiving an R01 award.

The case study proceeds as follows: First, we review the concept of causal inference and introduce *cfairer*, our R package[2] for causal analysis of unfairness in data. *cfairer* assists in identifying the causal model, estimating the magnitude of the causal relationships, and

[1] https://report.nih.gov/nih-supported/invstigators-and-trainees

[2] https://github.com/cfairer/

generating "fairer" counterfactuals. Later, we characterize the NIH grant dataset by racial identity, conditioning on several variables of interest. We then present preliminary results of the mediation analysis and conclude with their implications for reducing inequity in the grant awards.

METHODS

Causal Inference

The field of causal inference extends the scope of traditional statistics, which is limited to learning from associations among elements of a system, toward understanding of how an outcome would have changed had an element been different from what was observed [16]. It provides a paradigm to estimate the effect of hypothesized causes on an outcome of interest. This can then be considered an important tool for determining whether these outcomes are unfairly influenced by responses to sensitive attributes, such as identities that have been historically or ongoingly marginalized or underserved, such as in race, gender, sexual orientation, and socioeconomic class. Estimating the causal effect of a particular factor on elements in a system will then allow one to quantitatively describe what a fairer system would look like. The concept of path-specific counterfactual fairness, introduced by Chiappa et al., 2018, is the basis for the approach supported by *cfairer* [17]. Essentially, not only are we interested in the direct effect of discrimination against an identity attribute on an outcome, but we also care about the propagation of the causal effect of that identity attribute through other system variables to the outcome.

To move from association to causal relationships one must estimate a different quantity of interest, which requires stronger assumptions about the data-generating distribution. The observed association between a cause, or exposure, and outcome is

$$Y = E(outcome = o \mid exposure = e),$$

while the causal effect can be denoted as

$$Y(e) = E(outcome \mid do(exposure = e)).$$

Pearl coined this term as the do-operator, and it allows one to reason about interventions instead of just observed associations. These hypothetical outcomes under the do-operator are often referred to as *potential outcomes* or *counterfactuals*, where only one potential outcome can be observed per unit. Estimating the interventional effect of a cause requires at a minimum exchangeability between exposure groups, so that the outcome is unchanged regardless of choice of treatment and control subpopulations. Exchangeability implies the absence of a confounding factor, thus enabling identifiability, which means that the causal quantities may be estimated from observed associations. At its core, it implies the only difference between the treated and the untreated is precisely that their treatment differed. If the roles were switched, we would expect similar treatment responses when controlling for covariates. When we want to examine the causal effects of demographic characteristics like race, we cannot randomly assign race and also only observe it. Despite not controlling treatment assignment, we can treat a characteristic variable as a distribution—perhaps conditioned on other variables—and an observation's "treatment" as a draw from that distribution.

The most general estimate used to quantify the causal effect of an exposure (or treatment) on an outcome is the Average Treatment Effect (ATE). This is the difference in the pair of potential outcomes averaged over the entire population of individuals. Say we are interested in the causal effect of sex on NIH grants. In the NIH dataset, sex was defined as male or female, S={M,F}, and grants were either awarded or not awarded, G={1,0}, so the ATE of sex on whether a grant is awarded is,

$$\text{ATE} = \text{E}[\text{Y(M)} - \text{Y(F)}] = \text{E}[(\text{G} \mid \text{do}(\text{S} = \text{M})) - (\text{G} \mid \text{do}(\text{S} = \text{F}))] \quad (1)$$

While this is impossible to calculate at an individual level, given that a single unit can only be one sex at the time of grant decision, it can be estimated at a population level with

$$\text{ATE} = \text{E}[\text{G} \mid \text{do}(\text{S} = \text{M})] - \text{E}[\text{G} \mid \text{do}(\text{S} = \text{F})].$$

If one can assume exchangeability across the male and female candidates in the dataset (which is an untestable assumption), then the estimate simplifies to the association estimate, which can be calculated from the observed data:

$$\text{ATE} = \text{E}[\text{G} \mid \text{S} = \text{M}] - \text{E}[\text{G} \mid \text{S} = \text{F}] \quad (2)$$

It's often helpful to visualize the relationships defining a causal model with the aid of a Directed Acyclic Graph (DAG). Nodes represent features, while edges represent a causal relationship. A general DAG can be seen in Figure 1.

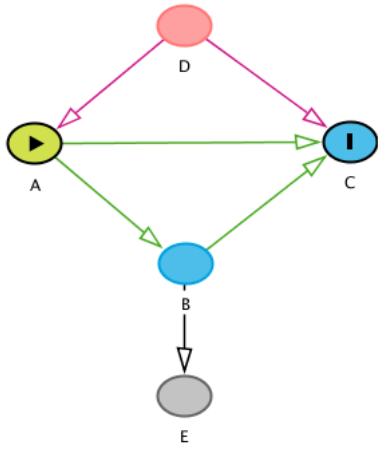

Figure 1: Components of a Directed Acyclic Graph (DAG)

The edge pointing from A to C means that A has a causal effect on C, A is a parent of C, and C is a child of A. Descendants of A are all nodes along directed paths originating from A, namely {B, C, E}. B is a mediator of the causal effect of A on C since it can explain, or mediate, the effect of A on C. D is a confounder of the causal effect of A on C since it has a causal effect on both A and C. Confounding is frequently considered the main shortcoming of observational studies owing to its influence on the exposure, A, and the outcome, C, and therefore needs careful adjustment. The term acyclic means that no directed path forms a closed loop. The acyclic requirement follows from the idea that time increases as a causal path is traversed, so a directed path returning to a node would imply that the future affected the past. For a more detailed overview on the makeup of the causal inference space, see [16]. For more on fairness with a causal lens, see [17, 18].

cfairer

The *cfairer* R package was used to assist in the causal analysis of inequity in data. The *cfairer* workflow was designed to facilitate three steps:

Identification: Through the combination of observed data, lived experience, and domain expertise, we can recover the causal structure that generated the process through structure learning (also called causal discovery).

Estimation: From the observed data and causal graph, we decompose and quantify the pathway effects using mediation analysis.

Mitigation: From the observed data, causal estimates, and user input, we make counterfactual predictions to "fix" the observed data where outcomes are *fairer* with respect to identity attributes of interest.

We will now give an overview of the approach of each of these 3 steps.

Causal Graph Identification through Structure Learning

In the first step of the *cfairer* process, the user works with the package to identify the causal structure of the system using the observed data. After specification of the preferred structure learning algorithm, the package outputs a suggested initial DAG. The user may then fine-tune the components of the DAG based on domain expertise by adding or deleting nodes and edges. Automatic generation of a *correct* DAG solely from data is typically impossible since observed data can at best only generate a Markov equivalence class graph [19]. That is, Causal Discovery generates a set of possible DAGs sharing the same set of conditional independencies, leaving it up to the user to make the final choice in edge directionality. Furthermore, there are usually sample-design factors, such as how the data was collected (e.g., censoring, selection bias, omitted variable bias), which will result in systematic differences between the true causal DAG and any DAG suggested by structure learning algorithms, thus requiring subject matter expertise for validation.

There are multiple approaches to structure learning, each with its own set of tradeoffs. Factors such as computational complexity, statistical assumptions, and classification errors such as false positive or negative rates play a role in algorithm choice. *cfairer* uses the *bnlearn R* package [20], which supports the three most common classes of structure learning algorithms: *constraint-based algorithms,* which rely on conditional independence tests; *score-based algorithms,* which optimize objective functions made up of goodness-of-fit scores; and *hybrid algorithms.* See [21] for a detailed discussion

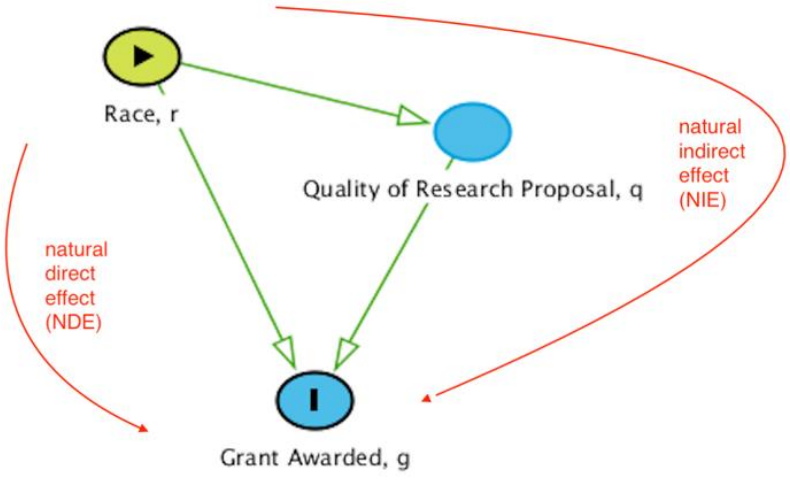

Figure 2: Direct and Indirect causal pathways of race, r, on NIH grant award, g

on the inner workings of these algorithms and the relative characteristic tradeoffs.

