NBER WORKING PAPER SERIES

FIXING MISALLOCATION WITH GUIDELINES: AWARENESS VS. ADHERENCE

Jason Abaluck
Leila Agha
David C. Chan Jr
Daniel Singer
Diana Zhu

Working Paper 27467
http://www.nber.org/papers/w27467

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
July 2020, Revised July 2021

For helpful comments and suggestions we thank Jonathan Gruber, Erzo Luttmer, Rachael Meager, David Molitor, Paul Novosad, Ziad Obermeyer, Stephen Ryan, Fiona Scott Morton, Jonathan Skinner, Sachin Shah, Frank Sloan, Doug Staiger, Mintu Turakhia, and Stefan Wager. We are grateful to Mohit Agrawal, Sophie Andrews, Samuel Arenberg, Liberty Greene, Johnny Huynh, Chris Lim, Uyseok Lee, and Natalie Nguyen for invaluable research assistance. We thank Carl vanWalraven and the Atrial Fibrillation Investigators for generously sharing the AFI database for reanalysis. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

At least one co-author has disclosed additional relationships of potential relevance for this research. Further information is available online at http://www.nber.org/papers/w27467.ack

Fixing Misallocation with Guidelines: Awareness vs. Adherence
Jason Abaluck, Leila Agha, David C. Chan Jr, Daniel Singer, and Diana Zhu
NBER Working Paper No. 27467
July 2020, Revised July 2021
JEL No. I11,I18,O33

ABSTRACT

Expert decisions often deviate from evidence-based guidelines. If experts are unaware of guidelines, dissemination may improve outcomes. If experts are aware of guidelines but continue to deviate, promoting stricter adherence has ambiguous effects on outcomes depending on whether experts have information not in guidelines. We study guidelines for anticoagulant use to prevent strokes among atrial fibrillation patients. By text-mining physician notes, we identify when physicians start using guidelines. After mentioning guidelines, physicians become more guideline-concordant, but adherence remains far from perfect. To evaluate whether nonadherence reflects physicians' superior information, we combine observational data on treatment choices with machine learning estimates of heterogeneous treatment effects from eight randomized trials. Most departures from guidelines are not justified by measurable treatment effect heterogeneity. Promoting stricter adherence to guidelines could prevent 24% more strokes, producing much larger gains than broader guideline awareness.

Jason Abaluck
Yale School of Management
Box 208200
New Haven, CT 06520-8200
and NBER
jason.abaluck@yale.edu

Daniel Singer
Harvard Medical School and
Massachusetts General Hospital
55 Fruit St # 148
Boston, MA 02114
DESINGER@mgh.harvard.edu

Leila Agha
Department of Economics
Dartmouth College
6106 Rockefeller Hall
Hanover, NH 03755
and NBER
leila.agha@dartmouth.edu

Diana Zhu
Yale University
dianayilunzhu@gmail.com

David C. Chan Jr
Center for Health Policy and
Center for Primary Care and Outcomes Research
117 Encina Commons
Stanford, CA 94305
and NBER
david.c.chan@stanford.edu

1 Introduction

In medicine, law, science, and many other settings, expert decision-makers frequently deviate from guideline recommendations (Arrowsmith et al., 2015; Grimshaw and Russell, 1993; Prior et al., 2008; Stevenson and Doleac, 2019). Experts may deviate because they lack awareness of guidelines. In this case, guideline dissemination may reduce deviations and improve outcomes as long as guidelines are well-crafted. If experts are aware of guidelines but continue to deviate, promoting stricter adherence may worsen or improve outcomes, depending on whether experts have additional information not encoded in guidelines.

Understanding deviations from guidelines is especially urgent in healthcare, where research documents large inefficiencies in care allocation.[1] Clinical guidelines have been the principal strategy used to encourage evidence-based care, with approximately 250,000 peer-reviewed papers about clinical scoring systems published over the past 50 years (Challener et al., 2019). Efforts to encourage greater guideline adherence have been criticized because they discount the role of physician expertise in tailoring individualized treatments (Basu et al., 2014; Costantini et al., 1999; Woolf et al., 1999). Important questions in health economics and policy center around whether greater guideline awareness or adherence would correct or exacerbate care misallocation.

In this paper, we study how physicians employ an existing clinical guideline and then evaluate their treatment choices using novel machine learning (ML) estimates of treatment effects that incorporate many patient characteristics absent from current guidelines. This approach allows us to test whether physician treatment choices depend on information about treatment effects not encoded in guidelines and to investigate how guideline awareness affects this dependence. We assess the impact on patient outcomes of both guideline "awareness," the change in decision-making when physicians begin to note a guideline in the clinical records, and stricter "adherence," the degree to which a physician follows guideline recommendations.

We focus on the clinical setting of atrial fibrillation, a common condition afflicting more than 5 million people in the US (Colilla et al., 2013). The principal risk in atrial fibrillation is debilitating or deadly ischemic stroke (hereafter, stroke); untreated patients have a 5% risk of stroke per year (Atrial Fibrillation Investigators, 1995). Anticoagulation (blood thinning) has been shown in clinical trials

[1] Abaluck et al. (2016) and Ribers and Ullrich (2019) show that physicians allocate testing inefficiently across patients. Mullainathan and Obermeyer (2019) and Chandra and Staiger (2020) show evidence of misallocation in heart attack testing and treatments. Similar misallocations have been shown in the setting of C-sections (Currie and MacLeod, 2017), depression (Currie and MacLeod, 2020), pneumonia (Chan et al., 2019), and emergency department care (Gowrisankaran et al., 2017).

to reduce stroke risk. However, anticoagulation also increases the risk of life-threatening hemorrhage (hereafter, bleed), including intracranial bleeding (Atrial Fibrillation Investigators, 1995). Therefore, the consequences of misallocating anticoagulation can be serious. In response to these stakes, researchers have developed the $CHADS_2$ score: a simple predictive score of stroke risk for patients with atrial fibrillation (Gage et al., 2001). Clinical guidelines recommend tailoring treatment decisions on the basis of patients' $CHADS_2$ scores (Fuster et al., 2006; Hirsh et al., 2008). The $CHADS_2$ score is among the most well known and widely used risk scores used for any clinical condition.[2]

We study guideline awareness and treatment decisions of nearly 5,800 physicians treating 113,000 newly diagnosed atrial fibrillation patients in the Veterans Health Administration (VHA) from 2002-2013. For each physician, we measure the date that the physician first incorporates the $CHADS_2$ guideline into their decision-making, by identifying the earliest mention of the $CHADS_2$ score in the physician's clinical notes. Following the publication of $CHADS_2$-based treatment guidelines in 2006, we see steady growth in physicians becoming aware of the guideline. Prior to awareness, physicians treat roughly 50% of patients with atrial fibrillation, and treatment probability is largely invariant to the $CHADS_2$ score. After the first $CHADS_2$ mention, practice patterns pivot towards $CHADS_2$-based recommendations: prescriptions to patients with low risk scores fall by 4.9 percentage points, while prescriptions to patients with high risk scores increase by 1.6 percentage points. Despite these changes, physicians who are aware of the $CHADS_2$ score still fail to adhere to guidelines in more than 40% of cases. Most non-adherence is not explained by a lack of guideline awareness.

To assess the benefits of guideline awareness as well as the possible benefits of stricter adherence, we need to understand how treatment effects vary across patients. To this end, we generate novel ML estimates of heterogeneous treatment effects from randomized control trial (RCT) microdata. To estimate heterogeneous treatment effects, we use detailed patient characteristics, clinical outcomes, and randomized treatment status of each patient from eight RCTs in the Atrial Fibrillation Investigators database (hereafter, AFI database) (van Walraven et al., 2009). Using a causal-forest model (Wager and Athey, 2017; Asher et al., 2016), we obtain estimates of conditional average treatment effects (CATEs) on strokes and bleeds that vary both with patient characteristics included in the $CHADS_2$ score and with other characteristics. We further compute best linear predictions (BLPs) of the underlying CATEs using a method described by Chernozhukov et al. (2018), which we use in our

[2]Researchers affiliated with the Mayo Clinic report that the $CHADS_2$ and its successor the CHA_2DS_2-VASc were the most common search queries in their internal clinical decision support tool (Challener et al., 2019). `MDCalc.com`, the popular website for calculating risk scores, currently lists CHA_2DS_2-VASc as second-highest in popularity and $CHADS_2$ as sixth-highest in popularity.

subsequent analysis.

We find substantial heterogeneity in stroke treatment effects. At the 90th percentile, warfarin reduces strokes by 10 percentage points, while at the 10th percentile, it reduces strokes by 4 percentage points. In contrast, we cannot detect conclusive evidence of heterogeneity in bleed treatment effects in the AFI database. The $CHADS_2$ score explains a substantial share of our estimated variation in stroke treatment effects ($R^2 = 0.67$). The median CATE for patients with the lowest $CHADS_2$ score is -0.03, while the median CATE for patients with the top highest three scores is -0.10. Nonetheless, there remains meaningful residual variation that we can validate across RCTs in the AFI database.

Exporting the ML-estimated treatment effects to the VHA, we estimate a model of clinician treatment decisions. Our goal is to understand the extent to which treatment deviations from $CHADS_2$ recommendations may be motivated by residual variation in stroke treatment effects that are not codified in the guideline. Our first finding is that new awareness of the $CHADS_2$ score (proxied by first mention of the score in a doctor's clinical notes) leads doctors to place greater decision weight on $CHADS_2$-related variation in stroke treatment effects. However, consistent with the reduced-form evidence above, responsiveness to variation in $CHADS_2$-related treatment effects remains low in absolute terms. Our second finding is that residual variation in stroke treatment effects (i.e., orthogonal to the $CHADS_2$ score) explains even less variation in treatment decisions.

These results suggest that departures from guidelines in this important setting generally worsen patient outcomes, but this interpretation relies on two key assumptions. First, our estimates of treatment effects in the AFI data must be externally valid in the VHA data. To assess the external validity of our estimates across trials, we demonstrate that treatment effects estimated on subsets of our trial database predict treatment effects in "out-of-bag" trials not used in estimation. Mean patient characteristics in the VHA data are "in the support" of mean patient characteristics across trials in the AFI data. Further, we show that treatment effects estimated from the AFI database are consistent with key patterns in our observational data: observational differences between treated and untreated patients are strongly correlated with RCT-estimated CATEs, and stroke outcomes for untreated patients are more frequent for patients with larger RCT-estimated CATEs.

Our second major assumption is that doctors are not making treatment decisions based on variation in treatment effects that cannot be predicted by the covariates in the AFI data. The attributes in the AFI data were included precisely because physicians plausibly believed they might impact the benefits and costs of anticoagulant treatment. Nonetheless, other attributes may also be relevant. To investigate this, we add to the model detailed patient characteristics that may relate to the benefits

of treatment despite not being available in the RCT data; these include variables that predict patient medication adherence, bleed risk, and fall risk. These characteristics explain very little of the variance in treatment choices given estimated treatment effects and do not impact the other coefficients in the model. This suggests that deviations from guidelines are not easily reconciled by other considerations raised by the clinical literature but not measured in the AFI database.

We consider a variety of counterfactual scenarios, using the model to simulate the impact of broader awareness or stricter adherence to the existing guideline. Our findings suggest that physicians allocate treatments to patients with atrial fibrillation about as well as a *random* decision rule. Extending awareness of the $CHADS_2$ score to all physicians would slightly improve treatment allocation, leading to a 1% improvement in strokes prevented per bleed induced. The benefits of strict adherence to current guidelines are much larger than the benefits of universal guideline awareness. Reallocating the same number of treatments in the observed allocation to patients with the highest $CHADS_2$ scores would prevent 18% more strokes per bleed induced than the status quo. Strict adherence to a guideline which incorporates all validated ML information about heterogeneity in stroke treatment effects would prevent 24% more strokes per bleed induced than the status quo. These results suggest that policies that aim to increase adherence may produce much larger improvements in patient outcomes than policies that only broaden guideline use by increasing awareness.

Our research relates to several strands of literature. First, we contribute to an active literature in economics studying the potential for machine-based algorithms to improve decision-making. Evidence from judge bail decisions (Kleinberg et al., 2018) and clinical care (Abaluck et al., 2016; Mullainathan and Obermeyer, 2019) suggests that human experts make frequent mistakes that could be corrected by optimal guidelines, but do not analyze how guidelines impact behavior in practice. Finkelstein et al. (2021) suggest that experts deviate from guidelines even for themselves and close relatives, but they do not directly link adherence to health outcomes. Hoffman et al. (2018) finds that managers in their hiring decisions frequently overrule a technology-driven hiring recommendation, but that doing so worsens outcomes. We build on this research by studying highly skilled experts making an important clinical decision. We analyze how new awareness of a guideline changes behavior and outcomes, in addition to simulating the effects of stricter adherence to existing and novel guidelines using estimates of heterogeneous treatment effects.

Prior papers comparing machine decisions to human discretion have typically relied on observational data and quasi-experimental assumptions to reach conclusions about misallocation.[3] A recent

[3]Many of these papers compare treatments and outcomes across decision-makers, usually assuming a common ranking

econometric literature has pointed out that many quasi-experiments recovering treatment effects averaged across subgroups of data may nevertheless fail to meet the stricter assumptions required for identifying *heterogeneous* treatment effects, estimated from within each subgroup of data (e.g., Kolesar et al., 2015; de Chaisemartin, 2017; Frandsen et al., 2019). We circumvent these concerns by using RCTs to develop measures of treatment effect heterogeneity. In our application of this method, we also propose methods to investigate the external validity of results from RCTs in observational data, addressing some of the concerns about extrapolation raised by (Manski, 2017).

Einav et al. (2019) and Oster (2020) both report that healthier patients are more likely to take up recommended health screenings and diets. This selection can bias observational estimates and may imply that those treated as a result of guidelines have smaller treatment effects than inframarginal patients. We contribute to this investigation of selection under evidence-based recommendations. In our setting, we study guidelines for physician decisions, rather than patient health behaviors. We find that the marginal treated patients due to the introduction of the $CHADS_2$ score have larger treatment effects, but physician decisions nevertheless achieve worse outcomes relative to strict adherence.

In the medical literature, research has focused on how treatment decisions relate to clinical guidelines. The literature has shown widespread lack of adherence, not only in the case of the $CHADS_2$ score, but also across many clinical risk scores and guidelines.[4] Our paper builds on this literature by documenting that even adopting physicians who are aware of the guideline continue to deviate frequently. The crucial difficulty in interpreting non-adherence is the lack of evidence on whether counterfactual treatment decisions promoted by guidelines will improve health outcomes. For example, Mehta et al. (2015) report that physicians who adhere more closely to guidelines have better outcomes but do not separate the impact of guidelines from other differences across physicians. We address this difficulty by combining evidence on guideline adherence with causal estimates of treatment effects.

The remainder of this paper is organized as follows. Section 2 provides clinical background. Section 3 describes our data. Section 4 provides reduced form evidence of the impact of $CHADS_2$ awareness on treatment. Section 5 presents our estimates of causal-forest treatment effects. Section 6

of cases across decision-makers (e.g., Abaluck et al., 2016; Kleinberg et al., 2018; Chandra and Staiger, 2020). Newer approaches in Chan et al. (2019) and Arnold et al. (2020) relax this assumption but restrict the direction of treatment effects. In another vein, recent papers have applied ML techniques to predict outcomes using observational data, again with necessary quasi-experimental assumptions because outcomes are selectively observed (e.g., Ribers and Ullrich, 2019; Mullainathan and Obermeyer, 2019).

[4] An older review article estimated that 40% of patients were not receiving guideline-recommended care for chronic conditions (Schuster et al., 1998). More recent research suggests non-adherence to guidelines continues to be widespread across a variety of clinical contexts (Lasser et al., 2016; Valle et al., 2015; Chen et al., 2015; Rosenberg et al., 2015). For the $CHADS_2$ score specifically, Chapman et al. (2017) find evidence of substantial non-adherence to the guideline.

models how guideline awareness impacts the relationship between treatment behavior and treatment effects. Section 7 considers counterfactual policies. Section 8 concludes with a discussion of policy implications.

2 Atrial Fibrillation and the $CHADS_2$ Score

Atrial fibrillation is the most common cardiac arrhythmia. It afflicts over 5 million Americans; for adults older than 40 years, one in four will develop the condition (Hsu et al., 2016). Atrial fibrillation increases stroke risk by five-fold and is responsible for 40% of strokes among patients older than 80 years (Piccini and Fonarow, 2016). The main treatment to reduce stroke risk among patients with atrial fibrillation is anticoagulation by warfarin.[5] While anticoagulation is effective in reducing stroke risk, by 68% on average, it has also been shown to increase the risk of major bleeding by more than twofold (Atrial Fibrillation Investigators, 1995; Kearon et al., 2012). Given the large potential benefits and risks of anticoagulation, an important task for clinicians evaluating patients with atrial fibrillation is to decide which patients to treat with anticoagulation.

Efforts to improve anticoagulation targeting have largely focused on predicting stroke risk, with the intuition that the benefits of anticoagulation are likely increasing in baseline stroke risk. Earlier studies re-analyzed data from the control arms of randomized trials of patients with atrial fibrillation to find hypertension and prior stroke as important risk factors of stroke (Atrial Fibrillation Investigators, 1995; Stroke Prevention in Atrial Fibrillation Investigators, 1995). Building on this work, the $CHADS_2$ score was first formulated by Gage et al. (2001), using registry data comprising 1,733 Medicare patients, and later validated for clinical practice by Gage et al. (2004). In 2006, the American College of Cardiology (ACC) became the first specialty society to issue a guideline recommending treatment decisions based on the $CHADS_2$ score (Fuster et al., 2006). Other professional societies followed, with the American College of Chest Physicians (ACCP) recommending $CHADS_2$-based treatment decisions in 2008 (Hirsh et al., 2008).

Designed to be easy to use, the $CHADS_2$ score is an index of five patient characteristics: "C" for congestive heart failure (1 point), "H" for hypertension (1 point), "A" for age $\geq$ 75 years (1 point), "D" for diabetes (1 point), and "S" for stroke (2 points) (Table 1). Since its introduction, the $CHADS_2$

[5] In our sample, fewer than 2% of patients are prescribed alternative anticoagulants. Novel oral anticoagulants (NOACs) were introduced near the end of our sample, with the FDA approval of dabigatran in 2010, rivaroxaban in 2011, and apixaban in 2012. Based on non-inferiority trials, they are similarly effective in preventing stroke with possibly lower risks of bleeding (Lane and Lip, 2012). Guideline recommendations for their use (vs. no anticoagulation) rely on the same stroke risk scores. Warfarin continues to be the mainstay drug for anticoagulation in atrial fibrillation (Hsu et al., 2016).

score has become one of the most widely recognized risk scores in clinical practice.[6] However, despite its widespread recognition, studies in a variety of settings have shown that adherence to the $CHADS_2$ score has been low, typically with only half of recommended patients being prescribed anticoagulation (Hsu et al., 2016; Piccini and Fonarow, 2016).

While poor adherence to $CHADS_2$-based treatment recommendations has been linked to many factors, physicians' concerns of increased bleeding risk due to frailty and multi-morbidity are commonly cited. Frailty and multi-morbidity often coexist with atrial fibrillation, since both conditions increase in prevalence with age. Less evidence has existed to guide physicians to assess bleeding risk. The first formal risk score for bleeding (HAS-BLED) was published toward the end of our study period (Lane and Lip, 2012). However, it remains unvalidated in the population of atrial fibrillation patients and is not as widely used in clinical practice. In recent years, more evidence has emerged to support the use of anticoagulation in frail and multi-morbid patients; nevertheless, the uptake of anticoagulation among patients who are frail and have a high-risk of stroke remains low (Fawzy et al., 2019).

3 Data

Our approach combines data from two main sources. We study treatment decisions in the context of guideline awareness using data from the Veterans Health Administration (VHA). We estimate heterogeneous treatment effects using RCT data from the Atrial Fibrillation Investigators (AFI) database.

3.1 Veterans Health Administration Data

Defining the atrial fibrillation cohort. To study initial treatment decisions, we identify patients with a new diagnosis of atrial fibrillation using electronic medical records from the Veterans Health Administration (VHA) from October 2002 to December 2013. Following a protocol developed in previous work, we err on the side of defining a narrower cohort of patients who are more likely both to have a new, confirmed atrial fibrillation diagnosis and to receive care at the VHA (Turakhia et al., 2013; Perino et al., 2017).

[6] In 2010, a modification of the $CHADS_2$ score, the CHA_2DS_2-VASc score, was introduced (Lip et al., 2010). The CHA_2DS_2-VASc score changes the weighting of age and introduces vascular disease as an additional risk factor. Due to the time period covered by our data, from 2003-2013, our analysis focuses principally on the original $CHADS_2$ score. We observe comparatively little use of the CHA_2DS_2-VASc score: while 23% of patient encounters in our data are by physicians who have previously mentioned the $CHADS_2$ score, fewer than 2% of patient encounters are by physicians who have previously mentioned the CHA_2DS_2-VASc in their notes. We do consider vascular disease in our causal-forest model of treatment effects, and in some simulations, we contrast the $CHADS_2$ and CHA_2DS_2-VASc scores.

