Detecting clinician implicit biases in diagnoses using proximal causal inference

Kara Liu†, Russ Altman, Vasilis Syrgkanis

Computer Science Department, Stanford University,
Stanford, CA 94305, USA
†*E-mail: karaliu@stanford.edu*

Clinical decisions to treat and diagnose patients are affected by implicit biases formed by racism, ableism, sexism, and other stereotypes. These biases reflect broader systemic discrimination in healthcare and risk marginalizing already disadvantaged groups. Existing methods for measuring implicit biases require controlled randomized testing and only capture individual attitudes rather than outcomes. However, the "big-data" revolution has led to the availability of large observational medical datasets, like EHRs and biobanks, that provide the opportunity to investigate discrepancies in patient health outcomes. In this work, we propose a causal inference approach to detect the effect of clinician implicit biases on patient outcomes in large-scale medical data. Specifically, our method uses proximal mediation to disentangle pathway-specific effects of a patient's sociodemographic attribute on a clinician's diagnosis decision. We test our method on real-world data from the UK Biobank. Our work can serve as a tool that initiates conversation and brings awareness to unequal health outcomes caused by implicit biases.*

Keywords: Implicit bias, proximal causal inference, fairness, healthcare

1. Introduction

Implicit bias refers to unconscious and automatic associations that affect how we perceive, evaluate, and interact with people from different social groups.[1] Outside of mere cognitive distortions, these biases held by healthcare professionals influence clinical decisions and alter a patient's quality of care. Implicit biases have been shown to be both harmful and pervasive in modern-day medicine, exacerbating existing inequality in the treatment and health outcomes of marginalized groups.[2,3] For instance, unconscious attitudes held by clinicians result in disparate outcomes where women are less likely then men to be diagnosed with myocardial infarction,[3] Black women in the UK and US experience higher maternal mortality than White women,[4] and low socioeconomic (SES) and non-White patients receive sub-optimal pain management treatment compared to high SES and White patients.[5,6]

The recent integration of machine learning (ML) models into clinical decision-making has highlighted the prevalence of biases in medicine. By replicating the patterns from real-world medical data, ML models perpetuate and risk amplifying existing disparities in the medical treatment of marginalized groups.[7,8] While much attention has been given to the statistical

*Our method is available at `https://github.com/syrgkanislab/hidden_mediators`

objectives of fairness and the development of fair models, there has been comparatively less focus on investigating the biases present in the underlying data. A method capable of detecting implicit clinician bias in observational datasets would prevent ML models from unintentionally perpetuating biased decisions.

However, measuring implicit bias is challenging. Existing methods for quantifying implicit bias rely on the Implicit Association Test (IAT)[9] and randomized psychological experiments like affective priming.[10] While these tests are useful for initiating dialogue, they only provide a snapshot of individual clinician attitudes and do not guarantee a causal link to behavior or larger systemic discrepancies of care.[10]

In this work, we propose a computational tool to detect clinician implicit bias in observational datasets by measuring the causal effect of patient attributes, like race, SES, and other social determinants of health (SDoH), on medical diagnoses. By decomposing the causal effect into two pathways, we can separate the *biological effect* (the influence of a demographic attribute on diagnosis as mediated by valid biological traits) from the *implicit bias effect* (how the patient's attribute affects a clinician's judgement independent of their actual health state). As it is unlikely to observe a patient's true health state, we use observed medical data as proxies using proximal causal inference.[11] To estimate the effect of implicit bias, we propose a novel proximal mediation method that guarantees identifiability under several assumptions. Using real patient data from the UK Biobank, we validate our method can robustly detect several clinician implicit biases identified from prior works. We aim for the proposed method to serve as a bias-detection tool in dataset audits and initiate discussion on reducing systemic discrimination in medicine.

Disclaimer: While we use the UK Biobank data for method validation, we emphasize that this work is not a commentary on specific examples of discrimination within the UK healthcare system. Additionally, it is crucial to clarify that our method of estimating implicit bias is not intended to target clinicians but rather reflect on clinician behaviors within the context of discriminatory healthcare systems.

2. Method

2.1. *Background*

2.1.1. *Overview*

According to the Hippocratic Oath, clinicians should base their diagnostic decisions on each patient's history and current health status, unaffected by biases or stereotypes of the perceived patient identity. However, even in the ideal scenario of unbiased treatment, patient sociodemographic attributes will still influence diagnosis. Attributes including race, sex, or SES have been shown to influence a patient's true health status via mechanisms like genetics, lifestyle, and weathering from systemic oppression.[12–14] These biologically-mediated effects increase the risk of certain medical conditions. For instance, patients from lower SES backgrounds experience higher levels of stress and reduced access to healthcare, increasing their risk of cardiovascular disease.[15] In light of these known biological influences, the causal effect of a patient's sociodemographic attribute on their diagnosis by a clinician is therefore comprised of two pathway effects: the *biological effect* and the *implicit bias effect*, the latter referring to

the clinician's subjective biases of the sociodemographic attribute not mediated through the patient's actual health state.

We present the assumed causal relationships between variables as the directed acyclic graph (DAG) in Figure 1. Dashed arrows denote optional edges, and bi-directional arrows denote indirect confounding paths through latent variables. Let D be the binary sociodemographic attribute and Y the diagnosis decision we wish to measure implicit bias with respect to. M represents the latent variables encoding a patient's true underlying health state. However, as M is typically unknown, we instead observe Z and X as multivariate proxies of M. We differentiate these proxies into the variables Z that do not affect the diagnostic decision Y but could be affected by the attribute D; and the variables X which are not directly affected by the attribute D but can influence diagnosis Y. For example, X could be recent lab reports a clinician uses to make their diagnosis, and Z could be a patient's survey responses to a sleep questionnaire (assuming the survey does not influence the clinician's diagnosis). Finally, let W be sociodemographic confounders to control for.

We can now reframe *biological* and *implicit bias effects* using pathway causal effects. The *biological effect* of attribute D on diagnosis Y is the indirect effect as mediated through the true underlying health state M: $D \to M \to Y$. The *implicit bias effect* we wish to measure is the direct effect of $D \to Y$ that flows through the edge θ and is defined as the residual of the biological effect. We formally define bias in terms of controlled direct effects in Equation (D.1).