Estimation of Pathway Effects through Mediation Analysis

In the second step of the *cfairer* process, the strengths of causal pathways mediating the effect of an exposure on an outcome are computed. In our simple example in Figure 1, this mediation step can measure the degree to which variable B explains the causal effect of A on C. *cfairer* relies on the *medflex* R package to evaluate this [22]. To illustrate *medflex's* mediation analysis, consider this example, where an NIH grant is not only awarded based on the quality of a research proposal but also on race. We are interested in understanding the causal effect of race, *r*, a sensitive attribute, on the decision to award a grant, *g*, while also accounting for the quality of the research proposal, *q*. *Medflex* calculates three quantities: the *Natural Direct Effect*, which quantifies the strength of the causal of effect of *r* on *g* while conditioning on *q;* the *Natural Indirect Effect*, which quantifies the strength of the *indirect* causal effect of *r* on *g* through the mediating variable, *q;* and the *Total Effect*, which combines these two causal effects.

Recall the *Average Treatment Effect (ATE)*, which measures the causal effect of an exposure on an outcome. The *Total Effect* of *r* on *g* is calculated using the *ATE*:

$$ATE = \mathrm{E}[(\mathrm{g} \mid \mathrm{r} = 1) - (\mathrm{g} \mid \mathrm{r} = 0)] \quad (3)$$

The *Natural Direct Effect* is calculated by estimating the exposure-induced change in the outcome while keeping the mediator fixed at the value that had naturally been observed if unexposed *q(r=0)*:

$$NDE = E[(g|r = 1, q(r = 0)) - (g|r = 0, q(r = 0))] \quad (4)$$

A similar approach yields the *Natural Indirect Effect*, except *r* is kept constant while *q* is varied, reflecting the expected difference in outcome, *g*, if all subjects were exposed (*r=1)* but their mediator value, *q*, had changed to the value it would take if unexposed *q(r=0):*

$$NIE = E[(g|r = 1, q(r = 1)) - (g|r = 1, q(r = 0))] \quad (5)$$

Each counterfactual quantity nested in equations (3)–(5) is estimated with a generalized linear model trained on data using the canonical set of covariates, i.e., all ancestors of the mediator and outcome minus descendants on causal paths, for adjustment. More information on this and alternative conditioning sets that can be produced from a given DAG can be seen in [23].

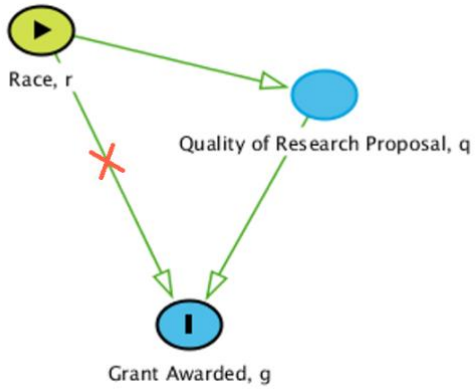

Figure 3: Removing the direct dependency of NIH grant award on applicant Race

Mitigation of Unfairness in Data

With a verified causal DAG and estimated mediating pathway effects, the third step of the *cfairer* process mitigates user-defined unfairness in the data. This unfairness is represented by certain undesirable pathways in the DAG, such as direct edges from

sensitive attributes to outcomes. This step removes an unfair edge by adjusting carefully chosen members of the outcome attribute's data through marginalization of the outcome over all possibilities of the protected attribute. For example, if NIH decides to award a grant based on the *Quality of the Research Proposal* submitted with *Race* as a confounder (see Figure 2), this step allows us to remove the deemed *unfair* direct edge between *Race* and *Awarded* by adjustment of the *Quality of Research Proposal* data column.

To understand how this works for this example DAG, it would be helpful to look at the joint probability distribution:

$$P(r,q,g) = P(g \mid \{r,q\}) \cdot P(q \mid r) \cdot P(r) \quad (6)$$

The first of the three terms, $P(g \mid \{r,q\})$, is considered unfair since the decision to award a grant is dependent on the sensitive *race* demographic attribute. To fix this, *race* is marginalized out of this term:

$$P(g \mid q) = \sum_{r} P(g \mid \{r,q\}) \cdot P(r) \quad (7)$$

so that the joint distribution in equation 6 is transformed to the fairer:

$$P(r,q,g) = P(g \mid q) \cdot P(q|r) \cdot P(r) \quad (8)$$

and the g data column has been fixed, or adjusted, according to equation (7). The conditional probability in equation (7), $P(g \mid \{r,q\})$, is estimated using a model for g trained using its parents $\{r,q\}$ as covariates. If g were to have descendant nodes, these would then be inconsistent with the newly adjusted g. *cfairer* recursively adjusts these descendant nodes of g using models trained on original data with their parent nodes as covariates.

Data Description

The de-identified dataset has 205,075 observations and includes applicants to NIH grants other than the R01. It also includes applicants to the R01 continuation grant, a grant for current R01 recipients to extend their funding. Additionally, the applicants

