Are Guidelines Worth Following? Treatment Decisions Under Scientific Uncertainty

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

August 19, 2025

Abstract

Physicians often deviate from medical guidelines, but the evidence base for guidelines is often incomplete and unrepresentative. 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 heterogeneity in the benefits of stroke reduction. However, the trials do not (and were not designed to) reveal heterogeneity in the harm of induced bleeding. In the observational data, we find that bleed risk is positively correlated with stroke reduction, raising the possibility that patients who benefit most from treatment may also suffer greater harm from it. Analyzing the performance of treatment rules based on current guidelines, we show that these guidelines may perform worse than randomly treating patients. Finally, we analyze new guidelines robust to scientific uncertainty, according to maximin and minimax-regret objectives, and we demonstrate that these robust guidelines may significantly improve welfare relative to observed treatment decisions. Accounting for scientific uncertainty, nearly all physicians perform better than strict adherence to existing guidelines.

*Abaluck: Yale University and NBER, jason.abaluck@yale.edu. Agha: Harvard University and NBER. agha@hcp.med.harvard.edu. Chan: UC Berkeley, Department of Veterans Affairs, and NBER, david.c.chan@berkeley.edu. Singer: Harvard University and Massachusetts General Hospital, desinger@mgh.harvard.edu. 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

Modern healthcare has witnessed a proliferation of medical therapies, accompanied by a corresponding growth in evidence and clinical guidelines supporting their use. However, it is well known that physicians often deviate from clinical guidelines. These facts have prompted two divergent and arguably reasonable reactions. On the one hand, many prominent thought leaders and professional societies emphasize that guidelines incorporate the best available evidence and that physicians should not only follow them but also be monitored and rewarded for adhering to them. On the other hand, others caution that guidelines are too narrowly focused on single diseases and are silent on large swaths of patients for whom evidence is missing. Randomized controlled trials (RCTs) often select patients who are more likely to benefit and less likely to be harmed by the treatment in question. As a result, large numbers of patients for whom clinical decision-making may be *a priori* the most difficult and consequential are the very ones missing from the evidence base.[1]

The disagreement implied between these two divergent viewpoints may not be resolvable with the current (and likely continuing) gaps in information. This begs two important and general questions. First, how should we use existing information when it is incomplete and unrepresentative? Second, how do we assess provider decisions in light of scientific uncertainty? In this paper, we investigate whether it is possible to integrate both known and missing information into a coherent framework. We consider a specific type of *scientific uncertainty* under which there is an asymmetry in information between patient-specific benefits and harms.[2] Under ideal settings, we would have causal evidence on both patient-specific benefits and harms of treatment. In practice, however, clinicians and policymakers must grapple with the common informational setting in which much less is known about potentially heterogeneous harms of treatment.

Our specific setting concerns a prominent guideline for treating atrial fibrillation (AF), a common condition afflicting more than 5 million people in the US (Colilla et al. 2013), with anticoagulation (or "blood thinning"). The principal risk in atrial fibrillation is debilitating or deadly embolic stroke

1. Twenty percent of Medicare beneficiaries have five or more chronic conditions, and half receive five or more medications. Most clinical guidelines do not discuss the applicability of their recommendations for multimorbid patients, who are already on multiple medications and may suffer more harm under additional treatment. Studies of clinical guidelines conclude that a representative elderly patient with extremely common diseases like chronic obstructive pulmonary disease, type 2 diabetes, osteoporosis, hypertension, and osteoarthritis would need to be on twelve or more medications (as of 2005) if clinical guidelines were to be followed. As noted by Tinetti, Bogardus, and Agostini (2004), "Viewing disease-specific medication guidelines from this perspective raises the question of whether what is good for the disease is always best for the patient" (p. 2870).

2. By scientific uncertainty, we refer to the concept in which the probabilities of states are unknown. The literature refers to this interchangeably by "ambiguity," "uncertainty" (as opposed to "risk"), and "Knightian uncertainty" in the literature. See Gilboa and Marinacci (2013) for a review.

(hereafter, stroke); untreated patients have a 5% risk of stroke per year (Atrial Fibrillation Investigators 1995). Anticoagulation 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 AF patients (Gage et al. 2001). Clinical guidelines recommend tailoring treatment decisions based on patients' $CHADS_2$ scores (Fuster et al. 2006; Hirsh et al. 2008). Following its publication in 2006, the $CHADS_2$ score has been among the most well-known and widely used risk scores for any clinical condition.[3]

We begin by examining physician responses to the introduction of the $CHADS_2$ score among nearly 5,800 physicians treating 113,000 newly diagnosed AF patients in the Veterans Health Administration (VHA) from 2002-2013. For each physician, we proxy awareness of the $CHADS_2$ guideline by identifying the earliest mention of the $CHADS_2$ score in the physician's clinical notes and find that at least 70% of physicians are aware of the guideline by the end of 2013. Prior to awareness, physicians are no more likely to treat patients with higher $CHADS_2$ scores, who are more likely to benefit from anticoagulation according to the guideline; if anything, they are slightly *less* likely to treat these patients. After a doctor's first mention of the $CHADS_2$ score, their practice patterns pivot towards $CHADS_2$-based recommendations, but mostly by a reduction in the treatment probability of patients with lower $CHADS_2$ scores. Despite these changes, physicians aware of the $CHADS_2$ score still deviate from the guideline in more than 40% of cases.

To understand how the $CHADS_2$ score and alternative guidelines affect health outcomes, we must understand how treatment effects (i.e., the benefits and harms of anticoagulation) vary across patients. To this end, we aim to generate novel machine learning (ML) estimates of heterogeneous treatment effects from randomized controlled trial (RCT) microdata. 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) (Walraven et al. 2009). Using a causal-forest model (Wager and Athey 2017; Asher et al. 2016), we estimate conditional average treatment effects (CATEs) on outcomes that vary with patient characteristics; importantly, we validate and shrink our CATEs in "hold-out" trials using a method of best linear prediction (BLP) described by

3. 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, Prokop, and Abu-Saleh 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.

Chernozhukov et al. (2018). While our approach finds substantial heterogeneity in stroke treatment effects, much of it predicted by the $CHADS_2$ score, we cannot validate any heterogeneity in bleed treatment effects.[4] In the absence of detectable heterogeneity in the harms of anticoagulation from the AFI data, we turn to measuring heterogeneity in the non-causal *risks* of bleeding, conditional on treatment, which we must obtain from the VHA data.[5] We examine the root of the problem by comparing stroke CATEs and bleed risks between patients in actual practice and those in the RCTs and show that AFI patients are slightly more likely to benefit from anticoagulation but have much lower bleeding risks, based on both observable and unobservable patient characteristics.[6]

Given the missing evidence on the heterogeneity of harms, we consider two canonical assumptions, both of which have a basis in current clinical and scientific thinking. First, under the scientific null hypothesis, reasonable clinicians may assume that unobserved heterogeneity in harms bears no relation with the heterogeneity in the benefits of treatment (Assumption *A*).[7] Second, clinical intuition and evidence from other settings may suggest that harms are proportional to untreated bleed risks (Assumption *B*).[8] Crucially, we show that, among VHA patients, bleed risks are strongly positively correlated with the stroke-reduction benefits of anticoagulation. Therefore, Assumptions *A* and *B* necessarily conflict with each other: Treating patients who would most likely benefit (as would be optimal under Assumption *A*) may also treat those who would most likely be harmed (under Assumption *B*). Despite our best efforts, we cannot reject either Assumption *A* or Assumption *B* by examining bleeds induced by randomized treatment in the AFI data according to VHA-estimated bleed risk, further highlighting the limitations of the RCT evidence base.

We then proceed by considering the implications of treatment rules on health outcomes, depending on the assumption about treatment harms (i.e., Assumption *A* vs. *B*). Conceptualizing treatment

4. At the 90th percentile, warfarin reduces stroke risk by six percentage points, double the reduction in stroke risk at the 10th percentile. A regression of CATEs on $CHADS_2$ score indicators has an R^2 of 0.49. The predictiveness of the $CHADS_2$ score is especially striking given its simplicity and the absence of regressions in its development (Gage et al. 2001).

5. We are also unable to validate meaningful heterogeneity in the AFI data in bleeding risks (i.e., heterogeneity in bleeding rates across patient characteristics among patient in the control arms of the RCTs).

6. Bleed risks predicted by patient characteristics are lower among untreated patients in the VHA data than among untreated patients in the AFI data. Further, the difference between observed and predicted bleed rates is higher in the VHA data than in the AFI data.

7. This assumption is implicit in the strong recommendations to follow the $CHADS_2$ score in treating patients. The $CHADS_2$ score is a prediction of the benefits of treatment; in the absence of evidence on harms, clinical guidelines recommend treating patients based on their presumed benefits without providing explicit guidance on how to incorporate harms.

8. Interestingly, as pointed out by Chandra, Flack, and Obermeyer (2024), this assumption also has substantial support from scientific evidence and among the clinical community, particularly in for cardiovascular diseases. Major randomized trials demonstrate larger absolute risk reductions from statins in groups with a higher predicted risk of heart disease, regardless of whether the risk is predicted by demographics, comorbidities, laboratory tests, or genetic markers (US Preventive Services Task Force 2016; Natarajan et al. 2017). The assumption underlies the $CHADS_2$ score itself, as prior medical studies have not estimated heterogeneity in stroke reduction; instead, they have made recommendations based on underlying stroke risk in the absence of treatment.

rules as an ordering of patients and a threshold above which treatment is recommended, our analytical framework compares stroke and bleed outcomes for a given ordering of patients while varying the share of patients treated. Akin to a production technology, an ordering yields a frontier of strokes prevented and bleeds induced, depending on the share of patients treated. An ordering may strictly outperform another one if, for any level of strokes prevented, it induces fewer bleeds.

Compared to treating a random half of the patients, ordering patients by the $CHADS_2$ score can induce 15.1% fewer bleeds while preventing the same number of strokes when we assume that the harms of treatment are independent of the benefits (Assumption *A*). The $CHADS_2$ score performs strikingly well compared to the optimal ranking under the presumed stroke CATEs: Under independent harms and the estimated stroke CATEs, the fewest possible bleeds a doctor can obtain (by ranking patients by their stroke CATEs) would be 21.5% fewer than the random benchmark, suggesting that the $CHADS_2$ score obtains about 70% of the maximum reduction in induced bleeds. However, under bleed harms proportional to bleed risks (Assumption *B*), we find that the $CHADS_2$ score performs *worse* than random treatment, inducing 6.6% more bleeds than random treatment. The updated and more granular CHA_2DS_2-VASc score performs no better than the $CHADS_2$ score.[9] Ordering patients by their stroke CATEs (optimal under Assumption *A*) also performs poorly under Assumption *B*: It reduces only 4.0% of induced bleeds compared to random treatment. Similarly, the performance of the optimal ranking under Assumption *B*—ranking by the ratio of stroke benefits to bleed risks—is highly sensitive to the assumption on harms.[10]

The large performance implications driven by unknown heterogeneity in harms motivate us to characterize novel orderings robust to scientific uncertainty. Following the literature on welfare criteria under scientific uncertainty, we consider two specific benchmarks (Manski 2004; 2011): We characterize a maximin treatment rule that maximizes the worst welfare, along the set of mixtures between Assumptions *A* and *B*, and a minimax-regret treatment rule minimizes the greatest regret between the best welfare and the welfare obtained along the same set. We show that both treatment rules empirically demonstrate monotonicity, such that patients can be ordered in order of marginal treatment as the utility cost of bleeds relative to strokes increases. The maximin ordering orders patients so that bleed risk is independent of the ranking and always equal to the overall bleed risk. This ordering may be equivalent to the optimal ranking under Assumption *A* (i.e., ranking by benefit) if bleed

9. Under Assumption *A*, the CHA_2DS_2-VASc score ranking obtains a similar number of bleeds for the same number of strokes as the $CHADS_2$ score score. It performs even worse under Assumption *B*, inducing 9.5% more bleeds than the random benchmark.

10. Under Assumption *B* this ranking would induce 32.9% fewer bleeds than random treatment, but under Assumption *A* it would induce only 3.0% fewer bleeds than random treatment.

risks are negatively correlated with stroke benefits. However, minimax-regret treatment rules always differ from both assumption-specific optimal treatment rules, splitting the difference so that regret is equalized between assumptions. Empirically, we demonstrate that these orderings perform very well regardless of the relationship between harms and benefits: The maximin ranking obtains 18.8% fewer bleeds compared to random treatment, regardless of the assumption on harms; the minimax-regret ordering obtains 14.8% to 26.4% fewer bleeds compared to random treatment.

Finally, we assess the health consequences of physician decisions. Depending on the assumptions about harms, we parameterize physician decisions with a simple skill-threshold structural model (Chan, Gentzkow, and Yu 2022), in which a physician's skill is the correlation between her ordering of patients and the assumption-specific optimal ranking. We find modest variation in skill and wide variation in treatment propensities across physicians. Importantly, the assumption about harms implies starkly different conclusions about the performance of physicians relative to that of established guidelines. Under independent harms, all physicians perform worse than the $CHADS_2$ score, and only one in five perform better than random. Under proportional harms, all physicians perform better than the $CHADS_2$ score, and nine in ten perform better than random. Under the worst-case criterion, almost all physicians perform better than the $CHADS_2$ score. We also predict via ML the within-physician probability of treatment and use the predicted probability of treatment as a "wisdom of the crowds" ordering. Interestingly, the crowd performs better than the average physician, suggesting that adhering to recommendations based on such AI implementation (see, e.g., Agarwal et al. 2023) may improve overall outcomes.

Our results contribute to an active literature in economics that studies the potential for machine-based algorithms to enhance decision-making. Evidence from clinical care (Abaluck et al. 2016; Mullainathan and Obermeyer 2022) to judge bail decisions (Kleinberg et al. 2018) suggests that human experts make frequent mistakes that could, in principle, be corrected by decision protocols.[11] To analyze these questions, the previous literature has assumed that it can measure causal effects on the outcomes of interest. We directly address the implications of scientific uncertainty, which may obscure expected outcomes due to missing information, for the formulation and evaluation of guidelines. In doing so, we apply important theoretical concepts from statistical decision theory (e.g., Savage 1951; Manski 2011; Gilboa and Marinacci 2013) to a concrete and rich empirical setting.

We develop an approach to using pooled RCT data to estimate treatment effect heterogeneity,

11. 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, Kahn, and Li (2018) finds that managers, in their hiring decisions, frequently overrule a technology-driven hiring recommendation but that doing so worsens outcomes in aggregate.

validating the estimates using held-out trials and relating their informational content relative to OLS estimates from the observational data.[12] The use of RCTs for the detection of treatment effect heterogeneity may provide important advantages for studying heterogeneity, 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; Chaisemartin 2017; Frandsen, Lefgren, and Leslie 2023). However, while internally valid heterogeneous causal effects may sometimes only be obtained from RCTs, our analysis addresses how decision-makers may account for scientific uncertainty (i.e., external validity) in the available evidence base.

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.[13] We show that physicians continue to deviate from a guideline of which they demonstrate awareness. Building on prior work studying the impact of guideline adherence (e.g., Cuddy and Currie 2020), we examine the performance of physicians and guidelines in the face of patient characteristics omitted from the guidelines and scientific uncertainty. To our knowledge, our work shows novel and stark empirical evidence that the verdict crucially depends on assumptions about unknown and potentially heterogeneous harms.[14]

The remainder of this paper is organized as follows. Section 2 describes the clinical setting, the VHA and AFI data, and basic facts about physician awareness of and adherence to the $CHADS_2$ score. Section 3 presents our estimates of causal-forest treatment effects. Section 4 analyzes known and optimal guidelines under scientific uncertainty about the distribution of treatment harms. Section 5 analyzes physician behavior and their performance relative to guidelines. Section 6 concludes with a discussion of policy implications.

12. In contrast, many prior papers comparing machine decisions to human discretion have relied on observational data and quasi-experimental assumptions to reach conclusions about misallocation. Many of these papers compare treatments and outcomes across decision-makers, usually assuming a common ordering of cases across decision-makers (Abaluck et al. 2016; Kleinberg et al. 2018; Chandra and Staiger 2020). Newer approaches in Chan, Gentzkow, and Yu 2022 and Arnold, Dobbie, and Hull 2022 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, Gentzkow, and Yu 2022).

13. An older review article estimated that 40% of patients were not receiving guideline-recommended care for chronic conditions (Schuster, McGlynn, and Brook 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.

14. 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), although this literature generally does not separate the impact of guidelines from other differences across physicians.

2 Setting, Data, and Descriptive Facts

2.1 Clinical Setting

Atrial fibrillation (AF) 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). AF 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 AF is anticoagulation by warfarin.[15] While anticoagulation is effective in reducing stroke risk by inhibiting the likelihood of blood to clot, by 68% on average, it unsurprisingly also increases the risk of major bleeding, by more than double (Atrial Fibrillation Investigators 1995; Kearon et al. 2012). Given the large potential benefits and harms of anticoagulation, an important task for clinicians evaluating AF patients is to incorporate information about the patients to decide whom 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 to increase with baseline stroke risk. Earlier studies re-analyzed data from the control arms of randomized trials of AF patients to find hypertension and prior stroke as important risk factors of stroke (Atrial Fibrillation Investigators 1995; Stroke Prevention in Atrial Fibrillation Investigators 1995). Using the risk factors identified from the trial data, the $CHADS_2$ score was first validated as a stroke prediction among Medicare registry data in 2001 and recommended for clinical practice in 2004 (Gage et al. 2001; Gage et al. 2004). The $CHADS_2$ score was subsequently adopted as a clinical guideline for treatment decisions, first by the American College of Cardiology (ACC) (Fuster et al. 2006), then by the American College of Chest Physicians (ACCP) (Hirsh et al. 2008).

Designed for ease of 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). Since its introduction, the $CHADS_2$ score has become one of the most widely recognized risk scores in clinical practice. Later, in 2010, researchers introduced a modified risk score—CHA_2DS_2-VASc score—accounting for additional characteristics that were suspected to predict strokes: vascular disease, age between 65 and 74 years, and female sex

15. 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 AF (Hsu et al. 2016).

(Lip et al. 2010). The CHA_2DS_2-VASc score has since become the standard of practice, though it was much less recognized during the period of our VHA cohort, from 2002 to 2013. Table 1 describes both the $CHADS_2$ score and its later CHA_2DS_2-VASc variant.

Despite its widespread recognition, studies in a variety of settings have shown that adherence to the $CHADS_2$ score and its later variant has been low, typically with only half of the recommended patients being prescribed anticoagulation according to the risk scores (Hsu et al. 2016; Piccini and Fonarow 2016). Physicians commonly cite concerns of increased bleeding risk due to frailty and multi-morbidity to justify non-adherence to the $CHADS_2$-based treatment recommendations (Fawzy, Olshansky, and Lip 2019). Since AF is a condition of the elderly, it frequently coexists with other comorbid conditions. Many of these conditions (e.g., unstable gait, renal insufficiency, hypertension, and older age in itself) may increase the risk of bleeding. The first formal risk score for bleeding (HAS-BLED) was published in 2010 (Pisters et al. 2010; Lane and Lip 2012). However, it remains unvalidated in the population of AF patients and has not yet been incorporated formally in society guidelines for anticoagulation among patients.

2.2 Data

Veterans Health Administration Data. To study AF treatment decisions, we identify patients with a new AF diagnosis using electronic medical records from the Veterans Health Administration (VHA) from October 2002 to December 2013. Following a protocol developed in previous work, we err on the side of defining a narrower cohort of patients who are more likely to have a new, confirmed AF diagnosis and to receive care at the VHA (Turakhia et al. 2013; Perino et al. 2017).[16]

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 and to match the other clinical characteristics recorded in the RCT data. To capture the anticoagulation decision, we rely on VHA records of prescriptions for warfarin

16. We first identify potentially new diagnoses of AF (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 AF) near the time of initial diagnosis and no prior anticoagulation. After an AF diagnosis, 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. 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 AF diagnosis. We also require each attributed physician to have at least 30 AF patients and to have prescribed warfarin for another patient. Our sample restrictions result in an analytic cohort of 113,270 patients. Appendix Table A.1 describes the cohort construction steps in detail.

or a novel oral anticoagulant (e.g., apixaban, dabigatran, edoxaban, rivaroxaban). Because noval oral anticoagulants were much less commonly used during our cohort period, we refer to warfarin and anticoagulation interchangeably.[17] The VHA records include prescriptions that are dispensed by the VHA as well as prescriptions that are paid for by the VHA.

We also use the VHA electronic health record to measure awareness of the $CHADS_2$ score guideline by searching physician visit notes for mentions of the $CHADS_2$ score in a broader cohort of AF patients up to six months before an index patient's initial AF diagnosis. This broader cohort contains 844,312 AF patients, most of whom have been previously diagnosed with AF. Our approach to measuring awareness of the $CHADS_2$ guideline represents a lower bound, as physicians may be aware of the guideline but do not mention it in their notes.[18]

Atrial Fibrillation Investigators Database. To evaluate the evidence base of heterogeneous benefits and harms of treatment, we rely on the Atrial Fibrillation Investigators database (hereafter, AFI database). We use the AFI database to analyze patient-level observations from eight trials in which patients were randomized to anticoagulants versus placebo or control.[19] Notably, the AFI database includes the same trials used to develop the clinical guidelines and, to our knowledge, represents the best causal evidence available on the effects of warfarin among AF patients. Details of the AFI database have been documented elsewhere (Walraven et al. 2009).

The AFI database was previously compiled by investigators to explore heterogeneity in stroke risk and treatment effects. Previous analyses using the database have selected patient characteristics heuristically (Walraven et al. 2002; Walraven et al. 2009). For each patient in the AFI database, we observe treatment randomization status and key 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

17. Fewer than 2% of patients are prescribed a novel oral anticoagulant. Among patients prescribed an anticoagulant, 4% are prescribed a NOAC.

18. Specifically, we search for non-case-sensitive occurrences of the string `chads`. While there may be false negatives, we found no evidence of false positives. 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.

