Are Guidelines Worth Following? Treatment Decisions Under Scientific Uncertainty

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

November 1, 2024

Abstract

Physicians often deviate from medical guidelines, but most guidelines focus on a single disease while ignoring complex interactions that render treatment harmful. We study prescribing decisions covered by a prominent guideline for treating atrial fibrillation in patients at higher risk for stroke. Following the guideline's introduction, physicians rapidly demonstrated awareness of it, yet they often declined to adhere to it, even for patients with the highest stroke risk. Using machine learning on eight randomized trials of anticoagulation, we find significant stroke treatment effect heterogeneity. However, the trials do not (and were not designed to) reveal heterogeneity in the bleeding side effect of anticoagulation. In the observational data, we find that bleed risk is positively correlated with stroke reduction, raising the possibility that the patients who benefit most from treatment may also suffer greater harm from it. Analyzing the performance of known and optimal treatment rules, we show that known guidelines may perform worse than randomly treating patients. Finally, we develop a guideline under an objective that maximizes the worst outcome under scientific ambiguity and show that this guideline may significantly improve welfare relative to status quo treatment decisions. Analyzing physician behavior under this benchmark, we find that ranking patients according to predicted physician treatment under the "wisdom of the crowds" performs better than strict adherence to known guidelines.

*Abaluck: Yale University and NBER, jason.abaluck@yale.edu. Agha: Harvard University and NBER. agha@hcp.med.harvard.edu. Chan: Stanford University, Department of Veterans Affairs, and NBER, david.c.chan@stanford.edu. Singer: Harvard University and Massachusetts General Hospital, desinger@mgh.harvard.edu. Zhu: Stellation Care, diana@stellationcare.com. 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, Noah Boden-Gologorsky, Liberty Greene, Johnny Huynh, Chris Lim, Uyseok Lee, Natalie Nguyen, Erick Ore Matos, Simone Pandit, Francis Peng, Jonatas Teixeira Prates, and Melinda Xu for invaluable research assistance. We thank Carl van Walraven and the Atrial Fibrillation Investigators for generously sharing the AFI database for reanalysis.

1 Introduction

Medical treatments typically target one outcome but may adversely impact others. For example, CT scans may detect cancer or blood clots, but they may also expose patients to radiation. Anticoagulants in patients with atrial fibrillation may reduce stroke risk, but they may also increase bleeding risk. From a scientific standpoint, the causal evidence on the benefits of treatment is typically better characterized than evidence on harm, as patients who may experience both benefit and harm from treatment are often excluded from clinical trials. As a result, medical guidelines are often one-sided, quantifying who is likely to benefit most from treatment but not explicitly weighing the benefits against the harms, and physicians are expected to use their discretion for decisions for which guidelines capture only partial information.

In this paper, we study how physicians respond to the introduction of a prominent clinical guideline and whether alternative treatment rules that account for *scientific uncertainty*—under which there is an asymmetry in information between patient-specific benefits and harms—could lead to better patient outcomes.[1] The guideline concerns 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 to reduce stroke risk. However, anticoagulation also increases the risk of life-threatening hemorrhage (hereafter, bleed), including gastrointestinal and 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 stroke risk score for patients with atrial fibrillation (Gage et al., 2001). Clinical guidelines recommend tailoring treatment decisions based on 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 for any clinical condition.[2]

We examine real-world treatment decisions among nearly 5,800 physicians treating 113,000 newly diagnosed atrial fibrillation patients in the Veterans Health Administration (VHA) from 2002-

[1] By scientific uncertainty, we refer to the concept in which the probabilities of states are unknown. This is referred to interchangeably by "ambiguity," "uncertainty" (as opposed to "risk"), and "Knightian uncertainty" in the literature. See Gilboa and Marinacci (2013) for a review.

[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.

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' awareness 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 mention of the $CHADS_2$ score, 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 aware of the $CHADS_2$ score still do not adhere to guidelines in more than 40% of cases.

To understand how the $CHADS_2$ score and alternative guidelines impact patients, we must understand how treatment effects vary across patients. To this end, we generate novel ML estimates of heterogeneous treatment effects from randomized controlled trial (RCT) microdata. To estimate heterogeneous treatment effects, we use detailed patient characteristics, clinical outcomes, and randomized treatment assignment 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).

Based on the RCT data, our approach finds substantial heterogeneity in stroke treatment effects. At the 90th percentile, warfarin reduces strokes by 6 percentage points, while at the 10th percentile, it reduces strokes by 3 percentage points. The $CHADS_2$ score explains a substantial share of our estimated variation in stroke treatment effects ($R^2 = 0.49$); nonetheless, there remains meaningful residual variation in stroke treatment effects we can validate across RCTs in the AFI database. In contrast, we cannot validate heterogeneity in bleed treatment effects in the AFI database. The asymmetry in evidence between benefits and risks of treatment is emblematic of many clinical scenarios: RCTs generally select patients who are more likely to benefit from a treatment than to be harmed by it, often excluding patients with significant comorbidities. Yet, clinicians must extrapolate this information to patients at higher risk for complications. Herein lies a fundamental problem in evidence-based guideline development (Tinetti et al., 2004).

Combining stroke treatment effect estimates from the RCT data and accounting for the uncertainty

about bleed treatment effects, we assess the implications of $CHADS_2$ guideline adherence on patient outcomes in our VHA sample. The aggregate status quo outcomes perform similarly to a random treatment rule: They prevent roughly as many strokes for a given number of bleeds induced as would assigning anticoagulation to patients at random. Under the assumption of homogeneous bleed treatment effects (or bleed treatment effects orthogonal to stroke treatment effects), better adherence to the $CHADS_2$ score would do considerably better: Treatment would be reallocated to patients whose strokes are more likely to be prevented. However, if we instead assume that bleed treatment effects are proportional to bleed risk, we reach the *opposite* conclusion: because stroke treatment effects and bleed risk are so highly correlated, strict adherence with the $CHADS_2$ score would perform *worse* than random treatment, with fewer strokes prevented per bleed induced. We arrive at a similar conclusion for an updated clinical guideline—the CHA_2DS_2-VASc score—introduced in 2010. Consistent with the prior literature (Fawzy et al., 2019), we find evidence that this concern may explain much of the non-adherence to known recommendations based on stroke risk: Physicians are more likely to adhere to the $CHADS_2$ guideline for patients with lower bleed risk.

This motivates us to study how scientific uncertainty about the heterogeneity of anticoagulation bleed effects may starkly impact the optimality of various treatment rules. Given the lack of evidence, we consider two plausible assumptions: First, reasonable clinicians may assume that bleed treatment effects are unrelated to stroke treatment effects. Second, clinical intuition and evidence from other settings may suggest that bleed treatment effects are proportional to untreated bleed risks (positively correlated with the stroke-prevention benefits of anticoagulation). Under the first assumption of homogeneous bleed treatment effects, the optimal treatment rule would rank patients by their stroke treatment effects, much like the $CHADS_2$ and CHA_2DS_2-VASc scores do. Under the second assumption of correlated bleed treatment effects proportional to bleed risks, the optimal treatment rule would treat patients according to the *ratio* of their stroke treatment effects and bleed risks. We show that the optimal treatment rules under these two assumptions may disagree for a wide set of patients, especially when the benefits and harms of treatment are positively correlated across patients. Importantly, adopting the wrong optimal guideline (based on the wrong assumption) may result in outcomes worse than random treatment. The circumstances under which this may occur are not too extreme: As long as the coefficient of variation of treatment harms is at least as large as the coefficient of variation of treatment benefits, a sufficient degree of positive correlation between treatment harms and benefits may result in (mistakenly) optimal guidelines performing worse than random treatment.

We further characterize a new optimal treatment rule that accounts for scientific ambiguity by

maximizing the welfare of the worse outcome between the two assumptions. We show that this "maxmin" treatment rule orders patients so that all sets of treated patients implied by the ranking have the same bleed risk as the overall population of patients, effectively orthogonalizing the ranking of patients to the possibility of heterogeneous treatment harms. In the setting of anticoagulation, we find that, while mistakenly optimal guidelines for anticoagulation may perform only slightly better than random treatment, the maxmin treatment rule may prevent 19.6% more strokes for the same number of bleeds in the status quo (or the random-treatment benchmark).

Finally, we reassess physician decisions by estimating a within-physician model of the probability of treatment according to patient characteristics. We show that ranking patients according to this "wisdom of the crowds" treatment rule could significantly outperform known guidelines based on the $CHADS_2$ score and its newer CHA_2DS_2-VASc variant. Under the criterion of the worse-case assumption, the crowd treatment rule performs only slightly worse than the optimal treatment rules under the two assumptions of treatment harms: All three rules perform about the same as the random benchmark under this worse-case criterion. Interestingly, this suggests that outcomes could be improved by recommending treatments to physicians based on what physicians do on average, much in the spirit of artificial intelligence (AI) implementations in other settings like radiology (see, e.g., Agarwal et al., 2023). The gains from such an approach (relative to the aggregate status quo outcomes, which are similar to random treatment) could derive from reducing symmetric decision errors and practice-style variation across physicians. Consistent with the idea of crowd wisdom, we find that crowd treatment rules estimated from cardiologists perform only slightly better than those estimated from primary care physicians; similarly, the crowd treatment rules from more-experienced physicians perform only slightly better than those from less-experienced ones.

Our results contribute to an active literature in economics studying the potential for machine-based algorithms to improve decision-making. Evidence from clinical care (Abaluck et al., 2016; Mullainathan and Obermeyer, 2022; Chan et al., 2022) and judge bail decisions (Kleinberg et al., 2018) suggests that human experts make frequent mistakes that could, in principle, be corrected by protocolized decision-making; we analyze how clinicians respond to the introduction of an important guideline 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 in aggregate. We build on this research by studying highly skilled experts in an important clinical decision. We analyze how new awareness

of a guideline changes behavior and outcomes, and we simulate the effects of adherence to known and novel guidelines using estimates of heterogeneous treatment effects. Importantly, we directly address the implications of scientific uncertainty for the performance and formulation of guidelines.

Developing an approach using pooled RCT data to estimate treatment effect heterogeneity, we provide some support for this heterogeneity in the observational data. By contrast, many prior papers comparing machine decisions to human discretion have relied on observational data and quasi-experimental assumptions to reach conclusions about misallocation.[3] The use of RCTs for the detection of treatment effect heterogeneity is particularly important given recent research in econometrics that suggests that quasi-experimental assumptions for recovering treatment effects may not hold within small cells of the data (e.g., Kolesar et al., 2015; de Chaisemartin, 2017; Frandsen et al., 2019); these stricter assumptions would be required to extract valid heterogeneous effect heterogeneity at a detailed level. Further, we directly investigate the external validity of our RCT-derived findings in the observational data, providing some tools to address concerns about extrapolation raised by Manski (2017). It is noteworthy that the vast majority of clinical guidelines are generated from RCT data or observational predictions, not from quasi-experimental evidence.

In the medical literature, research has focused on how treatment decisions relate to clinical guidelines. The literature has shown a 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 aware of the guideline continue to deviate frequently. There is little empirical evidence examining the implications of guideline non-adherence for health outcomes in practice. Previous research has investigated variation across physicians and found that physicians who adhere more closely to guidelines have better outcomes (Mehta et al., 2015; Currie and MacLeod, 2020; Cuddy and Currie, 2020), although this literature generally does not separate the impact of guidelines from other differences across physicians. We contribute to this literature by assessing physician behavior with estimates of the benefits of treatment based on detailed clinical characteristics potentially omitted from known guidelines and accounting for the possibility of

[3] Many of these papers compare treatments and outcomes across decision-makers, usually assuming a common ranking 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. (2022) 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; Chan et al., 2022).

[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 various 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.

scientific uncertainty on the harms of treatment.

The remainder of this paper is organized as follows. Section 2 describes the clinical setting and data. Section 3 provides evidence of the limited impact of $CHADS_2$ awareness on treatment decisions. Section 4 presents our estimates of causal-forest treatment effects. Section 5 analyzes, under scientific uncertainty about the distribution of treatment harms, the performance of aggregate status-quo decisions and the known $CHADS_2$ and CHA_2DS_2-VASc guidelines. Section 6 considers optimal guidelines depending on assumptions about treatment harms and analyzes the performance of optimal guidelines; it also considers the performance of treatment rules formed by "wisdom of the crowds" estimates of within-physician behavior. Section 8 concludes with a discussion of policy implications.

2 Setting and Data

2.1 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, 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 more than double the risk of major bleeding (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

[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 bleeding risks (Lane and Lip, 2012). Guideline recommendations for their use (vs. no anticoagulation) rely on the same stroke risk scores. Warfarin remains the mainstay drug for anticoagulation in atrial fibrillation (Hsu et al., 2016).

$CHADS_2$ 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 $CHADS_2$ score has become one of the most widely recognized risk scores in clinical practice. 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 the recommended patients being prescribed anticoagulation (Hsu et al., 2016; Piccini and Fonarow, 2016). 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 is 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 simulations, we contrast the $CHADS_2$ and CHA_2DS_2-VASc scores.

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 exists to guide physicians in assessing 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).

2.2 Veterans Health Administration Data

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 narrow cohort of patients who are likely to have a new, confirmed atrial fibrillation diagnosis and to receive care at the VHA (Turakhia et al., 2013; Perino et al., 2017).

We first identify potentially new diagnoses of atrial fibrillation (ICD9 codes 427.3X) 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 likely responsible for anticoagulation decisions, we require each patient to visit 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. 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 concerns not fully captured in the RCT data (e.g., bleed risk, fall risk, frailty, multimorbidity). To capture the anticoagulation decision, we rely on VHA records of prescriptions for warfarin or a novel oral anticoagulant (dabigatran, rivaroxaban, apixaban, edoxaban). The VHA records include prescriptions that are dispensed by the VHA as well as prescriptions that are paid for by the VHA. The vast majority of prescriptions in our sample are for warfarin: Fewer than 2% of patients are prescribed a novel oral anticoagulant (NOAC). Summary statistics on the VHA atrial fibrillation cohort are reported in Table 3, Column 1.

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.[6] We proxy the timing of a physician's first awareness of the guideline by the date of 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.[7]

As we will show in Section 3, 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 who have not yet written a note mentioning the $CHADS_2$ score as not currently aware of the guideline. As an important caveat, it is likely that many physicians whom we label as currently unaware 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 unaware physicians are less likely to follow guideline recommendations than physicians with observed guideline use. Although unaware physicians may have heard of the guideline, they are unlikely to be making explicit use of its recommendations to drive their clinical decision-making.

2.3 Atrial Fibrillation Investigators Database

To estimate heterogeneous treatment effects—which we use to assess physician decision-making and 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 placebo or control.[8] Details of the AFI database have been documented elsewhere (e.g., van

[6] 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.

[7] 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.

[8] There are a total of ten trials in the original AFI database. For our analysis, we define patients 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

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.3, we list the full set of characteristics that we use from the AFI database. In Appendix Table A.5, we report the results of balance tests, which suggest successful randomization in these clinical trials.

3 Guideline Awareness and Treatment Decisions in the VHA

3.1 Trends in Guideline Use and Prescription Patterns

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).[9] We, therefore, begin by describing trends in prescribing behavior for groups of patients defined by their $CHADS_2$ score.

Figure 2 Panel B displays trends in anticoagulation rates for patients with low risk of stroke ($CHADS_2$ score of 0 or 1) and patients with higher risk of stroke ($CHADS_2$ score of 2 and above). 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. 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

causal-forest implementation to increase power.

[9] 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 after our study period.

stroke risk ($CHADS_2$ score of 4 or greater) remain slightly less likely to be treated than patients with moderate stroke risk ($CHADS_2$ of 2 or 3). These findings establish that physicians had only weak adherence to recommended treatment patterns over the entire study period.

These trends towards reduced anticoagulant use for patients with low $CHADS_2$ scores coincide with the gradual diffusion of the $CHADS_2$ guideline into clinical practice. 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; 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. Similarly, we find the proportion of atrial fibrillation patients with the $CHADS_2$ score mentioned in their own doctor's notes rises from 0% to 17% over the same period.