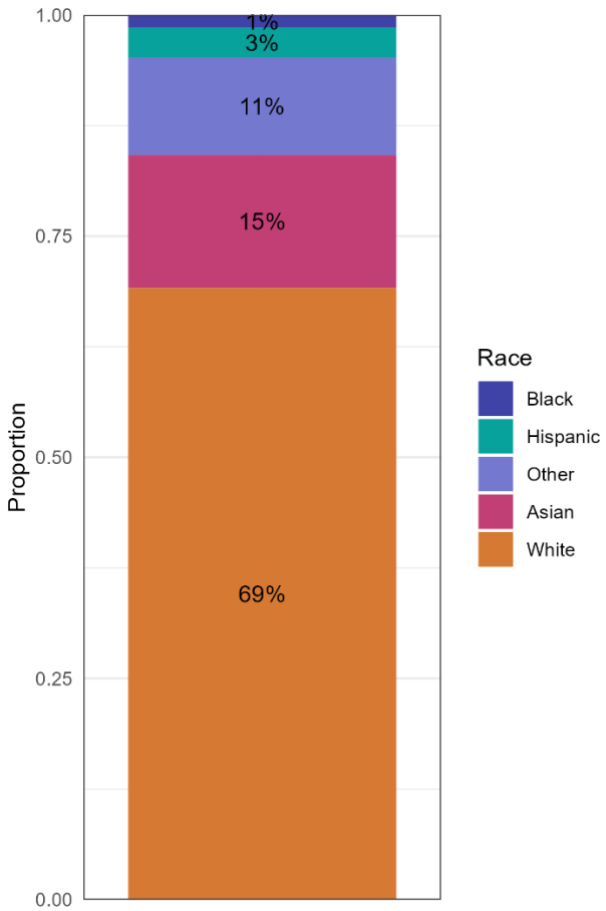

Figure 4: Race demographics of R01 applicants

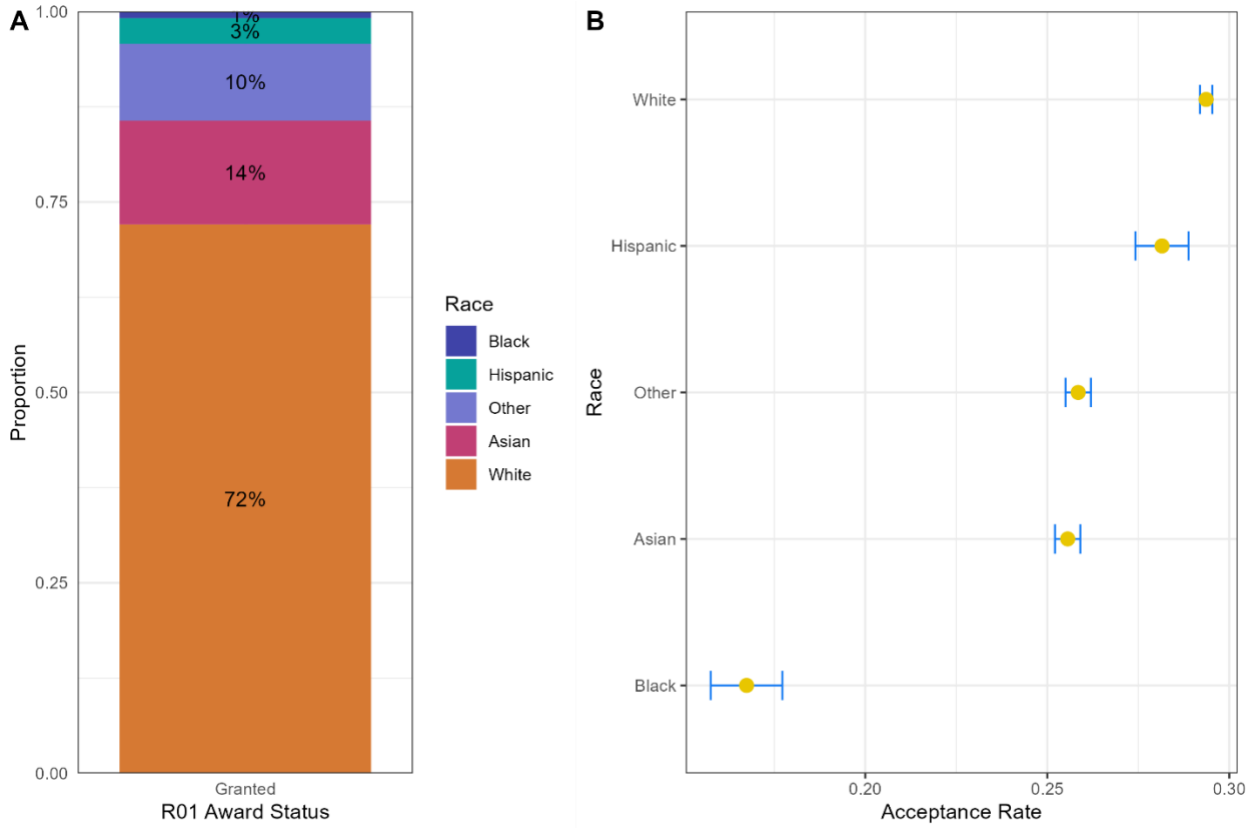

Figure 5: (A) R01 Awardees' demographics by race, (B) R01 acceptance rates by race

can by PhD holders, MDs, or some other degree. We filter out non-R01 applicants and R01 continuation applicants and have a total of 106,368 observations.

The racial demographic of this dataset is shown in Figure 4. Of the applicants, 69% are White, 15% are Asian, 3% are Hispanic, 1% are Black, and the remaining 11% are Other. In Figure 2, where we see applicants who successfully received an R01 grant, there is a roughly proportional distribution of applicants to those who applied, but with slightly more White applicants. Also in this figure, we see that acceptance rates across races are not equal, with Black applicants being rejected at a higher rate than others.

With respect to PhD-holding and non-PhD-holding applicants, the distribution of race is mostly similar, but there are more White applicants and Asian applicants holding PhDs and less applicants of Other race with PhDs proportionally. On the other hand, R01 award rates with respect to whether someone holds a PhD are similar.

We also check whether prior grants, productivity, and institution rankings differ with race and R01. We observe that having a prior NIH grant is associated with a higher chance of receiving an R01 grant, and White applicants make up more of those with a prior grant. With regard to institution ranks, we found that applicants' race demographics were similar across all institutions

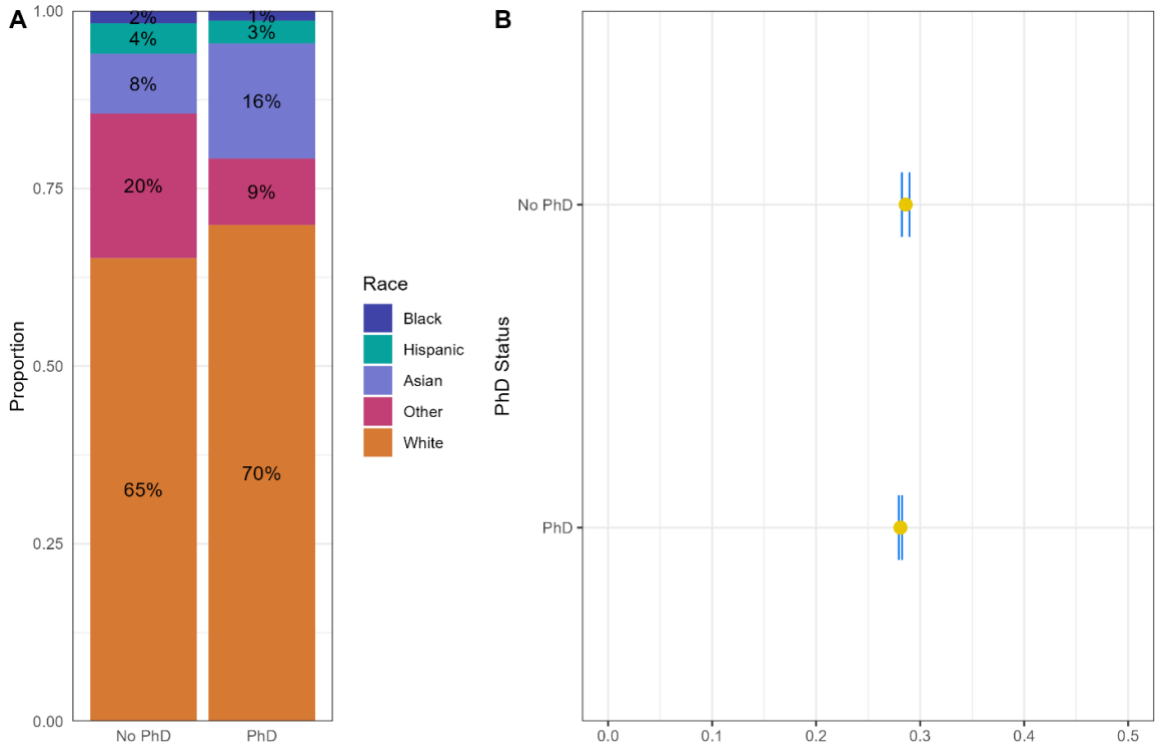

Figure 7: (A) Racial demographics of PhD and non-PhD holders, (B) R01 acceptance rates by PhD status

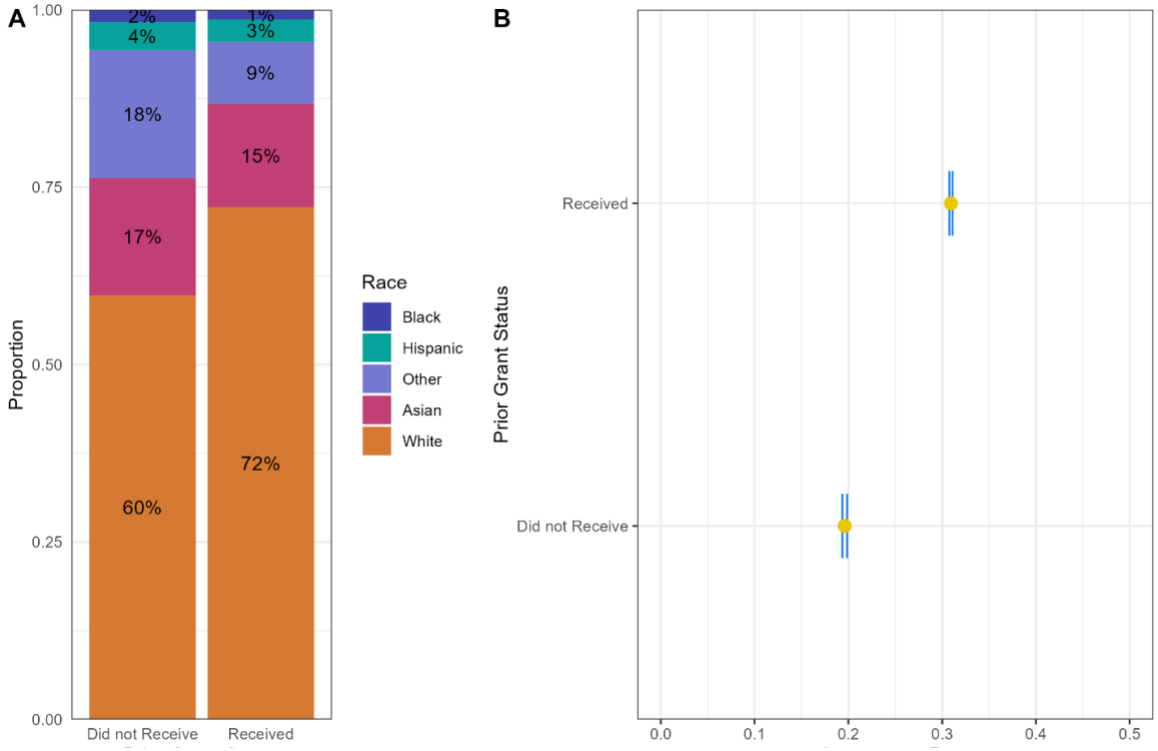

Figure 6: (A) Racial demographics of prior grant recipients, (B) R01 acceptance rates by prior grant status