19. 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 include their data to increase power for estimation of risk prediction models.

pressure, hemoglobin, smoking status, comorbidities, and history of transient ischemic attack (TIA), stroke, anginal symptoms, and myocardial infarction. In our analyses, we impute missing patient characteristics using a regression forest model. This model is trained using other characteristics as predictors and fitted on patients with observed values. If some predictors are missing, the model will consider their missingness as a potential "value" to be used for the prediction (Tibshirani, Athey, and Wager 2020). Then, the model is applied to patients who are missing specific characteristics, generating predicted values. Appendix Table A.2 lists the full set of characteristics that we use from the AFI database. Table 2 reports summary statistics for the VHA and AFI AF cohorts; Appendix Table A.3 reports further statistics for each trial in the AFI database. Finally, Appendix Table A.4 shows the results of balance tests, which suggest successful randomization in these clinical trials.

2.3 Descriptive Facts about Guideline Awareness and Adherence

Before assessing the RCT evidence base for heterogeneous benefits and harms and analyzing the implications of scientific uncertainty for actual practice, we use the rich electronic record of the VHA to establish some novel descriptive facts about guideline awareness and adherence. We focus on the $CHADS_2$ guideline rather than its later CHA_2DS_2-VASc variant because the former was much more commonly recognized during the time period of our cohort.[20]

Relying on treatment recommendations according to the 2006 ACC and 2008 ACCP guidelines based on the $CHADS_2$ score, we note that AF patients with a $CHADS_2$ score of 0 or 1 could forego anticoagulation with warfarin, while those with a $CHADS_2$ score of 2 were recommended anticoagulation (Hirsh et al. 2008; Fuster et al. 2006). Panel A of Figure 1 displays treatment trends for these two groups of patients. Before the 2006 guideline, patients with lower $CHADS_2$ scores were, if anything, *more* likely to be treated with anticoagulation. Following 2006, patients with lower $CHADS_2$ scores saw a relative reduction in anticoagulation. However, even among groups where anticoagulation is recommended, prescription rates remain below 55% for our sample period. Patients with high 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). Panel B of Figure 1 shows that, despite modest changes in anticoagulation among AF patients, a substantial share of physicians mention the $CHADS_2$ score after the $CHADS_2$ guideline is introduced. Nearly 70% of physicians will have mentioned the $CHADS_2$ score at least once before the end of our study period in 2013.

20. Specifically, while 23% of patient encounters in our data are with 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 link prescription patterns with awareness of the guideline. In Figure 2, we consider physicians who become aware of the guideline as being treated in an event study and plot their anticoagulation rates among high- and low-score patients after the physician's first note mentioning the $CHADS_2$ score. Following this event, if anything, the low-scoring patients see a modest reduction in anticoagulation, while any change in anticoagulation for higher-scoring patients is less noticeable. We describe this analysis further in Appendix A.1. Appendix Figure A.2 shows that while adherence varied substantively across physicians, few physicians reached an adherence rate of 80%. We find qualitative differences in adherence between groups of physicians, but these differences are quantitatively small.[21]

3 The Evidence Base

In this section, we will describe and assess the existing base of evidence upon which guidelines can be formed. Ideally, the clinical decisions should be based on the heterogeneous *benefits* and *harms* of treatment—for anticoagulation, reducing strokes but inducing bleeds, respectively. Both of these are causal objects that require a key unconfoundedness assumption of random assignment to treatment within all subsets of patients for whom guidelines may tailor recommendations. We will describe our machine learning and validation procedures to extract heterogeneous treatment effects from the AFI database. In the absence of causal harms that can be validated out of sample, we will consider heterogeneous bleed *risks*—the probability of bleeding conditional on treatment—as a key statistic that may relate to treatment harms.

3.1 RCTs for Extracting Benefits and Harms

Let $Y_i^s(d) \in \{0,1\}$ denote potential outcomes for whether patient i will have a stroke, depending on anticoagulation treatment $d \in \{0,1\}$, and let $Y_i^b(d) \in \{0,1\}$ denote a similar object for bleeding. Empirical estimates of treatment effects on strokes and bleeds (i.e., benefits and harms) based on patient observable characteristics are intuitively key inputs in policy decisions regarding which patients to treat.

Specifically, using the existing RCT data in the AFI database, we aim to estimate the following

21. 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.

conditional average treatment effects (CATEs) for one-year stroke and bleeding, respectively:

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

$$\tau^b(x) \equiv E\left[Y_i^b(1) - Y_i^b(0) \middle| X_i = x\right], \quad (2)$$

where X_i is a set characteristics belonging to patient i. We will describe how such CATEs play a key role in designing policies to maximize patient welfare in Section 4.

RCTs are considered the gold standard for estimating causal effects and occupy a preeminent place in the formulation of evidence for clinical guidelines. They may sometimes be the only source of evidence where the heterogeneous treatment effects of a drug may be identified. First, by design, RCTs randomize patients to treatment within all subsets of characteristics. Second, RCTs usually satisfy stricter requirements for extracting valid treatment effect heterogeneity—i.e., monotonicity and exclusion-restriction within all patient subgroups—whereas quasi-experiments may only be able to estimate treatment effects averaged over all compliers when violations may exist within subgroups but "average out" across subgroups (e.g., Kolesar et al. 2015; Frandsen, Lefgren, and Leslie 2023). Third, by explicitly randomizing treatment and by blinding participants to their treatment status, RCTs may more plausibly meet the exclusion criterion of impacting outcomes only through treatment. Fourth, RCTs rigorously collect information on infrequent yet important outcomes (e.g., stroke and bleeding), especially the timing of these events relative to initiation of treatment; doing the same in observational data can be very challenging, as we discovered in our setting with the VHA data.[22]

At the same time, RCTs suffer from important limitations in informing clinical guidelines. Because of their high cost, RCTs typically enroll hundreds to at most a few thousand patients, resulting in a ratio of thousands of patients per each RCT-enrolled patient.[23] RCTs are generally underpowered for heterogeneity analyses; their small sample size puts them at risk for both false negative and false positive findings related to heterogeneity (Kent et al. 2010). Finally, perhaps most importantly, RCTs are well known to suffer from external validity issues due to highly unrepresentative patient sampling (Rothwell 2005). Patients who are more likely to be harmed by treatment (e.g., elderly, multimorbid

22. Specifically, events such as strokes are recorded in visits with diagnosis codes, both in data generated from insurance claims and those from electronic health records. Records of these diagnosis codes can indicate either a new stroke event or a history of stroke. In our audit of the VHA data at the Palo Alto VA, we found that 40% of patients recorded in diagnosis codes to be experiencing a current stroke were revealed on detailed chart review to have only a history of stroke but no recurrence. This insight required us to drop any patient with a previous stroke diagnosis when we wished to detect new stroke events in the observational data.

23. As of 2004, when the $CHADS_2$ score was first recommended for clinical practice, fewer than 10,000 AF patients had been enrolled in any RCT studying anticoagulation, while the worldwide population of patients with AF likely exceeded 30 million individuals (Chugh et al. 2014).

patients) are underrepresented in RCTs; very few RCTs even bother to compare their trial patient characteristics with the overall patient population (Kennedy-Martin et al. 2015).[24] We detail the sample size and inclusion and exclusion criteria of each of the RCTs in the AFI database in Appendix Tables A.6 and A.7.

3.2 Machine Learning Procedure

Using patient-level data from the eight RCTs in the AFI database, we apply an ensemble machine learning (ML) approach that combines causal forest and LASSO to estimate CATEs with conditional random assignment. The LASSO method is adapted from Belloni, Chernozhukov, and Hansen (2014), and the causal forest methods from Wager and Athey (2017). The ensemble approach combines both predictions to improve overall predictive validity (Younas et al. 2022), leveraging the causal forest to discover predictive interactions among discrete variables and the LASSO to better fit predictions that smoothly vary with the continuous random variables.

Our machine learning model considers not only the variables underlying the $CHADS_2$ score, but also additional risk factors present in both the VHA data and the AFI database, as listed in Appendix Table A.2. In addition to including individual patient characteristics to predict treatment effects, we use machine learning predictions of stroke and bleed risk among control-group patients as inputs, allowing treatment effects to depend more parsimoniously on these predictions of risk. Appendix A.2.1 elaborates further details of our prediction procedure.

We subject our ML predictions to a validation exercise based on Chernozhukov et al. (2018) that relies on the eight-trial structure of the AFI database. In this exercise, we use observations in each of the eight trials in the AFI database as a "hold-out" sample to validate ML "proxy" predictions generated by the ensemble approach described above and applied to data from the remaining seven trials. This validation exercise both confirms that ML proxy predictions indeed transfer across trials and extracts consistent "best linear predictor" (BLP) estimates of the underlying CATEs. We use this approach to assess whether there is any valid heterogeneity in CATEs for strokes and bleeds. For stroke CATE proxy predictions, we also use a related approach to separately assess the validity of their $CHADS_2$-score projection and their residual components. We describe our validation procedure in further detail in Appendix A.2.2. Henceforth, we refer to the BLP estimates as our CATE estimates;

24. Kennedy-Martin et al. (2015) systematically reviewed 5,456 RCTs in cardiology, mental health, and oncology and found that only 1% of the RCTs compared trial patient characteristics with the overall patient population. Almost all of these trials concluded that their trial sample was unrepresentative of the overall patient population. Some RCTs excluded up to 90% of the overall patient population by eligibility criteria alone.

we use the term ML proxy predictions to refer to the ensemble predictions that are inputs in the validation exercise.

3.3 Validity and Representativeness

As described above, our ML procedure involves a BLP validation exercise that examines whether CATE proxy predictions apply across trials. Specifically, we ask whether predictions trained on observations from all but one hold-out trial indeed predict outcomes in the hold-out trial within a regression framework. We find that proxy predictions for stroke CATEs strongly transfer across trials. However, this is not the case for bleed CATE proxy predictions: In the BLP regressions, these proxy predictions do not predict bleed outcomes in hold-out trials.

Two analyses lend credence to this extrapolation of AFI treatment effects. We find that stroke outcomes among untreated patients in the VHA are highly correlated with CATEs estimated in the AFI data. Appendix Figure A.5 shows that estimated CATEs are strikingly proportional to risk and align closely with the 47% average reduction in strokes estimated in the original AFI analysis (Wallraven et al. 2002). Along similar reasoning as in the previous literature (Currie and MacLeod 2017; Mullainathan and Obermeyer 2022), this observational pattern is consistent with the possibility that patients with large estimated stroke CATEs suffer considerable harm from under-treatment, while patients with small stroke CATEs have low stroke incidence and thus small potential benefits from anticoagulation.

3.4 Estimates of Benefits and Risks

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 CHADS_2 score.[25]

Because we cannot detect bleed effect heterogeneity from the RCT data, we focus on estimating

25. 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.

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.

4 Guidelines Under Scientific Uncertainty

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 3. 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 consider the case of bleed treatment effects that are independent of stroke treatment effects. Second, we consider bleed treatment effects to be proportional to bleed risks. In Section 3, we showed 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.

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 have a better understanding of treatment benefits

or harms. Clinicians may go beyond the handful of variables encoded in the simple risk scores to identify patients with high stroke risk (and thus potentially greater treatment benefits), or they may better predict stroke treatment effects (which may not be strictly proportional to stroke risk, as the guideline presumes). A second way that physicians may outperform known guidelines is by explicitly considering the possibility that anticoagulation's bleed harms may be correlated with stroke benefits. If the benefits of treatment are positively correlated with its harms, then the guideline approach of prioritizing treatment for patients strictly based on their expected benefits may lead to excess harms.

4.1 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 ordering 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 AF 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.[26]

Figure 6 and Table **??** 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, suggesting that physicians are at least not using observable information that predicts treatment effects beyond what is in the $CHADS_2$ score.

4.2 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,[27] two guideline orderings can be welfare-ordered if one guideline prevents more strokes than the other guideline, for any fixed number of induced bleeds.

In Figure 6 Panel A, we see that under the assumption of independent bleed treatment effects

26. As we will discuss below, the outcomes in this second benchmark may not correspond to any individual 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.

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

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 prioritize 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 **??**. 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 mean-independent harms, 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.

4.3 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, Olshansky, and Lip 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).[28]

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 AF 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 of 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 **??** for a summary of predictor variables). In Appendix Figure **??**, 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 consistency, physicians adhere to the $CHADS_2$-based guideline only 60% of the time. Surprisingly, 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).

The findings in this section present a puzzle: why does guideline adherence remain persistently low, despite the fact that most providers are aware of the guideline? And, would patients' health outcomes improve if physicians followed the guideline more stringently? The rest of the paper is dedicated to exploring these two questions.

4.4 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

28. 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.

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/α_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 ρ defines a tangent to this frontier and could be used to select the optimal point.[29] 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 ρ.

In Table **??**, we report the outcomes following the optimal treatment rules, first assuming mean-independent harms 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 benefits and bleed risks τ_i^s/α_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 reductions in strokes but also

29. 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).

larger 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, 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.[30]

4.5 Uncertainty-Robust Maximin 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—mean-independent vs. proportional harms—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 independent of observable (α_i^b, τ_i^s), or $\tau_i^b = \overline{\tau}^b \equiv E\left[\tau_i^b\right]$ for all i. In the more complicated world B, we consider τ_i^b as proportional to α_i^b, or $\tau_i^b = \left(\overline{\tau}^b / \overline{\alpha}^b\right) \alpha_i^b$, where $\overline{\alpha}^b \equiv E\left[\alpha_i^b\right]$. Relative to no treatment, the expected utilities of treating patient i in the two worlds are, respectively,

$$u_A(i) = -\tau_i^s - \beta\overline{\tau}^b; \tag{3}$$

$$u_B(i) = -\tau_i^s - \beta\frac{\overline{\tau}^b}{\overline{\alpha}^b}\alpha_i^b. \tag{4}$$

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

$$W_\omega(D) = \int_0^1 u_\omega(i) D(i) di. \tag{5}$$

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

30. Appendix A.4 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.

0) and $D_B^*(i) = \mathbf{1}(u_B(i) > 0)$, respectively.

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

$$D^* = \arg\max_D(\min(W_A(D), W_B(D))). \tag{6}$$

In Appendix A.4, we show that maximin treatment rules take the form of a threshold that is linear in stroke treatment effects and bleed risks. Specifically, maximin treatment rules take the form $D^* = \mathbf{1}(\tau_i^s < -\beta(\overline{\alpha}^b + \chi/(1+\chi)(\alpha_i^b - \overline{\alpha}^b)))$, 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 (α_i^b, τ_i^s). In turn, it determines the slope of the maximin rule D^* in (α_i^b, τ_i^s) space (i.e., $-\beta\chi/(1+\chi)$).

The set of possible maximin treatment rules nest the optimal treatment rules under mean-independent and proportional harms.[31] Intuitively, the maximin treatment rule becomes the optimal treatment rule under mean-independent harms—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 maximin 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 maximin 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 A.4 examines the relative performance of D_A^*, D_B^*, and D^* under alternative correlations between α_i^b and τ_i^s.

In Appendix A.4, we further explore the idea of *monotonicity* in maximin treatment rules (i.e., as β decreases, patients will only be induced into treatment and not out of treatment).[32] When a treatment rule satisfies monotonicity, it may recommend an ordering of patients upon which doctors

31. The optimal treatment maximizing W_A under mean-independent harms is $D_A^*(i) = \mathbf{1}(u_A(i) > 0) = \mathbf{1}(\tau_i^s < \beta\overline{\alpha}^b)$; the optimal treatment maximizing W_B under proportional harms is $D_B^*(i) = \mathbf{1}(u_B(i) > 0) = \mathbf{1}(\tau_i^s < \beta\alpha_i^b)$. D^* approaches D_A^* when χ approaches 0, and it approaches D_B^* when χ approaches ∞.

32. D_A^* and D_B^* clearly satisfy monotonicity.

and policy-makers may choose an appropriate threshold, depending on β. We show that the joint distribution of (α_i^b, τ_i^s) in our setting implies monotonicity in the maximin rule and that a class of common distributions would similarly imply monotonicity.

4.6 Evaluating Maximin Guidelines

In Figure 11, we show the achievable decision frontier under optimal guidelines. In Panel A, we plot scenarios assuming mean-independent harms (Assumption *A*). 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 maximin 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 harms proportional to bleed risk (Assumption *B*). The maximin-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 **??**, we report that maximin 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 maximin 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 maximin frontier, the maximin rule could strictly outperform the status quo by preventing more strokes without increasing the number of bleeds (as illustrated in Figure 11).

5 Physician Decisions

As a final analysis, we reevaluate physician decisions to assess the performance of an algorithmic ordering 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. This model yields a patient ordering by aggregating physicians' revealed treatment propensity based on patient characteristics. This ordering appeals to a "wisdom of the crowds" philosophy that large groups may

choose better than individuals(Surowiecki 2005; Gillen, Plott, and Shum 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 A.9, we examine the features of our crowd ordering relative to known and optimal guidelines. On the x-axis, we arrange patients by their percentiles in each ordering, 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 maximin 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 4 to evaluate the performance of this "crowd" ordering. 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 mean-independent of stroke treatment effects (Assumption A), and (ii) bleed treatment effects are proportional to bleed risks (Assumption B). 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 ordering; in the status quo, physicians often deviate from the crowd ordering and vary in their treatment thresholds.

Figure A.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 independent harms, it performs substantially better than random treatment under 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 independent harms, they perform substantively worse than random treatment under proportional harms.

In Figure 9, 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 maximin 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 independent harms). 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.

6 Conclusion

We study the introduction of a prominent guideline to treat atrial fibrillation (AF) 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 accords with physician guideline adherence in many other settings. We harness a collection of RCTs studying anticoagulation for treating AF to analyze the implications of adherence, or lack thereof, for clinical decision rules. We apply machine learningmethods 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, both of which are implicit in (different) existing guidelines: (i) bleed treatment effects are homogeneous, or at least independent of 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 maximin 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. 2016. "The Determinants of Productivity in Medical Testing: Intensity and Allocation of Care." *American Economic Review* 106 (12): 3730–64.

Agarwal, Nikhil, Alex Moehring, Pranav Rajpurkar, and Tobias Salz. 2023. "Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology." NBER Working Paper 31422.

Amroze, Azraa, Kathleen Mazor, Sybil Crawford, Kevin O'Day, David D. McManus, and Alok Kapoor. 2019. "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* 48 (4): 629–637.

Arnold, David, Will Dobbie, and Peter Hull. 2022. "Measuring Racial Discrimination in Bail Decisions." *American Economic Review* 112 (9): 2992–3038.

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. 2018. "Electronic Physician Notifications to Improve Guideline-Based Anticoagulation in Atrial Fibrillation: A Randomized Controlled Trial." *Journal of General Internal Medicine* 33 (12): 2070–2077.

Asher, Sam, Denis Nekipelov, Paul Novosad, and Stephen P Ryan. 2016. "Classification trees for heterogeneous moment-based models." NBER Working Paper 22976.

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

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

Atrial Fibrillation Investigators. 1995. "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* 154 (3): 1449–1457.

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

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

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

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

Chan, David C, Matthew Gentzkow, and Chuan Yu. 2022. "Selection with Variation in Diagnostic Skill: Evidence from Radiologists." *The Quarterly Journal of Economics* 137 (2): 729–783.

Chandra, Amitabh, Evan Flack, and Ziad Obermeyer. 2024. "The Health Costs of Cost Sharing." *The Quarterly Journal of Economics,* qjae015.

Chandra, Amitabh, and Douglas O. Staiger. 2020. "Identifying Sources of Inefficiency in Healthcare." *Quarterly Journal of Economics* 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. 2017. "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* 17 (1): 127.

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

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

Chernozhukov, Victor, Ivan Fernandez-Val, and Alfred Galichon. 2010. "Quantile and Probability Curves Without Crossing." *Econometrica* 78 (3): 1093–1125.

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

Chugh, Sumeet S., Rasmus Havmoeller, Kumar Narayanan, David Singh, Michiel Rienstra, Emelia J. Benjamin, Richard F. Gillum, et al. 2014. "Worldwide Epidemiology of Atrial Fibrillation." Publisher: American Heart Association, *Circulation* 129 (8): 837–847.

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

Cuddy, Emily, and Janet Currie. 2020. "Rules vs. Discretion: Treatment of Mental Illness in U.S. Adolescents." NBER Working Paper 27890.

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

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

Fang, Margaret C., Alan S. Go, Yuchiao Chang, Leila Borowsky, Niela K. Pomernacki, and Daniel E. Singer. 2008. "Comparison of Risk Stratification Schemes to Predict Thromboembolism in People With Nonvalvular Atrial Fibrillation." *Journal of the American College of Cardiology* 51 (8): 810–815.

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

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

Frandsen, Brigham, Lars Lefgren, and Emily Leslie. 2023. "Judging Judge Fixed Effects." *American Economic Review* 113 (1): 253–277.

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. 2006. "ACC/AHA/ESC 2006 Guidelines for the Management of Patients with Atrial Fibrillation–Executive Summary." *European Heart Journal* 27 (16): 1979–2030.

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

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

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

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

Gottlieb, Lawrence K, and Susanne Salem-Schatz. 1994. "Anticoagulation in Atrial Fibrillation: Does Efficacy in Clinical Trials Translate into Effectiveness in Practice?" *Archives of Internal Medicine* 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. 2003. "Factors Influencing Physicians' Reported Use of Anticoagulation Therapy in Nonvalvular Atrial Fibrillation: A Cross-sectional Survey." *Clinical Therapeutics* 25 (6): 1750–1764.

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

Hoffman, Mitchell, Lisa B. Kahn, and Danielle Li. 2018. "Discretion in Hiring." *Quarterly Journal of Economics* 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. 2016. "Oral Anticoagulant Therapy Prescription in Patients With Atrial Fibrillation Across the Spectrum of Stroke Risk: Insights From the NCDR PINNACLE Registry." *JAMA Cardiology* 1 (1): 55–62.