3.2 Effect of Guideline Awareness on Adherence

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 Panel B of Figure 2, 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 lasso and scores 2 and higher. Figure A.1 displays estimation results of the θ_r coefficients and is broadly consistent with the raw treatment rates shown in Figure 2 Panel B. 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.[10]

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.2, we further show that, while adherence varied substantively across physicians, few physicians reached an adherence rate of 80%. Appendix Table A.2 shows that cardiologists were 4.6 percentage points more likely to adhere to the guideline than primary care or internal medicine physicians; adherence was also slightly higher for younger doctors, perhaps reflecting their more recent training.

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 A.1 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 $CHADS_2$ mentions coinciding with the publication of the ACC guidelines, as shown in Figure 2 Panel A, reducing concern about both false-positive and false-negative detection. An alternative approach to assessing the possible effects of guideline awareness is to study long-differences in prescription rates for high- vs. low-scoring patients from 2005 (1 year before publication of the first $CHADS_2$-based guideline) to 2013 (7 years after publication). Over this period, our measure of $CHADS_2$ awareness increased by 67 percentage points, while prescription rates fell by 9 percentage points for low-score patients, while rising by 3 percentage points for high-score patients (see Appendix Table **??**). If we attribute all over-time changes in anticoagulation prescription by $CHADS_2$ score to guideline awareness, the long difference results suggest larger responsiveness to the guideline than our more controlled estimates; yet even in 2013, doctors continue to deviate from the guideline recommendation for half of their patients.

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

3.3 Understanding Deviations from Guideline-Based Decisions

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). Our results further show that there is low adherence even among physicians who discuss the $CHADS_2$ score in their decision-making. 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 hesitation to prescribe anticoagulation (and thus failure to adhere to guidelines) 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).[11]

Prior 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). Another possible explanation for low guideline adherence is that physicians correctly predict that some patients will be unable to maintain their drug regimen, either by failing to refill their prescription or by failing to receive regular bloodwork and make dosage adjustments necessary to maintain the optimal level of anticoagulation. To test these hypotheses, we used patient characteristics in the VHA to predict two different measures of adherence among patients who filled at least one warfarin prescription: (1) prescription days supply (which measures the proportion of days covered with an anticoagulant prescription), and (2) time in therapeutic range (which measures what proportion the patient's lab values on their INR tests show dosing remained in the therapeutic range). Appendix Table **??** demonstrates that both measures can be predicted with patient characteristics, including adherence to other chronic medications and other measures of care utilization (see Appendix Table A.8 for a summary of predictor variables). In Appendix Figure A.3, we show that physicians are more likely to adhere to the $CHADS_2$-based guideline for patients who are predicted to have more consistent prescription refilling behavior, suggesting some reluctance to prescribe warfarin may be driven by patients' poor predicted adherence with the regimen. However, even among patients in the highest decile of predicted refill

[11] Specifically, physicians' estimates of rates of warfarin-associated intracerebral hemorrhage 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.

consistency, physicians adhere to the $CHADS_2$-based guideline only 60% of the time. Further, physicians are less likely to adhere to $CHADS_2$-based guideline for patients with higher predicted time in therapeutic range (i.e. patients who are more likely to maintain appropriate dosage).

4 The Effects of Anticoagulation

4.1 Setup and Design

To evaluate treatment decisions, including adhering or departing 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 use the AFI database of pooled RCTs to estimate CATEs. 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 describe later.

4.2 Estimation Procedure

Treatment effects are estimated in the pooled AFI database containing patient-level data from eight RCTs. We apply an ensemble machine learning approach that combines causal forest and lasso to estimate CATEs with conditional random assignment. The lasso method is adapted from Belloni et al. (2014) and the causal forest methods from Wager and Athey (2017). The ensemble approach combines both machine approaches to improve predictive validity (Younas et al., 2022), in this case, by leveraging the advantages of the causal forest for discovering predictive interactions among discrete variables and advantages of the lasso for using functional form restrictions to fit predictions that smoothly vary with the continuous random variables. 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 machine learning model considers every variable underlying the $CHADS_2$ score, as well many additional risk factors not currently encoded in clinical guidelines; the complete variable list is provided in Appendix Table A.3. We also use machine learning predictions of stroke and bleed risk among control-group patients as inputs, allowing treatment effects to depend on risk directly.

Our procedure begins by applying lasso methods to estimate CATEs that vary with continuous patient characteristics, including age, height, weight, blood pressure, and blood hemoglobin levels. Taking the residual variation in stroke and bleed outcomes not predicted by the lasso model, we then apply causal-forest methods to estimate CATEs that vary with discrete patient characteristics and their interactions. Our final CATE estimates sum the predictions from the lasso and causal-forest models. Throughout this estimation process, we account for differences across trials in patient risk and treatment probability. Trial-level adjustment is performed with fixed effects in the lasso model and with a re-centering procedure in the causal forest, as formally justified in Athey et al. (2019). We select parameters through cross-validation to minimize mean squared error. For further details on the CATE estimation process, see Appendix A.1

4.3 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, adapting the approach developed by Chernozhukov et al. (2018)

and applied in Athey and Wager (2019). Specifically, we test whether out-of-bag CATE predictions estimated in seven of the eight trials can predict stroke and bleed treatment effects in the eighth hold-out trial. To implement this approach, we construct out-of-bag CATE predictions for each trial in the AFI database, exclusively using 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). We also construct and control for risk predictions of $Y^o(x)$, based on other trials: $\hat{Y}^o_{-j}(x)$, using only control group data.[12]

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

$$\begin{aligned} Y^o_i &= \left[\delta^o_0 + \delta^o_1 \left(\hat{\tau}^o_{-j}(X_i) - \overline{\tau}^o\right)\right] W_i + \\ &\quad \varphi^o_1 \left(\hat{\tau}^o_{-j}(X_i) - \overline{\tau}^o\right) P_j + \varphi^o_2 \hat{Y}^o_{-j}(X_i) + \zeta^o_j + \varepsilon_i, \end{aligned} \tag{4}$$

where $\overline{\tau}^o \equiv \sum_i \hat{\tau}^o_{-j}(X_i)$ is the mean CATE prediction.[13] The coefficient δ^o_1 quantifies the predictive power of heterogeneous CATEs that are estimated in other trials. A coefficient value that is statistically indistinguishable from $\delta^o_1 = 0$ would suggest that the CATEs do not robustly predict variation in treatment effects in the hold-out trial; a coefficient value of $\delta^o_1 = 1$ would suggest that CATEs are well calibrated and that outcomes for a patient with characteristics x would increase one-for-one with treatment status, D_i, and the relevant CATE, $\hat{\tau}^o_{-j}(x)$. Intermediate values of $\delta^o_1 \in (0, 1)$ suggest the machine learning procedure may have over-fit the data. The re-scaled estimates $\delta^o_1 \hat{\tau}^o_{-j}(x)$ form the best linear predictors of the CATEs, as discussed in Chernozhukov et al. (2018).

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. This allows us to test whether we have recovered valid variation in stroke treatment effects beyond the dimensions recovered by the $CHADS_2$ score. Specifically, we first use a regression to project leave-out-trial stroke CATEs, $\hat{\tau}^s_{-j}(x)$, onto indicators

[12] We estimate two risk measures: one using a linear lasso model with all patient characteristics and one with the regression forest using only the discrete patient characteristics

[13] 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 $\hat{\tau}^o(x|j)$.

of the CHADS_2 score. We call the CHADS_2-projected component $\hat{\tau}^c_{-j}(x)$ and the residual component $\hat{\tau}^r_{-j}(x)$, noting that $\hat{\tau}^s_{-j}(x) = \hat{\tau}^c_{-j}(x) + \hat{\tau}^r_{-j}(x)$. We use these components to perform the following BLP projection:

$$\begin{aligned} Y_i^s &= \delta_0^s D_i + \sum_{\tilde{o} \in \{s(c), s(r)\}} \left[\delta_1^{\tilde{o}} \left(\hat{\tau}^{\tilde{o}}_{-j}(X_i) - \overline{\tau}^{\tilde{o}} \right) W_i + \varphi_1^{\tilde{o}} \left(\hat{\tau}^{\tilde{o}}_{-j}(X_i) - \overline{\tau}^{\tilde{o}} \right) P_j \right] + \\ & \varphi_2 \hat{Y}^s_{-j}(X_i) + \zeta_j^s + \varepsilon_i. \end{aligned} \tag{5}$$

Using Equation (5), we construct BLP-based adjustments to the stroke CATEs as follows:

$$\hat{\tau}^s_{BLP}(x) = \hat{\delta}_0^s + \hat{\delta}_1^c \hat{\tau}^c(x) + \hat{\delta}_1^r \hat{\tau}^r(x). \tag{6}$$

4.4 Validation Results

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.7. For stroke CATEs, using Equation (4), we estimate $\hat{\delta}_1^s = 0.520$ (standard error of 0.130), which is statistically distinguishable from 0 at the 1% level. This coefficient suggests that the CATEs are strongly predictive of treatment effects in out-of-bag trials, but also since the coefficient is smaller than 1, the raw CATEs may be over-fit or have statistical noise. Using Equation (5), we estimate the CHADS_2 component as $\hat{\delta}_1^c = 0.666$ (standard error of 0.250); for the residual component in this equation, we estimate $\hat{\delta}_1^r = 0.460$ (standard error of 0.142). In other words, both components significantly predict stroke treatment effect variation in held-out trials, with similar re-scaling for over-fitting. In our subsequent analyses, we will adjust CATEs predicted in Section 4.2 for stroke by using Equation (6) for $\hat{\tau}^s_{BLP}$.

We use the BLP framework to investigate one additional aspect of the stroke 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 observed by physicians but not reflected in the CATEs is of particular concern for our later counterfactual analyses. To test whether physicians observe and respond to additional treatment effect heterogeneity predicted by observable covariates, we estimate an index of treatment propensity in the observational VHA data and find that this treatment propensity, conditional on estimated CATEs, has no additional predictive power for stroke treatment effects.[14]

[14] Specifically, we test whether this treatment propensity index predicts treatment effects in the AFI data. Results are

In contrast to the stroke findings, the BLP regression suggests little reliably estimated variation in bleed CATEs across observable covariates; these results are reported in Appendix Table A.7, Column 4. The BLP coefficient is wrong-signed ($\hat{\delta}_1^b = -0.374$) and statistically insignificant, indicating that we cannot validate heterogeneity in bleed treatment effects. This failure to identify heterogeneity in bleed treatment effects may reveal either truly limited heterogeneity in bleed effects, or a lack of statistical power to detect bleed effects given that bleeds are a rarer outcome and the average bleed treatment effect is approximately half as large as the average stroke treatment effect.

In our subsequent analysis, we explicitly consider the scientific uncertainty about bleed treatment effects and model welfare under two different assumptions about bleed treatment effects: (1) a world where bleed effects are constant; (2) a world where bleed effects are proportional to bleed risk. When assuming bleed effects are proportional to bleed risk, we use with the ratio of average bleed treatment effect (0.019) to the average bleed risk (0.020) in the pooled RCT data set to pin down the proportionality constant (0.95). Our approach to directly considering the scientific uncertainty reflects the conflicting approaches of current practice guidelines. The widely used $CHADS_2$ and CHA_2DS_2-VASc guidelines focus on allocating treatment to patients with large expected benefits from stroke risk reduction (as if bleeding effects were orthogonal to stroke risk and stroke treatment effects). A less popular risk score known as the HAS-BLED recommends avoiding anticoagulant medication if a patient has sufficiently high bleed risk (as if bleeding effects were proportional to risk). There is no single score that considers these two effects jointly.

4.5 Stroke Treatment Effects and Bleed Risks in the VHA Data

In our main analysis, we take the BLP-adjusted stroke CATEs estimated in the RCTs and import these estimates to the observational VHA data, constructing a stroke treatment effect for each patient as a function of their observable characteristics. Figure 3, 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 -0.06, while the 90th percentile is -0.03. Patients with higher $CHADS_2$ scores have larger stroke reductions from anticoagulation implied by the CATEs, though there remains substantial variation in CATEs within each

reported in Appendix Table A.7, Column 3. The coefficient on the treatment propensity index is small at -0.046 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 exercise only relates to additional treatment effect heterogeneity predictable by variables covered in the AFI data.

$CHADS_2$ score.[15]

Because we cannot detect bleed effect heterogeneity from the RCT data, we focus on estimating bleed risk in the VHA data using a lasso regression. This lasso regression forces the inclusion of a variable to control for whether the patient takes an anticoagulant drug, which separates the average anticoagulant treatment effect from the bleed risk prediction. Figure 4 shows the relationship between bleed risk and stroke treatment effects among VHA patients. The heat map illustrates a downward-sloping relationship between bleed risk and stroke treatment effects; patients with higher bleed risks tend to have larger benefits in terms of stroke reduction. The $CHADS_2$ score also implies a potentially perverse relationship with bleed risk: As shown in Appendix Figure A.4, Panel B, Patients with higher $CHADS_2$ scores are at higher risk for bleeds. The welfare implications of treating patients by their $CHADS_2$ score (or by their stroke CATE) depend on the scientifically unknown relationship between bleeding risks (absent anticoagulation) and potentially heterogeneous anticoagulation-induced bleeding. If bleed treatment effects increase in proportion to bleed risks, then treating patients with greater benefits may also select patients with greater harms.

5 Assessing the Performance of Known Guidelines

We have established widespread physician awareness of the $CHADS_2$ score yet only modest changes in physician behavior. Even physicians aware of the guideline continued to deviate for approximately half of their treatment decisions. In this section, we will assess counterfactual outcomes under strict adherence to treatment rules following the $CHADS_2$ score and the later CHA_2DS_2-VASc variant.

For stroke outcomes, we use the ML-predicted stroke treatment effects in Section 4. These novel stroke treatment effects will allow us to assess reasonable but previously untested clinical assumptions that patients at higher risk for stroke will benefit more from anticoagulation. For bleed outcomes, our approach must account for scientific uncertainty in the joint distribution of stroke and bleed treatment effects. We do that by considering two plausible assumptions. First, we assume homogeneous bleed treatment effects or, equivalently, bleed treatment effects unrelated to stroke treatment effects. Second, we consider bleed treatment effects to be proportional to bleed risks. In Section 4, we show that bleed risks are negatively correlated with (negative) stroke treatment effects; thus, under this second assumption, the harms of treatment could be positively correlated with the benefits of treatment.

Given the lack of heterogeneous causal estimates of anticoagulation on strokes and bleeds in the

[15] See Appendix Figure A.4, Panel A, for the variation of CATEs within and between $CHADS_2$ scores. The R^2 from regressing BLP-adjusted stroke treatment effects, $\hat{\tau}^s_{BLP}$, on $CHADS_2$ score indicators is 0.49.

existing literature, it is unclear how strict adherence to the known guidelines for anticoagulation will perform relative to physician discretion. Well-known decision errors may cause actual physician decisions to perform worse than if they had strictly adhered to known guidelines. On the other hand, clinicians may perform better than the known guidelines if they accurately utilize more patient information to predict stroke treatment effects than those encoded in the simple risk scores. It is possible that the benefits of treatment may not be strictly proportional to stroke risk. Finally, physicians may be responding to the scientific uncertainty of heterogeneous bleed treatment effects. If the benefits of treatment are positively correlated with its harms, simply prioritizing patients by their benefits may lead to excess harms.

5.1 Framework and Benchmarks

To assess the outcomes of treatment rules, we will adopt a framework that characterizes treatment rules as combinations of an *ordering* of patients and a proportion of patients treated. In the case of the known $CHADS_2$- and CHA_2DS_2-VASc-based guidelines, patients can be ordered by their respective scores, and the proportion of patients treated under each guideline is determined by the threshold score at or above which treatment is recommended.

A given ordering can achieve a continuum of stroke and bleed outcomes, depending on the proportion of patients recommended for treatment. A treatment rule can dominate another if it achieves fewer strokes for (weakly) fewer bleeds. Within a given ranking of patients, treating more patients always (weakly) trades fewer strokes for more bleeds. Therefore, between two treatment rules formed from the same ordering, one can never achieve better outcomes both for strokes and for bleeds. In the limits where either no patient is treated or all patients are treated, all orderings converge to achieve the same outcomes.