and applicants from higher-ranked institutions have a lower likelihood of being rejected. Note, while applicant race is not visible to the grant committee, in practice race and other demographic

information may be inferred from institutional details [12]. Furthermore, while institutional rank is an important variable for understanding decision making, other factors such as public vs. private university, geography, and a host of other features may be relevant but are not available in the public dataset.

Table 1: Coefficients of original probit model(s) from Ginther et al. (2011). See Table A2 for variable descriptors.

	Model 1	Model 2	Model 3	Model 4	Model 5
raceAsian	-0.015 * (0.007)	-0.008 (0.007)	-0.011 (0.007)	-0.022 *** (0.007)	-0.009 (0.007)
raceBlack	-0.091 *** (0.023)	-0.089 *** (0.023)	-0.088 *** (0.023)	-0.070 ** (0.023)	-0.069 ** (0.023)
raceHispanic	-0.005 (0.013)	-0.000 (0.013)	-0.000 (0.013)	0.001 (0.013)	0.006 (0.013)
raceOther	-0.000 (0.008)	0.011 (0.008)	0.009 (0.008)	0.009 (0.008)	0.016 (0.008)
roleftk1		0.031 *** (0.005)	0.025 *** (0.005)	0.014 ** (0.005)	0.011 * (0.005)
fund_rank2			-0.025 *** (0.005)	-0.028 *** (0.005)	-0.028 *** (0.005)
fund_rank3			-0.039 *** (0.007)	-0.045 *** (0.007)	-0.047 *** (0.007)
fund_rank4			-0.103 *** (0.007)	-0.105 *** (0.008)	-0.112 *** (0.008)
cmte_c1				-0.015 ** (0.005)	-0.005 (0.005)
hs_y1				-0.066 *** (0.005)	-0.068 *** (0.005)
org_high1				-0.022 *** (0.007)	-0.025 *** (0.007)
priorgrant1				0.074 *** (0.006)	0.082 *** (0.006)
citq2					-0.027 *** (0.007)
citq3					-0.027 *** (0.008)
citq4					-0.018 (0.010)

pub_badmatch1					-0.027 *** (0.008)
pubq2					-0.030 *** (0.007)
pubq3					-0.055 *** (0.008)
pubq4					-0.060 *** (0.009)
Standard errors are heteroskedasticity robust. *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.					

RESULTS

We replicated the de-identified model estimates in R using a probit model and robust standard errors. To stay consistent with the authors (Table 1), we converted the coefficients into the average marginal effect of a variable, so the interpretation of the coefficient is a change in probability with respect to the independent variable. The first model looks at race and R01 award grant probability. The second model adds in whether applicants received an F, T, or K award prior. A third model includes the institution rank, and a fourth model adds if they served on an NIH committee, if the institution is a higher education institute, and if they've received a prior grant. The last fully defined model adds productivity with respect to citations and publications as well as whether the applicant was well-matched to productivity statistics. When just considering race, we see that being Black is associated with a 9% decrease in probability of getting the R01 grant awarded as compared to White applicants. As we add variables this figure decreases, but in the full model it is still a decrease of about 7%.

Using *cfairer* we create a DAG using the hill climbing algorithm and then alter the outputted DAG to orient all edges in a logically consistent manner. Within this DAG, the one step mediators are whether an applicant has received an NIH training grant in the past (F, T, or K grant), if they hold a PhD, if they've served on an NIH committee, their productivity in terms of

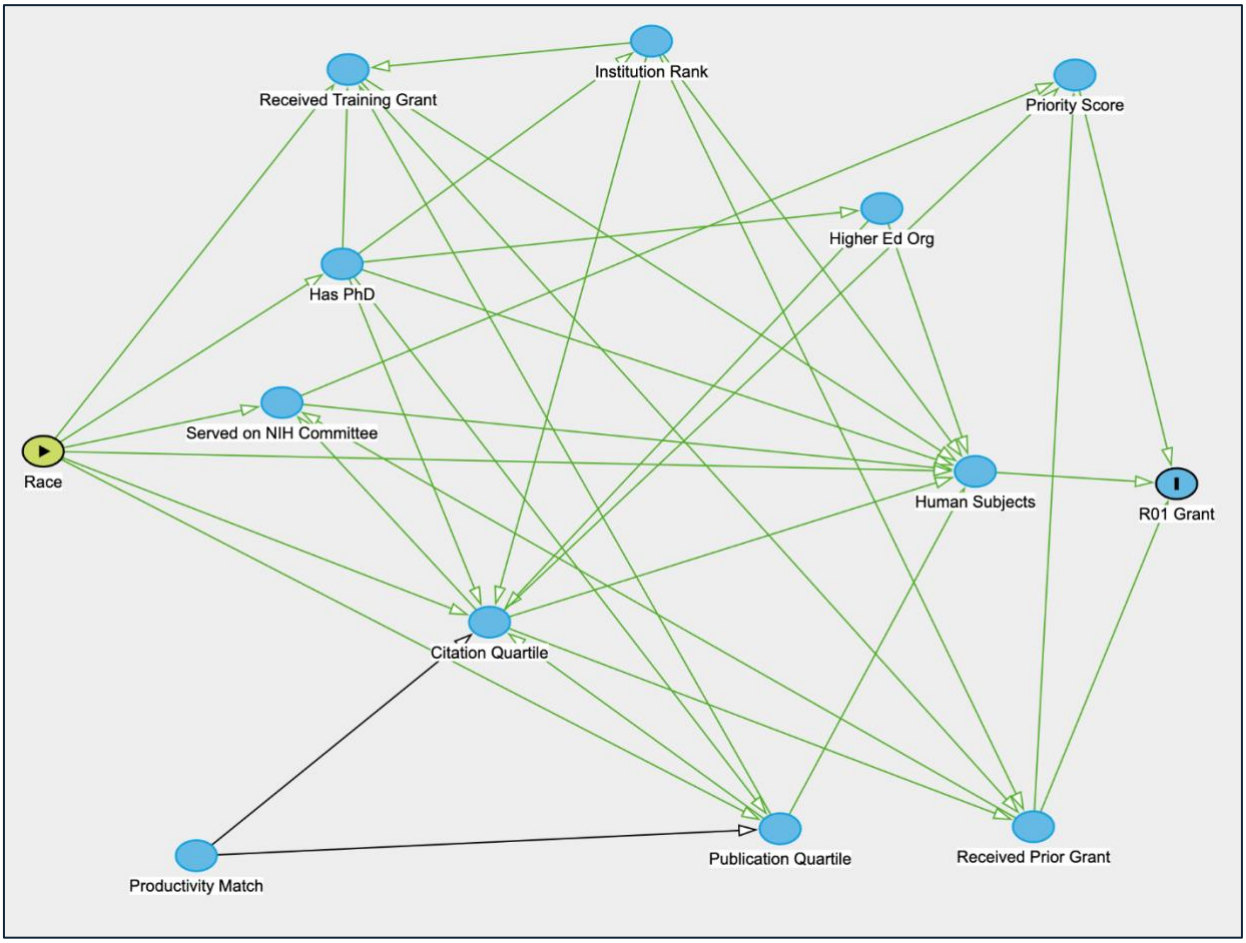

Figure 8: DAG of variables found in NIH R01 grant awards with Race as the treatment variable and R01 Grant as the outcome

citations and publications, and if their proposal uses human subjects. Note that there is no edge from race to R01 award outcome in this DAG—the DAG is only representing indirect effects—but the edge is implicitly included in the calculation of direct effects.