Kearon, Clive, Elie A. Akl, Anthony J. Comerota, Paolo Prandoni, Henri Bounameaux, Samuel Z. Goldhaber, Michael E. Nelson, et al. 2012. "Antithrombotic Therapy for VTE Disease: Antithrombotic Therapy and Prevention of Thrombosis, 9th ed: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines." *Chest* 141 (2, Supplement): e419S–e496S.

Kennedy-Martin, Tessa, Sarah Curtis, Douglas Faries, Susan Robinson, and Joseph Johnston. 2015. "A Literature Review on the Representativeness of Randomized Controlled Trial Samples and Implications for the External Validity of Trial Results." *Trials* 16:495.

Kent, David M, Peter M Rothwell, John PA Ioannidis, Doug G Altman, and Rodney A Hayward. 2010. "Assessing and Reporting Heterogeneity in Treatment Effects in Clinical Trials: A Proposal." *Trials* 11:85.

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

Koenker, Roger. 2005. *Quantile Regression.* Google-Books-ID: WjOdAgAAQBAJ. Cambridge University Press.

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

Lane, Deirdre A., and Gregory Y.H. Lip. 2012. "Use of the CHA2DS2-VASC and HAS-BLED Scores to Aid Decision Making for Thromboprophylaxis in Nonvalvular Atrial Fibrillation." *Circulation* 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. 2016. "Has Choosing Wisely Affected Rates of Dual-Energy X-ray Absorptiometry Use?" *Osteoporosis International* 27 (7): 2311–2316.

Lip, Gregory Y.H., Robby Nieuwlaat, Ron Pisters, Deirdre A. Lane, and Harry J.G.M. Crijns. 2010. "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* 137 (2): 263–272.

Manski, Charles F. 2011. "Choosing Treatment Policies Under Ambiguity." Publisher: Annual Reviews, *Annual Review of Economics* 3 (Volume 3, 2011): 25–49.

———. 2004. "Statistical Treatment Rules for Heterogeneous Populations." _Eprint: https://onlinelibrary.wiley.com/doi/p 0262.2004.00530.x, *Econometrica* 72 (4): 1221–1246.

Mehta, Rajendra H., Anita Y. Chen, Karen P. Alexander, E. Magnus Ohman, Matthew T. Roe, and Eric D. Peterson. 2015. "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* 131 (11): 980–987.

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

Natarajan, Pradeep, Robin Young, Nathan O. Stitziel, Sandosh Padmanabhan, Usman Baber, Roxana Mehran, Samantha Sartori, et al. 2017. "Polygenic Risk Score Identifies Subgroup With Higher Burden of Atherosclerosis and Greater Relative Benefit From Statin Therapy in the Primary Prevention Setting." Publisher: American Heart Association, *Circulation* 135 (22): 2091–2101.

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. 2017. "Treating Specialty and Outcomes in Newly Diagnosed Atrial Fibrillation: From the TREAT-AF Study." *Journal of the American College of Cardiology* 70 (1): 78–86.

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

Pisters, Ron, Deirdre A. Lane, Robby Nieuwlaat, Cees B. De Vos, Harry J.G.M. Crijns, and Gregory Y.H. Lip. 2010. "A Novel User-Friendly Score (HAS-BLED) to Assess 1-year Risk of Major Bleeding in Patients with Atrial Fibrillation: The Euro Heart Survey." *Chest* 138 (5): 1093–1100.

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

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

Rothwell, Peter M. 2005. "External Validity of Randomised Controlled Trials: To Whom Do the Results of This Trial Apply?â." Publisher: Elsevier, *The Lancet* 365 (9453): 82–93.

Savage, L. J. 1951. "The Theory of Statistical Decision." Publisher: Taylor & Francis Group, *Journal of the American Statistical Association.*

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

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

Stroke Prevention in Atrial Fibrillation Investigators. 1995. "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* 5 (3): 147–157.

Surowiecki, James. 2005. *The Wisdom of Crowds.* Reprint edition. Westminster: Vintage.

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

Tinetti, Mary E., Sidney T. Bogardus, and Joseph V. Agostini. 2004. "Potential pitfalls of disease-specific guidelines for patients with multiple conditions." *The New England Journal of Medicine* 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. 2013. "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* 165 (1): 93–101.

US Preventive Services Task Force. 2016. "Statin Use for the Primary Prevention of Cardiovascular Disease in Adults: US Preventive Services Task Force Recommendation Statement." *JAMA* 316 (19): 1997–2007.

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

Wager, Stefan, and Susan Athey. 2017. "Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests." *Journal of the American Statistical Association* 0 (ja): 0–0. eprint: https://doi.org/10.1080/01621459.2017.1319839.

Walraven, Carl van, Robert G. Hart, Stuart Connolly, Peter C. Austin, Jonathan Mant, F.D. Richard Hobbs, Peter J. Koudstaal, et al. 2009. "Effect of Age on Stroke Prevention Therapy in Patients with Atrial Fibrillation: The Atrial Fibrillation Investigators." *Stroke* 40 (4): 1410–1416.

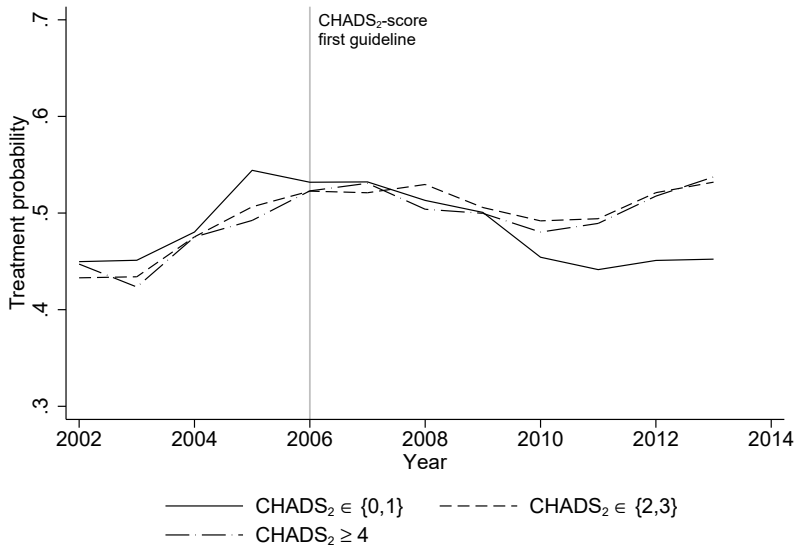
Walraven, Carl van, Robert G Hart, Daniel E Singer, Andreas Laupacis, Stuart Connolly, Palle Petersen, Peter J Koudstaal, Yuchiao Chang, and Beppie Hellemons. 2002. "Oral Anticoagulants vs Aspirin in Nonvalvular Atrial Fibrillation: An Individual Patient Meta-Analysis." *JAMA* 288 (19): 2441–2448.

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

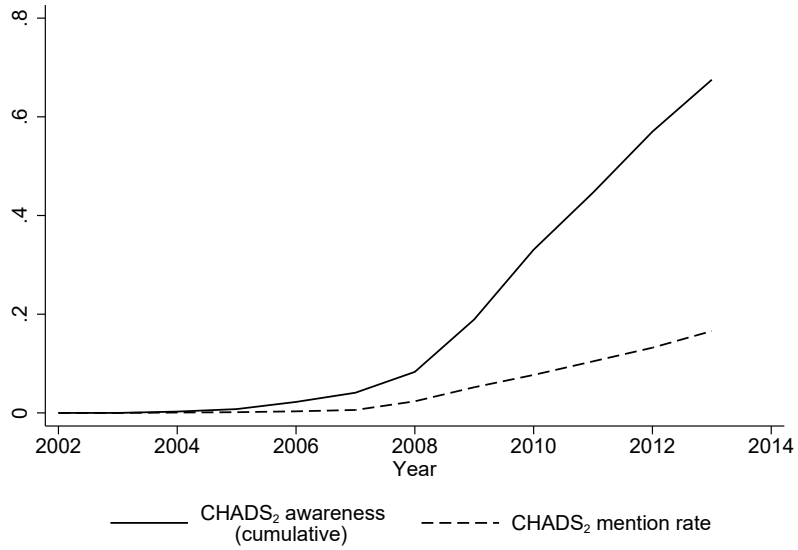
Younas, Neelam, Amjad Ali, Hafsa Hina, Muhammad Hamraz, Zardad Khan, and Saeed Aldahmani. 2022. "Optimal Causal Decision Trees Ensemble for Improved Prediction and Causal Inference." *IEEE Access* 10:13000–13011.

Figure 1: Time Trends in Treatment Decisions and Guideline Awareness

A. Treatment Decisions Around Guideline Introduction

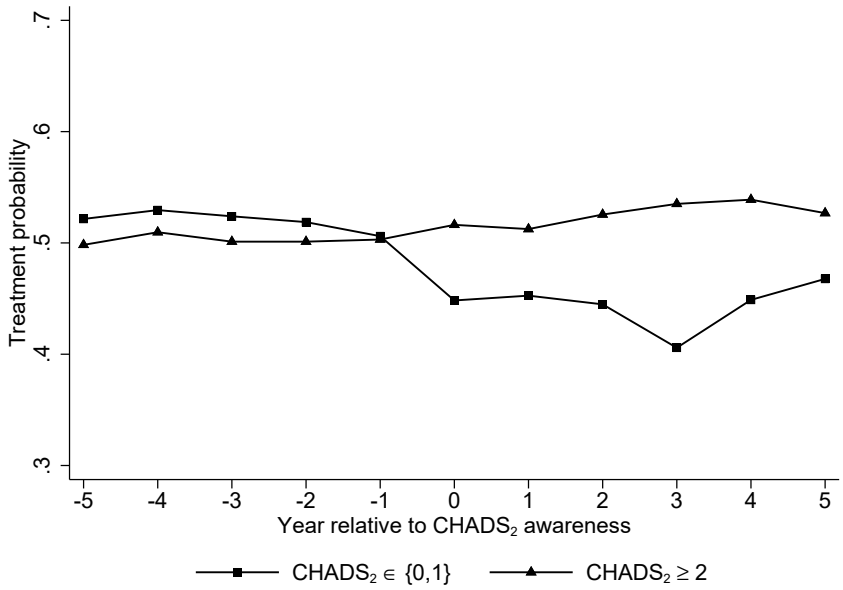


B. Guideline Awareness Over Time



Notes: This figure shows physician treatment decisions relative to the $CHADS_2$ score and changing awareness of the $CHADS_2$ score. Treatments are defined as anticoagulation prescriptions within 90 days of the initial AF diagnosis. Panel A shows the fraction of AF 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. Panel B 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.

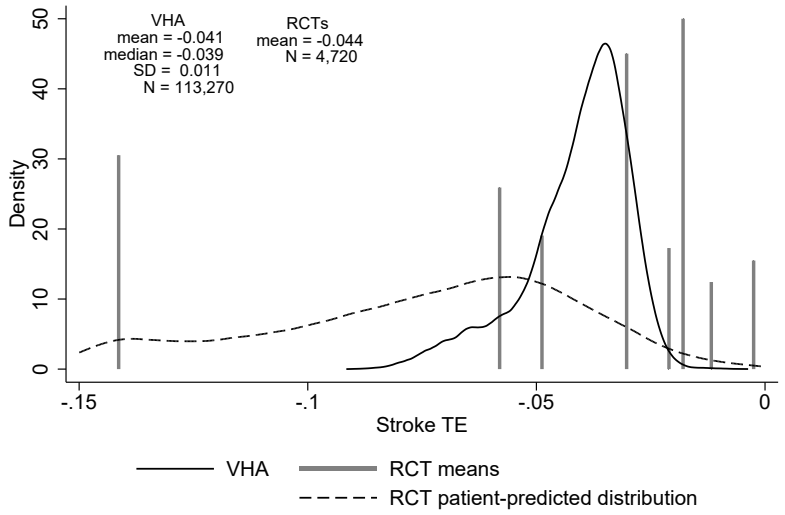
Figure 2: Changes in Treatment After Guideline Awareness



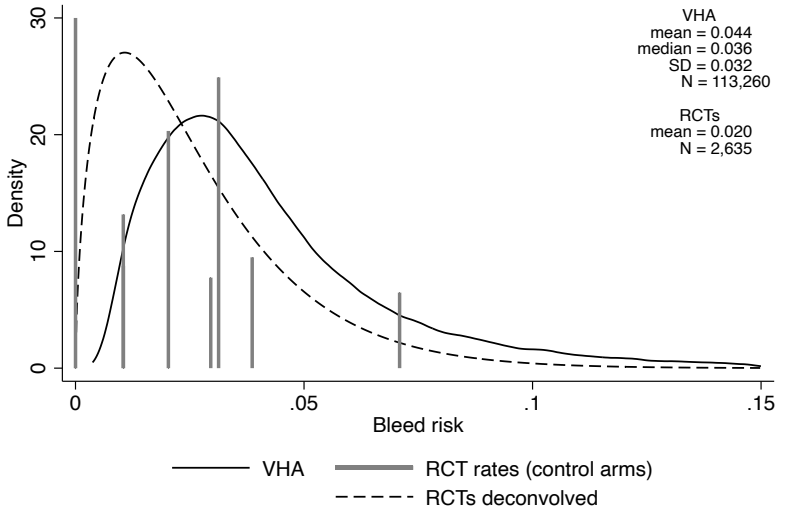
Notes: This figure shows the relationship between $CHADS_2$ score awareness and treatment decisions. It displays the fraction of patients treated with anticoagulation in each year relative to the year of first mentioning the $CHADS_2$ score for physicians who eventually become aware the $CHADS_2$ score. 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

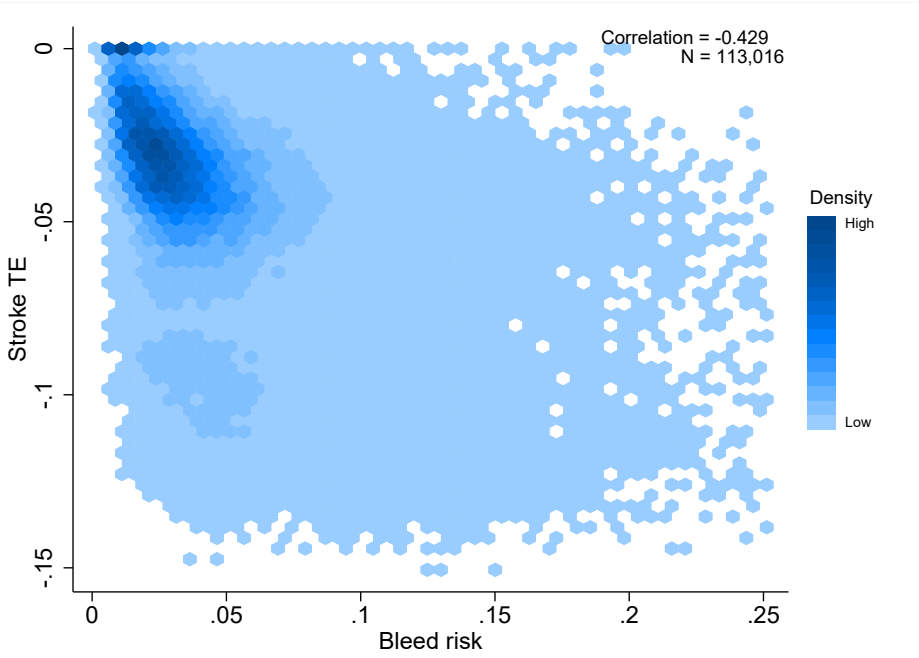


B. Bleed Risks



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 3. 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.3.1. 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.

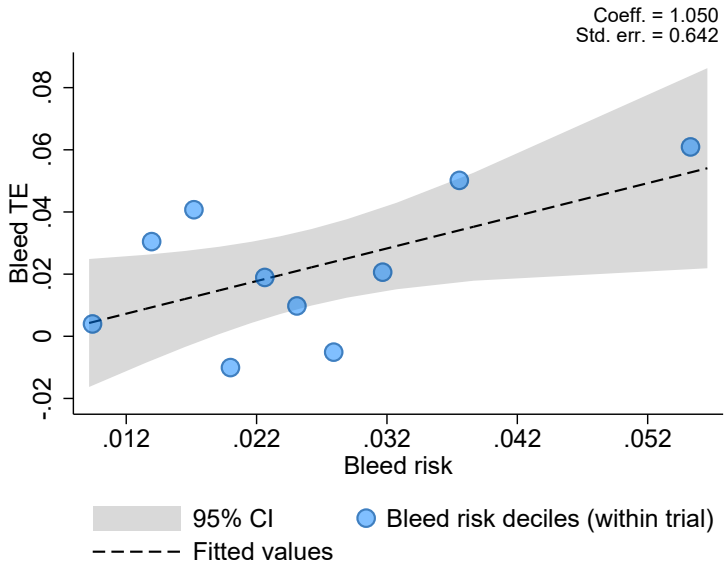
Figure 4: Joint Distribution



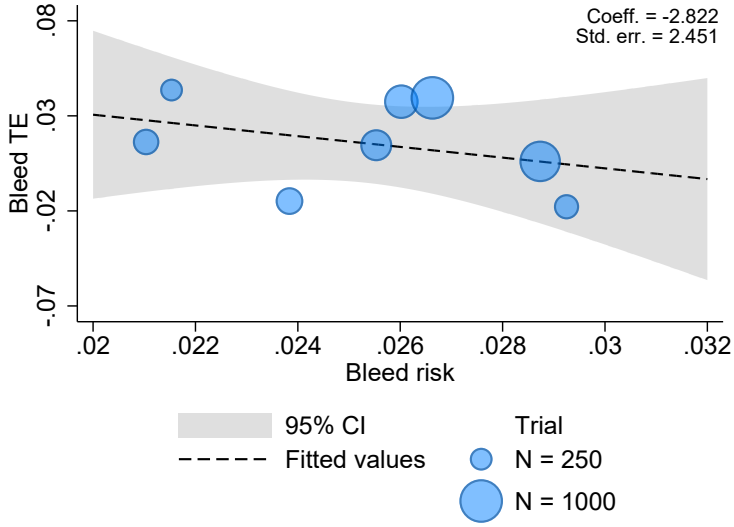
Notes: This figure displays a heat map showing the relationship between stroke treatment effects and bleed risk in the VHA sample. Stroke treatment effects are estimated from the RCT data as conditional average treatment effect (CATE) predictions are trained and validated by using causal-forest methods, described in Section 3. 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.3.1.

Figure 5: Trial Bleed Risks and Treatment Effects

A. Within-Trial Treatment Effects



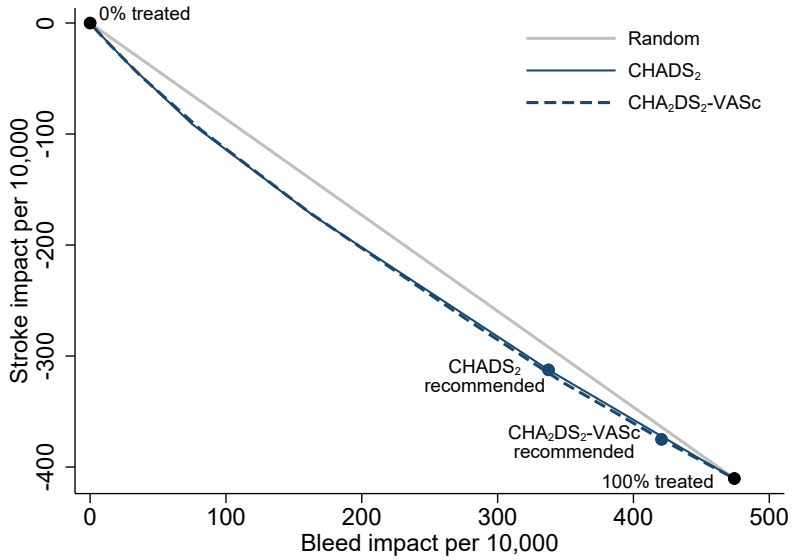
B. Across-Trial Treatment Effects



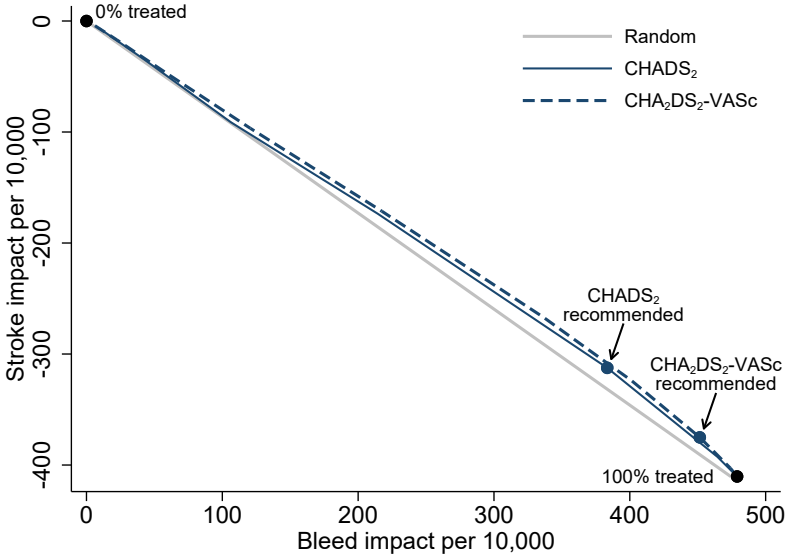
Notes: This figure shows the relationship between bleed risks and bleed treatment effects across and within trials in the RCT database. Bleed risks are calculated as the probability of bleeding in the control arm of each trial. Bleed treatment effects are calculated as the difference in the probability 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. Mean-Independent Harms

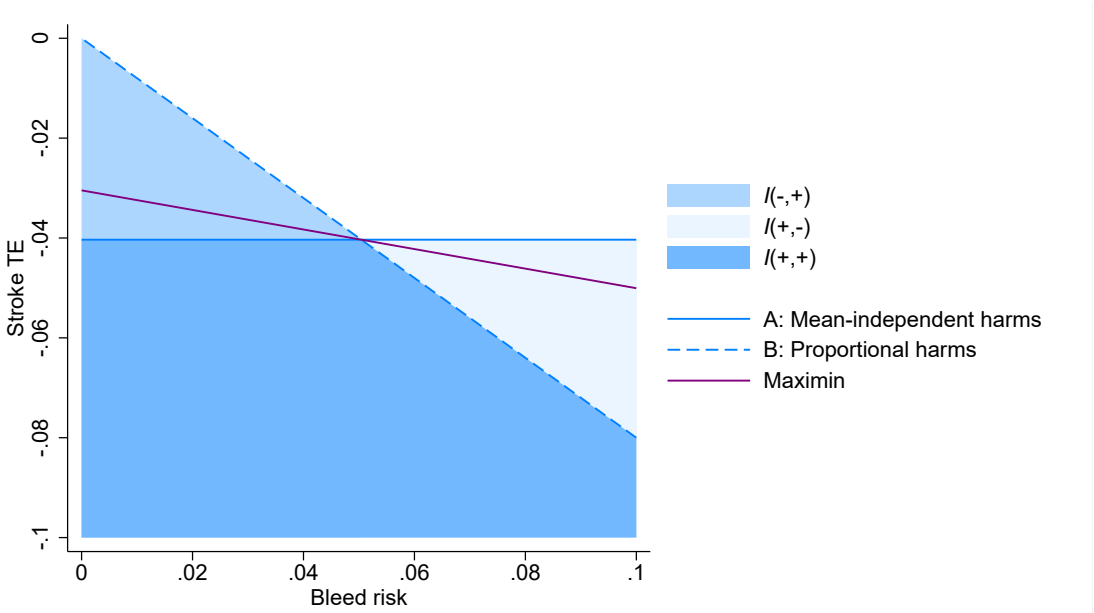


B. Proportional Harms



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 an ordering of patients, and outcomes depend on the ordering of patients and the share of patients treated under the guideline. Each line represents an ordering 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 bleed treatment effects independent of stroke treatment effects. Panel B shows outcomes under the assumption of bleed treatment effects proportional to bleed risks. For reference, outcomes under random treatment are shown as a straight gray line.