Based on average treatment effects, treating *all* patients with newly diagnosed atrial fibrillation in the VHA would prevent 409 strokes (hereafter, "preventable strokes") and induce 474 bleed events (hereafter, "inducible bleeds") per 10,000 patients after one year. A random ordering of patients would result in possible stroke and bleed outcomes that lie along the straight line connecting the 0%-treated and 100%-treated scenarios. While preventing all the preventable strokes without inducing any bleeds would be ideal, this outcome would only be feasible if all the patients with non-zero effects on stroke reductions also have zero effects on bleed effects. Nonetheless, among feasible decision rules, desirable treatment rules would lie along a frontier that pulls towards the lower left corner of the graph, forming a set of non-dominated treatment rules. The set of outcomes under random treatment

rules forms an important benchmark in our evaluation of treatment rules. The second benchmark is the set of stroke and bleed outcomes under observed treatment decisions in our dataset, aggregated across physicians.[16]

Figure 6 and Table 4 show stroke and bleed outcomes under these two benchmarks. Status-quo physician decisions perform similarly to the random-treatment benchmark. Physicians prescribe anticoagulation to 49.8% of patients and prevent 49.4% of preventable strokes. By construction, randomly treating 49.8% of patients would prevent 49.8% of preventable strokes and induce 49.8% of inducible bleeds. Assuming bleed effects are proportional to risk, the status quo allocation shows a very slight targeting of anticoagulation to patients with lower bleed risk, inducing only 48.3% of inducible bleeds.

The fact that observed anticoagulation decisions are about as effective as random may reveal the difficult tradeoffs physicians face, given that the patients who stand to benefit the most from treatment may also face the greatest risk of harm. We will explore this in further detail below. An additional explanation could be that physicians possess knowledge of additional predictors of stroke treatment effects that are not captured by our ML estimates, which would make their decisions more effective than they appear in this evaluation. However, in Appendix XX, we show that outcomes under status-quo physician treatment implied by heterogeneous stroke reductions in coarser patient bins are similar to those under richer bins. This evidence suggests that private information on stroke treatment effects is unlikely to explain deviations from following the $CHADS_2$ score.

5.2 Outcomes Under Strict Adherence to Known Guidelines

We now consider outcomes resulting from strict adherence to known guidelines. 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,[17] two guideline orderings can be welfare-ranked if one guideline prevents more strokes than the other guideline, for any fixed number

[16] As we will discuss below, the outcomes in this second benchmark may not correspond to any physician's treatment rule, as they are aggregated across physicians and will thus depend on practice variation. For example, if half of the physicians treated none of their patients and the other half treated all of their patients, the aggregated treatment decisions would involve treating some patients but not others.

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

of induced bleeds.

In Figure 6 Panel A, we see that under the assumption of homogeneous (or orthogonal) bleed treatment effects both the $CHADS_2$ guideline and its successor, the CHA_2DS_2-VASc, perform substantially better than a random allocation, by preventing additional strokes for any fixed quantity of induced bleeds. There is little difference in performance between these two guidelines. Turning to Figure 6 Panel B, we consider guideline allocations under the assumption that bleed effects are proportional to bleed risk. The pattern now reverses: Allocating treatment according to the $CHADS_2$ or CHA_2DS_2-VASc guidelines performs *worse* than random allocations at minimizing strokes per induced bleed. This suggests that while the guidelines do indeed prioritize patients with greater benefits of treatment (i.e. larger reductions in stroke risk) they also systematically prioritizes patients with very high bleed risk. Thus, if bleed effects are proportional to bleed risk, the guideline perversely targets patients for treatment with *smaller* magnitude benefit-to-cost ratios, i.e. patients with fewer strokes prevented per bleed induced.

This point is further demonstrated in Table 4. Allocating treatment to 49.8% of patients (matching the status quo treatment rate) with the highest $CHADS_2$ scores would prevent 56.3% of preventable strokes, compared to 49.5% of preventable strokes under status-quo treatment decisions. Assuming homogeneous bleed effects, treating 49.8% of patients according to any rule would only induce 49.8% of the inducible bleeds. However, if we instead assume bleed effects are proportional to bleed risk, these stroke-reducing benefits must be weighed against a dramatic increase in induced bleeds: Following the $CHADS_2$ would induce 59.7% of inducible bleeds, compared to 48.3% of inducible bleeds under the status quo.

6 Characterizing Optimal Guidelines

6.1 Optimal Guidelines Assuming Bleed Treatment Effects

If the decision-maker were fully informed of both stroke and bleed treatment effects (i.e., there exists no scientific uncertainty about bleed treatment effects), then the optimal decision rule maximizes the number of strokes prevented for a given number of induced bleeds. We could trace out the optimal decision frontier by varying the number of induced bleeds. The optimal rule prioritizes patients for treatment depending on the ratio τ_i^s/τ_i^b, where patients with the most favorable benefit-to-cost ratio are treated first. If bleed treatment effects are homogeneous (or independent of stroke treatment effects), the optimal decision rule would prioritize patients with the largest stroke treatment effects—

much like the $CHADS_2$ score does. The relative utility cost of a bleed vs. a stroke β defines a tangent to this frontier and could be used to select the optimal point.[18] Our analysis focuses on the orderings of patients and asks whether an ordering can generate greater reductions in strokes without increasing induced bleeds. An ordering that (weakly) prevents more strokes for any given level of induced bleeds (weakly) improves welfare regardless of β.

In Table 4, we report the outcomes following the optimal treatment rules, first assuming homogeneous bleed effects and then assuming bleed effects proportional to risk. If bleed effects are homogeneous, the optimal guideline orders patients by stroke treatment effects alone. This rule prevents 45 additional strokes per 10,000 patients relative to status quo decisions (a 22% improvement in strokes prevented), holding constant the number of treated patients. But if bleed treatment effects are proportional to bleed risk, this would also increase the number of induced bleeds by 49 per 10,000 (a 21% increase). Next, we consider the optimal treatment rule assuming bleed effects are proportional to bleed risk, which orders patients by the ratio of stroke and bleed treatment effects τ_i^s/τ_i^b. This rule would prevent only 7 additional strokes per 10,000 patients relative to status quo decisions (a 3% improvement in strokes prevented), holding constant the number of treated patients. But it would reduce the number of induced bleeds by 68 per 10,000 (a 30% improvement). Thus, a guideline constructed assuming homogeneous bleed treatment effects risks inducing large numbers of bleeds if bleed effects are truly proportional to risk. Conversely, a guideline constructed by assuming bleed treatment effects are proportional to risk may fall short of generating achievable reductions in strokes if bleed effects are homogeneous.

These results show that, in the case of anticoagulation, optimal guidelines constructed under incorrect assumptions about the joint distribution of harms and benefits may not perform much better than random treatment. The reason for this phenomenon is that the optimal treatment rules under the two assumptions about bleed treatment effects often disagree. As shown in Figure 7, patients for whom the two optimal treatment rules disagree are those with larger (smaller) reductions in strokes but also larger (smaller) bleed risks. A greater correlation between stroke reductions and bleed risks (i.e., a more negative correlation between (negative) stroke treatment effects and bleed risks) leads to greater gaps in ordering patients between the two optimal guidelines. As shown in Appendix Figure A.6, the negative correlation between stroke treatment effects and bleed risks results in a greater share of patients falling in the disagreement zone between the two possible optimal guidelines,

[18]There also exist possible corner solutions if the costs of stroke are sufficiently high relative to the costs of bleeds, such that it would be desirable to treat everyone (or conversely, if bleeds are much more costly than strokes, and it may be desirable to treat no one).

reducing the performance of either optimal guideline under the wrong assumption. Appendix Figure A.7 shows that it is even possible for mistakenly optimal guidelines to perform *worse* than random if the correlation is sufficiently large.[19]

6.2 Uncertainty-Robust Maxmin Guidelines

Given the true underlying scientific uncertainty about bleed treatment effects, it may be helpful to embed these tradeoffs directly during guideline construction. We now formalize an approach to constructing robust guidelines that could perform well in both worlds—homogeneous vs. heterogeneous bleed effects—by optimizing the welfare in the world under which patients would be worse off for any given treatment rule.

Specifically, we consider two worlds: In the simpler world A, τ_i^b is homogeneous, or $\tau_i^b = \kappa_b$ for all i. In the more complicated world B, τ_i^b is heterogeneous, such that $E[\tau_i^b] = \kappa_b$. Relative to no treatment, the utilities of treating patient i in the two worlds are, respectively,

$$u_A(i) = -\tau_i^s - \beta\kappa_b; \tag{7}$$

$$u_B(i) = -\tau_i^s - \beta\tau_i^b. \tag{8}$$

We consider a treatment rule as a function D that maps each patient i indexed in the unit interval $[0,1]$ to an indicator for treatment status: $D(i) \in \{0,1\}$, and define welfare associated with the treatment rule in world $\omega \in \{A,B\}$ as

$$W_\omega(D) = \int_0^1 u_\omega(i)D(i)di. \tag{9}$$

As discussed above, the (separately) optimal treatment rules in worlds A and B are $D_A^*(i) = \mathbf{1}(u_A(i) > 0)$ and $D_B^*(i) = \mathbf{1}(u_B(i) > 0)$, respectively.

The maxmin treatment rule maximizes the worse welfare between the two worlds:

$$D^* = \arg\max_D(\min(W_A(D), W_B(D))). \tag{10}$$

In Appendix XX, we show that maxmin treatment rules take the form of linear treatment rules in

[19] Appendix XX provides further details. This appendix analyzes this possibility with a joint bivariate normal distribution. As shown in Appendix Figure A.8, mistakenly optimal guidelines perform worse than random treatment if the correlation between stroke treatment effects and bleed risks is larger (in absolute value) than the ratio of the coefficient of variation in bleed risks and the coefficient of variation in stroke treatment effects.

$(\tau_i^b, \tau_i^s) \in [0,1] \times [-1,0]$ space passing through $(\kappa_b, -\beta\kappa_b)$ with slopes between $-\beta$ and 0. Specifically, maxmin treatment rules take the form $D^* = \mathbf{1}(\tau_i^s < -\beta(\kappa_b + \chi/(1+\chi)(\tau_i^b - \kappa_b)))$, where $\chi \in (0, \infty)$ is a constant representing the optimal $|u_A(i)/u_B(i)|$ for patients treated at the margin. χ depends on β and the joint distribution of (τ_i^b, τ_i^s). In turn, it determines the slope of the maxmin rule D^* in (τ_i^b, τ_i^s) space (i.e., $-\beta\chi/(1+\chi)$).

The set of possible maxmin treatment rules nest the optimal treatment rules under homogeneous and heterogeneous bleed effects.[20] Intuitively, the maxmin treatment rule becomes the optimal treatment rule under homogeneous bleed effects—targeting stroke treatment effects alone—if any heterogeneity in bleed treatment effects is independent of the stroke treatment effects.

Targeting stroke treatment effects remains the maxmin treatment rule when bleed treatment effects are negatively correlated in magnitude with stroke treatment effects (i.e., when patients benefiting the most from stroke reduction are also expected to be harmed the least from bleeds). Otherwise, targeting patients with higher stroke treatment effects may reduce welfare relative to random treatment. Further, under uncertainty about bleed treatment effect heterogeneity, adopting the optimal treatment rule that assumes world B may also not improve welfare compared to random treatment. We find a positive correlation between stroke treatment effects and potential bleed treatment effects (as proxied by bleed risk) in Figure 4. Given the growing complexity of managing patients with multiple comorbidities, this positive correlation between benefits and risks is likely to hold more generally. In many cases with such a positive correlation, the maxmin treatment rule balances recommendations between the two worlds, recommending treatment for a subset of patients who would be treated only under D_A^* or only under D_B^*. Appendix XX examines the relative performance of D_A^*, D_B^*, and D^* under alternative correlations between τ_i^b and τ_i^s.

In Appendix XX, we further explore the idea of *monotonicity* in maxmin treatment rules (i.e., as β decreases, patients will only be induced into treatment and not out of treatment).[21] When a treatment rule satisfies monotonicity, it may recommend a ranking of patients upon which doctors and policy-makers may choose an appropriate threshold, depending on β. We show that the joint distribution of (τ_i^b, τ_i^s) in our setting implies monotonicity in the maxmin rule and that a class of common distributions would similarly imply monotonicity.

[20] The optimal treatment maximizing W_A under homogeneous bleed effects is $D_A^*(i) = \mathbf{1}(u_A(i) > 0) = \mathbf{1}(\tau_i^s < \beta\kappa_b)$; the optimal treatment maximizing W_B under heterogeneous bleed effects is $D_B^*(i) = \mathbf{1}(u_B(i) > 0) = \mathbf{1}(\tau_i^s < \beta\tau_i^b)$. D^* approaches D_A^* when χ approaches 0, and it approaches D_B^* when χ approaches ∞.

[21] D_A^* and D_B^* clearly satisfy monotonicity.

6.3 Evaluating Maxmin Guidelines

In Figure 8, we show the achievable decision frontier under optimal guidelines. In Panel A, we plot scenarios assuming homogeneous bleed effects. The most convex curve is the decision rule that sorts patients solely by stroke treatment effects; this is the optimal frontier if we are certain that bleed effects are constant across patients. The maxmin ordering, which maximizes utility in the minimum utility world, performs nearly as well as the rule sorting patients based on stroke treatment effects only, and notably outperforms the $CHADS_2$ guideline. In Panel B, we plot scenarios assuming bleed treatment effects are proportional to risk. The maxmin-based decision rule still performs well, closing much of the difference between the random ordering and the best possible frontier (sorting patients by the ratio of stroke treatment effects to bleed risk).

In Table 4, we report that maxmin guideline prevents 40 additional strokes per 10,000 patients relative to status quo decisions (a 20% improvement in total strokes prevented), holding the number of treated patients constant at the status quo. It achieves this larger stroke reduction without any increase in induced bleeds (assuming constant bleed effects) or with only slight increases in induced bleeds (assuming bleed effects proportional to risk). If bleed effects are proportional to bleed risk, then the maxmin rule induces 7 additional bleeds per 10,000 patients relative to the status quo (a 3% increase). By reducing the number of treated patients along the maxmin frontier, the maxmin rule could strictly outperform the status quo by preventing more strokes without increasing the number of bleeds (as illustrated in Figure 8).

7 Wisdom of the Crowds

As a final analysis, we reevaluate physician decisions to assess the performance of an algorithmic ranking formed by the physicians' own decisions. To do this, we estimate a simple linear probability model of anticoagulation based on patient characteristics observed in the VHA data. We estimate this model using only within-physician variation by including physician fixed effects. As such, we interpret this model as a ranking of patients aggregating physicians' revealed ranking within each their patients based on patient characteristics. This ranking appeals to a "wisdom of the crowds" philosophy that studying the decisions of large groups of individuals may reveal better decision-making (Surowiecki, 2005; Gillen et al., 2017). Many artificial intelligence (AI) algorithms rely on a similar rationale to make recommendations based only on observed decisions without knowing causal effects or the ground truth.

In Figure 9, we examine the features of our crowd ranking relative to known and optimal guidelines. On the x-axis, we arrange patients by their percentiles in each ranking, ranging from most-preferred to least-preferred for treatment. In Panel A, we show the sequence of stroke treatment effects corresponding to each ranking, and in Panel B, we show the sequence of bleed risks. Regarding stroke treatment effects, the orderings according to $CHADS_2$ guideline, the maxmin solution, and stroke treatment effects (i.e., the optimal guideline under homogeneous bleed treatment effects) all show a strong preference for larger stroke treatment effects. In contrast, the orderings according to the crowd and the ratio between the stroke treatment effects and bleed risks (i.e., the optimal guideline under proportional bleed treatment effects) do not prefer treating patients with large stroke-reduction benefits. Regarding bleed risks, the orderings show a similar divide. Only the crowd and ratio orderings prefer to first treat patients with lower bleed risks, though this pattern is muted in the crowd ordering relative to the ratio ordering.

We then adopt the same framework as in Sections 5 and 6 to evaluate the performance of this "crowd" ranking. Specifically, we use the same ML-estimated stroke treatment effects to simulate strokes prevented along an ordering of patients. We also use the same bleed risks and consider the two assumptions about the relationship between bleed risks and bleed treatment effects: (i) bleed treatment effects are homogeneous (or orthogonal to stroke treatment effects), and (ii) bleed treatment effects are proportional to bleed risks. Note that outcomes in this analysis differ from the status-quo aggregate based on observed physician decisions we previously considered. In this analysis, we impose strict adherence to the crowd ranking; in the status quo, physicians often deviate from the crowd ranking and vary in their treatment thresholds.

