Are Guidelines Worth Following? Decision-Making Under Incomplete Evidence

Jason Abaluck
Leila Agha
David C. Chan*

September 2026

Abstract

Medical guidelines translate clinical evidence into treatment rules, but the evidence may identify treatment benefits more clearly than harms. We study anticoagulation for atrial fibrillation, combining randomized trials with Veterans Health Administration data. Trial evidence reveals substantial heterogeneity in stroke-prevention benefits but does not reliably identify how treatment-induced bleeding harms vary across patients. In the target population, patients with larger predicted stroke benefits also have higher untreated bleeding risk, making this ambiguity consequential. We evaluate existing guidelines and develop maximin and minimax-regret treatment rules robust to alternative harm structures. At a benchmark preventing 50% of preventable strokes, robust rules induce 33–41% fewer bleeds than random treatment and outperform existing guidelines. Finite-sample uncertainty generates noticeable instability in fine patient rankings but little instability in treatment assignments or welfare. Physician performance and the relationship between guideline adherence and skill also depend substantially on assumptions about how treatment-induced bleeding harms vary across patients.

*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. For helpful comments and suggestions on this paper, we thank Jacob Dorn, Ashvin Gandhi, Toru Kitagawa, Erzo Luttmer, Chuck Manski, Paul Novosad, Ziad Obermeyer, Alex Olssen, Stephen Ryan, Fiona Scott Morton, Sachin Shah, Advik Shreekumar, Jonathan Skinner, Doug Staiger, Mintu Turakhia, Stefan Wager, and Chris Walters. We thank Shubh Lashkary, Elliot Lee, Erick Ore Matos, Simone Pandit, Francis Peng, Jonatas Teixeira Prates, Melinda Xu, Sibo Zhou, Wyatt Zhou, and Jessie Zhu for invaluable research assistance. This paper builds on an earlier phase of the project, which was circulated as Abaluck, Agha, Chan, Singer, and Zhu (2021). We are grateful to Diana Zhu for her substantial contributions to that phase of the research. We thank Daniel Singer for his clinical and institutional guidance throughout the project and for connecting us with Carl van Walraven and the Atrial Fibrillation Investigators, who generously shared the AFI database for reanalysis.

1 Introduction

Modern medicine increasingly relies on clinical guidelines to translate scientific evidence into treatment rules. Yet the evidence base underlying many guidelines is incomplete in ways that matter for policy design. Randomized controlled trials are typically designed to detect average effects on outcomes and may reveal heterogeneity in benefits when patient-level data are pooled across multiple trials. By contrast, serious treatment harms are rarer and harder to study, especially when trials enroll healthier, less medically complex patients than those treated in practice.[1] As a result, evidence may identify patient-specific treatment benefits more clearly than patient-specific harms, especially for multimorbid patients for whom treatment decisions are most difficult and consequential.

When evidence is incomplete, determining the welfare-maximizing treatment rule becomes challenging. Optimal treatment depends on how patient-specific benefits and harms trade off across the population. When evidence is more informative about treatment benefits than harms, different plausible benefit-harm relationships can imply different treatment priorities and different welfare comparisons across policies. Finite trial samples can further limit the precision of patient-specific treatment-effect estimates. These challenges raise two fundamental questions: how should we use existing evidence when it leaves key welfare-relevant features of treatment unresolved, and how should we evaluate physician decisions in light of such incomplete evidence? In this paper, we develop an empirical framework to design and evaluate treatment rules under incomplete evidence.

Our empirical setting concerns anticoagulation guidelines for atrial fibrillation (AF), a common condition affecting more than 5 million people in the United States (Colilla et al. 2013). AF increases the risk of stroke approximately fivefold (Piccini and Fonarow 2016), and randomized trials show that anticoagulation reduces stroke risk while increasing the risk of life-threatening hemorrhage (Atrial Fibrillation Investigators 1995). Treatment, therefore, requires balancing stroke prevention against serious bleeding harms. To guide this decision, clinical guidelines recommend tailoring anticoagulation based on the $CHADS_2$ score and its successor, CHA_2DS_2-VASc (Gage et al. 2001; Fuster et al. 2006; Hirsh et al. 2008). These guidelines reflect heterogeneity in stroke risk but provide limited explicit guidance for incorporating heterogeneous bleeding harms.