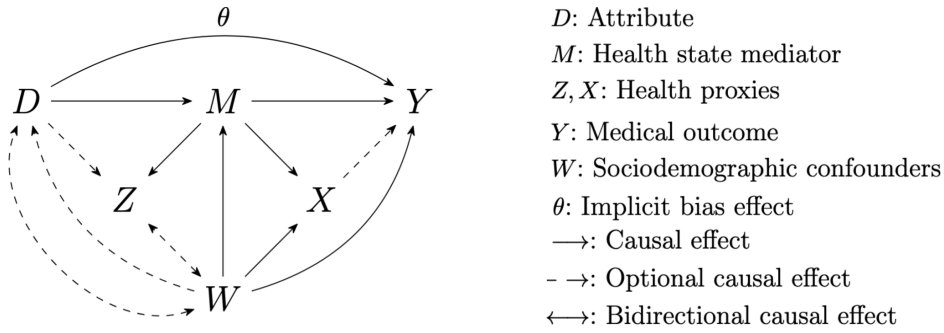

Fig. 1: Assumed causal graph.

2.1.2. *Related works*

Measuring implicit biases requires detecting the unconscious and automatic attitudes that shape behavior. The predominant method for implicit bias measurement thus far has been the Implicit Association Test (IAT),[9] a questionnaire developed in 1998 intended to measure group association through word categorization. To capture clinician biases, several works have linked clinician attitudes via their IAT score to behavioral manifestation.[10,16] Other methods for detecting implicit clinician bias include affective priming, which measures biased associations after stimulus priming; and the assumption method, which surveys clinicians' decisions after reading patient vignettes.[17] While association tests like the IAT have been integral in bringing awareness to medical biases, they are criticized for their arbitrary scoring system, inability to predict real-world patient outcomes, and context-dependency.[3,10,16,17] Furthermore,

administering these controlled tests in every clinical encounter is impractical and unscalable.

Computational methods present a promising and scalable alternative for detecting implicit bias in real-world medical data. While the field of ML fairness has explored bias detection, the focus has been on identifying and mitigating bias in models rather than the data.[7] In causal inference, disentangling a causal effect into natural indirect and direct pathway effects has led to methods that control for "fair" and "unfair" causal pathway effects. [18–21] propose metrics for measuring fair pathway influence on outcomes and develop methods that mitigate the effect of unfair pathways on the predicted outcome. [22] leveraged the Fairness-Aware Causal paThs (FACTS)[23] algorithm to quantify disparate pathway influence of SDoH attributes on mortality using real-world health data. While these methods recognize an attribute's influence on an outcome contains both fair and unfair effects, prior works are limited to simple scenarios where all variables are known and observed. Our work is the first to extend pathway inference to large-scale observational data with potentially unobserved variables.

Finally, a few recent methods have explored proximal mediation analysis, where pathway effects can be measured despite unobserved mediators by using proxy variables.[24,25] However, by relying on natural direct and indirect pathway effects, these works rely on more stringent assumptions, require learning complicated bridge functions, and limit their analysis to simple datasets. In comparison, our method makes several relaxations that enable application to observational data. First, we identify controlled instead of natural effects, which presents an equally good measurement of a biased decision yet lends to a much simpler statistical problem. Additionally, we assume partially linear equations instead of requiring the identification of a complex bridge function. Finally, we do not require uniqueness of the parameters unrelated to implicit bias (i.e., the nuisance parameters for the outcome bridge function). These relaxations enable our approach to be effective at analyzing large-scale real-world medical data.

2.2. *Our method*

Our goal is to identify and estimate the following controlled direct effect:

$$\theta = \int_{m,w} \mathbb{E}[Y(1,m) - Y(0,m) \mid W = w]\ p(m,w)\ dm\ dw \tag{1}$$

where $Y(d,m)$ is the potential (or counterfactual) outcome when we intervene on the attribute D and the mediator M and set them to values (d,m); and $p(m,w)$ is the natural probability distribution in the data. If the controlled direct effect is nonzero, then there exists a direct influence of the attribute D on the outcome Y, which is evidence of implicit bias.

If we observe M, the above controlled direct effect can be identified by a simple g-formula that "controls" for M and W: $\theta = \mathbb{E}[\mathbb{E}[Y \mid D = 1, M, W] - \mathbb{E}[Y \mid D = 0, M, W]]$. Unfortunately, this equation is intractable if M is unobserved. However, we show that under a few reasonable assumptions the controlled direct effect is still identifiable.

Theorem 1 (Identification). [a] *Consider a non-parametric structural causal model (SCM) that respects the causal relationships encoded in Figure 1 (see Appendix D.1) and assume there exists a "bridge function" q that solves $\mathbb{E}[Y \mid D, M, W] = \mathbb{E}[q(D, X, W) \mid D, M, W]$. Then*

[a]We present more intuitive interpretations of each theorem and lemma in the Appendix.

q also solves the Non-Parametric Instrumental Variable (NPIV) problem defined by the set of conditional moment restrictions

$$\mathbb{E}[Y - q(D, X, W) \mid D, Z, W] = 0 \tag{2}$$

and the controlled direct effect can be identified as $\theta = \mathbb{E}[q(1, X, W) - q(0, X, W)]$.

Identifying parameters θ using a bridge function q (where q also solves an NPIV problem) has been extensively studied in proximal causal inference literature.[26–38] However, these approaches rely on solving saddle-point problems with adversarial training or require learning conditional density functions, both of which are statistically daunting.

We can avoid these difficult statistical tasks if we assume that the bridge function is partially linear in D and X. The following lemma shows that partial linearity of q is implied by a more primitive assumption of partial linearity of two other functions (proof in Appendix D.4).

Lemma 1 (Identification under partial linearity). *Consider a non-parametric SCM that respects the constraints encoded in Figure 1 and assume that X has dimension p_X at least as large as the dimension p_M of M. Moreover, assume that the following functions are partially linear:*

$$\mathbb{E}[Y \mid D, M, X, W] = D\,c + M^T b + X^T g + f_Y(W) \tag{3}$$

$$\mathbb{E}[X \mid M, W] = F\,M + f_X(W) \tag{4}$$

where F is a $p_X \times p_M$ matrix, b is a p_M-dimensional vector, g is a p_X-dimensional vector and f_Y, f_X are arbitrary non-parametric functions. If we assume the matrix F has full column rank, then there exists a partially linear outcome bridge function

$$q(D, X, W) = D\,\theta + X^T h + f(W) \tag{5}$$

that satisfies Equation (2), where parameter $h = F^{+}b + g$[b] and $\theta = c$.

