DISPARATE TREATMENT AND OUTCOMES IN EMERGENCY DEPARTMENTS: EVIDENCE FROM FLORIDA

Minu Philip & Ozde Ozkaya*

November 3, 2024

Updated regularly, click here for the latest version

Abstract

Strokes are roughly twice more likely to be missed among Black patients, with most of the disparity arising from physicians testing Black patients less often. We develop a method to quantify the role of disparate treatment by physicians in driving this difference in testing. Specifically, we leverage a unique feature of strokes: whether a patient actually had a stroke can be inferred retrospectively even if initially misdiagnosed. This allows us to benchmark testing decisions against racially objective predictions of stroke risk made by a machine learning model trained on the true underlying stroke states. We decompose disparate treatment into two forces: an unjustified skill gap, where physicians make noisier risk assessments for Black patients; and racial prejudice, where physicians are less likely to test Black patients conditional on their risk assessment. Disparate treatment accounts for about 65% of the racial disparity in testing. Removing racial prejudice would lower testing disparities by half.

*Philip: New York University (email: minu.philip@nyu.edu); Ozkaya: Keystone Strategy, New York.
We are very grateful to Debraj Ray, Martin Rotemberg, Guillaume Fréchette, and Quang Vuong for their constant guidance and support. This project has also greatly benefited from conversations with Marco Morucci, Daniel Waldinger, Carlos Fernandez-Granda, Dr. Christopher Caspers (MD), Dr. Alexandra Ortego (MD), Rajeev Dehejia, Micheal Dickstein, Hamish Low, David Canning, Marcella Alsan, Sharon Traiberman, Raquel Fernandez, Christopher Flinn, Sahar Parsa, Pierre Bodéré, David Cesarini, Jaroslav Borovicka, Elena Manresa, Vishal Kamat, Jonathan Morduch, Willam Easterly, Arielle Bernhardt, Anna Vitali, Matthew Eisenberg, Petra Todd, Jerome Adda, and J Carter Braxton. We are also indebted to Anne Stubing and Geby Varughese for invaluable administrative and IT support. The data used in our analysis are sourced from Florida, State Emergency Department Databases (SEDD), State Inpatient Databases (SID), and State Ambulatory Surgery and Services Databases (SASD) of the Healthcare Cost and Utilization Project (HCUP), Agency for Healthcare Research and Quality. The views expressed herein are those of the authors and do not necessarily reflect the views of the Agency for Healthcare Research and Quality.

I. INTRODUCTION

Racial disparities in health outcomes are widely prevalent and persistent (Institute of Medicine, 2003). These disparities may result from multiple aspects; such as racial differences in the quality of facilities that patients go to (Chandra et al., 2024), differences in communication and self-advocacy (Alsan et al., 2019), differences in symptomatic presentations, or differential treatment by providers (Chandra & Staiger, 2010; Institute of Medicine, 2003; Balsa & McGuire, 2003). We can broadly classify these aspects into two groups: those that drive racial disparity via differential access to care, and those that drive disparity via their effect on clinical decision-making within a facility. With improvements in diagnostic technology and insurance coverage, Chandra et al. (2024) report a substantial decline in the quality difference between facilities that Black and non-Black patients seek care at. In this paper, therefore, we focus on the latter: disparity in outcomes that arise from clinical decision-making, and more specifically the role of *disparate treatment* by providers.

Disparate treatment refers to when a physician makes clinical decisions differently for one group of patients than for another, in a way that is not justified by medicine. Conceptually, it encompasses taste-based discrimination as well as the use of incorrect priors, beliefs, or stereotypes in decision-making.

In this paper, we describe a framework for studying disparity in clinical decision-making when the underlying state of whether the patient actually has the disease can be inferred by the econometrician. The 'outcome' that we focus on is the *quality* of physicians' decisions i.e. does it match the underlying state or not. We then ask whether the racial disparity in outcomes comes from *disparate treatment*, and undertake a quantitative assessment of its contribution relative to other relevant factors.

We apply this framework to study racial disparity in stroke diagnosis in emergency departments (EDs). Stroke is a cerebrovascular event during which blood supplied to a part of the brain gets interrupted, resulting in the damage and death of brain tissue. It is one of the leading causes of mortality and long-term neurological disability worldwide, and in particular in the United States (CDC, 2024). For instance, in 2021, the burden of stroke totaled 160.4 million disability-adjusted life years, and 7.44 million deaths worldwide (Institute for Health Metrics and Evaluation, 2024; Martin et al., 2024). Since strokes are characterized by rapid deterioration in brain function, patients who are misdiagnosed in their initial visit soon find themselves back again in the hospital. This life-threatening and non-self-recovering nature of strokes allows us to retrospectively infer whether an ED visit was indeed occasioned by a stroke, based on patient revisits, regardless of how it was initially diagnosed.

Using comprehensive administrative healthcare data from the Healthcare Cost and Utilization Project (HCUP), we track patient visits in Florida for the years 2016-2017 across emergency departments, in-patient settings, and ambulatory or out-patient settings.[1] Our primary sample spans 1,368,560 visits serviced in the ED in 2016, made by patients aged

[1]The data is sourced from the Florida State Emergency Department Databases (SEDD), State Ambulatory Surgery and Services Databases (SASD), and the State Inpatient Databases (SID) of HCUP; with unique patient identifiers that link patient visits within the state, across care-settings and time.

between 18-80 years and presenting with any symptom associated with stroke. For each visit, we code a binary indication of whether or not the visit was actually due to a stroke—which we call as the *latent stroke state*. It is 'latent' because at the time of physician decision-making, the underlying state is unknown to them. We label an ED visit as a stroke state if the patient is either diagnosed with stroke or is missed but subsequently diagnosed within 14 days on a revisit with similar symptoms.[2] Visits incorrectly diagnosed as stroke are identified by revisits within 60 days that result in a stroke mimic diagnosis, with no future mentions of personal stroke history. Importantly, we can infer latent stroke states for all ED visits, in a non-selective way, regardless of how they were diagnosed at initial presentation.

With latent stroke states known, we assess the quality of two sequential decisions of the physician—testing (i.e. whether neuroimaging was ordered) and diagnosis—and examine how it varies with patient race. Starting with diagnosis, we then trace back to testing decisions to pinpoint the origin of disparity. We find that strokes are missed twice as often for Black patients, with a 28% false negative rate compared to 16% for non-Black patients. The disparity remains this large even after patients' sex, age, comorbidity-profile, and insurance status are controlled for. The rate of false positives is low overall (about 0.03%), with no significant difference by race. In testing, we again find disparity, with Black stroke patients being 11.5 percentage points less likely to be tested in accordance with the stroke protocol, amounting roughly to a shortfall of 13.5%. Since testing decisions precede diagnosis, any disparity in testing decisions also contributes subsequently to disparity in diagnosis. Conditional on being tested adequately, however, we find no significant difference in misdiagnosis rate by patient race. Components analysis reveals that over 90% of the observed disparity in missed diagnosis rates can be attributed to differential testing, highlighting that the relevant decision to be examined for disparate treatment is that of testing.

Disparities in testing, however, are not necessarily indicative of disparate treatment. They can also arise from the potential selection of patients across differently skilled physicians and facilities, or from structural differences in the quality of information available based on which the physician is expected to infer stroke risk.[3] The quality of information matters because if it were harder to predict stroke risk for one group, a physician would make more errors in determining whom to test in that group despite being unbiased. Some ways in which the quality of information may differ across racial groups are through differences in the symptomatic presentation, differences in reporting of symptoms and history, or in the quality and availability of past medical records. Empirical inference of disparate treatment can therefore only be made if all the above features can be suitably controlled for.

Within-facility comparisons of outcomes, and the exploitation of quasi-random assignment of physicians to cases in the ED address the first of these concerns—that of selection into varying qualities of care. To address the second concern about the quality of information being different across groups, we define *objective risk assessments* or the best prediction of stroke risk for each visit *conditional on the information available to the physician at the time*

[2]Identification of missed stroke cases based on retrospective visit review or chart review is familiar in the medicine literature (Arch et al., 2016; Newman-Toker et al., 2014, 2022).

[3]The difference in information quality is synonymous to the *subgroup validity problem* in Ayres (2002).

of decision-making, whatever its quality.

We use machine learning to obtain the objective stroke risk predictions. Machine learning models are non-parametric and can identify non-linear and complex patterns in data, outperforming traditional statistical methods. We train a machine learning algorithm (XGBoost) to take as input the information available to physicians at the time of decision-making, including patient race, and predict the latent stroke state. The probabilistic predictions by the machine learning model provide the objective risk assessment for each ED visit. The features used in the algorithm cover all symptoms, co-morbidities, and risk factors from the current visit as well as from the patient's history. For this, we rely on comprehensive encounter-level information provided in HCUP datasets that include administrative details, patient demographics, and a detailed representation of the health encounter in the form of alphanumeric diagnosis codes and procedure codes.[4] We also obtain patient history and risk factors from all previous encounters made by the patient in the year, across different care-settings. The challenge here though is that we may not be observing all of which the physician observes about the patient. There may be some features, say patient demeanor, that are observed by the physician but are unobservable to the econometrician. To the extent that such unobservables are correlated with race and relevant to stroke, the inclusion of race as a feature assures that the algorithm learns from all race-correlated patterns in the data, including those mediated via unobserved factors. The predictions made by the algorithm therefore subsume race-specific and race-correlated differences in assessable stroke risk, allowing cross-group comparisons. Disparate treatment is then indicated by differential rates of testing *within* a facility and *conditional on the same objective stroke risk assessment*.

We find, conditional on the same decile of objective risk prediction, testing is around 24% lower on average for Black patients. The difference persists even among patients who have no contraindications to any imaging modality. It is significant particularly in the first decile of predicted stroke risk. Notably, disparate treatment by physicians explains 65% of the difference in testing rates between Black and non-Black patients.

Next, we model physicians' decision to test, and formalize two mechanisms driving disparate treatment in testing: *racial prejudice* and *unjustified skill gap*. Racial prejudice refers to physicians using different thresholds for different racial groups when deciding whether to test them. Unjustified skill gap refers to when the physician systematically makes noisier stroke risk assessments for one group relative to the other. We evaluate the noise in physicians' risk assessments relative to the objective risk predictions, to account for differences in the quality of information available. Any differential accuracy is therefore *unjustified* by statistical differences in risk or informativeness of the presentation, and originates specifically from physician error that systematically disfavors one group. Unjustified skill gap can be interpreted to be a consequence of low physician effort, the use of incorrect priors and stereotypes, or of race-insensitive medical protocols.[5]

[4]Diagnosis codes are detailed translations of written descriptions of diseases, symptoms, and risk factors, in the form of alphanumeric codes based on a standardized classification system. Likewise, procedure codes identify specific surgical, medical, or diagnostic interventions. HCUP datasets list up to 10-31 alphanumeric ICD-10-CM diagnoses codes, and 35 ICD-10-PCS/HCPS/CPT procedure codes for each encounter.

[5]Similar to *biased beliefs* in Hull (2021), or *inaccurate* statistical discrimination in Bohren et al. (2024).

We model physicians to be heterogeneous in their choice of threshold and in the quality of risk assessment they make; both of which vary with the "type" of the patient that the physician is interacting with. We define patient types based on a finite set of patient traits including race. Physicians' choice of threshold is determined by their preferences in how they trade-off disutility from false negatives relative to that from false positives. The quality of their risk assessment, or skill, determines the variance of their subjective assessments around the objective risk. The lower the physician's skill, the more false negative and false positive decisions are made. Threshold and skill are jointly identified by the size and ratio of the false negative and false positive decisions made by the physician. Conditional on same threshold, the size of the false negative and false positive rates determines the skill of the physician. The ratio of false negatives to false positives, conditional on a level of skill, pins the threshold. To estimate the model, we use Hierarchical Bayes and sample the parameters from the joint posterior using a Monte Carlo Markov Chain (MCMC) Gibbs sampler.

We find physicians to both use higher thresholds for testing and make noisier risk assessments for Black patients. On average, physicians' threshold for log odds risk is incrementally higher for Black patients by 0.280. Equalizing thresholds by race, all else same, closes the racial difference in false negatives by half. As for unjustified skill gap, the standard deviation in subjective log odds assessments is, on average, 19% higher for Black patients. The distinction between the two mechanisms is not only conceptual but also in how they inform policymaking. Policies that separately attend to either skill or threshold are insufficient to close the racial disparity in false negatives. For instance, combining physicians' testing decision with recommendations from a machine learning model guards patients against errors made by low-skilled physicians. An untargeted application of such a policy, however, reduces false negatives for both racial groups and more so for non-Black patients who are typically subject to lower thresholds. The policy could then in turn end up widening racial differences. Likewise, lowering costs of filing malpractice lawsuits (or more generally, imposing external costs on false negatives) lowers thresholds for all groups, while difference in skill still remains.

Related Literature. Our research builds on the empirical literature that tests for taste-based discrimination in decisions. Some settings studied in this literature include: stop and search for contraband by police officers (Knowles, Persico, & Todd, 2001; Anwar & Fang, 2006; Antonovics & Knight, 2009; Feigenberg & Miller, 2022), bail-setting by judges (Arnold, Dobbie, & Yang, 2018; Arnold, Dobbie, & Hull, 2022), loan approvals (Dobbie et al., 2021), screening for disability insurance (Low & Pistaferri, 2019), and academic journal evaluations (Card et al., 2020), among others. The empirical literature typically identifies prejudice or taste-based discrimination by comparing post-decision outcomes across groups at the decision margin, referred to as the *marginal outcome test*. This scholarship explores various ways to tackle the main empirical challenge in the implementation of marginal outcome tests—identification of the 'marginal' individuals from the decision-marker's perspective whose post-decision outcomes must be compared. Our work complements that of Low and Pistaferri (2019) examining gender differences in false negative rates of disability insurance screening. Our contribution to this literature is in identifying disparate treatment that arises not only from taste-based discrimination, but also from biased risk prediction. Arnold et al. (2022) also model decision-makers to vary in racial prejudice and in "skill" i.e. the informativeness

of the signals drawn for each racial group. Differently from their approach, we are able to use objective risk assessments to separate informativeness of the signal generated (or quality of the information set) from the decision-maker's ability to infer risk from it.

This paper also relates to the literature on variations in physician practice styles (Abaluck et al., 2016; Chandra & Staiger, 2020; Chan et al., 2022; Gowrisankaran et al., 2023). For example, Abaluck et al. (2016) assume physicians to vary in testing intensity but remain identical in potentially mis-weighting risk factors when evaluating patient risk; and Chan et al. (2022) model radiologists to be heterogeneous in skill and vary their diagnosis rates in response. Although this literature focuses on heterogeneity in decision-making and does not make between-group comparisons, our research borrows from this literature when modeling physician heterogeneity and extends it by examining its interaction with patient type. Our paper is closest to Chan et al. (2022) in its approach; the difference is that when making cross-group comparison of physician skill, we also allow the informativeness of the signal to differ by patient race or other race-correlated characteristics.

Methodologically, the paper also complements a growing literature that uses machine learning to evaluate the quality of human decisions. Most notably, Kleinberg et al. (2018) and Dobbie et al. (2021) test if machine learning predictions can achieve more favorable outcomes compared to human decision-makers in the contexts of bail setting and consumer lending, respectively. Similarly, Mullainathan and Obermeyer (2022) use machine learning predictions of heart attack risk to identify prevalence of under-testing and over-testing. Differently from these applications, we don't compare physicians against a machine learning model. Instead, we use machine learning predictions as a stand-in for variations in the quality of information that physicians have access to, enabling cross-group comparisons.

Finally, this paper contributes to a large multi-disciplinary literature on racial disparity in health. Disparities have been studied to be a consequence of underlying differences in the quality of care that one has access to (Chandra et al., 2024), differences in communication and self-advocacy (Alsan et al., 2019), differences in the symptomatic presentations, and of differences in care-seeking tendencies (Jayaraman et al., 2014). Operating simultaneously with these structural aspects, biases of healthcare providers have also been found to be a key contributing factor (Singh & Venkataramani, 2024; Institute of Medicine, 2003; Balsa & McGuire, 2003). Our research contributes to this scholarship by studying the mechanisms driving the disparate treatment.

The remainder of this paper is organized as follows. Section II provides some background on strokes, their diagnostic pathway, and the data used, including imputation of latent stroke states. Section III outlines the conceptual framework and empirical strategy. Section IV examines racial disparities in stroke diagnosis and testing, estimating disparate treatment using algorithmic stroke risk predictions. Two mechanisms of disparate treatment—racial prejudice and unjustified skill gap—are explored in Section V through a model of physician testing, which is then taken to data. Section V.E discusses counterfactuals and policy simulations. Finally, Section VI concludes.

II. SETTING AND DATA

II.A. Strokes

Stroke refers to a cerebrovascular event during which blood supply to a part of the brain gets interrupted. If the interruption is caused by blockages in blood vessels the event is called an *ischemic* stroke, and if caused by the rupture of blood vessels it is called a *hemorrhagic* stroke. Ischemic strokes are more common and constitute about $80-90\%$ of all stroke cases (Tsao et al., 2023). The abrupt interruption in blood supply cuts off the access of brain cells to oxygen and nutrients, thereby causing rapid damage to brain tissue. As a result, strokes come with a high risk of permanent loss of brain function, long-term disability, or even death. Strokes are therefore considered medical emergencies that warrant immediate medical attention.

The symptoms of stroke depend on the part of the brain that is affected. They range from more specific or typical indications such as weakness on one side of the body, facial drooping, difficulty in speech and comprehension, or even paralysis, to several non-specific and diffuse symptoms such as headaches, confusion, lack of balance, and dizziness. The symptomatic presentation of stroke is also found to vary by race, sex, and stroke subtype (Rathore et al., 2002). Common risk factors include high blood pressure, diabetes, atrial fibrillation, high cholesterol levels, and a personal or family history of stroke; some of which are differently prevalent across racial groups.

An accurate diagnosis of stroke calls for an ability to carefully examine the presenting symptoms, the patient's history, and neurological function; relying greatly on the physician's attentiveness and subjective risk assessment. Diagnostic errors are very costly and can potentially result in preventable deaths or serious long-term neurological disabilities. The longer a stroke episode goes untreated, the greater is the damage to the brain as more cells and tissue in the affected area continue to die. And yet, stroke is frequently misdiagnosed in the emergency department, and is one of the leading causes of death in the United States (Newman-Toker et al., 2022). Since EDs are fast-paced and high-volume environments, the scope for diagnostic errors and delays is even greater (Newman-Toker et al., 2014; Tarnutzer et al., 2017).

II.A.1. Diagnostic Pathway for Stroke

All visits to an ED are first triaged, registered, and then assigned to physicians available at the time on a priority basis. The assignment of physicians to cases is typically random conditional on the physicians' shift schedules. Physicians assess their patients, review vitals and relevant history, and order tests before diagnosis. If a physician evaluates a reasonably high stroke risk, the stroke protocol is activated.

The stroke protocol dictates that all suspected cases receive emergency neuroimaging, followed by immediate neurological assessment and prioritized blood work. If the patient is brought in already under suspicion for stroke, some of these procedures are done by the Emergency Medical Services (EMS) themselves and the ED is alerted in advance to rush the

patient for immediate neuroimaging upon arrival. Similarly, if the triage suspects stroke for a walk-in patient, the ED physician is notified and the patient is rushed for neuroimaging. Registration and initial blood-work are all prioritized at the bedside.

Neuroimaging is the principal step under any protocol for suspected stroke. It is performed to confirm stroke, assess the extent of brain injury, and to identify the type of stroke and its precise location. Although neuroimaging is ordered by the physician, the scans produced are generally read and interpreted by a radiologist whom the attending physician can consult with when necessary. Non-contrast computed tomography (CT) is the primary imaging modality recommended in the protocol. This is because CT is fast, taking approximately 20 minutes, and is widely available. CT, however, has low sensitivity for identifying ischemic strokes compared to other imaging procedures (Mullins et al., 2002; Chalela et al., 2007). For cases where the CT doesn't detect a stroke, physicians may order additional scans for higher quality of evidence.[6] If there are signs of hemorrhage or infarction on the initial CT images, follow-up imaging is unnecessary, and the patient is diagnosed and treated accordingly. In rare scenarios, stroke may be also diagnosed in the absence of neuroimaging, based on clinical presentation, risk factors, and neurological evaluation.

While scans from neuroimaging procedures are analyzed by radiologists, attending physicians are responsible for the overall diagnosis and treatment plan of the patient. Importantly, the decision to order imaging is typically made by the attending physician (Broder et al., 2016).

The monetary cost of neuroimaging is identical for all patients, but contraindications to testing may be differently prevalent across racial groups. Nevertheless, contraindications rarely cause a patient to receive no neuroimaging at all, since any of several neuroimaging modalities can be utilized to identify stroke. Appendices E.4 and E.5 lists the different testing modalities that can be used as well as their contraindications. Patients' ability to pay for the test shouldn't matter in case of stroke because the Emergency Medical Treatment and Active Labor Act (EMTALA) 1986 guarantees to all individuals who present to an ED with an emergency medical condition, the access to medical screening/exams and the necessary treatment for stabilization regardless of their ability to pay.

Once stroke is confirmed, patients are treated appropriately based on the stroke-type. The longer a stroke goes untreated, the more substantial in the loss in brain tissue and function; as is emphasized in the phrase "Time is Brain". The chances of complete recovery are the highest for patients diagnosed within 3-4 hours of the first symptoms (National Institute of Neurological Disorders and Stroke rt-PA Stroke Study Group, 1995; Bluhmki et al., 2009).

II.A.2. Transient Ischemic Attacks or 'Mini-Strokes'

Transient ischemic attack (TIA) or a 'mini-stroke' is an event where the interruption of blood supply is only temporary, and resolves by itself when the clot moves away within a few minutes or a few hours. TIAs still constitute a medical emergency and have the same

[6] Follow-up imaging modalities ordered by physicians include: CT Perfusion (CTP), CT Angiography (CTA), Magnetic Resonance Imaging (MRI), Magnetic Resonance Diffusion, Magnetic Resonance Imaging Perfusion, or Magnetic Resonance Angiography (MRA).

main symptoms as stroke, except that the symptoms are temporary. TIAs are however an important risk factor and warning sign for an impending stroke. They are often followed by an episode of acute stroke, with the highest risk of incidence being within the first few days and up to a week after.

Monitoring and testing TIA cases is therefore crucial for preventing impending acute strokes. Cases of TIA are hence recommended to be tested for blockages, plaque in arteries, or blood clots, and monitored for new or returning symptoms. Once again, neuroimaging is the principal step recommended for TIA cases. While TIAs are not caught on CT scans, the objective of neuroimaging is to rule out signs of an impending stroke. This is then followed by procedures such as contrast CT, MRI, Carotid Ultrasound, Transcranial Doppler, or 12-lead Echocardiogram, to identify the source of blood clots. When such a source is identified, the fact of high stroke risk is communicated to the patient and appropriate treatment is given to regulate blood pressure, lower cholesterol, and prevent new clots.

The analysis in this paper focuses specifically on how acute stroke is diagnosed in the emergency department. However, when defining latent stroke states in Section II.C.2, we also classify missed TIAs that subsequently result in a stroke, as cases of 'missed' stroke. This is both because strokes that occur after missed TIAs could have been prevented had the underlying cause been treated, and because missed TIAs are observationally equivalent to missed strokes. The incidence of ischemic strokes, as well the recurrence of stroke following TIA is substantially higher among Black adults compared to non-Black adults (Kamel et al., 2020). When assessing the quality of stroke diagnoses, however, we compare rates of false negatives and false negatives across groups, making this differences in incidence irrelevant.