From this DAG, we see that all variables have to go through having a human subject, having a priority score, and whether they've received a prior grant before having an effect on award grant probability (Figure 8).

Using *cfairer*, we run a mediation analysis upon this DAG to estimate the effect of being Black versus non-Black on R01 award. The model estimates the total effect, direct effect, and effect of the mediators on R01 award probability. The total effect is the effect of Race on award probability when no mediators are included and was estimated to be 0.52 with a direct effect; the effect with mediators included was 0.53. The effects of the first step mediators are all roughly 1.

These results are reported as an odds ratio, meaning the interpretation of a direct effect of 0.52 is that Black applicants have half the probability of getting awarded the R01 award as non-Black applicants (Figure 9). Similarly, for the mediators (role, publication quartile, prior grants, organization rank, funding rank, committee status, and citation quartiles), the interpretation for a value of 0.98 is that, in the presence of the mediator probability of getting awarded, the R01 mediator is 98% of what it would be in the absence of the mediator.

Because the difference between the effect estimates with and without mediators included is small

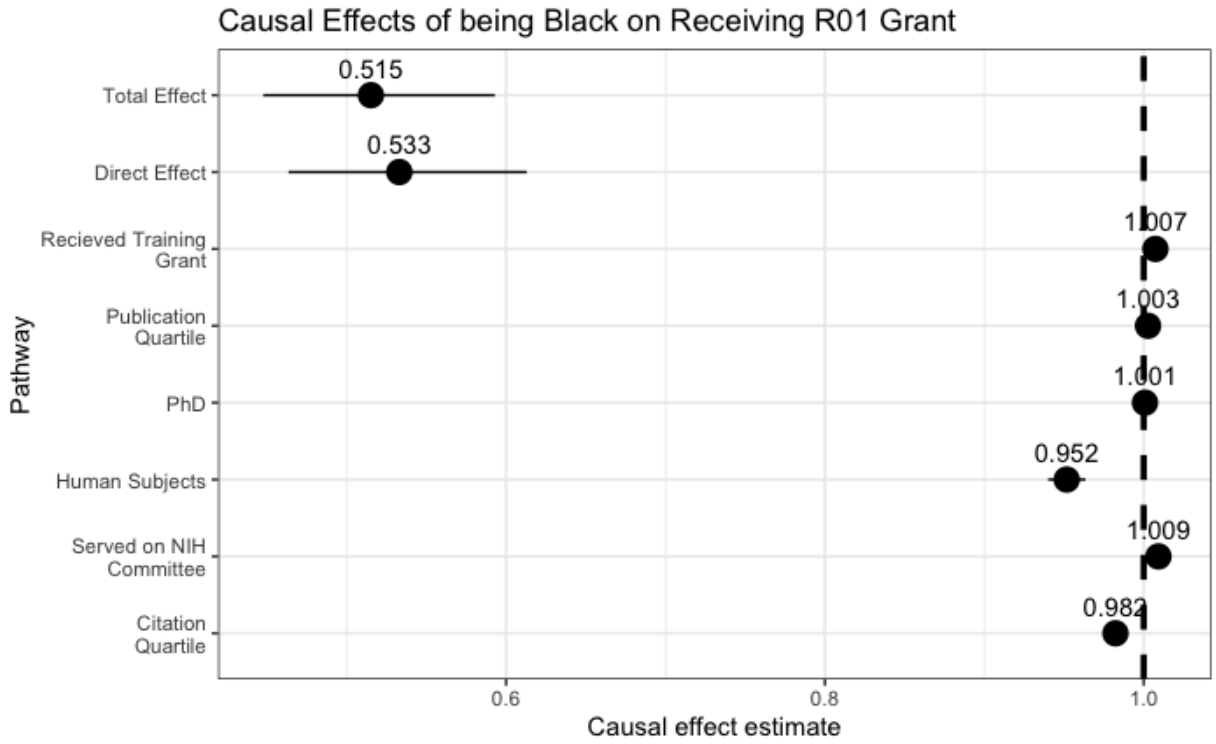

Figure 9: Causal Effect estimates from DAG. Encodes the total and direct effect of race on receiving an R01 grant as well as the mediating factors.

(although this statement has not been rigorously tested), we cannot conclude that the variables included in the dataset have an outsized role in mediating the effect of race upon R01 award probability, nor can they help explain the total effect. It appears that none of these causal pathways can explain this process whereby being Black lowers one's chance of receiving an R01 grant. The application is race- and name-blind, suggesting there are other mediators that are not included in this data and cannot be measured in this dataset. Some notable ones are gender and topic choice.

One can imagine that there is a different distribution of topics that applicants of different races apply to, for various reasons, and the NIH favors some topics over others. Similarly, gender may play a role, and the gender distributions underlying different racial groups applying to NIH grants can be different. One implication is that not all the variables in the dataset get causal estimates. This is because the design of our DAG means inclusion of these variables in our

statistical models does not and will not help get accurate causal estimates of the effect of race on R01 awards. Thus, despite not providing much intuition on the strength of effects along different pathways by which race has an effect, it does highlight which pathways are important due to our causal DAG.

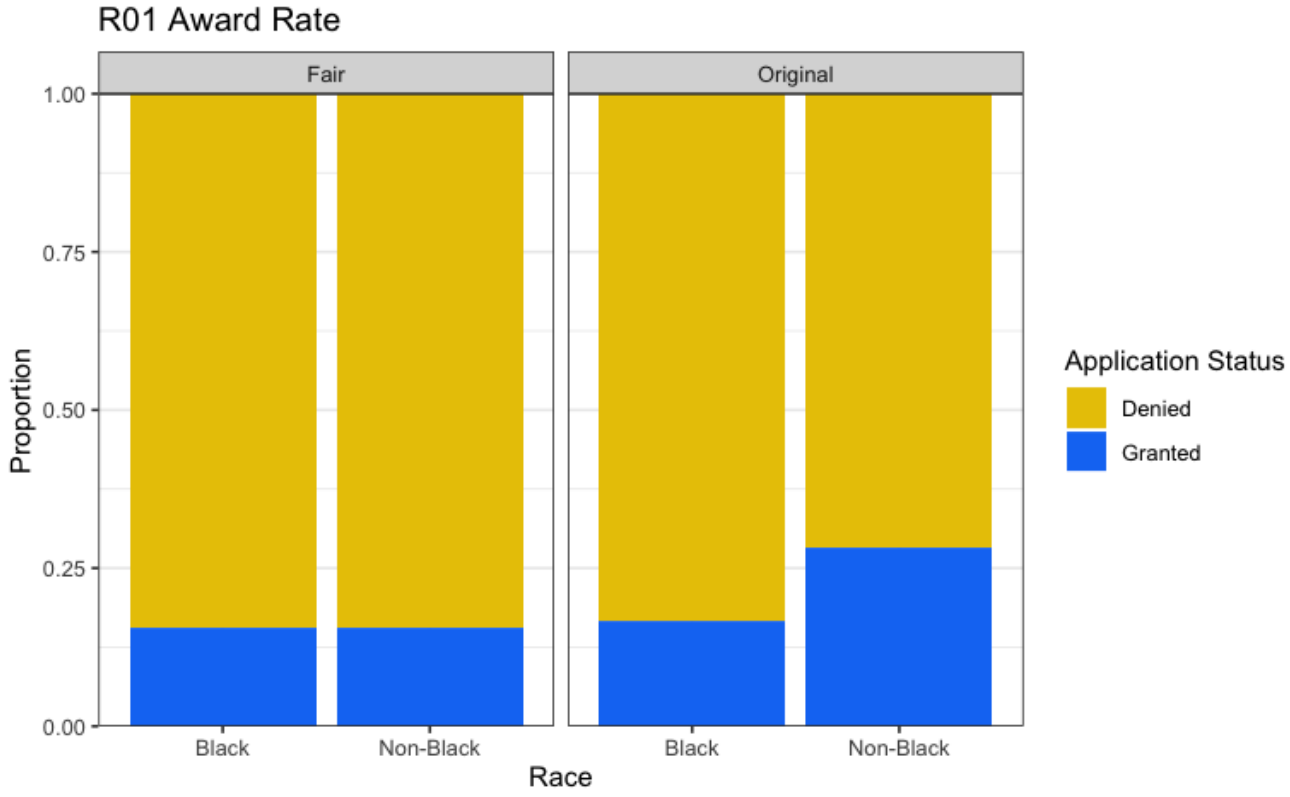

Figure 10: Award rate for Black and non-Black applicants under observed and fair regime