We study anticoagulation decisions using administrative and clinical-note data from the Veterans Health Administration (VHA), covering nearly 5,800 physicians treating more than 130,000 patients newly diagnosed with atrial fibrillation. These data allow us to compare real-world treatment behavior with

1. Clinical guidelines often provide limited guidance for multimorbid patients, for whom polypharmacy may increase treatment harms (Tinetti, Bogardus, and Agostini 2004; Boyd et al. 2005). More generally, clinical trials often exclude patients at elevated risk of treatment complications, implying that trial populations may understate the incidence and heterogeneity of treatment harms in clinical practice (Abaluck, Agha, and Shah 2025).

the experimental evidence underlying guideline recommendations. We combine them with patient-level data from eight randomized trials in the Atrial Fibrillation Investigators (AFI) database (van Walraven et al. 2009) and use machine-learning methods to estimate heterogeneous treatment effects. We detect substantial heterogeneity in stroke-reduction benefits: predicted benefits are more than five times larger for patients in the top decile than for those in the bottom decile. By contrast, the randomized evidence does not establish a reliable relationship between patient characteristics and treatment-induced bleeding harms. Comparing the AFI trials with the VHA population reveals that VHA patients have substantially higher untreated bleeding risk. Within the VHA population, untreated bleeding risk more than triples between the bottom and top deciles of predicted stroke benefit. With higher $CHADS_2$ scores, both predicted stroke benefits and untreated bleeding risks increase monotonically. This positive relationship implies that uncertainty about heterogeneous bleeding harms can have important consequences for which patients to prioritize for treatment.

To represent the unresolved heterogeneity in treatment-induced bleeding harms, we consider two benchmark assumptions. Under Assumption *A*, treatment-induced bleeding harms do not vary predictably across patients, providing a scientific-null benchmark in the absence of reliable evidence of harm heterogeneity. Under Assumption *B*, treatment-induced bleeding harms are proportional to untreated bleeding risk, capturing the possibility that patients at higher baseline bleeding risk also experience greater treatment-induced harms. These assumptions are not intended to exhaust the possible patterns of harm heterogeneity, but rather to provide two contrasting benchmarks for a policy-relevant feature left unresolved by the randomized evidence. Because patients with the largest predicted stroke benefits also tend to have the highest untreated bleeding risks, treatment priorities can differ substantially across these two assumptions.

To translate estimated benefits and risks into treatment policy, we use calibrated treatment-effect estimates within an empirical welfare maximization framework (Kitagawa and Tetenov 2018). For each harm assumption, we characterize the corresponding optimal patient ordering and compare rules by the bleeds they induce while preventing the same number of strokes. Ambiguity about harms changes optimal treatment recommendations for up to 41% of patients, demonstrating that plausible assumptions about harms can materially alter the ranking of patients for treatment. Under Assumption *A*, the existing $CHADS_2$ and CHA_2DS_2-VASc guidelines perform well despite their simplicity, substantially reducing bleeding relative to random treatment at a given level of stroke prevention. Their relative performance markedly deteriorates under Assumption *B*, because the same scores that identify patients with greater stroke benefit also tend to prioritize patients with higher untreated bleeding risk. Thus, the welfare performance of established clinical guidelines depends substantially on how treatment-induced bleeding harms vary across patients.

Next, we develop treatment rules robust to ambiguity about treatment harms. Following the decision-theoretic literature on decision-making under incomplete evidence (Wald 1950; Manski 2004; 2011), we

characterize maximin and minimax-regret rules that are robust across the two benchmark harm structures. These rules are especially valuable when patients who benefit most from treatment also face a greater risk of harm, as in our setting and more generally when disease severity, comorbidity, and treatment risk are positively related. Empirically, at the same stroke-prevention benchmark, the robust rules induce 33–41% fewer bleeds than random treatment across Assumptions *A* and *B*. More broadly, they demonstrate how to implement ambiguity-robust treatment rules empirically when randomized evidence is informative about treatment benefits but leaves important uncertainty about how treatment harms vary across patients.