We first identify potentially new diagnoses of atrial fibrillation (ICD9 code beginning with 427.3) among patients with no previous such diagnosis within three years, extending our data back to October 1999 to perform this exclusion. We also require an electrocardiogram (the primary means to diagnose atrial fibrillation) near the time of initial diagnosis and no anticoagulation prior to the initial diagnosis. After a diagnosis of atrial fibrillation, the anticoagulation decision is typically made by a physician who provides longitudinal care and makes prescription decisions for the patient. Therefore, to attribute patients to physicians who are likely responsible for anticoagulation decisions, we require each patient to have a visit with a VHA cardiologist or primary care physician (PCP) within 90 days after the initial diagnosis (to decide on treatment). Further, the patient must have received at least one drug (other than warfarin) prescribed by the attributed physician within one year before or after the atrial fibrillation diagnosis. We also require each attributed physician to have at least 30 other patients with atrial fibrillation and to have prescribed warfarin for another patient. Our sample restrictions result in an analytic cohort of 113,270 patients (see Table 2 for details).

For each patient in this cohort, we capture a broad array of characteristics that may influence the anticoagulation decision following an initial diagnosis. These characteristics include demographic information, comorbidities, laboratory test results, body measurements, and blood pressure readings. We use these characteristics to construct the $CHADS_2$ score, match the other clinical characteristics recorded in the RCT data, and proxy for other concerns not fully captured in the RCT data (e.g., bleed risk, fall risk, frailty, multi-morbidity).[7] To capture the anticoagulation decision, we rely on VHA records of prescriptions for warfarin or a novel oral anticoagulants (e.g., dabigatran, rivaroxaban, apixaban, edoxaban).[8] Summary statistics on the VHA atrial fibrillation cohort are reported in Table 3, Column 1.

Defining guideline awareness. We measure awareness of the $CHADS_2$ score guideline at the physician level by searching physician visit notes for mentions of the $CHADS_2$ score.[9] We proxy the tim-

[7]For further details on the additional patient comorbidities and risk factors we extracted from the VHA sample, see Appendix Section A.2.

[8]The VHA records include prescriptions that are dispensed by the VHA as well as prescriptions that are paid for by the VHA. Recall that the vast majority of prescriptions in our sample are for warfarin. Fewer than 2% of patients in our sample are prescribed a novel oral anticoagulant (NOAC). Among patients prescribed an anticoagulant, 4% are prescribed a NOAC.

[9]Recall that physicians in our analytic sample are cardiologists and PCPs who have each treated at least 30 atrial fibrillation patients. We increase our detection of $CHADS_2$ mentions for these physicians by using visit notes within 6 months of initial diagnosis in our broad cohort of 844,312 atrial fibrillation patients, who may be patients with previously established diagnoses (see Table 2), and search for non-case-sensitive occurrences of the string `chads`. We settled on this string after spot-checking several variants for false positives. Consistent with our spot-checking results, we find no positive mentions of this string in the first two years of our data, prior to diffusion of the $CHADS_2$ score. This suggests that false positives are very rare.

ing of a physician's first awareness of the guideline by her first clinical note mentioning the $CHADS_2$ score, and we define her as guideline-aware if she has previously mentioned the $CHADS_2$ score in a note. Our use of the term "awareness" here is an imperfect shorthand: physicians may be literally aware of a guideline without it ever impacting their practice; our goal is to understand how physicians change their behavior when they decide to incorporate the guideline into their decision-making.[10]

In Figure 2, we show that almost no physician mentioned the $CHADS_2$ score in the period prior to the ACC guideline in 2006, despite the fact that the $CHADS_2$ score was introduced in 2001 and validated in 2004. After 2006, we find a steady rise in the proportion of physicians who have previously mentioned the $CHADS_2$ score at least once, approaching 70% near the end of our study period in 2013. We note that this represents a lower bound to awareness of the $CHADS_2$ score, as physicians may be aware of the score yet not mention it in their notes.

As we will show in Section 4, for physicians who eventually mention the $CHADS_2$ score, treatment decisions become more guideline-concordant at the time of their first mention in clinical notes. We categorize physicians without any note mentioning the $CHADS_2$ score as never being aware of the guideline. As an important caveat, it is likely that many physicians whom we label as "never aware" have actually heard of the guideline, but do not meet the stringent definition of "awareness" that we impose here. Empirically, we will document that these "never aware" physicians deviate more frequently from guideline-based care. Thus, although 'never aware" physicians may have heard of the guideline, they are unlikely to be making explicit use of its recommendations to drive their clinical decision-making.

3.2 Atrial Fibrillation Investigators Database

To estimate heterogeneous treatment effects—which we use to assess physician decision-making and to evaluate the likely effects of counterfactual anticoagulation decisions on patient outcomes—we rely on the Atrial Fibrillation Investigators database (hereafter, AFI database). The AFI database contains patient-level observations from eight trials in which patients were randomized to anticoagulants versus a placebo or control.[11] Details of the AFI database have been documented elsewhere (e.g., van

[10] An earlier draft of the paper referred to this note-mentioning as "adoption" which is also imperfect. Our analysis below quantifies whether "awareness" in the sense above leads to behavior change.

[11] There are a total of ten trials in the original AFI database. For our analysis, we define patients who were treated with aspirin alone as being untreated with anticoagulation. We drop observations for patients on low Warfarin or low Warfarin plus aspirin therapy. After these modifications, eight trials remain with both treatment and control arms. In three of the eight trials, patients are divided into eligible versus ineligible groups for anticoagulation and then randomized among eligible patients. We treat the ineligible patients as separate trials (with only one treatment arm) and use data from all trials in the causal-forest implementation to increase power.

Walraven et al., 2009).

The AFI database was previously compiled by investigators to explore heterogeneity in risk and in treatment effects. Previous analyses using the database have selected patient characteristics heuristically (van Walraven et al., 2002, 2009). For each patient in the AFI database, we observe randomization status and subsequent stroke and bleeding events. In harmonizing data across the clinical trials, the investigators consistently recorded several important patient characteristics at the time of randomization, including all variables underlying the $CHADS_2$ score, as well as several additional variables, including further detail on demographics, height, weight, blood pressure, hemoglobin, smoking status, comorbidities, and history of transient ischemic attack (TIA), stroke, anginal symptoms, and myocardial infarction. In Appendix Table A.1, we list the full set of characteristics that we use from the AFI database. In Appendix Table A.3, we report the results of balance tests, which suggest successful randomization in these clinical trials.

4 $CHADS_2$ Guideline Awareness and Adherence

The 2006 ACC and 2008 ACCP guidelines recommended treating patients with a $CHADS_2$ score of 0 or 1 potentially with aspirin alone and treating patients with a $CHADS_2$ score of 2 or above with anticoagulation (Hirsh et al., 2008; Fuster et al., 2006).[12] We therefore begin our analysis by describing trends in prescribing behavior for groups of patients defined by their $CHADS_2$ score.

Figure 1 displays trends in anticoagulation rates for patients with low risk of stroke ($CHADS_2$ score of 0 or 1), patients with moderate risk of stroke ($CHADS_2$ score of 2 or 3), and patients with high risk of stroke ($CHADS_2$ score of 4 or greater). Prior to the 2006 guideline, patients with lower $CHADS_2$ scores were remarkably *more* likely to be treated with anticoagulation. This relationship held both between patients with low vs. moderate stroke risk and between patients with moderate vs. high stroke risk. We will later show that this pattern can be explained in part by physicians' reluctance to treat multimorbid patients with high mortality risk. In the years after 2006, we observe a gradual reduction in anticoagulation rates for low-risk patients for whom the guideline allowed for management without anticoagulation. There appears to be a small increase in treatment rates for patients at moderate or high risk. However, even among groups where anticoagulation is recommended, prescription rates remain below 55% for our sample period. Patients with high stroke risk ($CHADS_2$

[12] In the 2006 and 2008 guidelines, patients with a $CHADS_2$ score of 0 were recommended aspirin, while patients with a $CHADS_2$ score of 1 could be treated with either aspirin or anticoagulation. Later guidelines suggested anticoagulation for patients with a $CHADS_2$ score of 1 and some patients with a $CHADS_2$ score of 0 but a CHA_2DS_2-VASc score of 1 (Lip et al., 2010). This clinical consensus applied mostly to after our study period.

score of 4 or greater) remain slightly less likely to be treated than patients with moderate stroke risk ($CHADS_2$ of 2 or 3).

It is evident from Figure 1 that physicians had only weak adherence to recommendations to treat patients with high $CHADS_2$ risk scores and to leave patients with low $CHADS_2$ risk scores untreated. We proceed by estimating the causal impact of guideline awareness on prescription choice and investigating how guideline adherence changes after a physician incorporates the guideline into her clinical practice. In the Panel A of Figure 3, we first plot how anticoagulation rates among high- and low-score patients change following an adopting physician's first note mentioning the $CHADS_2$ score. Following this event, the anticoagulation rate for low-scoring patients drops by several percentage points, while the anticoagulation rate for higher-scoring patients increases slightly.

We further assess the effect of awareness on adherence by an event-study regression separately for the two guideline-relevant groups of patients:

$$W_i = \sum_{r=-5}^{5} \mathbf{1}\left(r\left(i\right) = r\right)\theta_r + \eta_{d(i)} + \xi_{t(i)} + \varepsilon_i. \tag{1}$$

$W_i \in \{0, 1\}$ indicates whether patient i was anticoagulated, and $r(i)$ is a function that returns the year of i's visit relative to the prescribing physician's becoming aware of the $CHADS_2$ score. The regression includes fixed effects for the prescribing physician, $d(i)$, and for the year, $t(i)$. We estimate Equation (1) separately for patients with $CHADS_2$ scores of 0 and 1 and for those with $CHADS_2$ scores 2 and higher. Panel B of Figure 3 displays estimation results of the θ_r coefficients and is broadly consistent with the raw treatment rates shown in Panel A. Treatment rates for low-score patients decline by 4.7 percentage points (standard error of 1.6 percentage points), while they increase by 1.7-percentage points (standard error of 1.1 percentage points) for higher-score patients.[13]

Taken together, these results suggest that although the $CHADS_2$ score was becoming widely known, adherence to anticoagulation recommendations increased only modestly with physician awareness. Prior to the 2006 guideline, almost no physician appeared to be using the $CHADS_2$ score in documented clinical decision-making, yet by the end of 2013, the vast majority had explicitly mentioned it in their notes. While our event-study results suggest a clear behavioral shift in prescribing at the time of becoming aware of the $CHADS_2$ score, most of this response is from avoiding treatment for low-risk patients, not increasing treatment for high-risk patients. In Figure A.1, we further show that, while adherence varied substantively across physicians, few physicians reached an adherence

[13]This estimate of changes in treatment rates comes from aggregating the results shown in Figure 3. Specifically, we calculate the difference between the average level in years 0 through 4 minus the average level in years −5 through −1.

rate of 80%.

Guideline awareness is more difficult to measure than adherence. However, two features of our setting lend greater confidence to our results on the effect of awareness and allow us to understand patterns of adherence in the context of guideline awareness. First, the lack of pre-trends in our event-study results in Figure 3 and the timing of the effect suggest that our measure based on clinical documentation coincides with a discrete change in physicians' consideration of the guideline. Second, we witness a dramatic shift in the pace of guideline awareness—from nearly no mentions of the $CHADS_2$ score prior to the ACCP guideline publication, growing to 70% of doctors having mentioned the score by the end of our study. Falsely classifying physicians who have adopted the guideline but have not mentioned the $CHADS_2$ score is thus less of a concern, and we can reconcile our event-study results with the overall shifts in adherence, shown in Figure 1, among all physicians.

These findings are consistent with a clinical literature that emphasizes the importance of reducing stroke risk by anticoagulation and documents widespread awareness of the $CHADS_2$ score among physicians, more than a decade after the 2006 ACC guideline (Ashburner et al., 2018; Amroze et al., 2019). Yet in numerous settings, only about half of patients with the highest stroke risks are treated with anticoagulation (Hsu et al., 2016). Our results further show that there is low adherence even among physicians who discuss the $CHADS_2$ score in their decision-making. Evidence suggests that reluctance to initiate anticoagulation mostly stems from physician decisions rather than patient preferences (Bungard et al., 2000); in one study, 93% of atrial fibrillation patients offered warfarin elected to take the treatment (Gottlieb and Salem-Schatz, 1994). In surveys, physicians report hesitance to anticoagulate patients who are elderly, frail, and multi-morbid, out of concern that anticoagulation may result in severe bleeding for these patients (Fawzy et al., 2019). The literature suggests that this hesitation may be in part driven by a mistaken assessment of bleeding risks and an overemphasis on avoiding adverse events of commission (i.e., due to initiating treatment) as opposed to those of omission (i.e., due to withholding treatment) (Gross et al., 2003).[14]

5 The Effects of Anticoagulation

The results in the previous section show a gap between $CHADS_2$ guideline awareness and full adherence to the $CHADS_2$-based recommendations. Our interpretation of this gap, as well as policy

[14] Specifically, physicians' estimates of rates of warfarin-associated intracerebral hemmorrhage were more than 10 times larger than research-based estimates (Gross et al., 2003). Further, Choudhry et al. (2006) show that physicians respond idiosyncratically to individual events, such as whether one of the physicians' other anticoagulated patients has recently experienced an adverse bleeding event.

recommendations for incorporating evidence from clinical trials into practice, will depend on the relative value of physician discretion versus guideline adherence for patient health outcomes.

In the case of atrial fibrillation, physicians may depart from $CHADS_2$-based recommendations for good reasons. The $CHADS_2$ score predicts stroke risk, while treatment decisions should be based on stroke treatment effects. Further, clinicians have access to more patient information than has been encoded in simple risk scores. Tailoring their treatment decisions to this additional information could lead to departures from $CHADS_2$-based guidelines.

Nevertheless, an emerging literature has documented widespread decision errors among physicians (Abaluck et al., 2016; Mullainathan and Obermeyer, 2019). Even with private information that is not encoded in guidelines, perfect recall of outcomes, and random variation, physicians will typically not have access to the numbers of patients needed to form their own reliable estimates of heterogeneous treatment effects from personal experience (Chandra et al., 2021).

5.1 Setup and Design

To evaluate treatment decisions and departures from guidelines, we need to characterize counterfactual patient stroke and bleeding outcomes. Let $Y_i^s(w) \in \{0,1\}$ denote whether patient i will have a stroke within one year, depending on anticoagulation $w \in \{0,1\}$, and let $Y_i^b(w) \in \{0,1\}$ denote a similar object for bleeding within one year. We then define conditional average treatment effects (CATEs) that are a function of patient characteristics, both those that are included in the $CHADS_2$ score and others that are omitted from it. Specifically, for one-year stroke and bleeding, respectively,

$$\tau^s(x) \equiv E\left[Y_i^s(1) - Y_i^s(0) \middle| X_i = x\right]; \tag{2}$$

$$\tau^b(x) \equiv E\left[Y_i^b(1) - Y_i^b(0) \middle| X_i = x\right], \tag{3}$$

where X_i is a set characteristics belonging to patient i.

We set about estimating these objects by applying "causal forest" ML techniques, as described in Wager and Athey (2017), to RCT-generated data in the AFI database (van Walraven et al., 2002). While many applications of machine learning methods use very large data sets, new work on causal forests apply related methods to estimate CATEs in sample sizes more typical of RCTs in medicine and social science; similar-scale RCT applications of causal-forest estimation have been previously demonstrated (Athey and Wager, 2019; Chernozhukov et al., 2018). We adapt the insights from these applications to our setting.

The experimental design of RCTs is well suited for estimating treatment effect heterogeneity, in contrast to many quasi-experimental designs commonly used in the economics literature. First, random assignment of treatment within each cell of patient characteristics x, a crucial requirement to estimate CATEs, is more plausible in RCTs. Second, many quasi-experiments involve monotonicity and exclusion-restriction violations within cells of the data, even if these violations "average out" in the entire sample (e.g., Kolesar et al., 2015; Frandsen et al., 2019). Third, RCTs carefully collect information on infrequent yet important outcomes (e.g., stroke and bleeding) for the disease and treatment being studied. Capturing these events, particularly their timing relative to initiation of treatment, may be challenging in even the most detailed of observational data, as we will note later.

5.2 Causal-Forest Implementation

We use the algorithm developed in Athey and Wager (2019) for estimating causal forests with conditional random assignment. The causal forest uses all the variables that overlap between the AFI and VHA data to predict treatment effect heterogeneity in the AFI RCTs (Appendix Table A.1).

The first step of the causal-forest procedure estimates risk among control-group patients as a function of patient characteristics x using regression forests. Specifically, the regression forests predict

$$Y^o(x) \quad = \quad E[Y_i^o(0)|X_i = x],\ o \in \{s, b\}. \tag{4}$$

$Y^o(x)$ is formed using observations in the control arm of each trial by estimating an "honest" regression forest.

Next, we estimate the causal forest to generate predictions of stroke and bleeding CATEs, $\tau^s(x)$ and $\tau^b(x)$ respectively, capturing heterogeneity in treatment effects along patient characteristics observable in the AFI database. The estimation procedure includes the regression forest predicted risk as one covariate along with the variables identified in Appendix Table A.1. Before fitting the causal forest, we first recenter both outcome and treatment variable (a formally justified procedure for removing trial fixed effects).[15] For more details on our implementation of causal forest, see Appendix A.3.

To provide insight into sources of treatment-effect heterogeneity, we summarize the "variable importance" of patient characteristics in predicting treatment effect heterogeneity.[16] Appendix Table A.2

[15]The "recentering" procedure is justified in Section 6.1.1 of Athey et al. (2019). Formally, we estimate these regression forests: $Y^o(x; j) = E[Y_i^o(0)|X_i = x, j(i) = j],\ o \in \{s, b\}$ and $W(x; j) = E[W_i|X_i = x, j(i) = j]$ where $j(i)$ indicates the RCT trial for individual i. This procedure resembles the construction of our risk measures, but includes trial fixed effects and uses all observations (not just control group observations).

[16]There are several methods to compute variable importance, and there is no clear consensus yet on the best method (Wei

lists variables ranked by importance in the stroke and bleed causal forests, as well as regression forests estimated to predict risk in the control groups. We also report the sign of each variable in a linear regression of treatment effects on the ten most important variables. The most important predictor of treatment effects in the stroke models is the regression-forest model of stroke risk. The variables in the $CHADS_2$ score generally rank highly in both the causal-forest and regression-forest models, but several variables not in the $CHADS_2$ score also matter: e.g., body weight, hemoglobin, and smoking behavior all predict stroke treatment effects.

5.3 Validation and Best Linear Predictors

In this section, we validate the estimated heterogeneity to assess concerns about potential over-fitting and external validity across trials. Following Athey and Wager (2019), we project outcomes onto leave-out-trial CATE predictions of treatment effects. Specifically, for observations in each trial in the AFI database, we make out-of-bag CATE predictions from causal forests grown exclusively from data in the *other* trials. Denote these leave-out-trial CATE predictions for stroke and bleeding as $\hat{\tau}^o_{-j(i)}(x)$, $o \in \{s, b\}$, for individual i in trial $j(i)$ with characteristics $X_i = x$ (we hereafter suppress the i in $j(i)$ to simplify notation). Using regression forests, we also construct and potentially control for predictions of $Y^o(x)$, based on other trials: $\hat{Y}^o_{-j}(x)$, using only control group data.

Following Chernozhukov et al. (2018), we assess external validity across trials by a "best linear predictor" regression of realized outcomes Y^o_i on treatment W_i interacted with demeaned out-of-bag CATE predictions, controlling for trial fixed effects ζ^o_j and treatment probability within each trial $P_j \equiv \Pr(W_i | j(i) = j)$ similarly interacted with demeaned out-of-bag CATE predictions:

$$\begin{aligned} Y^o_i &= \left[\delta^o_0 + \delta^o_1 \left(\hat{\tau}^o_{-j}(X_i) - \overline{\tau}^o\right)\right] W_i + \\ &\quad \gamma^o_1 \left(\hat{\tau}^o_{-j}(X_i) - \overline{\tau}^o\right) P_j + \gamma^o_2 \hat{Y}^o_{-j}(X_i) + \zeta^o_j + \varepsilon_i, \end{aligned} \tag{5}$$

where $\overline{\tau}^o \equiv \sum_i \hat{\tau}^o_{-j}(X_i)$ is the mean CATE prediction.[17] The coefficient δ^o_1 quantifies the predictive power of heterogeneous CATEs that are estimated in other trials; a coefficient value of $\delta^o_1 = 1$ would

et al., 2015). We use a measure from Athey et al. (2019) which ranks variables more highly in importance if the algorithm chooses to split trees in the forest earlier on those variables.

[17] The regression includes trial fixed effects to address a mechanical negative relationship between CATEs across trials when trials are few. To see this, consider the following decomposition: $\tau^o(x|j) = \overline{\tau}^o_x + \overline{\tau}^o_j + \tau^{o*}_{x,j}$, where $\tau^{o*}_{x,j}$ is by construction uncorrelated with $\overline{\tau}^o_x$ and $\overline{\tau}^o_j$. The heterogeneity of interest in the out-of-bag CATEs, $\hat{\tau}^o_{-j(i)}(x)$, is driven by variation in $\overline{\tau}^o_x$. If there are relatively few trials, then variation in $\overline{\tau}^o_j$ will bias downward the relationship between outcomes and CATEs, due to the small-sample negative correlation between $\hat{\tau}^o_{-j(i)}(x)$ and $\hat{\tau}^o(x|j)$.

suggest that the causal-forest estimates are well calibrated and that outcomes for a patient with characteristics x would increase one-for-one with treatment status, W_i, and the relevant CATE, $\hat{\tau}^o_{-j}(x)$.