II.B. Data

The primary sample for our analyses consists of all ED visits in Florida in 2016 where a patient between the ages of 18 and 80 presents with any symptom associated with stroke.

The data on ED visits is sourced from the State Emergency Department Databases (SEDD) of the Healthcare Cost and Utilization Project (HCUP). HCUP databases are derived from administrative data provided by participating states, and contain detailed visit-level abstracts of inpatient stays, ambulatory surgery or services visits, and emergency department encounters.[7] We combine the data on ED visits with the State Ambulatory Surgery and Services Databases (SASD) and the State Inpatient Databases (SID), to comprehensively record all hospital visits made in the years 2016 and 2017, across facilities and care-settings, linked by patient and physician identifiers. The cross-linking of ED visits to other care-settings is relevant for inferring the latent stroke state for each ED visit, as detailed under Section II.C.

From a total of 8,884,669 ED visits made in 2016, we narrow down to visits by patients in

[7]The Florida Agency for Health Care Administration (AHCA) provides HCUP with data on: ED visits to all licensed hospital-owned EDs in the state; inpatient stays in non-federal Florida hospitals, freestanding comprehensive rehabilitation facilities, and acute care psychiatric hospitals; and outpatient services from hospital-owned and non-hospital-owned (including physician-owned) ambulatory surgery centers, lithotripsy centers, and cardiac catheterization laboratories.

the specified age range, presenting with at least one of the many symptoms of stroke, and excluding visits obviously unrelated to stroke (caused by external causes of morbidity such as animal bites, poisoning, among others).[8] We also exclude visits to EDs that don't see any stroke patient visit during the year 2016, and those with missing information on the ED and visit time. This gives us the primary sample for our analysis, covering a total of 1,368,560 ED visits made over the year by 1,031,793 unique patients. Appendix E lists all the stroke symptoms used in our inclusion criteria, as well as the external causes of morbidity used for exclusion.

Each record in the data represents a unique patient visit. For each visit, the data provides information on the patient, the facility, the attending physician, and the specifics of the the visit. Patient information includes a patient identifier and associated patient demographics such as age, sex, race, patient location, and income quartile of the patient's zipcode within the state. Patients are categorized into racial groups based on the patient race available on record.[9] We refer to Black and Black Hispanic patients collectively as Black patients in this paper. Specifics of the visit include the hour of admission, quarter, facility identifier, physician identifier, duration of visit, and discharge information. The record also details the primary expected payor (insurance) for the visit, and up to 10-31 alphanumeric ICD-10-CM diagnoses codes and 35 ICD-10-PCS/HCPCS/CPT procedure codes.[10] The data on each visit consists of the specifics from the current visit, as well as the details from the patient's history of all previous healthcare encounters.

The diagnoses and procedure codes listed on each record offer a detailed representation of the particular health encounter. The merit of this dataset lies in this comprehensive description of each patient visit, which is ideal for our analysis. Diagnosis codes specify disorders, symptoms, abnormal findings, patient risk factors, and the nature of the encounter. Each diagnosis code can be potentially specified up to seven alpha-numeric characters, with each additional character hierarchically describing the general category of disease, etiology, anatomic site, severity, and episode of care. In the same way, procedure codes specify all medical tests, procedures, and services delivered during the patient encounter, including the use of any drugs, specialized services, or equipment. Together, the two sets of codes summarize the medical aspects of the patient's visit. It is in the interest of the provider to code patient visits accurately since procedure codes are used for medical billing, and the diagnosis codes on record attest to the medical necessity of the procedures performed.

[8] A visit is defined as a unique combination of the patient identifier, visit date, admission hour, discharge hour, and care-setting. This is to avoid counting separately billed instances from the same encounter as separate visits. In case of duplicates (there are 37 such cases), we keep the last record.

[9] Race information is missing for about 0.4% of the visits in the primary sample.

[10] ICD-10 is a comprehensive medical classification system managed by the World Health Organisation with over 70,000 ICD-10-PCS procedure codes and 69,000 ICD-10-CM diagnoses codes. CPT (Current Procedural Terminology) is a code set managed by the American Medical Association that describes medical procedures and services delivered, using more than 11,000 numeric codes. HCPCS (Healthcare Common Procedure Coding System) is another standardized coding system maintained by the Centers for Medicare and Medicaid Services with about 8000 alphanumeric codes used to identify services provided to Medicare and Medicaid patients. All these coding systems are regularly revised and updated to keep up with advances in medical knowledge and technology.

Over-billing and under-documentation are constrained by the risk of the insurer denying the claim.[11]

Further, we use facility identifiers to identify the level of stroke certification for each ED—none, primary, or comprehensive—as of January 1, 2016, using reports from the Florida Department of Health. Patient county FIPS code is also used to obtain county level data on the age adjusted rate of stroke hospitalizations and stroke deaths in the years 2013-15.

Table 1 provides a summary of the overall data sample as well as a breakdown by patient race. Visits by Black patients constitute about 25 percent of the primary sample. It is interesting to note at the outset that Black patients in our sample differ from non-Black patients on several counts. They are younger, have fewer comorbidities (age-adjusted), are predominantly female, and also less likely to be insured. Patients from the two racial groups also present with different types of symptoms, with Black patients somewhat more likely to experience non-specific symptoms like headache, weakness, or nausea.

II.C. Definition of Key Variables

II.C.1. Stroke Diagnosis

For every ED visit in the sample (indexed by i), we encode the attending physician $j(i)$'s diagnosis of stroke as D_{ij}. We code $D_{ij} = 1$ if either (1) the main diagnosis on the visit record is stroke, or (2) the ED record indicates a transfer of the patient to an in-patient facility, with the admit reason or diagnosis on the in-patient record listed as stroke. Otherwise, we code $D_{ij} = 0$.

II.C.2. Latent Stroke State

Since strokes are acute, symptomatic, and an emergency condition, a stroke episode that is missed by the physician at initial presentation will eventually need to be brought again to the health system very soon. We exploit this feature to assign the latent state of whether or not a visit was truly occasioned because of stroke (denoted by $S_i \in \{0, 1\}$), based on patient revisits. Identification of missed stroke cases based on retrospective visit review or chart review has been done previously in the medical literature; see for example Arch et al. (2016); Newman-Toker et al. (2014, 2022). It is possible for us to unambiguously do this because we use the unique patient identifier associated with each ED visit to track all other visits made by the specific patient in the year, across emergency departments, in-patient settings, and ambulatory or outpatient settings in the state of Florida.

If any patient is diagnosed with a stroke (or its sequelae) in any facility, and had visited an ED in the last 14 days prior to it with similar symptoms, the associated relevant prior ED visit is coded as $S_i = 1$ as long as it wasn't diagnosed as TIA. Our rationale for a 14-day interval is that it also captures cases of missed transient strokes (TIA) or mini-strokes that

[11] Codes are generally entered by physicians or medical coders based on physician documentation and patient chart. They are not only used for medical billing or medical documentation of patient history, but also to track disease statistics, utilization and cost.

TABLE 1 :
Summary Statistics of the Sample

	Visits by All	Visits by Non-Black patients	Visits by Black patients
	(1)	(2)	(3)
Patient Characteristics			
Black	0.2517	0.0000	1.0000
Age	45.5272	46.9634	41.2585
Female	0.6338	0.6226	0.6668
Hispanic	0.1740	0.2240	0.0254
Uninsured	0.1686	0.1553	0.2081
Charlson Comorbidity Index (age-adjusted)	1.3594	1.4243	1.1663
Personal History of Stroke	0.0387	0.0386	0.0392
Visit Characteristics			
Duration (Hours)	7.8029	7.9477	7.3734
Weekend	0.2585	0.2583	0.2593
Number of Procedure Codes	11.0945	11.4292	10.0997
Number of Diagnosis Codes	4.5262	4.6321	4.2112
Presenting Symptoms*			
General	0.7903	0.7862	0.8027
Sensory	0.0512	0.0524	0.0476
Speech	0.0062	0.0069	0.0043
Muscular	0.0095	0.0103	0.0070
Facial	0.0037	0.0041	0.0023
Visual	0.0221	0.0227	0.0207
Alertness/Consciousness	0.1679	0.1741	0.1494
Neuroimaging	0.2872	0.2959	0.2613
Total Visits	1,368,560	1,024,027	344,533

Notes: This table provides a summary of the primary sample. The primary sample consists of all ED visits in Florida in 2016 where a patient aged between 18-80 years presented with any symptom associated with stroke. It excludes visits caused due to accidents or events that are obviously unrelated to stroke but may have a symptom in common. *The specific ICD-10-CM codes for the symptoms under each category are specified under Appendix E.

later present as strokes which could have been prevented or prepared for. Visits with $S_i = 1$ and $D_{ij} = 0$ would therefore be those with missed acute strokes and missed TIAs that were followed by stroke within the next 14 days. We refer to these cases of false negatives as *missed diagnoses*. In our analyses later, we show that our findings are robust to interval choices that are different from the 14-day window.

For visits that are diagnosed with stroke in the ED, if (1) the patient returns to the health system within the next 60 days with similar symptoms and is instead diagnosed with a stroke mimic,[12] *and* (2) the personal history of stroke is not recorded in any of the patients' future visits, we code $S_i = 0$ even though $D_{ij} = 1$. These visits with a false positive diagnosis are referred to as cases of *incorrect diagnoses*. Patients who are incorrectly diagnosed with stroke end up returning to the health system for one or more of two reasons: the underlying stroke mimic that is not yet diagnosed recurs, or the unnecessary stroke treatment results in complications, such as angioedema or intracranial haemorrhage (Buck et al., 2021). For all other visits, we set $S_i = D_{ij}$.

In the case of ED visits following which the patient doesn't revisit the health system again in the year or even in 2017, there are two possibilities: either the patient doesn't need any medical care during this period, or the patient died relatedly or unrelatedly. There are 257,061 such visits in the sample.[13] If we set $S_i = D_{ij}$ for these visits, we possibly risk underestimating the rate of missed diagnoses for visits with $D_{ij} = 0$ and underestimating the rate of incorrect diagnoses for visits with $D_{ij} = 1$. Dropping these visits from the sample, on the other hand, would most likely overestimate the rate of incorrect and missed diagnoses since the denominator would fall. We report estimates from both these approaches, but decidedly err on the side of underestimating misdiagnosis in the rest of our analysis.

The ICD-10 CM codes used to identify diagnoses of stroke, stroke sequela, and stroke mimics, are detailed under Appendix E.

II.C.3. Neuroimaging and Test for Stroke

We use the term neuroimaging to refer to the use of any diagnostic brain-imaging technology used for strokes (such as CT, CTP, MRI, and others listed under Appendix E) during the patient's visit. However, neuroimaging is done not only for the diagnosis of stroke, but also to detect traumatic brain injuries, tumors, aneurysms, and epilepsy, among other brain disorders. To obtain a stronger indicator for whether or not the physician specifically suspected stroke and sought to test for it, we define the testing variable T_{ij}. We code T_{ij} based on whether the imaging ordered for the patient aligns with the stroke protocol.

We assign $T_{ij} = 1$ if (a) the patient gets a non-contrast CT and is diagnosed with stroke,

[12] Stroke mimics are conditions or acute neurological symptoms that present with symptoms that may be erroneously attributed to stroke. For example, seizure, migraine, or hypoglycemia are some stroke mimics (Fernandes et al., 2013; Anathhanam & Hassan, 2017).

[13] The share of visits following which the patient doesn't revisit the health system again in 2016-17 is roughly 20% for non-Black patients, and 13% for Black patients regardless of whether or not they are diagnosed with stroke or TIA during the visit. The difference is therefore unrelated to strokes or its diagnosis.

or (b) the patient gets a non-contrast CT combined with a follow-up imaging modality, or gets a high-stroke-sensitivity imaging modality such as MRI, or (c) the patient is diagnosed without imaging. We set $T_{ij} = 0$ otherwise. Conditions (a) and (b) are fairly obvious given stroke protocol guidelines and the differential sensitivity of non-contrast CT for ischemic strokes. Condition (c) allows to include the minority of cases in which a patient is evaluated for stroke without the use of neuroimaging.[14] Effectively, $T_{ij} = 1$ combines the physician's two separate decisions of ordering initial imaging and follow-up imaging (if necessary), into one composite decision variable. Note here that for visits where a stroke episode gets missed because the physician failed to order follow-up imaging, T_{ij} would be assigned 0 by this rule since the patient wasn't "sufficiently" tested.

II.C.4. Facility

For each ED visit in our primary sample, we define the 'facility' at the level of the specific ED, the admission hour of the visit, an indicator for weekend visits, and the quarter of the year. Treating each ED at a specific admission hour as a different 'facility' allows us to account for variations in patient volume, personnel staffing, availability of equipment or technicians, and other factors that may be relevant for diagnosis. We index facilities by the subscript f.

III. CONCEPTUAL FRAMEWORK

The diagnostic pathway for stroke can be summarized broadly as the physician making three key decisions sequentially: assessing the stroke risk for the case, deciding whether to test for stroke, and making the final diagnosis.

Stroke Risk Assessment	$\longrightarrow$	Neuroimaging / Testing	$\longrightarrow$	Diagnosis
$\pi_{ij} \in (0,1)$		T_{ij}		D_{ij}

The data permits us to directly observe the second and third of these, as described in Section II, but not the first. Our analysis of disparity must therefore rest on the testing and diagnostic decisions alone.

III.A. Disparity in Outcomes

The 'outcome' that we focus on is the *quality* of physicians' decisions. For binary decisions such as stroke diagnosis $D_{ij} \in \{0, 1\}$, the quality of the decision can be judged based on whether or not it matches the underlying latent stroke state $S_i \in \{0, 1\}$. False negatives (with $D_{ij} = 0$, $S_i = 1$) are *missed diagnoses*, and false positives (with $D_{ij} = 1$, $S_i = 0$) are *incorrect diagnoses*.

Racial disparity in the quality of diagnosis is the difference in diagnosis rates between Black

[14] For about 94% of the stroke diagnoses made in the sample, neuroimaging had been done. Diagnosis without testing may be done for patients who are contraindicated for neuroimaging but have fairly obvious stroke presentations and need immediate treatment.

($R_i = b$) and non-Black ($R_i = w$) patients who visit the ED with the same latent state.

$$\Delta_S^D \equiv \mathbb{E}\left[D \,|S, R = b\right] - \mathbb{E}\left[D \,|S, R = w\right] \tag{1}$$

For instance, $\Delta_{S=1}^D < 0$ means that the rate of missed diagnosis among Black patients is higher than among non-Black patients, and $\Delta_{S=0}^D > 0$ means that the rate of incorrect diagnosis is higher among Black patients than among non-Black patients. Likewise, $\Delta_{S=1}^D > 0$ and $\Delta_{S=0}^D < 0$ indicate a higher rate of missed diagnosis and a higher rate of incorrect diagnosis, respectively, among the non-Black patients instead.

Considering the sequential nature of physicians' decisions in the stroke diagnostic pathway, disparity in any of the physician preceding decisions of stroke assessment or testing contributes subsequently to disparity in diagnosis. Section IV estimates the racial disparity in the quality of stroke diagnoses delivered in emergency departments as well as disparity in testing decisions. To trace the origin of disparity along the sequence of physicians' decisions, we then test for disparity in the quality of diagnosis conditional on getting tested.

Even so, racial disparity in the quality of diagnosis or testing decisions are examples of disparity in *outcomes*, and cannot be conclusively inferred to stem from *disparate treatment* by physicians.[15] For example, the rate of missed diagnoses (when $S_i = 1, D_{ij} = 0$) may be higher for a group that systematically selects into EDs with a lower quality of care, despite EDs treating all groups equally—then there would be a disparity of outcomes, but not in treatment. Disparity in the quality of stroke diagnoses, therefore, represents the joint impact of all factors relevant to diagnoses. These factors may either be inherently different by race or effected by discriminatory practices at various points of the healthcare delivery system.

Disparate treatment can be determined only if racial differences in physicians' decision rates, of either diagnosis or testing, persist even after we account for *all* underlying race-correlated aspects medically relevant to the decision. Further, quantifying the role of disparate treatment informs of its contribution to the overall disparity in outcomes relative to other contributing factors such as inequitable access to care, variation in health behaviors, or social determinants of health. While race-correlated aspects such as insurance status, age, sex, co-morbidity profile, as well as facility and physician fixed effects, can be controlled for, the empirical challenge is in accounting for underlying statistical differences in stroke risk and any structural differences in the quality of information available based on which the physician is expected to infer stroke risk, as discussed next.

III.B. Disparate Treatment

The physician's decision to test a patient for stroke is based on whether or not they *think* the patient has stroke. Effectively, the physician first assesses the stroke risk for the case i.e. the probability that this patient may have stroke; and then orders a test if the risk assessment is high enough that it exceeds some threshold. We don't observe the risk assessments that physicians make for each patient, but any error made in assessing risk also translates to incorrect testing decisions, and in turn to poor quality of the final diagnoses.

[15] "Treatment" here refers to physicians' behavior towards the patient or their handling of the case, and not the remedial administration of medicine, therapy, or surgery.

Unlike disparate outcomes, we cannot define disparity in treatment by conditioning on the the latent state S_i. This is because the physician doesn't observe S_i, but must infer it based on the symptomatic presentations of the patient and the patients' history and risk factors. Stroke presentations with non-traditional or atypical symptoms such as generalized weakness, dizziness, or altered gait may be harder to diagnose and more likely to be missed in the ER (Lever et al., 2013). Underlying racial differences in the symptomatic presentations of stroke or in the symptom reporting behavior of patients can therefore affect the quality of information available to the physician and drive differential rates of testing by race. If, for example, stroke presented with diffuse and general symptoms such as headache or nausea in one group of patients as opposed to more typical symptoms such as slurred speech or facial drooping in another, inferring stroke for the former will be harder despite the two groups of patients sharing the same latent state.

An inference of disparity in treatment across racial groups must hence be made conditional on the probability of the visit being a stroke episode *given the information available to the physician*. Let $\mathcal{I}_i$ denote the information on visit i that is available to the physician. We define $P_i = \mathbb{P}(S_i = 1|\mathcal{I}_i)$ to be the *objective stroke risk assessment* for visit i. It represents the "best" stroke risk assessment that an unbiased or racially objective physician can make for the visit.

Any differences in the quality of information available to the physician, either due to different symptomatic presentation of stroke, behavioral differences in the reporting of symptoms, or even in the environmental risk factors indicated by the socio-economic status of the patient, are all reflected in the physician's information set $\mathcal{I}_i$. Basing cross-race comparisons conditional on P_i should therefore not only exclude the effect of all statistical differences by race, but also any structural differences in the quality of information available. Even though the distribution of stroke risk for patients in one group may be different from that of patients in another, at any given level of P_i the expected yield from diagnostic testing is identical and hence the patients must be tested with the same propensity. If they aren't, then it is disparate treatment.[16]

If P_i were known, disparity in treatment could be estimated using the following specification:

$$T_{ij} = \delta Black_i + \gamma_{P_i \text{ quantile}} + \gamma_{\text{facility(i)}} + \upsilon_{ij} \tag{2}$$

where $T_{ij} = 1$ if the physician tested the patient for stroke, and $Black_i = 1$ if $R_i = b$. Here, δ captures the differential testing of Black patients relative to others despite the same level (or range) of objective risk. The additive form in the specification assumes implicitly that the quality of care available *within* any facility is comparable across race. With our granular definition of a facility and the quasi-random assignment of physician, this assumption is not unreasonable. If $\delta \neq 0$, we can conclude that there is disparity in treatment of patients on the basis of their race when it comes to making testing decisions. Estimating the disparity in testing separately for different risk quantiles will additionally describe how this disparate treatment varies with the level of stroke risk.

[16]Note though that this definition of disparate treatment is defined only over the common support of the objective risk distributions of the two racial groups.

Physician-specific estimates of disparate treatment can likewise be estimated using the following regression specification

$$T_{ij} = \sum_j \theta_j Z_{ij} + \sum_j \delta_j Z_{ij} Black_i + \gamma_{P_i \text{ quantile}} + \gamma_{\text{facility(i)}} + \upsilon_{ij} \tag{3}$$

where Z_{ij} are indicators of physician assignment. $Z_{ij} = 1$ if physician j is assigned to visit i, and is 0 otherwise. The coefficient δ_j captures racial disparity in testing decisions made by physician j for patients who visit the same facility and have the same stroke risk quantile. We can interpret δ_j as physician j's disparate treatment by race, as long as υ_{ij} is uncorrelated with $Z_{ij} Black_i$.

The challenge in estimating disparity in treatment is that the objective risk P_i is unknown. More importantly, the information set $\mathcal{I}_i$ that the objective stroke risk assessment is based on, may not be completely observed by the econometrician. We therefore proceed as follows. We begin with the information that is in fact available to the econometrician, denoted by $\tilde{\mathcal{I}}_i$, where presumably $\tilde{\mathcal{I}}_i \subset \mathcal{I}_i$. It is a rich set which, as already detailed, contains information ranging from symptoms at presentation to all co-occurring conditions, but possibly excludes some features that might be available to the physician on the site, such as patient demeanor. We use the information $\tilde{\mathcal{I}}_i$ to train a machine learning algorithm that predicts stroke risk, generating a proxy $\hat{P}_i$ for unobserved objective risk assessment P_i. That is, $\hat{P}_i = \mathbb{P}(S_i = 1|\tilde{\mathcal{I}}_i)$.

Section III.C describes in detail the algorithm design used to obtain $\hat{P}_i$ and how we avoid algorithmic bias in our predictions. Importantly, the inclusion of race as a feature in the machine learning algorithm ensures that the algorithm identifies all patterns in the data that relate $Black_i$ to the target S_i—whether directly or indirectly (via unobservables). We have

$$\hat{P}_i = \mathbb{P}(S_i = 1|\tilde{\mathcal{I}}_i) = \mathbb{E}\left[\mathbb{E}\left[S_i|\mathcal{I}_i\right]|\tilde{\mathcal{I}}_i\right] = \mathbb{E}\left[P_i|\tilde{\mathcal{I}}_i\right]$$

Since $Black_i$ is included in (or measurable with respect to) $\tilde{\mathcal{I}}_i$, any additional information from unobservables that improves the risk prediction cannot be systematically related to patient race, i.e. $P_i - \hat{P}_i$ must be uncorrelated with patient race. This is only true if the algorithm is trained on a random non-selective sub-sample, esnsuring that the relationship between race and any unobservables is the same in the training sample as it is in the test sample. Note that we do not argue that machine learning predictions are equal to the objective risk despite being trained only on observables; but that the deviations would be similar on average for both the racial groups. In the event that certain unobservables were more informative in terms of predicting stroke risk for only one of the racial groups, the machine learning predictions for the other group would be noisier. But if such a symptom or characteristic that doctors could see had really been consistently used in assessing stroke risk, it would have been given an alphanumeric code, entered to the database of ICD-10 diagnosis codes, and included in our data.

Later, in Section V, we use the machine learning predictions again to separately quantify the mechanisms driving the estimated disparate treatment in testing.