Under the assumption of partial linearity, we can simplify the estimation problem by first removing the effect of W from all the remaining variables (see Appendix D.5), where for any variable V we define the residual $\tilde{V} = V - \mathbb{E}[V \mid W]$. Partial linearity of q from Equation (5), when combined with the NPIV Equation (2), implies that θ can be identified using linear instrumental variable (IV) regression where $(\tilde{Z}; \tilde{D})$[c] are the instruments and $(\tilde{X}; \tilde{D})$ are the treatments:

$$\mathbb{E}\left[(\tilde{Y} - \tilde{X}^T h - \tilde{D}\,\theta)\begin{pmatrix}\tilde{Z}\\ \tilde{D}\end{pmatrix}\right] = 0 \qquad \text{(Primal Equation)}$$

Unique identification of θ seemingly requires unique identification of the other "nuisance" parameters like h, which might be difficult to achieve as the covariance matrix $\mathbb{E}[(\tilde{X}; \tilde{D})\,(\tilde{Z}; \tilde{D})^T]$ is usually not full rank[d]. We invoke and simplify ideas from the recent proximal inference

[b] F^{+} is the Moore-Penrose pseudoinverse of F.
[c] We denote $(A; B)$ to be concatenation of vectors A and B.
[d] This could be the case if the number of proxies is much larger than the dimensionality of the latent mediator M.

literature[35,39] to show that θ can be point-identified even if h is not. To achieve this, we construct a moment restriction equation that is Neyman orthogonal to the nuisance parameters h but still point-identifies θ, given sufficient quality of the proxy Z. Intuitively, we learn a new instrument $V = (\tilde{D} - \gamma^\top \tilde{Z})$ such that V is uncorrelated with $\tilde{X}$ and thus estimation of θ is not sensitive to h. Existence of such a γ is sufficient for point-identification of θ. We provide the proof for the point identification of θ in Appendix D.7 and for Neyman orthogonality in D.8.

Theorem 2. *Let h_* be the minimum norm solution to the* (Primal Equation) *and assume that the following dual equation also admits a solution γ_*:*

$$\mathbb{E}[\tilde{X}\,(\tilde{D} - \gamma^T \tilde{Z})] = 0 \qquad \text{(Dual Equation)}$$

Furthermore, assume $\mathbb{E}[\tilde{D}\,(\tilde{D} - \gamma_*^T Z)] \neq 0$. Then the solution θ_* to the equation:

$$\mathbb{E}[(\tilde{Y} - \tilde{X}^T h_* - \tilde{D}\,\theta)\,(\tilde{D} - \gamma_*^T \tilde{Z})] = 0 \tag{6}$$

uniquely identifies the controlled direct effect θ. Furthermore, this moment restriction is Neyman orthogonal with respect to nuisance parameters γ_*, h_*.

Theorem 2.2 allows us to invoke the general framework of [40] to construct an estimate and confidence interval for the controlled direct effect θ. The full estimation algorithm is presented in Appendix D.9.

2.3. *Testing and Removing Weak Instruments*

Our method for uniquely identifying the controlled direct effect θ relies on several assumptions, e.g., $(\tilde{Z}; \tilde{D})$ are good instruments for $(\tilde{X}; \tilde{D})$. To assess the validity of these assumptions, we developed a suite of tests that must pass for the estimate θ to be valid and can be used as validity checks by practitioners. These tests are further described in Appendix C:

(1) *Primal equation violation* - We develop a χ^2-test to check if the primal equation admits a solution, i.e., $\mathbb{E}[(\tilde{Y} - \tilde{X}^T h_* - \tilde{D}\,\theta_*)(\tilde{Z}; \tilde{D})] \approx 0$. Intuitively, violation of the primal test implies either the variables X are insufficient proxies of the health state M or the residual proxy $\tilde{Z}$ has a direct path to $\tilde{Y}$.

(2) *Dual equation violation* - We develop a χ^2-test to check if the dual equation admits a solution, i.e., $\mathbb{E}[\tilde{X}(\tilde{D} - \gamma_*^T \tilde{Z})] \approx 0$. Violation of the dual implies the variables Z are insufficient proxies of the health state M or that the residual proxy $\tilde{X}$ has a direct path from $\tilde{D}$.

(3) *Strength of identification* - We perform two tests to check if $V = (\tilde{D} - \gamma^T \tilde{Z})$ is a good instrument for (i.e., retains enough information about) $\tilde{D}$. (a) We develop an effective F-test[41,42] to check the correlation strength of V with $\tilde{D}$. (b) We develop a z-test to check if the quantity $\mathbb{E}[\tilde{D}(\tilde{D} - \gamma_*^T \tilde{Z})]$ is substantially bounded away from zero (see assumption in Theorem 2.2). Intuitively, these tests will fail if the hidden mediator is a very deterministic function of the attribute D.

(4) *Proxy covariance rank test* - To ensure the health proxies are sufficiently related, we check the rank of the covariance matrix of $\tilde{X}$ and $\tilde{Z}$ by identifying the number of statistically significant singular values. This rank can be viewed as an upper bound on the dimensionality of the hidden mediator M that we can control for.

2.3.1. *Proxy selection algorithm*

In practice, the initial selection of proxies X, Z may violate key assumptions, which can be detected by the failure of one or more of the aforementioned tests. In Appendix B, we provide an algorithm for identifying subsets of X and Z that satisfy the necessary assumptions and thus produce valid estimates. This proxy selection algorithm should be performed on a separate dataset from the one used to estimate θ.

3. Experiments

3.1. *Data*

To validate our approach, we use the UK Biobank, a rich and accessible repository containing genomic, imaging, and tabular health data from over 500,000 patients. Our work uses its tabular data, which includes survey questions and biometrics collected upon an individual's enrollment into the biobank. In addition, several health outcomes, including medical diagnoses via ICD10 codes, have been linked to most patients. We note and discuss the caveats of applying our method to biobank data in Section 5.2.

		Prevalence in UK Biobank (n=502411)	Prior works on implicit bias
Sociodemographic attribute D	*Race* - Asian	2.4%	43–45
	Race - Black	1.8%	3–5,46,47
	Gender - Female	54.4%	3,48
	Disability status - On disability allowance	6.2%	49,50
	Income - Household income <18,000£	20.3%	5,14,15,51
	Education - No post-secondary education	67.3%	5,51
	Weight - BMI >30	24.3%	52,53
	Insurance - Not on private insurance	31.4%	54
Medical diagnosis Y	Osteoarthritis	18.0%	47,49
	Rheumatoid arthritis	1.9%	55
	Chronic kidney disease	5.0%	56,57
	Complications during labor	2.4%	3,4
	Heart disease	10.7%	3,15,48
	Depression	6.0%	46,58
	Melanoma	1.2%	59,60

Table 1: Selected sociodemographic attributes D and diagnoses Y

Prior works have proposed sociodemographic attributes that might bias clinical decisions. For example, [48] showed that clinicians exhibited greater uncertainty when diagnosing coronary heart disease in women compared to men. We list in Table 1 most of the attributes D and diagnoses Y we test for implicit bias, and present the full list of the 102 (D, Y) pairs in Appendix E.2. To highlight the influence of clinician subjectivity, we concentrate on diagnoses that require clinician interpretation of patient-reported symptoms, e.g., chronic pain.