For strokes, we also consider a modified procedure where we decompose treatment effects into two dimensions: stroke treatment effects that vary with the $CHADS_2$ score, and stroke treatment effects that are orthogonal to the $CHADS_2$ score. Specifically, we first use a regression to project leave-out-trial stroke CATEs, $\hat{\tau}^s_{-j}(x)$, onto indicators of the $CHADS_2$ score. We call the $CHADS_2$-projected component $\hat{\tau}^{s(c)}_{-j}(x)$ and the residual component $\hat{\tau}^{s(r)}_{-j}(x)$, noting that $\hat{\tau}^s_{-j}(x) = \hat{\tau}^{s(c)}_{-j}(x) + \hat{\tau}^{s(r)}_{-j}(x)$. We use these components to perform the following BLP projection:

$$\begin{aligned} Y_i^s &= \delta_0^s W_i + \sum_{\tilde{o}\in\{s(c),s(r)\}} \left[\delta_1^{\tilde{o}}\left(\hat{\tau}^{\tilde{o}}_{-j}(X_i) - \overline{\tau}^{\tilde{o}}\right) W_i + \gamma_1^{\tilde{o}}\left(\hat{\tau}^{\tilde{o}}_{-j}(X_i) - \overline{\tau}^{\tilde{o}}\right) P_j\right] + \\ &\quad \gamma_2 \hat{Y}^s_{-j}(X_i) + \zeta_j^s + \varepsilon_i. \end{aligned} \tag{6}$$

Using Equation (6), we construct BLP-based adjustments to the stroke CATEs as follows:

$$\hat{\tau}^s_{BLP}(x) = \hat{\delta}_0^s + \hat{\delta}_1^{s(c)}\hat{\tau}^{s(c)}(x) + \hat{\delta}_1^{s(r)}\hat{\tau}^{s(r)}(x). \tag{7}$$

The BLP estimation validates the performance of our causal-forest procedure on held-out trials. Results of the BLP procedure are reported in Appendix Table A.4. All of our BLP coefficients on the recovered stroke causal-forest treatment effects are significantly different from 0, and we cannot reject the null that each coefficient equals 1. For stroke CATEs, using Equation (5), we estimate $\hat{\delta}_1^s = 0.823$ (standard error of 0.230). Using Equation (6), we estimate the $CHADS_2$ component as $\hat{\delta}_1^{s(c)} = 1.329$ (standard error of 0.408); for the residual component in this equation, we estimate $\hat{\delta}_1^{s(r)} = 0.645$ (standard error of 0.265). In other words, both components predict stroke treatment effect variation in held-out trials, with coefficients not statistically distinguishable from 1.

By contrast, the BLP regression suggests that there is little reliably estimated variation in bleed CATEs across observable covariates. The BLP coefficient is close to zero ($\hat{\delta}_1^b = -0.37$) and not statistically significant. In our subsequent analyses, we will adjust CATEs predicted in Section 5.2 for stroke by using Equation (7) for $\hat{\tau}^s_{BLP}$. Because both of the stroke BLP coefficients are close to 1, the BLP adjustment makes very little difference in practice. For bleeds, the BLP regression indicates that we cannot validate heterogeneity in bleed treatment effects. In our subsequent analysis, we assume bleed treatment effects are constant at the estimated average treatment effect. This echoes the approach taken by current practice guidelines which focus on allocating treatment to patients with

large expected benefits from stroke risk reduction.

We use the BLP framework to investigate one additional aspect of the causal-forest estimates. Limited sample size may lead the causal forest to pool subgroups that in truth have different treatment effects because there is limited statistical power to detect treatment effect differences in small subsamples. Treatment effect heterogeneity that is observed by physicians but not reflected in the CATEs is of particular concern for our later counterfactual analyses. To further test whether physicians observe treatment effect heterogeneity predicted by observable covariates but not reflected in the estimated CATEs, we estimate an index of treatment propensity in the observational VHA data.[18] We then test whether this treatment propensity index predicts treatment effects in the AFI data, beyond the variation predicted by the causal-forest CATE estimates. Results are reported in Appendix Table A.4, Column 3. The coefficient on the treatment propensity index is small at -0.038 and not statistically distinguishable zero, suggesting that physicians' decisions based on observable patient characteristics in the VHA data do not reveal any additional signal of treatment effect heterogeneity. Of course, this cannot rule out the possibility that physicians have private information about treatment effect heterogeneity that is predicted by variables that are not covered in the AFI data. We discuss this possibility at length in Section 6.4.

5.4 Implied Treatment Effects in the VHA Data

In our main analysis, we take causal-forest prediction rules trained and validated in the AFI database, in the form of causal-forest splits and leaf values, and apply these rules to patients in the VHA data. For the remainder of the paper, we use BLP-adjusted CATEs, with weights defined by our validation exercise in Section 5.3, Equations (6) and (7). For brevity, we will hereafter sometimes refer to these objects as "treatment effects" or "CATEs."

Figure 4, Panel A, shows variability in the distribution of estimated stroke (BLP-adjusted) CATEs when applied to the VHA data. The 10th percentile stroke treatment effect (corresponding to the largest reductions in stroke risk) is -0.101, while the 90th percentile is -0.040. Given that the best linear predictor regression found no statistically significant relationship between predicted CATEs and treatment effects in holdout trials, we assume constant bleed treatment effects of 0.019 for the rest of our analysis.

We additionally relate our CATE estimates with the $CHADS_2$ score. Figure 4, Panel B, shows that stroke treatment effects increase roughly monotonically with the $CHADS_2$ score. The median value

[18] Thanks to David Molitor for this suggestion.

of $\hat{\tau}^s_{BLP}$ for patients with a $CHADS_2$ of 0 is -0.027, while the median value of $\hat{\tau}^s_{BLP}$ for patients with a $CHADS_2$ of 4 to 6 is -0.101. Although stroke treatment effects and the $CHADS_2$ score are highly correlated, we find substantial residual variation in stroke treatment effects after conditioning on the $CHADS_2$ score. The R^2 from regressing BLP-adjusted stroke treatment effects, $\hat{\tau}^s_{BLP}$ on $CHADS_2$ score indicators is 0.67.

5.5 External Validity

While the BLP validation exercise suggests that our causal forest predictions are externally valid across trials within the AFI database, we ultimately seek to use treatment effects estimated in the AFI database to evaluate counterfactuals in the VHA data. To assess whether this extrapolation is reasonable, we conduct three additional analyses that map patient characteristics and implied CATEs to key features in the VHA data.

First, we compare mean observable attributes of patients in each trial in the AFI database to those in the VHA data to assess whether, at least with respect to mean characteristics, the VHA data lies roughly within the support of mean characteristics across the AFI trials. Table 3 compares summary statistics in the VHA and AFI data. The clearest difference in average patient characteristics is in the share of male patients. We estimate CATEs for both male and female patients from the AFI database, allowing the causal forest to use gender to predict treatment effect heterogeneity. Our analysis of variable importance, shown in Appendix Table A.2, finds that patient gender is unimportant in predicting CATEs. Among other patient characteristics, the AFI database has a larger share of the population over 65, a lower incidence of hypertension and diabetes, and a higher rate of congestive heart failure than the VHA data on average. However, in Appendix Table A.5, we can see that for all patient characteristics (including gender), the characteristic mean in the VHA data is within the range of characteristic means across AFI trials. This mitigates concerns that the types of patients seen in the VHA are not represented in the AFI database.

Second, we compare our estimated AFI stroke CATEs with observational stroke "treatment effects" in the VHA data, or regression-adjusted differences in stroke outcomes between treated and untreated VHA patients.[19] In this process, we noted a major limitation of observational data with respect to recording the timing of events. Specifically, in contrast to the AFI RCT data generated with

[19] OLS estimates interact anticoagulation treatment with each of the patient characteristics that are covered in both the AFI and VHA data (see Appendix Table A.1 for the complete list). In addition to controlling for this variable set, the OLS specification also controls for a complete set of Elixhauser comorbidities, history of hemorrhage, family history of stroke, and a 3-knot spline in the predicted mortality index. For more details on construction of the predicted mortality index see Appendix A.3.

the express purpose of measuring predefined events, ex post measures of stroke diagnosis in observational data—both those generated by insurance claims and those from electronic health records—can indicate either a history of stroke or a new event of stroke.[20] Thus, to reliably capture the outcome of stroke, we restricted estimation of observational stroke treatment effects to patients in the VHA data who had no prior history of stroke. In Figure 5, Panel A, we show that CATEs imputed for these patients from the AFI database are quite predictive of observational treatment effect estimates in the VHA data. The correlation coefficient between these two measures of treatment effects, estimated from different data sets, is 0.68. We also note that OLS treatment effects are on average slightly *positive*, which suggests selection bias in the VHA data: physicians tend to treat patients with higher stroke risk.

Finally, in Figure 5, Panel B, we investigate stroke outcomes among untreated patients in the VHA. Due to the difficulty distinguishing current from past strokes, we again exclude patients with stroke history from this analysis. VHA patients with large predicted CATEs who nevertheless go untreated suffer much higher rates of stroke within one year, compared to patients with smaller predicted CATEs. Along similar reasoning in Mullainathan and Obermeyer (2019), this observational pattern is consistent with the possibility that patients with large estimated stroke CATEs likely suffer considerable harm from under-treatment, while patients with small stroke CATEs have low stroke incidence and thus small potential benefits from anticoagulation.

6 Assessing Physician Decisions in the VHA

With ML predictions of treatment effects in hand, we turn to assessing how physician treatment decisions in the VHA relate to treatment effects. We introduce a model to characterize how treatment decisions respond to treatment effect variation, separately considering variation that is and is not captured by guidelines. This model allows us to consider the relative effects of guidelines awareness and adherence on patient outcomes. In particular, we will evaluate whether guidelines may lead physicians to neglect information relevant for health outcomes but not incorporated into a guideline. Finally, we use the model to simulate the counterfactual impact of adopting guidelines that incorporate more information about how treatment effects vary across patients.

[20] This concern is not hypothetical—in our audit of the VHA data at the Palo Alto VA, we found that 40% of patients recorded in diagnosis codes to be experiencing a current stroke were revealed on detailed chart review to have only a history of stroke but no recurrence.

6.1 Stylized Model of Treatment Decisions

We model how treatment decisions depend on variation in treatment effects as well as the physician's guideline awareness status. Guideline awareness status is denoted $g \in \{\text{never}, \text{pre}, \text{post}\}$, corresponding to decisions made by physicians who never adopt the CHADS_2 score, decisions made by physicians before adopting the CHADS_2 score, and decisions made by physicians after adopting the CHADS_2 score. With respect to guideline awareness, we focus on perceived stroke treatment effects rather than bleed treatment effects for two reasons: First, we detect no meaningful variation in bleed treatment effects in Section 5, and second, the CHADS_2 score relates to stroke treatment effects.

Anticoagulation decisions W_i in state g are made as follows:

$$W_i = \mathbf{1}\left\{\beta\tilde{\tau}^s_{i,g} + f_g(X_i) + v_{i,g} > 0\right\}. \tag{8}$$

This model includes two components. First, physicians consider beliefs about patient-specific stroke treatment effects, $\tilde{\tau}^s_{i,g}$, with preference weight β. Guideline awareness may improve physicians' information about treatment effects, giving them more accurate posterior beliefs. The second component consists of other factors—including both observable characteristics (to the econometrician) $f_g(X_i)$ and otherwise unobservable factors $v_{i,g}$—which may impact treatment decisions. This component may capture beliefs about bleed treatment effects that we cannot detect in our AFI data or other concerns, such as frailty, that have been discussed in the literature (Fawzy et al., 2019).

For the first component—beliefs about stroke treatment effects—we consider a simple Bayesian model in Appendix A.1. The model implies that posterior beliefs are a linear function of true treatment effects and prior beliefs:[21]

$$\tilde{\tau}^s_{i,g} \quad = \quad \lambda^{s(c)}_g \tau^{s(c)}_i + \lambda^{s(r)}_g \tau^{s(r)}_i + \mu_g + v_{i,g}. \tag{9}$$

True stroke treatment effects, τ^s_i, are decomposed into a CHADS_2-related component $\tau^{s(c)}_i$ and a residual component $\tau^{s(r)}_i$, such that $\tau^s_i = \tau^{s(c)}_i + \tau^{s(r)}_i$. The parameter $\lambda^{\tilde{o}}$, $\tilde{o} \in \{s(c), s(r)\}$, correspond to the signal-to-noise ratio of posterior beliefs $\tilde{\tau}^s_{i,g}$ with respect to $\tau^{s(c)}_i$ and $\tau^{s(r)}_i$. Physicians may have more precise signals for $\tau^{s(c)}_i$ than for $\tau^{s(r)}_i$, and the precision of their signals may change with guideline awareness status g. In the model, CHADS_2 awareness increases the precision of the

[21] This linear projection can be exactly microfounded by a standard Bayesian model with normal true treatment effects and normal noise, which we detail in Appendix A.1. However, absent a joint-normal model of signals and noise, Equation (9) is a linear approximation of Bayesian updating, common in empirical Bayes applications (e.g., Chetty et al., 2014).

doctor's signal of $\tau_i^{s(c)}$, the $CHADS_2$-related variation in stroke treatment effects. μ_g is a constant within g (which depend on physician's priors), and $\upsilon_{i,g}$ is a noise term with variance that depend on the precision of the signals that physicians receive.

Our framework also allows for the possibility of "distraction effects," whereby guideline awareness leads physicians to place less weight on other decision-relevant factors. In the Bayesian language of the model, distraction effects correspond to physicians forming less precise beliefs about $\tau_i^{s(r)}$. By the same token, however, guidelines may also lead physicians to place less weight on any consideration, including those in $f_g(X_i)$ and $v_{i,g}$ that may not strictly align with reducing strokes or bleeds.

6.2 Estimating Equation

To take our behavioral model in Equation (8) to the data, we estimate the following probit model:

$$W_i = \mathbf{1}\left\{\frac{1}{\sigma_{\varepsilon,g}}\left(\alpha_g^{s(c)}\tau^{s(c)}(X_i) + \alpha_g^{s(r)}\tau^{s(r)}(X_i) + f_g(X_i) + \theta_g + \varepsilon_{i,g}\right) > 0\right\}, \tag{10}$$

where $\varepsilon_{i,g} = \beta^s \upsilon_{i,g} + v_{i,g}$ is normally distributed with variance $\sigma^2_{\varepsilon,g}$. Because we allow all components in this model to vary by g, $\sigma^2_{\varepsilon,g}$ is not identified, and we specify for estimation an error term of $\varepsilon_{i,g}/\sigma_{\varepsilon,g}$ with a normalized variance of 1 by construction.

In Equations (8) and (9), the coefficient $\alpha_g^{\tilde{o}}$, for treatment effect $\tilde{o} \in \{s(c), s(r)\}$ measures the responsiveness of decisions to treatment effect variation and can be interpreted as $\alpha_g^{\tilde{o}} = \beta^s \lambda_g^{\tilde{o}}$. Recall that β^s is the preference weight that physicians place on preventing strokes, and $\lambda_g^{\tilde{o}}$ is the signal-to-noise ratio describing how well-informed the physician is about variation in treatment effect $\tilde{o}$.

The term $f_g(X_i)$ includes controls for time trends and a three-knot spline index of patient mortality risk, all interacted with guideline adoption status. We include predicted mortality risk to capture physicians' reluctance to treat older and frailer patients (Fawzy et al., 2019). For more details on the construction of our mortality risk index in an external sample of VA patients without diagnosed atrial fibrillation see Appendix A.3. θ_g are fixed effects indicating a doctor's guideline adoption status.

Guideline awareness may change the relative responsiveness to treatment effects in three ways. First, $CHADS_2$ awareness may give physicians more precise information about $CHADS_2$-related variation in stroke treatment effects. This effect increases the responsiveness to treatment effects $\alpha^{s(c)}$ by increasing the doctor's signal-to-noise ratio for the $CHADS_2$-related variation in stroke treatment effects, i.e. increasing $\lambda^{s(c)}$. Second, $CHADS_2$ awareness may lead physicians to place less weight on residual variation in stroke treatment effects ($\lambda^{s(r)}$), thereby leading to a smaller estimated coefficient

$\alpha^{s(r)}$. Third, guideline awareness may change the variance of $\varepsilon_{i,g}$, either by reducing the noisiness of assessment incorporated into $v_{i,g}$ or by reducing the emphasis on other concerns in $v_{i,g}$.[22] A reduction in the variance of $\varepsilon_{i,g}$ will increase the relative responsiveness to stroke treatment effects.

To recover and interpret these objects, we require the following assumptions, which we will assess in our results below. First, to recover unbiased (and correctly signed) estimates of coefficients $\alpha_g^{\tilde{o}}$, we require that $\varepsilon_{i,g}$ is uncorrelated with $\tau^{\tilde{o}}(X_i)$. Second, in order to interpret deviations from treating according to treatment effects as worsening strokes or bleeds, we assume that v_i does not include unmeasured variation in treatment effects. We will examine these assumptions further in Section 6.4.

In estimating Equation (10), we use ML predictions of treatment effects, $\hat{\tau}^{s(c)}_{BLP}(x)$ and $\hat{\tau}^{s(r)}_{BLP}(x)$, in place of true treatment effects, $\tau^{s(c)}(x)$, $\tau^{s(r)}(x)$. These estimates are measured with error in the sense that they differ from the treatment effects we could obtain with infinite data. One concern is that differential measurement error of $\hat{\tau}^{s(c)}(x)$ and $\hat{\tau}^{s(r)}(x)$ may differentially attenuate $\alpha_g^{s(c)}$ and $\alpha_g^{s(r)}$. However, our BLP-adjustment in Section 5.3 of $CHADS_2$-related and residual stroke treatment effects provides a means to ensure that $\alpha_g^{s(c)}$ and $\alpha_g^{s(r)}$ can be interpreted on the same scale. This approach follows a "regression calibration" literature that addresses measurement error (George and Foster, 2000) and also resembles the first stage of "split-sample" instrumental variables approaches in economics (Angrist and Krueger, 1995).

6.3 Results

Table 4 reports estimates of average marginal effects, from Equation (10). To facilitate statistical comparisons across awareness states, the estimation takes $\alpha^{\tilde{o}}_{\text{pre}}$ as the baseline, and uses interaction terms to report differences between $\alpha^{\tilde{o}}_{\text{post}}$ or $\alpha^{\tilde{o}}_{\text{never}}$ and this baseline. Since stroke events are undesirable, we expect the marginal effects of stroke treatment effects to be negative. In other words, all else equal, physicians should be more likely to treat patients with larger reductions in stroke risk. Our baseline specification controls for year fixed effects and cubic splines in patient age. Column 1 shows these results. Column 2 allows for differential time trends in the sensitivity to treatment effects. Column 3 allows for controls, in particular the splines of predicted mortality, $f_g(X_i)$ and year fixed effects, to differ by awareness status.

The main finding from this analysis is that $CHADS_2$ awareness substantially increases physi-

[22] If guideline awareness sufficiently increases the precision of physician signals, then the variance of $v_{i,g}$ may decrease. However, because $v_{i,g}$ includes both noisy assessments and the weight that physicians' place on them, it is possible that variance of $v_{i,g}$ may increase if better information causes physicians to place more weight on their signals, including the noisy component of them. This point follows formally from the microfoundation in Appendix A.1.

cians' sensitivity to $CHADS_2$ variables in their treatment decisions. Conditional on patient's mortality risk index, physicians are more likely to treat patients with larger treatment effects predicted by the $CHADS_2$ score ($\hat{\tau}^{s(c)}_{BLP}$). This is true for both g = never and g = pre.[23] Prior to awareness, physicians are about 4 percentage points more likely to treat a patient for whom treatment effects are one percentage point larger in magnitude, and physicians who never adopt are slightly less sensitive. Following awareness of the $CHADS_2$ score, physicians become about twice as sensitive to these treatment effects.

In addition, this analysis reveals that physicians are much less sensitive to the residual component of treatment effects than to the $CHADS_2$ score prior to adoption. Following adoption, physicians if anything become slightly more sensitive to the residual component of treatment effects, rather than being distracted.

6.4 Interpreting the Model

Effects of guidelines on physician information and decisions. Under the lens of our model, these findings suggest that the $CHADS_2$ score improves physicians' responsiveness to $CHADS_2$-related stroke treatment effects. In contrast, physicians appear to have much less knowledge of residual stroke treatment effects, and adopting the $CHADS_2$ score if anything makes them more likely to attend to these residual factors. Regardless of guideline awareness, it appears that the precision of physicians' information about $CHADS_2$-related stroke treatment effects is much greater than the precision of their information about other sources of treatment effect variation, i.e. $\lambda^{s(c)}_g \gg \lambda^{s(r)}_g$. These estimates imply that doctors make relatively little use of variation in stroke treatment effects that is not captured by the $CHADS_2$ score to improve treatment allocation.

While guideline awareness may provide important informational benefits, these results also suggest the limits of increasing guideline awareness without encouraging stricter adherence. Although we see a large relative increase in physicians' responsiveness to $CHADS_2$-related treatment effects, much of the variation in treatment choice remains unexplained by the model. These deviations from guideline-based care are largely unexplained by the substantial variation in residual stroke treatment effects that we can detect and validate across trials.

[23] As noted in Figure 3, doctors do not treat patients with low $CHADS_2$ scores less than high score patients, prior to becoming aware of the guideline. However, once we condition on the patient mortality risk index to capture physician's distaste for treating frail patients, we find that even in the pre-awareness period, physicians are considering the $CHADS_2$ variables.