III.C. Stroke Risk Benchmarking using Machine Learning

To facilitate between-race comparisons of physician decisions, we defined P_i as the racial objective conditional probability of stroke given the information available with the physician at the time of the visit i. Effectively, P_i already accounts for underlying differences in the prevalence and presentation of stroke and hence can serve as an objective benchmark against which physicians' decisions can be compared. Since P_i is unknown, we obtain predictions of the objective risk, denoted by $\hat{P}_i$, using machine learning.

III.C.1. Algorithm Design

To obtain $\hat{P}_i$, we train a machine learning algorithm to build a stroke risk predictor. The algorithm takes as input, data that would have been available to the physician at the time they were making testing decisions. It includes patient characteristics such as sex, age, race, stroke hospitalization rates and death rates in the patient's ZIP; and visit-specific information such as the visit reason, symptoms at the time of the visit, external causes of morbidity, patient vitals, and any co-occurring illnesses. All this information is taken from the patient record of the current visit, and from the records of the patient's previous care-seeking encounters in the year. We don't blind the algorithm to patient race so that the algorithm can learn from any systematic race-correlated patterns in the symptomatic presentation of stroke or in the quality of chart documentation (arising either from differences in patient reporting behavior or provider bias during documentation).

The target variable that the algorithm is supervised to learn to predict is the latent stroke state (S_i) and transient ischemic attack (TIA). The inclusion of TIA in our target is due to two reasons. First, our definition of a missed stroke diagnosis also includes cases of undiagnosed TIA that eventually recur as full blown stroke. Second, and more importantly, TIA presents with similar symptoms as an acute stroke. Thus, from the physician's point of view, the prediction of S or TIA is an identical prediction problem.

We train the predictor model using an ensemble method of extreme gradient boosting, also called XGBoost (Friedman, 2001; Chen & Guestrin, 2016). It begins with simple decision trees or base learners to make state predictions,[17] and then sequentially generates and combines new base learners to incrementally improve upon the predictive performance of previous learners using gradient descent on the specified objective. By focusing on misclassified instances when training the new base learners, the algorithm forces the model to improve its predictive accuracy even on the challenging samples. The ensemble model thus obtained captures all underlying patterns in the data, however complex and non-linear, when making predictions. XGBoost's strength is that combines sequential learning with parallel computation in the construction of individual learners for efficiency.

[17] Decision trees start at a root node containing the entire dataset. The algorithm then chooses an input feature and splits the data into two subsets based on a condition defined on the feature. At each of the new nodes created, the algorithm then chooses another feature and splits the subsets further. This process continues until it reaches a stopping condition. At each of the final nodes, the algorithm then determines the prediction for all the data points in that node.

We train the algorithm on a random sixty percent of the primary sample, and then use the model to predict stroke risk for visits in the remaining forty percent of the *test sample* that are unseen by the algorithm.

We avoid algorithmic bias in the following ways. First, the algorithm uses as target the latent stroke state that we eventually want to predict, instead of a proxy.[18] Importantly, the target variable is inferred retrospectively and are not based on potentially biased physicians' decisions. For example, if physicians selectively misdiagnosed patients of one race, any algorithm trained on physicians' diagnoses would also mispredict risk for that race. Algorithms that otherwise learn from past discriminatory human decisions, or that which are trained on select data generated as a consequence of potentially biased decisions, typically reproduce the bias in their predictions; as is summarized by the popular phrase "bias in, bias out". Second, since the latent stroke states are non-selectively known, the random train-test split ensures that we train our algorithm on a random sample of patient visits regardless of whether they were tested or diagnosed for stroke. This allows for the joint-distribution of observables and unobservables to be different among patients who are diagnosed from among those who aren't. Third, the inclusion of race as an input feature allows the algorithm to identify all patterns that relate patient race to the target—whether directly or indirectly via unobservables. Appendix C discusses additional considerations in our algorithm design.

The random split between training and test data is target label-stratified. Label imbalance of minority stroke states is improved by synthetic minority oversampling (SMOTE) such that positive latent stroke states constitute at least 5 percent of the training data. The features used as input include: patient characteristics, one-hot encodings of diagnoses codes (ICD-10-CM) representing co-morbidities, symptoms, health risk factors, and external causes of morbidity, vector embeddings of the ICD-10-CM code listed as the visit reason,[19] and frequency counts for each ICD-10-CM code listed in any of the patients' previous visits. As part of feature selection, ICD-10-CM codes listed in fewer than 0.05 percent of the visits for both stroke and non-stroke states are excluded. A total of 1,729 features is finally used for training.

Since we are interested only in probabilistic predictions, we train the algorithm to minimize log loss or cross-entropy loss. Hyperparameters are tuned with stratified 3-fold cross-validation to maximize the area under the receiver operating characteristic curve (ROC AUC),[20] and the F1-score.[21] The probability predictions are then calibrated by fitting a non-parametric isotonic regressor.

[18] Obermeyer et al. (2019) describe one such algorithmic bias born out of using medical expenditures as the proxy target variable for predicting health needs. An algorithm trained on medical expenditures as a proxy for health needs would under-predicted the needs of Black adults because of their lower health spending overall. As a consequence, the algorithm failed to target Black adults for enrollment in high-risk care-management programs.

[19] Embeddings are vector representations of non-numeric values/objects. We use 10-dimensional embeddings for ICD-10-CM codes from Kane et al. (2023) that are generated using BioGPT Large Language Model.

[20] Receiver operating characteristic (ROC) curve plots the true positive rate (TPR) against the false positive rate (FPR) for each threshold used for classification.

[21] F1 score is the harmonic mean of the precision and recall of the binary classification.

III.C.2. Evaluating the Stroke Risk Predictions

The algorithm trains a model that predicts the likelihood of the target variable when given any input data. We use the trained model to predict stroke risk for the visits in the remaining forty percent of the primary sample. Appendix Figure A1 plots the distribution of the ML predicted stroke risks by patient race and the latent stroke state. Note that the supports of the predicted stroke risks are largely overlapping for the two racial groups, thus making our comparisons that are conditional on predicted risk a valid exercise.

The quality of the probabilistic stroke risk predictions made by the ML trained model can be evaluated based on how well the predicted stroke risk probabilities match the true share of stroke incidences, i.e. by evaluating whether the probability predictions are well-calibrated or not. Appendix Figure A2 plots the calibration curve for patient visits in the test data—both overall, and separately by patient race. The calibration curves in each case are close to the 45 degree identity line, suggesting that our risk predictions do in fact reflect true stroke likelihoods on average. Importantly, the algorithm does not systematically over-predict or under-predict stroke risk for patients of either race.[22] In Appendix Figure A3, we also present the calibration curves for visits where a diagnosis of stroke is made, separately from visits where it is not. The figure illustrates that the algorithm makes reasonable risk predictions regardless of the physician diagnosis.

Another way of gauging the performance of the trained model is to check the quality of its binary predictions of the target. The performance of a binary classifier is typically measured by the area under the receiver operating characteristic curve (ROC AUC) that represents the classifier's ability to discriminate between the positive (stroke) and negative (non-stroke) labels when thresholds are varied. ROC AUC ranges from 0 to 1, where random guessing would produce a score of 0.5, and perfect prediction would result in 1. Our ML stroke risk predictor has an ROC AUC of 0.79 which would be considered reasonably good. The ROC AUC for non-Black patients is 0.79, which is comparable to the 0.74 for Black patients.

It is also relevant for our analysis to check if physicians' decisions agree with the algorithmic risk predictions. We verify this by testing if physicians' decision rates of testing and diagnosis are increasing in predicted risk. Appendix Figure A4 plots the proportion of patients for whom physicians order neuroimaging and whom physicians diagnose with stroke, at varying deciles of the algorithmic stroke risk prediction. The decisions rates are increasing in the predicted risk confirming that physicians' perceptions of stroke risk are broadly consistent with the algorithmic risk predictions.

Lastly, we compare distributions of machine learning predicted risk for visits that were misdiagnosed. As shown in Appendix Figure A5, stroke episodes missed in the ED were mostly those with low predicted stroke risk. Visits for which the algorithm infers high levels of risk, were indeed correctly diagnosed as stroke. Similarly, non-stroke episodes that are incorrectly diagnosed as stroke typically have high levels of predicted stroke risk.

[22] We also confirm that the standard error in the algorithmic risk predictions is low (~ 3 percentage points) and comparable by race. We measure it as the standard deviation in predictions made across individuals trees in our ensemble model.

When training the algorithm, we included patient race as an input feature so that the algorithm can learn from any underlying racial differences in symptomatic presentations for stroke and non-stroke visits. To check if indeed patient race is relevant in predicting stroke risk given clinical symptoms, we look at the importance of race as a feature in the model algorithm. Race is the eighth most importance feature according to Shapley value—a measure of each feature's contribution in the prediction of an instance. Based on gain, another metric of feature importance that captures the improvement in the accuracy at all tree-based nodes where the feature is used to split, race is one of the twenty five most important features.

III.D. Quasi-Experimental Assignment of Physicians

An important assumption in our empirical estimation of physician-specific parameters of disparate treatment is that within each facility the assignment of physicians to cases is random. In other words, conditional on the patient visiting an ED at a particular hour of a specific day in the week during a given quarter, physicians' decisions rates and the quality of these decisions should be independent of patient characteristics.

Quasi-experimental assignment is a plausible assumption in our setting because after patients are triaged, physician assignment is typically based on availability conditional on shift schedules. Additionally, the shift and staffing schedules of attending physicians are made separately from those of triage nurses, radiologists, or other laboratory technicians. We can therefore study the effects of physicians' decisions separately from its joint effects with other heath workers or lab/imaging services. At the very least, we can interpret the effects of physicians' decisions as the physicians' average effect taking as given the other healthcare staff on duty. Even if a healthcare worker were tied specifically to a unique physician in the facility, we would effectively be examining the decisions of a 'care-giving unit'.[23]

To validate the plausibility of quasi-random assignment, we first calculate leave-out measures of stroke diagnosis rate, rate of false negatives in stroke diagnosis, rate of false positives in stroke diagnosis, and the propensity to test for stroke—averaged across all *other* patients seen by the attending physician. We then test how well patient characteristics predict these leave-out measures *within* each facility, using OLS regressions. All continuous patient covariates are standardized. Appendix Figure A6 shows the coefficients from our regressions of various leave-one-out measures on patient covariates. The panels also report the respective F-statistic and p-value of the joint F-test of all patient covariates. We find all coefficients to be small and not significantly different from zero statistically. We fail to reject the null of quasi-random assignment at conventional levels of statistical significance only in a few specifications though.

[23]HCUP SEDD for Florida specifies up to three physician identifiers per visit. In our primary sample, a unique physician is listed for about 87.35% of the visits. For visits diagnosed with stroke in particular, the equivalent shares disaggregated by race are 86.24% for Black patients and 84.45% for non-Black patients.

IV. DISPARITY IN OUTCOMES AND IN TREATMENT

In this section, we document racial disparity in two outcomes: the quality of stroke diagnosis, and testing rates. Considering the sequential nature of physicians' decisions, with testing decisions preceding diagnosis, any disparity in testing would also translate to disparity in diagnosis. Therefore, to then trace the origin of disparity along the sequence of physicians' decisions, we first examine disparity in diagnosis decisions and then move backwards to examine testing decisions.

IV.A. Disparate Outcomes

IV.A.1. Disparity in the quality of stroke diagnosis and testing

Disparity in diagnosis quality is estimated using the following specification that compares diagnosis rates across race, given a latent stroke state.

$$D_{ij} = const + \beta_1 S_i + \beta_2 Black_i + \beta_3 S_i Black_i + error_{ij} \tag{4}$$

Here, β_2 records the racial difference in diagnosis rates for visits with latent stroke state $S_i = 0$ (i.e. false positives), and β_3 records the racial difference in diagnosis rates for visits with latent stroke state $S_i = 1$ (and hence the difference in false negatives). Table 2 specification (1) reports the estimates. While 16% of the stroke episodes among non-Black patients are missed in the ED, roughly 28% of stroke cases are missed for Black patients—slightly short of twice as higher.[24] The rate of false positives i.e. non-stroke cases being incorrectly diagnosed as stroke is very low (about 0.03 %) with no significant difference by race. Racial disparity in the quality of stroke diagnosis is hence driven mostly by the large gap in the rate of missed diagnoses.[25] The estimated disparity is also robust to the choice of time interval used in inference of the latent stroke state; as is shown in Appendix Table B1 which reports the estimates of racial disparity when latent stroke states are inferred from patient revisits over 10-days, 14-days, 20-days, and 30-days since the initial encounter.

To rule out the possibility that the disparity in quality of diagnosis is driven solely by underlying differences between the two race groups, specification (2) in Table 2 revisits the comparison by including facility fixed effects and other controls such as the patients' sex, age, insurance status, income quartile of zip in state, and the co-morbidity profile of the patient quantitated by the Charlson Comorbidity Index. Our estimates of the disparity in diagnosis quality are robust to these controls, and also to physician fixed effects included under specification (3) in Table 2. The robustness of the estimated racial disparity to facility and insurance level controls strengthens our motivation of examining disparity arising from physician's decisions.

With a non-linear specification such as the logistic, the odds of a missed diagnosis are about 2.7 times higher for Black patients; see Figure 1. This difference by race is separate from the

[24]The rate of missed diagnoses among White patients specifically, is also 16%, and that among non-White patients is around 23%—about 1.5 times higher.

[25]In terms of raw counts, the number of visits that are missed (false negatives) are roughly 4 times as large as the count of visits that are incorrectly diagnosed as stroke (false positives) for Black patients. The ratio is about 2.5 for non-Black patients.

TABLE 2 :
Racial disparity in the quality of stroke diagnosis

	Linear Probability Model for Stroke Diagnosis ($D = 1$)		
	(1)	(2)	(3)
Latent Stroke State ($S = 1$)	0.8395***	0.8373***	0.8355***
	(0.005)	(0.005)	(0.005)
Black × Latent Stroke State ($S = 1$)	−0.1202***	−0.1211***	−0.1205***
	(0.014)	(0.013)	(0.013)
Black	−0.0000***	0.0002***	0.0002***
	(0.000)	(0.000)	(0.000)
Constant	0.0003***		
	(0.000)		
Observations	1,368,560	1,367,438	1,360,817
Facility FE		×	×
Controls		×	×
Physician FE			×
Adjusted/Within R^2	0.756	0.754	0.752

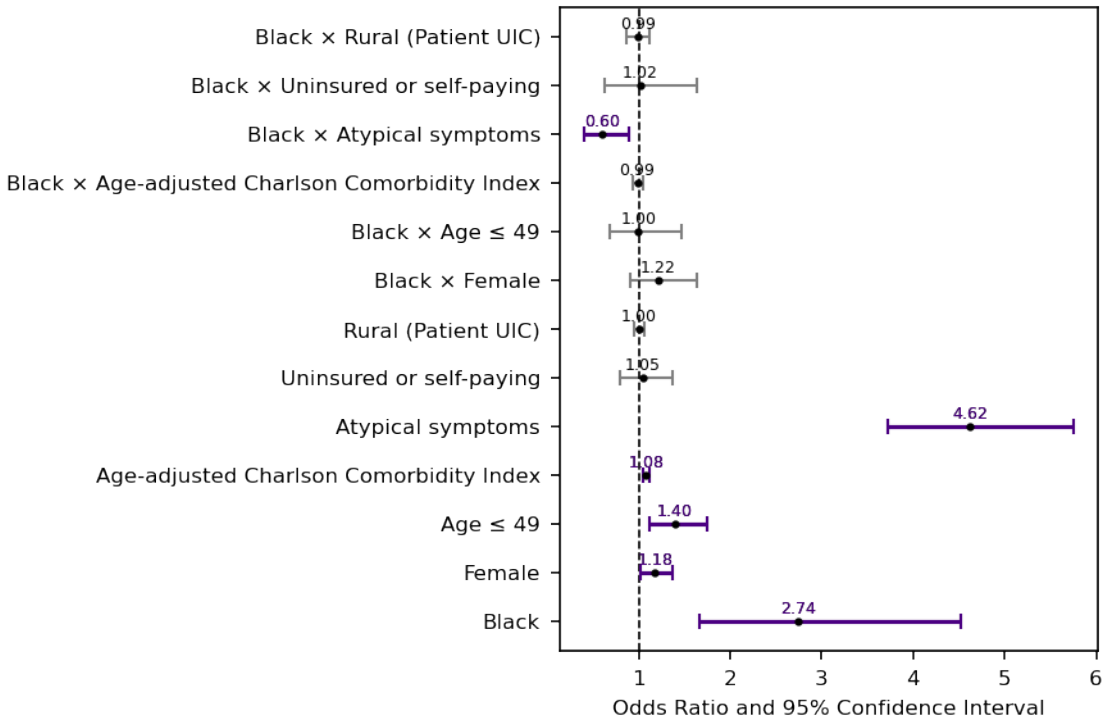
*p<0.1; **p<0.05; ***p<0.01

Notes: This table reports the estimates of racial disparity in the quality of stroke diagnosis based on the specification in Equation (4). Columns (2) and (3) subsequently add controls to this specification. Controls for columns (2) and (3) include: patients' age and sex, income quartile of patient's zipcode in state, the primary expected payer (insurance), the Charlson Comorbidity Index at the time of the visit, and facility fixed effects, where a facility is defined at the level of the specific ED, the quarter of the visit, indicator for weekends, and the admission hour. Additionally, column (3) also includes fixed effects for the attending ED physician. Heteroskedasticity-consistent MacKinnon and White (1985) HC3 standard errors are reported in parentheses.

one mediated by other factors potentially correlated with race, such as sex, age, insurance status, or the symptomatic presentation. To account more carefully for the differences in quality of insurance coverage, we also estimate disparity in diagnosis quality for patients covered by Medicare and Medicaid in Appendix Table B2, and consistently find equally large differences in missed diagnosis rates by patient race. Our findings are also robust to the inclusion of visits by patients who don't make any subsequent visits during year; see Appendix Table B3. We consistently find the missed diagnosis rate to be almost twice as large for Black patients under all these specifications.

These diagnostic errors are costly. Stroke patients who are missed at their initial presentation

FIGURE 1 :
Odds ratio for missed stroke diagnosis given patient characteristics



Notes: This figure shows the odds ratios and 95% confidence intervals from a logistic specification for a stroke episode being missed in the ED. The error bars indicate the 95% confidence intervals.

make roughly 2.5 additional visits within the year on average, costing an excess of about $\$78,263$ over these future visits (excluding professional fees and non-covered charges). In the case of non-stroke visits incorrectly diagnosed as stroke, patients make 0.6 additional visits on average, costing roughly $\$161,073$ more.

Having documented a racial disparity in the quality of diagnosis, we proceed backwards to examine testing (or neuroimaging) decisions. Table 3 compares the rate at which neuroimaging is ordered, disaggregated by patient race and the underlying latent stroke state. Neuroimaging is ordered for roughly 30% of the visits in our primary sample. The neuroimaging rate is however lower for Black patients by a statistically significant difference of 3.5 percentage points. Specifically among patients who present with stroke, Black patients are roughly five percentage points less likely to get any diagnostic neuroimaging at all. Most ($\sim$ 94%) of the visits that get neuroimaging are ordered a non-contrast computed tomography (CT), which is the primary imaging modality recommended for the evaluation of patients with suspected stroke. Due to low sensitivity of non-contrast CT for ischemic strokes, physicians that strongly suspect stroke order follow-up imaging such as contrast CT or MRI. Once again, we note that there is a racial disparity in how often this follow-up imaging is ordered for the ischemic stroke-type. Black patients presenting with an ischemic type are about eight

percentage points (roughly 20 percent) less likely to get additional follow-up imaging; see Table 3. On the hand, for patients presenting with a hemorrhagic stroke, follow-up imaging is ordered only 9% of the times, and the difference by race is not statistically significant.

TABLE 3 :
Neuroimaging rates, by patient race

	Non-Black	Black	Difference
	(1)	(2)	(1) - (2)
Panel A			
Any neuroimaging at all			
for all	0.2959	0.2613	0.0346***
	(0.000)	(0.000)	
for visits with latent stroke state $S = 1$	0.8808	0.8334	0.0474***
	(0.004)	(0.011)	
for visits with latent stroke state $S = 0$	0.2929	0.2593	0.0336***
	(0.000)	(0.000)	
Follow-up imaging, in addition to non-contrast CT			
for visits with latent stroke state $S = 1$			
of Ischemic[§] type	0.3521	0.2734	0.0787***
	(0.009)	(0.017)	
of Hemorrhagic[§] type	0.0926	0.1124	-0.0198
	(0.007)	(0.017)	
Panel B			
Testing for Stroke ($T = 1$)			
for all	0.0385	0.0250	0.0135***
	(0.000)	(0.000)	
for visits with latent stroke state $S = 1$	0.8510	0.7364	0.1146***
	(0.005)	(0.012)	
for visits with latent stroke state $S = 0$	0.0343	0.0224	0.0119***
	(0.000)	(0.000)	
Two-sample t-test			*p<0.1; **p<0.05; ***p<0.01

Notes: Panel A reports differences in rates of initial and follow-up neuroimaging by patient race. Panel B reports the differences in the composite decision variable 'Test for Stroke' or T_i that combines these two decisions, as described in Section II.C.3. Standard errors are in parenthesis. [§]Stroke type is as identified upon the patient's revisit when the stroke gets diagnosed.

As a summary, in Panel B of Table 3, we compare rates of testing based on our composite variable of stroke testing T_{ij} that combines the two decisions of initial neuroimaging and follow-up imaging, as described in Section II.C.3. We find that Black patients are less likely to be tested for stroke in general, and in the case of stroke patients in particular, the difference is of 11.46 percentage points, corresponding roughly to a shortfall of 13.5% for Black patients.

IV.A.2. Contribution of disparity in testing to the disparity in missed diagnosis rates

So far we have documented that the rates of stroke diagnosis and that of testing differ significantly across race. Considering the sequential nature of physicians' decisions, with testing decisions preceding diagnosis, any disparity in the quality of physicians' testing decisions also contributes subsequently to disparity in diagnosis. Alternately, there may also be factors that contribute to disparity in diagnosis *after* testing decisions are made. For example, it may be that testing modalities are differently accurate for some patients, or that attending physicians or radiologists discriminate when reading or interpreting the test scans. Consequently, it is critical to quantify the extent to which testing disparities contribute to disparity in the quality of diagnosis. If disparity in testing rates account for most of the disparity in diagnosis, then the relevant decision of the physician that we must examine for disparate treatment should be that of testing.

Table 4 reports the rates of missed diagnoses (or false negatives) conditional on testing status, by patient race. For hemorrhagic stroke types—that non-contrast CT scans do detect with high sensitivity—missed diagnosis rate is around 6-8% with no significant difference by race (Table 4). Ischemic strokes, on the other hand, have higher missed diagnosis rates if non-contrast CT is the only modality used and no additional imaging is ordered. Missed diagnosis rates for ischemic sub-types is significantly higher for Black patients—with a difference of about 12 percentage points—possibly due to differential rates of follow-up imaging. Among patients who do receive follow-up imaging and those who are screened by high-sensitivity modalities other than CT, the difference in missed diagnosis rates by race is insignificant. These findings make clear that there is no significant racial disparity in diagnosis *after* a patient is appropriately tested for stroke.