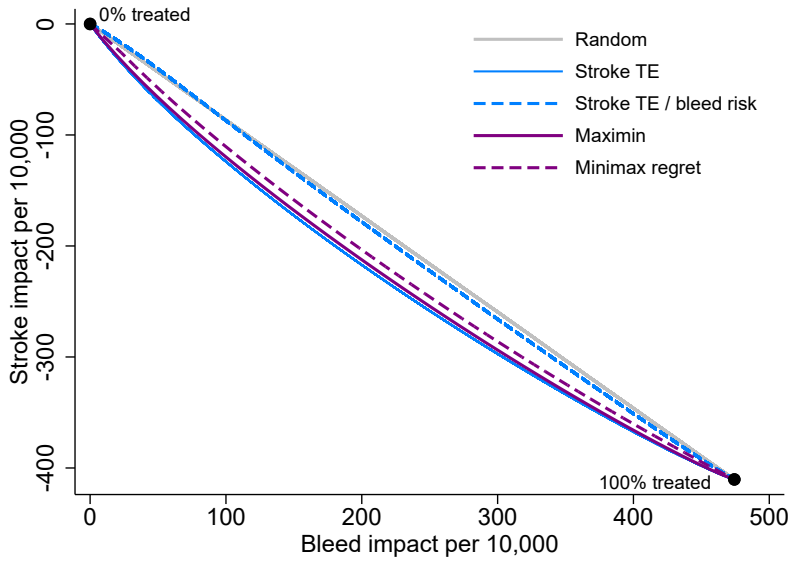
Figure 7: Optimal Treatment Rules in Benefit-Risk Space



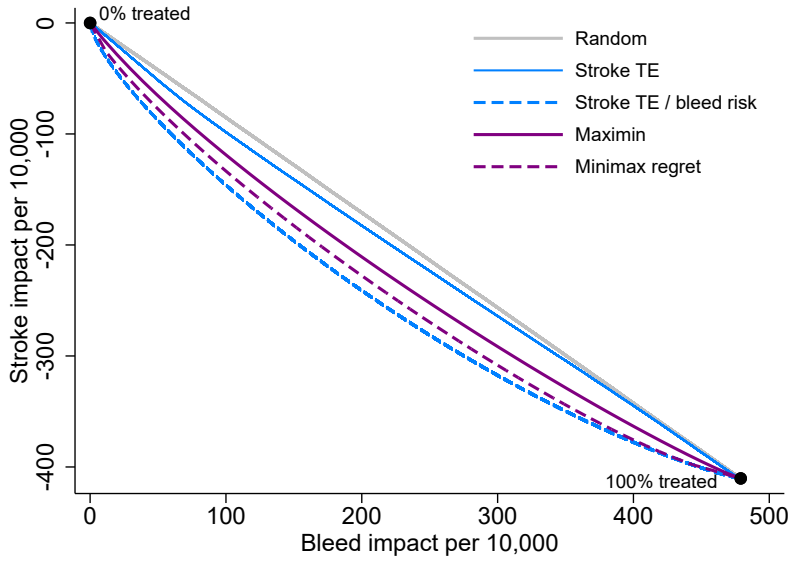
Notes: This figure represents optimal guidelines in the space of stroke-reduction benefits (i.e., treatment effects) and bleed risks. Patients fall within this space via a joint distribution, shown in Figure 4. The underlying joint distribution for this figure has an average bleed treatment effect of $\overline{\alpha}^b = 0.05$. 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 the assumption on bleed harms. $\mathcal{I}(+,-)$ and $\mathcal{I}(-,+)$ are the spaces including patients who should only be treated under independent bleed harms and proportional bleed harms, respectively. The solid blue and dashed blue lines indicate treatment rules for marginal patients treated under independent and proportional bleed harms, respectively. The purple line represents a possible maximin treatment threshold in which average bleed risk conditional on treatment equals the average (unconditional) bleed risk. Further details of the theory are discussed in Appendix A.4.

Figure 8: Outcomes Following Optimal Guidelines

A. Mean-Independent Harms

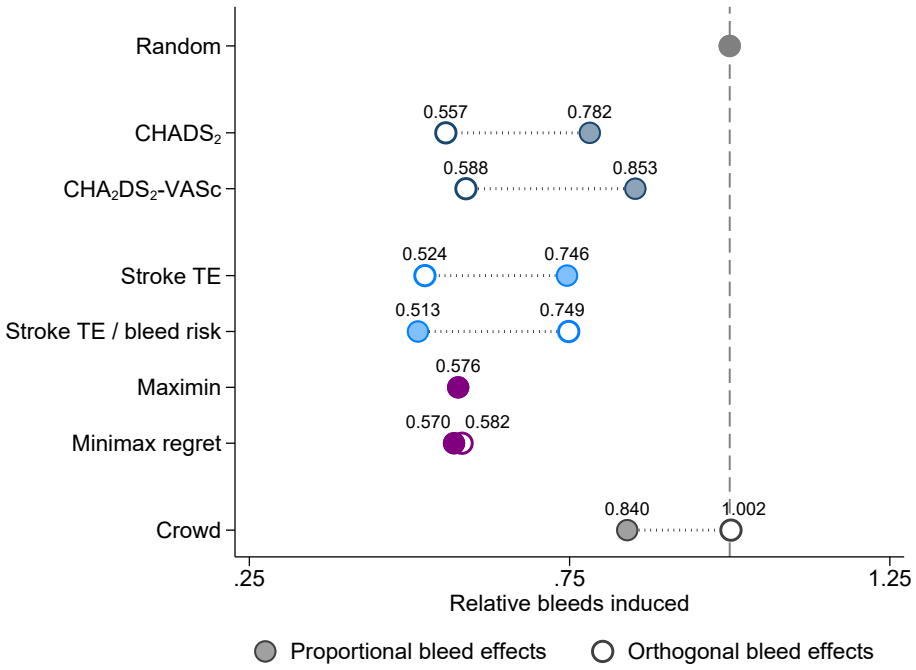


B. Proportional Harms



Notes: This figure shows outcomes in terms of strokes prevented and bleeds induced by anticoagulation under different optimal treatment rules. Treatment rules imply orderings of patients, and outcomes depend on the ordering of patients and the share of patients treated under the treatment rule. Each line represents an ordering 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 bleed treatment effects independent of stroke treatment effects. Panel B shows outcomes under the assumption of bleed treatment effects proportional to bleed risks. For reference, outcomes under random treatment are shown as a straight gray line.

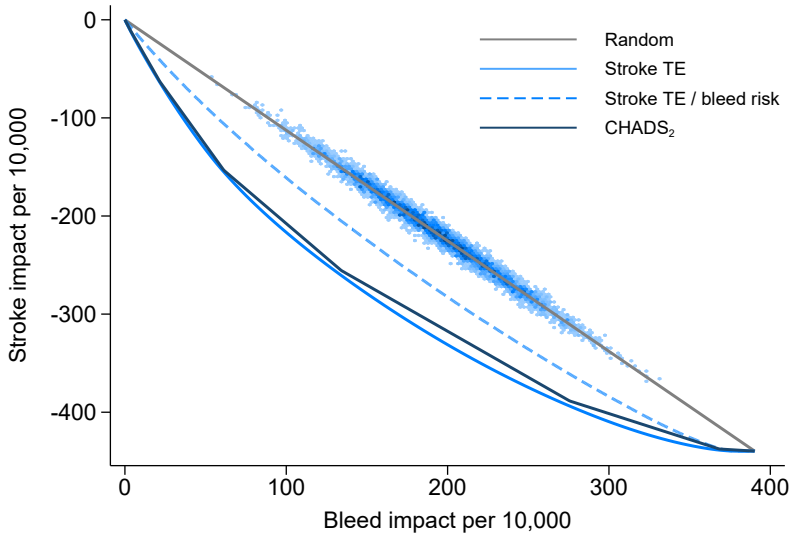
Figure 9: Performance Relative to Random Treatment



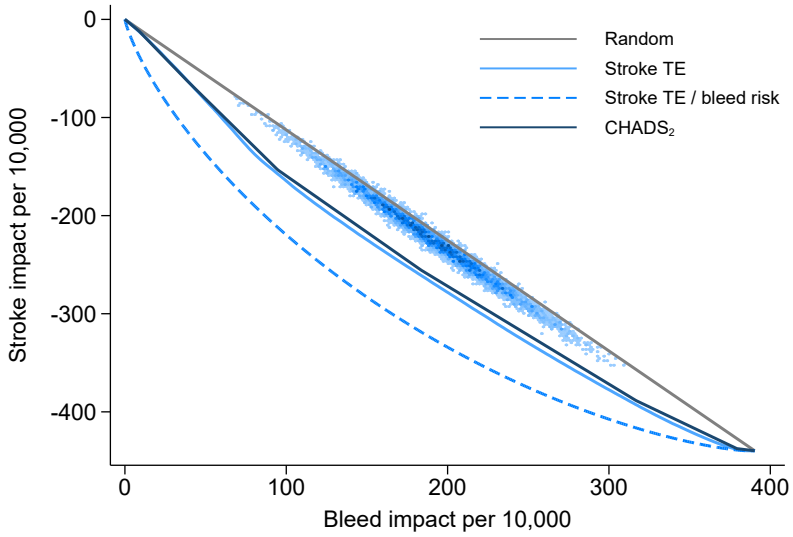
Notes: This figure illustrates summarizes outcomes under random treatment, known guidelines, optimal guidelines, and treatment rules implied by "wisdom of the crowds". All outcomes hold fixed strokes prevented at the same level as treating 50% of patients randomly. Bleeds induced are shown relative to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

Figure 10: Variation in Physician Outcomes

A. Mean-Independent Harms



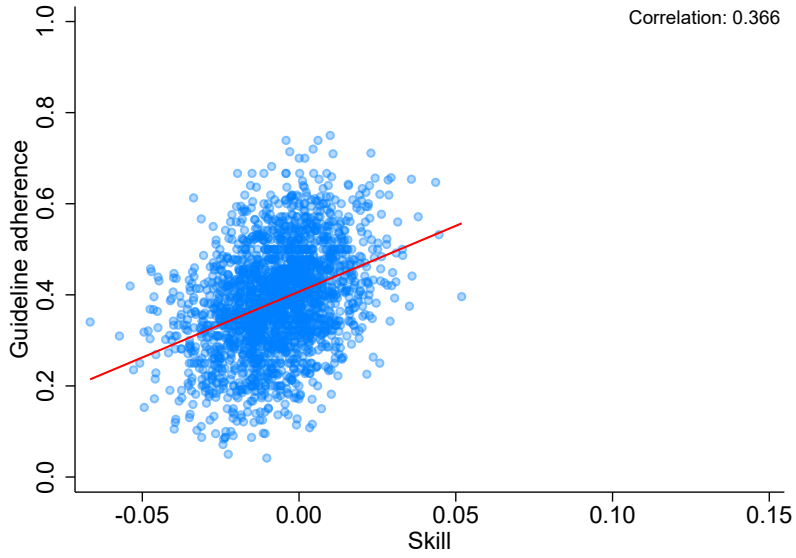
B. Proportional Harms



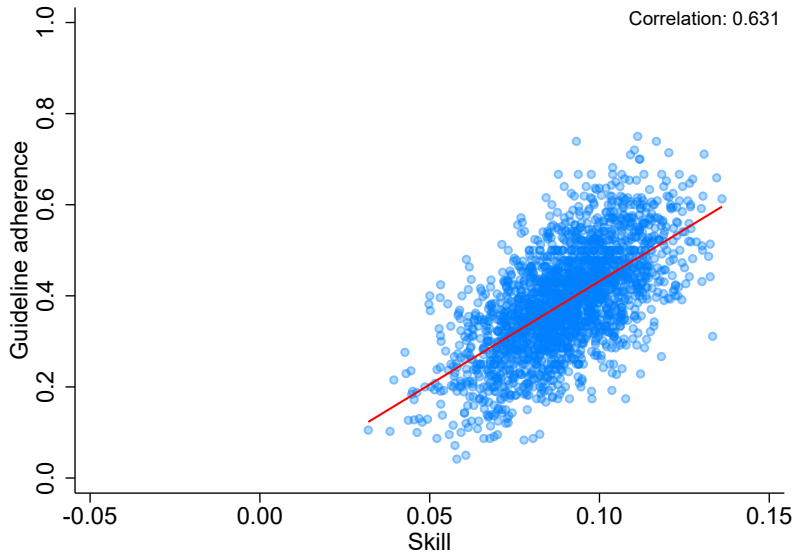
Notes:

Figure 11: Physician Guideline Adherence and Skill

A. Mean-Independent Harms



B. Proportional Harms



Notes:

Table 1: $CHADS_2$ and CHA_2DS_2-VASc Guidelines

Guideline	Risk factor	Points	Treatment recommendations
$CHADS_2$	**C**ongestive heart failure / LV dysfunction	1	0: no treatment; 1: aspirin or OAC ≥ 2: OAC
	Hypertension	1	
	Age ≥ 75 years	1	
	Diabetes	1	
	Stroke / TIA / thromboembolism history	2	
CHA_2DS_2-VASc	**C**ongestive heart failure / LV dysfunction	1	0: no treatment; 1: aspirin or OAC; ≥ 2: OAC
	Hypertension	1	
	Age ≥ 75 years	2	
	Diabetes	1	
	Stroke / TIA / thromboembolism history	2	
	Vascular disease (prior MI, PAD, aortic plaque)	1	
	Age ≥ 65 years (but < 75 years)	1	
	Sex **c**ategory female	1	

Notes: This table describes the risk scoring system and treatment recommendations implied by the $CHADS_2$ and CHA_2DS_2-VASc guidelines. The original $CHADS_2$ guideline was developed by Gage et al. (2001) and Gage et al. (2004). Treatment recommendations initially placed patients with a score of 2 in the intermediate group (consider aspirin or warfarin); they were later revised by Fang et al. (2008) too place these patients in the warfarin treatment group. The CHA_2DS_2-VASc guideline was developed by Lip et al. (2010). Table 1 of that paper describes other clinical guidelines and risk-stratification schemes for anticoagulation in AF. Abbreviations: LV = left ventricular; TIA = transient ischemic attack; OAC = oral anticoagulant (i.e., warfarin).

Table 2: 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 A.1. 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.

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

Jason Abaluck
Leila Agha
David Chan
Daniel Singer
Diana Zhu

December 13, 2025

A.1 Details of Event Study of Awareness

This appendix provides details of our descriptive analysis of prescription choice following awareness of the $CHADS_2$ score. We assess adherence by focusing on treatment choices for AF patients according to their $CHADS_2$ score: those with a score of 0 or 1, for whom anticoagulation is not necessarily recommended, and those with a score of at least 2, for whom anticoagulation is recommended.

Specifically, we estimate the following regression:

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

where $D_i \in \{0,1\}$ indicates whether patient i was treated with anticoagulation, $r(i)$ is a function that returns the year of i's visit relative to the prescribing physician's first mention (i.e., becoming aware) of the $CHADS_2$ score, $\eta_{j(i)}$ is a physician fixed effect, and $\xi_{t(i)}$ is a year fixed effect. We consider physicians who eventually mention the $CHADS_2$ score as being treated and physicians who never mention the score as being controls in this event-study setup. We estimate Equation (A.1) separately for patients with $CHADS_2$ scores of 0 and 1 and for those with scores 2 and higher.

Figure A.1 displays estimation results of the θ_r coefficients and echoes the patterns shown in raw treatment rates displayed in Figure 2. 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.[1]

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 and persistent 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 Panel A of Figure 2, reducing concern about both false-positive and false-negative detection. 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 and rose by 3 percentage points for high-score patients.

A.2 Machine Learning Procedure

We estimate conditional average treatment effects (CATEs) using an ensemble machine learning (ML) method that combines LASSO (least absolute shrinkage and selection operator) and causal-forest predictions. We combine the predictions from LASSO and causal forest in what amounts to equally weighting the two predictions. In order to obtain consistent estimates of CATEs, we utilize a "best linear predictor" approach to validate the ML predictions to be "proxy" predictions to be validated in hold-out trials.

1. 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.

A.2.1 ML Predictions

Lasso Prediction. We leverage LASSO to estimate predictions based on 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.2, 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.2 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:

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

$\tilde{Y}_i^o$ 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.2) 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.2) and rerun a post-LASSO OLS regression:

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

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.5, 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 $\tilde{X}_i^{o,\text{lasso}}$ and $\widetilde{XW}_i^{o,\text{lasso}}$: the LASSO-predicted risk score and its interaction with W_i. This prediction can be written as $\tilde{X}_i^{o,\text{lasso}}\hat{\alpha}^o_{\text{post-lasso}} \odot I_{\text{cont}} + \widetilde{XW}_i^{o,\text{lasso}}\hat{\tau}^o_{\text{post-lasso}} \odot I_{\text{cont}}$, 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.[2]

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 $X_i^{o,\text{cf}}$. Stating the causal-forest CATE estimand from Equation (5) of Athey and Wager (2019), we have

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

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

We use the `grf` package developed by Tibshirani, Athey, and Wager (2020) to run causal and regression forests. Details about the algorithm can be found at `https://github.com/grf-labs/grf`. In brief, both causal and regression forests form predictions by creating 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

2. Following Section 6.1.1 of Athey, Tibshirani, and Wager (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.

(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 $\hat{\tau}^o\left(X_i\right) = \hat{\tau}^o_{\text{lasso}}\left(X_i^{o,\text{lasso}}\right) + \hat{\tau}^o_{\text{cf}}\left(X_i^{o,\text{cf}}\right)$.

Appendix Table A.5 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.[3] 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-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.2.2 Validation Across RCTs

We validate the estimated heterogeneity across trials, following best practice to assess potential overfitting. Specifically, borrowing from an approach developed by Chernozhukov et al. (2018) and applied in Athey and Wager (2019), we test whether CATE predictions estimated in seven of the eight trials can predict stroke and bleed treatment effects in the hold-out trial.

To implement this approach, we will validate the ML CATE "proxy" predictions described in the section above in a hold-out trial in the AFI database that was not used to construct the CATE proxy. In other words, we use CATE proxy predictions constructed exclusively using data in the seven *other* trials besides the trial used for validation. Denote these leave-out-trial CATE proxy predictions for stroke and bleeding as $\hat{\tau}^o_{-j(i)}\left(x\right)$, $o \in \{s, b\}$, for individual i in trial $j(i)$ with characteristics $X_i = x$ (we

3. There are several methods to compute variable importance, and there is no clear consensus yet on the best method (Wei, Lu, and Song 2015). We use a measure from Athey, Tibshirani, and Wager (2019) which ranks variables more highly in importance if the algorithm chooses to split trees in the forest earlier on those variables.

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.[4]

Validating Predictions. To assess the validity of predictions across trials, we estimate a "best linear predictor" (BLP) regression of realized outcomes Y_i^o on treatment D_i interacted with demeaned CATE proxy predictions, controlling for trial fixed effects ζ_j^o and treatment probability within each trial $P_j \equiv \Pr(D_i|j(i)=j)$ similarly interacted with demeaned CATE proxy predictions:

$$\begin{aligned} Y_i^o &= \left[\delta_0^o + \delta_1^o\left(\hat{\tau}^o_{-j}(X_i) - \overline{\tau}^o\right)\right] W_i + \\ &\quad \varphi_1^o\left(\hat{\tau}^o_{-j}(X_i) - \overline{\tau}^o\right) P_j + \varphi_2^o \hat{Y}^o_{-j}(X_i) + \zeta_j^o + \varepsilon_i, \end{aligned} \tag{A.3}$$

where $\overline{\tau}^o \equiv \sum_i \hat{\tau}^o_{-j}(X_i)$ is the mean CATE prediction.[5] The coefficient δ_1^o quantifies the predictive power of heterogeneous CATEs that are estimated in other trials. A coefficient value that is statistically indistinguishable from $\delta_1^o = 0$ would suggest that the CATEs do not robustly predict variation in treatment effects in the hold-out trial; a coefficient value of $\delta_1^o = 1$ 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 suggest the machine learning procedure may have over-fit the data. The re-scaled estimates $\delta_1^o \hat{\tau}^o_{-j}(x)$ form the best linear predictors of the CATEs, as discussed in Chernozhukov et al. (2018).