Figure 10 shows the performance of the crowd ordering in terms of strokes prevented and bleeds induced depending on the assumption about bleed effects. While the crowd rule does about the same (or slightly worse) than random treatment under homogeneous (orthogonal) bleed effects, it performs substantively better than random treatment under heterogeneous (proportional) bleed effects. As also shown in Figure 6, ordering patients according to the known guidelines of the $CHADS_2$ or CHA_2DS_2-VASc scores produces disparate outcomes depending on the bleed-effect assumption. While they perform much better than random treatment under homogeneous bleed effects, they perform substantively worse than random treatment under heterogeneous bleed effects.

In Figure 11, we show the performance of all orderings in terms of bleeds induced at the point where 50% of preventable strokes are prevented. Evaluating policies by the assumption that yields more bleeds, the known guidelines of the $CHADS_2$ and CHA_2DS_2-VASc scores perform the worst

among all of the orderings, substantially worse than random treatment. Only the maxmin ordering achieves substantially fewer bleeds than random treatment under both assumptions—by construction, bleeds are equal between the two assumptions under this ordering. Optimal guidelines, by construction, would perform the best under the correct assumptions, but they perform only slightly better than random treatment under the wrong assumptions. Finally, the crowd ordering outperforms the known guidelines under the worse-case assumption (of homogeneous bleed effects). Consistent with the idea of crowd wisdom, crowd treatment rules estimated from cardiologists perform only slightly better than those estimated from primary care physicians; similarly, the crowd treatment rules from more-experienced physicians perform only slightly better than those from less-experienced ones.

8 Conclusion

We study the introduction of a prominent guideline to treat atrial fibrillation with anticoagulation: the $CHADS_2$ score. While the $CHADS_2$ guideline was particularly well-accepted among guidelines, in many respects, the score and its newer CHA_2DS_2-VASc variant are quite representative of clinical guidelines in general. They were developed by specialists focused on the more tractable goal of treating a single disease, rather than characterizing potentially heterogeneous harms for patients with other complex medical conditions. The RCT evidence base for the effects of anticoagulation, like many other treatments, generally skews towards patients more likely to benefit from it and less likely to be harmed by it.

Despite the vast majority of physicians rapidly acknowledging the $CHADS_2$ score in their notes, we document that these physicians do not adhere to the $CHADS_2$-score recommendations about 40% of the time. This finding concords with physician guideline adherence in many other settings. We harness a collection of RCTs studying anticoagulation for treating atrial fibrillation to analyze the implications of adherence, or lack thereof, to clinical decision rules. We apply machine learning (ML) methods to characterize and validate heterogeneous treatment effects of anticoagulation. We confirm that existing guidelines target patients with greater benefits of stroke prevention. However, while the RCT evidence base demonstrates clear stroke treatment effect heterogeneity, it does not permit us to validate any heterogeneity in bleed treatment effects. We further show that, across patients in our real-world sample, bleed risks are strongly correlated with greater stroke treatment effects—in other words, if bleed risks portend treatment harms, patients more likely to benefit from treatment could also be the ones more likely harmed by treatment.

Gaps in the evidence from RCTs leave scientific uncertainty about the joint distribution of the effects on benefits and harms of treatment. We approach this uncertainty by considering two plausible assumptions about the heterogeneity of bleed treatment effects: (i) bleed treatment effects are homogeneous, or at least orthogonal to stroke treatment effects; and (ii) bleed treatment effects are proportional to (untreated) bleed risks. We show potentially stark differences in the performance of optimal guidelines depending on these assumptions. Optimal guidelines constructed under the wrong assumption may perform worse than random treatment. This striking possibility is more than academic: We show that both known risk scores (the $CHADS_2$ and CHA_2DS_2-VASc scores) perform worse than random treatment if bleed treatment effects are proportional to bleed risks. We then characterize a maxmin treatment rule that is robust to this uncertainty by equalizing welfare between the two assumptions about the treatment harms. If physicians adhered to this robust treatment rule, they could achieve substantially better outcomes than the status quo.

In aggregate, physicians in the status quo achieve roughly no better outcomes than treating patients randomly. However, remarkably, we show that a treatment rule formed only from the within-physician likelihood of treatment performs substantially better than strict adherence to guidelines based on the known $CHADS_2$ and CHA_2DS_2-VASc risk scores. This finding suggests that physicians could be making tradeoffs—based on the possibility that treatment harms could be correlated with comorbid conditions—that could achieve better outcomes than following existing guidelines. Our finding that adhering to a "wisdom of the crowds" rule formed from physicians' own decisions could also perform better than the status quo suggests potential gains from reducing practice variation while also accounting for detailed patient heterogeneity.

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.

Agarwal, Nikhil, Alex Moehring, Pranav Rajpurkar, and Tobias Salz, "Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology," July 2023.

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.

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.

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.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen, "High-Dimensional Methods and Inference on Structural and Treatment Effects," *Journal of Economic Perspectives*, May 2014, *28* (2), 29–50.

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.

Centers for Medicare and Medicaid Services, "2022 Quality Rating System Measure Technical Specifications," 2021.

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," *The Quarterly Journal of Economics*, May 2022, *137* (2), 729–783.

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

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.

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.

Cuddy, Emily and Janet Currie, "Rules vs. Discretion: Treatment of Mental Illness in U.S. Adolescents," October 2020.

Currie, Janet M. and W. Bentley MacLeod, "Understanding Doctor Decision Making: The Case of Depression Treatment," *Econometrica*, 2020, *88* (3), 847–878.

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

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," *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.

Gilboa, Itzhak and Massimo Marinacci, "Amiguity and the Bayesian Paradigm," in Daron Acemoglu, Manuel Arellano, and Eddie Dekel, eds., *Advances in Economics and Econometrics: Theory and Applications, Tenth World Congress of the Econometric Society*, New York: Cambridge University Press, 2013.

Gillen, Benjamin J., Charles R. Plott, and Matthew Shum, "A Pari-Mutuel-Like Mechanism for Information Aggregation: A Field Test inside Intel," *Journal of Political Economy*, August 2017, *125* (4), 1075–1099.

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.

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.

Health Resources and Services Administration, "UDS Mapper."

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.

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, "Diagnosing Physician Error: A Machine Learning Approach to Low-Value Health Care," *The Quarterly Journal of Economics*, May 2022, *137* (2), 679–727.

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.

Pharmacy Quality Alliance, "PQA Adherence Measures," 2022.

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.

Prieto-Merino, David, Amy Mulick, Craig Armstrong, Helen Hoult, Scott Fawcett, Lina Eliasson, and Sarah Clifford, "Estimating proportion of days covered (PDC) using real-world online medicine suppliersâ datasets," *Journal of pharmaceutical policy and practice*, 2021, *14* (1), 1–14.

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.

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.

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.

Surowiecki, James, *The Wisdom of Crowds*, reprint edition ed., Westminster: Vintage, August 2005.

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

Tinetti, Mary E., Sidney T. Bogardus, and Joseph V. Agostini, "Potential pitfalls of disease-specific guidelines for patients with multiple conditions," *The New England Journal of Medicine*, December 2004, *351* (27), 2870–2874.

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.

U.S. Census Bureau, "Understanding and Using American Community Survey Data: What All Data Users Need to Know," 2020.

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.

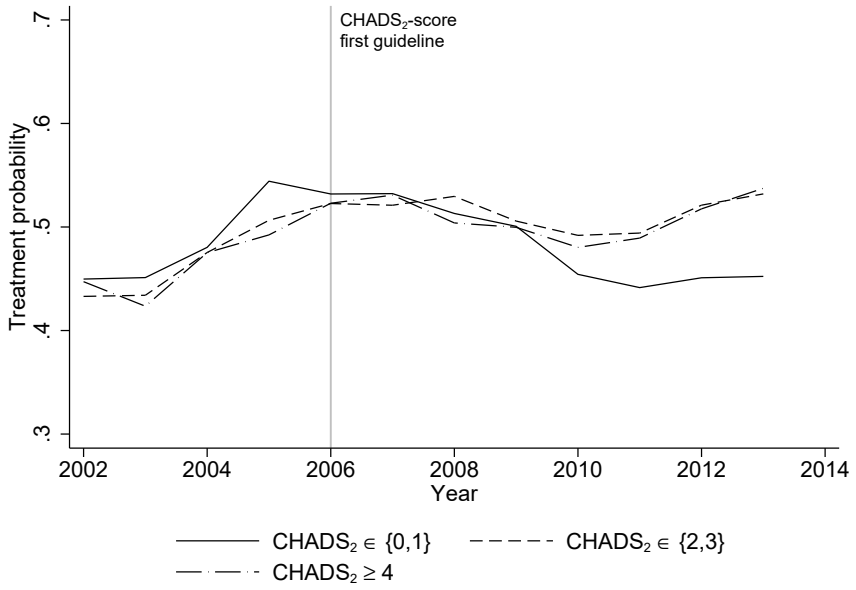
_ , Robert G. Hart, Stuart Connolly, Peter C. Austin, Jonathan Mant, F.D. Richard Hobbs, Peter J. Koudstaal, Palle Petersen, Francisco Perez-Gomez, J. Andre Knottnerus, Beppie Boode, Michael D. Ezekowitz, and Daniel E. Singer, "Effect of Age on Stroke Prevention Therapy in Patients with Atrial Fibrillation: 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.

Younas, Neelam, Amjad Ali, Hafsa Hina, Muhammad Hamraz, Zardad Khan, and Saeed Aldahmani, "Optimal causal decision trees ensemble for improved prediction and causal inference," *IEEE Access*, 2022, *10*, 13000–13011.

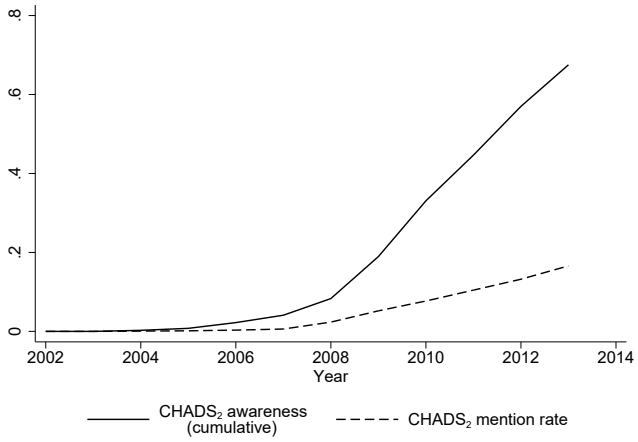
Figure 1: Treatment Decisions and the $CHADS_2$ Score



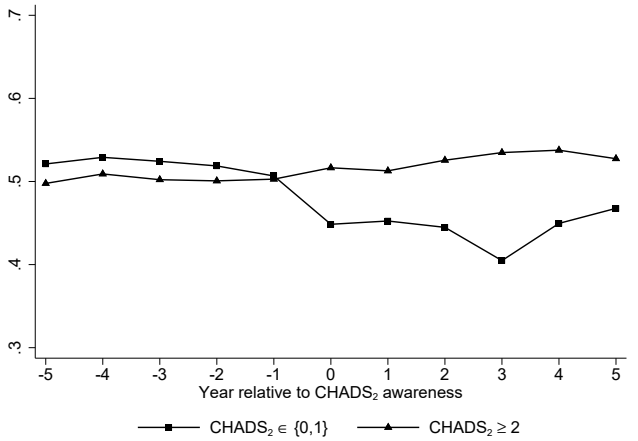
Notes: This figure shows physician treatment decisions relative to the $CHADS_2$ score. Treatments are defined as anticoagulation prescriptions within 90 days of initial diagnosis of atrial fibrillation. Panel A shows the fraction of atrial fibrillation patients treated with anticoagulation over time, for three groups of patients by $CHADS_2$ score. $CHADS_2$-score guideline unambiguously recommends treatment for patients with a score of at least 2.

Figure 2: $CHADS_2$-Score Awareness

A. Trends Relative to $CHADS_2$ Awareness



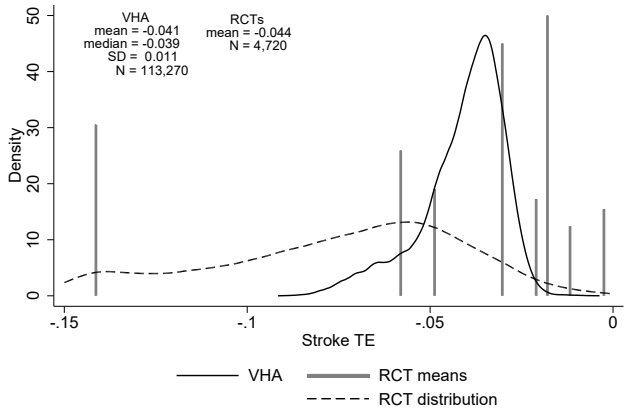
B. Trends Relative to $CHADS_2$ Awareness



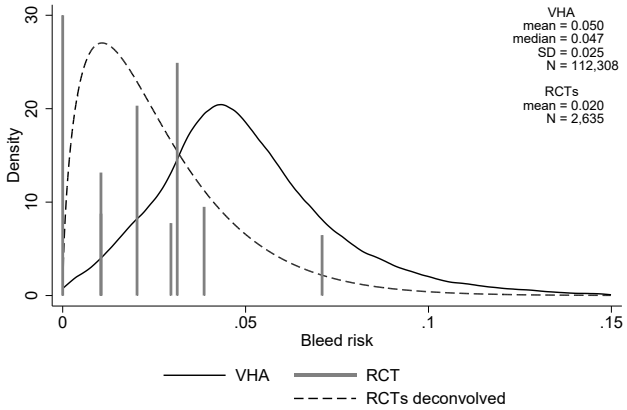
Notes: This figure shows awareness of the $CHADS_2$ score and its relationship with treatment decisions. Panel A shows the fraction of patients in a given year with physicians who either mention the $CHADS_2$ score in the note for the index patient or have mentioned the $CHADS_2$ score in either the 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 $CHADS_2$ 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 notes in a given year. Panel B displays the fraction of patients treated with anticoagulation in each year relative to the year of first mentioning the $CHADS_2$ score for physicians who eventually become aware the $CHADS_2$ score. Event-study results of the underlying data in this panel are shown in Appendix Figure A.1.