Considering the role of physicians' private information. We now consider the possibility that physicians may deviate from treatment decisions based on estimated CATEs because they might have private information about treatment effect heterogeneity. The model we have estimated cannot directly test whether physicians are responding to other sources of heterogeneity using characteristics that are unmeasured in the AFI data. However, three patterns in the data suggest that physicians may not be effectively using information beyond that encoded in the $CHADS_2$ score to improve treatment allocations.

First, physician are only slightly more likely to treat patients with large residual stroke treatment effects. By design, the AFI studies record many clinical characteristics (beyond those explicitly incorporated into the $CHADS_2$ score) that expert clinicians believe might drive variation in risk and treatment effects. We demonstrate that these factors indeed predict wide variation in stroke treatment effects. Nevertheless, treatment decisions by VHA physicians are largely unresponsive to this variation in stroke treatment effects, beyond those factors codified into the $CHADS_2$.

Second, as discussed in Section 5.3 and reported in Appendix Table A.4, we similarly find that physician treatment propensity does not predict any heterogeneity in treatment effects beyond the variation predicted by the estimated CATEs. In an infinitely large sample, this would follow by construction because we could nonparametrically estimate CATEs for every set of observables. However, in a finite sample, this is a substantive test of whether physician decisions have information about how treatment effects vary with observables not recovered by the causal forest. We find that they do not.

Third, we consider the possibility that physician decisions are responding to other variables that might predict treatment effect heterogeneity but are not available in the AFI data. We can use the rich set of covariates in the VHA data to assess this possibility. Recall that our baseline model already accounts for the role of many variables covered in the AFI that predict treatment effect heterogeneity and may enter physician decision-making, including salient biomarkers (blood pressure, hemoglobin), patient history (stroke, heart attack, angina), key comorbidities and demographic variables. In Appendix A.2, we describe many additional patient characteristics from the VHA electronic health record that may influence anticoagulation decisions, which we now add to our analysis. These variables extensively cover the factors suggested by clinicians and prior researchers to influence treatment decisions. Importantly, these include many variables related to frailty and fall risk, including past reports of dizziness, muscle weakness, prior injuries (fractures, head injuries), and other conditions (Parkinson's Disease, neuropathy, arthritis, vision problems). We also include variables that were later included into the HAS-BLED guideline to assess bleeding risk (liver disease, renal failure,

alcohol abuse, prior bleeds). Other variables include a full set of Elixhauser comorbidities, physician specialization, and variables that predict patients' ability to comply with warfarin monitoring.

In Panel A of Figure A.3, we investigate the robustness of our findings to these additional control variables. Enriching the control variables does not substantively change the estimated effect of $CHADS_2$ awareness on treatment decisions. Specifically, the change in physicians' responsiveness to $CHADS_2$-related variation in stroke treatment effects after guideline awareness, $\alpha^{s(c)}_{\text{pre}}/\alpha^{s(c)}_{\text{post}}$, does not vary much as we progressively add these additional covariates to the model.

In Panel B of Figure A.3, we explore whether the residual variation in treatment decisions (conditional on estimated treatment effects) can be explained by the variables described above that might predict treatment decisions but are available only in the observational VHA data and not in the RCT data from AFI. These additional variables do not explain a large fraction of treatment decisions and therefore cannot explain the high rates guideline non-adherence.[24] Even as we control for detailed patient characteristics, we find little increase in the explained share of variance in treatment decisions.

Taken together, these three pieces of evidence suggest that departures from treating according to measured treatment effects is unlikely to be explained by unmeasured variation in treatment effects. Instead, these deviations might represent practice style variation across physicians or idiosyncratic decision-making within each physician.[25] These findings are broadly consistent with earlier analysis of physician survey responses to clinical vignettes by Gross et al. (2003). The survey found that there was no relationship between physicians' perceived benefits of warfarin and their clinical decisions to recommend its use. Although perception of bleeding risk was an important determinant of prescription choice in the survey, physicians had quantitatively large mistakes in their perceptions of bleeding risk.

7 Counterfactual Awareness and Adherence

Based on our ML-predicted treatment effects in Section 5 and our analysis of physician decision-making in Section 6, we simulate outcomes under counterfactual scenarios of guideline awareness

[24] To the degree that one interprets the results in Panel A as a nonlinear analogue of the test in Altonji et al. (2008), one might argue that Panel B suggests that this test has limited power in the sense of Oster (2019). The test in Panel A is of direct interest because the covariates we include account for specific normative justifications that physicians give for non-adherence, but it is not especially informative about other unobservable characteristics in the Oster (2019) sense.

[25] Finally, note that selection on unobservable determinants of treatment effects is immaterial for our counterfactual analyses comparing strict adherence with random treatment decisions. These analyses consider treatment rules based only on observable characteristics, for which CATEs are the relevant objects. If doctors did have private information about treatment effects, our counterfactuals will understate the benefits of the status quo, but still correctly assess the impact of guideline adherence relative to random treatment.

and adherence. When discussing counterfactual outcomes, it is useful to compare outcomes to a few benchmarks. Treating *all* patients with newly diagnosed atrial fibrillation in the VHA would prevent 745 strokes (hereafter, "preventable strokes") and induce 186 bleeding events (hereafter, "inducible bleeds") per 10,000 patients after one year.

In Figure 6 and Table 5, we show key results on prevented strokes and induced bleeds under counterfactual scenarios. We first show that status quo physician decisions are approximately equivalent to *random* anticoagulation of atrial fibrillation patients: physicians prescribe anticoagulation to 49.8% of patients and prevent 49.8% of preventable strokes.

Universal guideline awareness. Next, we consider counterfactual outcomes under scenarios varying the extensive margin of guideline awareness. Awareness of the $CHADS_2$ score had relatively muted effects on outcomes. Under the counterfactual scenario of no $CHADS_2$ awareness, 51.7% of patients would be treated, preventing 51.4% of preventable strokes. Universal $CHADS_2$ awareness reduces the treatment rate to 48.0%, slightly reducing the rate of induced bleeds accordingly, and averting 48.8% of preventable strokes; this is a 1% improvement in the number of strokes prevented per bleed induced, relative to the status quo.

Strict guideline adherence. We then turn to scenarios involving strict adherence to a guideline. Treatment decisions under these scenarios strictly follow an ordering according to guideline recommendations. Each guideline implies a "score" that we use to order patients; patients with the same score (e.g., patients with the same $CHADS_2$ score for the $CHADS_2$ guideline) are randomly ordered. We evaluate the performance of adhering to each guideline-implied ordering by a *set* of counterfactual outcomes, moving from no patients treated to all patients treated. Under the assumption that the costs of treatment and monitoring are negligible relative to the clinical benefits,[26] two guideline orderings can be welfare-ranked if one guideline prevents more strokes than the other guideline, for any fixed number of induced bleeds.

Compared to expanding awareness of guidelines, policies that achieve strict adherence to guidelines produce much better outcomes. Holding treatment rates fixed at the status quo level, strict adherence to the $CHADS_2$ score prevents 59% of preventable strokes, which is 18% more than were prevented under the status quo. Adherence to a score based on full stroke treatment effects performs better still, preventing 62% of preventable strokes, or 24% more strokes than those prevented under

[26] This assumption is standard in the existing medical literature on anticoagulation, e.g. Singer et al. (2009).

the status quo. Thus, strict adherence to an optimal guideline can yield 24 times greater improvement in the number of strokes prevented per bleed induced, relative to universal $CHADS_2$ awareness.

Guideline revisions. In recent years, the $CHADS_2$ score has been replaced with the enriched CHA_2DS_2-VASc score as the basis for anticoagulation recommendations. In Appendix Figure A.4, we show that adherence to the CHA_2DS_2-VASc guideline prevents slightly fewer strokes at any given number of bleeds than strict adherence to the $CHADS_2$ score: i.e. it performs slightly worse than $CHADS_2$ score. Although our causal-forest estimates corroborate that vascular disease is an important predictor of stroke treatment effects, the CHA_2DS_2-VASc guideline gives vascular disease too much weight relative to other variables.[27] Thus, more comprehensive guidelines are not always better. More comprehensive guidelines can improve outcomes, but only if the relevant factors are weighted appropriately given treatment effects.

The role of patient frailty. We confirm previous reports that physicians are reluctant to treat older and more frail patients (Fawzy et al., 2019).[28] In Appendix Figure A.5 and Appendix Table A.6, we explore the extent to which these treatment patterns can account for the large returns of treatment reallocation. In these counterfactual analyses, we reallocate treatment only within ventiles of predicted mortality risk. We find that status quo decisions slightly outperform a random benchmark within ventiles of predicted mortality, preventing 1% more strokes per bleed induced. Notably, the benefits of guideline adherence within mortality risk bins are only slightly attenuated compared to what we find in the unconstrained counterfactuals. Adherence to an optimal guideline within mortality risk bins could still prevent 19% more strokes per bleed than in the status quo.

8 Conclusion

Our findings suggest that evidence-based clinical guidelines have the potential to improve patient health outcomes. The $CHADS_2$ score shifted physician behavior and likely prevented a small number of additional strokes while inducing an even smaller number of additional bleeds. Awareness of more comprehensive guidelines that incorporates all of the variables that predict stroke treatment effects

[27] In the interest of simplicity, most of the existing $CHADS_2$ weights were unchanged in the CHA_2DS_2-VASc score (all variables other than age), but additional variables were added. Vascular disease was given a weight of "1", the lowest weight available in the score. This weight was still too large and reduced the performance of the score.

[28] A similar pattern has been documented in the setting of heart attack care by (Currie et al., 2016): physicians avoid treating older patients, even when they would benefit from treatment.

would have larger benefits. Stricter adherence to existing or novel treatment rules produces much larger gains than awareness with discretionary adherence. Strict adherence to an optimal treatment rule that minimizes strokes can prevent 24% more strokes without increasing the number of induced bleeds.

Our results suggest important lessons for the use of guidelines in clinical care. First, the extensive margin of guideline *awareness*, in which physicians are aware of the guideline but exercise discretion in how to use it, achieves only a fraction of the benefits of greater intensive margin adherence. Second, incorporating more variables into guidelines can improve outcomes, but only if they are weighted properly. The CHA_2DS_2-VASc score provides a cautionary tale of a more complex score that does not outperform the simpler $CHADS_2$ score due to misweighting.

Many policy instruments are available to promote adherence. On the less invasive side, in-person campaigns to educate and persuade physicians to adhere to guidelines, as well as order sets and electronic reminders can make more salient the costs of departing from guidelines (Piccini et al., 2019). Alerting physicians if their adherence rates are low relative to peers could also shift behavior (Sacarny et al., 2018). More directly, pay for performance incentives could reward physicians whose treatment behavior accords with guidelines (Werner et al., 2011), and insurers could impose hassle costs on physicians to justify treatment decisions which do not comply with guidelines (e.g. failing to treat patients with higher stroke to bleed ratios than other treated patients) (Dillender, 2018). Best practices for implementing guidelines in clinical decision support (CDS) IT systems call for generating evidence for their external validity (Bates et al., 2020). An alternative way to increase adherence to new and existing guidelines may be to generate better evidence for their validity as we seek to do here, increasing the strength of the signal that guidelines provide.

Our results incorporate more information to estimate treatment effects than has been previously considered, but they only scratch the surface of what is possible. Machine-based algorithms could continue to learn both from additional trials and from observational data, in order to create more powerful predictors of treatment effects. While there remain logistical challenges to the widespread integration of machine-based algorithms into health IT systems (Kawamoto and McDonald, 2020), these are likely to be lessened as data integration and methods of validation in healthcare becomes more commonplace. Important avenues for future research include refining techniques used to build and validate clinical decision rules, as well as identifying best practices for encouraging effective guideline use.

References

Abaluck, Jason, Leila Agha, Chris Kabrhel, Ali Raja, and Arjun Venkatesh, "The Determinants of Productivity in Medical Testing: Intensity and Allocation of Care," *American Economic Review*, December 2016, *106* (12), 3730–64.

Altonji, Joseph G., Todd E. Elder, and Christopher R. Taber, "Using Selection on Observed Variables to Assess Bias from Unobservables When Evaluating Swan-Ganz Catheterization," *American Economic Review*, 2008, *98* (2), pp. 345–350.

Amroze, Azraa, Kathleen Mazor, Sybil Crawford, Kevin O'Day, David D. McManus, and Alok Kapoor, "Survey of Confidence in Use of Stroke and Bleeding Risk Calculators, Knowledge of Anticoagulants, and Comfort With Prescription of Anticoagulation in Challenging Scenarios: Support-Af II Study," *Journal of Thrombosis and Thrombolysis*, November 2019, *48* (4), 629–637.

Angrist, Joshua D. and Alan B. Krueger, "Split-Sample Instrumental Variables Estimates of the Return to Schooling," *Journal of Business & Economic Statistics*, April 1995, *13* (2), 225–235.

Arnold, David, Will Dobbie, and Peter Hull, "Measuring Racial Discrimination in Bail Decisions," Working Paper 2020-33, University of Chicago, Becker-Friedman Institute for Economics April 2020.

Arrowsmith, Cheryl H, James E Audia, Christopher Austin, Jonathan Baell, Jonathan Bennett, Julian Blagg, Chas Bountra, Paul E Brennan, Peter J Brown, Mark E Bunnage et al., "The Promise and Peril of Chemical Probes," *Nature Chemical Biology*, 2015, *11* (8), 536–541.

Ashburner, Jeffrey M., Steven J. Atlas, Shaan Khurshid, Lu-Chen Weng, Olivia L. Hulme, Yuchiao Chang, Daniel E. Singer, Patrick T. Ellinor, and Steven A. Lubitz, "Electronic Physician Notifications to Improve Guideline-Based Anticoagulation in Atrial Fibrillation: A Randomized Controlled Trial," *Journal of General Internal Medicine*, December 2018, *33* (12), 2070–2077.

Asher, Sam, Denis Nekipelov, Paul Novosad, and Stephen P Ryan, "Classification trees for heterogeneous moment-based models," Technical Report, National Bureau of Economic Research 2016.

Athey, Susan and Stefan Wager, "Estimating Treatment Effects with Causal Forests: An Application," *Observational Studies*, 2019, *5*, 36–51.

_ , **Julie Tibshirani, and Stefan Wager**, "Generalized Random Forests," *Annals of Statistics*, 2019, *47* (2), 1148–1178.

Atrial Fibrillation Investigators, "Risk Factors for Stroke and Efficacy of Antithrombotic Therapy in Atrial Fibrillation: Analysis of Pooled Data from Five Randomized Controlled Trials," *Archives of Internal Medicine*, 1995, *154* (3), 1449–1457.

Basu, Anirban, Anupam B. Jena, Dana P. Goldman, Tomas J. Philipson, and Robert Dubois, "Heterogeneity in Action: The Role of Passive Personalization in Comparative Effectiveness Research," *Health Economics*, 2014, *23* (3), 359–373.

Bates, David W., Andrew Auerbach, Peter Schulam, Adam Wright, and Suchi Saria, "Reporting and Implementing Interventions Involving Machine Learning and Artificial Intelligence," *Annals of Internal Medicine*, June 2020, *172* (11_Supplement), S137–S144.

Bungard, Tammy J, William A Ghali, Koon K Teo, Finlay A McAlister, and Ross T Tsuyuki, "Why Do Patients with Atrial Fibrillation Not Receive Warfarin?," *Archives of Internal Medicine*, 2000, *160* (1), 41–46.

Challener, Douglas W., Larry J. Prokop, and Omar Abu-Saleh, "The Proliferation of Reports on Clinical Scoring Systems: Issues about Uptake and Clinical Utility," *JAMA*, 2019, *321* (24), 2405–2406.

Chan, David C., Matthew Gentzkow, and Chuan Yu, "Selection with Variation in Diagnostic Skill: Evidence from Radiologists," Working Paper 26467, National Bureau of Economic Research November 2019.

Chandra, Amitabh and Douglas O. Staiger, "Identifying Sources of Inefficiency in Healthcare," *Quarterly Journal of Economics*, 2020, *135* (2), 785–843.

_ , **Evan Flack, and Ziad Obermeyer**, "The Health Costs of Cost-Sharing," Working Paper 28439, National Bureau of Economic Research February 2021.

Chapman, Scott A., Catherine A. St Hill, Meg M. Little, Michael T. Swanoski, Shellina R. Scheiner, Kenric B. Ware, and May N. Lutfiyya, "Adherence to Treatment Guidelines: The Association between Stroke Risk Stratified Comparing CHADS2 and CHA2DS2-VASc Score Levels and Warfarin Prescription for Adult Patients with Atrial Fibrillation," *BMC Health Services Research*, 2017, *17* (1), 127.

Chen, Jonathan H., Daniel Z. Fang, Lawrence Tim Goodnough, Kambria H. Evans, Martina Lee Porter, and Lisa Shieh, "Why Providers Transfuse Blood Products Outside Recommended Guidelines in Spite of Integrated Electronic Best Practice Alerts," *Journal of Hospital Medicine*, 2015, *10* (1), 1–7.

Chernozhukov, Victor, Mert Demirer, Esther Duflo, and Ivan Fernandez-Val, "Generic Machine Learning Inference on Heterogenous Treatment Effects in Randomized Experiments," Working Paper 24678, National Bureau of Economic Research June 2018.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff, "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates," *American Economic Review*, September 2014, *104* (9), 2593–2632.

Choudhry, Niteesh K, Geoffrey M Anderson, Andreas Laupacis, Dennis Ross-Degnan, Sharon-Lise T Normand, and Stephen B Soumerai, "Impact of Adverse Events on Prescribing Warfarin in Patients with Atrial Fibrillation: Matched Pair Analysis," *BMJ*, 2006, *332* (7534), 141–145.

Colilla, Susan, Ann Crow, William Petkun, Daniel E. Singer, Teresa Simon, and Xianchen Liu, "Estimates of Current and Future Incidence and Prevalence of Atrial Fibrillation in the US Adult Population," *American Journal of Cardiology*, 2013, *112* (8), 1142–1147.

Costantini, Otto, Klara K. Papp, Julie Como, John Aucott, Mark D. Carlson, and David C. Aron, "Attitudes of Faculty, Housestaff, and Medical Students Toward Clinical Practice Guidelines," *Academic Medicine: Journal of the Association of American Medical Colleges*, 1999, *74* (10), 1138–1143.

Currie, Janet M. and W. Bentley MacLeod, "Diagnosing Expertise: Human Capital, Decision Making, and Performance among Physicians," *Journal of Labor Economics*, 2017, *35* (1), 1–43.

_ and _, "Understanding Doctor Decision Making: The Case of Depression Treatment," *Econometrica*, 2020, *88* (3), 847–878.

Currie, Janet, W Bentley MacLeod, and Jessica Van Parys, "Provider Practice Style and Patient Health Outcomes: The Case of Heart Attacks," *Journal of Health Economics*, 2016, *47*, 64–80.

de Chaisemartin, Clement, "Tolerating Defiance? Local Average Treatment Effects Without Monotonicity," *Quantitative Economics*, 2017, *8* (2), 367–396.

Dillender, Marcus, "What Happens When the Insurer Can Say No? Assessing Prior Authorization as a Tool to Prevent High-Risk Prescriptions and to Lower Costs," *Journal of Public Economics*, 2018, *165*, 170–200.

Einav, Liran, Amy Finkelstein, Tamar Oostrom, Abigail J. Ostriker, and Heidi L. Williams, "Screening and Selection: The Case of Mammograms," Working Paper 26162, National Bureau of Economic Research August 2019.

Fawzy, Ameenathul M., Brian Olshansky, and Gregory Y. H. Lip, "Frailty and Multi-Morbidities Should Not Govern Oral Anticoagulation Therapy Prescribing for Patients With Atrial Fibrillation," *American Heart Journal*, 2019, *208*, 120–122.

Finkelstein, Amy, Petra Persson, Maria Polyakova, and Jesse Shapiro, "A Taste of Their Own Medicine: Guideline Adherence and Access to Expertise," 2021.

Frandsen, Brigham R., Lars J. Lefgren, and Emily C. Leslie, "Judging Judge Fixed Effects," Working Paper 25528, National Bureau of Economic Research February 2019.

Fuster, Valentin, Lars E Rydén, David S Cannom, Harry J Crijns, Anne B Curtis, Kenneth A Ellenbogen, Jonathan L Halperin, Jean-Yves Le Heuzey, G Neal Kay, and Task Force on Practice Guidelines, American College of Cardiology/American Heart Association, Committee for Practice Guidelines, European Society of Cardiology, European Heart Rhythm Association, Heart Rhythm Society, "ACC/AHA/ESC 2006 Guidelines for the Management of Patients with Atrial Fibrillation–Executive Summary: A Report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines and the European Society of Cardiology Committee for Practice Guidelines (Writing Committee to Revise the 2001 Guidelines for the Management of Patients with Atrial Fibrillation) Developed in Collaboration with the European Heart Rhythm Association and the Heart Rhythm Society," *European Heart Journal*, 2006, *27* (16), 1979–2030.

Gage, Brian F., Amy D. Waterman, William Shannon, Michael Boechler, Michael W. Rich, and Martha J. Radford, "Validation of Clinical Classification Schemes for Predicting Stroke: Results from the National Registry of Atrial Fibrillation," *JAMA*, 2001, *285* (22), 2864–2870.

_ , Carl van Walraven, Lesly Pearce, Robert G. Hart, Peter J. Koudstaal, B.S.P. Boode, and Palle Petersen, "Selecting Patients with Atrial Fibrillation for Anticoagulation: Stroke Risk Stratification in Patients Taking Aspirin," *Circulation*, 2004, *110* (16), 2287–2292.