To then quantify the role of disparity in testing in driving the disparity in diagnosis, we do components analysis as detailed in Kitagawa (1955). By representing aggregate missed diagnosis rates as a weighted sum of the testing-status-specific missed diagnosis rates, we can decompose the aggregate difference by patient race into the effects of differential testing and disparity in testing-status-specific missed rates. The Kitagawa (1955) decomposition calculates counterfactual missed diagnosis rates for Black and non-Black patients had testing rates and missed diagnosis rate conditional on testing, been separately equalized across the two racial groups. When any component is equalized across groups, it is set equal to the average value of the component between the two groups. Difference in counterfactual missed diagnosis rates when only one component is allowed to vary, captures how much of the total difference in aggregates can be attributed to that component. Table 5 reports the counterfactual rates from this method.

TABLE 4 :
Missed diagnosis rates, by patient race and testing status

	Missed Diagnosis Rate		
	Non-Black	Black	Difference
Panel A: Crude rates of missed stroke diagnoses			
	0.1602	0.2805	−0.1203***
	(0.005)	(0.013)	
Panel B: Missed stroke diagnoses conditional on testing			
For stroke patients who get any neuroimaging			
all§ types	0.1032	0.1913	−0.0881***
	(0.004)	(0.012)	
Ischemic§ type	0.1363	0.2538	−0.1175***
	(0.006)	(0.016)	
Hemorrhagic§ type	0.0618	0.0831	−0.0213
	(0.006)	(0.015)	
For stroke patients who get non-contrast CT with follow-up imaging, or get MRI			
Ischemic§ type	0.0527	0.0766	−0.0238
	(0.007)	(0.018)	
Panel C: Missed stroke diagnoses conditional on 'Test for Stroke'			
when $T_{ij} = 1$	0.0131	0.0229	−0.0097**
	(0.002)	(0.005)	
Two-sample t-test			*p<0.1; **p<0.05; ***p<0.01

Notes: Panel A reports crude missed diagnosis rates, and Panel B reports them conditional on initial and follow-up testing. Panel C reports missed diagnosis rates conditional on the composite decision variable 'Test for Stroke' or T_{ij} that combines the two decisions as described in Section II.C.3. Standard errors are in parenthesis. §Stroke type is as identified upon the patient's revisit when stroke diagnosis is made.

We find that a sizable 93.5% of the observed disparity in the quality of diagnosis can in fact be explained by differences in stroke testing. The remaining 6.5% can be attributed to other factors such as, say, racial differences in composition of strokes of different types or etiology that can be detected on easily on CT, or in the likelihood that stroke is suspected or diagnosed even before the physician orders the test.

TABLE 5 :
Decomposition of racial difference in missed stroke diagnosis

	Missed Diagnosis Rate			
	Non-Black (1)	Black (2)	Difference (1)-(2)	Share of total
Crude Missed Diagnosis Rate	0.1602 (0.005)	0.2805 (0.013)	−0.1203***	100%
Counterfactual Missed Diagnosis Rates				
Crude missed diagnosis rate is the sum of stroke testing status (T) and respective test-status specific missed rates. If these components were equalized for the two racial groups, and only the				
stroke testing rates were unequal	0.1643	0.2769	-0.1126	93.5%
category-specific miss rates were unequal	0.2167	0.2245	-0.0078	6.5%
Two-sample t-test			*p<0.1; **p<0.05; ***p<0.01	

Notes: This table reports the aggregate missed diagnosis rates by patient race, and their corresponding counterfactual values based on the Kitagawa (1955) decomposition.

With more than ninety percent of disparity in missed diagnosis stemming from disparate rates of testing, the relevant decision to examine for disparate treatment is that of testing—which we turn to next.

IV.B. Disparate Treatment

Disparate treatment in testing occurs when the physician tests patients differently by race despite the same statistical risk of stroke. Assessment of stroke risk, however, relies on the quality of the information available to the physician, which in turn depends on the symptomatic presentation, patients' reporting behavior, and on how well the patients' medical history is recorded—all of which might vary by race. Therefore, we estimate disparate treatment by comparing testing rates for patients with the same predicted objective risk of stroke, where these risk predictions are made *conditional on the information available to the physician at the time*, whatever its quality. Equation (2) gives the specification we use to estimate disparate treatment. Table 6 reports the coefficient δ, which represents the average disparate treatment within any facility and given a decile of predicted stroke risk.

Relative to the 3.86 percent of non-Black patients who are tested for stroke, the testing rate conditional on the same level of objective stroke risk is, on average, 0.88 percentage points lower for Black patients (i.e. about 23 percent lower). Of the simple difference in stroke

TABLE 6 :
Disparate treatment by race in physicians' testing decisions

	Linear Probability Model for Test for Stroke ($T = 1$)			
	(1)	(2)	(3)	(4)
Black	−0.0088***	−0.0090***	−0.0107***	−0.0123***
	(0.001)	(0.001)	(0.003)	(0.003)
Sample	All§	No contra-indications	Predicted risk ≥ 1%	Predicted risk ≥ 1% and no contraindications
Mean (T) for non-Black patients	0.0386	0.0379	0.1219	0.1202
Observations	545,116	523,365	78,212	73,039
Facility FE	X	X	X	X
Risk Decile FE	X	X	X	X

*p<0.1; **p<0.05; ***p<0.01

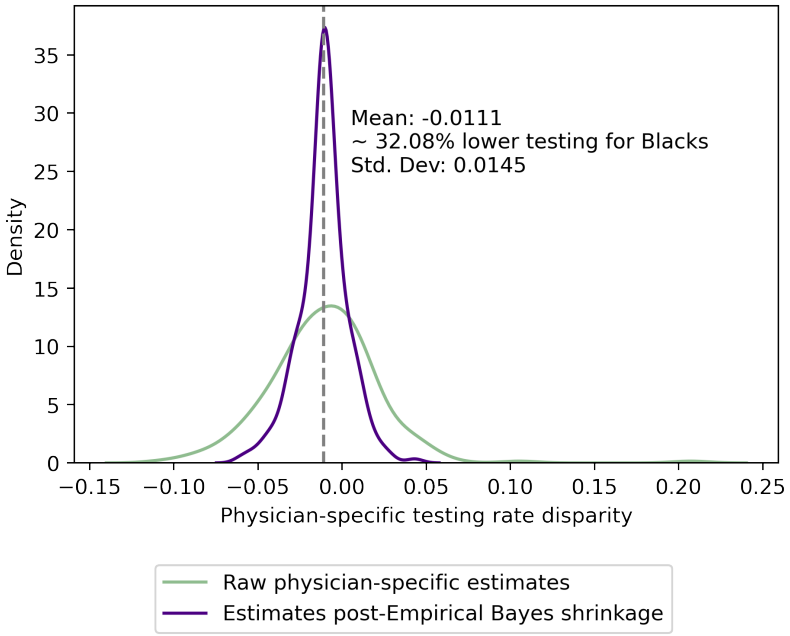
Notes: This table reports estimates of disparate treatment, or δ from the specification given in Equation (2). §The estimates are based on visits in the test data i.e. the 40% of the primary sample unused in training the machine learning algorithm. The ICD-10-CM codes that are taken as contraindications to neuroimaging are listed in Appendix E. Heteroskedasticity-consistent MacKinnon and White (1985) HC3 standard errors are reported in parentheses.

testing rates by patient race, as recorded in Panel B of Table 3, disparate treatment accounts for about sixty five percent of it. Even among patients with no contraindications for any neuroimaging modality, stroke testing is 24% lower on average for Black patients within any facility and risk decile. Across various risk deciles, racial disparity is testing is significant only in the first decile of predicted stroke risk; see Appendix Figure A7.

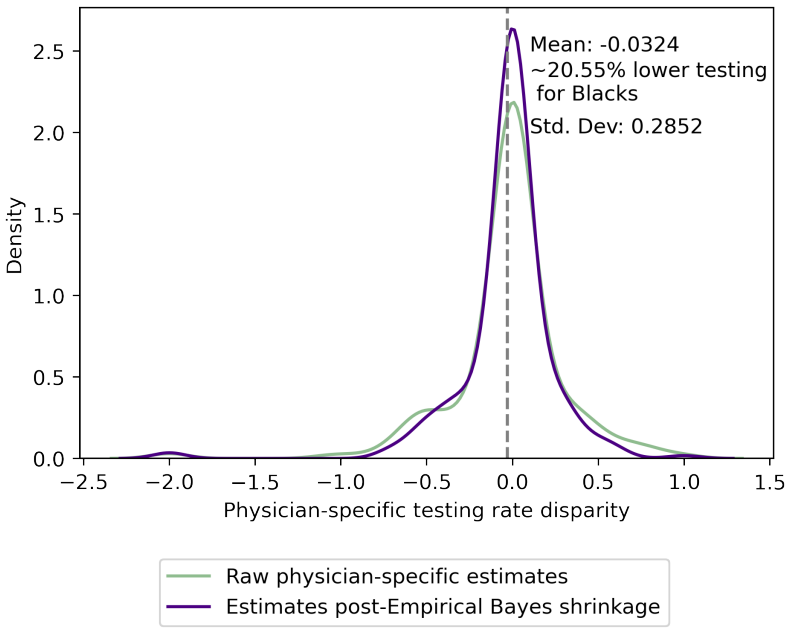
There is, however, considerable variation across physicians in their disparate treatment by patient race. We estimate physician-specific estimates of disparate treatment, δ_j, using the specification described in Equation (3). Figure 2(a) plots the distribution of physician-specific estimates of disparate treatment by patient race, as well as their posteriors after empirical Bayes shrinkage. On average, physicians test Black patients 1.11 percentage points less often when compared to non-Black patients, amounting to roughly a difference of thirty two percent by race. The standard deviation across physicians is about 1.45 percentage points. Specifically among patients with predicted risk greater than 1% and those with no contraindications, the disparate treatment in testing is about 3.2 percentage points ($\sim$ 20.55 percent lower) on average, with a standard deviation of 0.28.

FIGURE 2 :
Physician-specific estimates of racial disparity in testing for stroke

(A)
Based on all[§] visits



(B)
Based on visits with risk$\geq$ 1% and no contraindications



Notes: This figure plots the distribution of δ_j from Equation (3) conditional on risk deciles, as well as their posteriors after empirical Bayes shrinkage. [§]The estimates are based on visits in the test data i.e. the 40% of the primary sample unused in training the machine learning algorithm. Of these, we exclude visits to physicians who see fewer than 50 patients of either race and test fewer than 5 patients for stroke, since the physician-specific estimates for them would be unreliable.

While we do not observe physician characteristics in the data, we aggregate visit-specific and patient-specific variables at the level of the physician to look for suggestive evidence on what drives this variation in disparate treatment. Appendix Figure A8 separately plots correlations between physician-specific estimates of disparate treatment and some aggregate descriptives of their patient pool, such as: whether they come from rural counties, or share of patients from high income ZIPs within the state, and the share of Black patients. We do not find any statistically significant relations.

IV.B.1. Possible Mechanisms

Disparity in testing conditional on a given objective stroke risk, stems from two sources. First, physicians may be applying a different risk threshold when making testing decisions for each group. The use of different thresholds by race when the costs of testing are the same must be interpreted as *racial prejudice*. It is as though the physician is willing to accept a higher risk of missing strokes among patients of a particular group only in order to avoid testing them.

Second, physicians may be making testing decisions based on *subjective* assessments that differ from the objective risk P_i. Deviations in the physicians' subjective risk assessments, denoted by π_{ij}, from the objective value P_i, indicate the reduced quality of risk assessments made by physicians. Such deviations are reflective of both physicians' skill and the quality of care at the facility that this interaction takes place at. Since P_i accounts for all statistical differences in risk and information quality by race, larger deviations of π_{ij} from P_i for say Black patients, is indicative of risk being assessed differently for them in a way that is not justified by underlying differences in their presentations. In other words, two patients may have the same P_i and yet the physicians' subjective assessment for one race may be further away from P_i on average than the other. We refer to this differential accuracy in stroke risk assessment as the *unjustified skill gap*. Common reasons why this might happen are if physicians have implicit biases, hold incorrect stereotypes, disregard or minimize Black patients' symptoms, expend low effort in assessing Black patients, or use race-insensitive medical protocols.[26]

The two mechanisms can be viewed as discriminatory because despite having the same stroke risk, patients would end being differentially tested on the basis on their race solely because of physician error. Any 'accurate' statistical discrimination, on the other hand, is subsumed within the objective risk P_i and is not captured in our estimates of disparate treatment.

Section V formalizes these two mechanisms with a model of physicians' testing decision.

[26] Physician interactions with Black patients are typically shorter, with low information exchange, affected by stereotype threat, summarized in clinician notes with more negative patient descriptions, and likely to result in fewer positive outcomes (Tang et al., 2024; Sun et al., 2022; Beach et al., 2021; Alsan et al., 2019; Cooper et al., 2012; Penner et al., 2012). Physicians have also been found to be less likely to recommend treatments to Black patients due to implicit bias (Drwecki, Moore, Ward, & Prkachin, 2011; Green et al., 2007), perceptions or stereotypes about their failure to comply with medical advice (Calabrese et al., 2014; van Ryn et al., 2006), and false beliefs about biological differences by race (Hoffman et al., 2016; Todd et al., 2000).

V. STRUCTURAL ANALYSIS

In this section, we model physicians' testing decisions to formalise the two mechanisms of disparate treatment: the use of different testing thresholds for different racial groups i.e. *racial prejudice*, and the varying quality of subjective risk assessments by race, i.e. the *unjustified skill gap*. We then take the model to data in order to quantify the effect of these mechanisms.

V.A. Model Framework

Upon patient visit i to a facility, physician j is quasi-randomly assigned to the case of that patient.[27] The physician assesses the probability that the visit is a stroke episode, and then determines whether or not to test the patient by comparing the stroke risk assessment against the physician's threshold for testing. The quality of the physician's risk assessment and the choice of threshold applied by the physician, however, vary by *patient type* z. Patient type is determined by a vector of patient traits $\boldsymbol{Z_i}$ including race $R_i \in \{b, w\}$ and other non-race characteristics such as sex, insurance status, income quartile, and age. Thus, $z \equiv (r, c)$ where r indicates the patient's race, and c indicates other non-race characteristics.

Let $\pi_{ij} \in (0, 1)$ denote the physicians' subjective stroke risk assessment for visit i, and $\tau_j^z \in (0, 1)$ be the physician's threshold for testing patients of type z. The physician orders a test for patient i, indicated by $T_{ij} \in \{0, 1\}$, as long as the risk assessment exceeds the threshold.

$$\begin{aligned} T_{ij} &= \mathbb{1}\left\{\pi_{ij} \geq \tau_j^{z(i)}\right\}, \text{ or, equivalently} \\ &= \mathbb{1}\left\{\log\frac{\pi_{ij}}{1-\pi_{ij}} \geq \log\frac{\tau_j^{z(i)}}{1-\tau_j^{z(i)}}\right\} \end{aligned} \tag{5}$$

The threshold τ_j^z captures the physician's relative cost of testing patients of type z.[28] If the physician's threshold differs based on patient race when all other non-race characteristics are identical, then it constitutes racial prejudice (Becker, 1957). Consider for example, the case of racial prejudice against Black patients, i.e. $\tau_j^{z'=(b,c)} > \tau_j^{z=(w,c)}$. Here, by applying a higher threshold for Black patients, the physician is choosing to bear a higher risk of false negatives to avoid testing them.

The physician's subjective stroke risk assessment π_{ij}, on the other hand, depends on the details of the presenting case and the physician's skill. As before, let P_i denote the objective

[27] Section III.D verifies the validity of quasi-random physician assignment for our sample. Quasi-random assignment allows us to attribute any variation in testing within a facility and conditional on a given level of stroke risk, to variations in physician-specific characteristics of threshold or skill.

[28] Instead of assuming thresholds explicitly, Appendix D.1 describes how thresholds can also conceptualized to be derived from physicians' preference to match testing decisions to the underlying latent state, therefore minimizing classification errors. The two types of classification errors in this context are: false negatives (not testing a stroke case), and false positives (testing a non-stroke case). The higher the physician's relative cost of false negatives to false positives, the lower the threshold for testing.

probability of stroke associated with patient visit i based on the information available with the physician at the time the testing decision is made. We model the physician's subjective risk assessment to be such that the log odds of stroke as perceived by the physician, are given by

$$\log \frac{\pi_{ij}}{1-\pi_{ij}} = \log \frac{P_i}{1-P_i} + \zeta_{ij} \tag{6}$$

where ζ_{ij} captures the noise in the physician's assessment. We assume ζ_{ij} to be distributed normally with mean zero,[29] and standard deviation given by

$$\sqrt{\text{var}(\zeta_{ij})} = \sigma_j^{z(i)} \exp(\gamma_{f(i)}) \tag{7}$$

where $\gamma_{f(i)}$ captures the effect of the facility, i.e. its infrastructure, training of staff, and equipment availability; and σ_j^z represents the physician's skill. Physicians' skill is also modeled to vary with the patient type z to account for variations in the quality of the physician's interactions with the patient (such as symptom discounting, for example) and in the efforts expended by the physician on the case. The larger the value of σ_j^z, the greater are the deviations in the physician's assessment relative to the true risk, and higher are the chances of the physician making false negative or false positive testing decisions. Note that even though we don't explicitly model physician skill to vary with physician characteristics such as years of experience or area of specialty, any such relationships are still allowed to drive our estimates since we specify physician-specific parameters.

Since the objective risk P_i already accounts for any statistical differences in risk inference that can be made given the quality of information on visit i, large deviations from the objective odds are indicative of risk being assessed differently for the patient type in a way that is not justified by underlying differences in their presentations. The difference in the quality of risk assessments by patient race, given all other non-race characteristics identical, therefore represents the unjustified skill gap. That is, if $\sigma_j^{z'=(b,c)} > \sigma_j^{z=(w,c)}$, then the physician is worse at assessing risk for Black patients in a way that is not justified by underlying differences in presentation by race subsumed under P_i.

V.A.1. Some Model Considerations

Role of the facility. By modeling testing thresholds to be physician-specific and invariant across all facilities the physician works in, we implicitly assume that there are no facility-specific factors such as non-availability of neuroimaging equipment or technicians, that could potentially constraint testing. To support this assumption, we look at the variance in testing rates at the level of each physician, and at the level of each facility. Appendix Figure A9 plots the distributions of variance in testing rates of a physician across facilities, and the distributions of variance in testing rates across physicians working at the same facility. The variances in testing rates for each physician across the different facilities they work at, are

[29] If the mean of ζ_{ij} were to be some $\mu \neq 0$, then that would be observationally equivalent to the physician shifting the threshold for log odds by $-\mu$ instead. This systematic shift in risk assessment cannot be identified independently from the threshold.

very low and close to zero on average. Most of the variation in testing rates, as the figure suggests, actually comes from variation *across physicians* at a given facility.

Other objectives of the physician. We allow the threshold to also vary by the patient's non-race characteristics, thereby accommodating alternative objectives of the physician. For example, our framework allows physicians to practice *defensive medicine*, wherein physicians have a distaste for malpractice lawsuits and may therefore set lower thresholds for patient types who are more likely to file one. Similarly, physicians may also be selectively cautious and set lower thresholds for patient types with increased patient self-advocacy. No financial incentives are modeled in testing decisions because physicians in the emergency department are not given additional payments based on the volume of diagnostic imaging ordered, though radiologists who interpret those images may be.

Relation between risk threshold and skill. It is plausible that less experienced or low skilled doctors generally set lower thresholds to avoid false negatives. While our specification does not assume a relation between physician preference and skill, it also does not impose that there isn't any. By not assuming an explicit relationship between the physician's choice of threshold and their skill, we allow for the possibility that the physician may be unaware of their skill in making risk assessments, and therefore might not necessarily adjust their threshold in response.

V.B. Identification

Given any distribution of stroke risk for patients of type z, a choice of threshold τ_j^z produces a rate of false negatives (FNR_j^z) and false positives (FPR_j^z); see Appendix Figure A10 for an illustration. The physician's choice of threshold originates from the physician's disutility of false negatives relative to false positives. The higher the threshold is set, the greater is the associated rate of false negatives and smaller is the associated rate of false positives. The physician's choice of threshold rests therefore on their preferred ratio of false negatives to false positives.

With variation in physician skill, however, the distribution of risk *as perceived* by the physician differs from the distribution of objective risk assessments. A less skilled physician is less able to separate stroke episodes from non-stroke visits, and is hence more likely to make false negative *and* false positive decisions. We modeled a less skilled physician as one who makes noisier risk assessments, i.e. has a higher σ_j^z. Between physicians who make use the same threshold τ, a physician with higher σ_j^z produces higher FPR_j^z and FNR_j^z; see Appendix D.4 for a formal demonstration.

For each physician j and patient type z, we observe the physician's testing decisions given different levels of risk, and the physician's overall rates of false negatives and false positives. Threshold and skill are jointly identified by the size and ratio of FNR_j^z to FPR_j^z made by the physician. Conditional on same threshold, the size of the false negative and false positive rates determines the skill of the physician. The ratio of false negatives to false positives, conditional on a level of skill, pins the threshold.

V.C. Parameterisation, and Estimation using Hierarchical Bayes

V.C.1. Physician Heterogeneity

We specify heterogeneity in physicians thresholds, by positing heterogeneity in their relative disutility from false negatives to false positives. We model that disutility as $-\exp(b_j+\boldsymbol{\beta}_j'\boldsymbol{Z_i})$, yielding a risk threshold given by

$$\tau_j^z = \frac{1}{1+\exp(b_j+\boldsymbol{\beta}_j'\boldsymbol{z})} \tag{8}$$

for patients with $\boldsymbol{Z_i}=\boldsymbol{z}$. Appendix D.1 details how the threshold is derived. This specification renders the associated threshold in the log odds space to be $-b_j-\boldsymbol{\beta}_j'\boldsymbol{Z_i}$. We include in $\boldsymbol{Z_i}$ indicators of patient race, sex, age, insurance status, and income quartile of the patient's zipcode in the state. The threshold τ_j^0 for the reference patient type with $\boldsymbol{Z_i}=\boldsymbol{0}$ is given by $1/(1+\exp(b_j))$.

Likewise, heterogeneity in physician skill and its relation to patient type is specified as

$$\sigma_j^z = \exp(c_j+\boldsymbol{\varsigma}_j'\boldsymbol{z}) \tag{9}$$

Rewriting Equation (7) with the parameterisation above, and normalizing physician skill for the reference patient type ($\boldsymbol{Z_i}=\boldsymbol{0}$) at the reference facility ($f=1$) to 1, we obtain

$$\sqrt{var(\zeta_{ij})} = \exp(\boldsymbol{\varsigma}_j'\boldsymbol{z}+\sum_{f=2}^{F}\tilde{\gamma}_f D_{if}) \tag{10}$$

where D_{if} is an indicator for whether patient visit i was at facility f or not, and $\tilde{\gamma}_f \equiv \gamma_f-\gamma_1$, $\forall f>1$. For estimation, we define facility f at the level of stroke certification of the ED, an indicator for weekend visits, and an indicator visits during the night.[30] EDs that have the same level of stroke certification are expected to be identical in the infrastructure and training required to treat stroke patients. We designate EDs with no stroke certification visited during daylight hours on weekdays as the 'reference' facility.