We were interested in seeing what happens if things were "fairer" within the current grant-selection paradigm—if the mediating variables remained the determinants of grant selection, but

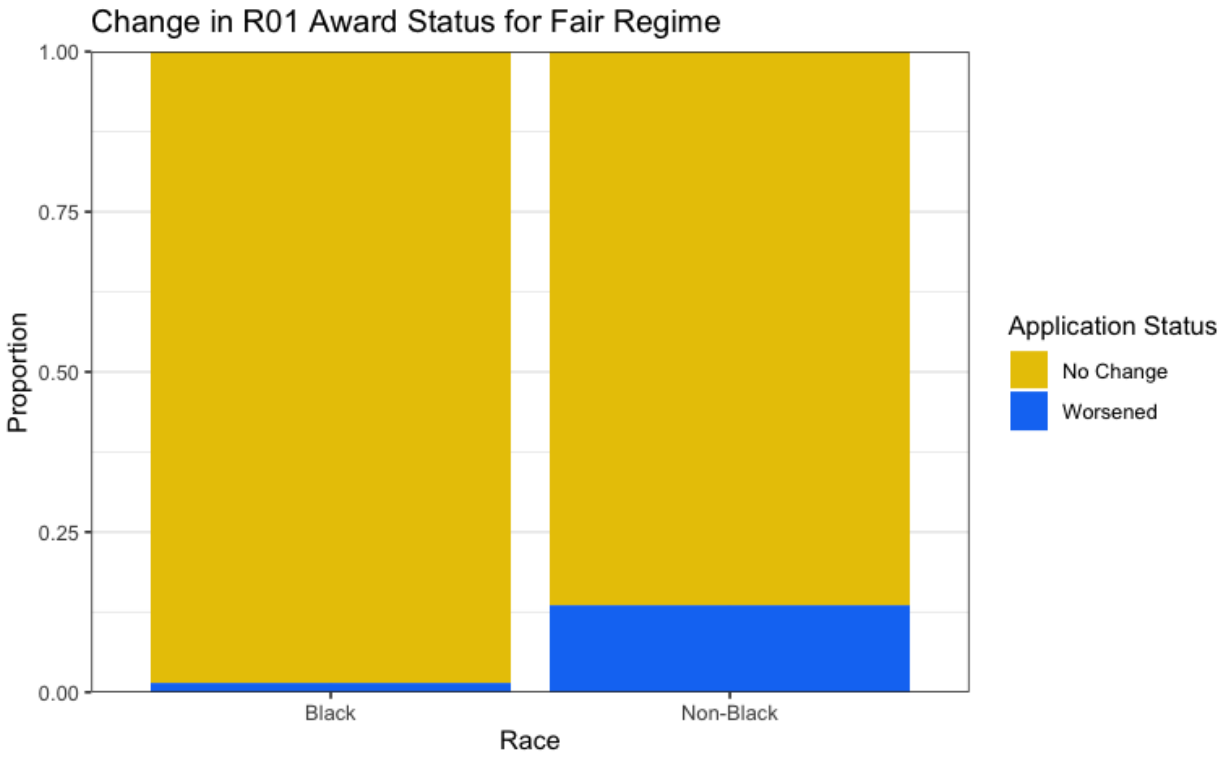

Figure 11: Change in award status in fair regime

without the effect of race. Using *cfairer*, we construct a fairer dataset by marginalizing out Race from the probability of getting the R01 award (Figures 10 – 11). This removes dependence on any other variables that are also descendants of Race. In other words, we changed the dataset to what it may look like if there were no effect of race on receiving the grant. Now we can look at how removing the relationship between race and grant awards changes the acceptance rates. The new data set shows that Black and non-Black applicants are awarded R01 grants at the same rate by way of rejecting additional non-Black applicants, with mostly no change to Black applicants. In other words, when we remove the relationship between race and receiving the R01 grant, there are not any accepted Black applicants that were not previously accepted, but there are rejected non-Black applicants who were not previously rejected.

Sensitivity Analysis

For robustness, we run the mediation analysis on other potential DAGs. The first DAG was the original output from *cfairer* unmanipulated. The structure is largely the same except for the orientation of some edges. The estimates for the total, direct, and mediator effects are largely the same (Figures 12–13).