Estimating Equation (A.3) for strokes, we obtain $\hat{\delta}_1^s = 0.520$ (standard error of 0.130) (Appendix Table A.8, Column 1). This result suggests that the ML CATE proxies are strongly predictive of outcomes in hold-out trials, though the CATE proxies may still be over-fit or have statistical noise. For bleeds, however, the BLP regression suggests little reliably estimated variation in bleed CATE proxies across observable covariates (Appendix Table A.8, Column 4): The BLP coefficient is wrong-signed ($\hat{\delta}_1^b = -0.374$) and statistically insignificant.

Decomposing Benefits into $CHADS_2$ and Residual Components. 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 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

4. 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.

5. 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 is the leave-out CATE proxy predictions, $\hat{\tau}^o_{-j(i)}(x)$, 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 $\hat{\tau}^o(x|j)$.

$\hat{\tau}^s_{-j}(x) = \hat{\tau}^c_{-j}(x) + \hat{\tau}^r_{-j}(x)$. We use these components to perform the following BLP projection:

$$\begin{aligned} Y_i^s &= \delta_0^s D_i + \sum_{\tilde{o} \in \{s(c), s(r)\}} \left[\delta_1^{\tilde{o}} \left(\hat{\tau}^{\tilde{o}}_{-j}(X_i) - \overline{\tau}^{\tilde{o}} \right) W_i + \varphi_1^{\tilde{o}} \left(\hat{\tau}^{\tilde{o}}_{-j}(X_i) - \overline{\tau}^{\tilde{o}} \right) P_j \right] + \\ & \varphi_2 \hat{Y}^s_{-j}(X_i) + \zeta_j^s + \varepsilon_i. \end{aligned} \tag{A.4}$$

Using Equation (A.4), we construct BLP-based adjustments to the stroke CATEs as follows:

$$\hat{\tau}^s_{BLP}(x) = \hat{\delta}_0^s + \hat{\delta}_1^c \hat{\tau}^c(x) + \hat{\delta}_1^r \hat{\tau}^r(x). \tag{A.5}$$

Estimating Equation (A.4), we obtain $\hat{\delta}_1^c = 0.666$ (standard error of 0.250) and $\hat{\delta}_1^r = 0.460$ (standard error of 0.142) (Appendix Table A.8, Column 2). 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 3.2 for stroke by using Equation (A.5) for $\hat{\tau}^s_{BLP}$.

A.3 Variable Construction and Prediction Details

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

A.3.1 Predicted 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 comorbidity set, which include: HIV/AIDS, alcohol abuse, blood loss anemia, cardiac arrhythmia, congestive heart failure, chronic pulmonary disease, coagulopathy, deficiency anemia, depression, diabetes (with and without chronic complications), drug abuse, hypertension, hypothyroidism, liver disease, lymphoma, fluid and electrolyte disorders, metastatic 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.3.2 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 AF. 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 an AF history. 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.

A.3.3 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.2. We then take this model to calculate $\Pr_{\text{VHA}}(D_i \mid X_i)$ for each patient in the AFI database.

A.4 Optimal Treatment Rules

A.4.1 Setup and Definitions

Consider a continuum of patients in set $\mathcal{I}$ indexed in the unit interval $[0,1]$. For each patient $i \in \mathcal{I}$, the net expected utility of treatment depends on the benefits and harms of treatment. These are each patient's heterogeneous treatment effects for strokes and bleeds, respectively: $\tau_i^s \in [-1,0]$ and $\tau_i^b \in [0,1]$. The heterogeneous harms are unknown, but we observe heterogeneous bleed risks: $\alpha_i^b \in [0, \alpha_{\max}^b]$.

Under the current information, we consider two canonical assumptions:

Assumption *A* (Independent Harms). *Harms are mean-independent of benefits and risks. That is,* $E_{i\in\mathcal{I}}\left[\tau_i^b \middle| (\tau_i^s, \alpha_i^b) = (\tau^s, \alpha^b)\right] = E_{i\in\mathcal{I}}\left[\tau_i^b\right]$.

Assumption *B* (Proportional Harms). *Harms are proportional to risks. That is,* $E_{i\in\mathcal{I}}\left[\tau_i^b \middle| \alpha_i^b = \alpha^b\right] = \alpha^b \cdot E_{i\in\mathcal{I}}\left[\tau_i^b\right] / E_{i\in\mathcal{I}}\left[\alpha_i^b\right]$.

To simplify notation, we assume (without loss of generality) that $E_{i\in\mathcal{I}}\left[\tau_i^b\right] = E_{i\in\mathcal{I}}\left[\alpha_i^b\right]$ and denote $\overline{\alpha}^b \equiv E_{i\in\mathcal{I}}\left[\alpha_i^b\right]$. Thus, the expected net utilities of treatment under Assumptions A and B, for some utility cost of a bleed relative to a stroke of $\rho > 0$, can respectively be stated as follows:[6]

$$u_A(i;\rho) = -\tau_i^s - \rho\overline{\alpha}^b; \tag{A.6}$$

$$u_B(i;\rho) = -\tau_i^s - \rho\alpha_i^b. \tag{A.7}$$

Considering a treatment rule $D(i) \in \{0,1\}$ as a function mapping each patient i, we can construct a welfare under the treatment rule and bleed cost parameter ρ:

$$W_A(D;\rho) = E_{i\in\mathcal{I}}\left[u_A(i;\rho)\,D(i)\right]; \tag{A.8}$$

$$W_B(D;\rho) = E_{i\in\mathcal{I}}\left[u_B(i;\rho)\,D(i)\right]. \tag{A.9}$$

We can also consider welfare under a mixture of Assumptions A and B. Specifically, if harms are independent (Assumption A) with probability θ and proportional (Assumption B) with probability $1-\theta$, then we may write the expected welfare as

$$W(D;\rho,\theta) = E_{i\in\mathcal{I}}\left[\left(\theta u_A(i;\rho) + (1-\theta)u_B(i;\rho)\right)D(i)\right]. \tag{A.10}$$

Note that $W(D;\rho,\theta) = \theta W_A(D;\rho) + (1-\theta)W_B(D;\rho)$.

The optimal treatment rules under Assumptions A and B treat i if and only if the net expected utility of treatment under the relevant assumption is positive. Under Assumption A, the optimal treatment rule is

$$D_A^*(i;\rho) = \mathbf{1}\left(-\tau_i^s/\overline{\alpha}^b > \rho\right). \tag{A.11}$$

Under Assumption B, the optimal treatment rule is

$$D_B^*(i;\rho) = \mathbf{1}\left(-\tau_i^s/\alpha_i^b > \rho\right). \tag{A.12}$$

Similarly, if θ is known, the optimal treatment rule would be

$$D_\theta^*(i;\rho) = \mathbf{1}\left(\frac{-\tau_i^s}{\theta\overline{\alpha}^b + (1-\theta)\alpha_i^b} > \rho\right). \tag{A.13}$$

In the remainder of this appendix, we will consider the possibility that θ is unknown and may vary in the interval $[0,1]$, without loss of generality. We will analyze the following two standards for optimality under scientific uncertainty about θ.

Definition. The *maximin* treatment rule maximizes the lowest welfare $W(D;\rho,\theta)$ among any $\theta \in$

6. The assumption that $E_{i\in\mathcal{I}}\left[\tau_i^b\right] = E_{i\in\mathcal{I}}\left[\alpha_i^b\right]$ is without loss of generality. If $E_{i\in\mathcal{I}}\left[\tau_i^b\right] \neq E_{i\in\mathcal{I}}\left[\alpha_i^b\right]$, we can interpret ρ as the utility cost of a bleed relative to a stroke, scaled by $E_{i\in\mathcal{I}}\left[\tau_i^b\right]/E_{i\in\mathcal{I}}\left[\alpha_i^b\right]$.

$[0,1]$:

$$D^*_{\max}(\cdot;\rho) = \arg\max_D \min_{\theta\in[0,1]} W(D;\rho,\theta). \tag{A.14}$$

Definition. The *minimax-regret* treatment rule minimizes the greatest regret between the optimal welfare and the realized welfare, i.e., $\max_{D'} W(D';\rho,\theta) - W(D;\rho,\theta)$, among any $\theta \in [0,1]$:

$$D^*_{\text{mr}}(\cdot;\rho) = \arg\min_D \max_{\theta\in[0,1]} \left\{ \max_{D'} W(D';\rho,\theta) - W(D;\rho,\theta) \right\}. \tag{A.15}$$

A.4.2 Ordering-Threshold Formulation of Treatment Rules

We can define treatment rules as a combination of an ordering of patients and a threshold below which ordered patients are treated. Beginning here, we will (temporarily) hold ρ fixed and suspend reference to it in notation.

Definition. An *ordering* is a bijective mapping $\Omega : \mathcal{I} \to [0,1]$.

The ordering allows us to define a treatment rule with a scalar threshold π.

Definition. For a given ordering Ω, a *threshold* π defines a treatment rule D such that $D(i) = \mathbf{1}(\Omega(i) \leq \pi)$.

A given ordering Ω gives rise to a continuum of treatment rules depending on the threshold $\pi \in [0,1]$, where π can be interpreted as the measure of patients treated. While the combination (Ω,π) defines a single treatment rule, an infinite number of orderings may yield D. Specifically, a treatment rule defined by (Ω,π) can be alternatively defined by any $\left(\tilde{\Omega},\pi\right)$, where $\tilde{\Omega} : \{i : D(i) = 1\} \to [0,\pi]$ and $\tilde{\Omega} : \{i : D(i) = 0\} \to (\pi,1]$.

Using this formulation, we can restate welfare under Assumption $\omega \in \{A,B\}$ as

$$W_\omega(D(\Omega,\pi)) = \int_0^\pi u_\omega(i)\, d\Omega(i). \tag{A.16}$$

We proceed by declaring a set of axioms for welfare, which should hold under the definition of Lebesque integrals and an always finite utility function $u_\omega \in \{u_A, u_B\}$, as defined by Equations (A.6) and (A.7).

Axiom 1. *Consider an ordering Ω and a set of treatment rules defined by $D(\Omega,\pi)$. Then $W(D)$ is continuous in π.*

Axiom 2. *Consider any treatment rules D_1 and D_2, such that $D_1(i) = D_2(i)$ for all $i \in \mathcal{I}$. Then $W(D_1) = W(D_2)$.*

Axiom 3. *Consider any finite number K treatment rules $D_1,\ldots,D_K$, such that $D(i) \equiv \sum_{k=1}^K D_k(i) \in \{0,1\}$ for all $i \in \mathcal{I}$. Then $W(D) = \sum_{k=1}^K W(D_k)$.*

Axiom 1 allows for convenient equality conditions based on welfare for optimal treatment rules in the next subsection. Axiom 2 allows us to state welfare equivalently for any reordering of patients as long as the treatment rule remains unchanged. Axiom 3 allows us to break welfare into component parts from different subsets of treated patients, as long as the union of treated patients remains the same.

Definition. A *subordering* is a bijective mapping $\Omega : \mathcal{I}' \to [0,1]$, where $\mathcal{I}' \subset \mathcal{I}$. The subordering and threshold π defines a treatment rule D such that

$$D(i) = \begin{cases} \mathbf{1}(\Omega(i) \leq \pi), & i \in \text{dom}(\Omega), \\ 0, & i \notin \text{dom}(\Omega). \end{cases}$$

With the above definition, we may also use treatment rules defined by orderings within subsets, particularly $\mathcal{I}_{+,-}$ and $\mathcal{I}_{-,+}$.[7]

A.4.3 Equality Conditions for Uncertainty-Robust Optimality

Relying on assumption-specific optimal treatment rules in Equations (A.11) and (A.12), we divide $\mathcal{I}$ into four subsets:

$$\begin{aligned} \mathcal{I}_{+,+} &= \left\{i : D_A^*(i) = 1, D_B^*(i) = 1\right\}; \\ \mathcal{I}_{+,-} &= \left\{i : D_A^*(i) = 1, D_B^*(i) = 0\right\}; \\ \mathcal{I}_{-,+} &= \left\{i : D_A^*(i) = 0, D_B^*(i) = 1\right\}; \\ \mathcal{I}_{-,-} &= \left\{i : D_A^*(i) = 0, D_B^*(i) = 0\right\}. \end{aligned}$$

Figure 7 shows these four regions in the joint space of stroke benefits and bleed risks.

It is clear that any optimal rule must treat all patients in $i \in \mathcal{I}_{+,+}$.

Remark 1. $D_{\max}^*(i) = D_{\text{mr}}^*(i) = 1$ if $i \in \mathcal{I}_{+,+}$; $D_{\max}^*(i) = D_{\text{mr}}^*(i) = 0$ if $i \in \mathcal{I}_{-,-}$.

Treating any $i \in \mathcal{I}_{+,+}$ raises welfare regardless of the assumption on harms. Similarly, treating any $i \in \mathcal{I}_{-,-}$ lowers welfare regardless of the assumption on harms. If $\mathcal{I}_{+,-}$ and $\mathcal{I}_{-,+}$ are both null, then the maximin and minimax-regret solutions are trivial as there are no patients over which D_A^* and D_B^* disagree.[8] For the remainder of this appendix, we assume that $\mathcal{I}_{+,-}$ and $\mathcal{I}_{-,+}$ are never null.

Assumption 1. $\mathcal{I}_{+,-} \neq \varnothing$; $\mathcal{I}_{-,+} \neq \varnothing$.

7. Specifically, consider $\Omega_{+,-} : \mathcal{I}_{+,-} \to [0,1]$ and $\pi_{+,-}$, which can define a treatment rule within $\mathcal{I}_{+,-}$. The component treatment rule would be $D\left(i; \Omega_{+,-}, \pi_{+,-}\right) = \mathbf{1}\left(\Omega_{+,-}(i) \leq \pi_{+,-}\right)$; note that $D\left(i; \Omega_{+,-}, \pi_{+,-}\right) = 0$ for any $i \notin \mathcal{I}_{+,-}$. We can define a similar subordering and threshold within $\mathcal{I}_{+,-}$: $\Omega_{+,-} : \mathcal{I}_{+,-} \to [0,1]$ and $\pi_{+,-}$. An overall treatment rule can then be stated, for example, as $D(i) = D\left(i; \Omega_{+,-}, \pi_{+,-}\right) + D\left(i; \Omega_{-,+}, \pi_{-,+}\right) + D_{+,+}(i)$. Note that welfare under a treatment rule defined by a subordering would be modified from Equation (A.16) to $W_\omega(D(\Omega, \pi)) = \Pr(i \in \text{dom}(\Omega)) \int_0^\pi u_\omega(i) d\Omega(i)$.

8. If $\mathcal{I}_{+,-}$ and $\mathcal{I}_{-,+}$ are both null, then $W_A(D_A^*) = W_A(D_B^*)$ and $W_B(D_A^*) = W_B(D_B^*)$. This implies that $\min(W_A, W_B)$ takes the same value at D_A^* and D_B^*. There would also be no regret. Reflecting that D_A^* and D_B^* are equivalent treatment rules, they would both satisfy the maximin and minimax-regret criteria.

We next note that the maximin solution may be the same as D_A^* or D_B^*. These may be considered as corner solutions in optimizing the maximin criterion in Equation (A.14) in that $W_A(D_{\max}^*) \neq W_B(D_{\max}^*)$.

Remark 2. If $W_A\left(D_A^*\right) < W_B\left(D_A^*\right)$, then $D_{\max}^* = D_A^*$. If $W_A\left(D_B^*\right) > W_B\left(D_B^*\right)$, then $D_{\max}^* = D_B^*$.

If $W_A(D_A^*) < W_B(D_A^*)$, then by the definition of D_A^*, $W_A(D) \leq W_A(D_A^*) < W_B(D_A^*)$ for all D. The maximin objective $\min(W_A, W_B) = W_A$ is maximized at $D_{\max}^* = D_A^*$. By similar reasoning, if $W_B(D_B^*) < W_A(D_B^*)$, then $D_{\max}^* = D_B^*$.

We then ask when an uncertainty-robust optimal treatment rule is different from D_A^* or D_B^*.

Definition. An *interior* optimal treatment rule is a solution D^* to Equation (A.14) or (A.15), such that $D^* \neq D_A^*$ and $D^* \neq D_B^*$.

Proposition 1 (Maximin Equality of Welfare). *If $W_A\left(D_A^*\right) > W_B\left(D_A^*\right)$ and $W_A\left(D_B^*\right) < W_B\left(D_B^*\right)$, then the maximin treatment rule D_{max}^*, as defined by Equation (A.14), is an interior treatment rule that satisfies $W_A\left(D_{max}^*\right) = W_B\left(D_{max}^*\right)$.*

Proof. Note that $\min_{\theta\in[0,1]} W(D;\theta) = \min_{\omega\in\{A,B\}} W_\omega(D)$. So we can focus on maximizing the latter minimum between $W_A(D)$ and $W_B(D)$ instead of the more general minimum over $\{W(D;\theta)\}_{\theta\in[0,1]}$.

Consider arbitrary suborderings $\Omega_{+,-} : \mathcal{I}_{+,-} \to [0,1]$ and $\Omega_{-,+} : \mathcal{I}_{-,+} \to [0,1]$. By Remark 1 and Axioms 2 and 3, we may use these suborderings to construct a continuum of treatment rules between D_A^* and D_B^*.

Suppose $D_{\max}^* = D_A^*$. $\Omega_{+,-}$ and $\Omega_{-,+}$ define possible treatment rules between D_A^* and D_B^*. By Axiom 1, $W_B(D)$ strictly and continuously increases from $W_B(D_A^*)$ to $W_B(D_B^*)$, while the welfare $W_A(D)$ strictly and continuously decreases from $W_B(D_A^*)$ to $W_B(D_B^*)$. So, there exists a treatment rule D such that $W_A(D_A^*) > W_A(D) > W_B(D) > W_B(D_A^*)$, improving upon the maximin criterion $\min(W_A(D), W_B(D))$ and contradicting $D_{\max}^* = D_A^*$. A similar contradiction can be reached when supposing $D_{\max}^* = D_B^*$ and using $W_A\left(D_B^*\right) < W_B\left(D_B^*\right)$.

Suppose $W_A(D_{\max}^*) > W_B(D_{\max}^*)$. By similar reasoning, we can use a combination of suborderings passing through $D_{\max}^*$ in $\mathcal{I}_{+,-}$ and $\mathcal{I}_{-,+}$. By the Intermediate Value Theorem and Axiom 1, there exists a D along the continuum of treatment rules through $D_{\max}^*$ and between D_A^* and D_B^* for any $W_A(D)$ and $W_B(D)$ such that $W_A(D_{\max}^*) > W_A(D) > W_B(D) > W_B(D_{\max}^*)$, which strictly improves on the maximin criterion. A similar contradiction can be reached when supposing $W_A(D_{\max}^*) < W_B(D_{\max}^*)$. □

Proposition 2 (Minimax-Regret Equality of Regret). *Any minimax-regret treatment rule D_{mr}^*, as defined by Equation (A.15), is an interior treatment rule that satisfies $W_A(D_A^*) - W_A\left(D_{mr}^*\right) = W_B(D_B^*) - W_B\left(D_{mr}^*\right)$.*

Proof. As in Proposition 1, we focus on $\min_{\omega\in\{A,B\}} W_\omega(D)$. We also can use arbitrary suborderings in $\mathcal{I}_{+,-}$ and $\mathcal{I}_{-,+}$ and Axioms 1 to 3 to define an arbitrary continuum of treatment rules between D_A^* and D_B^* with continuous W_A and W_B. Define regret functions $R_\omega(D) \equiv W_\omega(D_\omega^*) - W_\omega(D)$. $R_\omega(D_\omega^*) = 0$

and $R_\omega(D) > 0$ for $D \neq D^*_\omega$ for $\omega \in \{A, B\}$. The minimax-regret criterion $\max(R_A(D), R_B(D))$ cannot be minimized when $D = D^*_A$ as the criterion continuously decreases as D approaches $D = D^*_B$ and vice versa.

Similarly, suppose $R_A(D^*_{\text{mr}}) \neq R_B(D^*_{\text{mr}})$. Without loss of generality, consider the case where $R_A(D^*_{\text{mr}}) - R_B(D^*_{\text{mr}}) > 0$. By the existence of an arbitrary continuum of treatment rules running through D^*_{mr} and between D^*_A and D^*_B, we can find D approaching D^*_A such that $R_B(D) = R_A(D) < R_A(D^*_{\text{mr}})$, which contradicts the definition of D^*_{mr}. The case where $R_A(D^*_{\text{mr}}) - R_B(D^*_{\text{mr}}) < 0$ proceeds similarly. □

A.4.4 Ranking Patients by $|u_A/u_B|$

We next consider the optimality of ranking patients by $|u_A/u_B|$ in $\mathcal{I}_{+,-}$ and in $\mathcal{I}_{-,+}$ in the following sense: Ranking patients by decreasing $|u_A/u_B|$ in $\mathcal{I}_{+,-}$ gives the highest increase in W_A for a given decrease in W_B, and ranking patients by increasing $|u_A/u_B|$ in $\mathcal{I}_{-,+}$ gives the lowest decrease in W_A for a given increase in W_B.

Relying on the ordering-threshold formulation of a treatment rule, $D(\cdot; \Omega, \pi)$, we similarly denote welfare associated with the treatment rule as $W_\omega(\Omega, \pi) \equiv W_\omega(D(\cdot; \Omega, \pi))$ for $\omega \in \{A, B\}$. The continuum of treatment rules defined by a fixed Ω and $\pi \in [0, 1]$ implies a curve in (W_A, W_B) space. Consider a treatment rule for patients in $\mathcal{I}_{+,-}$. We will now state the optimality of ranking patients by $|u_A/u_B|$ in $\mathcal{I}_{+,-}$. We omit the $(+,-)$ subscript in the notation for orderings and thresholds for simplicity.