V.C.2. Substitution of objective odds with machine learning predictions

When substituting for the objective log odds in Equation (6) with machine learning predictions, it is important to account for prediction errors. We assume

$$\log\frac{\hat{P}_i}{1-\hat{P}_i} = \frac{1}{\rho}\log\frac{P_i}{1-P_i}+\upsilon_i$$

where $\upsilon_i \sim \mathcal{N}(0,\sigma_\upsilon^2)$. Substituting this relation in Equation (6), the deviations in risk assessments from algorithmic log odds are given by $\zeta_{ij}-\rho\upsilon_i$. The noise in assessments is

[30] Stroke care certificates verify that the respective facility's stroke care programs, staffing, and infrastructure meet the national standards of stroke care delivery. We use each facility's stroke certification status as of January 1, 2016 to categorize it into one with no certification, primary stroke certification, or comprehensive stroke certification.

therefore $var(\zeta_{ij} - \rho\upsilon_i)$. Since the noise in algorithmic predictions is common to all physicians, it does not affect the ordering of physicians based on their skill. Therefore, accounting for such prediction errors, we interpret our estimates of physician skill as the dispersion in noise of physicians' risk assessment net of algorithmic prediction errors.

V.C.3. Parameters of Interest

In $\boldsymbol{Z_i}$, Z_{1i} is an indicator for the patient race being Black. Racial prejudice is hence indicated by $\beta_{j1} \neq 0$. In particular, $\beta_{j1} < 0$ suggests prejudice against Black patients. Similarly, unjustified skill gap by patient race is indicated by $\varsigma_{j1} \neq 0$, where $\varsigma_{j1} > 0$ suggests unjustified skill gap disfavoring Black patients.

V.C.4. Estimation using Hierarchical Bayes

When taking the model to the data, we first define our *modeling sample* for estimation. The modeling sample is based on the 40 percent of the visits that were not used in the training of the machine learning algorithm. To estimate physician-specific preference and skill parameters, it is important that we observe physician decisions for sufficiently many Black and non-Black patients with varying stroke risks. Therefore, to construct our modeling sample, we exclude visits to physicians who see fewer than 50 patients of either race and test fewer than five patients for stroke. We also exclude visits by patients with any contraindications to testing. Appendix D.2 reports the number of observations dropped at each of these steps. The modeling sample thus obtained covers 73,006 visits to 239 physicians. Appendix Figure A11 illustrates the partition of the primary sample into a training sample for machine learning prediction, and the subsequent filter on the remaining test sample to obtain the modeling sample for estimation.

To improve power, we assume a common population distribution for the physician-specific parameters that are governed by hyper-parameters. Specifically, we assume

$$\begin{aligned} \begin{bmatrix} b_j & \boldsymbol{\beta}_j \end{bmatrix}' &\sim \mathcal{N}\left(\mu_\beta, \Omega_\beta\right) \\ \boldsymbol{\varsigma_j}' &\sim \mathcal{N}\left(\mu_\varsigma, \Omega_\varsigma\right) \end{aligned} \tag{11}$$

This generates a hierarchy among the parameters—the physician-specific parameters drive decisions and thereby the data generated in the process, while the hyper-parameters govern the physician-specific parameters. The advantage of assuming such a structure is that physician-specific parameters are then disciplined by the hyper-parameters that in-turn learn from the pooled data generated by all physicians. Therefore, by considering dependencies in data across different levels, we can fit parameters more accurately even if physician-level data were limited.

We estimate the parameters using Bayesian procedures. First, we define the likelihood of the observed physicians' decisions and outcomes. It then, combined with appropriate priors on the hyper-parameters and the priors on the physician-specific parameters as specified in Equation (11), produces a joint posterior distribution of all parameters.

Let $\boldsymbol{\Theta}$ denote the vector of hyper-parameters, and $\boldsymbol{\Lambda}$ denote the vector of remaining parameters, partitioned into physician-specific parameters $\boldsymbol{\Lambda_1}$ and fixed parameters $\boldsymbol{\Lambda_2}$. From the physician's testing decision modeled in Equation (5) and our parameterizations, we have

$$\begin{aligned} P(T_{ij} = 1|\boldsymbol{\Lambda}) &= \mathbb{P}\left(\log\frac{P_i}{1-P_i} + \zeta_{ij} \geq \log\frac{\tau_j^{z(i)}}{1-\tau_j^{z(i)}}\right) \\ &= 1 - \Phi\left(\frac{-\tilde{\rho}_j \log\frac{\hat{P}_i}{1-\hat{P}_i} - \tilde{b}_j - \tilde{\beta}_j' Z_i}{\exp\left(\tilde{\zeta}_j' Z_i + \sum_{f>1}\tilde{\gamma}_f D_{if}\right)}\right) \end{aligned} \tag{12}$$

where $\tilde{\rho}_j = \dfrac{\rho}{\exp(c_j+\gamma_1)}$, $\tilde{b}_j = \dfrac{b_j}{\exp(c_j+\gamma_1)}$, $\tilde{\boldsymbol{\beta}}_j = \dfrac{\boldsymbol{\beta}_j}{\exp(c_j+\gamma_1)}$, and $\tilde{\gamma}_f = \gamma_f - \gamma_1$. The physician-specific parameters can therefore be identified only up to scale, as can only the differences in effects of all facilities relative to the reference facility.

The likelihood of physician j's decisions are then given by

$$\mathcal{L}(\mathcal{D}_j|\boldsymbol{\Lambda}) = \prod_{i\in I_j} \mathbb{P}(T_{ij}|\boldsymbol{\Lambda})^{w_{ij}} \tag{13}$$

where I_j is the set of patient visits assigned to the physician and w_{ij} is the class weight proportional to the inverse of the share of j's patients with the same testing decision as T_{ij}.

With $\boldsymbol{\Lambda}_1$ drawn from population distributions with hyper-parameters in $\boldsymbol{\Theta}$, and a prior $k(\boldsymbol{\Theta}, \boldsymbol{\Lambda_2})$, the joint-posterior of all parameters is given by

$$K(\boldsymbol{\Theta}, \boldsymbol{\Lambda}|\mathcal{D}) \propto \prod_{j\in J} \mathcal{L}(\mathcal{D}_j|\boldsymbol{\Lambda})\, g(\boldsymbol{\Lambda}_1|\boldsymbol{\Theta})\, k(\boldsymbol{\Theta}, \boldsymbol{\Lambda_2}) \tag{14}$$

where g is the density of the common distribution specified in Equation (11).

To approximate the joint posterior we draw samples from it using Gibbs sampling, a Monte Carlo Markov Chain (MCMC) algorithm that sequentially draws from conditional posteriors of each parameter given values of all the other parameters. Physician-level parameters are sampled from the conditional posteriors using Metropolis–Hastings algorithm. The hyper-parameters, on the other hand, are drawn from tractable posteriors obtained using conjugate priors. Appendix D.3 details our choice of priors and the Gibbs sampling procedure.

The stationary distribution of the Markov chain of the samples drawn using the Gibbs procedure approximates the joint posterior distribution of all parameters. By Bernstein-von Mises theorem, under certain conditions,[31] the posterior distribution converges to the asymptotic sampling distribution of the classical maximum likelihood estimator, with the mean of the posterior converging to the maximum likelihood estimator. Bayesian credible sets obtained from the posterior distribution also have frequentist coverage properties.

[31] Specifically, it holds for fixed-dimensional problems under the assumption that the prior is continuous, and (strictly) positive in a neighborhood of the population parameter.

V.D. Results

Table 7 reports our estimates for the mean of the physician-specific threshold and skill parameters. Panel A reports the hyper-parameter μ_β that denotes the mean of the preference parameters β_j that determine the threshold. A negative coefficient in the vector μ_β indicates that the incremental log disutility of physicians from false negatives relative to false positives among patients with the associated trait, is lower on average. It translates to the physician applying a higher threshold for them. Our estimates suggest that physicians use higher thresholds when determining whether to test Black and female patients. On average, physicians' threshold for log odds risk is incrementally higher for Black patients by 0.280. Higher thresholds are also used for patients aged younger than 50 years, among whom stroke is less common. There is suggestive evidence of physicians' also lowering thresholds for patients from richer ZIPs.

TABLE 7 :
Select Estimates from Hierarchical Bayes

Parameter Description	Notation	Estimate	Std. Error
Panel A: Preference parameters that determine threshold			
Incremental log disutility from false negatives to false positives for			
Black patients	$\mu_{\beta 1}$	-0.280^{***}	(0.041)
Female patients	$\mu_{\beta 2}$	-0.126^{***}	(0.004)
Patients from high income quartile ZIPs	$\mu_{\beta 3}$	0.378^{*}	(0.209)
Patients aged < 50	$\mu_{\beta 4}$	-0.200^{***}	(0.060)
Uninsured patients	$\mu_{\beta 5}$	-0.027	(0.040)
Panel B: Skill parameters			
Incremental noise in subjective assessments of log odds risk for			
Black patients	$\mu_{\varsigma 1}$	0.190^{***}	(0.004)
Female patients	$\mu_{\varsigma 2}$	0.585^{***}	(0.003)
Patients from high income quartile ZIPs	$\mu_{\varsigma 3}$	0.236	(0.206)
Patients aged < 50	$\mu_{\varsigma 4}$	0.485^{***}	(0.005)
Uninsured patients	$\mu_{\varsigma 5}$	0.276	(0.206)

*p<0.1; **p<0.05; ***p<0.01

Notes: The table reports the estimates of select model parameters. The estimates are obtained from the mean of the marginal posterior of the respective parameter, and the standard errors (in parentheses) are the corresponding standard deviations. The Hierarchical Bayes method and the Gibbs sampling procedure are described under Appendix D.3.

Panel B reports estimates of the hyper-parameter μ_ς representing the mean of the physician-specific threshold parameters ς_j. Recall that ς_{jk} indicates the incremental increase in log standard deviation of the subjective log odds assessments made by physician j for patients with trait k. Positive and significant coefficients indicate greater noise in assessing risk for patients with said trait, indicating lower physician skill. Yet again, we find physicians' risk assessments to be noisier for Black, female and young patients. In particular, the standard deviation of subjective log odds assessments is, on average, 19% higher for Black patients.

We also find variance in assessments to be lower in primary and comprehensive certified facilities, however the gains are sometimes offset on weekdays and daytime hours probably due to high patient volume (Appendix Table B4).

V.E. Counterfactuals and Policy Simulations

Lastly, we use our estimates to quantify the role played by racial prejudice and unjustified skill gap in disparate testing. We also simulate the outcomes of a few policies aimed at potentially reducing racial disparity. We evaluate the policy simulations on how they affect the racial gap in false negatives i.e. the share of stroke visits that are not tested. In the modeling sample, the rate of false negatives is around 5.55 percentage points higher for Black patients. Our estimates produce a baseline difference of 6.56 percentage points by race, against which we compare the different scenarios and policy simulations.

First, we consider some simple counterfactuals where physicians separately equalize thresholds and the quality of their risk assessments for the two racial groups. Our findings are given in Panel A of Table 8. Under Counterfactual 1, we simulate testing decisions if testing thresholds were equalized by race, i.e. if β_{j1} were zero for all physicians. Since our estimates of β_{j1} are negative, this counterfactual effectively lowers the threshold of Black patients. As a consequence, testing rate among Black patients increases offsetting partially the racial difference in false negatives. Equalizing thresholds, all else the same, the racial difference in false negatives is reduced by half.

Likewise, under Counterfactual 2, we simulate testing decisions for when the quality of risk assessments are equalized by race; i.e. ς_{j1} becomes zero for all j. Since our estimates of ς_{j1} are positive, this counterfactual reduces the variance of physicians' subjective odds risk assessments around the objective odds risk. This improves precision in testing decisions for Black patients, as is indicated by the narrower difference in precision across racial groups. However, the racial difference in false negatives is widened. The widening occurs because while Black patients are less likely to under-assessed in this counterfactual, they are also less likely to be over-assessed. Overall, we find the testing rate for Black patients to fall, and consequently the difference in false negatives increases.

Panel B in Table 8 examines the effect of different policy simulations. The first of these combines physicians' testing decisions with machine learning recommendations. Several studies have examined the usefulness of artificial intelligence (AI) as complements to physician diagnoses and found mixed results (Farzaneh et al., 2023; Gallo et al., 2024). While AI outperforms physicians *on average* for straightforward cases, AI accuracy is lower than physicians for complex case presentations. Farzaneh et al. (2023) examine various physician-AI collaboration strategies to see which of them improve diagnostic accuracy. They find using AI as the primary diagnosis tool, with its diagnoses replaced with that of the physician only when the AI is uncertain, to statistically dominate other strategies. Reducing the autonomy of healthcare providers who are able to observe nuances of each individual patient is however unlikely to be acceptable to physicians and patients. We explore a feasible alternative in Counterfactual 3.

TABLE 8 :
Counterfactuals and Policy Simulations

			False Negative	False Positive	Precision
			T=0|S=1	T=1|S=0	S=1|T=1
			(rates, in percentages)		
	Data Baseline	Non-Black	7.03	2.55	26.93
		Black	12.58	1.96	22.53
		Difference	-5.55	0.59	4.40
	Counterfactual Baseline	Difference	-6.56	-0.04	2.10
			False Negative	False Positive	Precision
			(differences by race)		
	Panel A: Simple Counterfactuals				
1	No racial prejudice ($\beta_{j1} = 0 \, \forall j$)		-3.25	-1.36	2.23
2	No unjustified skill gap by race ($\varsigma_{j1} = 0 \, \forall j$)		-7.22	3.08	1.13
	Panel B: Policy Simulations				
3	Combine physicians' decision with ML recommendation		-6.26	0.73	1.90
4	Lower thresholds by 10% by externally raising costs of false negatives		-6.61	0.49	1.76
5	Redirect patients to a stroke-certified facility within 10 miles from the patient's ZIP, if any		-9.10	0.18	2.20

Notes: The table reports baseline outcomes of false negative rate (share of stroke visits that aren't tested), false positive rate (share of non-stroke visits that are tested), and precision rate (share of stroke visits among those tested) of stroke testing in the modeling sample. Panels A and B report what these outcomes would change to under various counterfactual scenarios and policy simulations, respectively.

Under Counterfactual 3, both physician and AI assess the patient's risk for stroke. Only if the physician's assessment is lower than the algorithmic prediction, testing is based on the maximum of the two risk predictions. If not, the physician makes the testing decision as they did at baseline, and the choice of threshold is still physician-specific.[32] At first glace, this policy guards patients against errors made by low-skilled physicians and improve outcomes for both racial groups. However, an untargeted application of such a policy reduces false negatives for both racial groups and more so for non-Black patients who are subject to lower thresholds. The policy therefore could end up widening racial differences depending on the difference in thresholds. In our sample, the two effects offset and the racial difference in false negatives remains comparable to the baseline.

Counterfactual 4 simulates another untargeted policy that externally imposes a cost on false negative testing decisions lowering all thresholds by, say, 10 percent. An example of this external imposition of cost is facilitating the filing of malpractice lawsuits. As before, the policy mechanically raises testing rates for both groups, but since the skill gap is still unaddressed, the racial difference persists. Policies that separately attend to either skill or threshold are therefore insufficient to close the large racial disparity in false negatives.

Finally, Counterfactual 5 examines a scenario where patients are redirected to stroke certified facilities within 10 miles of their ZIP, if there are any.[33] Appendix Figure A12 illustrates the distribution of stroke-certified EDs across ZIPs. At the facility, a physician is assigned at random. Once again, testing rates improve for both racial groups but the increase is larger for non-Black patients. Consequently, the racial difference increases.

The common drawback in the set of policy simulations we examine is that they are untargeted. Alternatively, remedying disparate treatment among prejudiced physicians is a tall order. We know however that racial concordance matters for health outcomes (Alsan et al., 2019). Policies that promote diversity in the healthcare workforce and improve cultural competence via medical education might perhaps be more successful in closing the racial gap in physician care.

VI. CONCLUSION

This paper estimates a large racial disparity in the quality of stroke diagnoses delivered in emergency departments (EDs). We then also identify and quantify the role of disparate treatment by physicians in driving this disparity. Our identification of disparate treatment relies on obtaining predictions of stroke risk for each ED visit, that are racially objective and subsume the quality of information available to the physician on patients from either groups. We use machine learning to obtain these risk predictions. The usage of machine learning predictions is justified because we observe a detailed representation of the patient's

[32] Davenport (2023) cautions against a discretionary version of this policy where physicians could simply choose to consult the AI prediction if they wanted. Davenport (2023) finds that prejudiced decision-makers selectively use algorithmic recommendations to discreetly discriminate against minorities to not appear discriminatory.

[33] This counterfactual is based on visits only made by local patients i.e. those with a Florida-based ZIP Code.

chart and are able to train the algorithm on the underlying latent stroke states of all visits non-selectively. The inclusion of race as a feature in the algorithm assures that the machine learning model learns from all race-correlated patterns in the data, whether via observed or unobserved factors. The risk predictions made by the algorithm therefore subsume race-specific and race-correlated differences in the quality of information available or levels of stroke risk, thereby allowing cross-group comparisons.

We find disparate treatment in testing to account for roughly 60% of the racial disparity in missed diagnoses. The disparate treatment is realized via two mechanisms: unjustified skill gap, wherein physicians make noisier risk assessments for Black patients relative to the objective risk predictions, and racial prejudice in the canonical sense, where physicians apply differential thresholds. Untargeted policies that separately attend to either skill or threshold separately are therefore insufficient to close the racial disparity in the quality of testing decisions.

With considerable variation across physicians in disparate treatment by race, another important question that follows is of what drives physician heterogeneity. Whether it driven by physician-level biases or from broader systemic factors is an important and policy-relevant area for future research.

REFERENCES

Abaluck, J., Agha, L., Kabrhel, C., Raja, A., & Venkatesh, A. (2016). The Determinants of Productivity in Medical Testing: Intensity and Allocation of Care. *American Economic Review*, *106(12)*, 3730-3764.

Albert, J. H., & Chib, S. (1993). Bayesian Analysis of Binary and Polychotomous Response Data. *Journal of the American Statistical Association*(422), 669-679.

Alsan, M., Garrick, O., & Grazianiy, G. (2019). Does Diversity Matter for Health? Experimental Evidence from Oakland. *American Economic Review*, *109*(12), 4071–4111.

Alvarez, R. M., & Brehm, J. (1995). American Ambivalence Towards Abortion Policy: Development of a Heteroskedastic Probit Model of Competing Values. *American Journal of Political Science*(4), 1055–82.

Anathhanam, S., & Hassan, A. (2017). Mimics and chameleons in stroke. *Clinical medicine (London, England)*, *17(2)*, 156–160.

Antonovics, K., & Knight, B. G. (2009). A new look at racial profiling: Evidence from the Boston police department. *The Review of Economics and Statistics*, *91*(1), 163-177.

Anwar, S., & Fang, H. (2006). An Alternative Test of Racial Bias in Motor Vehicle Searches: Theory and Evidence. *American Economic Review*, *96*, 127–151.

Arch, A. E., Weisman, D. C., Coca, S., Nystrom, K. V., WiraIII, C. R., & Schindler, J. L. (2016). Missed Ischemic Stroke Diagnosis in the Emergency Department by Emergency Medicine and Neurology Services. *Stroke*, *47*(3), 668–673. doi: https://doi.org/10.1161/STROKEAHA.115.010613

Arnold, D., Dobbie, W., & Hull, P. (2022). Measuring Racial Discrimination in Bail Decisions. *American Economic Review*, *112(9)*, 2992-3038.

Arnold, D., Dobbie, W., & Yang, C. (2018). Racial Bias in Bail Decisions. *Quarterly Journal of Economics*, *133(4)*, 1885-1932.

Ayres, I. (2002). Outcome Tests of Racial Disparities in Police Practices. *Justice Research and Policy*, *4*.

Balsa, A. I., & McGuire, T. G. (2003). Prejudice, clinical uncertainty and stereotyping as sources of health disparities. *Journal of Health Economics*, *22*(1), 89–116.

Beach, M. C., Saha, S., Park, J., Taylor, J., Drew, P., Plank, E., ... Chee, B. (2021). Testimonial Injustice: Linguistic Bias in the Medical Records of Black Patients and Women. *Journal of General Internal Medicine*, *36*(6), 1708-1714.

Becker, G. S. (1957). *The Economics of Discrimination*. University of Chicago Press, Chicago.

Bernstein, S. N. (1917). *Theory of Probability*. Moscow.

Bluhmki, E., Ángel Chamorro, Dávalos, A., Machnig, T., Sauce, C., Wahlgren, N., ... Hacke, W. (2009). Stroke treatment with alteplase given 3·0–4·5 h after onset of acute ischaemic stroke (ecass iii): additional outcomes and subgroup analysis of a randomised controlled trial. *The Lancet Neurology*(12), 1095 - 1102.

Bohren, J. A., Haggag, K., Imas, A., & Pope, D. G. (2024). Inaccurate Statistical Discrimination: An Identification Problem. *Review of Economics and Statistics*, forthcoming.

Broder, J. S., Bhat, R., Boyd, J. P., Ogloblin, I. A., Limkakeng, A., Hocker, M. B., ... Repplinger, M. D. (2016). Who Explicitly Requests the Ordering of Computed To-

mography for Emergency Department Patients? A Multicenter Prospective Study. *Emergency Radiology*, *23*(3), 221-227.

Buck, B. H., Akhtar, N., Alrohimi, A., Khan, K., & Shuaiba, A. (2021). Stroke mimics: incidence, aetiology, clinical features and treatment. *Annals of Medicine*, *53*(1), 420-436.

Calabrese, S. K., Earnshaw, V. A., Underhill, K., Hansen, N. B., & Dovidio, J. F. (2014). The impact of patient race on clinical decisions related to prescribing HIV pre-exposure prophylaxis (PrEP): assumptions about sexual risk compensation and implications for access. *AIDS and Behavior*, *18*(2), 226–240.

Card, D., DellaVigna, S., Funk, P., & Iriberri, N. (2020). Are Referees and Editors in Economics Gender Neutral? *The Quarterly Journal of Economics*(1), 269–327.

CDC. (2024). *Stroke Facts.* Retrieved Accessed: August 6, 2024, from https://www.cdc.gov/stroke/data-research/facts-stats/index.html

Chalela, J. A., Kidwell, C. S., Nentwich, L. M., Luby, M., Butman, J. A., Demchuk, A. M., ... Warach, S. (2007). Magnetic resonance imaging and computed tomography in emergency assessment of patients with suspected acute stroke: a prospective comparison. *The Lancet*, *369*(9558), 293–298.

Chan, D. C., Gentzkow, M., & Yu, C. (2022). Selection with variation in diagnostic skill: Evidence from radiologists. *The Quarterly Journal of Economics*, *137(2)*, 729-783.

Chandra, A., Kakani, P., & Sacarny, A. (2024). Hospital Allocation and Racial Disparities in Health Care. *The Review of Economics and Statistics*, *106*(4), 924-937.

Chandra, A., & Staiger, D. O. (2010). Identifying Provider Prejudice in Healthcare. *National Bureau of Economic Research Working Paper #16382*.

Chandra, A., & Staiger, D. O. (2020). Identifying Sources of Inefficiency in Healthcare. *The Quarterly Journal of Economics*, *135*(2), 785–843.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In (p. 785–794). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/2939672.2939785 doi: 10.1145/2939672.2939785

Cooper, L. A., Roter, D. L., Carson, K. A., Beach, M. C., Sabin, J. A., Greenwald, A. G., & Inui, T. S. (2012). The associations of clinicians' implicit attitudes about race with medical visit communication and patient ratings of interpersonal care. *American Journal of Public Health*, *102*(5), 979–987.