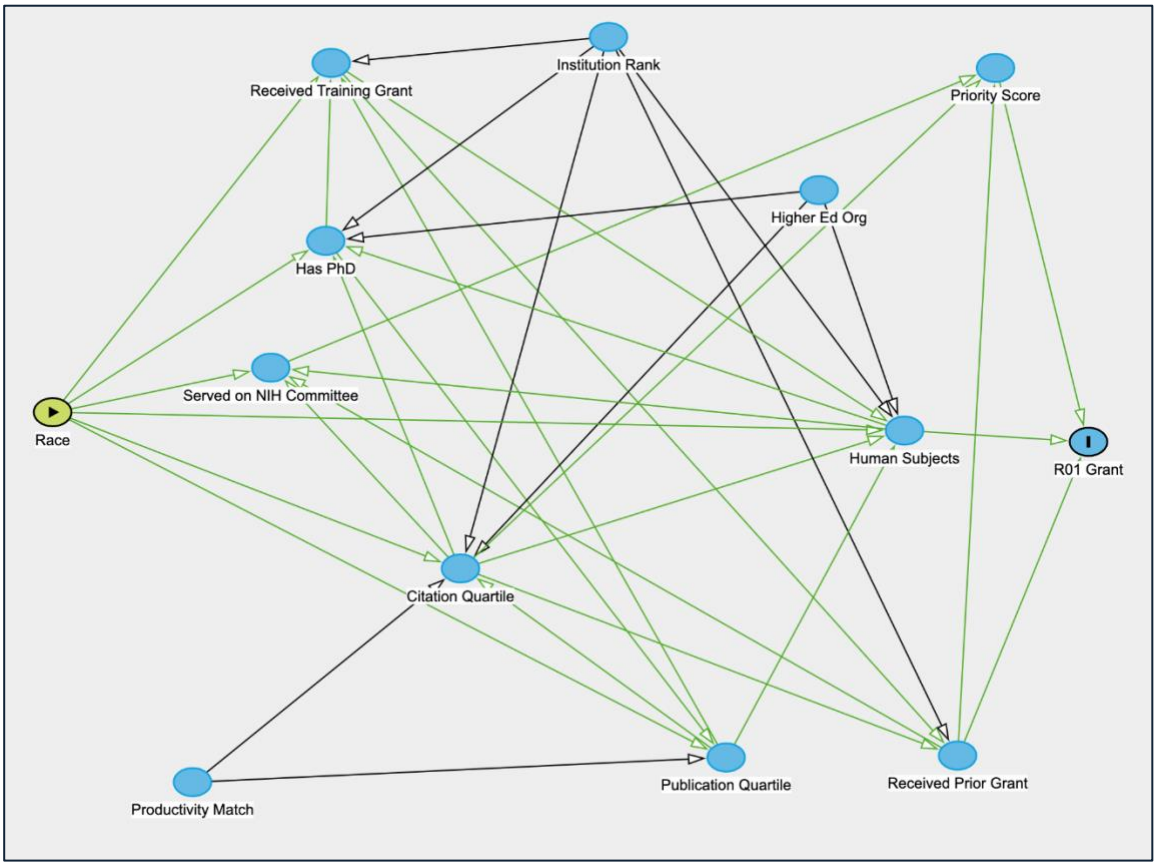

Figure 13: Alternative formulation of DAG from unmanipulated cfairer *output*

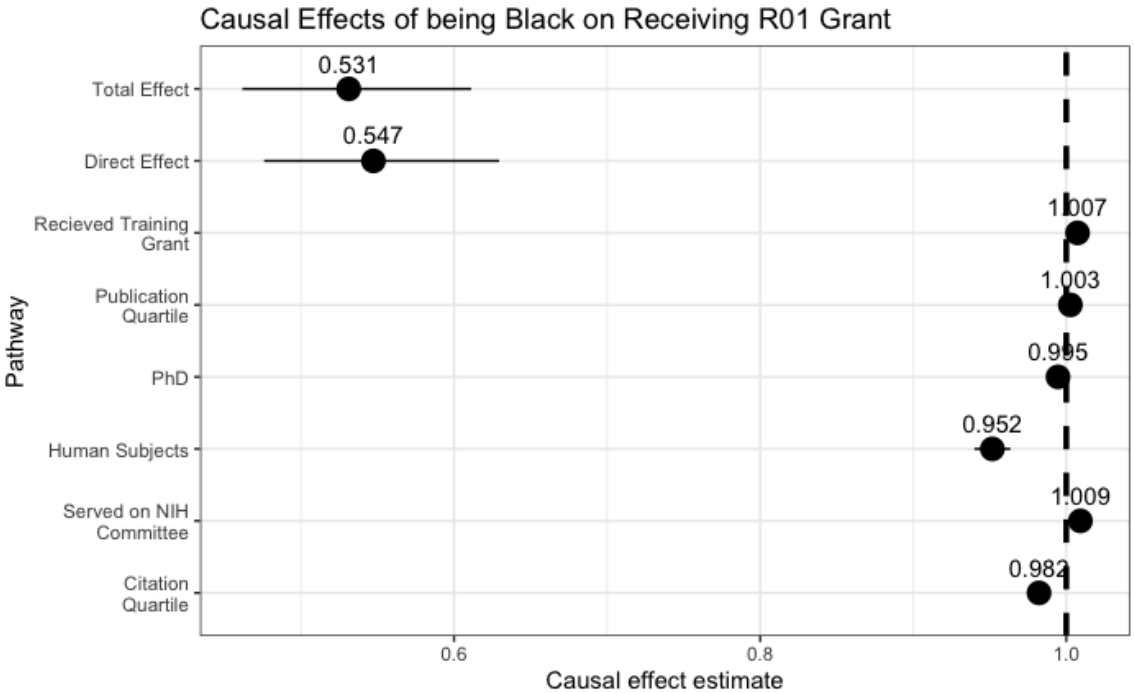

Figure 13: Causal estimates of alternative DAG

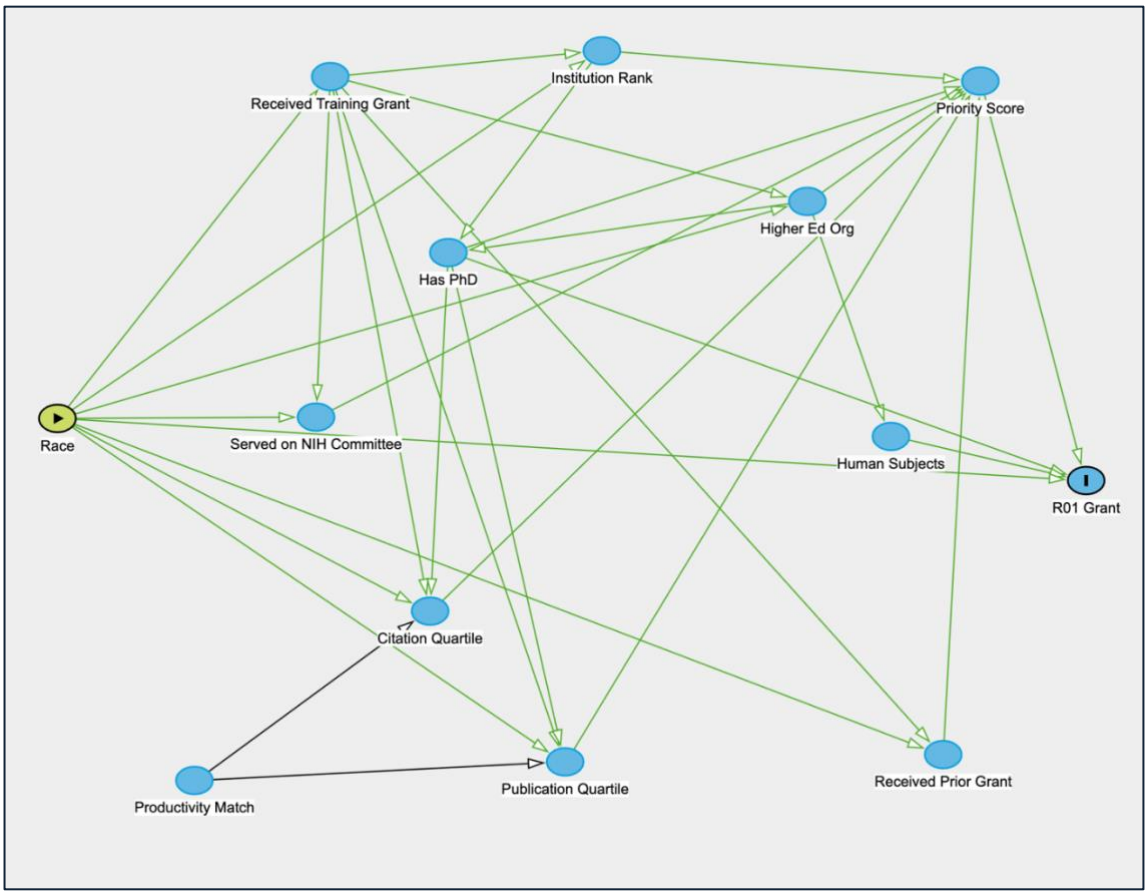

Figure 14: Second alternative DAG formulation

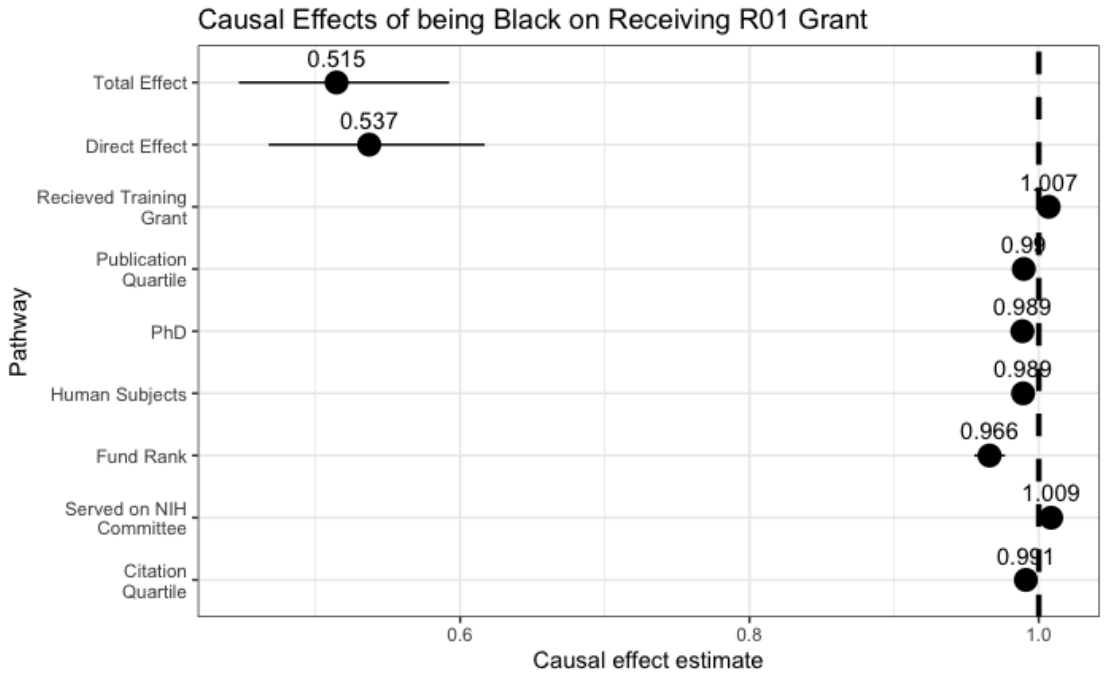

Figure 15: Causal estimates of second alternative DAG

The second alternative DAG that we tried was constructed from our ad-hoc evaluation of the data generation process and differs more from the DAG that the main results were reported from (Figures 14–15). In this DAG, the one-step mediators are now the institution rank, whether the applicant received a training grant, if they are a higher education organization applicant, if they served on an NIH committee, if they have prior grants, and their productivity. We add a direct edge from race to R01 award. However, the mediation analysis on this graph reveals much of the same results and is coherent with the other two graphs, indicating our results are not sensitive to choice in graph.