Figure 3: Distribution of Stroke Treatment Effects and Bleed Risks

A. Stroke Treatment Effects

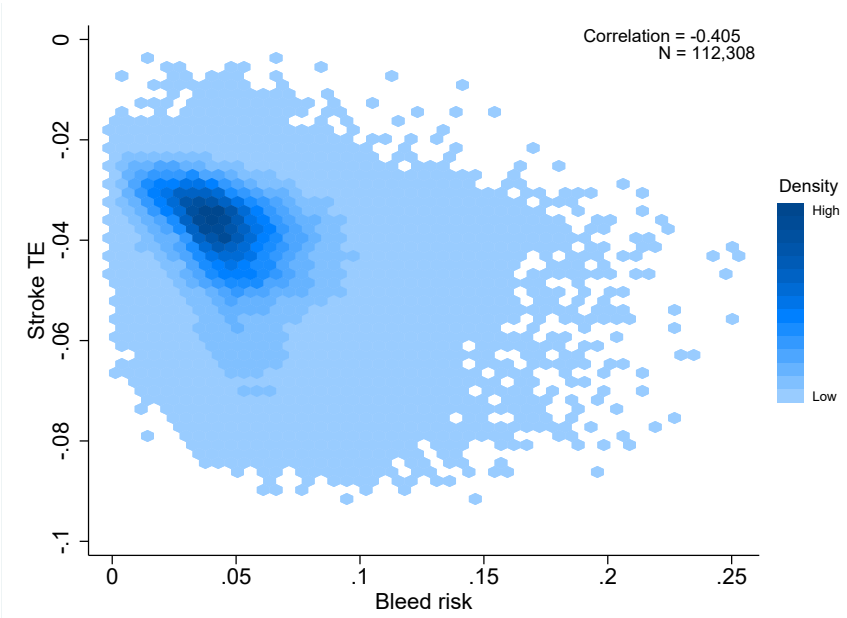


B. Bleed Risks



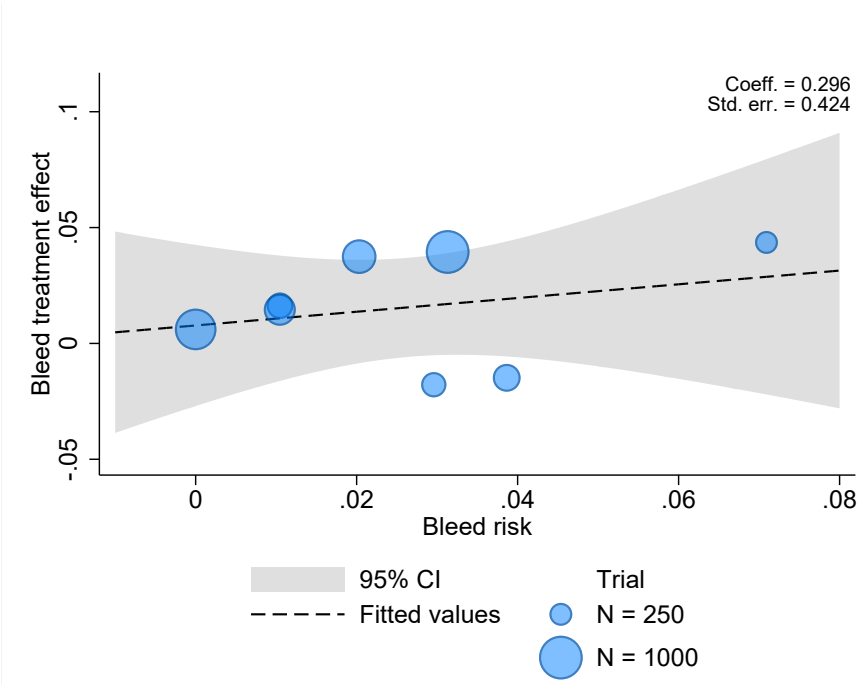
Notes: This figure shows the distribution of stroke treatment effects and bleed risks across patients in the VHA and RCT datasets. Panel A displays stroke treatment effects, and Panel B displays bleed risks. Stroke treatment effects are estimated from the RCT data as conditional average treatment effect (CATE) predictions are trained and validated by using causal-forest methods, described in Section 4. We use the causal-forest rules to calculate CATEs as a function of patient characteristics for each patient in the VHA data. Bleed risks are estimated using LASSO on patients in the VHA data with patient characteristics detailed in Appendix A.2. Because we cannot validate bleed treatment effects in the RCT data, we only display trial-specific average risks among patients in the control arms and the deconvolved distribution of trial-specific risks across trials. We show the distributions in the VHA data as kernel densities with solid black lines, the distributions in the RCT data as dashed lines, and trial-specific estimates as vertical gray lines.

Figure 4: Joint Distribution



Notes: This figure displays a heat map showing the relationship between stroke treatment effects and bleed risk in the VHA sample. Stroke treatment effects are estimated from the RCT data as conditional average treatment effect (CATE) predictions are trained and validated by using causal-forest methods, described in Section 4. We use the causal-forest rules to calculate CATEs as a function of patient characteristics for each patient in the VHA data. Bleed risks are estimated using LASSO on patients in the VHA data with patient characteristics detailed in Appendix A.2.

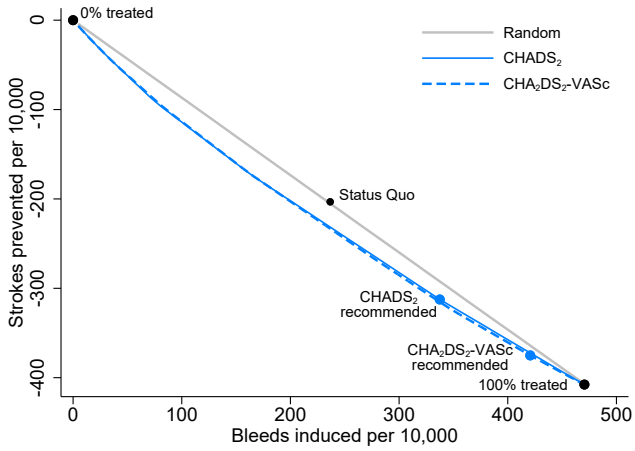
Figure 5: Trial Bleed Risks and Treatment Effects



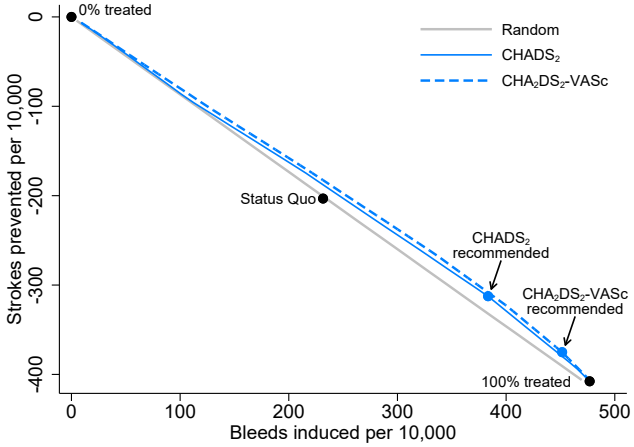
Notes: This figure shows the relationship between bleed risks and bleed treatment effects across trials in the RCT database. Bleed risks are calculated as the probability of bleeding in the control arm of each trial. Bleed treatment effects are calculated as the difference in the probability of bleeding in the treatment arm vs. the control arm of each trial. The line of best fit and the 95% confidence interval of the fit is shown as a dashed line and gray area, respectively.

Figure 6: Outcomes Following Known Guidelines

A. Homogeneous Bleed Effects

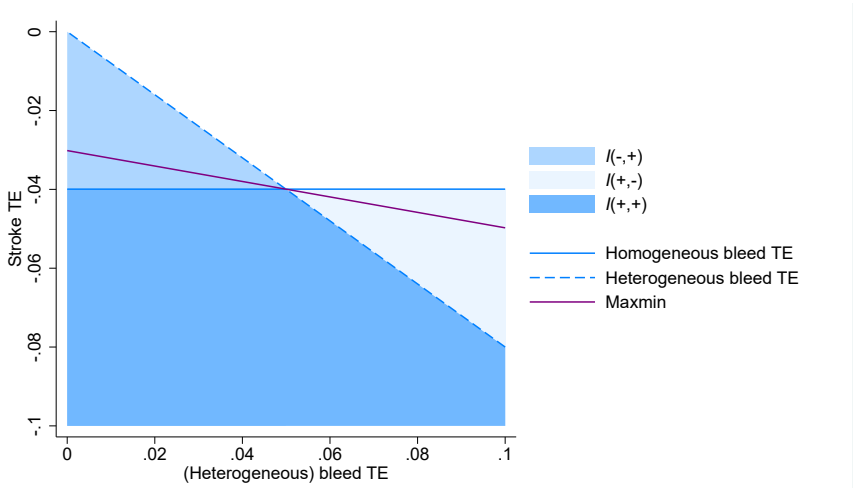


B. Heterogeneous Bleed Effects



Notes: This figure shows outcomes in terms of strokes prevented and bleeds induced by anticoagulation under strict adherence to two known guidelines: the $CHADS_2$ score and its successor, the CHA_2DS_2-VASc score. A guideline implies a rankings of patients, and outcomes depend on the ranking of patients and the share of patients treated under the guideline. Each line represents a ranking of patients corresponding to a guideline, extending from 0% to 100% shares of patients treated. 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. Panel A shows outcomes under the assumption of homogeneous bleed treatment effects or, equivalently, bleed treatment effects orthogonal to stroke treatment effects. Panel B shows outcomes under the assumption of heterogeneous bleed treatment effects proportional to bleed risks. For reference, outcomes under random treatment are shown as a straight gray line.

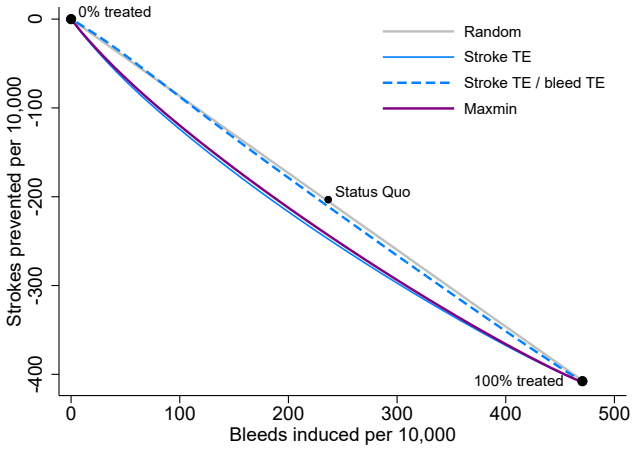
Figure 7: Optimal Treatment Rules in Treatment-Effect Space



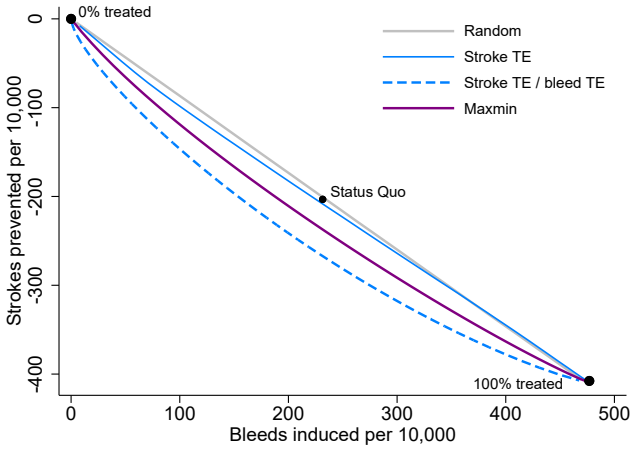
Notes: This figure represents optimal guidelines in the space of stroke treatment effects and potentially heterogeneous bleed treatment effects. Patients fall within this space via a joint distribution. Figure 4 illustrates the potential joint distribution (when bleed treatment effects are proportional to bleed risks) in the VHA data. The underlying joint distribution for this figure has an average bleed treatment effect of $\kappa_b = 0.05$. The treatment spaces and treatment rules correspond to the utility cost of a bleed being 80% of that of a stroke. $\mathcal{I}(+,+)$ is the space in which patients should be treated regardless of whether the bleed treatment effect is homogeneous (orthogonal to stroke treatment effects) or heterogeneous. $\mathcal{I}(+,-)$ and $\mathcal{I}(-,+)$ are the spaces including patients who should only be treated under homogeneous bleed effects and heterogeneous bleed effects, respectively. The solid blue and dashed blue lines indicate treatment rules for marginal patients treated under homogeneous and heterogeneous bleed treatment effects, respectively. The purple line represents a possible maxmin treatment threshold in which average bleed effect conditional on treatment equals the average (unconditional) bleed effect. Further details of the theory are discussed in Appendix XX.

Figure 8: Outcomes Following Optimal Guidelines

A. Homogeneous Bleed Effects

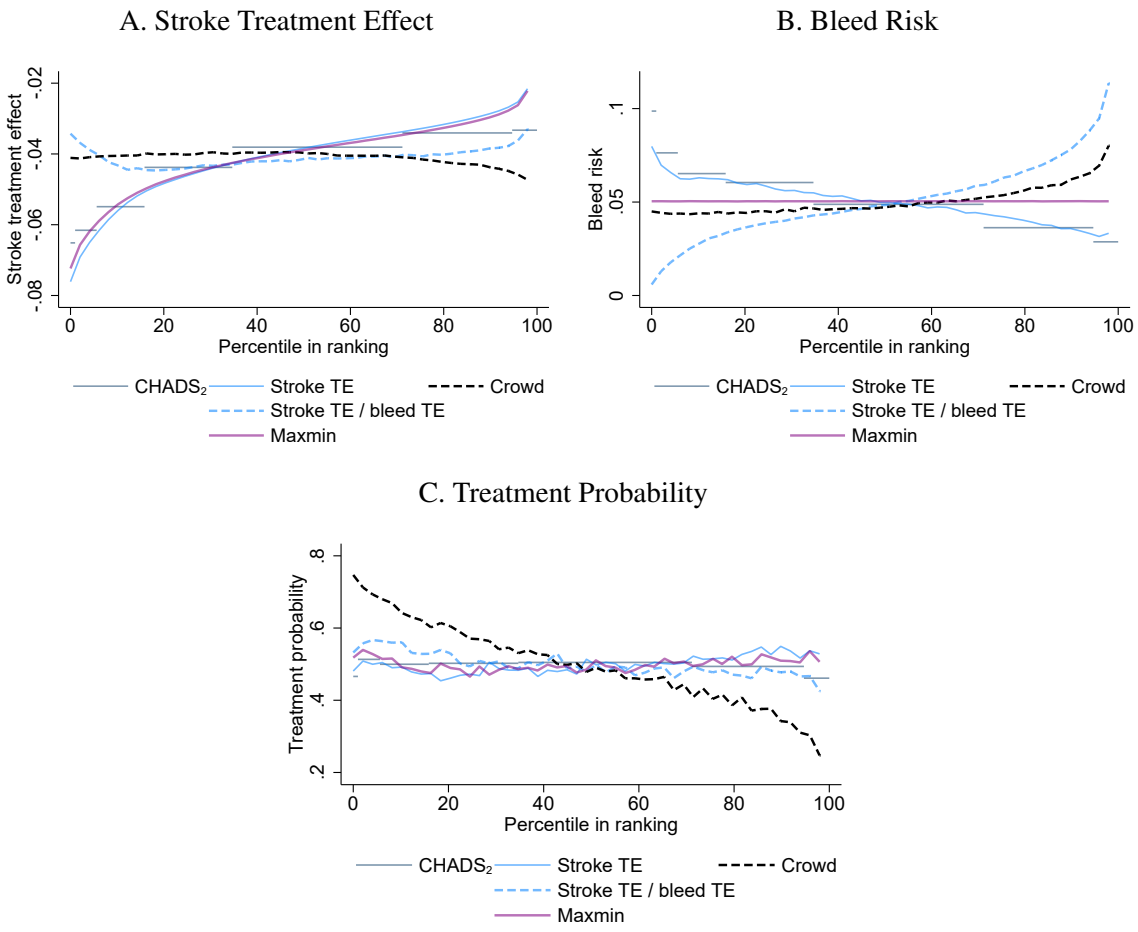


B. Heterogeneous Bleed Effects



Notes: This figure shows outcomes in terms of strokes prevented and bleeds induced by anticoagulation under different optimal treatment rules. Treatment rules imply rankings of patients, and outcomes depend on the ranking of patients and the share of patients treated under the treatment rule. Each line represents a ranking of patients corresponding to a treatment rule, extending from 0% to 100% shares of patients treated. 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. Panel A shows outcomes under the assumption of homogeneous bleed treatment effects or, equivalently, bleed treatment effects orthogonal to stroke treatment effects. Panel B shows outcomes under the assumption of heterogeneous bleed treatment effects proportional to bleed risks. For reference, outcomes under random treatment are shown as a straight gray line.