Davenport, D. (2023). Discriminatory Discretion: Theory and evidence from use of pretrial algorithms. *Working Paper*.

Dobbie, W., Liberman, A., Paravisini, D., & Patania, V. (2021). Measuring Bias in Consumer Lending. *Review of Economic Studies*, *88*, 2799–2832.

Drwecki, B. B., Moore, C. F., Ward, S. E., & Prkachin, K. M. (2011). Reducing racial disparities in pain treatment: the role of empathy and perspective-taking. *Pain*, *152*(5), 1001–1006.

Farzaneh, N., Ansari, S., Lee, E., Ward, K. R., & Sjoding, M. W. (2023). Collaborative strategies for deploying artificial intelligence to complement physician diagnoses of acute respiratory distress syndrome. *npj Digital Medicine*, *6*(62). doi: 10.1038/s41746-023-00797-9

Feigenberg, B., & Miller, C. (2022). Would Eliminating Racial Disparities in Motor Vehicle Searches have Efficiency Costs? *The Quarterly Journal of Economics*, *137*(1), 49–113.

Fernandes, P. M., Whiteley, W. N., Hart, S. R., & Salman, R. A.-S. (2013). Strokes: Mimics and Chameleons. *Practical Neurology*, *13*, 21-28.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*(5), 1189-1232.

Gallo, R. J., Shieh, L., Smith, M., Marafino, B. J., Geldsetzer, P., Asch, S. M., ... Li, R. C. (2024). Effectiveness of an Artificial Intelligence-Enabled Intervention for Detecting Clinical Deterioration. *JAMA Internal Medicine*, *184*(5), 557-562. doi: 10.1001/jamainternmed.2024.0084

Geman, S., & Geman, D. (1984). Stochastic relaxation gibbs distributions and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 721–741.

Gowrisankaran, G., Joiner, K., & Léger, P. T. (2023). Physician Practice Style and Healthcare Costs: Evidence from Emergency Departments. *Management Science*, *69*(6), 3202–3219.

Green, A. R., Carney, D. R., Pallin, D. J., Ngo, L. H., Raymond, K. L., Iezzoni, L. I., & Banaji, M. R. (2007). Implicit Bias among Physicians and its Prediction of Thrombolysis Decisions for Black and White Patients. *Journal of General Internal Medicine*, *22*(9), 1231–1238.

Healthcare Cost and Utilization Project (HCUP). (2016 and 2017). State Emergency Department Databases (SEDD), State Inpatient Databases (SID), and State Ambulatory Surgery and Services Databases (SASD). Agency for Healthcare Research and Quality, Rockville, MD. www.hcup-us.ahrq.gov/sidoverview.jsp.

Hoffman, K. M., Trawalter, S., Axt, J. R., & Oliver, M. N. (2016). Racial bias in pain assessment and treatment recommendations, and false beliefs about biological differences between blacks and whites. *PNAS Proceedings of the National Academy of Sciences of the United States of America*, *113*(16), 4296–4301.

Hsiao, C. (2014). *Analysis of Panel Data* (Third ed.). Cambridge University Press.

Hull, P. (2021). What Marginal Outcome Tests Can Tell Us about Racially Biased Decision-Making. *NBER Working Paper 28503*.

Iakovlev, A., & Liang, A. (2024). The Value of Context: Human versus Black Box Evaluators.

Institute for Health Metrics and Evaluation. (2024). *Global Burden of Disease 2021: Findings from the GBD 2021 Study.* Retrieved Accessed: August 6, 2024, from https://www.healthdata.org/research-analysis/library/global-burden-disease-2021-findings-gbd-2021-study

Institute of Medicine. (2003). *Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care.* Brian D. Smedley, Adrienne Y. Stith and Alan R. Nelson, editors. Washington (DC): National Academies Press (US). doi: https://doi.org/10.17226/12875

Jayaraman, R., Ray, D., & Wang, S.-Y. (2014). Engendered access or engendered care? Evidence from a major Indian hospital. *Economic and Political Weekly*, *49*(25), 47–53.

Kamel, H., Zhang, C., Kleindorfer, D. O., Levitan, E. B., Howard, V. J., Howard, G., ... Johnston, S. C. (2020). Association of black race with early recurrence after minor ischemic stroke or transient ischemic attack: Secondary analysis of the point randomized clinical trial. *JAMA Neurology*(5), 601–605.

Kane, M. J., King, C., Esserman, D., Latham, N. K., Greene, E. J., & Ganz, D. A. (2023). A Compressed Large Language Model Embedding Dataset of ICD 10 CM Descriptions.

Keele, L., & Park, D. K. (2006). Difficult Choices: An Evaluation of Heterogeneous Choice Models. *Presented at the 2004 Meeting of the American Political Science Association*.

King, G., & Zeng, L. (2001). Logistic Regression in Rare Events Data. *Political Analysis*, *9*(2), 137-163.

King, J., & Matthews, B. (2024). DasGuptR: standardisation and decomposition following Prithwis Das Gupta's 1991 method [Computer software manual]. (R package version 2.0.1)

Kitagawa, E. M. (1955). Components of a Difference Between Two Rates. *Journal of the American Statistical Association*, *50*(272), 1168-1194.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human Decisions and Machine Predictions. *The Quarterly Journal of Economics*, *133(1)*.

Knowles, J., Persico, N., & Todd, P. (2001). Racial Bias in Motor Vehicle Searches: Theory and Evidence. *Journal of Political Economy*, *109*, 203-229.

Laplace, P.-S. (1820). *Théorie analytique des probabilités* (Third ed.). Paris.

Lever, N. M., Nyström, K. V., Schindler, J. L., III, C. W., & Funk, M. (2013). Missed Opportunities for Recognition of Ischemic Stroke in the Emergency Department. *Journal of Emergency Medicine*, *39*(5), 434-39.

Low, H., & Pistaferri, L. (2019). Disability Insurance: Error Rates and Gender Differences. *NBER Working Paper 26513*.

MacKinnon, J. G., & White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, *29*(3).

Martin, S. S., Aday, A. W., Almarzooq, Z. I., Anderson, C. A., Arora, P., Avery, C. L., ... Subcommittee, S. S. (2024). 2024 heart disease and stroke statistics: A report of us and global data from the american heart association. *Circulation*, *149*(8), e347-e913. doi: 10.1161/CIR.0000000000001209

Mullainathan, S., & Obermeyer, Z. (2022). Diagnosing Physician Error: A Machine Learning Approach to Low-Value Health Care. *The Quarterly Journal of Economics*, *132*(2), 679-727.

Mullins, M. E., Schaefer, P. W., Sorensen, A. G., Halpern, E. F., Ay, H., He, J., ... Gonzalez, R. G. (2002). CT and conventional and diffusion-weighted MR imaging in acute stroke: study in 691 patients at presentation to the emergency department. *Radiology*, *224*(2), 353-60.

National Institute of Neurological Disorders and Stroke rt-PA Stroke Study Group. (1995). Tissue plasminogen activator for acute ischemic stroke. *The New England Journal of Medicine*(24), 1581-7.

Newman-Toker, D. E., Moy, E., Valente, E., Coffey, R., & Hines, A. L. (2014). Missed diagnosis of stroke in the emergency department: a cross-sectional analysis of a large population-based sample. *Diagnosis (Berlin, Germany)*, *1(2)*, 155–166.

Newman-Toker, D. E., Peterson, S. M., Badihian, S., Hassoon, A., Nassery, N., Parizadeh, D., ... Robinson, K. A. (2022). Diagnostic Errors in the Emergency Department: A Systematic Review. Comparative Effectiveness Review No. 258. (Prepared by the Johns Hopkins University Evidence-based Practice Center under Contract No. 75Q80120D00003.).December 2022. Errata and Addendum, August 2023. *AHRQ Pub-*

lication No. 22(23)-EHC043. Rockville, MD: Agency for Healthcare Research and Quality.. doi: 10.23970/AHRQEPCCER258

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*, 447-53.

Penner, L. A., Eggly, S., Griggs, J. J., III, W. U., Orom, H., & Albrecht, T. L. (2012). Life-threatening disparities: The treatment of Black and White cancer patients. *Journal of Social Issues*, *68*(2), 328–357.

Rathore, S. S., Hinn, A. R., Cooper, L. S., Tyroler, H. A., & Rosamond, W. D. (2002). Characterization of incident stroke signs and symptoms: Findings from the atherosclerosis risk in communities study. *Stroke*, *33(11)*, 2718–2721.

Singh, M., & Venkataramani, A. (2024). Rationing by Race. *NBER Working Paper 30380*.

Sun, M., Oliwa, T., Peek, M. E., & Tung, E. L. (2022). Negative Patient Descriptors: Documenting Racial Bias In The Electronic Health Record. *Health Affairs*, *41*(2).

Tang, M., Mishuris, R. G., Payvandi, L., & Stern, A. D. (2024). Differences in Care Team Response to Patient Portal Messages by Patient Race and Ethnicity. *JAMA Network Open*, *7*(3).

Tarnutzer, A. A., Lee, S.-H., Robinson, K. A., Wang, Z., Edlow, J. A., & Newman-Toker, D. E. (2017). ED misdiagnosis of cerebrovascular events in the era of modern neuroimaging: A meta-analysis. *Neurology*, *88(15)*, 1468–1477.

Todd, K. H., Deaton, C., D'Adamo, A. P., & Goe, L. (2000). Ethnicity and analgesic practice. *Annals of Emergency Medicine*, *35*(1).

Train, K. E. (2009). *Discrete Choice Methods with Simulation.* Cambridge University Press.

Tsao, C. W., Aday, A. W., Almarzooq, Z. I., Anderson, C. A., Arora, P., Avery, C. L., ... Subcommittee, S. S. (2023). Heart Disease and Stroke Statistics—2023 Update: A Report From the American Heart Association. *Circulation*, *147:e93–e621*.

van Ryn, M., Burgess, D., Malat, J., & Griffin, J. (2006). Physicians' Perceptions of Patients' Social and Behavioral Characteristics and Race Disparities in Treatment Recommendations for Men With Coronary Artery Disease. *American Journal of Public Health*, *96*(2), 351-357.

von Mises, R. (1931). *Wahrscheinlichkeitsrechnung.* Springer, Berlin.

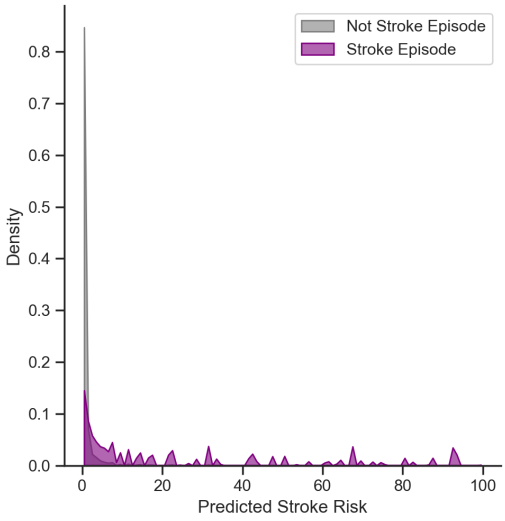
Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data* (Second ed.). MIT Press.

APPENDIX

A. FIGURES

FIGURE A1 :
Distribution of algorithmic stroke risk predictions

(A)
By stroke label



(B)
By stroke label, and patient race

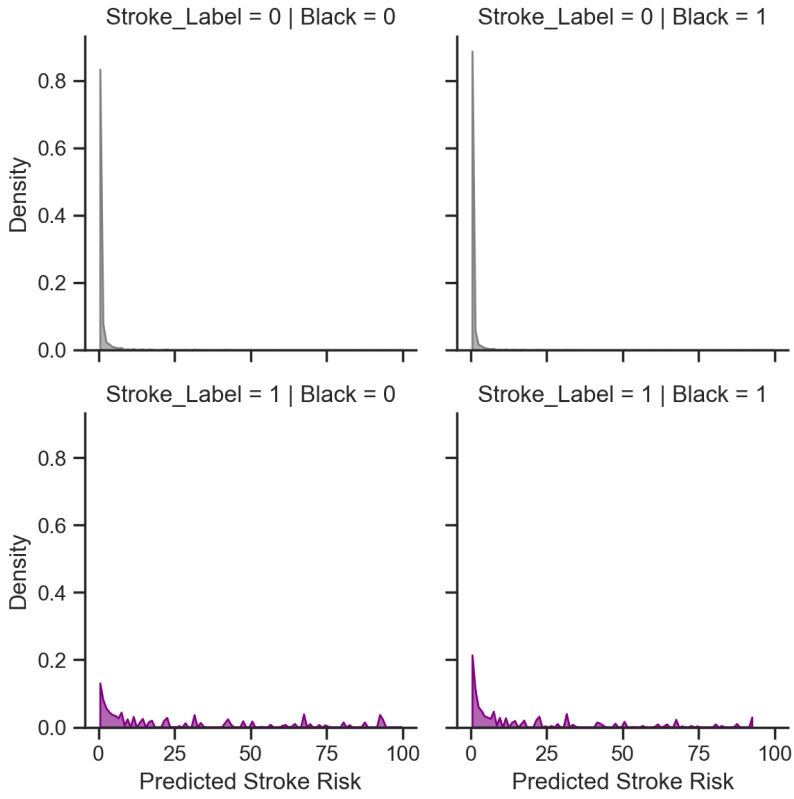
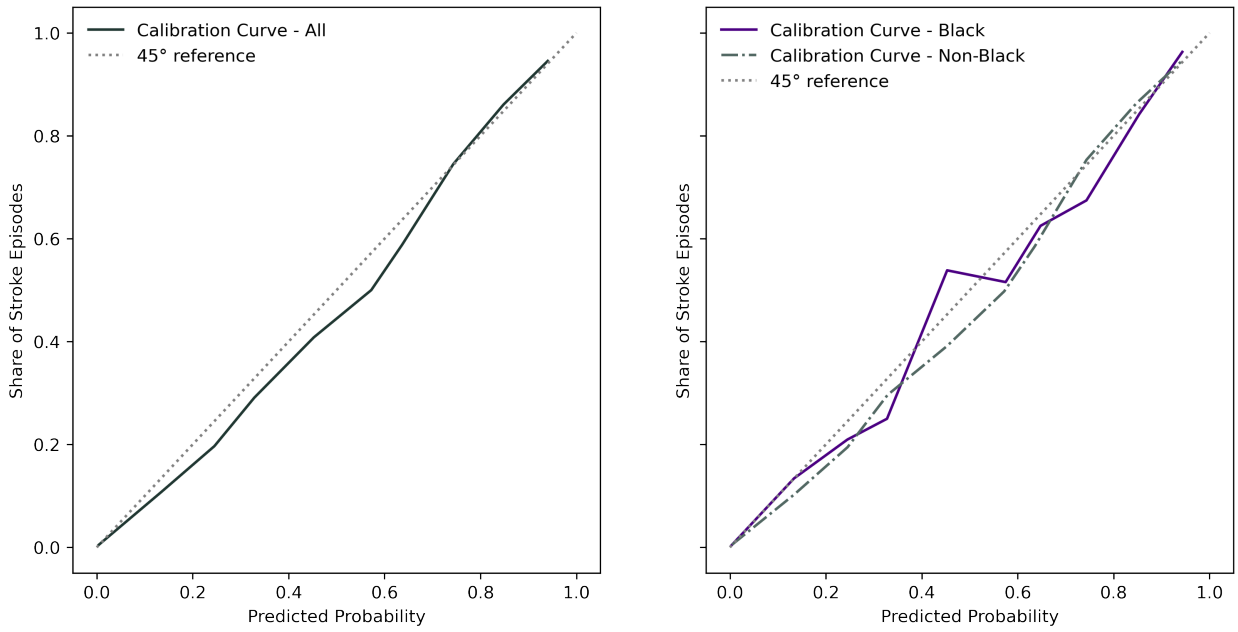
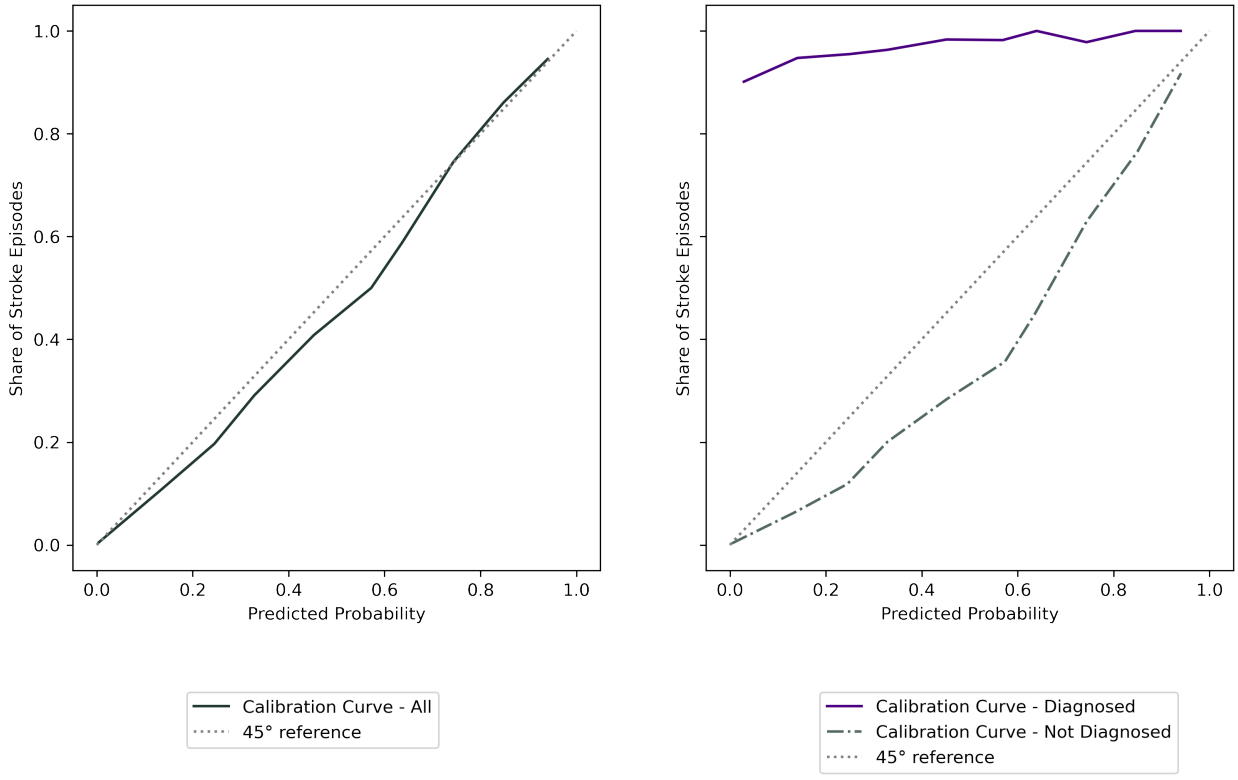


FIGURE A2 :
Calibration Plots (by Race)



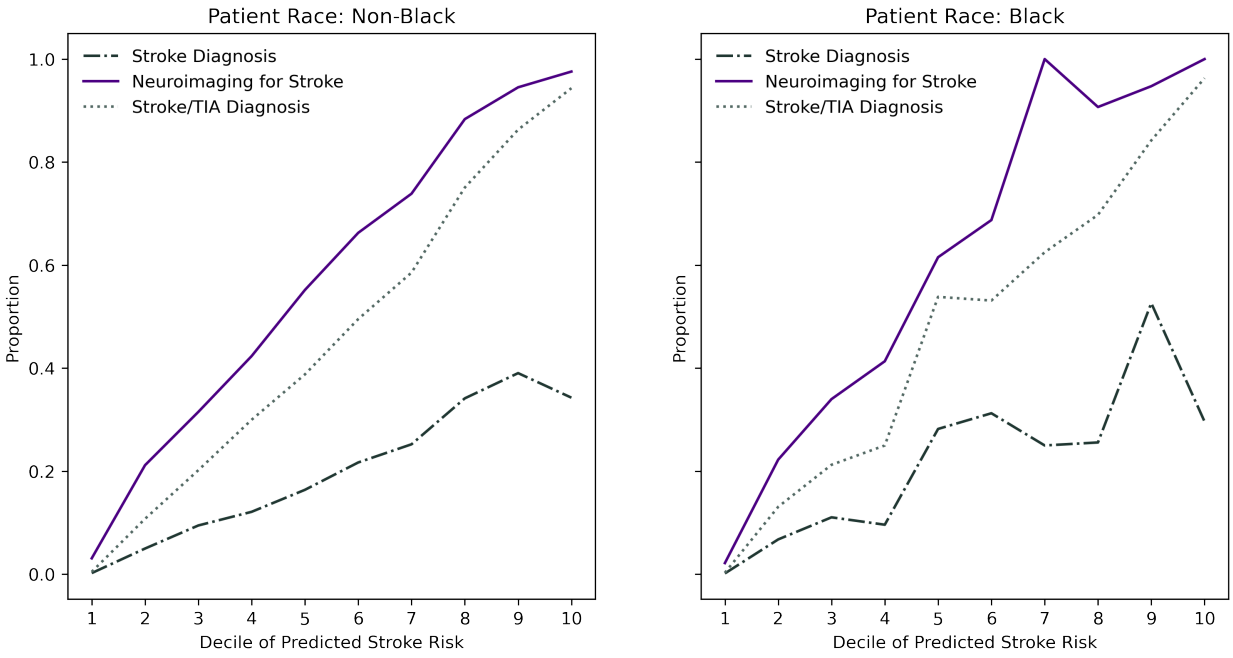
Notes: This figure compares the predicted probabilities generated by the machine learning model against the actual observed frequencies of the latent stroke states, based on all visits in the test set (left panel) and based on visits grouped by patient race (right panel). The algorithmic predictions of risk are well-calibrated for both racial groups.

FIGURE A3 :
Calibration Plots (by Stroke Diagnosis)



Notes: This figure compares the predicted probabilities generated by the machine learning model against the actual observed frequencies of the latent stroke states, based on all visits in the test set (left panel), and based on visits grouped by stroke diagnosis by physicians (right panel).

FIGURE A4 :
Concordance of algorithmic predictions with physicians' decisions



Notes: This figure plots physicians' decision rates of stroke diagnosis and neuroimaging for each decile of the predicted stroke risk, separately for non-Black patients (left panel) and Black patients (right panel). The decisions rates are increasing in the predicted risk confirming that physicians' perceptions of stroke risk are broadly consistent with the algorithmic risk predictions.