We also examine finite-sample uncertainty in estimated treatment benefits. Using repeated resampling and re-estimation of the randomized evidence, we find noticeable instability in fine patient rankings: 14.2% of patient pairs under Assumption *A* and 20.2% under Assumption *B* cannot be ordered the same way in at least 90% of estimated benefit realizations. Yet this ranking instability has much smaller consequences for policy and welfare. Incorporating sampling uncertainty changes the minimax-regret treatment assignment for only 0.12% of patients. At the 50%-stroke-prevention benchmark, the welfare cost of uncertainty about the realized benefit estimates is 1.56–2.06 per 10,000 patients, compared with about 17 per 10,000 from using the sampling-aware minimax-regret rule rather than the assumption-specific optimum under the mean benefit estimates. Thus, substantial uncertainty in fine patient rankings need not translate into comparable instability in treatment assignments or welfare; in this application, ambiguity about treatment harms is substantially more consequential for welfare than finite-sample uncertainty in estimated benefits.

Finally, we evaluate physician decision-making under the same uncertainty. Physician decisions provide both an object of evaluation and a source of potential decision rules. Motivated by the wisdom-of-crowds idea that aggregating many individual judgments can reveal useful information (Surowiecki 2005; Gillen, Plott, and Shum 2017), we construct a physician-consensus rule based on observed treatment patterns. Because this rule depends only on observable patient characteristics, we can evaluate it within the same trial-based welfare framework as clinical guidelines and algorithmic treatment rules. To ask how actual physician behavior relates to optimal guidelines, we estimate a skill-threshold model that separates ordering skill from treatment propensity (Chan, Gentzkow, and Yu 2022). This model lets us compare physicians as if they faced the same patient population. Both the consensus rule and the median-skill physician perform worse than random treatment under Assumption *A* but better under Assumption *B*, consistent with physicians collectively behaving as if they account for bleeding risk. At the individual level, physicians exhibit modest variation in ordering skill but wide variation in treatment propensities. The share of physicians who outperform existing guidelines depends substantially on the assumed harm structure: no simulated physician outperforms the $CHADS_2$ guideline under Assumption *A*, but 44.5% outperform it under Assumption *B*.

Our paper contributes to several strands of literature at the intersection of decision theory, heterogeneous

treatment effects, algorithmic decision-making, and clinical guideline evaluation. First, we contribute to the literature on decision-making under ambiguity and incomplete evidence. Statistical decision theory has long studied how decision-makers should act when key features of the environment are unknown, including through maximin and minimax-regret criteria (Wald 1950; Savage 1951; Manski 2004; 2011; Gilboa and Marinacci 2013). In clinical and public-health settings, this problem arises because guidelines and empirical studies must often support decisions under partial knowledge about patient risks, treatment response, and feasible identifying assumptions (Manski 2013; Mullahy et al. 2021). More broadly, internally valid causal evidence need not identify the policy-relevant objects needed for decision-making when treatment effects and welfare tradeoffs are heterogeneous (Deaton 2010). We provide an empirical implementation of decision-making under ambiguity in a setting with patient-level randomized evidence, existing clinical guidelines, and physician treatment decisions. In this rich setting, we characterize robust treatment rules and evaluate formal guidelines and physician behavior within a common welfare framework.

Second, we contribute to the literature on heterogeneous treatment effects and policy learning. A growing body of work uses machine-learning methods to estimate conditional average treatment effects and to construct welfare-improving treatment rules (Wager and Athey 2018; Athey, Tibshirani, and Wager 2019; Athey and Wager 2021). Empirical welfare maximization provides a direct link between estimated treatment effects and policy choice (Kitagawa and Tetenov 2018). We extend this literature by studying a setting in which randomized evidence supports policy-relevant heterogeneity in benefits but does not identify the corresponding heterogeneity in harms. We further show that finite-sample uncertainty in estimated benefits can generate noticeable instability in fine patient rankings while producing relatively little instability in treatment assignments or welfare.