CONCLUSION

The original study by Ginther et al. found significant associations between race and R01 award probability, as well as significant associations with whether someone received a training grant, their productivity, their institution ranking and type, if they received a prior grant, and if they use human subject [12]. Using *cfairer*, we created a DAG to represent the causal relationships between these variables, with particular focus on the modeling of the mechanisms by which race affects R01 award probability.

We found that many of the significant covariates identified by the original paper were one-step mediators in this model. However, our mediation analysis with *cfairer* did not suggest that these variables may explain why race influences R01 award probability. We found the total effect of being Black on R01 award was 0.51 as opposed to being non-Black and that the direct effect was 0.54. The effects of the mediators were all roughly 1, indicating that they do not adequately explain why race is causing R01 award probability to change, suggesting there may be other mediators at play that are not recorded in this dataset. As a retrospective case study, we were limited to the variables in the dataset; in an active equity assessment, ideally the involved parties would iterate on DAGs based on lived experience and external domain expertise, and collect additional data to interrogate the different hypothesized mediators [24].

When looking at a “fairer” dataset—keeping the outcome dependencies on the mediators but averaging out the effect from race—additional Black applicants are not accepted but rather additional non-Black applicants are rejected.

Ultimately, using causal modeling can reveal the underlying mechanisms of a system of variables beyond surface-level associations and, in this case study, identify when a racially disparate outcome is being caused through pathways other than the ones noted as potential explanations for the disparity.

DATA AND CODE AVAILABILITY

cfairer is an open-source package available on GitHub: https://github.com/cfairer/. The de-identified dataset can be found at https://report.nih.gov/nih-supported/invstigators-and-trainees.

CONTRIBUTION STATEMENT

Saimun Habib and Joshua Stadlan conceived of the case study. Mary Munro developed the causal inference tool, provided technical guidance, and wrote the Method section. Saimun Habib performed the data preparation, executed the analysis, and wrote the Data Characterization, Results, and Conclusion sections. Joshua Stadlan provided the social equity assessment framing and wrote the Introduction. Lilly Boyer and Tamey Habtu contributed background research to the case study. All authors reviewed the final manuscript.

ACKNOWLEDGMENTS

We thank Allen Ross for guiding the *cfairer* project and providing initial ideas for this case study, Ant Ngo for his meticulous work developing the tool, Dr. Erica Taylor for her expert input on the equity assessment framing, and Dr. Hannah De los Santos for her careful technical review and help with finalizing the report. We thank Dr. Arya Farahi at University of Texas at Austin, who, through our collaboration with Good Systems: A UT Grand Challenge, provided feedback on statistical elements of the report. We also thank Jenine Patterson and Dr. Laura Leets for supporting the project via MITRE's Social Justice Platform and MITRE's Independent Research and Development Program, respectively. Any remaining errors are our own.

This research was funded by MITRE's Independent Research and Development Program and MITRE's Social Justice Platform. The authors declare no conflicts of interest.

REFERENCES

[1] T. C. Freeman, J. Patterson and K. Jiang, "A Framework for Assessing Equity in Federal Programs and Policies: Advancing Racial Equity and Support for Underserved Communities Through the Federal Government (EO 13985).," McLean, VA, 2021.

[2] E. J. Kennedy, "Can data drive racial equity?," *MIT Sloan Management Review,* vol. 62, no. 2, 2021.

[3] J. Patterson, J. Z. Stadlan, R. Bergstein, T. Sinha, S. Habib, A. Johnson and R. Josephs, "The Racial Wealth Gap in Washington, D.C.: Research and Analysis in Support of the Council Office of Racial Equity, Council of the District of Columbia," McLean, VA, 2021.

[4] K. Cox and K. Edwards, "Black Americans have a clear vision for reducing racism but little hope it will happen," 2022.

[5] R. Richardson, J. M. Schultz and K. Crawford, "Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, And Justice," *New York University Law Review,* 2019.

[6] L. Arellano, "Questioning the Science: How Quantitative Methodologies Perpetuate Inequity in Higher Education," *Education Sciences,* vol. 12, no. 2, 2022.

[7] N. Krieger, "Structural Racism, Health Inequities, and the Two-Edged Sword of Data: Structural Problems Require Structural Solutions," *Frontiers in Public Health,* vol. 9, 2021.

[8] P. B. Adkins-Jackson, T. Chantarat, Z. D. Bailey and N. A. Ponce, "Measuring Structural Racism: A Guide for Epidemiologists and Other Health Researchers," *American Journal of Epidemiology,* vol. 191, no. 4, 2022.

[9] Supreme Court of the United States, "Griggs v. Duke Power Co., 401 US 424," 1971.

[10] Office of the Assistant Secretary for Fair Housing and Equal Opportunity, "Reinstatement of HUD's Discriminatory Effects Standard: Proposed rule.," Washington, DC, 2021.

[11] NIH National Institute of Allergy and Infections Diseases, "Comparing Popular Research Project Grants – R01, R03, and R21," 6 2021.

[12] D. K. Ginther, W. T. Schaffer, J. Schnell, B. Masimore, F. Liu, L. L. Haak and R. Kington, "Race, ethnicity, and NIH research awards," *Science,* vol. 333, no. 6045, 2011.

[13] National Institutes of Health, "Racial Disparities in NIH Funding," *National Institutes of Health,* 8 2022.

[14] M. Lauer, K. Patel and D. Roychowdhury, "RPG and R01-Equivalent Funding and Success Rates by Race-Ethnicity FY2010-FY2021".

[15] Working Group on Diversity in the Biomedical Research Workforce (WGDBRW) and The Advisory Committee to the Director (ACD), "Draft Report of the Advisory Committee to the Director Working Group on Diversity in the Biomedical Research Workforce," 2012.

[16] J. Pearl, "An Introduction to Causal Inference," *The International Journal of Biostatistics,* vol. 6, no. 2, 2010.

[17] S. Chiappa and T. P. S. Gillam, "Path-Specific Counterfactual Fairness," *arXiv,* 2018.

[18] J. R. Loftus, C. Russell, M. J. Kusner and R. Silva, "Causal Reasoning for Algorithmic Fairness," *arXiv,* 2018.

[19] C. Heinze-Deml, M. H. Maathuis and N. Meinshausen, Causal Structure Learning, vol. 5, Annual Review of Statistics and Its Application, 2018, pp. 371-391.

[20] M. Scutari, "Learning Bayesian Networks with the bnlearn R Package," *Journal of Statistical Software,* 2010.

[21] M. Scutari, "Who Learns Better Bayesian Network Structures: Accuracy and Speed of Structure Learning Algorithms," *Proceedings of Machine Learning Research,* pp. 416-427, 2018.

[22] J. Steen, "medflex: An R Package for Flexible Mediation Analysis using Natural Effect Models," 2017.

[23] J. Textor, "Robust causal inference using directed acyclic graphs: the R package 'dagitty'," *International Journal of Epidemiology,* vol. 45, no. 6, pp. 1887-1894, 2016.

[24] C. Fox, S. Gellen, S. Morris, J. Ozan and J. Crockford, "Impact Evaluation With Small Cohorts: Methodology Guidance," 2022.

APPENDIX

Table A2: Variables included in de-identified dataset used in the Ginther et al. (2011) paper.

Column Name	Data Type	Notes	Used
r01awd	Boolean	Whether the applicant was granted the R01 award	Yes
org_high	Boolean	Whether the applicant was from a higher education organization	Yes
hs_y	Boolean	Whether the proposal uses human subjects	Yes
prior_grant	Boolean	Whether the applicant received a prior grant	Yes
cmte_c	Boolean	Whether the applicant has served on an NIH committee	Yes
pub_badmatch	Boolean	If the applicant was matched to publication history with high probability	Yes
Phdsamp	Boolean	Whether the applicant has a PhD	Yes
Scored	Boolean	Whether the applicant had a high-priority subject proposal	Yes
Roleftk	Boolean	Whether the applicant received a prior training grant	Yes
Race	Categorical	Applicant Race	Yes
Fund_rank	Categorical	Quartile of applicant's home institution	Yes
Pubq	Categorical	Quartile of applicant's number of publications	Yes
Citq	Categorical	Quartile of applicant's number of citations	Yes
R01	Boolean	Whether the applicant was applying to the R01 grant	No
Rpgawd	Boolean	Whether the applicant was awarded a research grant for the proposal (includes non-R01 grants)	No
R01_type1	Boolean	Whether this would be the first time the applicant receives an R01 grant	No