Selecting health proxies for Z and X relies user intuition and medical expertise to determine which variables have a direct relationship with attribute D and outcome Y, respectively. In general, proxies X could be observed by the clinician during their diagnostic decision, and proxies Z are not accessible during diagnosis but might have a direct causal relationship with attribute D. In the UK Biobank, we select X to be the biometric variables collected by the biobank at patient enrollment, which includes lab results and blood pressure readings. For Z, we use survey responses of self-reported pain levels, mental health, and sleep. We list all variables, including the sociodemographic confounders W, in Appendix E.1. Note our data

contains a mix of binary, integer, and continuous variable types.

3.2. *Evaluation metrics*

3.2.1. *Semi-synthetic data validation*

We test if our method can retrieve a known implicit bias effect using semi-synthetic data. We use real data from the UK Biobank for attribute D, confounders W, and health proxies X,Z. We develop a model that computes M and a synthetic diagnosis Y with a known implicit bias effect $\theta = 0.5$ using linear structural equations. We test against fully continuous (Experiment 1) and both binary and continuous (Experiment 2) semi-synthetic data, the latter being more realistic in real-world medical data. Our semi-synthetic data generation method is fully described in Appendix A. As a baseline, we compare two variants of ordinary least squares (OLS): (a) given we know M, we fit an OLS model over W, D, X, and M to predict Y; (b) in the more realistic scenario where M isn't known, we learn over W, D, X, and Z. We compute the average effect estimate and confidence interval based on $\pm 1.96\ \sigma$ where the average and standard deviation σ is taken over K=100 iterations.

3.2.2. *Calculating the implicit bias effect in the UK Biobank*

We next run our method on the full UK Biobank data. We compute the residuals of Z, X, Y, D fitted on W using Lasso regression. For all models, the regularization term is chosen via semi-cross fitting[61,62] over 3 splits. We fit all models using the `scikit-learn` Python package. For nuisance parameters h_* and γ_* we used regularized adversarial IV estimation[35,63] with linear functions and a theoretically driven penalty choice that decays faster than the root of the number of samples.

In cases where the data may not meet the method's assumptions, we developed a proxy selection algorithm (see Section 2.3.1) that identifies an optimal subset of X, Z proxies for each (D, Y) pair using the assumption tests from Section 2.3. Although we recommend separate data splits for proxy selection and effect estimation, we use the same dataset as our intent is method demonstration rather than robust effect estimates. Details of the hyperparameters used for the selection algorithm are provided in Appendix B.

For each of the 102 pairs of attribute D and diagnosis Y, we report seven metrics: the implicit bias effect θ, the 95% confidence interval, as well as our five proposed tests from Section 2.3: (1) the primal and (2) dual violation, (3-4) the strength of identification, and (5) the $\tilde{Z}, \tilde{X}$ covariance rank test. In addition, we also run the following five analyses:

Weak identification confidence interval - If the instrument identification tests from 2.3 are violated, then effect estimation can be unstable and normality-based confidence intervals inaccurate. We thus compute an alternative confidence interval[62] developed under the assumption of weak instruments (see Appendix C.5 for the description).

Bootstrapping analyses - We perform several bootstrapping analyses to test the sensitivity of the estimate. In the first analysis, given the computational complexity of recomputing the full estimate, we compare K=10 bootstrapped iterations re-estimating the full pipeline (stage 1); K=100 iterations using the pre-computed residuals but re-estimating all other parameters

(stage 2); and K=1000 iterations re-computing only the final Equation (6) (stage 3). Each iteration samples 50% of the data without replacement. In the second analysis, we compare sampling $10\%, 25\%, 50\%$ or 75% of the original data for K=10 bootstrapped iterations, re-estimating over full pipeline (stage 1). Finally, we compare different sample sizes for $K = 1000$ iterations re-estimating from stage 3 of the pipeline.

Influence points - Inspired by [64], we analyze influence scores, which measure how influential each data point is in the effect estimate. A significant change to the estimate after removing a small set of highly-influential points indicates the implicit bias calculation is highly sensitive to a few (potentially) outlier patients. We also include a preliminary interpretability analysis that explores the distinguishing phenotypes of highly influential patients, which could aid in determining if these subsets of patients correspond to some interpretable outlier group. We describe how we calculate the influence score and identify highly-influential patient sets in Appendix C.6.

Income stratification - To investigate intersectionality in implicit biases, we perform a stratified effect estimate over different income groups where $D \neq$Income.

Partial non-linearity of W - Our identification theorem allows for partial non-linearity in the effect of W. We thus re-compute the point estimate allowing for non-linear interactions with W using XGBoost[65] models instead of Lasso.

4. Results

4.1. *Synthetic data validation*

The results in Table 2 demonstrate that our method is able to retrieve the true implicit bias effect $\theta = 0.5$ with high certainty for both fully continuous and mixed-type data, with comparable performance to the best-case OLS where M is known. We report our method's coverage, RMSE, bias, standard deviation, mean confidence interval, and performance on our five tests (from Section 2.3), as well as testing other values of θ, in Appendix F.1.

	θ	Our method	OLS(D, W, M, X)	OLS(D, W, Z, X)
Experiment 1: Continuous	0.5	0.54 ± 0.003	0.5 ± 0.01	1.10 ± 0.01
Experiment 2: Continuous and binary	0.5	0.53 ± 0.003	0.5 ± 0.01	1.385 ± 0.01

Table 2: Semi-synthetic data estimates θ and confidence interval over K=100 iterations.

4.2. *Calculating the implicit bias effect in the UK Biobank*

In Appendix F.2, we show the effect estimates for the (D, Y) pairs using all proxies Z, X, adjusting the confounders W by excluding the column corresponding to the attribute D. However, as evidenced by the failure of the dual and primal tests, we found the initial sets of proxies Z, X did not meet our method's necessary assumptions. As discussed further in Appendix F.3, we believe these test failures indicate there might exist some features in X with a causal path from D that does not go through M or features within Z with a causal path to Y that doesn't flow through M. Such paths invalidate the resulting effect estimates.