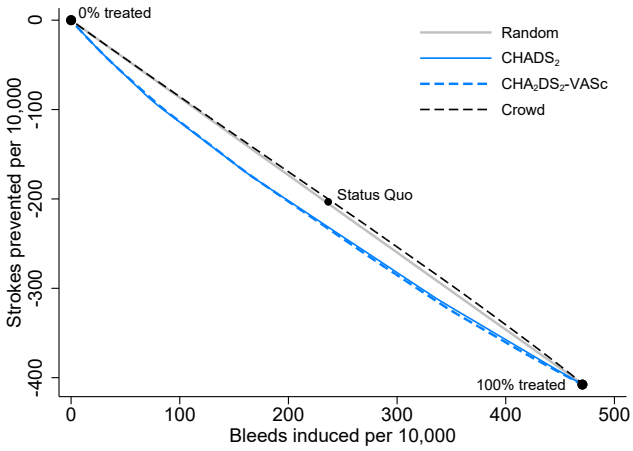
Figure 9: Patient Characteristics by Different Rankings



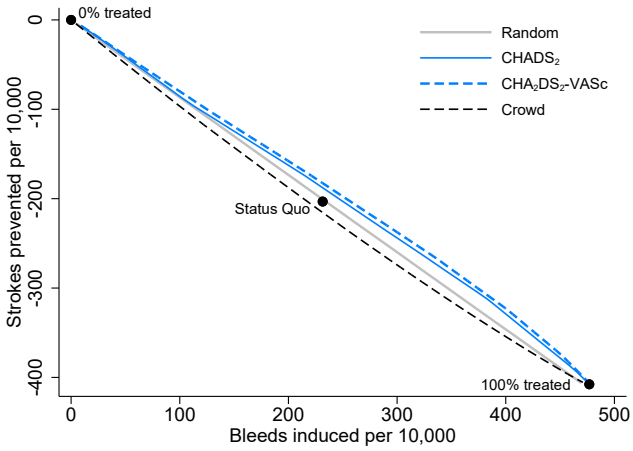
Notes: This figure shows patient characteristics and provider treatment probability of patients under several different rankings. The figure ranks patients by the $CHADS_2$ score, stroke treatment effects (the optimal homogeneous rule), the ratio of stroke and bleed treatment effects (the optimal heterogeneous rule), the maxmin rule, and a "wisdom of the crowds" rule. The crowd rule orders patients by their predicted probability of treatment from a within-physician regression of treatment on patient characteristics. Panel A represents patients by their stroke treatment effects, Panel B represents patients by their bleed risks, and Panel C represents patients by their treatment probability.

Figure 10: Wisdom of the Crowds

A. Homogeneous Bleed Effects

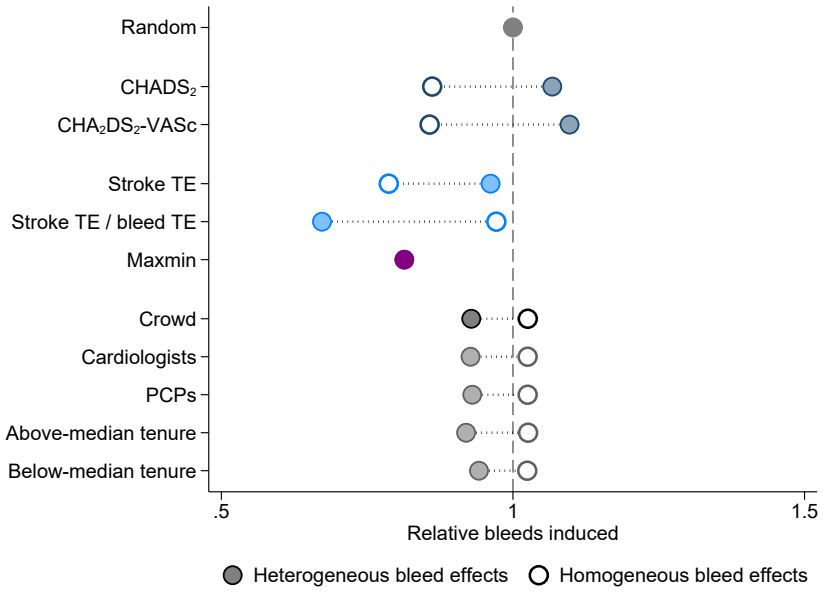


B. Heterogeneous Bleed Effects



Notes: This figure shows outcomes in terms of strokes prevented and bleeds induced by anticoagulation under strict adherence to two known guidelines (the $CHADS_2$ and CHA_2DS_2-VASc scores) and the "wisdom of the crowds" formed by predicting treatment in a within-physician regression. Each treatment rule implies a rankings of patients, and outcomes depend on the ranking of patients and the share of patients treated under the guideline. Each line represents a ranking of patients corresponding to a guideline, extending from 0% to 100% shares of patients treated. 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. Panel A shows outcomes under the assumption of homogeneous bleed treatment effects or, equivalently, bleed treatment effects orthogonal to stroke treatment effects. Panel B shows outcomes under the assumption of heterogeneous bleed treatment effects proportional to bleed risks. For reference, outcomes under random treatment are shown as a straight gray line.

Figure 11: Performance Relative to Random Treatment



Notes: This figure illustrates summarizes outcomes under random treatment, known guidelines, optimal guidelines, and treatment rules implied by "wisdom of the crowds". All outcomes hold fixed strokes prevented at the same level as treating 50% of patients randomly. Bleeds induced are shown relative to the bleeds induced under 50% random treatment. Bleeds induced under homogeneous bleed treatment effects are shown with hollow circles. Bleeds induced under heterogeneous bleed treatment effects proportional to bleed risks are shown with shaded circles.

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	−0.04	−0.04	−0.08	−0.03
Bleed risk	0.05	0.02	0.00	0.07
$CHADS_2$ components:				
Congestive heart failure	0.15	0.30	0.00	0.70
Hypertension	0.84	0.45	0.32	0.59
Age ≥ 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 the mean characteristics of patients in the VHA data and in the AFI database. Column 1 shows the characteristics of patients in the VHA data, specifically in the sample created by the steps described in Table 2. Column 2 shows the 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: Counterfactual Treatment Decisions and Outcomes

		Strokes prevented		Bleeds induced			
				Homogeneous effects		Heterogeneous effects	
	Percent of patients treated	Per 10K patients	Percent of maximum	Per 10K patients	Percent of maximum	Per 10K patients	Percent of maximum
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
A: Benchmarks							
Status quo	49.8%	202	49.4%	236	49.8%	229	48.3%
Randomly assigned treatment	49.8%	204	49.8%	236	49.8%	236	49.8%
All patients treated	100%	409	100%	474	100%	474	100%
B: Strict guideline adherence							
$CHADS_2$ guideline	49.8%	231	56.3%	236	49.8%	283	59.7%
Stroke TE guideline	49.8%	247	60.3%	236	49.8%	278	58.7%
Stroke TE / bleed TE guideline	49.8%	209	51.2%	236	49.8%	161	34.0%
Maxmin guideline	49.8%	242	59.2%	236	49.8%	238	50.2%

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 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

November 2, 2024

A.1 Estimating Heterogeneous Treatment Effects

We estimate conditional average treatment effects (CATEs) using an ensemble method that combines lasso (least absolute shrinkage and selection operator) and causal-forest predictions. This approach leverages advantages of the lasso for using functional form restrictions to fit predictions that smoothly vary with continuous random variables and advantages of the causal forest for discovering predictive interactions among discrete random variables. We combine the predictions from lasso and causal forest in what amounts to equally weighting the two predictions and validating the sum to match out-of-bag treatment effects.

Lasso Prediction. We leverage lasso to estimate predictions based continuous variables. Prior to running lasso, we apply a "within-transformation" to outcomes and each patient characteristic, which involves subtracting the mean of each variable within each trial from variable observations from that trial; this procedure is equivalent to including unpenalized dummy variables for each trial. Using control group observations and all patient characteristics listed in Appendix Table A.3, we run a lasso model to predict the outcome Y_i^o, for $o \in \{s, b\}$, and call this prediction the lasso-predicted risk score for outcome o.

We then proceed to predict treatment effects with lasso, using both treatment and control group observations, all patient characteristics listed in Appendix Table A.3 and the previously lasso-predicted risk score for outcome o. We denote the row vector of these characteristics (including the lasso-predicted risk score) $X_i^{o,\text{lasso}}$. We interact $X_i^{o,\text{lasso}}$ with the treatment indicator W_i. As before, we used outcome and patient characteristics to which we have applied the within-transformation described above to account for trial fixed effects. We additionally apply a within transformation to each interaction $x_{ik}W_i$ involving patient characteristic k and the treatment indicator.

We use lasso to select characteristics and treatment-interacted characteristics with non-zero coefficients in the following minimization problem:

$$\left(\hat{\alpha}^o_{\text{lasso}}, \hat{\tau}^o_{\text{lasso}}\right) = \arg\min_{\alpha,\varphi} \left\{ \sum_{i=1}^{N} \left(\tilde{Y}_i^o - \tilde{X}_i^{o,\text{lasso}}\alpha - \widetilde{XW}_i^{\,o,\text{lasso}}\tau \right)^2 + \lambda^o \left(\|\Psi_X \alpha\|_1 + \|\Psi_{XW}\tau\|_1 \right) \right\}. \quad \text{(A.1)}$$

$\tilde{Y}_i^o$ is the within-transformed outcome variable; $\tilde{X}_i^{o,\text{lasso}}$ is a row vector of within-transformed patient characteristics (including the lasso-predicted risk score); $\widetilde{XW}_i^{\,o,\text{lasso}}$ is a row vector of within-transformed interactions between W_i and elements of $X_i^{o,\text{lasso}}$. λ^o is the penalty parameter for the lasso model for outcome o; $\|\cdot\|_1$ is the L_1 norm; and Ψ_X and Ψ_{XW} are predictor-specific penalty "loading" vectors, where the elements are the standard deviations of elements $\tilde{X}_i$ and $\widetilde{XW}_i$, respectively, across observations.

We arrive at the optimal λ^o by ten-fold cross-validation. In each "fold," for a given λ^o, we estimate $\left(\alpha^o_{\text{lasso}}, \tau^o_{\text{lasso}}\right)$ by minimizing Equation (A.1) in a training sample comprising 90% of the observations. We compute the corresponding objective associated with λ^o in the validation sample, comprising the remaining 10% of the observations. We estimate this for ten folds of the data, averaging the objective

across the ten validation samples. We repeat this process to find the objective-minimizing λ^o.

Given the optimal λ^o, we take the non-zero elements of (α, τ) that minimize Equation (A.1) and rerun a post-lasso OLS regression:

$$\left(\hat{\alpha}^o_{\text{post-lasso}}, \hat{\tau}^o_{\text{post-lasso}}\right) = \arg\min_{\alpha,\tau} \left\{ \sum_{i=1}^{N} \left(\tilde{Y}^o_i - \tilde{X}^{o,\text{lasso}}_i \alpha \odot I_{\hat{\alpha}^o_{\text{lasso}}} - \widetilde{XW}_i^{\,o,\text{lasso}} \tau \odot I_{\hat{\tau}^o_{\text{lasso}}} \right)^2 \right\},$$

where $\odot$ indicates element-wise multiplication, and $I_{\alpha^o_{\text{lasso}}}$ and $I_{\tau^o_{\text{lasso}}}$ are vectors of indicators corresponding to whether the kth element of $\hat{\alpha}^o_{\text{lasso}}$ or $\hat{\tau}^o_{\text{lasso}}$, respectively, is non-zero or not. We set any element of $\hat{\alpha}^o_{\text{post-lasso}}$ or $\hat{\tau}^o_{\text{post-lasso}}$ to 0 if the corresponding element in $\hat{\alpha}^o_{\text{lasso}}$ or $\hat{\tau}^o_{\text{lasso}}$ is 0. As reported in Appendix Table A.6, variables corresponding to non-zero elements of $\hat{\alpha}^s_{\text{lasso}}$ are diabetes history and lasso-predicted stroke risk; variables corresponding to non-zero elements of $\hat{\tau}^s_{\text{lasso}}$ are stroke history, female gender, white race, and lasso-predicted stroke risk.

Causal-Forest Prediction. We use a causal forest to predict residual variation in risk and treatment effects, focusing on discrete variables. We residualize stroke and bleed outcomes by subtracting from observed outcomes the post-lasso predictions due to the only selected continuous variables in $\tilde{X}^{o,\text{lasso}}_i$ and $\widetilde{XW}_i^{\,o,\text{lasso}}$: the lasso-predicted risk score and its interaction with W_i. This prediction can be written as $\tilde{X}^{o,\text{lasso}}_i \hat{\alpha}^o_{\text{post-lasso}} \odot I_{\text{cont}} + \widetilde{XW}_i^{\,o,\text{lasso}} \hat{\tau}^o_{\text{post-lasso}} \odot I_{\text{cont}}$, where I_{cont} is a vector of indicators corresponding to whether a given variable is continuous. We further subtract trial-specific averages of these residual outcomes and the treatment variable.[1]

Using the ten discrete variables—gender, smoking status, history of angina, congestive heart failure, diabetes, hypertension, myocardial infarction, peripheral vascular disease, transient ischemic attack (TIA), and stroke—we fit a regression forest to predict risk of stroke and bleed. We then use this regression-forest prediction as an eleventh predictor in the causal forest. We denote the 1×11 row vector of characteristics for observation i used in the causal forest as $X^{o,\text{cf}}_i$. Stating the causal-forest CATE estimand from Equation (5) of Athey and Wager (2019), we have

$$\hat{\tau}^o_{\text{cf}}(\cdot) = \arg\min_{\tau} \left\{ \sum_{i=1}^{N} \left(\left(Y^o_i - \hat{m}^o_{(-i)} \left(X^{o,\text{cf}}_i \right) \right) - \tau \left(X^{o,\text{cf}}_i \right) \left(W_i - \hat{e}_{(-i)} \left(X^{o,\text{cf}}_i \right) \right) \right)^2 + \Lambda_n \left(\tau(\cdot) \right) \right\},$$

where $\Lambda_n(\tau(\cdot))$ is a regularizer that controls the complexity of the learned function $\tau(\cdot)$, $\hat{m}^o_{(-i)}(x) = \hat{E}_{(-i)} \left[Y^o_i \middle| X^{o,\text{cf}}_i = x \right]$ is the empirical expectation of Y^o_i conditional on $X^{o,\text{cf}}_i = x$ formed on observations leaving out i, and $\hat{e}_{(-i)}(x) = \widehat{\Pr}_{(-i)} \left(W_i \middle| X^{o,\text{cf}}_i = x \right)$ is the corresponding leave-out-i empirical treatment propensity conditional on $X^{o,\text{cf}}_i = x$.

We use the `grf` package developed by Tibshirani et al. (2020) to run causal and regression forests.

[1] Following Section 6.1.1 of Athey et al. (2019), we compute these averages by a leave-one-out procedure (i.e., not counting the index observation). Section 6.1.1 describes this procedure as "local centering" and subtracts an expectation that accounts for covariates that may differ across observations. We treat the trial identity as the only covariate by which outcomes and treatment are expected to covary; we ignore other covariates as they are randomly assigned between treatment and control within each trial.

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 many 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 to maximize differences in treatment effects—the difference between average outcomes among treated and untreated observations—between child nodes. Regression forests split nodes to maximize 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 set the following parameter values in our algorithm. We use cross-validation to select the minimum node size for causal and regression forests. 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 2000. 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.

Combining Predictions. We combine the treatment effects predicted from lasso and causal forest—i.e., $\hat{\tau}^o_{\text{lasso}}\left(X_i^{o,\text{lasso}}\right) = X_i^{o,\text{lasso}}\hat{\tau}^o_{\text{post-lasso}}$ and $\hat{\tau}^o_{\text{cf}}\left(X_i^{o,\text{cf}}\right)$, respectively. We sum the two predictions for a single observation to arrive at the treatment effect $\hat{\tau}^o\left(X_i\right) = \hat{\tau}^o_{\text{lasso}}\left(X_i^{o,\text{lasso}}\right) + \hat{\tau}^o_{\text{cf}}\left(X_i^{o,\text{cf}}\right)$. In order to validate this prediction, we first fit the prediction using out-of-bag observations from seven trials other than the index trial. We then form a best linear predictor (BLP) to estimate a shrinkage factor matching out-of-bag CATEs to realized outcomes. We discuss this procedure in Section 4.3. We show in that section that our bleed treatment effects are not validated across trials and therefore focus on stroke treatment effects for the remainder of our analysis.

Appendix Table A.6 summarizes the sources of stroke treatment effect heterogeneity. The table lists the selected predictors of treatment effect heterogeneity in our lasso model and the ranking of discrete characteristics by their "variable importance" in the causal-forest model.[2] 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 predicted stroke risk. The variables in the CHADS2 score generally rank highly in both the causal-forest and regression-

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

forest models, but several variables not in the CHADS2 score also matter: e.g., history of angina and smoking behavior also predict stroke treatment effects.