George, Edward I. and Dean P. Foster, "Calibration and Empirical Bayes Variable Selection," *Biometrika*, December 2000, *87* (4), 731–747.

Gottlieb, Lawrence K and Susanne Salem-Schatz, "Anticoagulation in Atrial Fibrillation: Does Efficacy in Clinical Trials Translate into Effectiveness in Practice?," *Archives of Internal Medicine*, 1994, *154* (17), 1945–1953.

Gowrisankaran, Gautam, Keith A. Joiner, and Pierre-Thomas Léger, "Physician Practice Style and Healthcare Costs: Evidence from Emergency Departments," Working Paper 24155, National Bureau of Economic Research December 2017.

Grimshaw, Jeremy M. and Ian T. Russell, "Effect of Clinical Guidelines on Medical Practice: A Systematic Review of Rigorous Evaluations," *Lancet*, 1993, *342* (8883), 1317–1322.

Gross, Cary P, Eric W Vogel, Abhay J Dhond, Cheryl B Marple, Roger A Edwards, Ole Hauch, Elizabeth A Demers, and Michael Ezekowitz, "Factors Influencing Physicians' Reported Use of Anticoagulation Therapy in Nonvalvular Atrial Fibrillation: A Cross-sectional Survey," *Clinical Therapeutics*, 2003, *25* (6), 1750–1764.

Hirsh, Jack, Gordon Guyatt, Gregory W. Albers, Robert Harrington, and Holger J. Schünemann, "Executive Summary: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines (8th Edition)," *Chest*, June 2008, *133* (6), 71S–109S.

Hoffman, Mitchell, Lisa B. Kahn, and Danielle Li, "Discretion in Hiring," *Quarterly Journal of Economics*, 2018, *133* (2), 765–800.

Hsu, Jonathan C., Thomas M. Maddox, Kevin F. Kennedy, David F. Katz, Lucas N. Marzec, Steven A. Lubitz, Anil K. Gehi, Mintu P. Turakhia, and Gregory M. Marcus, "Oral Anticoagulant Therapy Prescription in Patients With Atrial Fibrillation Across the Spectrum of Stroke Risk: Insights From the NCDR PINNACLE Registry," *JAMA Cardiology*, 2016, *1* (1), 55–62.

Kawamoto, Kensaku and Clement J. McDonald, "Designing, Conducting, and Reporting Clinical Decision Support Studies: Recommendations and Call to Action," *Annals of Internal Medicine*, June 2020, *172* (11_Supplement), S101–S109.

Kearon, Clive, Elie A. Akl, Anthony J. Comerota, Paolo Prandoni, Henri Bounameaux, Samuel Z. Goldhaber, Michael E. Nelson, Philip S. Wells, Michael K. Gould, Francesco Dentali, Mark Crowther, and Susan R. Kahn, "Antithrombotic Therapy for VTE Disease: Antithrombotic Therapy and Prevention of Thrombosis, 9th ed: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines," *Chest*, February 2012, *141* (2, Supplement), e419S–e496S.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan, "Human Decisions and Machine Predictions," *Quarterly Journal of Economics*, 2018, *133* (1), 237–293.

Kolesar, Michal, Raj Chetty, John Friedman, Edward Glaeser, and Guido W. Imbens, "Identification and Inference With Many Invalid Instruments," *Journal of Business and Economic Statistics*, October 2015, *33* (4), 474–484.

Lane, Deirdre A. and Gregory Y.H. Lip, "Use of the CHA2DS2-VASC and HAS-BLED Scores to Aid Decision Making for Thromboprophylaxis in Nonvalvular Atrial Fibrillation," *Circulation*, 2012, *126* (7), 860–865.

Lasser, Elyse C., Elizabeth R. Pfoh, Hsien-Yen Chang, Kitty S. Chan, Justin Bailey, Hadi Kharrazi, Jonathan P. Weiner, and Sydney Morss Dy, "Has Choosing Wisely Affected Rates of Dual-Energy X-ray Absorptiometry Use?," *Osteoporosis International*, 2016, *27* (7), 2311–2316.

Lip, Gregory Y.H., Robby Nieuwlaat, Ron Pisters, Deirdre A. Lane, and Harry J.G.M. Crijns, "Refining Clinical Risk Stratification for Predicting Stroke and Thromboembolism in Atrial Fibrillation Using A Novel Risk Factor-Based Approach: The Euro Heart Survey on Atrial Fibrillation," *Chest*, 2010, *137* (2), 263–272.

Manski, Charles F., "Improving Clinical Guidelines and Decisions under Uncertainty," Working Paper 23915, National Bureau of Economic Research October 2017.

Mehta, Rajendra H., Anita Y. Chen, Karen P. Alexander, E. Magnus Ohman, Matthew T. Roe, and Eric D. Peterson, "Doing the Right Things and Doing Them the Right Way: Association between Hospital Guideline Adherence, Dosing Safety, and Outcomes among Patients with Acute Coronary Syndrome," *Circulation*, 2015, *131* (11), 980–987.

Mullainathan, Sendhil and Ziad Obermeyer, "A Machine Learning Approach to Low-Value Health Care: Wasted Tests, Missed Heart Attacks and Mis-Predictions," Working Paper 26168, National Bureau of Economic Research August 2019.

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

_ , "Health Recommendations and Selection in Health Behaviors," *American Economic Review: Insights*, June 2020, *2* (2), 143–60.

Perino, Alexander C., Jun Fan, Susan K. Schmitt, Mariam Askari, Daniel W. Kaiser, Abhishek Deshmukh, Paul A. Heidenreich, Christopher Swan, Sanjiv M. Narayan, and Paul J. Wang, "Treating Specialty and Outcomes in Newly Diagnosed Atrial Fibrillation: From the TREAT-AF Study," *Journal of the American College of Cardiology*, 2017, *70* (1), 78–86.

Piccini, Jonathan P. and Gregg C. Fonarow, "Preventing Stroke in Patients With Atrial Fibrillation-A Steep Climb Away From Achieving Peak Performance," *JAMA Cardiology*, 2016, *1* (1), 63–64.

_ **, Haolin Xu, Margueritte Cox, Roland A. Matsouaka, Gregg C. Fonarow, Butler Javed, Curtis Anne B., Desai Nihar, Fang Margaret, McCabe Pamela J., Page II Robert L., Turakhia Mintu, Russo Andrea M., Knight Bradley P., Sidhu Mandeep, Hurwitz Jodie L., Ellenbogen Kenneth A., and Lewis William R.**, "Adherence to Guideline-Directed Stroke Prevention Therapy for Atrial Fibrillation Is Achievable," *Circulation*, March 2019, *139* (12), 1497–1506.

Prior, Mathew, Michelle Guerin, and Karen Grimmer-Somers, "The Effectiveness of Clinical Guideline Implementation Strategies–A Synthesis of Systematic Review Findings," *Journal of Evaluation in Clinical Practice*, 2008, *14* (5), 888–897.

Ribers, Michael A. and Hannes Ullrich, "Battling Antibiotic Resistance: Can Machine Learning Improve Prescribing?," Discussion Paper 1803, DIW Berlin May 2019.

Rosenberg, Alan, Abiy Agiro, Marc Gottlieb, John Barron, Peter Brady, Ying Liu, Cindy Li, and Andrea DeVries, "Early Trends among Seven Recommendations from the Choosing Wisely Campaign," *JAMA Internal Medicine*, 2015, *175* (12), 1913–1920.

Sacarny, Adam, Michael L Barnett, Jackson Le, Frank Tetkoski, David Yokum, and Shantanu Agrawal, "Effect of peer comparison letters for high-volume primary care prescribers of quetiapine

in older and disabled adults: a randomized clinical trial," *JAMA Psychiatry*, 2018, *75* (10), 1003–1011.

Schuster, Mark A., Elizabeth A. McGlynn, and Robert H. Brook, "How Good Is the Quality of Health Care in the United States?," *Milbank Quarterly*, 1998, *76* (4), 517–563.

Singer, Daniel E, Yuchiao Chang, Margaret C Fang, Leila H Borowsky, Niela K Pomernacki, Natalia Udaltsova, and Alan S Go, "The Net Clinical Benefit of Warfarin Anticoagulation in Atrial Fibrillation," *Annals of Internal Medicine*, 2009, *151* (5), 297–305.

Stevenson, Megan T and Jennifer L Doleac, "Algorithmic Risk Assessment in the Hands of Humans," *Available at SSRN*, 2019.

Stroke Prevention in Atrial Fibrillation Investigators, "Risk Factors for Thromboembolism During Aspirin Therapy in Patients with Atrial Fibrillation: The Stroke Prevention in Atrial Fibrillation Study," *Journal of Stroke and Cerebrovascular Diseases*, 1995, *5* (3), 147–157.

Tibshirani, Julie, Susan Athey, and Stefan Wager, *grf: Generalized Random Forests* 2020. R package version 1.2.0.

Turakhia, Mintu P., Donald D. Hoang, Xiangyan Xu, Susan Frayne, Susan Schmitt, Felix Yang, Ciaran S. Phibbs, Claire T. Than, Paul J. Wang, and Paul A. Heidenreich, "Differences and Trends in Stroke Prevention Anticoagulation in Primary Care vs Cardiology Specialty Management of New Atrial Fibrillation: The Retrospective Evaluation and Assessment of Therapies in AF (TREAT-AF) Study," *American Heart Journal*, 2013, *165* (1), 93–101.

Valle, Christopher W., Helen J. Binns, Maheen Quadri-Sheriff, Irwin Benuck, and Angira Patel, "Physicians' Lack of Adherence to National Heart, Lung, and Blood Institute Guidelines for Pediatric Lipid Screening," *Clinical Pediatrics*, 2015, *54* (12), 1200–1205.

van Walraven, Carl, Robert G Hart, Daniel E Singer, Andreas Laupacis, Stuart Connolly, Palle Petersen, Peter J Koudstaal, Yuchiao Chang, and Beppie Hellemons, "Oral Anticoagulants vs Aspirin in Nonvalvular Atrial Fibrillation: An Individual Patient Meta-Analysis," *JAMA*, 2002, *288* (19), 2441–2448.

_ , Robert G. Hart, Stuart Connolly, Peter C. Austin, Jonathan Mant, F.D. Richard Hobbs, Peter J. Koudstaal, Palle Petersen, Francisco Perez-Gomez, J. Andre Knottnerus, Beppie

Boode, Michael D. Ezekowitz, and Daniel E. Singer, "Effect of Age on Stroke Prevention Therapy in Patients with Atrial Fibrillation: The Atrial Fibrillation Investigators," *Stroke*, 2009, *40* (4), 1410–1416.

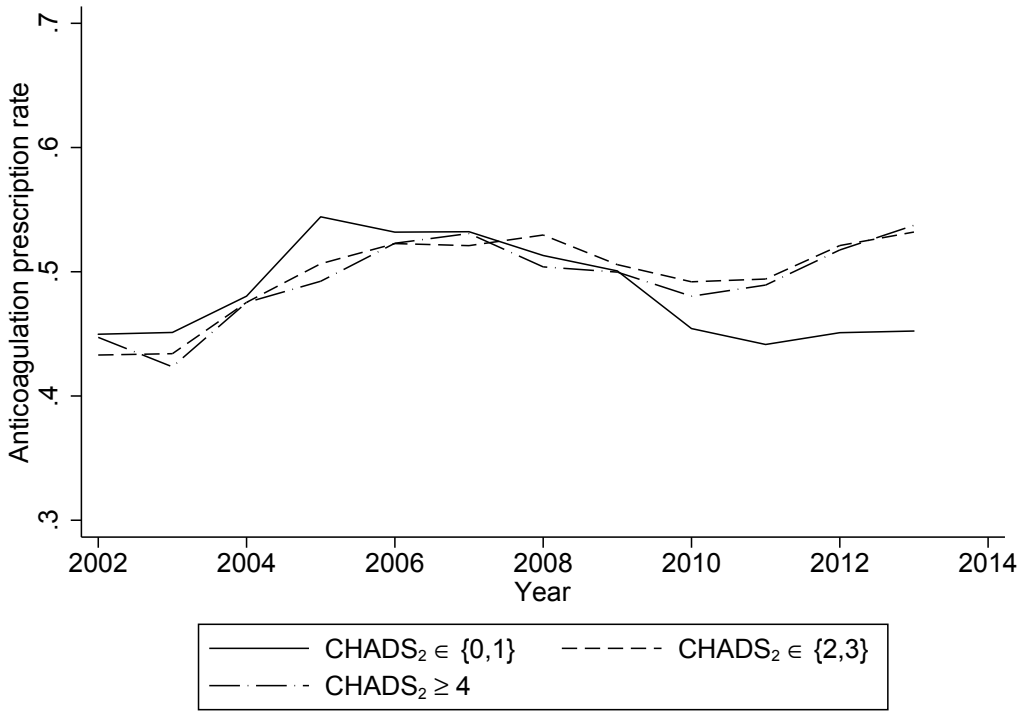
Wager, Stefan and Susan Athey, "Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests," *Journal of the American Statistical Association*, 2017, *0* (ja), 0–0.

Wei, Pengfei, Zhenzhou Lu, and Jingwen Song, "Variable Importance Analysis: A Comprehensive Review," *Reliability Engineering & System Safety*, 2015, *142*, 399–432.

Werner, Rachel M, Jonathan T Kolstad, Elizabeth A Stuart, and Daniel Polsky, "The Effect of Pay-for-Performance in Hospitals: Lessons for Quality Improvement," *Health Affairs*, 2011, *30* (4), 690–698.

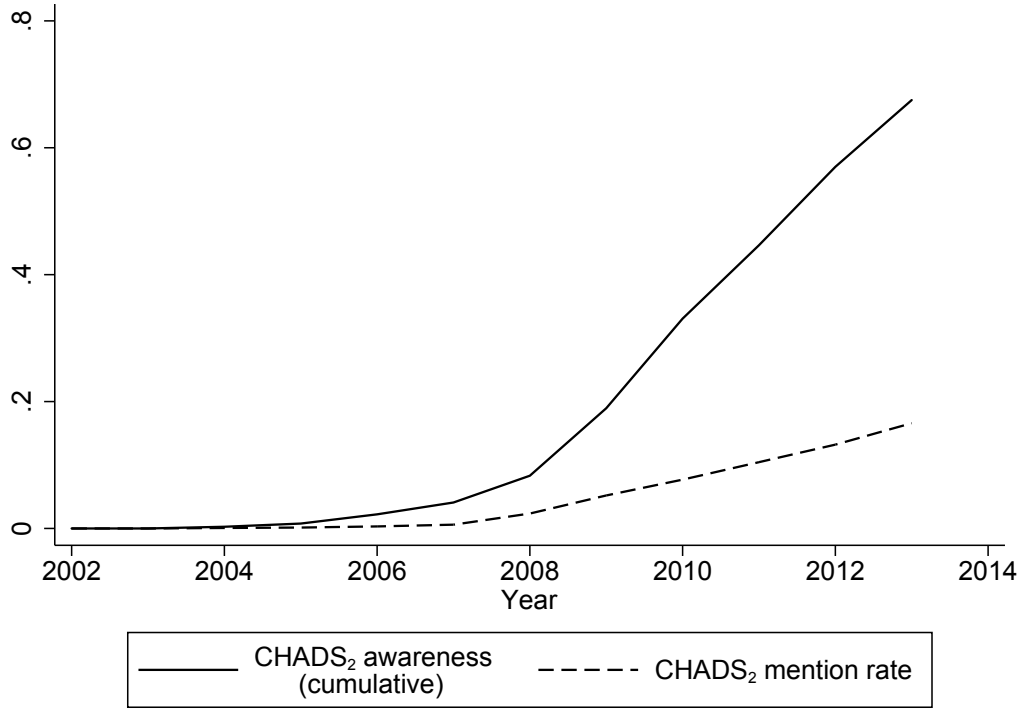
Woolf, Steven H., Richard Grol, Allen Hutchinson, Martin Eccles, and Jeremy Grimshaw, "Potential Benefits, Limitations, and Harms of Clinical Guidelines," *BMJ*, 1999, *318* (7182), 527–530.

Figure 1: Anticoagulation Trends by $CHADS_2$ Score



Notes: This figure shows the fraction of atrial fibrillation patients treated with anticoagulation over time, for three groups of patients by $CHADS_2$ score. The sample reflect patients with newly diagnosed atrial fibrillation in the VHA, and anticoagulation treatments are defined as prescriptions within 90 days of initial diagnosis. Table 2 provides further details about the sample selection.

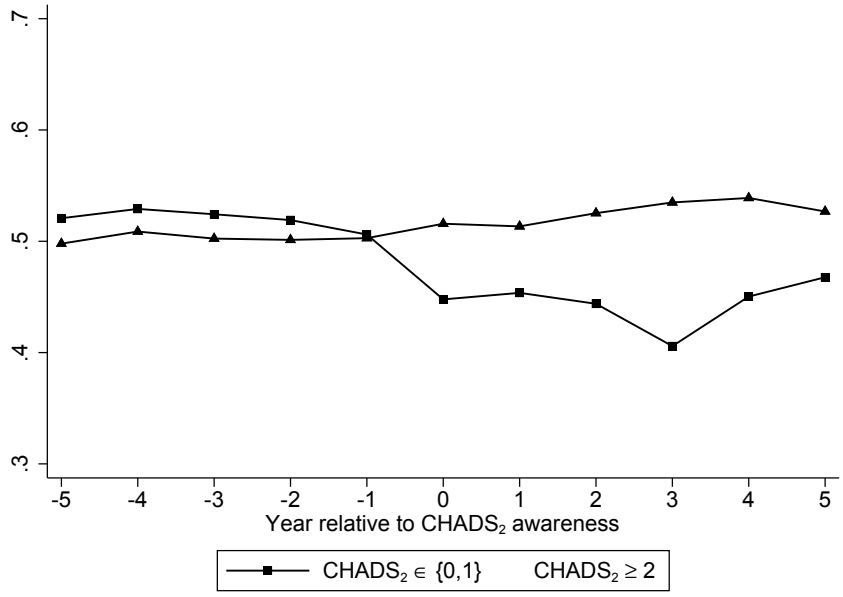
Figure 2: Diffusion of the $CHADS_2$ Score



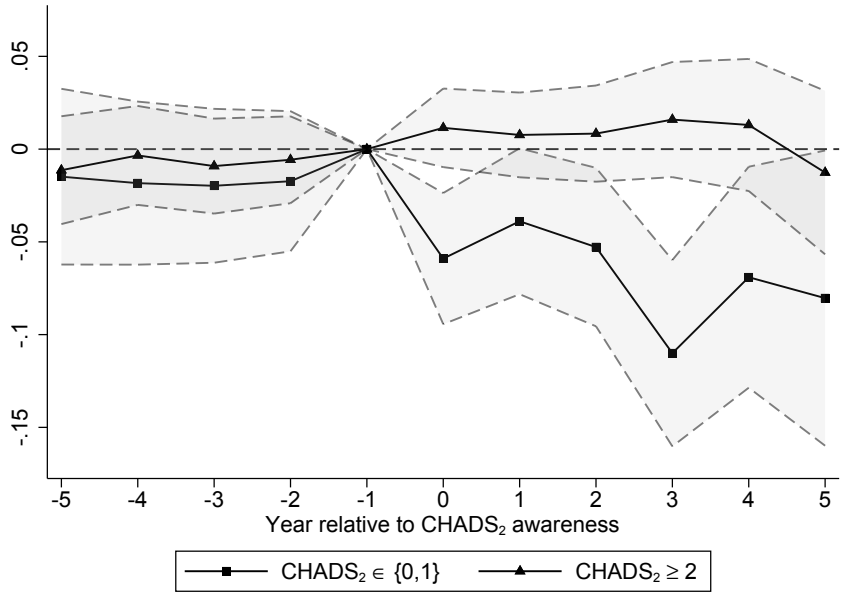
Notes: This figure shows the fraction of patients in a given year with physicians who either mention the $CHADS_2$ score in the note for the index patient or have mentioned the $CHADS_2$ score in either the note for the index patient or in a previous note. We identify mentions by searching the note text for the phrase `chads` (not case-sensitive). We consider any physician who has mentioned the the $CHADS_2$ score in the current note or in a previous note as being *aware* of the $CHADS_2$ guideline, shown in the solid line. The dashed line reflects the rate of mentions in the index patient's note. The sample reflect patients with newly diagnosed atrial fibrillation in the VHA. Table 2 provides further details about the sample selection.