We thus found applying our proxy selection algorithm (see 2.3.1) necessary for producing valid effect estimates. After running the algorithm to select subsets of admissible X, Z proxies (the description and interpretation of the selected proxies can be found in Appendix F.3), we found 34 (D, Y) pairs that pass all tests with narrow confidence intervals. We report six in Table 3 and include the remaining estimates in Appendix F.4. Note that $\theta > 0$ implies a patient with D is more likely to be diagnosed with Y due to clinician bias, and conversely $\theta < 0$ implies a patient is less likely to be diagnosed. In Section 5.2, we offer a framework for interpreting the implications of these results.

4.2.1. *Weak instrument confidence interval*

As shown in Figure 2A, the confidence interval predicted under the weak instrument regime consistently aligns with the interval under our method, thus indicating our estimate's robustness to weak instruments.

(D, Y)	$\theta \pm 95\%$ **CI**	**(1) Primal statistic < critical**	**(2) Dual statistic < critical**	**(3)** $\mathbb{E}[\tilde{D}V] \neq 0$ **statistic > critical**	**(4)** V **strength F-test statistic > critical**	**(5)** $\mathrm{Cov}(\tilde{X}, \tilde{Z})$ **rank**
Low income, Depression	0.03 ± 0.02	59.9<60.5	31.9<40.1	84.1>0.4	3332.1>23.1	3
Disability insurance, Rh. Arthritis	0.06 ± 0.0	67.3<75.6	3.4<11.1	29.2>0.4	801.1>23.1	3
Female, Heart disease	-0.19 ± 0.06	115.8<118.8	23.3<23.7	18.8>1.3	92.5>23.1	4
Black, Chronic kidney disease	0.14 ± 0.03	56.9<58.1	10.6<21.0	9.8>0.3	23.3>23.1	4
Obese, Osteoarthritis	0.09 ± 0.02	90.5<100.7	24.8<28.9	76.5>1.7	254.9>23.1	3
Asian, Osteoarthritis	-0.06 ± 0.03	94.7<101.9	33.1<33.9	13.9>0.3	74.6>23.1	5

Table 3: Six of the 34 valid UK Biobank implicit bias effect estimates after applying our X, Z proxy selection algorithm. Tests (1-5) are detailed in 2.3, where *statistic* is the given data's statistic and *critical* is the necessary critical value to be greater or less than to pass. $V = \tilde{D} - \gamma^T \tilde{Z}$.

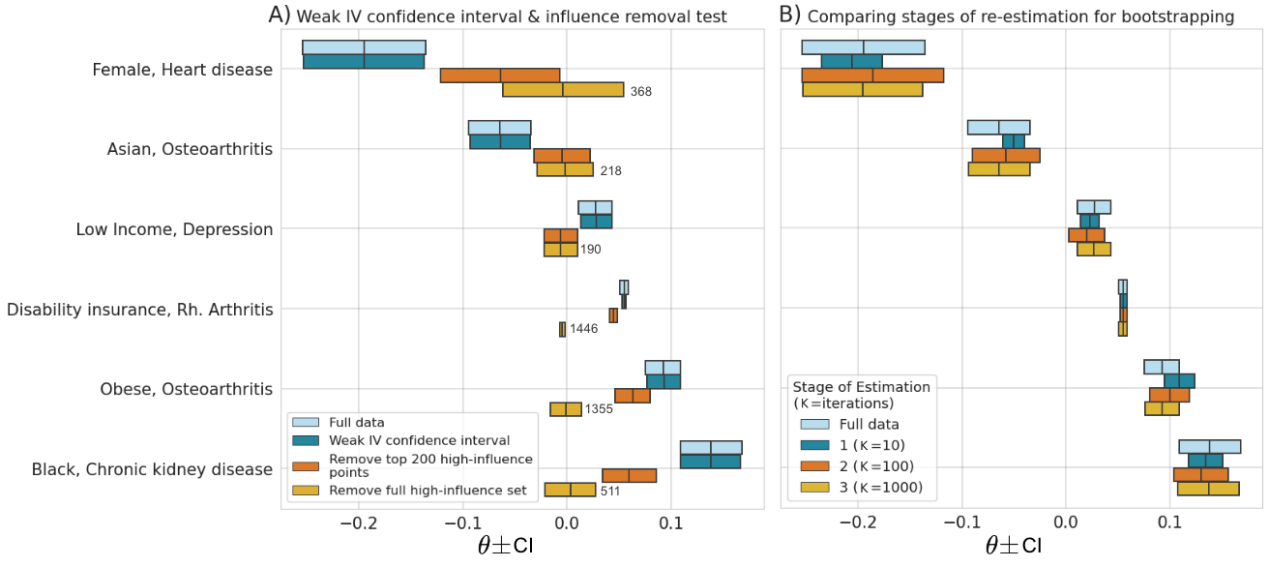

Fig. 2: Comparing effect estimates for six (D, Y) pairs using all data with: A) weak instrument and influence set removal (where the numbers next to the yellow bar reflect the set size of high-influence points); B) bootstrapped subsampling 50% of the data at different stages of re-estimation.

4.2.2. *Bootstrapping analyses*

In Figure 2B, we show the results of the first bootstrap analysis comparing different stages of re-estimation. We observe that, regardless of the estimation stage, bootstrapped estimates are consistent with the estimate from the full dataset. The consistency of the bootstrapped estimates over different sample sizes, as shown in Appendix F.6, further support the robustness of our method.

4.2.3. *Influence points*

In Figure 2A, we see that removing only a few highly-influential points leads to a significant decrease in the magnitude of the estimated effect. To investigate, we run a preliminary interpretabilty analysis where we analyze the univariate differences between patients with high influence and those with low influence. In Figure 3A patients that strongly influence the negative implicit bias estimate for (D=Female, Y=Heart disease) are more likely to be low income, unemployed due to disability, and suffer from depression. It is plausible such patients are the "outliers" driving the strong negative bias estimate.