FIGURE A5 :
Distribution of the predicted risk for the misdiagnosed

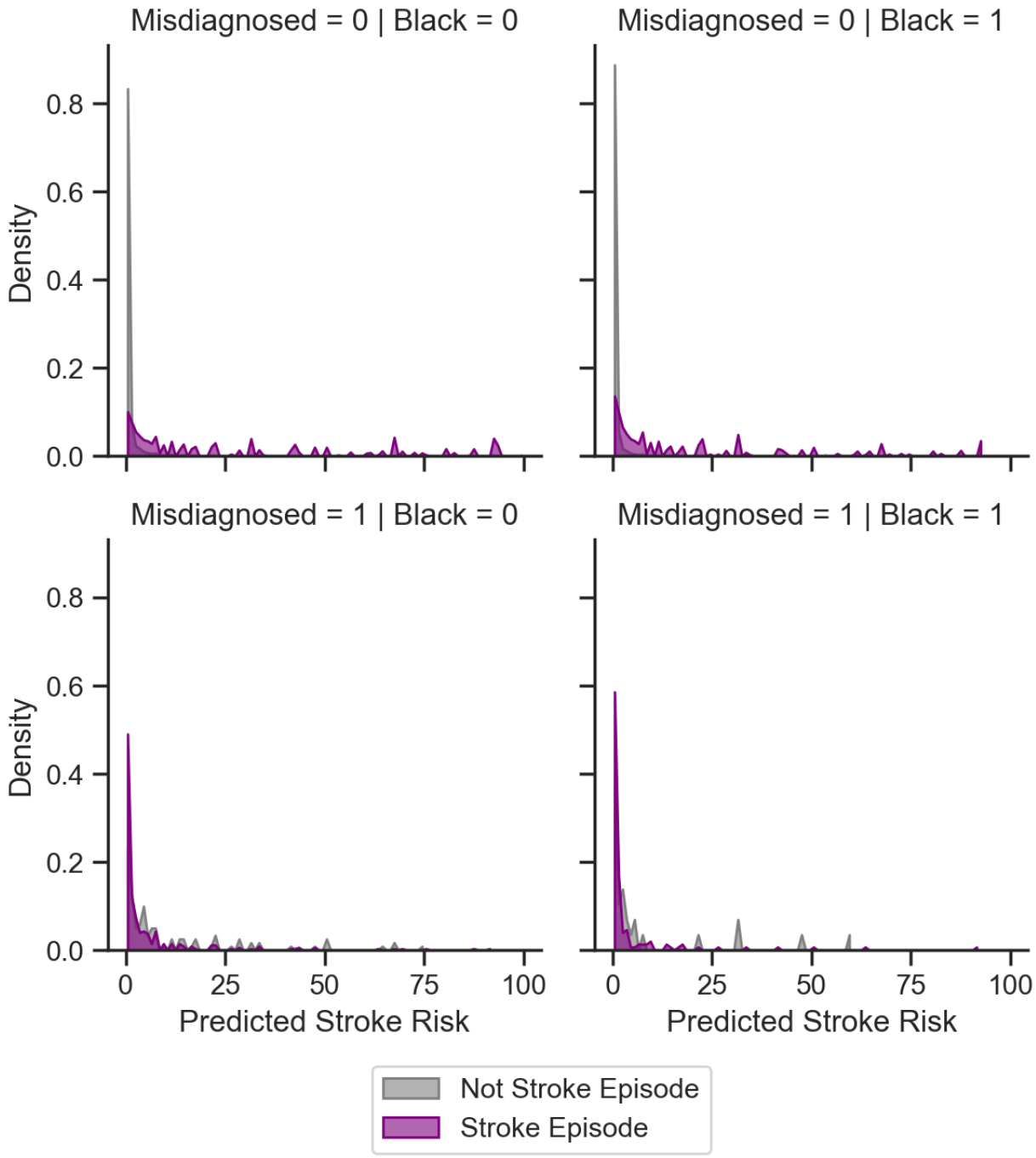
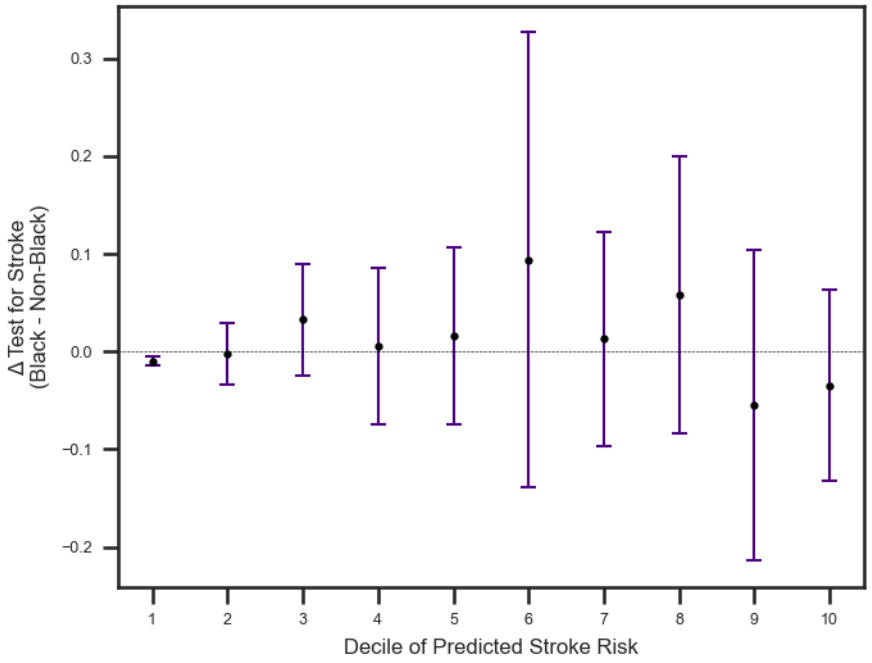


FIGURE A6 :
Quasi-experimental assignment of physicians

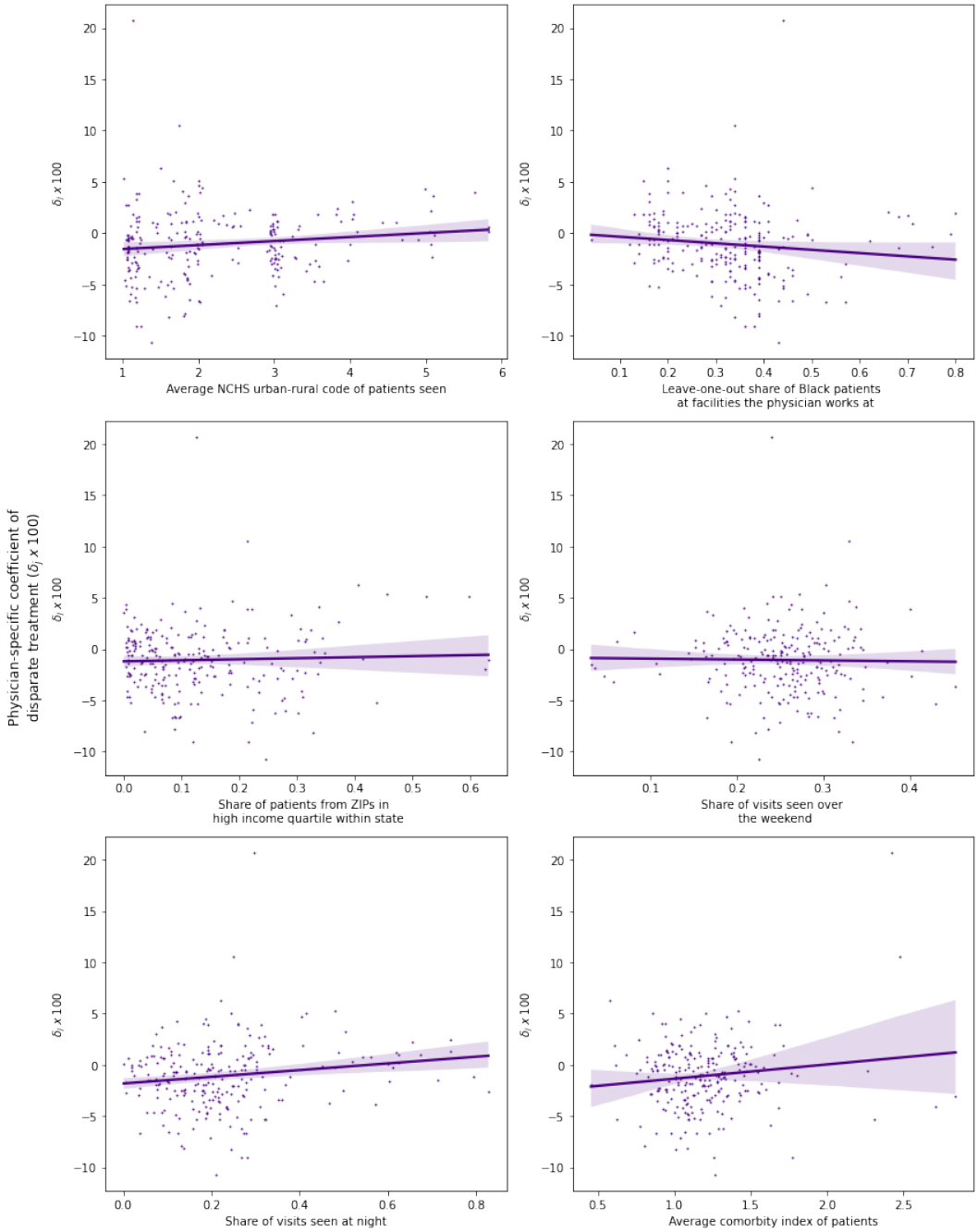
Notes: Coefficients and 95% confidence intervals from regressing the assigned physicians leave-one-out measures of decision propensity and decision quality, on patient covariates conditional on facility fixed effects. Patient covariates that are on the continuous scale are standardised. F-statistic and p-value from the joint F-test of patient covariates are reported at the bottom of each panel after adjusting for the multiple hypothesis testing using Bonferroni correction.

FIGURE A7 :
Disparate treatment in testing by predicted risk decile



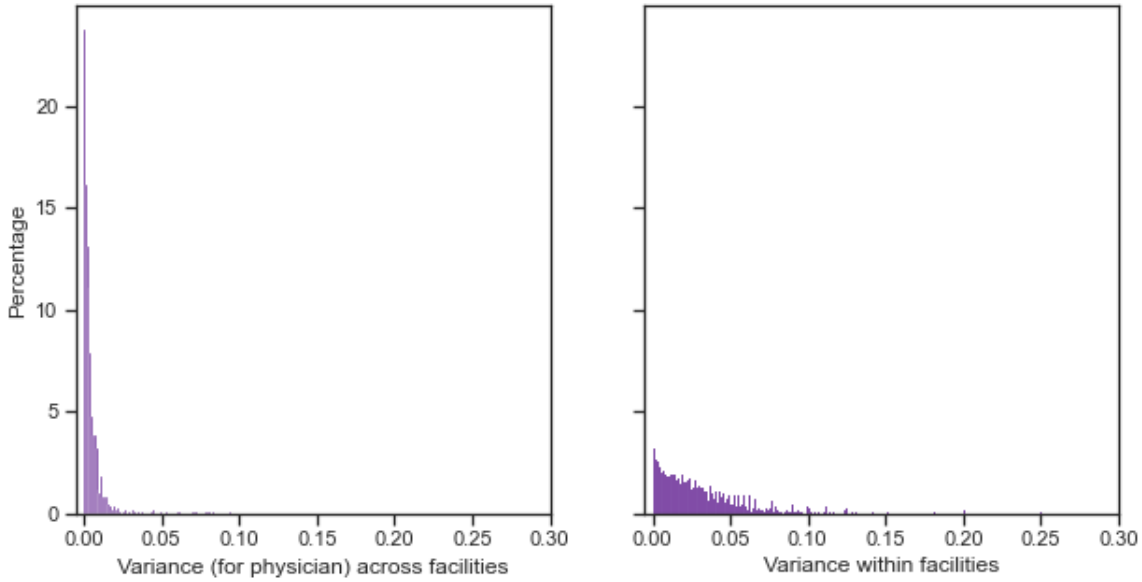
Notes: Coefficients and 95% confidence intervals from a linear probability specification of being tested for stroke on patient race. The error bars indicate the 95% confidence intervals.

FIGURE A8 :
Heterogeneity in Disparate Treatment



Notes: This figure plots correlations between physician-specific estimates of disparate treatment (δ_j) from Equation (3) with physician-level aggregates of visit and patient characteristics.

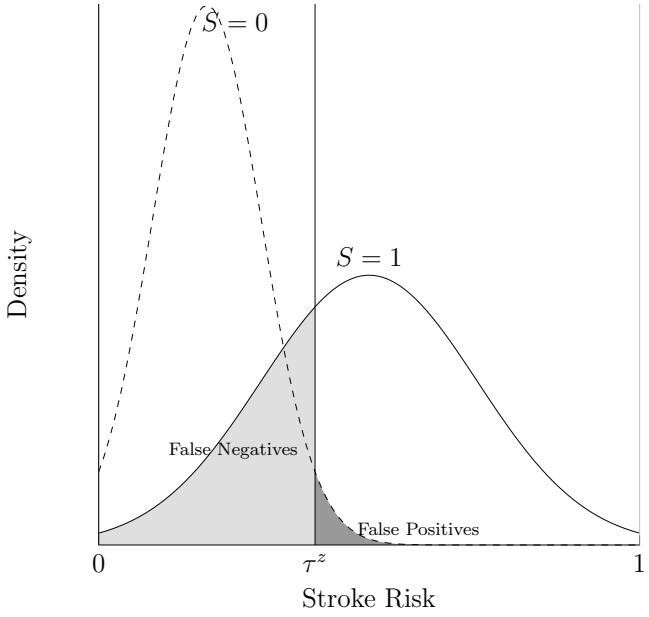
FIGURE A9 :
Distribution of variances in testing rates



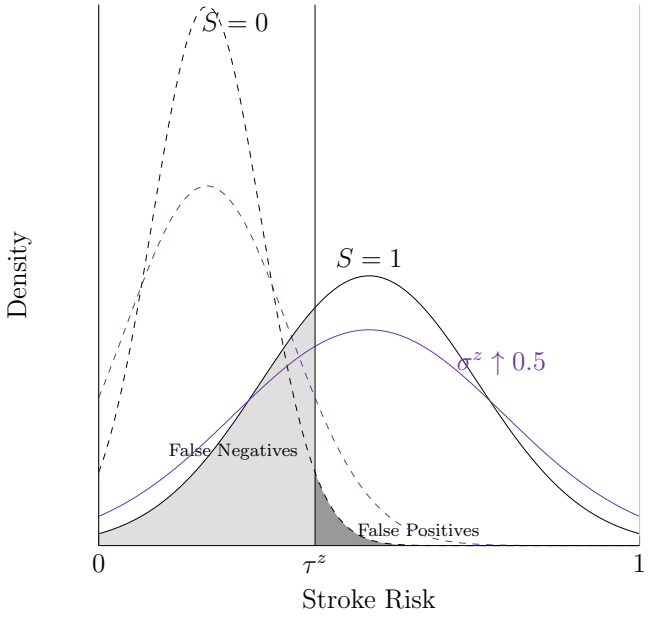
Notes: The left panel plots the distribution of variances in each physician's testing rates across the facilities they work at, and the right panel plots the distribution of variances in testing rates at each facility across the physicians who work at the facility. Facility is a defined as a unique combination of the ED, weekend/weekday, and admission during the day/night.

FIGURE A10 :
Illustration of threshold, false negatives, and false positives

(A)
Risk distribution given a level of skill



(B)
Skill variation distorts perceived distribution



Notes: The figure plots the risk distribution of patients of any type z, based on their latent stroke state $S \in \{0, 1\}$ separately. Given any level of skill, the choice of a specific threshold τ^z produces a unique ratio of false negatives and false positives; as illustrated in Panel (A). Physicians with lower level of skill perceive the risk distribution to have larger variance. Given any choice of threshold, lower skill produces are a larger volume of false negatives and false positives; as illustrated in Panel (B).

FIGURE A11 :
Partition of the primary sample into the training and modeling samples.

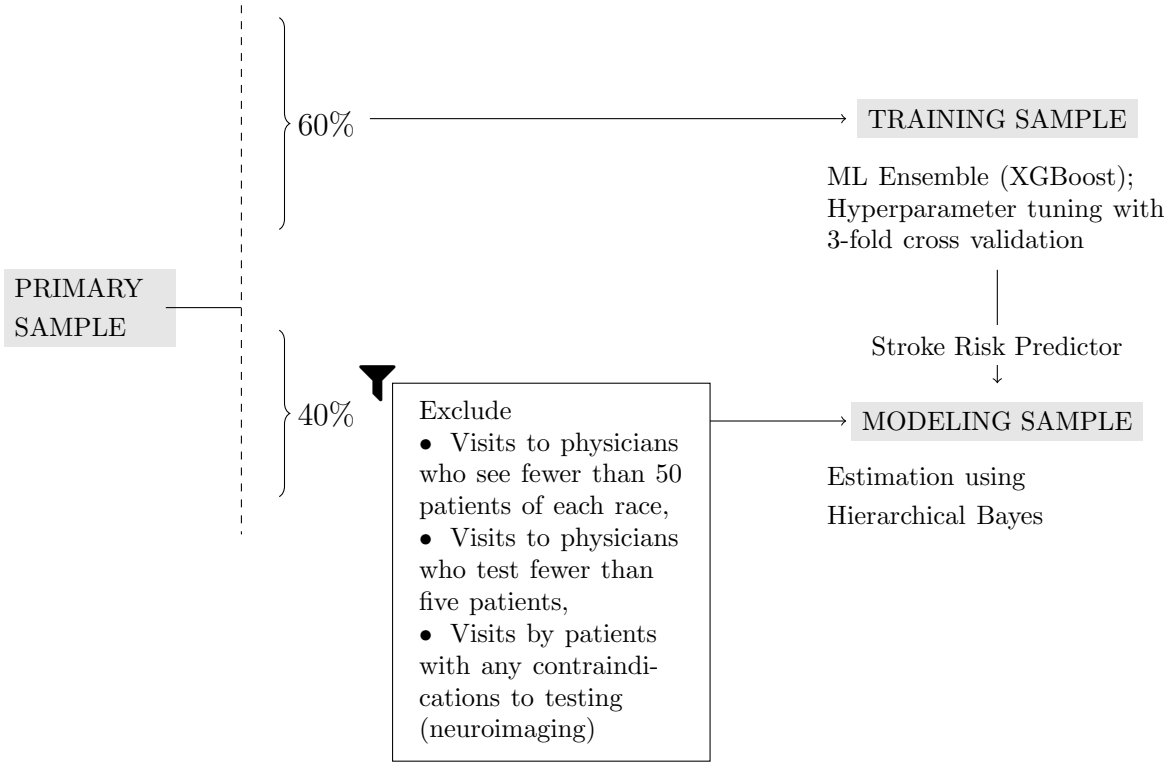
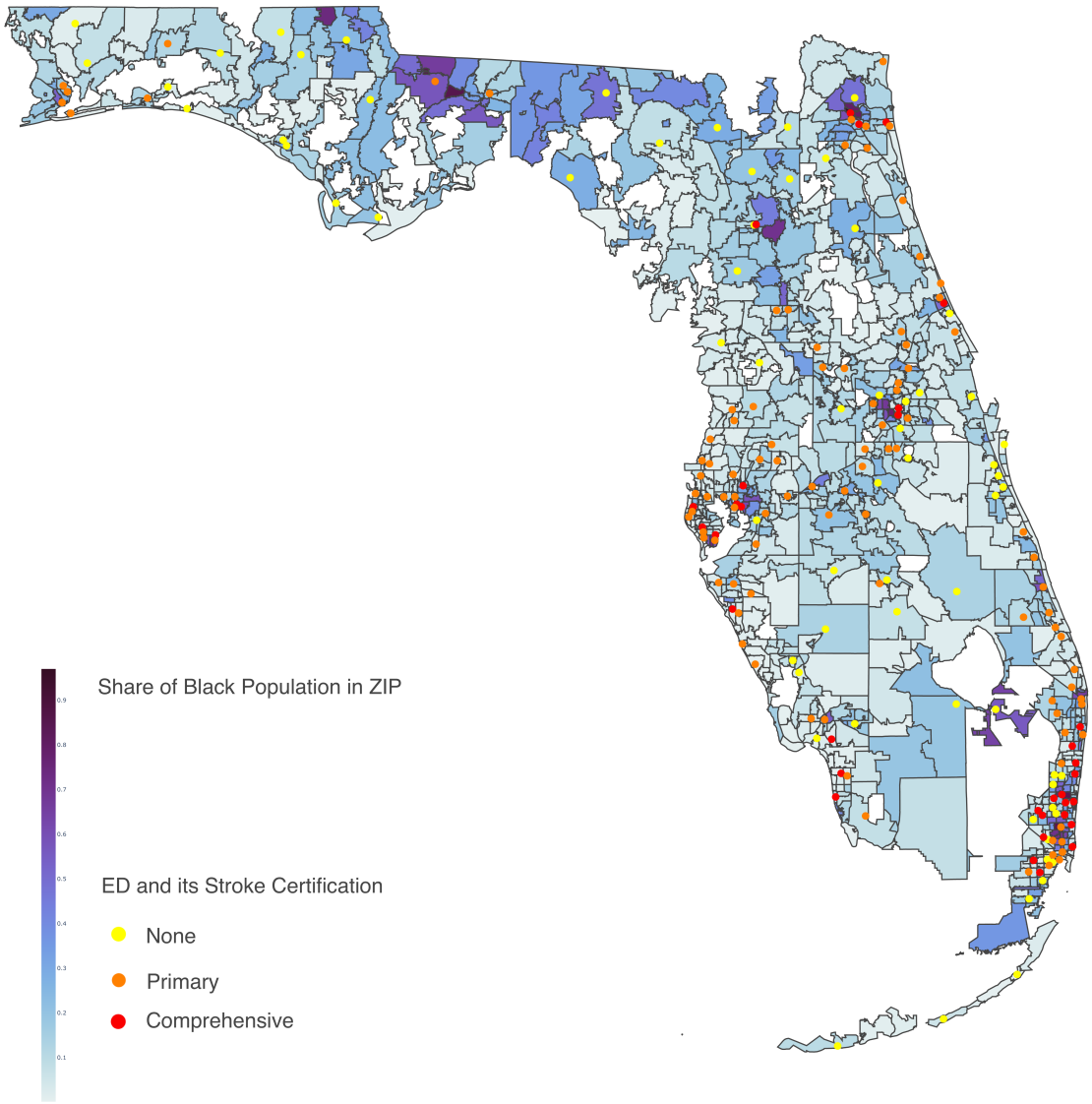


FIGURE A12 :
Distribution of stroke-certified EDs across ZIPs



Notes: This figure illustrates the distribution of stroke-certified emergency departments across ZIPs in Florida. The color gradient represents the share of Black population in the ZIP.

B. TABLES

TABLE B1 :
Racial disparity in the quality of stroke diagnosis, based on latent states as inferred from patient revisits over different time intervals.

	Linear Probability Model for Stroke Diagnosis ($D = 1$)			
	30-day	20-day	14-day	10-day
	(1)	(2)	(3)	(4)
Latent Stroke State ($S = 1$)	0.7865***	0.8184***	0.8395***	0.8603***
	(0.005)	(0.005)	(0.005)	(0.005)
Black × Latent Stroke State ($S = 1$)	−0.1407***	−0.1347***	−0.1202***	−0.1015***
	(0.014)	(0.014)	(0.014)	(0.013)
Black	−0.0000***	−0.0000***	−0.0000***	−0.0000***
	(0.000)	(0.000)	(0.000)	(0.000)
Constant	0.0003***	0.0003***	0.0003***	0.0003***
	(0.000)	(0.000)	(0.000)	(0.000)
Observations	1,368,680	1,368,680	1,368,560	1,368,650
Adjusted R^2	0.704	0.734	0.756	0.778

*p<0.1; **p<0.05; ***p<0.01

Notes: Estimates of racial disparity in the quality of stroke diagnosis based on Equation (4), using latent stroke states S_i inferred off of different time intervals. No. of observations differ across columns because the exclusion criteria for visits to facilities that don't see any stroke patients depends on S_i labels and hence on the choice of the interval. Heteroskedasticity-consistent MacKinnon and White (1985) HC3 standard errors are reported in parentheses.