Figure 3: Treatment Decisions and $CHADS_2$ Awareness

A. Trends Relative to $CHADS_2$ Awareness



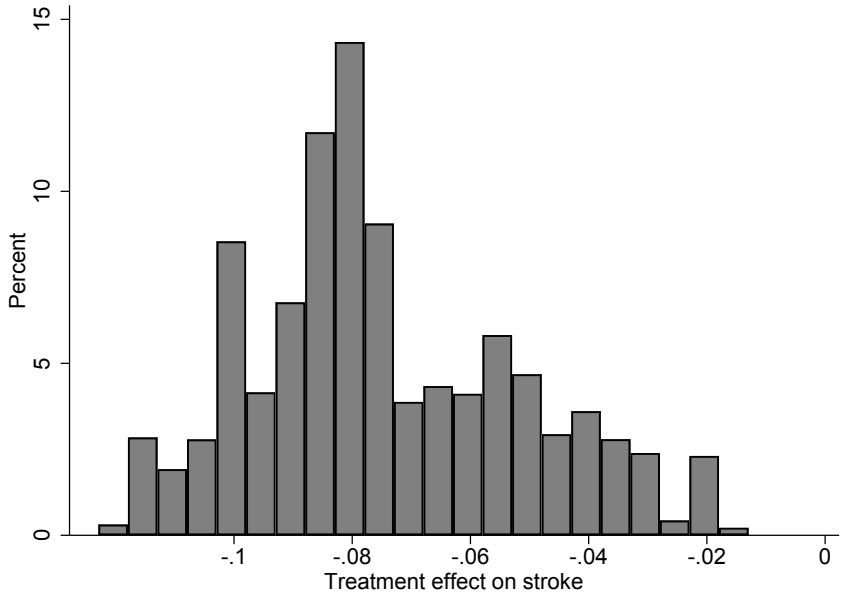
B. Event Study Estimates



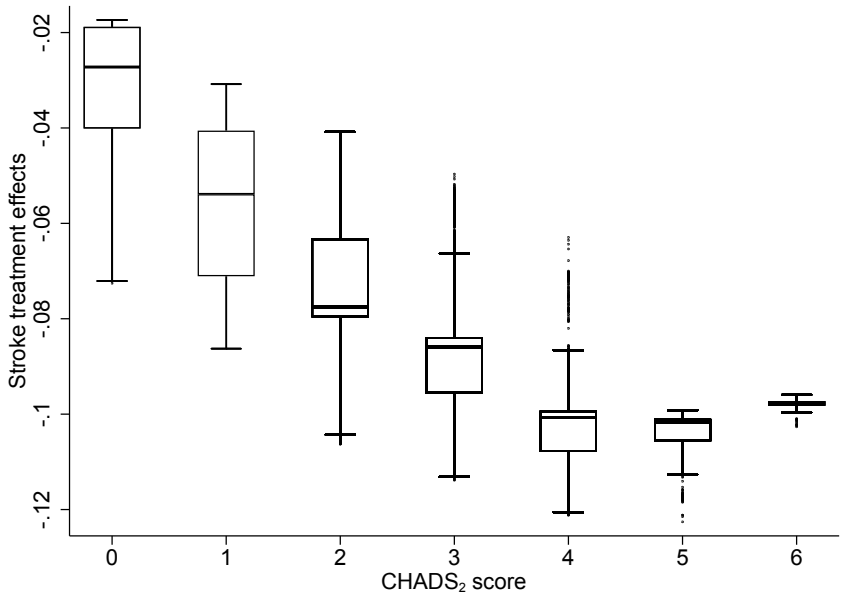
Notes: Panel A displays the fraction of atrial fibrillation patients treated with anticoagulation in each year relative to $CHADS_2$ awareness for physicians who eventually adopt the $CHADS_2$ score. Panel B shows regression coefficients and 95% confidence intervals from Equation (1), run separately for patients with $CHADS_2 \in \{0,1\}$ and for patients with $CHADS_2 \geq 2$. The 12-month period prior to the physician's first $CHADS_2$ mention is normalized to 0. The regression sample includes 104,585 VHA patients who either are treated within 5 years of their physician's observed $CHADS_2$ awareness or are treated by a never-aware physician.

Figure 4: Distribution of Stroke Treatment Effects Across VHA Patients

A. Histogram of Stroke Treatment Effects



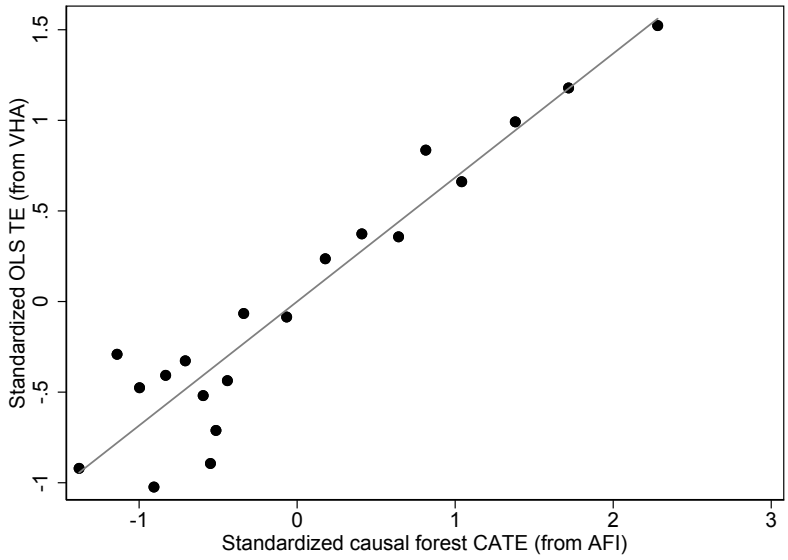
B. Stroke Treatment Effects by $CHADS_2$ Score



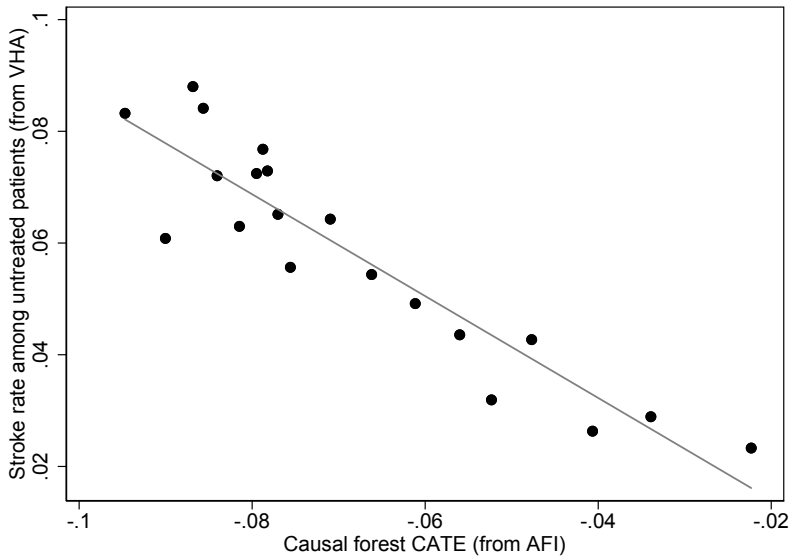
Notes: Panel A displays a histogram of stroke treatment effects in the VHA sample. Panel B shows a box plot for the distribution of treatment effects by $CHADS_2$ score in the VHA data. Bounds on the box plot are at the 25th and 75th percentile, with the median marked with a horizontal line. Whiskers extend to the 5th and 95th percentiles. For both panels, conditional average treatment effect (CATE) predictions are trained and validated by using causal-forest methods, described in Section 5, applied to RCT data in the AFI database. We use the causal-forest rules to calculate CATEs as a function of patient characteristics for each patient in the VHA data.

Figure 5: Evidence on the External Validity of AFI CATEs

A. AFI CATEs and Observational Treatment Effects



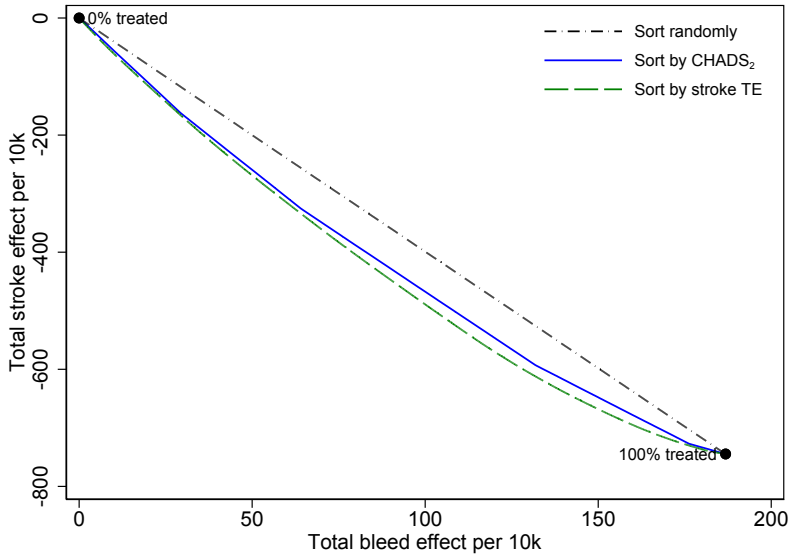
B. Stroke Outcomes Among Untreated Patients and AFI CATEs



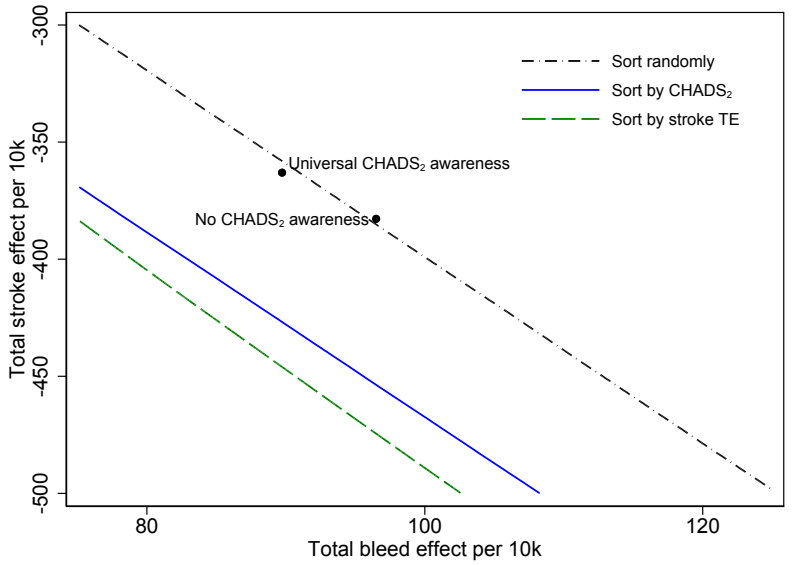
Notes: Panel A displays a binned scatterplot of the relationship between z-standardized stroke treatment effects estimated by causal forest in the AFI RCT data and z-standardized stroke treatment effects estimated by OLS in the VHA observational data. The correlation coefficient is 0.68. OLS estimates interact anticoagulation treatment with each of the patient characteristics that are covered in both the AFI and VHA data (see Appendix Table A.1 for the complete list). In addition to controlling for this variable set, the OLS specification also controls for a complete set of Elixhauser comorbidities, history of hemorrhage, family history of stroke, and a 3-knot spline in the predicted mortality index. In Panel B, we show a binned scatterplot of the relationship between 1-year stroke incidence for untreated patients in the VHA data and CATE estimates from the AFI RCT data. Due to data limitations that prevent us from differentiating new stroke events from repeated coding of a prior stroke, both panels are estimated in a restricted sample of 91,797 patients with no stroke history at the time of atrial fibrillation diagnosis.

Figure 6: Counterfactual Outcomes

A. Strict Guideline Adherence



B. Guideline Awareness (Inset)



Notes: This figure shows strokes prevented and bleeds induced by anticoagulation in counterfactual scenarios. Strokes prevented per 10,000 patients are shown on the *y*-axis, and bleeds induced per 10,000 patients are shown on the *x*-axis. Both panels display outcomes under a random treatment allocation, ranging from 0% of patients treated (top-left corner of Panel A) to 100% of patients treated (bottom-right corner of Panel A). Panel A shows outcomes under counterfactual strict adherence to various guideline rules. Each rule implies a patient sorting, and the curves indicate counterfactual outcomes ranging from treating 0% to 100% of patients. Patients with the same score or treatment effect are randomly sorted into treatment. Panel B shows an inset area and plots outcomes for counterfactual awareness (i.e., imperfect adherence) scenarios.

Table 1: $CHADS_2$ Score and Treatment Recommendations

$CHADS_2$ Components	Points
History of congestive heart failure	1
History of hypertension	1
History of diabetes mellitus	1
Aged 75 or older	1
Previous stroke or transient ischemic attack	2

Treatment Recommendation
Score of 2 or greater: high risk of stroke; oral anticoagulant recommended
Score of 1: moderate risk of stroke; oral anticoagulant considered
Score of 0: low risk of stroke; oral anticoagulant not recommended

Notes: This table describes the $CHADS_2$ score used to assess stroke risk among patients with atrial fibrillation. The score is based on evidence developed by Gage et al. (2001, 2004). In the bottom panel, the table also summarizes the 2006 ACC and 2008 ACCP guideline treatment recommendations based on the $CHADS_2$ score, published in Fuster et al. (2006) and Hirsh et al. (2008).

Table 2: VHA Sample Selection

		Observations	
Sample step	Description	Dropped	Remaining
1. Identify potentially new atrial fibrillation patients	Identify candidate patients with a diagnosis of atrial fibrillation not previously diagnosed in the last three years.		844,312
2. Prescription restriction	Keep patients who had a prescription filled at the VA within the last year. Drop patients who had a prior anticoagulation prescription.	290,214	554,098
3. Confirmed atrial fibrillation diagnosis	Keep patients who had an EKG within 30 days before or after initial diagnosis and a second atrial fibrillation diagnosis recorded 30-365 days after index visit.	254,164	299,934
4. PCP or cardiologist visit	Keep patients who had a PCP or cardiologist visit up to 90 days after the index visit. The earliest such visit identifies the attributed physician. Require that the attributed physician wrote at least one non-warfarin prescription for the patient within one year (before or after).	135,395	164,539
5. Physicians with sufficient sample	Keep patients attributed to a physician with at least 30 atrial fibrillation patients and has written at least one warfarin prescription in the unrestricted sample defined in step #1.	32,868	131,671
6. Drop observations with missing variables	Keep patients with non-missing demographics, comorbidities, and clinical information.	18,401	113,270

Note: This table describes key VHA sample selection steps, the observations dropped, and the observations remaining after each step.

Table 3: Summary Statistics

	VHA Data	AFI Database		
	Mean	Overall	Smallest Trial Mean	Largest Trial Mean
Characteristic	(1)	(2)	(3)	(4)
Treated with anticoagulation	0.50	0.44	0.34	0.50
Male	0.99	0.67	0.46	1.00
Age	74.05	70.37	67.75	73.67
Stroke treatment effect	−0.07	−0.04	−0.02	−0.07
Bleed treatment effect	0.02	0.02	0.02	0.02
$CHADS_2$ components:				
Congestive heart failure	0.15	0.30	0.00	0.70
Hypertension	0.84	0.45	0.32	0.59
Age $\geq$ 65	0.52	0.76	0.63	0.90
Diabetes	0.36	0.14	0.08	0.19
Previous stroke	0.15	0.11	0.00	0.76
Number of physicians	5,752			
Number of patients	113,270	4,720		

Notes: This table reports mean and standard deviations of characteristics of patients in the VHA data and in the AFI database. Column 1 shows characteristics of patients in the VHA data, specifically in the sample created by the steps described in Table 2. Column 2 shows characteristics of patients in the overall AFI database. Columns 3 and 4 show the smallest and largest trial means, respectively, for the patient characteristics.

Table 4: Average Marginal Effects of Probit Model

	Dependent Variable: Anticoagulant Prescription		
	(1)	(2)	(3)
CHADS$_2$-related stroke treatment effect, $\hat{\tau}^{s(c)}_{BLP}(x)$			
Pre-awareness baseline, $\alpha^{s(c)}_{\text{pre}}$	−3.952***	−4.035***	−4.685***
	(0.361)	(0.367)	(0.416)
Post-awareness difference, $\alpha^{s(c)}_{\text{post}} - \alpha^{s(c)}_{\text{pre}}$	−4.488***	−4.410***	−3.173***
	(0.539)	(0.543)	(0.621)
Never-aware difference, $\alpha^{s(c)}_{\text{never}} - \alpha^{s(c)}_{\text{pre}}$	0.807*	0.841*	1.363**
	(0.477)	(0.479)	(0.546)
Residual stroke treatment effect, $\hat{\tau}^{s(r)}_{BLP}(x)$			
Pre-awareness baseline, $\alpha^{s(r)}_{\text{pre}}$	−0.815	−1.001*	−1.391**
	(0.498)	(0.535)	(0.547)
Post-awareness difference, $\alpha^{s(r)}_{\text{post}} - \alpha^{s(r)}_{\text{pre}}$	−1.687**	−1.509*	−0.428
	(0.767)	(0.794)	(0.823)
Never-aware difference, $\alpha^{s(r)}_{\text{never}} - \alpha^{s(r)}_{\text{pre}}$	0.580	0.833	1.153
	(0.662)	(0.690)	(0.711)
Year fixed effects, predicted mortality spline controls	Yes	Yes	Yes
Differential trends on treatment effects	No	Yes	Yes
Controls interacted with CHADS_2 awareness status	No	No	Yes
Number of observations	113,270	113,270	113,270

Notes: This table reports average marginal effects from probit regressions of anticoagulation treatment decisions, as specified in Equation (10). Key regressors of interest are causal-forest predictions of CATEs: CHADS_2-related stroke treatment effects, or $\hat{\tau}^{s(c)}_{BLP}(x)$; residual stroke treatment effects, or $\hat{\tau}^{s(r)}_{BLP}(x)$; and bleed treatment effects, or $\hat{\tau}^{b}_{BLP}(x)$. All specifications include calendar year fixed effects and a 3-knot spline in predicted mortality. Column 2 includes linear trends interacted with each of these treatment effects. Column 3 includes interactions of CHADS_2 awareness status with both year fixed effects and the predicted mortality spline variables. Standard errors are clustered at the physician level. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 5: Counterfactual Treatment Decisions and Outcomes

		Strokes prevented		Bleeds induced	
	Percent of patients treated	Per 10K patients	Percent of maximum	Per 10K patients	Percent of maximum
A: Benchmarks					
Status quo	49.8%	371	49.8%	93	49.8%
Randomly assigned treatment	49.8%	371	49.8%	93	49.8%
All patients treated	100%	745	100%	187	100%
B: Guideline Awareness					
No $CHADS_2$ awareness	51.7%	383	51.4%	96	51.7%
Universal $CHADS_2$ awareness	48.0%	363	48.8%	90	48.0%
C: Strict Guideline Adherence					
$CHADS_2$ guideline	49.8%	439	59.0%	93	49.8%
Stroke TE guideline	49.8%	460	61.7%	93	49.8%

Notes: This table reports treatment rates and patient outcomes in counterfactual awareness and adherence scenarios for patients in the VHA data. Outcomes of strokes prevented and bleeds induced are reported per 10,000 patients and as a percent of the maximum number of preventable strokes or inducible bleeds. Panel A reports treatment rates and outcomes under the status quo, which we observe in our data, and in a counterfactual assignment of the same number of treatments to random patients. Panel B reports treatment rates and outcomes under counterfactual awareness scenarios, assuming adherence implied by our structural model in Equation (10). Panel C reports treatment rates and outcomes under patient orderings according to scores implied by counterfactual strict adherence to different guidelines. Patients with the same score are randomly ranked. "$CHADS_2$ guideline" orders patients by their $CHADS_2$ score. "Stroke TE guideline" orders patients by $\hat{\tau}^s_{BLP}(x)$.

Online Appendix
"Fixing Misallocation with Guidelines: Awareness vs. Adherence"

Jason Abaluck
Leila Agha
David Chan
Daniel Singer
Diana Zhu

July 9, 2021

A.1 Bayesian Model of Decision-Making

In this appendix, we describe in greater detail the Bayesian model of decision-making specified in Equation (8), which we restate here:

$$W_i = \mathbf{1}\left\{\beta\tilde{\tau}^s_{i,g} + f_g(X_i) + v_{i,g} > 0\right\},$$

focusing on the Bayesian posterior beliefs about stroke treatment effects: $\tilde{\tau}^s_{i,g}$. Recall that g denotes the awareness status of the physician. Awareness status may change the informativeness of physician beliefs about treatment effects.

A.1.1 Component Treatment Effects, Signals, and Beliefs

Treatment effects and physician beliefs depend on patient characteristics, which we may orthogonalize into components $k \in \mathcal{K}$. We can conceptualize each principal component as implying additional (orthogonal) information about treatment effects. Specifically, assume that stroke treatment effects are normally distributed and comprise component treatment effects that are also normally distributed:

$$\tau^s_{i,k} \sim N\left(\overline{\tau}^s_k, \sigma^2_{\tau(s),k}\right), \tag{A.1}$$

for each $k \in \mathcal{K}$. We assume that physicians know the moments of each component treatment effect $\left(\overline{\tau}^s_k, \sigma^2_{\tau(s),k}\right)_{k=1}^{K}$.[1]

For each component k, physicians receive a noisy signal of the underlying treatment effect, $\dot{\tau}^s_{i,k}$:

$$\dot{\tau}^s_{i,g,k} = \tau^s_{i,k} + \epsilon^s_{i,g,k}, \tag{A.2}$$

where $\epsilon^s_{i,g,k}$ is a normally distributed noise term with variance $\sigma^2_{\epsilon(s),g,k}$, or $\epsilon^s_{i,g,k} \sim N\left(0, \sigma^2_{\epsilon(s),g,k}\right)$. Note the dependence of signals on g. This models the possibility that awareness status may change the quality of information that physicians receive about treatment effects.

Given prior beliefs and the noisy signals, physicians form posterior beliefs, $\tilde{\tau}^s_{i,k}$. Specifically,

$$\tilde{\tau}^s_{i,g,k} = \lambda^s_{g,k}\dot{\tau}^s_{i,g,k} + (1-\lambda^s_{g,k})\overline{\tau}^s_k, \tag{A.3}$$

where $\lambda^s_{g,k} = \frac{\sigma^2_{\tau(s),k}}{\sigma^2_{\tau(s),k}+\sigma^2_{\epsilon(s),g,k}}$ is the signal-to-noise ratio of the kth component.