Third, our results relate to the literature on algorithmic decision-making and expert discretion. Prior work shows that structured decision rules can improve decisions relative to human judgment in settings such as clinical care and bail (Abaluck et al. 2016; Mullainathan and Obermeyer 2022; Kleinberg et al. 2018). Other work studies how experts respond to algorithmic recommendations and when human discretion complements or overrides algorithmic advice (Hoffman, Kahn, and Li 2015; Agarwal et al. 2023; Angelova, Dobbie, and Yang 2026). Our contribution is to emphasize that comparisons between algorithms and experts depend on the welfare mapping from predictions to outcomes. When the relationship between patient characteristics and treatment-induced harms is unresolved, whether a decision rule improves welfare can depend on the maintained harm structure.

Finally, we contribute to the medical and health economics literature on clinical guidelines and physician behavior. Prior research documents substantial non-adherence to clinical guidelines (Schuster, McGlynn, and Brook 1998), including atrial fibrillation risk scores (Chapman et al. 2017). This paper builds on an earlier working paper from the same clinical setting that examined physicians' awareness of and adherence

to the $CHADS_2$ guideline (Abaluck et al. 2020). The present paper asks a different question: how guidelines and physician decisions should be evaluated when the evidence identifies heterogeneous treatment benefits more clearly than heterogeneous harms.[2] Related economics research studies physician decision-making and the role of formal rules versus clinical discretion (Currie and MacLeod 2020; Boone 2025; Cuddy and Currie 2026). We evaluate both formal guidelines and physician behavior within a unified benefit–harm framework. Our results show that assessments of guideline adherence and physician performance depend substantially on how treatment-induced harms vary across patients.

The remainder of this paper is organized as follows. Section 2 describes the clinical setting and data. Section 3 characterizes the evidence on treatment benefits, bleeding harms, and sampling uncertainty. Section 4 evaluates existing, optimal, and robust treatment rules under alternative harm assumptions and finite-sample uncertainty. Section 5 evaluates physician decisions under these assumptions. Section 6 concludes.

2 Setting, Data, and Guideline Use

2.1 Clinical Setting

Atrial Fibrillation. Atrial fibrillation (AF) is the most common cardiac arrhythmia. It affects more than 5 million Americans (Colilla et al. 2013); for adults older than 40, one in four will develop the condition (Hsu et al. 2016). AF increases stroke risk by fivefold 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.[3] While anticoagulation reduces stroke risk by inhibiting clot formation by 68% on average, it also increases the risk of major bleeding by more than double (Atrial Fibrillation Investigators 1995; Kearon et al. 2012). Given these substantial potential benefits and harms, clinicians evaluating AF patients must decide which patients are sufficiently likely to benefit from anticoagulation to justify the accompanying bleeding risk.

Guidelines for Treating AF. Efforts to improve anticoagulation targeting have largely focused on predicting stroke risk, with the intuition that patients at higher baseline stroke risk have more to gain from stroke prevention. Earlier studies re-analyzed data from the control arms of randomized trials of AF patients to

2. The earlier paper evaluated broader guideline awareness and stricter adherence under a maintained assumption of homogeneous treatment-induced bleeding harms (i.e., the mean-independent harms we now consider as Assumption A). The present paper treats uncertainty about heterogeneous harms as central and develops new ambiguity-robust treatment rules and analyses of physician performance.

3. In our sample, fewer than 2% of patients are prescribed alternative anticoagulants. Novel oral anticoagulants (NOACs) were introduced near the end of our sample period, with FDA approval of dabigatran in 2010, rivaroxaban in 2011, and apixaban in 2012. Non-inferiority trials show they are similarly effective at preventing stroke, with possibly lower bleeding risk (Lane and Lip 2012). Warfarin remains the mainstay drug for anticoagulation in AF (Hsu et al. 2016).