Lemma 1. A treatment rule for patients in $\mathcal{I}_{+,-}$ that maximizes W_A for any W_B can be stated as $D(\cdot; \Omega^*, \pi)$, where $\Omega^*(i) < \Omega^*(i')$ for any (i, i') such that $|u_A(i)/u_B(i)| > |u_A(i')/u_B(i')|$.

Proof. Consider an alternative treatment rule $D(\cdot; \Omega, \pi)$ that yields $W_A(\Omega, \pi') > W_A(\Omega^*, \pi)$ and $W_B(\Omega, \pi') = W_B(\Omega^*, \pi)$. Define three treatment rules:

$$D_1(i) = D(i; \Omega^*, \pi) D(i; \Omega, \pi'); \tag{A.17}$$

$$D_2(i) = D(i; \Omega^*, \pi)(1 - D(i; \Omega, \pi')); \tag{A.18}$$

$$D_3(i) = (1 - D(i; \Omega^*, \pi)) D(i; \Omega, \pi'). \tag{A.19}$$

$D_1(\cdot)$ treats patients that would be treated under both $D(\cdot; \Omega^*, \pi)$ and $D(\cdot; \Omega, \pi')$; $D_2(\cdot)$ treats patients that would be treated under $D(\cdot; \Omega^*, \pi)$ but not under $D(\cdot; \Omega, \pi')$; and $D_3(i)$ treats patients that would be treated under $D(\cdot; \Omega, \pi')$ but not under $D(\cdot; \Omega^*, \pi)$. Denote the share of patients in each treatment rule, respectively, as π_1, π_2, and π_3, where $\pi_1 + \pi_2 = \pi$ and $\pi_1 + \pi_3 = \pi'$.

By Axiom 3, $W_B(\Omega^*, \pi) = W_B(D_1) + W_B(D_2)$ and $W_B(\Omega, \pi') = W_B(D_1) + W_B(D_3)$. Thus, we have $W_B(D_2) = W_B(D_3)$.

To evaluate $W_A(D_2)$ and $W_A(D_3)$, we define respective reorderings $\Omega_2 : \{i : D_2(i) = 1\} \to [0, \pi_2]$ and $\Omega_3 : \{i : D_3(i) = 1\} \to [0, \pi_3]$ and state welfare in the corresponding ordering-threshold formula-

tion (Axiom 2) to show the following inequality:

$$\begin{aligned} W_A(D_2) &= \int_0^{\pi_2} u_A(i)\, d\Omega_2(i) \\ &= \int_0^{\pi_2} \frac{u_A(i)}{u_B(i)} u_B(i)\, d\Omega_2(i) \\ &\geq \min_{i:D_2(i)=1} \left|\frac{u_A(i)}{u_B(i)}\right| |W_B(D_2)| \\ &\geq \max_{i:D_3(i)=1} \left|\frac{u_A(i)}{u_B(i)}\right| |W_B(D_3)| \\ &\geq \int_0^{\pi_3} \frac{u_A(i)}{u_B(i)} u_B(i)\, d\Omega_3(i) \\ &= W_A(D_3) \end{aligned}$$

By the ordering of $\left|\frac{u_A(i)}{u_B(i)}\right|$ in Ω^*, $\min_{i:D_2(i)=1}\left|\frac{u_A(i)}{u_B(i)}\right| > \max_{i:D_3(i)=1}\left|\frac{u_A(i)}{u_B(i)}\right|$. This is because the maximizing treatment rule was built by ranking patients on exactly this metric. Furthermore, since $W_B(D_2) = W_B(D_3)$, we have $W_A(D_2) > W_A(D_3)$. By Axiom 3, then, $W_A(\Omega^*, \pi) = W_A(D_1) + W_A(D_2) > W_A(D_1) + W_A(D_3) = W_A(\Omega, \pi')$, which contradicts the assumption of the alternative treatment rule improving W_A. Note that, by definition of $i \in I_{+,-}$, $\frac{u_A(i)}{u_B(i)} < 0$ and $u_B(i) < 0$, justifying the use of absolute values. □

A similar finding for patients in $\mathcal{I}_{-,+}$ can be stated as follows (now, omitting $(-,+)$ from the subscripts for simplicity):

Corollary 1. A treatment rule for patients in $\mathcal{I}_{-,+}$ that maximizes W_B for any W_A can be stated as $D(\cdot; \Omega^*, \pi)$, where $\Omega^*(i) < \Omega^*(i')$ for any (i, i') such that $|u_B(i)/u_A(i)| > |u_B(i')/u_A(i')|$.

Lemma 1 and Corollary 1 can be used to prove that any optimal treatment in which some but not all patients in $\mathcal{I}_{+,-}$ or $\mathcal{I}_{-,+}$ are treated (i.e., $D^*(i) = D\left(i; \Omega^*_{+,-}, \pi^*_{+,-}\right) + D\left(i; \Omega^*_{-,+}, \pi^*_{-,+}\right) + D_{+,+}(i)$, where $\pi^*_{+,-} \in (0,1)$ or $\pi^*_{-,+} \in (0,1)$) is characterized by an ordering based on $|u_A(i)/u_B(i)|$ within subsets in which a share of patients are treated.

Proposition 3 (Optimality of Ranking by $|u_A/u_B|$). *Consider a maximin or minimax-regret treatment rule D^* as defined by Equation (A.14) or (A.15). For any $\{i, i'\} \subset \mathcal{I}_{+,-}$, $D^*(i) \geq D^*(i')$ if and only if $|u_A(i)/u_B(i)| > |u_A(i')/u_B(i')|$. For any $\{i, i'\} \subset \mathcal{I}_{-,+}$, $D^*(i) \geq D^*(i')$ if and only if $|u_B(i)/u_A(i)| > |u_B(i')/u_A(i')|$.*

Proof. Using Axioms 2 and 3, we state D^* as component treatment rules in $\mathcal{I}_{+,-}$ and $\mathcal{I}_{-,+}$ defined by orderings and thresholds $\left(\Omega^*_{+,-}, \pi^*_{+,-}\right)$ and $\left(\Omega^*_{-,+}, \pi^*_{-,+}\right)$, respectively. Restating the proposition, we seek to show that, if $\pi^*_{+,-} \in (0,1)$, then $\Omega^*_{+,-}(i) < \Omega^*_{+,-}(i')$ for any (i, i') such that $|u_A(i)/u_B(i)| > |u_A(i')/u_B(i')|$.

Consider that this is not the case. By Lemma 1, we can improve W_A while holding fixed W_B by substituting $D\left(\cdot; \Omega^*_{+,-}, \pi^*_{+,-}\right)$ with an alternative (component) treatment rule $D\left(\cdot; \tilde{\Omega}_{+,-}, \tilde{\pi}_{+,-}\right)$ such that $\tilde{\Omega}_{+,-}$ follows the ranking by $|u_A(i)/u_B(i)|$. That is, if we consider $\tilde{D}(\cdot) = D\left(\cdot; \tilde{\Omega}_{+,-}, \tilde{\pi}_{+,-}\right) +$

$D\left(\cdot;\Omega^*_{-,+},\pi^*_{-,+}\right)+D_{+,+}\left(\cdot\right)$, we will have $W_A\left(\tilde{D}\right)>W_B\left(\tilde{D}\right)=W_B\left(D^*\right)=W_A\left(D^*\right)$. By either decreasing $\tilde{\pi}_{+,-}$, under Axiom 1, we can obtain $\min\left(W_A\left(\tilde{D}\right),W_B\left(\tilde{D}\right)\right)>W_A\left(D^*\right)=W_B\left(D^*\right)$. This contradicts the optimality of D^* under both Equations (A.14) and (A.15), so $\Omega^*_{+,-}$ must follow the ranking by $\left|u_A\left(i\right)/u_B\left(i\right)\right|$.

By similar reasoning, using Corollary 1, we can show that if $\pi^*_{-,+}\in(0,1)$, $\Omega^*_{-,+}$ must follow the ranking by $\left|u_B\left(i\right)/u_A\left(i\right)\right|$. □

Remark 3. D^*_A and D^*_B, as potential maximin solutions under Remark 2, also satisfy Proposition 3.

Patients treated under D^*_A satisfy $|u_A/u_B|>0$ in $\mathcal{I}_{+,-}$ (all patients) and $|u_B/u_A|>\infty$ in $\mathcal{I}_{-,+}$ (no patients); patients treated under D^*_B satisfy $|u_A/u_B|>\infty$ in $\mathcal{I}_{+,-}$ (no patients) and $|u_B/u_A|>0$ in $\mathcal{I}_{-,+}$ (all patients).

A.4.5 Equality of Utility-Ratio Cutoffs in $\mathcal{I}_{+,-}$ and $\mathcal{I}_{-,+}$

To summarize, Remark 1 and Proposition 3 characterize maximin and minimax-regret treatment rules as follows:

$$D^*\left(i\right)=\begin{cases}1, & i\in\mathcal{I}_{+,+};\\ \mathbf{1}\left(\left|u_A\left(i\right)/u_B\left(i\right)\right|>\chi_{+,-}\right), & i\in\mathcal{I}_{+,-};\\ \mathbf{1}\left(\left|u_A\left(i\right)/u_B\left(i\right)\right|<\chi_{-,+}\right), & i\in\mathcal{I}_{-,+};\\ 0, & i\in\mathcal{I}_{-,-}.\end{cases}\tag{A.20}$$

Proposition 4 (Equal Utility-Ratio Cutoffs). *Any maximin or minimax-regret treatment rule has equal utility-ratio cutoffs in Equation (A.20) corresponding to $\mathcal{I}_{+,-}$ and $\mathcal{I}_{-,+}$ (i.e., $\chi_{+,-}=\chi_{-,+}$).*

Proof. First, consider the maximin criterion. Remark 3 notes that this condition holds for the corner solutions at D^*_A and D^*_B, so we focus our proof on interior optimal treatment rules. If there exists an interior optimal treatment rule, by Proposition 1, $W_A\left(D^*\right)=W_B\left(D^*\right)$. Under the optimal orderings $\Omega^*_{+,-}$ and $\Omega^*_{-,+}$, by continuity (Axiom 1), there exist a continuum of thresholds $\left\{\left(\pi_{+,-},\pi_{-,+}\right)\right\}$, such that $W_A\left(D\right)=W_B\left(D\right)$ for $D\left(\cdot\right)=D_{+,+}+D\left(\cdot;\Omega^*_{+,-},\pi_{+,-}\right)+D\left(\cdot;\Omega^*_{-,+},\pi_{-,+}\right)$. Under the constraint $W_A\left(D\right)=W_B\left(D\right)$, increasing the share of patients treated in $\mathcal{I}_{+,-}$ increases the share of patients treated in $\mathcal{I}_{-,+}$, so $\pi_{+,-}$ increases with $\pi_{-,+}$.

From Proposition 3, $\chi_{+,-}$ weakly decreases in $\pi_{+,-}$, and $\chi_{-,+}$ weakly increases in $\pi_{-,+}$. Thus, under the constraint $W_A\left(D\right)=W_B\left(D\right)$, $\chi_{+,-}$ decreases in $\chi_{-,+}$.

For ease of notation, consider an index share $\sigma=\Pr\left(i\in\mathcal{I}_{+,-}\right)\pi_{+,-}+\Pr\left(i\in\mathcal{I}_{-,+}\right)\pi_{-,+}\in\left[\underline{\sigma},\overline{\sigma}\right]$ for a pair $\left(\pi_{+,-},\pi_{-,+}\right)$ satisfying $W_A\left(D\right)=W_B\left(D\right)$. We can then refer to $\pi_{+,-}(\sigma)$, $\pi_{-,+}(\sigma)$, $\chi_{+,-}(\sigma)$, and $\chi_{-,+}(\sigma)$. Appealing to Axiom 3, we focus on components $W^k_\omega\left(\sigma\right)\equiv W_\omega\left(D\left(\cdot;\Omega^*_k,\pi_k\left(\sigma\right)\right)\right)$, for $\omega\in\{A,B\}$ and $k\in\{(+,-),(-,+)\}$. Now consider perturbations in welfare moving from σ to σ', e.g.,

$$\Delta W^k_\omega\equiv W^k_\omega\left(\sigma'\right)-W^k_\omega\left(\sigma\right)=\int_{\pi_k(\sigma)}^{\pi_k(\sigma')}u_\omega\left(i\right)d\Omega^*_k\left(i\right).\tag{A.21}$$

Making use of the constraint,

$$\Delta W = \Delta W_A^{+,-} + \Delta W_A^{-,+} = \Delta W_B^{+,-} + \Delta W_B^{-,+}. \tag{A.22}$$

1. *Case 1:* $\chi_{+,-}(\sigma) > \chi_{-,+}(\sigma)$ and nearby index σ' such that $\chi_{+,-}(\sigma') > \chi_{-,+}(\sigma')$. Then $\underline{\chi}_{+,-} \equiv \min(\chi_{+,-}(\sigma), \chi_{+,-}(\sigma')) > \max(\chi_{-,+}(\sigma), \chi_{-,+}(\sigma')) \equiv \overline{\chi}_{-,+}$. From Proposition 3 and Equation (A.21), $\left|\Delta W_A^{+,-}/\Delta W_B^{+,-}\right| \geq \underline{\chi}_{+,-} > \overline{\chi}_{-,+} \geq \left|\Delta W_A^{-,+}/\Delta W_B^{-,+}\right|$ if and only if $\sigma' > \sigma$. Combining this with Equation (A.22), $\Delta W > 0$ if and only if $\sigma' > \sigma$.

   We can show this by contradiction. Assume that $\Delta W \leq 0$, then it follows that $|\Delta W_A^{-,+}| \geq |\Delta W_A^{+,-}|$, and $|\Delta W_B^{+,-}| \geq |\Delta W_B^{-,+}|$. However then re-arranging the above inequality, $1 \geq \frac{|\Delta W_B^{-,+}|}{|\Delta W_B^{+,-}|} > \frac{|\Delta W_A^{-,+}|}{|\Delta W_A^{+,-}|} \geq 1$, which is a contradiction.

2. *Case 2.* $\chi_{+,-}(\sigma) < \chi_{-,+}(\sigma)$ and nearby index σ' such that $\chi_{+,-}(\sigma') < \chi_{-,+}(\sigma')$. Thereby $\overline{\chi}_{+,-} \equiv \max(\chi_{+,-}(\sigma), \chi_{+,-}(\sigma')) < \min(\chi_{-,+}(\sigma), \chi_{-,+}(\sigma')) \equiv \underline{\chi}_{-,+}$ Applying similar logic, $\Delta W > 0$ if and only if $\sigma' < \sigma$.

Welfare increases whenever the gap $\left|\chi_{+,-} - \chi_{-,+}\right|$ decreases under the constraint $W_A(D) = W_B(D)$. By this logic, Equation (A.20) cannot be an optimal treatment rule unless $\chi_{+,-} = \chi_{-,+}$. Furthermore, if $|u_A(i)/u_B(i)|$ is continuous, $\sup|u_A(i)/u_B(i)| = \infty$, and $\inf|u_A(i)/u_B(i)| = 0$ in $\Omega_{+,-}^*$ and $\Omega_{-,+}^*$, $\chi_{+,-}(\sigma)$ and $\chi_{-,+}(\sigma)$ will cross for some $\sigma \in \left[\underline{\sigma}, \overline{\sigma}\right]$.

The proof follows similarly for the minimax-regret criterion. Instead of using the constraint $W_A(D) = W_B(D)$, we use the constraint $R_A(D) = R_B(D)$ as implied by Proposition 2, where $R_\omega(D) \equiv W_\omega(D_\omega^*) - W_\omega(D)$ for $\omega \in \{A, B\}$. □

A.4.6 Representation in $\left(\alpha^b, \tau^s\right)$ Space

The optimal treatment rules defined by Equation (A.19) and Proposition 4 can be represented as threshold rules in $\left(\alpha^b, \tau^s\right)$ space. While we will now re-complicate the notation by referring to the parameter ρ to highlight its relationship with the threshold rules, we will simplify the notation by referring to $\left(\alpha^b, \tau^s\right)$ as (X, Y) (for random variables) and (x, y) (for realizations). Similarly, we will refer to $\overline{\alpha}^b$ and $\alpha_{\max}^b$ as $\overline{x}$ and $x_{\max}$, respectively.

To aid the analysis, we will assume the following regularity condition on the joint distribution of (X, Y) and consider uncertainty-robust rules for the infinite population of patients under this distribution.

Assumption 2. *The density of the joint distribution of* (X, Y) *is absolutely continuous and everywhere positive in* $[0, x_{max}] \times [-1, 0]$.

Proposition 5 (Linear Optimal Treatment Rules). *Any maximin or minimax-regret treatment rule* $D^*(\cdot; \rho)$ *can be represented as* $D^*(i; \rho) = \mathbf{1}(Y_i < \ell^*(X_i; \rho))$*, where* $\ell^*(\cdot; \rho)$ *is a linear function passing through* $(\overline{x}, -\rho\overline{x})$ *with a slope* $m \in [-\rho, 0]$*. Further, each* ρ *implies exactly one maximin treatment rule and exactly one minimax-regret treatment rule.*

Proof. Starting with Equation (A.19) and the condition $\chi_{+,-} = \chi_{-,+}$ in Proposition 4, we substitute the definitions of $u_A(i)$ and $u_B(i)$ in Equations (A.6) and (A.7) into the ratio $u_A(i)/u_B(i)$. With some algebraic manipulation, we can state this condition as a threshold:

$$D^*(i;\rho) = \mathbf{1}\left(Y_i < -\rho\overline{x} - \rho\chi/(1+\chi)\left(X_i - \overline{x}\right)\right). \tag{A.23}$$

This is an equation for a straight line passing through $(\overline{x}, -\rho\overline{x})$ with slope $-\rho\chi/(1+\chi)$. For $\chi \in (0,\infty)$, the slope is $m = -\rho\chi/(1+\chi) \in (-\rho, 0)$.

Assumption 2 implies the existence of a maximin treatment rule and a minimax-regret treatment rule. Further, under Assumption 2, χ is uniquely determined by ρ and the joint distribution of (X,Y) such that $W_A(D^*) = W_B(D^*)$ or $R_A(D^*) = R_B(D^*)$, as described in the proof of Proposition 4. Therefore, the line is uniquely determined by ρ. So each ρ implies exactly one maximin treatment rule and exactly one minimax-regret treatment rule.

Note that $D_A^*(\cdot;\rho)$ and $D_B^*(\cdot;\rho)$ correspond to the same linear rules in Equation (A.23) with slope $m = 0$ and $m = -\rho$, respectively. □

Definition. A set of treatment rules $\{D_k : \mathcal{I} \to \{0,1\}\}_{k\in\mathcal{K}}$ satisfies *monotonicity* if, for any $\{k,k'\} \in \mathcal{K}$ and any $\{i,i'\} \in \mathcal{I}$, $D_k(i) > D_{k'}(i)$ implies $D_k(i') \geq D_{k'}(i')$.

Monotonicity of Assumption-Specific Optimal Treatment Rules. It is straightforward to recognize that the corner solutions to the maximin treatment rules—i.e., $D_A^*(\cdot;\rho)$ and $D_B^*(\cdot;\rho)$—satisfy monotonicity.

Remark 4. The set of treatment rules $\left\{D_A^*(\cdot;\rho)\right\}_{\rho\in[0,\infty)}$ in Equation (A.11) satisfies monotonicity. The set of treatment rules $\left\{D_B^*(\cdot;\rho)\right\}_{\rho\in[0,\infty)}$ in Equation (A.12) satisfies monotonicity.

Proof. Monotonicity of $\left\{D_A^*(\cdot;\rho)\right\}_{\rho\in[0,\infty)}$ and $\left\{D_B^*(\cdot;\rho)\right\}_{\rho\in[0,\infty)}$ follows directly from Equations (A.11) and (A.12), respectively. Without loss of generality, consider $\rho' < \rho$. $D_A^*(i;\rho) < D_A^*(i;\rho')$ if and only if $Y_i \in [-\rho\overline{x}, -\rho'\overline{x}]$. Otherwise, $D_A^*(i;\rho) = D_A^*(i;\rho')$. $D_B^*(i;\rho) < D_B^*(i;\rho')$ if and only if $Y_i \in [-\rho X_i, -\rho' X_i]$. Otherwise, $D_B^*(i;\rho) = D_B^*(i;\rho')$ □

Monotonicity of Maximin Treatment Rules.

Remark 5. Any interior maximin treatment rule $D_{\max}^*(\cdot;\rho)$ satisfies

$$E\left[X_i \middle| D_{\max}^*(i;\rho) = 1\right] = \overline{x}. \tag{A.24}$$

Any treatment rule that satisfies Equation (A.24) is a maximin treatment rule.

Proof. Suppressing ρ and using $W_A\left(D^*_{\max}\right) = W_B\left(D^*_{\max}\right)$ from Proposition 1,

$$\begin{aligned}
\Pr\left(D^*_{\max}(i)=1\right) E\left[u_A(i)\middle|D^*_{\max}(i)=1\right] &= \Pr\left(D^*_{\max}(i)=1\right) E\left[u_B(i)\middle|D^*_{\max}(i)=1\right] \\
E\left[-Y_i-\rho\overline{x}\middle|D^*_{\max}(i)=1\right] &= E\left[-Y_i-\rho X_i\middle|D^*_{\max}(i)=1\right] \\
\overline{x} &= E\left[X_i\middle|D^*_{\max}(i)=1\right].
\end{aligned}$$

□

Figure 7 shows the maximin rule satisfying the condition in Remark 5 in the space of stroke benefits and bleed risks. As noted in Proposition 5, the maximin rule is linear and falls in between the two assumption-specific optimal rules (i.e., $D^*_A(\cdot;\rho)$ and $D^*_B(\cdot;\rho)$), within the regions $\mathcal{I}_{+,-}$ and $\mathcal{I}_{-,+}$.

Remark 6. Suppose that the conditional quantile functions (CQFs) of Y given $X = x$ are linear for any quantile π. Then the set of maximin treatment rules $\left\{D^*_{\max}(\cdot;\rho)\right\}_{\rho\in[0,\infty)}$ in Equation (A.14) satisfies monotonicity.