TABLE B2 :
Racial disparity in the quality of stroke diagnosis based on insurance type

	Linear Probability Model for Stroke Diagnosis ($D = 1$)	
	Medicare	Medicaid
	(1)	(2)
Stroke Episode ($S = 1$)	0.8496***	0.7936***
	(0.006)	(0.018)
Black × Stroke Episode ($S = 1$)	−0.1210***	−0.1450***
	(0.020)	(0.035)
Black	0.0005***	0.0000
	(0.000)	(0.000)
Observations	337,015	285,784
Facility FE	X	X
Controls	X	X
Adjusted/Within R^2	0.773	0.698

*p<0.1; **p<0.05; ***p<0.01

Notes: Estimates of racial disparity in the quality of stroke diagnosis for patients covered by Medicare and Medicaid. Controls include: patients' age and sex, income quartile of patient's zip in state, insurance status with primary expected payer, and Charlson Comorbidity Index at the time of the visit. Facility fixed effects are for combinations of emergency room facility, quarter, indicator for weekends, and the admission hour. Heteroskedasticity-consistent MacKinnon and White (1985) HC3 standard errors in parenthesis.

TABLE B3 :

Racial disparity in the quality of stroke diagnosis, based on samples with different inclusion criteria on visits following which the patient doesn't revisit the health system again in the year and even in 2017

	Linear Probability Model for Stroke Diagnosis ($D = 1$)	
	includes visits with no subsequent revisits in 2016 or 2017	excludes visits with no subsequent revisits in 2016 or 2017
	(1)	(2)
Stroke Episode ($S = 1$)	0.8395***	0.8216***
	(0.005)	(0.006)
Black × Stroke Episode ($S = 1$)	−0.1202***	−0.1244***
	(0.014)	(0.015)
Black	−0.0000**	−0.0001
	(0.000)	(0.000)
Constant	0.0003***	0.0004***
	(0.000)	(0.000)
Observations	1,368,560	1,111,499
Adjusted R^2	0.756	0.731

*p<0.1; **p<0.05; ***p<0.01

Notes: In the case of visits following which the patient doesn't revisit the health system again in the year or even in the year 2017, there are two possibilities: either the patient doesn't need any medical care during this period, or the patient died relatedly or unrelatedly sometime after the ED visit. If we set $S_i = D_{ij}$ for these visits, we could possibly risk underestimating the rate of missed diagnoses for visits with $D_{ij} = 0$ and underestimating the rate of incorrect diagnoses or false positives for visits with $D_{ij} = 1$. Dropping these visits from the sample, on the other hand, overestimates the rate of incorrect and missed diagnoses. The estimates from both these approaches are broadly similar, but we decidedly err on the side of underestimating misdiagnosis in the rest of our analysis. Heteroskedasticity-consistent MacKinnon and White (1985) HC3 standard errors are reported in parentheses.

TABLE B4 :
Facility-Specific Parameter Estimates from Hierarchical Bayes

Parameter		Notation	Estimate	Standard Error
Incremental noise in subjective assessment of log odds risk due to the facility, relative to the reference				
Certification-Weekend-Night				
None	- 0 - 0	γ_1	— reference —	
None	- 0 - 1	$\tilde{\gamma}_2$	-0.215^{***}	(0.027)
None	- 1 - 0	$\tilde{\gamma}_3$	-0.116^{***}	(0.025)
None	- 1 - 1	$\tilde{\gamma}_4$	-0.262^{***}	(0.048)
Primary	- 0 - 0	$\tilde{\gamma}_5$	-0.053^{***}	(0.016)
Primary	- 0 - 1	$\tilde{\gamma}_6$	-0.213^{***}	(0.024)
Primary	- 1 - 0	$\tilde{\gamma}_7$	-0.046	(0.025)
Primary	- 1 - 1	$\tilde{\gamma}_8$	-0.221^{***}	(0.045)
Comprehensive	- 0 - 0	$\tilde{\gamma}_9$	-0.022^{***}	(0.010)
Comprehensive	- 0 - 1	$\tilde{\gamma}_{10}$	-0.212^{***}	(0.020)
Comprehensive	- 1 - 0	$\tilde{\gamma}_{11}$	-0.070^{***}	(0.019)
Comprehensive	- 1 - 1	$\tilde{\gamma}_{12}$	-0.237^{***}	(0.047)

*p<0.1; **p<0.05; ***p<0.01

Notes: The table reports the estimates of facility-specific parameters. For model estimation, we define a facility at the level of the stroke certification of the ED, an indicator for weekend visits, and an indicator visits during the night. The estimates are obtained from the mean of the marginal posterior of the respective parameter, and the standard errors (in parentheses) are the corresponding standard deviations. The Hierarchical Bayes method and the Gibbs sampling procedure are described under Appendix D.3.

C. SOME CONSIDERATIONS FOR THE ML APPLICATION

- *Racial biases in the input data*
  The chart data from each patient's record that we use as input is coded based on physician documentation, which could arguably also be contaminated with racial bias. The codes listed in the patient's previous visits may also be selectively detailed or biased for some patients, depending on the number of times they visit the health system or on the quality of facilities they visit. Nevertheless, the machine learning algorithm takes as input the *same data* as what is also available to the physician; taking as given any discriminatory actions made in the past. The algorithm could potentially "de-bias" the manifestations of past systemic biases when making risk predictions, but only as long as they systematically and meaningfully correlate with *both* stroke state and the patient's race. It is therefore not unreasonable to also expect practicing physicians to acknowledge and factor-in the existence of any such distinct systematic biases relating race and stroke.

- *The physicians' "problem" vs the algorithm's*
  The algorithm is effectively making a binary class predication of stroke vs not stroke. The physician, on the other hand, might also have in mind other illnesses relevant to the presentation.[34] This distinction is important and matters a lot more in the decision of determining the final diagnosis. At the level of testing however, which is where we use the algorithmic predictions, it is reasonable to assume that physicians' objectives are also to simply match testing decision to the underlying state i.e. to minimize classification errors. Since neuroimaging (testing) is a critical step in the stroke protocol, any patient fairly suspected of stroke will likely be tested. The decision to test then really depends only on whether the physician suspects stroke or not.

- *Advantages or disadvantages of the machine learning algorithm over the physician*
  The algorithm has the following advantages:
  - Physicians, unlike the machine learning algorithm, have cognitive constraints on the dimension of input vector that they can process.
  - The algorithm trains on the latent stroke state S that we have the benefit of inferring retrospectively. It is not obvious that physicians also learn of all diagnostic errors that they or other physicians make.

  On the other hand, the algorithm is also disadvantaged in some ways:
  - The algorithm and the physician begin with observing the same set of observables, but physicians have the advantage of being able to incorporate nuanced context in the case of non-standard or complex presentations, gather additional information in a tailored way, and also use unobservable cues. (Iakovlev & Liang, 2024).
  - ML models are only as good as the data they are trained on. While our algorithm trains only on visits made in 2016, physicians can learn from and update their "prediction model" over several years.

[34] This relates to the issue of *omitted-payoffs bias* discussed in Kleinberg et al. (2018).

D. DETAILS OF STRUCTURAL ANALYSIS

D.1. Model of Physicians' Testing Decision and Extensions

Physicians' testing thresholds can be conceptualized to come from the minimization of false negative and false positive errors (Chan et al., 2022; Arnold et al., 2022). Suppose that the utility cost is given by

$$U_{ij} = -T_{ij}(1 - S_i) - C_j^{z(i)}(1 - T_{ij})S_i$$

where the cost of a false positive (testing a non-stroke case) is normalized to -1, and $-C_j^z$ denotes the physician's disutility from a false negative relative to a false positive. The utility-maximizing physician would test if and only if the expected cost of testing a patient (from making a false positive decision) was lower than expected cost of not testing (from making false negative decision), where the expectations are defined over the stroke risk assessment. The physician would therefore order a test only if they assess the patient's stroke risk to be greater than the threshold $\tau_j^z = \dfrac{1}{1 + C_j^z}$

D.2. Modeling Sample Selection

TABLE D1 :
Sample Selection for physician-level modeling

Step	Description	No. of Observations
0	Primary sample of patients visits to the ED with at least one stroke symptom	1,368,560 visits 11,279 physicians
1	40 percent of the primary sample (randomly sampled) not used in training the risk prediction algorithm	547,424 visits 9,331 physicians
2	Exclude visits by patients with any contraindications to testing (neuroimaging)	525,788 visits 9,150 physicians
3	Exclude visits to physicians who see fewer than 50 patients of each race in the remaining sample	224,129 visits 839 physicians
4	Exclude visits to physicians who test fewer than five patients	73,006 visits 239 physicians

D.3. Estimation using Bayesian Procedures: Gibbs Sampling

The data $\mathcal{D}$ observed are physician-level testing decisions made by physician $j \in J$ for patient visits $i \in I_j$. Let $\mathbf{\Theta}$ denote the vector of hyper-parameters in Equation 11; and $\mathbf{\Lambda}$ be the vector of other parameters, partitioned into random coefficients $\mathbf{\Lambda_1}$ and fixed parameters $\mathbf{\Lambda_2}$.

The likelihood of physician j's testing decisions, conditional on Λ, is given by

$$\mathcal{L}(\mathcal{D}_j|\Lambda) = \prod_{i \in I_j} \mathbb{P}(T_{ij}|\mathbf{\Lambda})^{w_{ij}}$$

where I_j is the set of patient visits assigned to the physician and w_{ij} the class weight proportional to the inverse of the share of j's patients with the same testing decision as T_{ij}.

With $\mathbf{\Lambda}_1$ drawn from population distributions with hyper-parameters in $\mathbf{\Theta}$, the mixed model likelihood of decisions made by all physicians $j \in J$ is

$$\begin{aligned}\mathcal{L}(\mathcal{D}|\mathbf{\Theta}, \mathbf{\Lambda}_2) &= \prod_{j \in J} \mathcal{L}(\mathcal{D}_j|\mathbf{\Theta}, \mathbf{\Lambda}_2) \\ &= \prod_{j \in J} \int \mathcal{L}(\mathcal{D}_j|\mathbf{\Lambda}_1, \mathbf{\Lambda}_2)\, g(\mathbf{\Lambda}_1|\mathbf{\Theta})\, d\mathbf{\Lambda}_1 \end{aligned} \tag{15}$$

where g is the density of the population distribution specified in Equation (11).

The joint posterior for $\mathbf{\Theta}, \mathbf{\Lambda}_2$ is

$$K(\mathbf{\Theta}, \mathbf{\Lambda}_2|\mathcal{D}) \propto \mathcal{L}(\mathcal{D}|\mathbf{\Theta}, \mathbf{\Lambda}_2)\, k(\mathbf{\Theta}, \mathbf{\Lambda}_2) \tag{16}$$

where $k(\mathbf{\Theta}, \mathbf{\Lambda}_2)$ is the prior. Drawing from this posterior is computationally challenging because of the integral in the likelihood does not have a closed form. Alternatively, we draw from the joint posterior of $\mathbf{\Theta}, \mathbf{\Lambda}_2$ and $\mathbf{\Lambda}_1$, given by

$$K(\mathbf{\Theta}, \mathbf{\Lambda}|\mathcal{D}) \propto \prod_{j \in J} \mathcal{L}(\mathcal{D}_j|\mathbf{\Lambda})\, g(\mathbf{\Lambda}_1|\mathbf{\Theta})\, k(\mathbf{\Theta}, \mathbf{\Lambda}_2) \tag{17}$$

To draw from this posterior, we use the Gibbs sampling procedure that sequentially draws one parameter at a time, conditional on the values of all the other parameters. One iteration of the Gibbs sampler sequentially draws from the respective conditional posterior of all parameters. The resulting draws, over several iterations, converge to draws from the joint posterior.

We use conjugate priors to obtain the conditional posterior of the hyper-parameters.

$$\begin{aligned}\begin{bmatrix} b_j & \boldsymbol{\beta_j} \end{bmatrix}' &\sim \mathcal{N}\left(\boldsymbol{\mu}_\beta, \mathbf{\Omega}_\beta\right) \\ \boldsymbol{\varsigma_j} &\sim \mathcal{N}\left(\boldsymbol{\mu}_\varsigma, \mathbf{\Omega}_\varsigma\right)\end{aligned} \tag{18}$$

With $\tilde{\boldsymbol{\beta}}_j = \begin{bmatrix} b_j & \boldsymbol{\beta}_j \end{bmatrix}' \sim \mathcal{N}\left(\boldsymbol{\mu}_\beta, \boldsymbol{\Omega}_\beta\right)$, we assume a diffuse multivariate normal prior on $\boldsymbol{\mu}_\beta$, and a diffuse inverted Wishart prior on $\boldsymbol{\Omega}_\beta$ with K degrees of freedom and $\mathbb{I}_K$ scale matrix, where K is the dimension of $\tilde{\boldsymbol{\beta}}_j$. The posterior for $\boldsymbol{\mu}_\beta$ conditional on $\boldsymbol{\Omega}_\beta$ and $\{\tilde{\boldsymbol{\beta}}_j\}_{j\in J}$ is then $\mathcal{N}\left(\sum_j \tilde{\boldsymbol{\beta}}_j/J,\, \boldsymbol{\Omega}_\beta/J\right)$; and the posterior for $\boldsymbol{\Omega}_\beta$ conditional on $\boldsymbol{\mu}_\beta$ and $\{\tilde{\boldsymbol{\beta}}_j\}_{j\in J}$ is IW $\left(K+J, \dfrac{K\mathbb{I}_K + J\overline{\boldsymbol{S}}}{K+J}\right)$ where $\overline{S} = \sum_j (\tilde{\boldsymbol{\beta}}_j - \boldsymbol{\mu}_\beta)(\tilde{\boldsymbol{\beta}}_j - \boldsymbol{\mu}_\beta)'/J$.

Likewise, we assume a diffuse multivariate normal prior on $\boldsymbol{\mu}_\varsigma$, and diffuse inverted Wishart prior on $\boldsymbol{\Omega}_\varsigma$ with (K-1) degrees of freedom and $\mathbb{I}_{(K-1)}$ scale matrix to obtain the respective conditional posteriors, where $K-1$ is the dimension of $\boldsymbol{\varsigma}_j$.

The conditional posterior of the parameters in $\boldsymbol{\Lambda}_1$, given $\boldsymbol{\Theta}$, are obtained using the hyper-priors $g(\boldsymbol{\Lambda}_1|\boldsymbol{\Theta})$, and the parameters are drawn using Metropolis-Hastings algorithm. For the fixed parameters $\boldsymbol{\Lambda}_2$, we assume an uninformative prior and again use Metropolis-Hastings algorithm to make draws from the conditional posterior.

A total of 1000 sample draws are obtained from 20,000 iterations by burning-in 10,000 draws and retaining every tenth draw from the remaining iterations.

D.4. Identification of Skill Parameters

Lemma 1 *For any choice of threshold $\tau \in (0,1)$, the rate of false positives (FPR_j^z) and the rate of false negatives (FNR_j^z) are increasing in σ_j^z.*

Proof: Consider two physicians, j and j'. Without loss of generality, suppose $\sigma_j^z > \sigma_{j'}^z$. This implies that the distribution of log odds of stroke risk for patients without stroke ($S_i = 0$), as faced by physician j, is a mean-preserving spread of the distribution faced by physician j'.

Let G_j^{0z} and $G_{j'}^{0z}$ denote the corresponding cumulative distributions of log odds as faced by the physicians j and j' respectively; and F_j^{0z}, $F_{j'}^{0z}$ denote the respective cumulative distributions of stroke risk. Since the transformation from probability risk to log odds is strictly monotonic, $\forall x \in (0,1)$ and $y = \log \dfrac{x}{1-x}$, we have $G_j^{0z}(y) = F_j^{0z}(x)$ and $G_{j'}^{0z}(y) = F_{j'}^{0z}(x)$.

Given some risk threshold τ,

$$\text{FPR}_j^z = 1 - F_j^{0z}(\tau)$$
$$\text{FPR}_{j'}^z = 1 - F_{j'}^{0z}(\tau)$$

Since the risk distribution faced by j is a mean preserving spread of the distribution faced by physician j', the cumulative distributions F_j^{0z} and $F_{j'}^{0z}$ must intersect only once. Let k_0 denote this point of intersection. We know then that $\forall\, \tau \geq k_0$, $F_j^{0z}(\tau) \leq F_{j'}^z(\tau) \implies \text{FPR}_j^z \geq \text{FPR}_{j'}^z$.

Likewise, let F_j^{1z} and $F_{j'}^{1z}$ denote the corresponding cumulative distributions of stroke risk

for patients with stroke ($S_i = 1$). Given some threshold τ,

$$\text{FNR}_j^z = F_j^{1z}(\tau)$$

$$\text{FNR}_{j'}^z = F_{j'}^{1z}(\tau)$$

As before, if $\sigma_j^z > \sigma_{j'}^z$, the distribution faced by physician j is a mean-preserving spread of the distribution faced by j'. Then, F_j^{1z} and $F_{j'}^{1z}$ intersect only once, say at point k_1. We know then that $\forall\, \tau \leq k_1$, $F_j^{1z}(\tau) \geq G_{j'}^{1z}(\tau) \implies \text{FNR}_j^z \geq \text{FNR}_{j'}^z$.

We rely here on the thresholds τ being greater than k_0, and lower than k_1. This is a reasonable assumption since in the data we observe the FPR_j^rs and FNR_j^rs to be well below 0.5.

E. RELEVANT CODES FROM MEDICAL CLASSIFICATION SYSTEM

E.1. ICD-10-CM codes/hierarchies that indicate stroke diagnosis

- I60 Nontraumatic subarachnoid hemorrhage
- I61 Nontraumatic intracerebral hemorrhage
- I62 Other and unspecified nontraumatic intracranial hemorrhage
- I63 Cerebral infarction
- I64 Stroke, not specified as haemorrhage or infarction
- G46.3 Brain stem stroke syndrome
- G46.4 Cerebellar stroke syndrome
- I6782 Cerebral ischemia
- I6789 Other cerebrovascular disease
- I679 Cerebrovascular disease, unspecified

Note: Excludes transient cerebral ischemic attacks (TIA) and sequelae of cerebrovascular disease (I69)

E.2. ICD-10-CM codes/hierarchies for stroke-related symptoms

1. General Symptoms
    - R0602 Shortness of breath
    - R51 Headache
    - R52 Pain, unspecified
    - R53.1 Weakness
    - R53.8 Other malaise and fatigue
    - R55 Syncope and collapse
    - R11.0 Nausea
    - R11.10 Vomiting, unspecified
    - R11.2 Nausea with vomiting, unspecified
2. Muscular/Neurological Symptoms
    - R27 Other lack of coordination
    - R29.5 Transient paralysis
    - R29.818 Other symptoms and signs involving the nervous system
    - R29.898 Other symptoms and signs involving the musculoskeletal system
    - R29.9 Unspecified symptoms and signs involving the nervous and musculoskeletal

systems

- M62.81 Muscle weakness (generalized)
- R569 Unspecified convulsions

3. Facial (Muscle weakness, Numbness, Drooping) Symptoms
    - R29.810 Facial weakness
4. Symptoms related to Consciousness/Alertness
    - R40.4 Transient alteration of awareness
    - R41.0 Disorientation, unspecified
    - R41.82 Altered mental status, unspecified
    - R41.9 Unspecified symptoms/signs involving cognitive function and awareness
    - R42 Dizziness and giddiness
    - R45.1 Restlessness and agitation
5. Sensory (Pins/Needles, Low sensation of touch, Numbness) Symptoms
    - R20 Disturbances of skin sensation
6. Symptoms related to Speech (Difficulty speaking, slurred speech, or speech loss)
    - R47 Speech disturbances, not elsewhere classified
7. Symptoms related to Vision (Blurred vision, double vision, sudden visual loss, or temporary loss of vision in one eye)
    - H53.12 Transient visual loss
    - H53.13 Sudden visual loss
    - H53.14 Visual discomfort
    - H53.2 Diplopia
    - H53.8 Other visual disturbances
    - H53.9 Unspecified visual disturbance
    - H54.7 Unspecified visual loss

E.3. ICD-10-CM codes/hierarchies for Stroke Mimics

1. G40: Epilepsy and recurrent seizures
2. G43: Migraine
3. G93: Other disorders of brain
4. Tumor in the Brain

- C71: Malignant neoplasm of brain
- D43: Neoplasm of uncertain behavior of brain and central nervous system

5. G58: Other mononeuropathies
6. Toxic or metabolic disorders
    - E87.0: Hyperosmolality and hypernatremia
    - E87.1: Hypo-osmolality and hyponatremia
7. E15: Non-diabetic hypoglycemic coma
8. E16: Other disorders of pancreatic internal secretion (Hypoglycemia)
9. H81: Disorders of vestibular function
10. I65-I68: Other cerebrovascular diseases

E.4. CPT/HCPCS/ICD-10-PCS codes for testing/ neuroimaging

- Computed Tomography (CT) head with/without contrast: 70450, 70460, 70470
- CT Perfusion: 0042T
- CT Angiogram (Head): 70496
- Magnetic Resonance Imaging (MRI) of head: 70551, 70552, 70553
- Magnetic Resonance Angiography: 70544, 70545, 70546, 70447, 70548, 70549
- Carotid Ultrasound (93880, 93882), Transcranial Doppler (93886), 12 Lead Echocardiogram (93000, 93005, 93010)

E.5. ICD-10-CM codes of contraindications to neuroimaging:

Since CT involves exposure to radiation, it is generally contraindicated for patients in the first trimester of their pregnancy. Contrast CT, in particular, is contraindicated for patients who are allergic to the contrast dye, have lower kidney functionality, or have active hyperthyroidism. On the other hand, magnetic resonance modalities use powerful magnetic fields and are contraindicated for patients with electronic or magnetic implants (such as pacemakers) or other metallic foreign bodies (such as bullet fragments, aneurysm clips, piercings, prosthetic limbs) that aren't marked MR safe by the manufacturers. Severely obese patients who exceed the weight capacity of the machines or the circumference of the scanner are also contraindicated.

- Pregnant state, incidental/gestational carrier (Z33.1, Z33.3)
- Radiographic dye allergy (Z91.041)
- Procedure and treatment not carried out because of other contraindication (Z53.09)
- Disorder of kidney and ureter, unspecified (N28.9)
- Presence of implants or devices (Z95-Z97), Retained foreign body fragments (Z18)

- BMI greater than 40 (Z68.4)
- Thyrotoxicosis or hyperthyroidism (E05)

E.6. ICD-10-CM codes of external causes of co-morbidity excluded from the primary sample

- V90 - V97: Water, air, or space transport accidents
- W3 - W9: Contact with machinery, animal attacks or other contact with animate mechanical forces, explosion, exposure to electric current or radiation, or accidental drowning
- X0 - X3: Exposure to smoke, heat, or forces of nature
- X7 - X9: Intentional self-harm
- Y35 - Y38: Legal intervention, operations of war, military operations, and terrorism