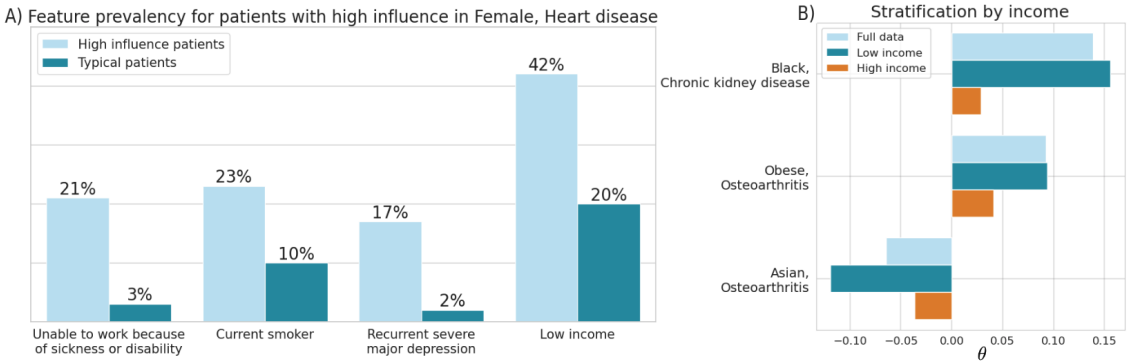

Fig. 3: A) Interpretability into high influence points. B) Income stratification

4.2.4. *Income stratification*

In Figure 3B we analyze the effect of stratification based on income. We see a general increase in bias effect estimate for the low income strata and a corresponding decrease in effect for high income strata, demonstrating potential evidence of intersectional discrimination.[5,66]

4.2.5. *Partial non-linearity of W*

In Appendix F.9, we show our implicit bias estimate with non-linear W interactions leads to a similar effect estimates of θ.

5. Discussion

5.1. *Limitations*

In this work, we propose a robust causal inference method designed to detect clinician implicit bias by estimating pathway-specific causal effects. We demonstrate the applicability of our approach to large-scale medical data by validating on both semi-synthetic and real-world datasets.

However, our work contains several limitations. First, while the UK Biobank is a rich and accessible source of medical data, most patient information is collected once upon signing up for the biobank. Although UK Biobank has synced their records to a handful of outcomes provided by EHR data (like ICD10 codes), it is unclear to what extent the available proxies for X (which were collected at patient enrollment) are used by clinicians for diagnoses. Additionally, the synced ICD10 codes are from hospital records, thus excluding primary care visits. We plan to validate our method with time-series EHR data in follow-up work.

Second, while the assumption of partially linear structural equations is crucial for enabling better identifiability of the outcome bridge function under minimal conditions, it is possible the ground truth equations are non-linear.

Finally, it is well known that intersectional identities shape complex patterns of discrimination in healthcare.[5,66] A more comprehensive analysis on the effect of implicit bias from intersectional attributes on patient treatment would be valuable for improving equity in healthcare outcomes.

5.2. *Interpretation and application of results*

While we re-iterate the intent of this work is not to diagnose specific cases of implicit bias in the UK Biobank, our method did flag several areas of clinical inequity that have been reported in literature. For instance, many works have reported gender-based inequality in cardiovascular health,[67] and we similarly detected an estimate of θ = -0.19 indicating clinicians are less likely, due to implicit biases, to diagnose D=Females with Y=heart disease. In another example, our estimate θ = -0.06 suggested clinicians are less likely to diagnose D=Asian patients with Y=osteoarthritis, and many works have highlighted both patient- and clinician-stigmas regarding pain-associated disorders, like osteoarthritis, in Asians.[68–70]

However, we did find several estimates contrary to what we expected. For example, our estimate $\theta = 0.14$ indicated clinicians are *positively* biased towards diagnosing Black patients with chronic kidney disease. However, at the time of UK Biobank data collection, many doctors relied on a race-based equation for kidney function now known to have under-detected kidney disease in Black patients.[71]

To understand a discrepancy between a produced estimate and literature (or user intuition), we recommend (1) ensuring the data used contains sufficient health proxies and satisfy all assumptions (e.g., see biobank data limitations in 5.1); (2) investigating all mechanisms creating the medical outcome Y (e.g., hospital-specific diagnosis protocol); and (3) exploring how the discovered bias estimate fits in context, rather than opposed, to those found in literature. While our method does not offer a solution on *how* to tackle implicit biases, by bringing awareness to potential areas of discrimination within a given healthcare system, detecting biases is the first step towards creating systemic-level change through interdisciplinary collaboration and targeted anti-bias training programs.

6. Appendix

The appendix can be found at `https://github.com/syrgkanislab/hidden_mediators`.

References

1. J. Holroyd, J. Sweetman, M. Brownstein and J. Saul, The heterogeneity of implicit bias, *Implicit bias and philosophy* **1**, 80 (2016).
2. M. B. Vela, A. I. Erondu, N. A. Smith, M. E. Peek, J. N. Woodruff and M. H. Chin, Eliminating explicit and implicit biases in health care: evidence and research needs, *Annual review of public health* **43**, 477 (2022).
3. D. P. Gopal, U. Chetty, P. O'Donnell, C. Gajria and J. Blackadder-Weinstein, Implicit bias in healthcare: clinical practice, research and decision making, *Future healthcare journal* **8**, 40 (2021).
4. B. Saluja and Z. Bryant, How implicit bias contributes to racial disparities in maternal morbidity and mortality in the united states, *Journal of women's health* **30**, 270 (2021).
5. T. M. Anastas, M. M. Miller, N. A. Hollingshead, J. C. Stewart, K. L. Rand and A. T. Hirsh, The unique and interactive effects of patient race, patient socioeconomic status, and provider attitudes on chronic pain care decisions, *Annals of Behavioral Medicine* **54**, 771 (2020).
6. J. A. Sabin and A. G. Greenwald, The influence of implicit bias on treatment recommendations for 4 common pediatric conditions: pain, urinary tract infection, attention deficit hyperactivity disorder, and asthma, *American journal of public health* **102**, 988 (2012).
7. D. Ueda, T. Kakinuma, S. Fujita, K. Kamagata, Y. Fushimi, R. Ito, Y. Matsui, T. Nozaki, T. Nakaura, N. Fujima *et al.*, Fairness of artificial intelligence in healthcare: review and recommendations, *Japanese Journal of Radiology* **42**, 3 (2024).
8. S. R. Pfohl, A. Foryciarz and N. H. Shah, An empirical characterization of fair machine learning for clinical risk prediction, *Journal of biomedical informatics* **113**, p. 103621 (2021).
9. A. G. Greenwald, D. E. McGhee and J. L. Schwartz, Measuring individual differences in implicit cognition: the implicit association test., *Journal of personality and social psychology* **74**, p. 1464 (1998).
10. S. A. Arif and J. Schlotfeldt, Gaps in measuring and mitigating implicit bias in healthcare, *Frontiers in Pharmacology* **12**, p. 633565 (2021).
11. E. J. T. Tchetgen, A. Ying, Y. Cui, X. Shi and W. Miao, An introduction to proximal causal learning, *arXiv preprint arXiv:2009.10982* (2020).
12. A. T. Forde, D. M. Crookes, S. F. Suglia and R. T. Demmer, The weathering hypothesis as an explanation for racial disparities in health: a systematic review, *Annals of epidemiology* **33**, 1 (2019).
13. D. R. Williams, Stress and the mental health of populations of color: Advancing our understanding of race-related stressors, *Journal of health and social behavior* **59**, 466 (2018).
14. X. Cui and C.-T. Chang, How income influences health: decomposition based on absolute income and relative income effects, *International Journal of Environmental Research and Public Health* **18**, p. 10738 (2021).
15. A. M. K. Minhas, V. Jain, M. Li, R. W. Ariss, M. Fudim, E. D. Michos, S. S. Virani, L. Sperling and A. Mehta, Family income and cardiovascular disease risk in american adults, *Scientific reports* **13**, p. 279 (2023).
16. A. G. Greenwald, N. Dasgupta, J. F. Dovidio, J. Kang, C. A. Moss-Racusin and B. A. Teachman, Implicit-bias remedies: Treating discriminatory bias as a public-health problem, *Psychological Science in the Public Interest* **23**, 7 (2022).
17. C. FitzGerald and S. Hurst, Implicit bias in healthcare professionals: a systematic review, *BMC medical ethics* **18**, 1 (2017).
18. S. Chiappa, Path-specific counterfactual fairness, **33**, 7801 (2019).
19. R. Nabi and I. Shpitser, Fair inference on outcomes, **32** (2018).
20. L. Zhang, Y. Wu and X. Wu, A causal framework for discovering and removing direct and indirect discrimination, *arXiv preprint arXiv:1611.07509* (2016).