A.1.2 Regression Interpretation

The relationship between posterior beliefs and signals in Equation (A.3) can be interpreted as a regression of posterior beliefs on signals. This relationship may also be interpreted as a regression

[1] Our model in Equation (8) allows for potentially non-Bayesian beliefs that can shift decision-making via $f_g(X_i)$ and $v_{i,g}$. In order to study the effect of information in a Bayesian framework, we compartmentalize the two components of the model and consider the first component, described in this appendix, as Bayesian.

of posterior beliefs on true treatment effects, since the noise component of signals is orthogonal to treatment effects:

$$\begin{aligned}\tilde{\tau}^s_{i,g,k} &= \lambda^s_{g,k}\dot{\tau}^s_{i,g,k} + (1-\lambda^s_{g,k})\overline{\tau}^s_k \\ &= \lambda^s_{g,k}\tau^s_{i,k} + (1-\lambda^s_{g,k})\overline{\tau}^s_k + \lambda^s_{g,k}\epsilon^s_{i,g,k},\end{aligned}$$

where the second line uses the definition of the signal in Equation (A.2). In other words, a unit increase in the treatment effect $\tau^s_{i,k}$ should increase posterior beliefs by $\lambda^s_{g,k}$.

We may use this framework to consider the relationship between overall treatment effects, overall signals, and overall posterior beliefs, aggregated across components $k \in \mathcal{K}$. These overall objects are, respectively, $\tau^s_i \equiv \sum_{k\in\mathcal{K}}\tau^s_{i,k}$; $\dot{\tau}^s_{i,g} = \sum_{k\in\mathcal{K}}\dot{\tau}^s_{i,g,k}$; and $\tilde{\tau}^s_{i,g} = \sum_{k\in\mathcal{K}}\tilde{\tau}^s_{i,g,k}$. Substituting the definition of the component signals from Equation (A.3), we may also state the overall posterior belief as

$$\tilde{\tau}^s_{i,g} = \sum_{k\in\mathcal{K}}\left(\lambda^s_{g,k}\dot{\tau}^s_{i,g,k} + (1-\lambda^s_{g,k})\overline{\tau}^s_k\right). \tag{A.4}$$

We now consider the overall signal-to-noise ratio in a regression predicting the overall posterior belief using the signal:

$$\tilde{\tau}^s_{i,g} = \lambda^s_g\dot{\tau}^s_{i,g} + (1-\lambda^s_g)\overline{\tau}^s. \tag{A.5}$$

Using Equation (A.4) for $\tilde{\tau}^s_{i,g}$ and the definition of the overall signal for $\dot{\tau}^s_{i,g}$, the coefficient λ^s_g in this regression is

$$\lambda^s_g = \frac{\text{Cov}\left(\tilde{\tau}^s_{i,g}, \dot{\tau}^s_{i,g}\right)}{\text{Var}\left(\dot{\tau}^s_{i,g}\right)} = \frac{\sum_{k\in\mathcal{K}}\lambda^s_{g,k}\text{Var}\left(\dot{\tau}^s_{i,g,k}\right)}{\sum_{k\in\mathcal{K}}\text{Var}\left(\dot{\tau}^s_{i,g,k}\right)} \tag{A.6}$$

$$= \frac{\sum_{k\in\mathcal{K}}\sigma^2_{\tau(s),g,k}}{\sum_{k\in\mathcal{K}}\left(\sigma^2_{\tau(s),g,k} + \sigma^2_{\epsilon(s),g,k}\right)}. \tag{A.7}$$

Equation (A.6) reveals that the overall signal-to-noise ratio, λ^s_g, can be thought of as a variance-weighted average of the component signal-to-noise ratios, $\lambda^s_{g,k}$. Equation (A.7) shows that a posterior belief formed directly from the aggregate signal, as in Equation (A.5), will have the same signal-to-noise ratio as a posterior belief aggregated from component posterior beliefs, as in Equation (A.4).

A.1.3 CHADS$_2$ and Residual Treatment Effects

We are now in a position to state posterior beliefs as in Equations (9). For strokes, we can separate the set of components $\mathcal{K}_c$ that predict CHADS$_2$-related treatment effects and $\mathcal{K}\setminus\mathcal{K}_c$ components that predict residual treatment effects. We expect that the component posterior beliefs related to the CHADS$_2$ score should increase in informativeness. That is, we expect that $\lambda^s_{g,k}$ should increase with g = post, for $k \in \mathcal{K}_c$. We first define the two components of stroke treatment effects: $\tau^{s(c)}_i \equiv$

$\sum_{k\in\mathcal{K}_c}\tau^s_{i,k}$, and $\tau^{s(r)}_i \equiv \sum_{k\notin\mathcal{K}_c}\tau^s_{i,k}$. Restating Equation (9) as

$$\tilde{\tau}^s_{i,g} = \lambda^{s(c)}_g \tau^{s(c)}_i + \lambda^{s(r)}_g \tau^{s(r)}_i + \mu^s_g + \upsilon^s_{i,g},$$

we can then interpret the signal-to-noise coefficients in the equation as follows:

$$\begin{aligned}
\lambda^{s(c)}_g &= \frac{\sum_{k\in\mathcal{K}_c}\sigma^2_{\tau(s),g,k}}{\sum_{k\in\mathcal{K}_c}\left(\sigma^2_{\tau(s),g,k}+\sigma^2_{\epsilon(s),g,k}\right)};\\
\lambda^{s(r)}_g &= \frac{\sum_{k\notin\mathcal{K}_c}\sigma^2_{\tau(s),g,k}}{\sum_{k\notin\mathcal{K}_c}\left(\sigma^2_{\tau(s),g,k}+\sigma^2_{\epsilon(s),g,k}\right)}.
\end{aligned}$$

If we conceptualize the posterior belief as directly formed from $\dot{\tau}^{s(c)}_i \equiv \tau^{s(c)}_i + \sum_{k\in\mathcal{K}_c}\epsilon^s_{i,g,k}$ and $\dot{\tau}^{s(r)}_i \equiv \tau^{s(r)}_i + \sum_{k\notin\mathcal{K}_c}\epsilon^s_{i,g,k}$, then we can interpret the constant, μ^s_g, and error term, $\upsilon^s_{i,g}$ as

$$\begin{aligned}
\mu^s_g &= \sum_{k\in\mathcal{K}}\left(1-\mathbf{1}\left(k\in\mathcal{K}_c\right)\lambda^{s(c)}_g - \mathbf{1}\left(k\notin\mathcal{K}_c\right)\lambda^{s(r)}_g\right)\overline{\tau}^s_k;\\
\upsilon^s_{i,g} &= \sum_{k\in\mathcal{K}}\left(\mathbf{1}\left(k\in\mathcal{K}_c\right)\lambda^{s(c)}_g + \mathbf{1}\left(k\notin\mathcal{K}_c\right)\lambda^{s(r)}_g\right)\epsilon^s_{i,g,k}.
\end{aligned}$$

Unlike $\lambda^{s(c)}_g$ and $\lambda^{s(r)}_g$, μ^s_g and $\text{Var}\left(\upsilon^s_{i,g}\right)$ are not exactly invariant to the level of aggregation with which posterior beliefs are formed.[2] Nevertheless, regardless of this level of aggregation, qualitative interpretations are unchanged: μ^s_g is a function of the signal-to-noise ratio and prior beliefs, and $\upsilon^s_{i,g}$ is a function of signal-to-noise ratio and noise. If $\lambda^s_{g,k} = 1$ for all $k\in\mathcal{K}$, there is no noise, and $\upsilon^s_{i,g} = 0$. At the other extreme, if $\lambda^s_{g,k} = 0$ for all $k\in\mathcal{K}$, there is no meaningful signal. In this case, physicians will ignore all $\dot{\tau}^s_{g,k}$, and we will also have $\upsilon^s_{i,g} = 0$.

A.2 Predicting Physician Treatment Decisions

Observable variation in treatment effects, patient age, and time trends explain a relatively small fraction of the total variation in treatment decisions. In this section, we explore other factors that might drive physician treatment decisions. Specifically, we consider the following additional variables, which may influence physicians' treatment decisions. None of these variables are available in the AFI database, and so estimated treatment effects are not a direct function of these variables.

1. **Variables related to frailty and fall risk.** We include indicators for neurologic disorder (including Parkinson's Disease), fall risk (neuropathy, muscle weakness, dizziness), vision problems, arthritis, head injury, fracture. Frailty and fall risk are frequently cited clinical explana-

[2] For μ^s_g to be invariant, we require λ^s_g to be a different weighted average of $\lambda^s_{g,k}$, with weights proportional to $\overline{\tau}^s_k$ rather than $\text{Var}\left(\dot{\tau}^s_{i,g,k}\right)$. For $\text{Var}\left(\upsilon^s_{i,g}\right)$ to be invariant, we require $(\lambda^s_g)^2$ to be a weighted average of $(\lambda^s_{g,k})^2$, with weights proportional to $\text{Var}\left(\epsilon^s_{i,g,k}\right)$ rather than $\text{Var}\left(\dot{\tau}^s_{i,g,k}\right)$.

tions for not prescribing warfarin to patients with high $CHADS_2$ scores. Patients with high fall risk may be more likely to suffer intracranial bleeds if they are taking warfarin.

2. **Elixhauser comorbidities that are not in the AFI database.** We include indicators for HIV/AIDS, deficiency anemia, hypothyroidism, tumor, metastasis, lymphoma, obesity, weight loss, paralysis, pulmonary circulation disorders, ulcer, valvular disease. These are additional patient characteristics that have been shown to predict health care spending and mortality.

3. **Variables included in the HAS-BLED score to predict bleeding risk if anticoagulated.** We include indicators for liver disease, renal failure, alcohol abuse, history of bleeds. These variables are included in the HAS-BLED score, which is a predictive risk score that aims to inform physicians of the risk of induced bleed, if the patient is anticoagulated. The HAS-BLED score incorporates three variables that we have already included into our predictions of bleed treatment effect heterogeneity, including age, hypertension, and stroke history; we do not consider these variables separately here, since included bleed treatment effects may already depend on these variables. The HAS-BLED also includes a measure of a measure of unstable or high INRs among treated patients, which is not observed prior to treatment, and so not included here. Finally, HAS-BLED score also includes medication usage that predisposes patients to bleeding, such as aspirin or NSAIDS. Unfortunately, we do not consistently observe the use of these medications because they are widely available over the counter, without a prescription.

4. **Variables related to patient's ability to comply with warfarin monitoring.** We include indicators for drug abuse, depression, psychoses, number of years of military service. Appropriate management of patients on warfarin requires blood work repeated at regular intervals (typically every 2-4 weeks) to ensure the dosing is appropriate. Optimal dosing can depend on a patient's diet and other medications, and may need to be adjusted from time to time as those factors change. If the warfarin dosage is too low, the patient will not reap the benefits of anticoagulation for stroke reduction; if the dosage is too high, the patient will be at elevated risk of bleeds. These variables included here are related to the likelihood that the patient can comply with the monitoring regimen.

5. **Physician characteristics.** We include indicators for the physician's specialty code, specifically for cardiology, internal medicine, and primary care. This specialty coding variable indicates the physician's training and role at the VHA.

Controlling for these variables in our model estimation does not materially change the conclusions of our analysis. Figure A.3 reports the results of regressions that permute the control variable sets to cover every possible combination of the above list. In Panel A, we find a similar increase in sensitivity to the $CHADS_2$-component of stroke treatment effects after guideline awareness in each model, regardless of the set of included controls. In Panel B, we show that the unexplained variance in treatment propensity does not change substantially, even after we control for these detailed patient and physician characteristics.

A.3 Construction and Prediction Details

In this appendix, we provide further details of various constructed objects and predictions.

Predicted Mortality. We construct a measure of predicted mortality based on patient characteristics, using mortality outcomes among patients in the VHA who do not have atrial fibrillation. To select this sample, we first take all patients at the VHA who are administratively linked to a primary care physician in 2010 for the first time. We exclude patients who are administratively linked in a primary care relationship to nurse practitioners or physician assistants; we also exclude patients linked to primary care doctors whose service section falls under the following categories: psychiatry, geriatrics, orthopedics, surgery, infectious diseases, rheumatology, neurology, renal failure, spinal conditions, cardiology, oncology, sleep, and behavioral health. We exclude patients whose age is below 18 years or above 100 years. Importantly, we exclude patients with a history of atrial fibrillation. This leaves us with 833,298 patients. We construct an OLS prediction of 3-year mortality using patient age, weight, height, vital signs, hemoglobin, gender, the indicator for whether a patient is white, and indicators for Elixhauser comorbidities.

Treatment Propensity. We construct the propensity of treatment as a function of patient characteristics in the VHA data and denote this object as $\Pr_{\text{VHA}}(W_i \mid X_i)$. This treatment propensity model is constructed as the OLS linear probability of being treated with anticoagulants given patient characteristics from the VHA data using patient covariates that also exist in the AFI database, as reported in Table A.1. We then take this model to calculate $\Pr_{\text{VHA}}(W_i \mid X_i)$ for each patient in the AFI database, in order to test whether physician treatment decisions in the VHA reveal additional signal about treatment effects heterogeneity in the AFI database, in Section 5.3 and Appendix Table A.4.

Causal- and Regression-Forest Predictions. We use the `grf` package developed by Tibshirani et al. (2020) to run causal and regression forests. Details about the algorithm can be found at `https://github.com/grf-labs/grf`. In brief, both causal and regression forests form predictions by creating a number of decision trees, each trained on a random sample of observations. In each tree, nodes are split recursively by a random subset of characteristics. This process occurs until no node can be split any further, as determined by parameters of the algorithm that we discuss below. Causal forests split nodes with the objective of maximizing differences in treatment effects—the difference between average outcomes among treated and untreated observations—between child nodes. Regression forests split nodes with the objective of maximizing differences in average outcomes between child nodes. After a decision tree is formed, predicted treatment effects (or outcomes) for a given vector of characteristic values are determined by the average treatment effect (or outcome) in the terminal node that contains those characteristic values. The prediction of the forest is the average prediction over each tree in the forest.

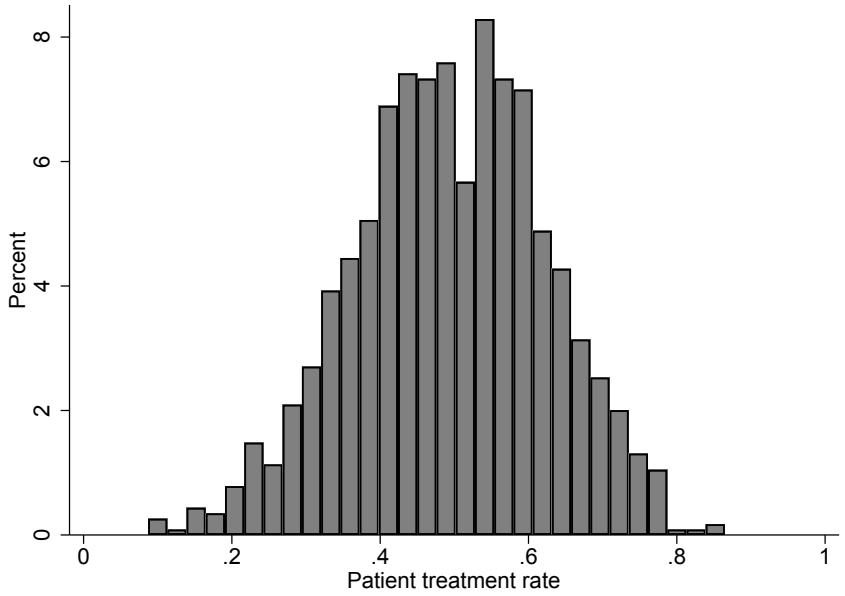
In training both causal and regression forests, we use the default `honesty` option. This ensures

that separate random samples of the data are used to determine splits and to compute average treatment effects or outcomes in the nodes (Athey and Wager, 2019).

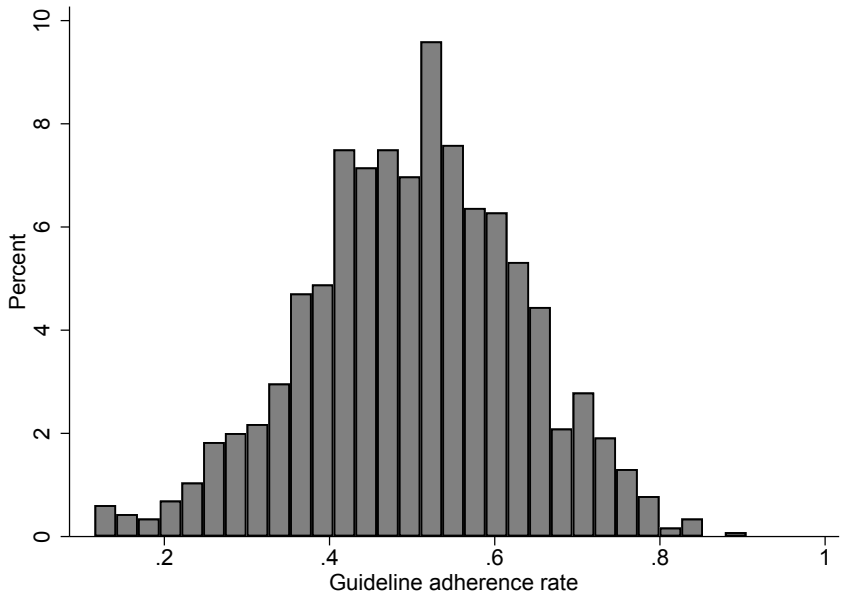
We use the following parameter values in our algorithm. In the causal forest, we set minimum node size (`min.node.size`) such that all nodes must contain at least 150 treated observations and 150 untreated observations. In the regression forest, we set this parameter so that each node must have at least 300 observations. We set these constraints so that each node will have a sufficient number of realized strokes, as the outcome of stroke is relatively uncommon, at 6.5%, as shown in Appendix Table A.4. We set the parameter `alpha` to 0.2, which restricts imbalance of splits so that each child node required to be greater than 20% of the size of its parent node. We set the number of trees grown in the forest (`num.trees`) at 800. We set `sample.fraction`, or the fraction of the data used to build each tree, at 0.75, within the default range. We set `honesty.fraction`, or the fraction of the training sample used for determining splits, at 0.75, also within the default range. We left the number of variables considered in each split (`mtry`) at the default value, which implied that all variables were considered in each split.

Figure A.1: Distribution of Physician Treatment Decisions

A. Treatment

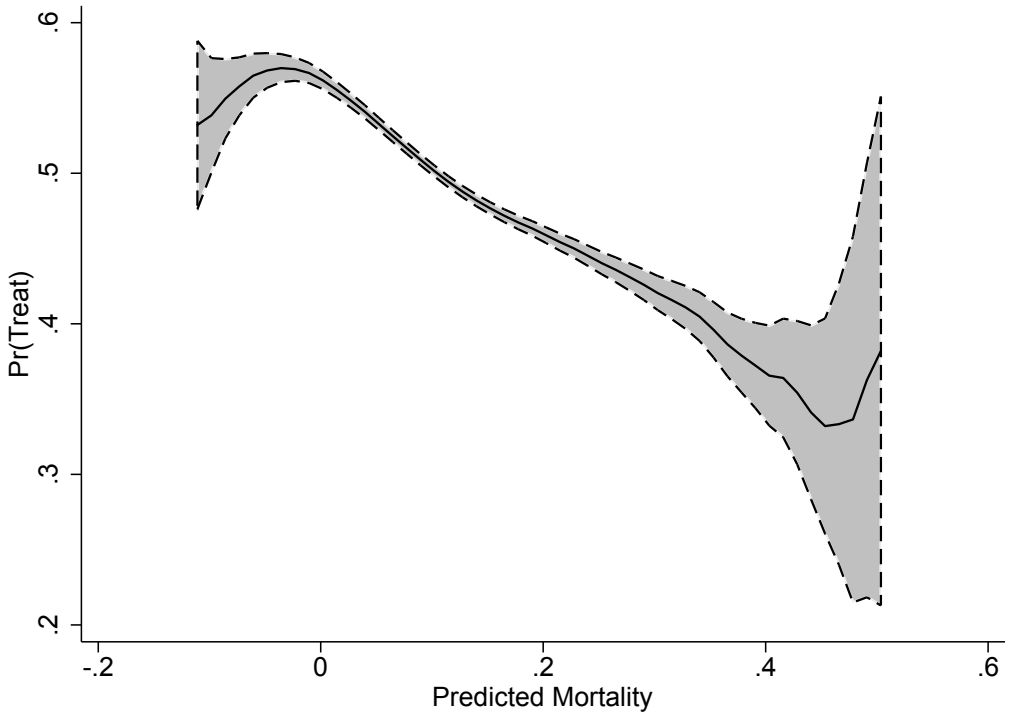


B. $CHADS_2$ Adherence



Notes: These figures show the distribution of treatment rates and $CHADS_2$ adherence rates across physicians. They cover the subsample of 1,146 physicians treating at least 30 patients in our final analysis sample. This covers 50,426 patients treated by the higher volume doctors, or a little less than half of the VHA sample defined in Table 2. Panel A shows the distribution of treatment rates. Panel B shows the distribution of $CHADS_2$ adherence rates. We define $CHADS_2$-adherent anticoagulation decisions as follows: No anticoagulation for patients with a $CHADS_2$ of 0 and anticoagulation for patients with a $CHADS_2$ score greater than or equal 2; we omit patients with a $CHADS_2$ score of 1 from this calculation, since the ACC and ACCP guideline allowed for either anticoagulation or aspirin for these patients.