A.2 Predicting Bleed Risk

To predict bleed risk, we use data from the VHA, which includes a richer set of clinical variables associated with bleed risk than the AFI data and a larger sample size to improve power. We estimate a lasso model, where the outcome is a gastrointestinal hemorrhage or intracranial hemorrhage within one year. We use the following variables from the data in our lasso model:

1. **Frailty, Fall Risk, and Bleeding History.** We include indicators for neurologic disorders (including Parkinson's Disease), fall risk (neuropathy, muscle weakness, dizziness), vision problems, arthritis, head injury, fracture, and osteoporosis. Frailty and fall risk are frequently cited clinical explanations 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.** We include indicators for the Elixhauser comoribidity set, which include: HIV/AIDS, alcohol abuse, blood loss anemia, cardiac arrhythmia, congestive heart failure, chronic pulmonary disease, coagulapathy, deficiency anemia, depression, diabetes (with and without chronic complications), drug abuse, hypertension, hypothyroidism, liver disease, lymphoma, fluid and elctrolyte disorders, metastastic cancer, obesity, other neurologic disorders, paralysis, pulmonary circulation disorders, and psychoses, renal failure, solid tumor without metastasis, peptic ulcer disease, valvular disease, and weight loss.

3. **Medical History.** We include indicators for other important medical history, including: brain hemorrhage, gastrointestinal hemorrhage, ischemic stroke, pulmonary embolism, deep venous thrombosis, and myocardial infarction.

A.3 Predicting Patient Warfarin Adherence

In this appendix, we describe how we construct a measure of each patient's predicted adherence to a prescribed warfarin regime. Physicians may be reluctant to prescribe warfarin to patients who will not maintain consistency with medication dosing, since excessive anticoagulation increases the patient's vulnerability to bleeds, while insufficient anticoagulation may fail to deliver stroke-reducing benefits. To predict medication adherence, we use measures of patients' past health care utilization, medication adherence and socioeconomic characteristics; these are all factors physicians could plausibly consider when deciding on an anticoagulation treatment course. We then assess the predictive power of these metrics on various measures of warfarin medication adherence. These measures will be described in detail below, and are further summarized in Table A.8.

A.3.1 Past Health Care Utilization

In the VHA analytic cohort, we count the number of hospitalizations, emergency department, primary care, specialist and all outpatient visits in the year prior to Atrial Fibrillation diagnosis. Table A.8 displays the averages of these measures.

A.3.2 Zip-Code Income

We obtain zip code level median household income data from the 2013 5-year American Community Survey (ACS). This survey, conducted by the U.S. Census Bureau, provides estimates that draw on the previous 5 years of income data. The benefits of using 5-year estimates include availability and increased precision in areas with lower and higher population density, respectively (U.S. Census Bureau, 2020). The U.S. Census Bureau collects income information by ZIP Code Tabulation Areas (ZCTAs), which are general aerial representations of USPS zip codes. We map median ZCTA income to zip codes using the 2013 UDS Mapper ZCTA-zip code crosswalk (Health Resources and Services Administration, n.d.).

We determine VHA cohort zip codes by selecting the zip code reported closest to the end of day on the date of Atrial Fibrillation diagnosis. We then match these zip codes to the census income data via the UDS ZCTA-zip code crosswalk. We find a high matching rate, with over 99% of the cohort showing non-missing income data.

A.3.3 Medication Adherence

We measure medication adherence using the "proportion of days covered" (PDC), a metric endorsed and utilized by the Pharmacy Quality Alliance (PQA) and CMS Quality Ratings System (QRS) (Pharmacy Quality Alliance, 2022; Centers for Medicare and Medicaid Services, 2021). While methods to calculate PDC vary somewhat across studies, we adopt the conventional approach as described by Prieto-Merino et al. (2021). Non-warfarin PDC measures are calculated for the period between a patient's first prescription fill (after one-year prior to Atrial Fibrillation diagnosis) and the day before diagnosis. PDC of warfarin is calculated from the first fill after diagnosis, until one year after diagnosis. For each period of interest, we count the number of days a patient was covered by a fill. We calculate PDC as,

$$\text{PDC} = \min\left(1, \frac{\sum_{r\in\mathcal{R}(i)} \min(s(r), \bar{t} - t(r))}{\bar{t}(i) - \min\mathcal{R}(i)}\right) \tag{A.2}$$

where r is the prescription, $\mathcal{R}(i)$ is the set of prescriptions for patient i, $s(r)$ and $t(r)$ are the days supplied and fill date of prescription r, respectively, and $\bar{t}(i)$ is the date of diagnosis.

A.3.4 Predicting Medication Adherence

We test the predictive power of the pre-diagnosis measures on post-diagnosis anticoagulant prescription adherence using the following OLS regression:

$$Y_i = \alpha_0 + \alpha_1(\text{PDC}_i) + \alpha_2(\text{Util}_i) + \alpha_3(\text{Income}_i) + \alpha_4(\mathbf{X}) + \epsilon_i \tag{A.3}$$

In Equation A.3, Y_i represents various measures of anticoagulant medication adherence shown in Columns 1 - 4 of Table A.9. We assess the predictive power of all measures in Table A.8 on post-diagnosis medication adherence. In addition to these measure, we also include binary variables representing whether the patient recorded a fill in the year prior to diagnosis, denoted as $\mathbf{X}$ in Equation A.3.

Table A.9 shows the results of Equation A.3. We find strong evidence that adherence to prior chronic medications increases the probability that patients will refill their warfarin prescriptions and dose the medication appropriately to remain in therapeutic range. Patients with more intensive past health care utilization have lower adherence to warfarin prescription regime, perhaps due to the challenges of managing other co-occuring health conditions. Finally, patients residing in wealthier zip codes are less likely to maintain their prescription fills, but are more likely to maintain dosing with the therapeutic range.

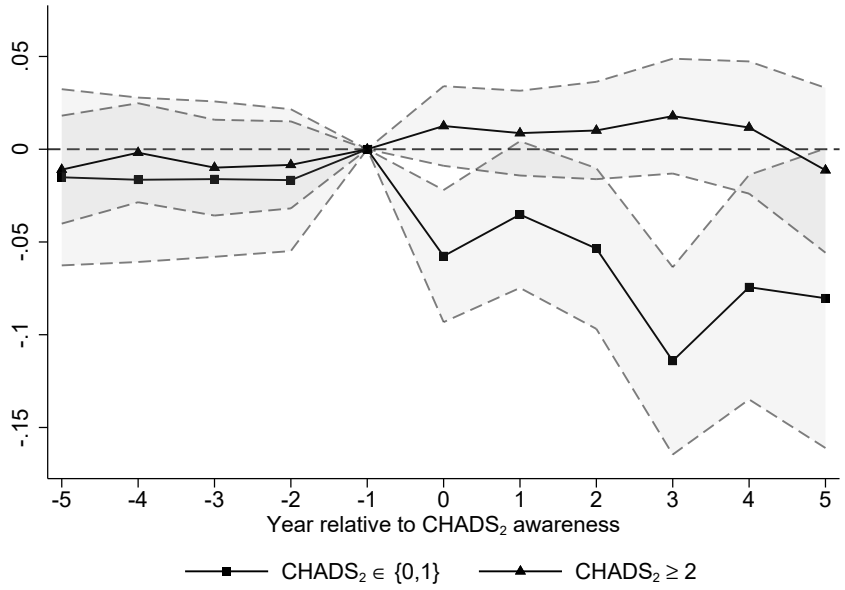
A.4 Construction and Prediction Details

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 $\Pr_{\text{VHA}}(D_i \mid X_i)$. 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.3. We then take this model to calculate $\Pr_{\text{VHA}}(D_i \mid X_i)$ 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 4.3 and Appendix Table A.7.

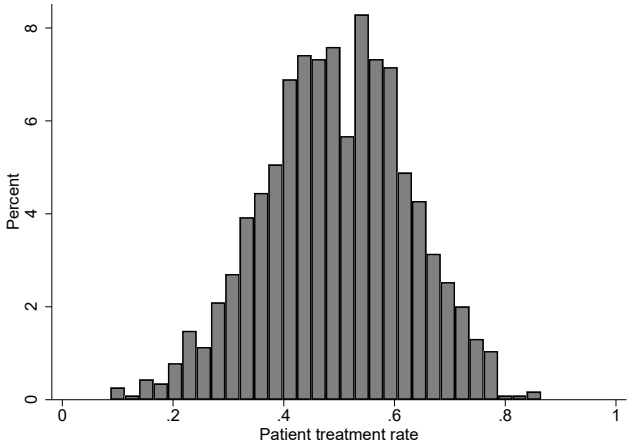
Figure A.1: Event-Study Estimates



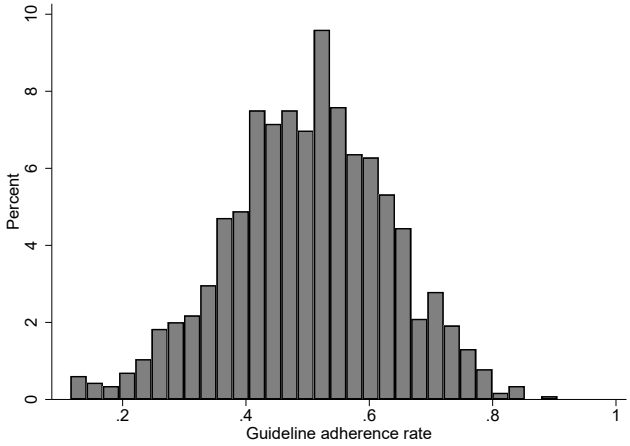
Notes: This figure shows regression coefficients and 95% confidence intervals from Equation (1), run separately for patients with $\text{CHADS}_2 \in \{0,1\}$ and for patients with $\text{CHADS}_2 \geq 2$. The 12-month period prior to the physician's first CHADS_2 mention is normalized to 0. The regression 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 A.2: Physician Treatment and $CHADS_2$ Adherence Rates

A. Treatment

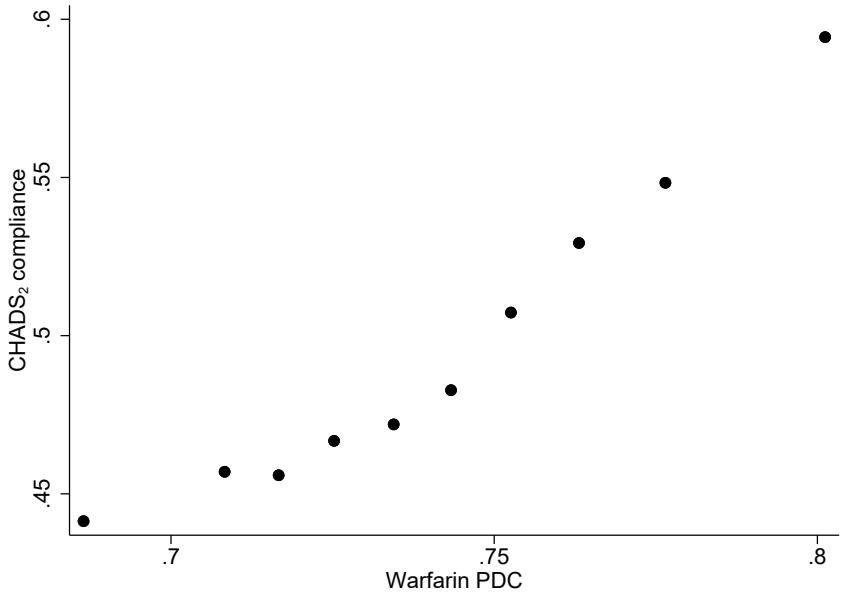


B. $CHADS_2$ Adherence



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.

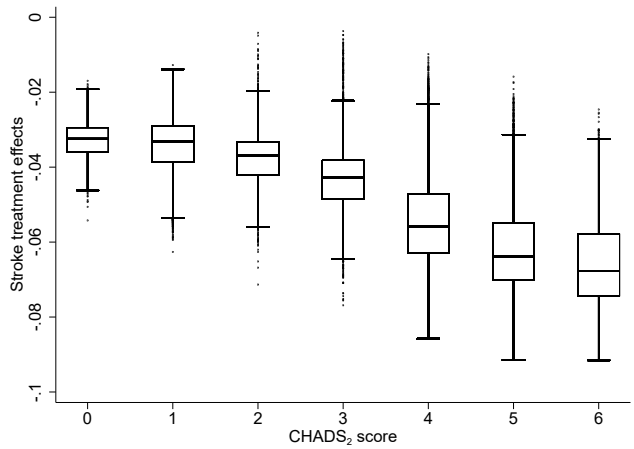
Figure A.3: $CHADS_2$ Prescribing Adherence by Predicted Patient Medication Adherence



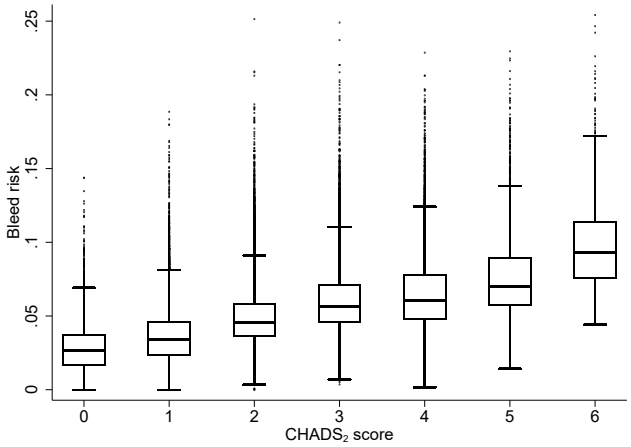
Notes: This figure shows the relationship between physician prescribing adherence to the $CHADS_2$-score guideline and a patient-level prediction of medication adherence as measured by the proportion of days covered (PDC) metric. Further details on this metric and the patient-level prediction can be found in Appendix A.3.

Figure A.4: Stroke Treatment Effects and Bleed Risks by $CHADS_2$ Score

A. Stroke Treatment Effect



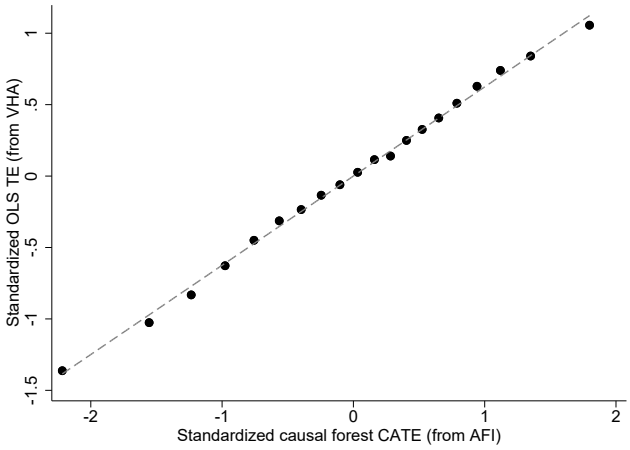
B. Bleed Risk



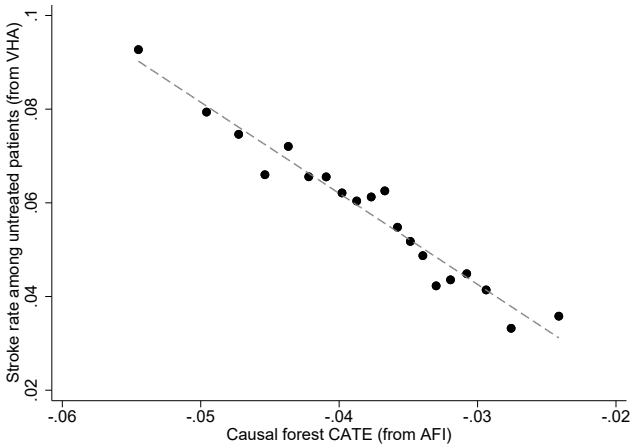
Notes: This figure shows box plots for the distribution of stroke treatment effects and bleed risk by $CHADS_2$ score across patients in the VHA data. Panel A shows stroke treatment effects; Panel B shows bleed risk. Bounds on the box plot are at the 25th and 75th percentile, with the median marked with a horizontal line. Whiskers extend to the 5th and 95th percentiles.