21. A. I. Naimi, M. E. Schnitzer, E. E. Moodie and L. M. Bodnar, Mediation analysis for health disparities research, *American journal of epidemiology* **184**, 315 (2016).
22. I. Jun, S. E. Ser, S. A. Cohen, J. Xu, R. J. Lucero, J. Bian and M. Prosperi, Quantifying health outcome disparity in invasive methicillin-resistant staphylococcus aureus infection using fairness algorithms on real-world data, in *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2024*, 2023.
23. W. Pan, S. Cui, J. Bian, C. Zhang and F. Wang, Explaining algorithmic fairness through fairness-aware causal path decomposition, in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021.
24. A. Ghassami, A. Yang, I. Shpitser and E. T. Tchetgen, Causal inference with hidden mediators, *arXiv preprint arXiv:2111.02927* (2021).
25. A. Ghassami, I. Shpitser and E. T. Tchetgen, Partial identification of causal effects using proxy variables, *arXiv preprint arXiv:2304.04374* (2023).
26. X. Chen and D. Pouzo, Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals, *Econometrica* **80**, 277 (2012).
27. C. Ai and X. Chen, Efficient estimation of models with conditional moment restrictions containing unknown functions, *Econometrica* **71**, 1795 (2003).
28. C. Ai and X. Chen, The semiparametric efficiency bound for models of sequential moment restrictions containing unknown functions, *Journal of Econometrics* **170**, 442 (2012).
29. G. Lewis and V. Syrgkanis, Adversarial generalized method of moments, *arXiv preprint arXiv:1803.07164* (2018).
30. N. Dikkala, G. Lewis, L. Mackey and V. Syrgkanis, Minimax estimation of conditional moment models, *Advances in Neural Information Processing Systems* **33**, 12248 (2020).
31. A. Bennett, N. Kallus and T. Schnabel, Deep generalized method of moments for instrumental variable analysis, *Advances in neural information processing systems* **32** (2019).
32. A. Bennett and N. Kallus, The variational method of moments, *arXiv preprint arXiv:2012.09422* (2020).
33. W. Miao and E. T. Tchetgen, A confounding bridge approach for double negative control inference on causal effects (supplement and sample codes are included), *arXiv preprint arXiv:1808.04945* (2018).
34. Y. Cui, H. Pu, X. Shi, W. Miao and E. T. Tchetgen, Semiparametric proximal causal inference, *arXiv preprint arXiv:2011.08411* (2020).
35. A. Bennett, N. Kallus, X. Mao, W. Newey, V. Syrgkanis and M. Uehara, Inference on strongly identified functionals of weakly identified functions, *arXiv preprint arXiv:2208.08291* (2022).
36. A. Bennett, N. Kallus, X. Mao, W. Newey, V. Syrgkanis and M. Uehara, Source condition double robust inference on functionals of inverse problems, *arXiv preprint arXiv:2307.13793* (2023).
37. A. Bennett, N. Kallus, X. Mao, W. Newey, V. Syrgkanis and M. Uehara, Minimax instrumental variable regression and l_2 convergence guarantees without identification or closedness, *arXiv preprint arXiv:2302.05404* (2023).
38. J. Zhang, W. Li, W. Miao and E. T. Tchetgen, Proximal causal inference without uniqueness assumptions, *Statistics & Probability Letters* **198**, p. 109836 (2023).
39. Q. Chen, Robust and optimal estimation for partially linear instrumental variables models with partial identification, *Journal of Econometrics* **221**, 368 (2021).
40. V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen and W. Newey, Double/debiased/neyman machine learning of treatment effects, *American Economic Review* **107**, 261 (2017).
41. J. L. M. Olea and C. Pflueger, A robust test for weak instruments, *Journal of Business & Economic Statistics* **31**, 358 (2013).
42. D. W. Andrews and J. H. Stock, *Inference with Weak Instruments*, Working Paper 313, National Bureau of Economic Research (August 2005).