Proof. If all CQFs of Y given $X = x$ are linear, a linear conditional quantile function $Q_Y(\pi(\rho)|X=x)$ passes through $(\overline{x}, -\rho\overline{x})$, where its quantile π is a function of ρ that we denote as $\pi(\rho)$. We now consider a linear treatment rule $D(i;\rho) = \mathbf{1}(Y_i < \ell(X_i;\rho))$, where $\ell(x;\rho) = Q_Y(\pi(\rho)|X=x)$.

Because $F_{Y|X}(\ell(x;\rho)|X=x) = \pi(\rho)$ for all x, $\Pr(D(i;\rho)=1) = \pi(\rho))$. The treatment rule satisfies Equation (A.24):

$$\begin{aligned}
E\left[X_i\,|D(i)=1\right] &= \frac{1}{\pi(\rho)}\int_0^{x_{\max}} xF_{Y|X}(\ell(x;\rho)|X=x)dF_X(x) \\
&= \int_0^{x_{\max}} xdF_X(x) \\
&= \overline{x},
\end{aligned}$$

for all ρ. Therefore, the interior maximin treatment rules are CQFs. Because CQFs do not cross by definition, the treatment rules $\left\{D^*_{\max}(\cdot;\rho)\right\}_{\rho\in[0,\infty)}$ satisfy monotonicity. □

We can generalize this remark by noting that the maximin treatment rules can be viewed as a special case of distributional regression that targets specific quantile-level thresholds using conditional distribution information by minimizing squared loss in CDF space.

Remark 7. The maximin treatment rule $D^*_{\max}(\cdot;\rho)$ in Equation (A.23) is equivalent to a population linear distributional regression $\hat{q}_\pi(X) = \beta_0(\pi) + \beta_1(\pi)X$ targeting the following objective:

$$\min_{\beta_0,\beta_1} E\left[\left(F_{Y|X}(\beta_0+\beta_1 X|X) - \pi\right)^2\right],$$

where $\pi \in [0,1]$ is a function of ρ.

Proof. A treatment rule defined by the threshold $D(i) = \mathbf{1}(Y_i < \hat{q}_\pi(X_i))$ will satisfy the condition for maximin treatment rules in Equation (A.24). Let $\pi = \Pr(D(i)=1) \in [0,1]$. The proof proceeds

similarly to the one in Remark 6:

$$\begin{aligned} E\left[X_i \,|D\left(i\right)=1\right] &= \frac{1}{\pi}\int_0^{x_{\max}} x F_{Y|X=x}(\hat{q}_\pi(x))dF_X(x) \\ &= \int_0^{x_{\max}} x dF_X(x) + \frac{1}{\pi}E[X\cdot(F_{Y|X}(\beta_0+\beta_1 X|X)-\pi)] \\ &= \overline{x}, \end{aligned}$$

where the last equality follows from properties resulting from minimizing the expected squared CDF-residual loss: $E[r(X)] = 0$ and $\text{Cov}(X \cdot r(X)) = 0$, where $r(X) \equiv F_{Y|X}(\beta_0 + \beta_1 X|X) - \pi$, so that $E[X \cdot r(X)] = \overline{x} \cdot 0 = 0$. From Proposition 5, each ρ has exactly one maximin treatment rule. So π is a well-defined function of ρ. □

Remark 5 is useful because it relates the maximin treatment rules to projections of the CQFs for specific π, very much like quantile regression fits are. There is a longstanding literature acknowledging that, in general, quantile regression fits may cross (e.g., Koenker 2005, Chapter 2.5). Projections of CQFs may cross even though the underlying CQFs do not cross (by construction) because the projections are simplifications of the underlying functions. When X is bounded, as in our case, it is intuitive that projections are less likely to cross when CQFs are "close" to being parallel or linear (i.e., when the projections are better approximations).[9] We empirically show that our maximin treatment rules obey monotonicity. Nonetheless, methods to ensure that quantile projections obey monotonicity (e.g., rearrangement and monotonization) are an area of rich research (Chernozhukov, Fernandez-Val, and Galichon 2010).

Monotonicity of Minimax-Regret Treatment Rules. The minimax-regret treatment rule can be stated as a moment condition generalized from Equation (A.24) in Remark 5.

Remark 8. Any interior minimax-regret treatment rule $D^*_{\text{mr}}(\cdot;\rho)$ satisfies

$$E\left[X_i \left| D^*_{\text{mr}}(i;\rho) = 1\right.\right] = \overline{x} + g(\rho), \tag{A.25}$$

where $g(\rho) \equiv W_B(D^*_B(\cdot;\rho)) - W_A(D^*_A(\cdot;\rho))$. Any treatment rule that satisfies Equation (A.25) is a minimax-regret treatment rule.

Proof. This follows from Proposition 2 and similar reasoning as the proof for Remark 5. □

Based on the definition of $g(\rho)$, we have $g(0) = g(\infty) = 0$; $g(\rho)$ is also a continuous and smooth function of ρ since $W_A(D^*_A(\cdot;\rho))$ and $W_B(D^*_B(\cdot;\rho))$ are continuous and smooth functions of ρ under Assumption 2. To build intuition for the conditions for monotonicity of minimax treatment rules $\left\{D^*_{\text{mr}}(\cdot;\rho)\right\}_{\rho\in[0,\infty)}$, we first note a sufficient condition for monotonicity in a neighborhood of ρ (similar to Remark 6) and connect minimax-regret treatment rules to a projection of CQFs (similar to Remark 7).

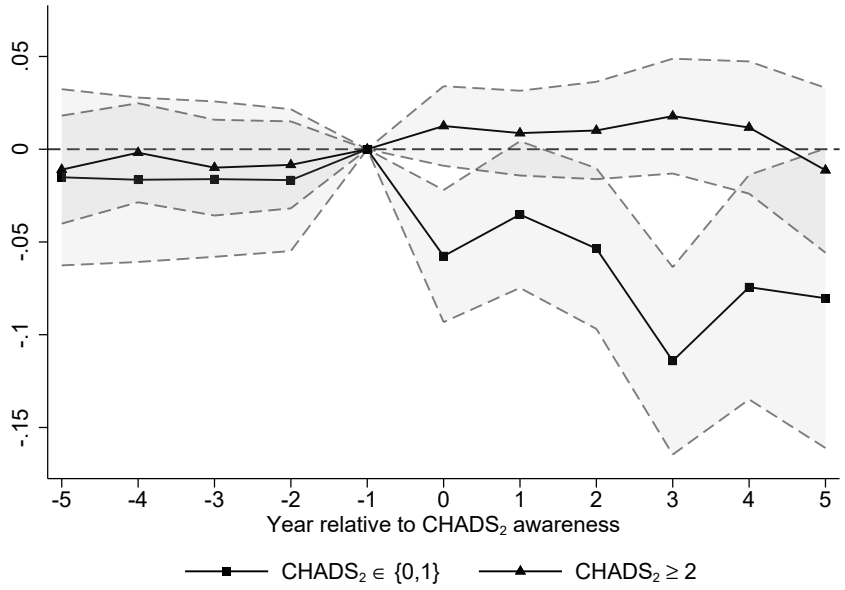
9. When CQFs are parallel or linear (as we note in Remark 6), then the projections will not cross. Note that joint normal distributions have parallel CQFs. Of course, CQFs may only be nonparallel and linear when the domain X is bounded.

Remark 9. Consider a neighborhood $(\rho - \Delta_\rho, \rho + \Delta_\rho)$ around ρ and the minimax-regret treatment rule $\ell^*(\cdot;\rho)$ for ρ as described in Proposition 5. If $g(\cdot)$ is constant in the neighborhood and if $F_{Y|X}(\ell^*(x;\rho) + \Delta_Y | X = x) - F_{Y|X}(\ell^*(x;\rho)|X = x)$ is a constant across all x, for $|\Delta_Y| < \left|\Delta_\rho\right|\overline{x}$, then the set of minimax-regret treatment rules in the neighborhood around ρ obeys monotonicity.

Proof. If the conditional CDF along a candidate linear treatment rule shifted by $-\Delta_\rho\overline{x}$ is the same across x but for a constant, then $E\left[X_i | Y_i < \ell^*(X_i;\rho)\right] = E\left[X_i | Y_i < \ell^*(X_i;\rho) - \Delta_\rho\overline{x}\right]$. If $g(\rho) = g(\rho + \Delta_\rho)$, then the minimax-regret treatment rule $D^*_{\text{mr}}(\cdot;\rho + \Delta_\rho)$ satisfying Equation (A.25) can be characterized by $\ell^*(\cdot;\rho + \Delta_\rho) = \ell^*(\cdot;\rho) - \Delta_\rho\overline{x}$. That is, $\ell^*(\cdot;\rho)$ and $\ell^*(\cdot;\rho + \Delta_\rho)$ are parallel. □

The moment condition in Equation (A.25) may be viewed as another type of CQF projection for minimax-regret treatment rules (as we note in Remark 7 for maximin treatment rules). Within the finite space of $x \in [0, x_{\max}]$, linear treatment rules in (X, Y) need not be parallel to result in monotonicity. Therefore, $g(\rho)$ need not be exactly constant, and the conditional CDF need not shift by a constant across x (e.g., the CQFs need not be parallel) for the minimax-regret treatment rules to satisfy monotonicity. However, it is intuitive that the slower that $g(\rho)$ changes or the less that the change in $F_{Y|X}$ varies across x along the minimax-regret treatment line, the more that monotonicity is likely to be satisfied.

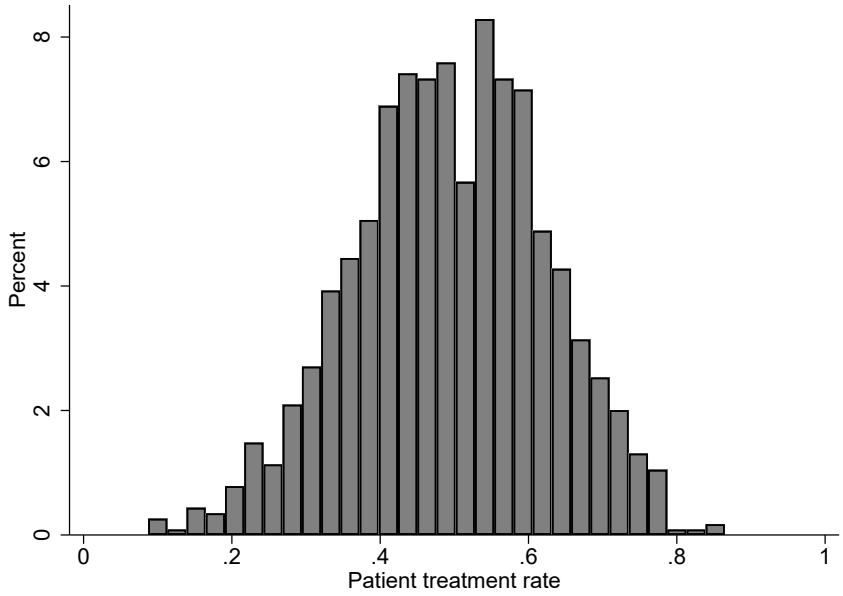
Figure A.1: Event-Study Estimates



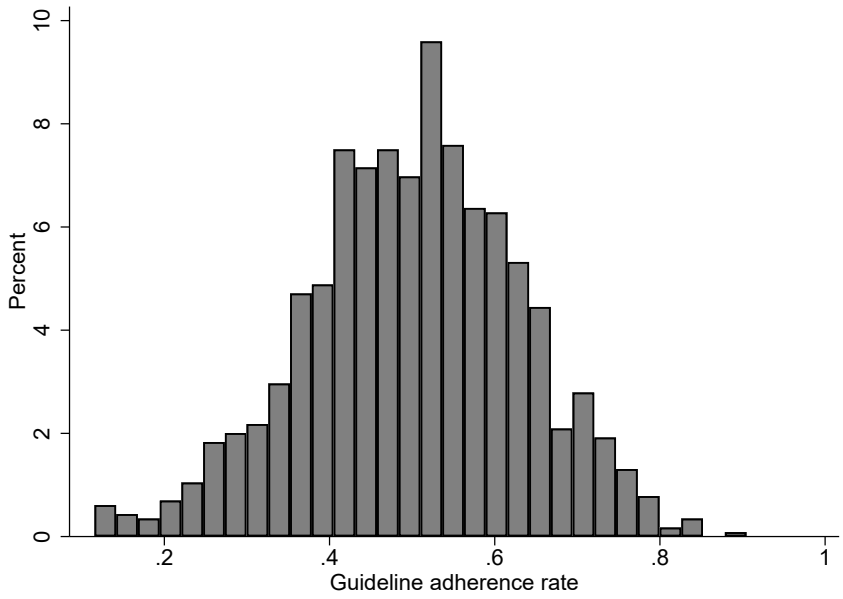
Notes: This figure shows regression coefficients and 95% confidence intervals from Equation (A.1), run separately for patients with $\text{CHADS}_2 \in \{0,1\}$ and for patients with $\text{CHADS}_2 \geq 2$. The 12-month period prior to the physician's first CHADS_2 mention is normalized to 0. The regression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



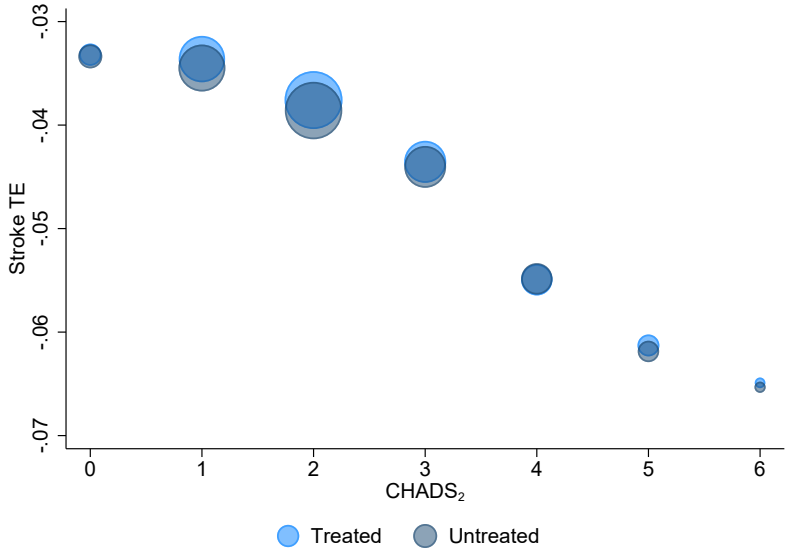
B. $CHADS_2$ Adherence



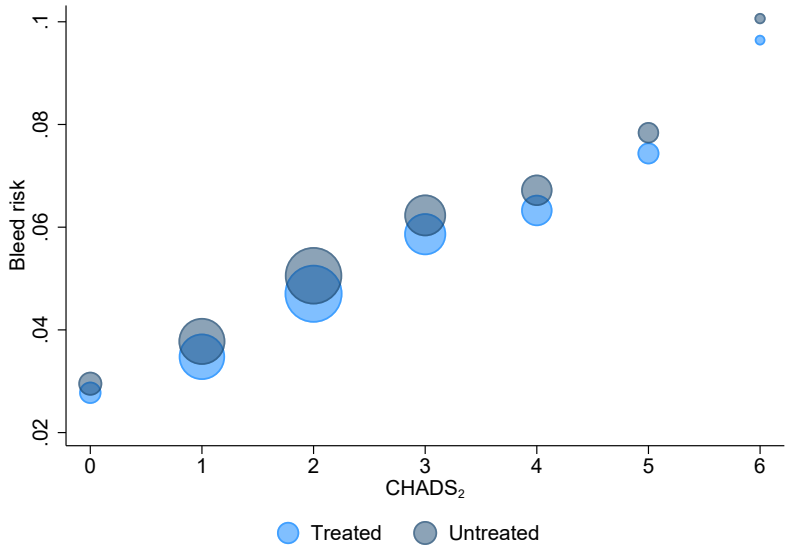
Notes: These figures show the distribution of treatment rates and $CHADS_2$ adherence rates across physicians. They cover the subsample of 1,146 physicians treating at least 30 patients in our final analysis sample. This covers 50,426 patients treated by the higher volume doctors, or a little less than half of the VHA sample defined in Table A.1. 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: Patient Characteristics by $CHADS_2$ Score and Treatment Status

A. Stroke Treatment Effect



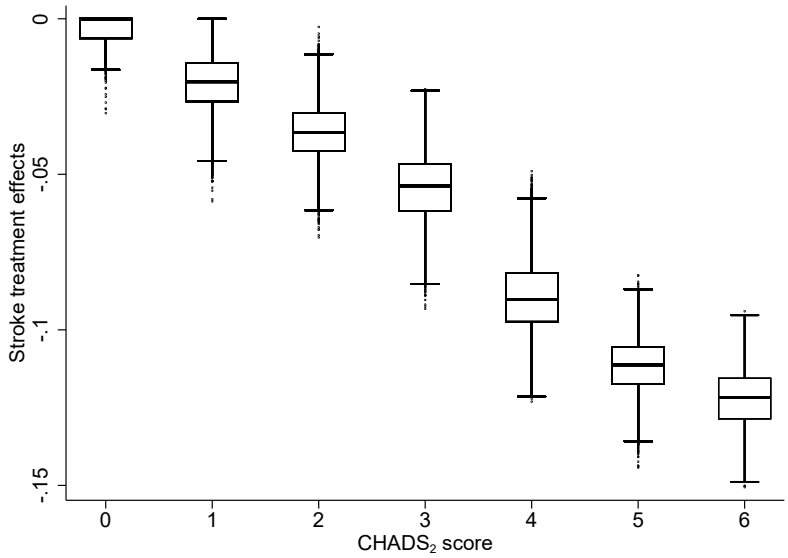
B. Bleed risk



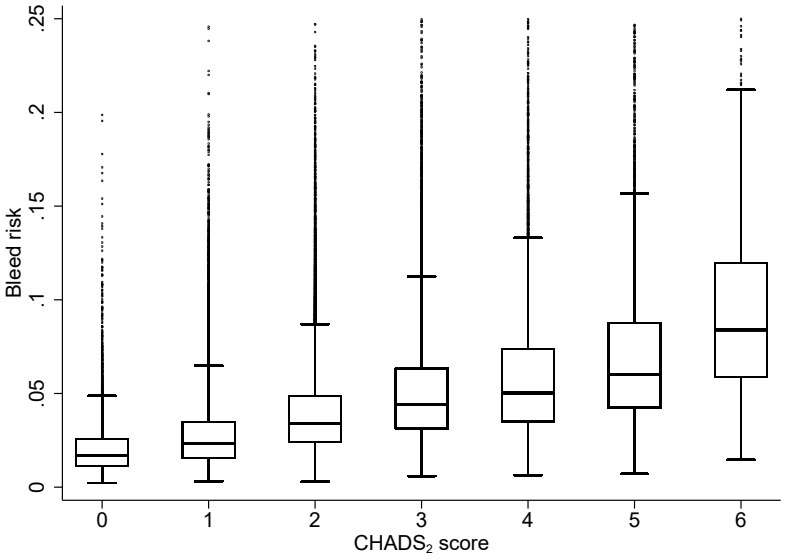
Notes:

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

A. Stroke Treatment Effect



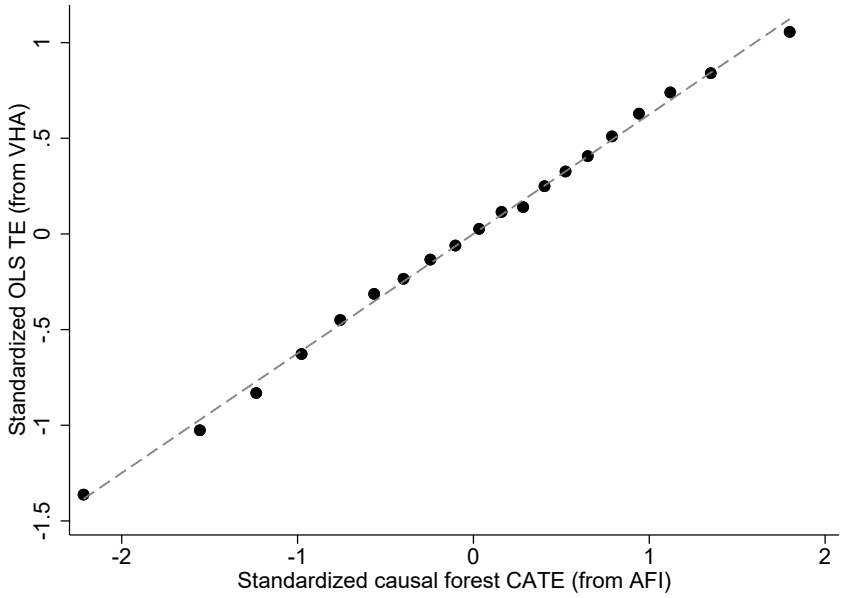
B. Bleed Risk



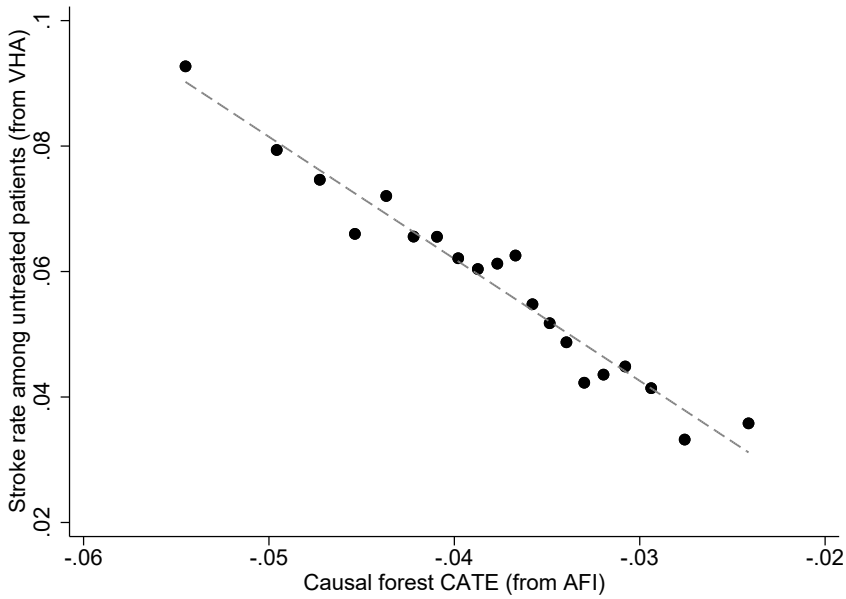
Notes: This figure shows box plots for the distribution of stroke treatment effects and bleed risk by $CHADS_2$ score across patients in the VHA data. Panel A shows stroke treatment effects; Panel B shows bleed risk. Bounds on the box plot are at the 25th and 75th percentile, with the median marked with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