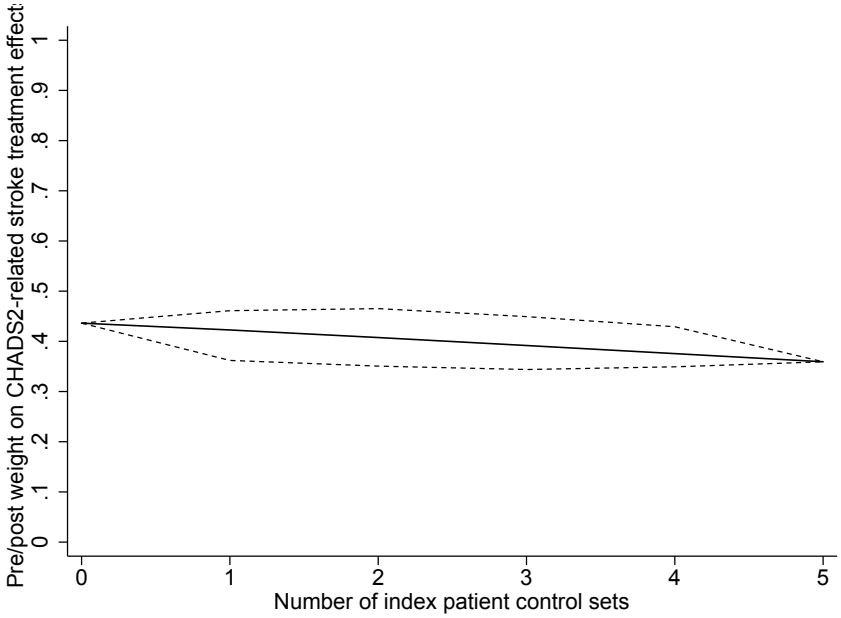
Figure A.2: Treatment Probability by Predicted Mortality Risk



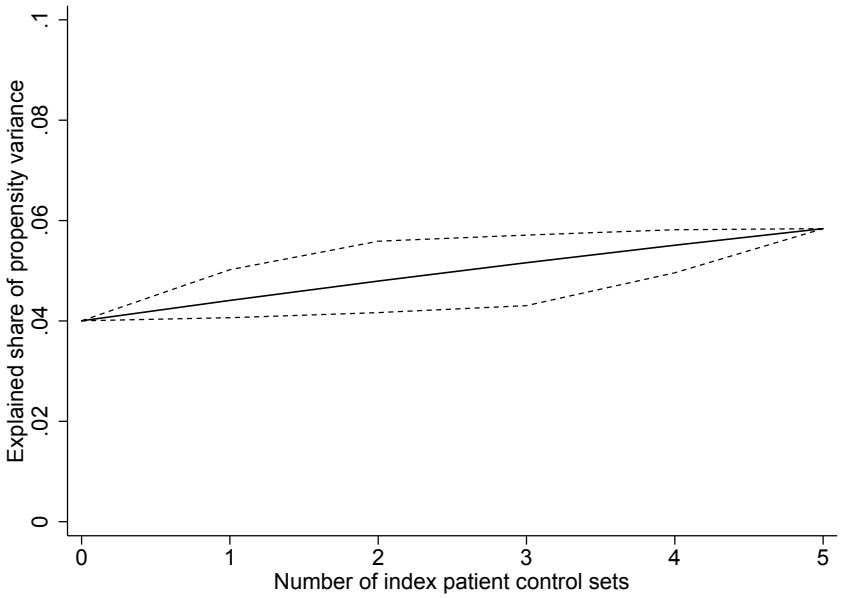
Notes: This figure shows the probability of anticoagulation as a function of predicted mortality risk in the VHA sample. The statistical model predicting mortality risk is calculated in a separate sample of patients receiving primary care at the VHA who have no diagnosis of atrial fibrillation. The curve fits the observed data with a kernel weighted local polynomial. The shaded area represents the 95% confidence interval.

Figure A.3: Stability of the Structural Results

A. Increased weight on $\tau^{s(c)}$, $\alpha_{\text{pre}}^{s(c)}/\alpha_{\text{post}}^{s(c)}$



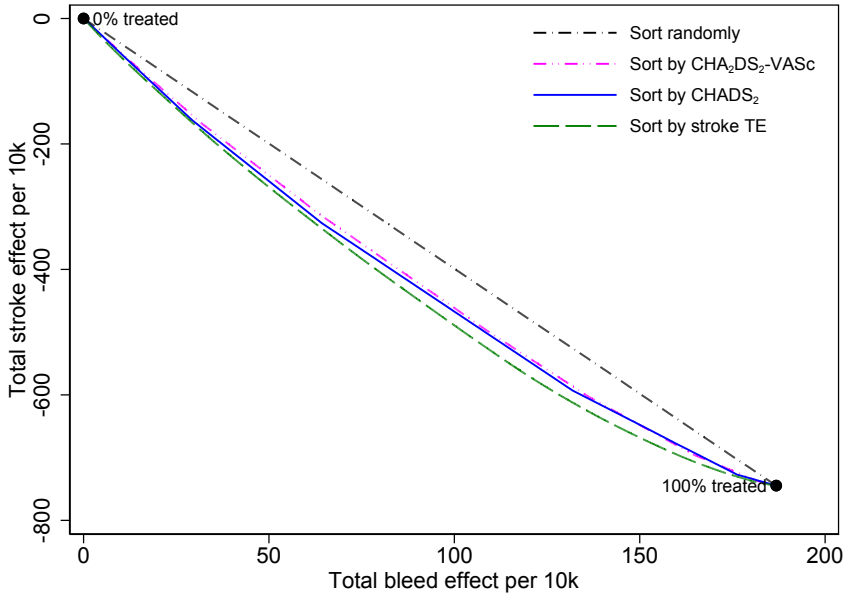
B. Explained Share of Latent Variable



Notes: These graphs illustrate how key results of our structural model vary as we include various sets of control variables in its estimation. Panel A examines the increased decision weight physicians place on $CHADS_2$-related stroke treatment effects with $CHADS_2$ awareness, or $\alpha_{\text{pre}}^{s(c)}/\alpha_{\text{post}}^{s(c)}$. Panel B examines the proportion of variance in the latent variable that we can explain with observable characteristics (i.e., the complement of the share explained by $\sigma^2_{\varepsilon,g}$). In each panel, we include varying sets of patient characteristics in $f(X_i)$ in our structural model stated in Equation (10). We estimate the baseline specification, shown in Column 1 of Table 4. The solid line shows the mean value of the statistic among specifications with the indicated number of control sets; the top (bottom) dashed line shows the maximum (minimum) of the statistic. The control variables are detailed in Appendix Section A.2.

Figure A.4: Counterfactual Outcomes with CHA_2DS_2-VASc Guideline

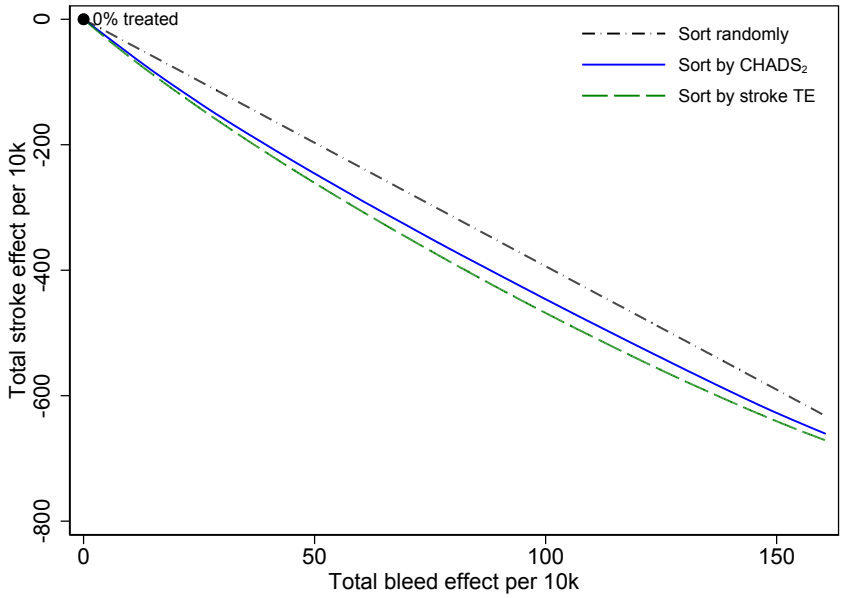
A. Strict Guideline Adherence



Notes: Relative to Figure 6, this figure includes an additional set of counterfactual outcomes under strict adherence to the CHA_2DS_2-VASc guideline. Like other counterfactuals of strict adherence, strict adherence to this guideline implies ranking patients by their CHA_2DS_2-VASc score. The CHA_2DS_2-VASc score assigns one point for congestive heart failure, hypertension, age 65-74 years, female sex, vascular disease, and diabetes; it assigns two points for age 75 years or older, and for stroke, transient ischemic attack, or thromboembolism. Details for this figure are otherwise described in the notes for Figure 6.

Figure A.5: Counterfactual Outcomes, Fixed Predicted Mortality Distribution

A. Strict Guideline Adherence



Notes: Relative to Figure 6, this figure shows counterfactual outcomes for patient rankings that hold fixed the predicted mortality distribution of treated patients at every point along the curve. The fraction of treated patients in each 5-year age bin matches the fraction observed treated in our sample. Within each predicted mortality group, patients are ranked according to scores in each guideline. In order to maintain a predicted mortality distribution of treated patients, it is not possible to treat 100% of patients. This is reflected by curves for counterfactual outcomes not reaching the same bottom-right corner of Figure 6. Only counterfactual outcomes for strict adherence and for random sorting are changed in this figure; outcomes for awareness scenarios are unchanged from Figure 6. For more details, see notes to Figure 6.

Table A.1: AFI Database Covariates and VHA Definitions

AFI variable	VHA definition
Age	Coded demographic at time of AF diagnosis
Female	Coded demographic
Height	Coded vital closest to AF diagnosis
Weight	Coded vital closest to AF diagnosis
White race	Coded demographic
Smoker at all	Smoking record ±1 year of AF diagnosis
Systolic blood pressure	Vital closest to post AF diagnosis, with valid reading in range [70,300]
Diastolic blood pressure	Vital closest to post AF diagnosis, with valid reading in range [25,200]
Hemoglobin	Lab value closest to post AF diagnosis
History of angina	ICD9: 433.X, 434.X, 436.X
History of congestive heart failure	ICD9: 398.X, 402.01, 402.11, 402.91, 428.X
History of diabetes	ICD9: 249.X, 250.X, 357.20
History of hypertension	ICD9: 401.X-405.X
History of myocardial infarction	ICD9: 410.X, 412.X
History of peripheral vascular disease	ICD9: 093.0, 437.3, 440.X, 441.X, 443.1-443.9, 447.1, 557.1, 557.9, V43.4
History of TIA or stroke	ICD9: 433.XX, 434.XX, 435.XX, 436.XX
History of TIA	ICD9: 435.XX

Note: This table lists all the covariates from the AFI database that are used in our causal-forest implementation. We set the treatment variable to be 1 for patients treated with warfarin and 0 for patients on control or ASA therapy (aspirin). Observations are dropped for those on low warfarin and low warfarin plus ASA therapy. Patients with especially relevant disease histories missing are also excluded from the sample. These disease histories include Transient Ischemic Attack (TIA), stroke, diabetes and hypertension. Patients whose date of birth predates the date for inclusion in the trial randomization are also excluded. We regrouped the variable "Race" so that it equals 1 when the patient is white and 0 otherwise. The variable "Smoker at all" equals 1 if the subject is a current or past smoker and 0 otherwise. Otherwise, missing variables are imputed using the regression forest method.

Table A.2: Variable Importance

Stroke Causal Forest	Stroke Regression Forest	Bleed Causal Forest	Bleed Regression Forest
Stroke Risk (−)	History of Stroke or TIA (+)	Systolic Blood Pressure (+)	Age (+)
History of Stroke or TIA (+)	Systolic Blood Pressure (+)	Weight (−)	History of TIA (+)
Age (+)	History of TIA (−)	History of TIA (+)	History of PVD (−)
History of TIA (−)	White (−)	Bleed Risk (−)	Smoker At All (−)
Weight (−)	Hemoglobin (−)	Smoker At All (+)	White (−)
Systolic Blood Pressure (−)	Age (+)	Hemoglobin (−)	Hemoglobin (−)
Hemoglobin (+)	Weight (−)	Diastolic Blood Pressure (+)	Diastolic Blood Pressure (−)
Height (+)	Height (−)	Age (+)	Systolic Blood Pressure (−)
Smoker At All (−)	History of Hypertension (+)	Height (+)	Weight (+)
Diastolic Blood Pressure (+)	Diastolic Blood Pressure (−)	Female (+)	Height (+)

Note: Each column of this table shows the top 10 most important variables for each forest in descending order of importance. Risks computed from regression forest using only the control sample are then used as an input into causal forests. The +/− signs following each variable indicate the sign of its coefficient in a bivariate linear regression with the corresponding forest output as the dependent variable.

Table A.3: Balance Table

Patient Characteristics	Control Group Mean	Treatment Group Mean	Coefficient
Age	70.4	70.3	0.18
			(0.47)
Congestive Heart Failure	0.30	0.29	-0.006
			(0.012)
Age above 65	0.77	0.76	0.003
			(0.012)
History of Hypertension	0.45	0.46	-0.008
			(0.015)
History of Stroke	0.17	0.12	-0.006
			(0.008)
History of Diabetes	0.14	0.14	-0.006
			(0.010)
Male	0.67	0.68	-0.016
			(0.013)

Notes: This table shows the unadjusted means of each patient characteristics in the treatment and control group. The last column shows results of a regression of each patient characteristics on trial fixed effects and treatment indicator in AFI database. Standard errors are shown in parentheses.

Table A.4: Causal-Forest BLP Validation Regressions

	Stroke			Bleed
	(1)	(2)	(3)	(4)
Treatment, W_i	−0.043***	−0.042***	−0.043***	0.019***
	(0.007)	(0.007)	(0.007)	(0.005)
Treatment effect interactions				
$W_i \times \left(\hat{\tau}^o_{-j}(X_i) - \overline{\tau}^o\right)$	0.823***		0.843***	−0.374
	(0.230)		(0.230)	(0.691)
$W_i \times \left(\hat{\tau}^{s(c)}_{-j}(X_i) - \overline{\tau}^{s(c)}\right)$		1.329***		
		(0.408)		
$W_i \times \left(\hat{\tau}^{s(r)}_{-j}(X_i) - \overline{\tau}^{s(r)}\right)$		0.645**		
		(0.265)		
$W_i \times \left(\Pr_{\text{VHA}}(W_i \mid X_i) - \overline{\Pr_{\text{VHA}}(W_i)}\right)$			−0.038	
			(0.088)	
Outcome mean	0.065	0.065	0.065	0.030
Observations	4,720	4,720	4,720	4,720
Trial fixed effects	Yes	Yes	Yes	Yes
Predicted outcome controls				
$\hat{Y}^o_{-j}(X_i)$	Yes	Yes	Yes	Yes
$P_j \times \left(\hat{\tau}^o_{-j}(X_i) - \overline{\tau}^o\right)$	Yes		Yes	Yes
$P_j \times \left(\hat{\tau}^{s(c)}_{-j}(X_i) - \overline{\tau}^{s(c)}\right)$		Yes		
$P_j \times \left(\hat{\tau}^{s(r)}_{-j}(X_i) - \overline{\tau}^{s(r)}\right)$		Yes		
$\Pr_{\text{VHA}}(W_i \mid X_i)$			Yes	

Notes: This table reports the coefficients of best linear predictor (BLP) validation regressions of stroke and bleed outcomes. Columns 1 and 4 corresponds to Equation (5), which interacts treatment with the demeaned full treatment effect $\left(\hat{\tau}^o_{-j}(X_i) - \overline{\tau}^o\right)$ for stroke and bleed, respectively. Column 2 corresponds to Equation (6), which interacts treatment with the demeaned CHADS_2 and residual components of the stroke treatment effect, or $\left(\hat{\tau}^{s(c)}_{-j}(X_i) - \overline{\tau}^{s(c)}\right)$ and $\left(\hat{\tau}^{s(r)}_{-j}(X_i) - \overline{\tau}^{s(r)}\right)$. Column 3 reports the specification in Column 1 with the additional interaction between treatment and demeaned VHA propensity to treat, or $\left(\Pr_{\text{VHA}}(W_i \mid X_i) - \overline{\Pr_{\text{VHA}}(W_i)}\right)$. The VHA propensity to treat, $\Pr_{\text{VHA}}$, is constructed as the OLS probability of treatment in the VHA data given patient characteristics and exported to the AFI databse, detailed further in Appendix A.3. All specifications control for trial fixed effects and regression forest predictions of the outcome estimated in the control groups of leave-out trials, or $\hat{Y}^o_{-j}(X_i)$. Additionally, the specifications in Column 1, 3 and 4 control for treatment probability in each trial interacted with demeaned treatment effect, or $P_j \times \left(\hat{\tau}^o_{-j}(X_i) - \overline{\tau}^o\right)$, where P_j denotes the treatment probability in each trial. Analogously, the specification in Column 2 controls for the treatment probability in each trial interacted with demeaned CHADS_2 and residual components of the stroke treatment effect, or $P_j \times \left(\hat{\tau}^{s(c)}_{-j}(X_i) - \overline{\tau}^{s(c)}\right)$ and $P_j \times \left(\hat{\tau}^{s(r)}_{-j}(X_i) - \overline{\tau}^{s(r)}\right)$. The specification in Column 3 additionally controls for the treatment propensity main effect. The sample used in the validation regression exclude patients ineligible for Warfarin, reducing the sample size to 4,720. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table A.5: Patient Characteristics in the VHA Data and Across Trials

	VHA Data	AFI Database Trial							
		AFASAK1	BAATAF	CAFA	SPINAF	SPAF2	AFASAK2	EAFT Group 1	PATAF Group 1
Age	74.0	73.0	67.8	68.0	67.8	70.2	73.7	70.7	70.5
Stroke treatment effect, $\hat{\tau}^s$	−0.07	−0.05	−0.03	−0.03	−0.02	−0.03	−0.04	−0.07	−0.04
Bleed treatment effect, $\hat{\tau}^b$	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
Congestive heart failure	0.15	0.52	0.26	0.22	0.31	0.21	0.70	0.09	0.00
Hypertension	0.84	0.32	0.51	0.39	0.59	0.53	0.44	0.44	0.36
Age $\geq$ 65	0.52	0.84	0.67	0.68	0.69	0.73	0.90	0.81	0.79
Diabetes	0.36	0.08	0.15	0.12	0.19	0.16	0.12	0.13	0.17
Stroke or TIA history	0.15	0.06	0.03	0.03	0.08	0.06	0.07	1.00	0.01
Male	0.99	0.54	0.73	0.75	1.00	0.70	0.61	0.59	0.46
Number of patients	113,270	985	417	378	567	1,094	339	668	272

Note: This table shows the mean of each listed patient characteristic in the VHA data, as well as each of the eight trials with both control and treatment arms in the AFI database. Note that there are a total of ten trials in the original AFI database. For our analysis, we define patients who were treated with aspirin alone as being untreated with anticoagulation. We drop observations for patients on low warfarin or low warfarin plus aspirin therapy. After these modifications, eight trials remain with both treatment and control arms. In three of the eight trials, patients are divided into eligible versus ineligible groups for anticoagulation. This table also excludes patients ineligible for anticoagulation. AFASAK1: Atrial Fibrillation, Aspirin, and Anticoagulation Study 1; BAATAF: Boston Area Anticoagulation Trial for Atrial Fibrillation; CAFA: Canadian Atrial Fibrillation Anticoagulation; SPINAF: Stroke Prevention in Non-rheumatic Atrial Fibrillation; SPAF2: Stroke Prevention in atrial Fibrillation; AFASAK2: Atrial Fibrillation, Aspirin, and Anticoagulation Study 1; EAFT Group 1: European Atrial Fibrillation Trial; PAATAF Group 1: Primary Prevention of Arterial Thromboembolism in Atrial Fibrillation.

Table A.6: Counterfactual Outcomes, Fixed Predicted Mortality Risk Distribution

		Strokes prevented		Bleeds induced	
	Percent of patients treated	Per 10K patients	Percent of maximum	Per 10K patients	Percent of maximum
A: Benchmarks (from Table 5)					
Observed treatment choices	49.8%	371	49.8%	93	49.8%
Randomly assigned treatment	49.8%	371	49.8%	93	49.8%
All patients treated	100%	745	100%	187	100%
B: Assignment within Predicted Mortality Bins					
Randomly assigned treatment	49.8%	366	49.2%	93	49.8%
Adherence to $CHADS_2$ guideline	49.8%	419	56.3%	93	49.8%
Adherence to stroke TE guideline	49.8%	441	59.2%	93	49.8%

Notes: This table reports counterfactual outcomes that hold the fixed the predicted mortality risk distribution of treated patients. For comparison, Panel A reproduces benchmark results from Table 5, in which treatment probability is not held fixed within predicted mortality bins. In Panel B, the fraction of treated patients in each ventile of predicted mortality bin matches the fraction observed treated in our sample. We also hold the overall percentage of treated patients fixed at 49.8%. For adherence counterfactuals, within each predicted mortality group, patients are treated according to rankings implied by the noted guideline. Figure A.5 shows counterfactual outcomes varying the overall percentage of treated patients. For more details, see notes to Table 5 and Figure A.5.