Figure A.5: Evidence on the External Validity of AFI CATEs

A. AFI CATEs and Observational Treatment Effects

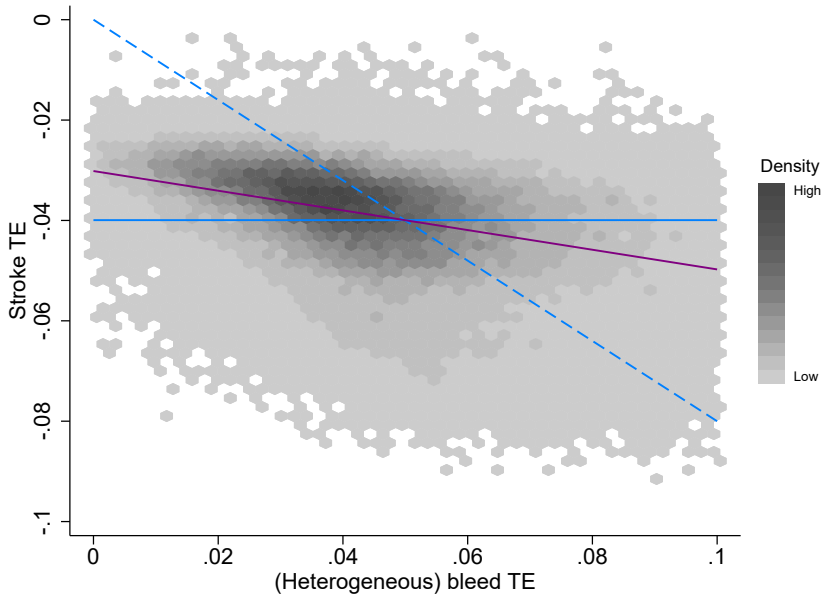


B. Stroke Outcomes Among Untreated Patients and AFI CATEs



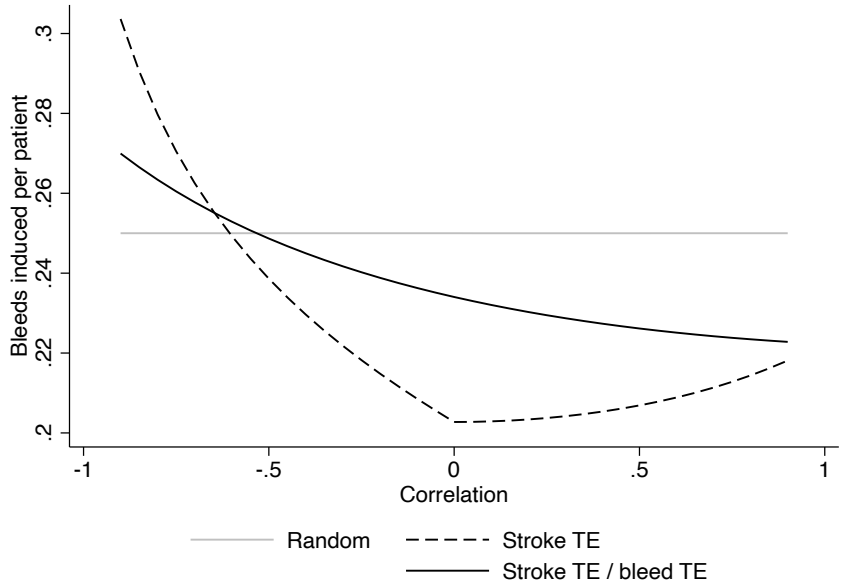
Notes: Panel A displays a binned scatterplot of the relationship between z-standardized stroke treatment effects estimated by a causal forest in the AFI RCT data and z-standardized stroke treatment effects estimated by OLS in the VHA observational data. The correlation coefficient is 0.62. OLS estimates interact with anticoagulation treatment with each of the patient characteristics that are covered in both the AFI and VHA data (see Appendix Table A.3 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 A.6: Optimal Treatment Rules for the VHA Sample



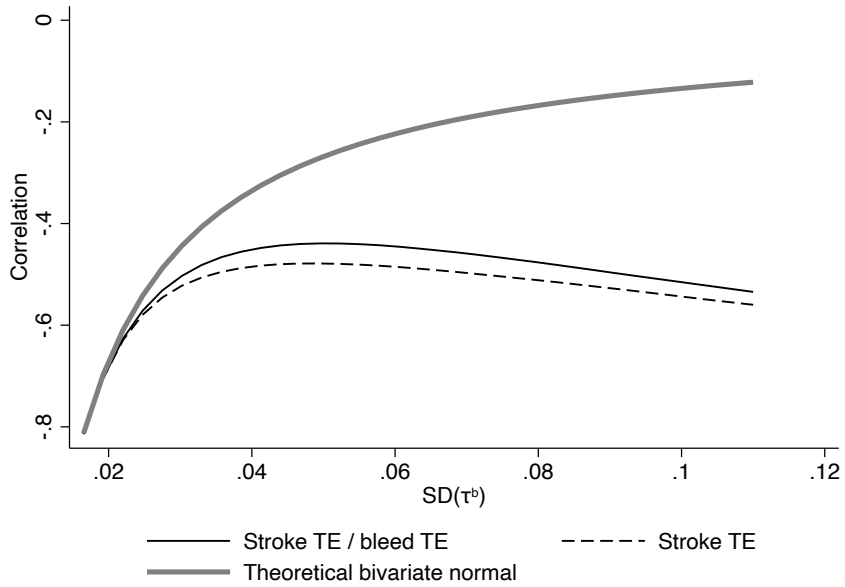
Notes: This figure represents optimal treatment rules based on stroke treatment effects and bleed risks in the VHA data. It builds on Figure 4, which illustrates the joint distribution of stroke treatment effects and bleed risks in the VHA data, and Figure 7, which illustrates optimal treatment rules in treatment-effect space. The underlying joint distribution when bleed treatment effects are heterogeneous and proportional to bleed risks is calculated by multiplying bleed risks by 0.95 (as the overall bleed treatment effect in the RCT data is approximately 0.95 times the overall bleed risk). As in Figure 7, this figure considers the utility cost of a bleed being 80% of that of a stroke. The optimal homogeneous rule orders patients by stroke effects and is shown in a solid blue line. The optimal heterogeneous rule orders patients by the ratio of stroke and bleed effects and is shown in a dashed blue line. The maxmin rule equalizes the average bleed effect conditional on treatment and the average unconditional bleed effect. See Figure 7 for further notes.

Figure A.7: Worse-Case Outcomes by Correlation



Notes: This figure illustrates worse-case outcomes under three optimal treatment rules depending on the correlation between stroke and bleed treatment effects. The joint distribution of stroke treatment effects and potentially heterogeneous bleed treatment effects is a truncated bivariate normal distribution, with stroke treatment effects ranging between -1 and 0 and bleed treatment effects ranging between 0 and 1. The means for stroke and bleed treatment effects are set to -0.041 and 0.050, respectively; the standard deviations of the untruncated stroke and bleed treatment effects are set to 0.011 and 0.025, respectively. The correlation shown on the x-axis is the correlation between stroke and bleed treatment effects in an untruncated bivariate normal distribution. Outcomes are shown for bleeds induced per patient, when strokes are held fixed at the rate that would occur when 50% of patients are randomly treated. The worse-case for the optimal homogeneous rule (order by stroke effects) occurs when bleed effects are actually heterogeneous; the worse-case for the optimal heterogeneous rule (order by the ratio of stroke and bleed effects) occurs when bleed effects are actually homogeneous.

Figure A.8: Achieving Worse Outcomes Than Random Treatment



Notes: This figure illustrates the point at which a potential joint distribution of bleed and stroke treatment effects would imply mistakenly optimal treatment rules, formed under the wrong assumption about bleed treatment effects, would perform worse than random treatment of 50% of patients. Potential joint distributions are truncated bivariate normal distributions, with stroke treatment effects ranging between -1 and 0 and bleed treatment effects ranging between 0 and 1. The (untruncated) standard deviation of stroke treatment effects are fixed at 0.011 to reflect the marginal distribution of stroke treatment effects in the VHA data, shown in Panel A of Figure 3. The x-axis shows the (untruncated) standard deviation of the marginal distribution of bleed treatment effects. The lines show the correlations between stroke and bleed treatment effects below which the mistakenly optimal treatment rule would perform worse than random treatment. The thick gray line represents the theoretical point for an (untruncated) bivariate normal distribution, when the ratios of the coefficients of variation between bleed and stroke treatment effects is greater than 1. This point is the same for both the optimal homogeneous rule (ordering by stroke effects) and the optimal heterogeneous rule (ordering by the ratio of stroke and bleed effects). The dashed and solid black lines represent the points, simulated under the corresponding truncated bivariate normal distribution, for the optimal homogeneous and heterogeneous rules, respectively.

Table A.1: Prescription Rates

	Prescription rate $CHADS_2 \leq 1$	Prescription rate $CHADS_2 \geq 2$	$CHADS_2$ awareness	Number of observations
2005	54.42%	50.36%	0.77%	9,693
2013	45.23%	53.34%	67.52%	10,647
Change over time	−9.19	2.98	66.75	

Notes: This table displays the warfarin prescription rate and $CHADS_2$ awareness pre- and post-$CHADS_2$ guideline existence (2005 and 2013, respectively).

Table A.2: Provider Characteristics

Provider characteristics	Adherence
Female	−0.010**
	(0.004)
Cardiologist	0.046***
	(0.007)
Age	−0.001**
	(0.000)
Tenure	0.000
	(0.000)
Number of observations	61,588

Notes: This table shows the relationship between selected provider characteristics and $CHADS_2$ guideline adherence. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table A.3: 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.4: 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	72.6	67.8	68.0	67.8	70.2	73.7	70.7	70.5
Stroke treatment effect, $\hat{\tau}^s$	−0.04	−0.03	−0.03	−0.04	−0.04	−0.04	−0.04	−0.08	−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.5: 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.6: Stroke Causal-Forest Variable Importance and Lasso-Selected Variables

Causal forest	Regression forest	Lasso TE model selected variables	Lasso risk model selected variables
Causal-forest risk (−)	History of stroke or TIA (+)	Lasso risk (−)	History of stroke or TIA (+)
History of TIA (+)	White (−)	White (-)	Systolic blood pressure (+)
White (+)	History of TIA (+)	History of stroke or TIA (−)	Age (+)
History of PVD (+)	Smoker at all (−)	Female (−)	History of MI (+)
Smoker at all (+)	History of hypertension (+)		Female (−)
History of hypertension (+)	History of PVD (−)		History of diabetes (+)
Female (−)	History of diabetes (+)		History of hypertension (+)
History of CHF (+)	History of angina (+)		History of angina (+)
History of stroke or TIA (−)	History of MI (+)		Weight (−)
History of angina (−)	Female (+)		Hemoglobin (−)

Note: Each column of this table shows the top 10 most important variables for the causal-forest, regression-forest, and lasso treatment-effect and risk models in descending order of importance with stroke as the outcome. For lasso-selected variables, we rank them in the order of the absolute magnitude of their coefficient when we regress the lasso treatment effect or lasso risk on these selected variables after standardization and select the top 10 to include in this table. The +/− signs following each variable indicate the sign of its coefficient in a linear regression with the corresponding lasso and forest output as the dependent variable.

Table A.7: Causal-Forest BLP Validation Regressions

	Stroke			Bleed
	(1)	(2)	(3)	(4)
Treatment, W_i	−0.042***	−0.041***	−0.042***	0.018***
	(0.007)	(0.007)	(0.007)	(0.005)
Treatment effect interactions				
$W_i \times \left(\hat{\tau}^o_{-j}(X_i) - \overline{\tau}^o\right)$	0.520***		0.519***	−0.374
	(0.130)		(0.123)	(0.691)
$W_i \times \left(\hat{\tau}^c_{-j}(X_i) - \overline{\tau}^c\right)$		0.666***		
		(0.250)		
$W_i \times \left(\hat{\tau}^r_{-j}(X_i) - \overline{\tau}^r\right)$		0.460**		
		(0.142)		
$W_i \times \left(\text{Pr}_{\text{VHA}}(W_i \mid X_i) - \overline{\text{Pr}_{\text{VHA}}(W_i)}\right)$			−0.046	
			(0.061)	
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
$P_j \times \left(\hat{\tau}^c_{-j}(X_i) - \overline{\tau}^c\right)$		Yes		
$P_j \times \left(\hat{\tau}^r_{-j}(X_i) - \overline{\tau}^r\right)$		Yes		
$\text{Pr}_{\text{VHA}}(W_i \mid X_i)$			Yes	

Notes: This table reports the coefficients of best linear predictor (BLP) validation regressions of stroke and bleed outcomes. Columns 1 and 4 correspond to Equation (4), which interacts treatment with the demeaned full treatment effect $\left(\hat{\tau}^o_{-j}(X_i) - \overline{\tau}^o\right)$ for stroke and bleeding, respectively. Column 2 corresponds to Equation (5), which interacts treatment with the demeaned CHADS_2 and residual components of the stroke treatment effect, or $\left(\hat{\tau}^c_{-j}(X_i) - \overline{\tau}^c\right)$ and $\left(\hat{\tau}^r_{-j}(X_i) - \overline{\tau}^r\right)$. Column 3 reports the specification in Column 1 with the additional interaction between treatment and demeaned VHA propensity to treat, or $\left(\text{Pr}_{\text{VHA}}(D_i \mid X_i) - \overline{\text{Pr}_{\text{VHA}}(D_i)}\right)$. The VHA propensity to treat, Pr_{VHA}, is constructed as the OLS probability of treatment in the VHA data given patient characteristics and exported to the AFI database, detailed further in Appendix A.4. 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 Columns 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 $P_j \times \left(\hat{\tau}^c_{-j}(X_i) - \overline{\tau}^c\right)$ and $P_j \times \left(\hat{\tau}^r_{-j}(X_i) - \overline{\tau}^r\right)$. 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.8: Medication Adherence Summary Statistics

	Mean	Share with prescription
Proportion of days covered		
Statin	0.86	0.59
Beta-adrenergic blockers	0.86	0.42
ACE inhibitors	0.86	0.44
Angiotensin II receptor blockers	0.87	0.11
Calcium channel blockers	0.86	0.33
Diuretics	0.84	0.40
Diabetes	0.89	0.23
Utilization		
Emergency department	0.31	
Primary care	3.34	
Specialist	4.66	
Outpatient visits	19.64	
Hospitalizations	0.16	
Income by zip code		
Median household income	$52,500	

Notes: This table reports the average values of the data used to predict medication adherence. 113,263 VHA patients were included in this sub-analysis - encompassing 99.99% of the analytic cohort. For the PDC measurement, we also report the share of the analytic sample that was prescribed the specified drug class in the year prior to Atrial Fibrillation diagnosis.

Table A.9: Predictive Power of Medication Adherence

	(1)	(2)
	Warfarin PDC	TTR365
Statin PDC	0.071***	0.077***
	(0.009)	(0.010)
Beta-adrenergic blockers PDC	0.044***	0.036***
	(0.010)	(0.011)
ACE inhibitors PDC	0.036***	0.042***
	(0.010)	(0.011)
Angiotensin II receptor blockers PDC	0.064***	0.053**
	(0.021)	(0.023)
Calcium channel blockers PDC	0.060***	0.031**
	(0.012)	(0.013)
Diuretics PDC	0.027***	0.018*
	(0.009)	(0.010)
Diabetes PDC	0.043***	0.066***
	(0.014)	(0.015)
Emergency department	-0.002	-0.016***
	(0.002)	(0.002)
Primary care	0.000	-0.002***
	(0.000)	(0.000)
Specialist	0.000*	-0.002***
	(0.000)	(0.000)
Outpatient visits	-0.000*	-0.001***
	(0.000)	(0.000)
Hospitalizations	-0.009*	-0.012*
	(0.005)	(0.007)
Median household income (in $10,000s)	-0.005***	0.004***
	(0.001)	(0.001)
Mean dep. var	0.745	0.455
S.D. dep. var	0.293	0.287
Adjusted R^2	0.013	0.058
Number of observations	49,203	39,684

Notes: This table reports resykts from Equation A.3. In addition to the regressors shown, the specification also includes indicator variables for whether the patient was prescribed each drug listed above. Robust standard errors are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.