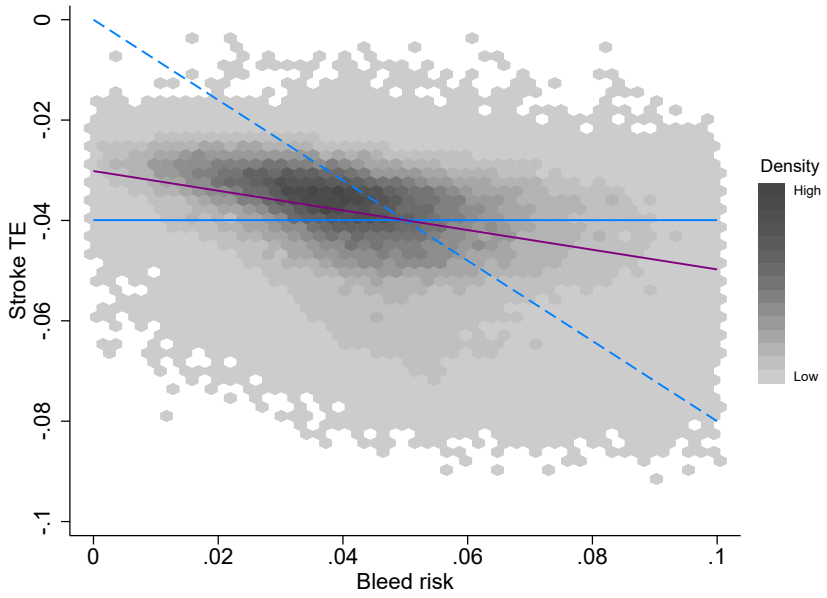


B. Stroke Outcomes Among Untreated Patients and AFI CATEs



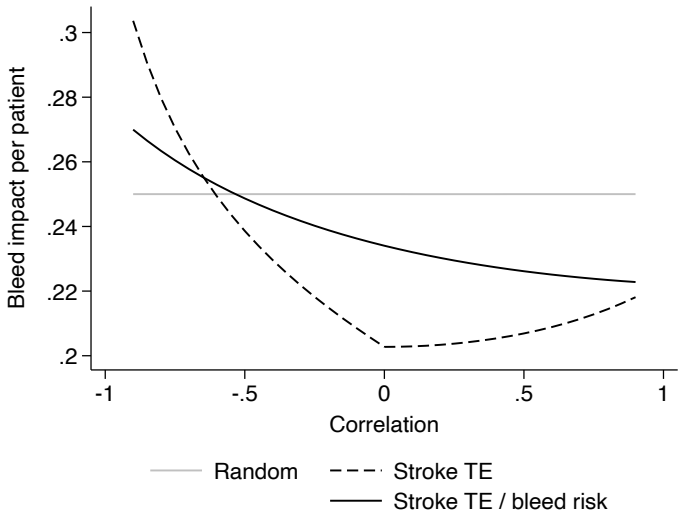
Notes: Panel A displays a binned scatterplot of the relationship between z-standardized stroke treatment effects estimated by a causal forest in the AFI RCT data and z-standardized stroke treatment effects estimated by OLS in the VHA observational data. The correlation coefficient is 0.62. OLS estimates interact with anticoagulation treatment with each of the patient characteristics that are covered in both the AFI and VHA data (see Appendix Table A.2 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 AF diagnosis.

Figure A.6: Optimal Treatment Rules for the VHA Sample



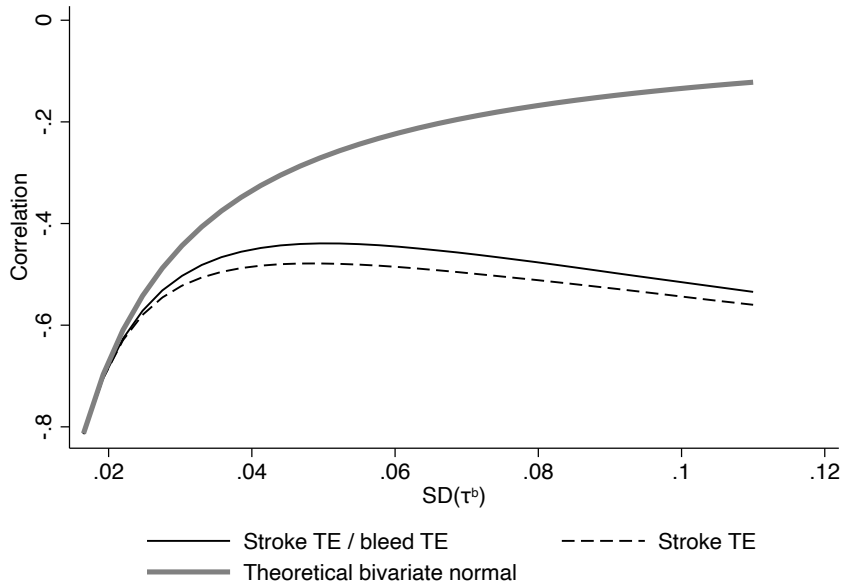
Notes: This figure represents optimal treatment rules based on stroke treatment effects and bleed risks in the VHA data. It builds on Figure 4, which illustrates the joint distribution of stroke treatment effects and bleed risks in the VHA data, and Figure 7, which illustrates optimal treatment rules 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 maximin 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



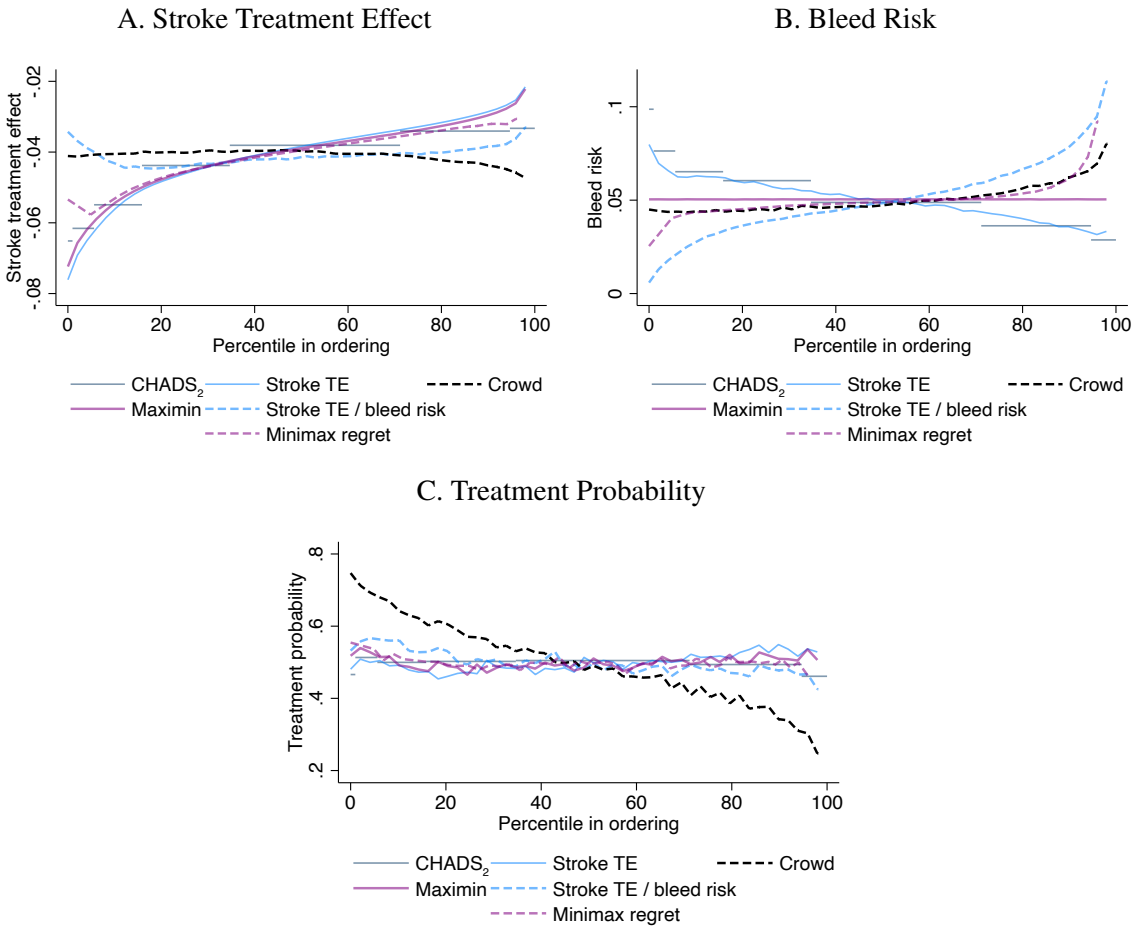
Notes: This figure illustrates worse-case outcomes under three optimal treatment rules depending on the correlation between stroke and bleed treatment effects. The joint distribution of stroke treatment effects and potentially heterogeneous bleed treatment effects is a truncated bivariate normal distribution, 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.

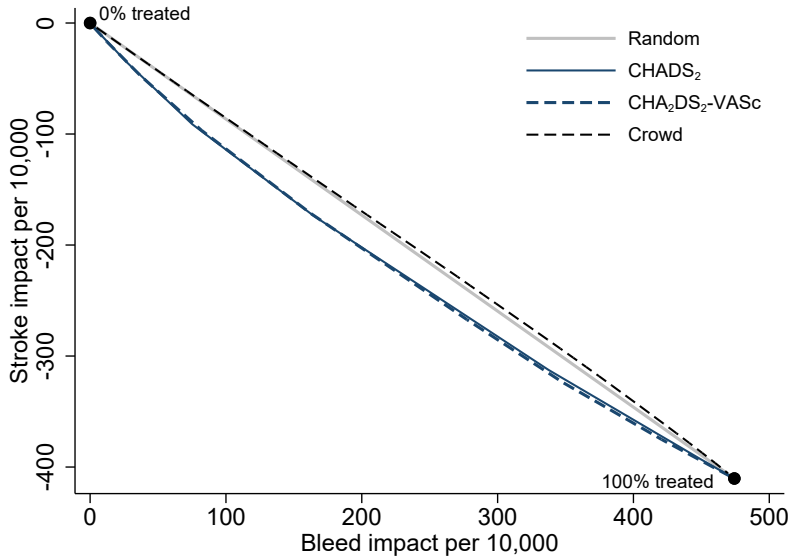
Figure A.9: Patient Characteristics by Different Rankings



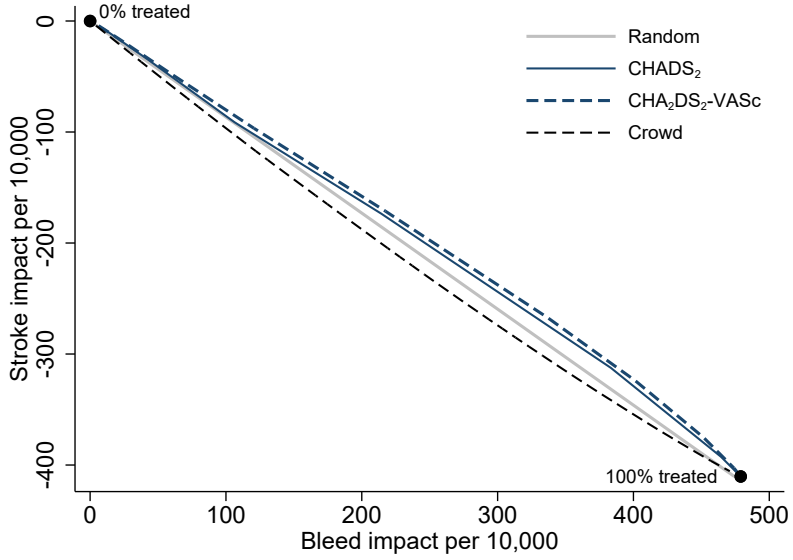
Notes: This figure shows patient characteristics and provider treatment probability of patients under several different rankings. The figure orders patients by the $CHADS_2$ score, stroke treatment effects (the optimal homogeneous rule), the ratio of stroke and bleed treatment effects (the optimal heterogeneous rule), the maximin 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 A.10: Wisdom of the Crowd

A. Mean-Independent Harms



B. Proportional Harms



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 ranking 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 independent bleed treatment effects. Panel B shows outcomes under the assumption of bleed treatment effects proportional to bleed risks. For reference, outcomes under random treatment are shown as a straight gray line.

Table A.1: VHA Sample Selection

Sample step	Description	Observations Dropped	Observations Remaining
1. Identify potentially new atrial fibrillation (AF) patients	Identify candidate patients with a diagnosis of AF 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 AF diagnosis	Keep patients who had an EKG within 30 days before or after initial diagnosis and a second AF 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 AF 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

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

Table A.2: 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

Notes: 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.3: 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 ≥ 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

Notes: 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.4: Balance in Patient Characteristics

	Means			Within-trial difference	
Patient characteristic	Treatment	Control	Missing (%)	Non-imputed	Imputed
Panel A: Demographics					
Age	70.3	70.4	0.00%	0.180	0.180
				(0.247)	(0.247)
Age ≥ 65	0.76	0.77	0.00%	0.003	0.003
				(0.012)	(0.012)
Male	0.68	0.67	0.00%	-0.016	-0.016
				(0.013)	(0.013)
White	0.94	0.95	23.01%	0.003	0.002
				(0.007)	(0.006)
Panel B: Patient medical history					
Congestive heart failure	0.288	0.305	0.00%	-0.006	-0.006
				(0.012)	(0.012)
Hypertension	0.459	0.447	0.00%	-0.008	-0.008
				(0.015)	(0.015)
Diabetes	0.138	0.136	0.00%	-0.006	-0.006
				(0.010)	(0.010)
Stroke	0.125	0.174	0.00%	-0.006	-0.006
				(0.008)	(0.008)
Myocardial infarction	0.106	0.110	0.00%	-0.011	-0.011
				(0.009)	(0.009)
Angina	0.168	0.166	0.00%	0.003	0.003
				(0.011)	(0.011)
Transient ischemic attack	0.283	0.231	85.28%	0.051	0.007
				(0.034)	(0.005)
Peripheral vascular disease	0.085	0.070	29.75%	0.011	0.007
				(0.009)	(0.007)
Panel C: Lifestyle factor					
Smoker	0.500	0.500	43.03%	-0.005	-0.006
				(0.018)	(0.011)
Panel D: Clinical measures					
Weight	175.9	173.3	16.44%	-1.198	-1.123
				(1.240)	(1.040)
Systolic blood pressure	142.4	144.4	0.06%	-0.405	-0.378
				(0.565)	(0.565)
Diastolic blood pressure	82.8	83.6	0.08%	-0.267	-0.243
				(0.287)	(0.287)
Hemoglobin	14.8	14.7	30.28%	-0.008	-0.017
				(0.049)	(0.035)

Notes: This table shows the means of each patient characteristic in the treatment and control groups of the AFI trials. It shows the share of observations that are missing a value for a given characteristic. The last two columns report regression estimates of the difference in means between the treatment and control groups, adjusting for trial fixed effects. The "non-imputed" column shows the difference when using the original data, dropping observations with missing values. The "imputed" column shows the difference when using data with missing values substituted with imputations from a regression forest model based on non-missing characteristics. Standard errors of these estimates are shown in parentheses.

Table A.5: Machine Learning Variable Importance and Selection

Variable importance		Lasso selected variables	
Causal forest	Regression forest	Stroke risk	Stroke TE
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 (−)

Notes: Each column of this table shows the top 10 most important variables for the causal forest, regression forest, and LASSO risk and treatment-effect 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.6: Trial-Specific Information (Part 1)

Trial	Dates	Inclusion criteria	Exclusion criteria	Arms and sample size
Atrial Fibrillation, Aspirin, and Anticoagulation Study 1 (AFASAK-1)	Nov 1985 – Jun 1988	> 18 years old, ECG-verified AF	Previous anticoagulation (> 6 months), cerebrovascular event (< 1 month), contraindications for aspirin/warfarin, previous side effects from aspirin/warfarin, current treatment with aspirin/warfarin, pregnancy or breastfeeding, BP > 180/100, psychiatric diseases (including alcoholism), heart surgery with valve replacement, sinus rhythm, rheumatic heart disease, refusal to participate	Warfarin (N = 335), INR target 2.8-4.2; aspirin (N = 336), 75 mg once daily; placebo (N = 336)
Atrial Fibrillation, Aspirin, and Anticoagulation Study 2 (AFASAK-2)	May 1993 – Oct 1996	> 18 years old, AF on > 1 ECG (at least 1 month apart)	Mitral stenosis, lone AF (< 60 years), SBP > 180 mmHg, DBP > 100 mmHg, stroke/TIA (< 6 months), bleeding risk factors, previous indication to OAC/AP, already on OAC	Mini-dose warfarin (N = 167), 1.25 mg/day; Adjusted-dose warfarin (N = 170), INR target 2.0-3.0; aspirin (N = 169), 300 mg daily; aspirin + warfarin (N = 171), 300 mg daily + 1.25 mg daily
Boston Area Anticoagulation Trial for Atrial Fibrillation (BAATAF)	Not specified	Chronic AF (ECG-proven), 2 separate AF ECGs	Mitral stenosis, transient AF, planned cardioversion, intracardiac thrombus, ventricular aneurysm, severe heart failure, prosthetic valve, stroke (< 6 months), predisposition to intracranial hemorrhage, abnormal thyroid function, previous indication/contraindication for OAC or AP	Warfarin (N = 212), INR target 1.5-2.7; control (N = 208), allowed aspirin
Canadian Atrial Fibrillation Anticoagulation (CAFA)	Jun 1987 – Apr 1990	Chronic AF (> 1 month), paroxysmal AF (> 2 episodes in 3 months), > 18 years old	Mitral valve prosthesis, mechanical aortic valve prosthesis, mitral valve stenosis, previous indication/contraindication for OAC, stroke/TIA (< 1 year), hyperthyroidism, uncontrolled hypertension, myocardial infarction (< 1 month)	Warfarin (N = 187), INR target 2.0-3.0; placebo (N = 191), no aspirin

Table A.7: Trial-Specific Information (Part 2)

Trial	Dates	Inclusion criteria	Exclusion criteria	Arms and sample size
European Atrial Fibrillation Trial (EAFT)	Sept 1989 – Apr 1993	> 25 years old, TIA or minor stroke (< 3 months), AF (ECG-proven or paroxysmal)	Mitral stenosis, contraindications for OAC/AP, abnormal thyroid function, prosthetic valve, cardiac aneurysm, atrial myxoma, cardiothoracic ratio > 0.65, myocardial infarction (< 3 months), blood coagulation disorders, scheduled coronary surgery or carotid endarterectomy	OAC (N = 225), INR target 2.5-4.0; Aspirin (N = 230), 300 mg daily; Placebo (N = 214)
Primary Prevention of Arterial Thromboembolism in Atrial Fibrillation (PAATAF)	Jan 1990 – Dec 1996	> 60 years old, AF on ECG	Treatable causes of AF, previous stroke, rheumatic heart disease, myocardial infarction or cardiovascular surgery (< 1 year), Cardiomyopathy (EF < 40%), congestive heart failure, cardiac aneurysm, previous systemic or retinal embolism, OAC use (<3 months), contraindications to AP/OAC, pacemaker, life expectancy < 2 years	Standard Coumarin (N = 131), INR target 2.5-3.5; Low Coumarin (N = 122), INR target 1.1-1.6; Aspirin (N = 141), 150 mg daily
Stroke Prevention in Atrial Fibrillation 2 (SPAF-2)	Sept 1985 – Jun 1991	AF in previous 12 months	Rheumatic heart disease, mitral stenosis, prosthetic valve, previous indications or contraindications for OAC/AP, stroke/TIA (< 2 years), Age < 60 years without overt cardiovascular disease	Warfarin (N = 555), INR target 2.0-4.5; aspirin (N = 545), 325 mg daily
Stroke Prevention in Non-rheumatic Atrial Fibrillation (SPINAF)	May 1990 – Mar 1991	Male, AF (2 ECGs > 4 weeks apart)	OAC (<6 months), intermittent AF, prosthetic heart valve, mitral stenosis, active thromboembolic disease, coronary-artery bypass surgery, intracardiac thrombus, alcoholism, hemostasis disorder, PR interval time outside normal range, history of GI or intracranial hemorrhage, planned surgery, previous contraindication for OAC, rheumatic heart disease	Warfarin (N = 260), INR target 1.4-2.8; placebo (N = 265), aspirin withdrawn if approved

Table A.8: BLP Validation

	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 (A.3), 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 (A.4), 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.3. $\hat{Y}^o_{-j}(X_i)$ denotes regression forest predictions of the outcome estimated in the control groups of leave-out trials. $P_j \times \left(\hat{\tau}^o_{-j}(X_i) - \overline{\tau}^o\right)$ denotes the treatment probability in each trial interacted with demeaned treatment effect. $P_j \times \left(\hat{\tau}^c_{-j}(X_i) - \overline{\tau}^c\right)$ and $P_j \times \left(\hat{\tau}^r_{-j}(X_i) - \overline{\tau}^r\right)$ denote, respectively, the treatment probability in each trial interacted with demeaned CHADS_2 and residual components of the stroke treatment effect. The sample used in the validation regression excludes patients ineligible for warfarin, reducing the sample size to 4,720. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$