43. C. L. McMurtry, M. G. Findling, L. S. Casey, R. J. Blendon, J. M. Benson, J. M. Sayde and C. Miller, Discrimination in the united states: Experiences of asian americans, *Health services research* **54**, 1419 (2019).
44. O. Bougie, M. I. Yap, L. Sikora, T. Flaxman and S. Singh, Influence of race/ethnicity on prevalence and presentation of endometriosis: a systematic review and meta-analysis, *BJOG: An International Journal of Obstetrics & Gynaecology* **126**, 1104 (2019).
45. C. Wu, Y. Qian and R. Wilkes, Anti-asian discrimination and the asian-white mental health gap during covid-19, in *Race and Ethnicity in Pandemic Times*, (Routledge, 2021) pp. 101–117.
46. H. N. Garb, Race bias and gender bias in the diagnosis of psychological disorders, *Clinical Psychology Review* **90**, p. 102087 (2021).
47. J. L. W. Taylor, C. M. Campbell, R. J. Thorpe Jr, K. E. Whitfield, M. Nkimbeng and S. L. Szanton, Pain, racial discrimination, and depressive symptoms among african american women, *Pain Management Nursing* **19**, 79 (2018).
48. N. N. Maserejian, C. L. Link, K. L. Lutfey, L. D. Marceau and J. B. McKinlay, Disparities in physicians' interpretations of heart disease symptoms by patient gender: results of a video vignette factorial experiment, *Journal of women's health* **18**, 1661 (2009).
49. J. McClendon, U. R. Essien, A. Youk, S. A. Ibrahim, E. Vina, C. K. Kwoh and L. R. Hausmann, Cumulative disadvantage and disparities in depression and pain among veterans with osteoarthritis: the role of perceived discrimination, *Arthritis Care & Research* **73**, 11 (2021).
50. L. VanPuymbrouck, C. Friedman and H. Feldner, Explicit and implicit disability attitudes of healthcare providers., *Rehabilitation psychology* **65**, p. 101 (2020).
51. I. Stepanikova and G. R. Oates, Perceived discrimination and privilege in health care: the role of socioeconomic status and race, *American journal of preventive medicine* **52**, S86 (2017).
52. S. M. Phelan, D. J. Burgess, M. W. Yeazel, W. L. Hellerstedt, J. M. Griffin and M. van Ryn, Impact of weight bias and stigma on quality of care and outcomes for patients with obesity, *Obesity Reviews* **16**, 319 (2015), Open Access, Citations: 737.
53. M. Fulton, S. Dadana and V. N. Srinivasan, Obesity, stigma, and discrimination, in *StatPearls [Internet]*, (StatPearls Publishing, 2023)
54. X. Han, K. T. Call, J. K. Pintor, G. Alarcon-Espinoza and A. B. Simon, Reports of insurance-based discrimination in health care and its association with access to care, *American journal of public health* **105**, S517 (2015).
55. C. A. McBurney and E. R. Vina, Racial and ethnic disparities in rheumatoid arthritis, *Current rheumatology reports* **14**, 463 (2012).
56. K. Evans, J. Coresh, L. D. Bash, T. Gary-Webb, A. Köttgen, K. Carson and L. E. Boulware, Race differences in access to health care and disparities in incident chronic kidney disease in the us, *Nephrology Dialysis Transplantation* **26**, 899 (2011).
57. K. A. Jenkins, S. Keddem, S. B. Bekele, K. E. Augustine and J. A. Long, Perspectives on racism in health care among black veterans with chronic kidney disease, *JAMA Network Open* **5**, e2211900 (2022).
58. M. Kim, Racial/ethnic disparities in depression and its theoretical perspectives, *Psychiatric Quarterly* **85**, 1 (2014).
59. Z. Rizvi, V. Kunder, H. Stewart, P. Torres, S. Moon, N. Lingappa, M. Kazaleh, V. Mallireddigari, J. Perez, N. John *et al.*, The bias of physicians and lack of education in patients of color with melanoma as causes of increased mortality: a scoping review, *Cureus* **14** (2022).
60. L. Krueger, E. Hijab, J.-A. Latkowski and N. Elbuluk, Clinical decision-making bias in darker skin types: a prospective survey study identifying diagnostic bias in decision to biopsy., *International Journal of Dermatology* **62** (2023).
61. A. Ahrens, C. B. Hansen, M. E. Schaffer and T. Wiemann, ddml: Double/debiased machine learning in stata, *The Stata Journal* **24**, 3 (2024).

62. V. Chernozhukov, C. Hansen, N. Kallus, M. Spindler and V. Syrgkanis, Applied causal inference powered by ml and ai, *arXiv preprint arXiv:2403.02467* (2024).
63. N. Dikkala, G. Lewis, L. Mackey and V. Syrgkanis, Minimax estimation of conditional moment models, *Advances in Neural Information Processing Systems* **33**, 12248 (2020).
64. T. Broderick, R. Giordano and R. Meager, An automatic finite-sample robustness metric: when can dropping a little data make a big difference?, *arXiv preprint arXiv:2011.14999* (2020).
65. T. Chen and C. Guestrin, XGBoost: A scalable tree boosting system, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16 (ACM, New York, NY, USA, 2016).
66. O. Ogungbe, A. K. Mitra and J. K. Roberts, A systematic review of implicit bias in health care: A call for intersectionality, *IMC Journal of Medical Science* **13**, 5 (2019).
67. J. Bosomworth and Z. Khan, Analysis of gender-based inequality in cardiovascular health: An umbrella review, *Cureus* **15** (2023).
68. S.-Y. Yang, E. Y. S. Woon, K. Griva and B. Y. Tan, A qualitative study of psychosocial factors in patients with knee osteoarthritis: insights learned from an asian population, *Clinical Orthopaedics and Related Research®* **481**, 874 (2023).
69. G. C. Gee, M. S. Spencer, J. Chen and D. Takeuchi, A nationwide study of discrimination and chronic health conditions among asian americans, *American journal of public health* **97**, 1275 (2007).
70. K. Kumar, R. J. Stack, A. Adebajo and J. Adams, Health-care professionals' perceptions of interacting with patients of south asian origin attending early inflammatory arthritis clinics, *Rheumatology Advances in Practice* **3**, p. rkz042 (2019).
71. M. A. Marzinke, D. N. Greene, P. M. Bossuyt, A. B. Chambliss, L. R. Cirrincione, C. R. McCudden, S. E. Melanson, J. H. Noguez, K. Patel, A. E. Radix *et al.*, Limited evidence for use of a black race modifier in egfr calculations: a systematic review, *Clinical chemistry* **68**, 521 (2022).
72. J. H. Stock, Weak instruments, weak identification, and many instruments: Part 1 and part 2 (2018), NBER Summer Institute Methods Lectures.
73. H. Weyl, Das asymptotische verteilungsgesetz der eigenwerte linearer partieller differentialgleichungen (mit einer anwendung auf die theorie der hohlraumstrahlung), *Mathematische Annalen* **71**, 441 (1912).
74. I. Montagni, T. Cariou, C. Tzourio and J.-L. González-Caballero, "i don't know","i'm not sure","i don't want to answer": a latent class analysis explaining the informative value of non-response options in an online survey on youth health, *International Journal of Social Research Methodology* **22**, 651 (2019).