identify hypertension and prior stroke as important risk factors for stroke (Atrial Fibrillation Investigators 1995; Stroke Prevention in Atrial Fibrillation Investigators 1995). Using risk factors identified from the trial data, the $CHADS_2$ score was first validated as a stroke prediction rule in 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 assigns one point each for congestive heart failure, hypertension, age $\geq$ 75 years, and diabetes, and two points for prior stroke. 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—that added vascular disease, age between 65 and 74 years, and female sex (Lip et al. 2010). The CHA_2DS_2-VASc score has since become the standard of practice, though it was much less widely recognized during our study period, from 2002 to 2013. These scores define simple orderings of patients by stroke-related risk factors, but whether such orderings imply welfare-improving treatment rules depends on how treatment-induced bleeding harms vary among patients placed earlier versus later in the same ordering. We return to this evidence question in Section 3.

Despite the widespread recognition of these stroke-risk scores, 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 about bleeding risk due to frailty and multi-morbidity to justify non-adherence to $CHADS_2$-based treatment recommendations (Fawzy, Olshansky, and Lip 2019).[4] AF frequently coexists with comorbid conditions that may increase bleeding risk, including unstable gait, renal insufficiency, hypertension, and older age itself. Although formal bleeding risk scores such as HAS-BLED were developed to summarize bleeding risk, they do not enter anticoagulation guidelines symmetrically with stroke-risk scores. Among patients who are candidates for anticoagulation, stroke-risk scores define explicit treatment thresholds, whereas bleeding scores guide risk-factor modification, monitoring, and clinical judgment rather than definitive treatment recommendations.[5] This matters

4. More generally, disease-specific guidelines can generate complex treatment burdens for multimorbid patients. For example, Boyd et al. (2005) conclude that applying guidelines for common chronic diseases to a representative older patient with chronic obstructive pulmonary disease, type 2 diabetes, osteoporosis, hypertension, and osteoarthritis would require twelve or more medications. As Tinetti, Bogardus, and Agostini (2004) note, this raises the question of whether "what is good for the disease is always best for the patient" (p. 2870).

5. HAS-BLED was published in 2010 (Pisters et al. 2010; Lane and Lip 2012). Current U.S. guidelines retain this asymmetry: they state that bleeding risk scores should not be used in isolation to determine oral anticoagulation eligibility, but rather to identify and modify bleeding risk factors and to inform medical decision-making (Joglar et al. 2024). The clinical consensus of "contraindication" is thus narrower than high bleeding risk alone: it refers to specific clinical circumstances that preclude anticoagulation, such as nonreversible serious bleeding or serious bleeding related to recurrent falls when the cause is not treatable

because stroke-risk and bleeding-risk assessments share several risk factors. Older age, hypertension, and prior stroke, for example, enter both $CHADS_2$ and HAS-BLED; a patient with all three risk factors would be recommended anticoagulation and would also be classified as having high bleeding risk.

2.2 Data

Veterans Health Administration Data. To study real-world AF treatment decisions, we identify patients with a new AF diagnosis using electronic medical records from the Veterans Health Administration from October 2002 to December 2013.[6] 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).[7]

For each patient, we observe characteristics that may influence the anticoagulation decision, including demographics, comorbidities, laboratory test results, body measurements, and blood pressure readings. We use these characteristics to construct the $CHADS_2$ score and to match the clinical characteristics recorded in the randomized trial data. We measure anticoagulation using VHA prescription records for warfarin or a novel oral anticoagulant (e.g., apixaban, dabigatran, edoxaban, rivaroxaban). Because novel oral anticoagulants were much less commonly used during our cohort period, we refer to warfarin and anticoagulation interchangeably.[8] The VHA records include prescriptions that are dispensed by the VHA as well as prescriptions that are paid for by the VHA.

Atrial Fibrillation Investigators Database. To characterize the randomized evidence base for anticoagulation, we use patient-level data from the Atrial Fibrillation Investigators database (hereafter, AFI database) (van Walraven et al. 2009). The AFI database includes the same trials used to develop the clinical guidelines and, to our knowledge, represents the best available causal evidence on warfarin effects in patients with AF. We analyze patient-level observations from eight trials in which patients were randomized to anticoagulants versus placebo or control.[9]

For each patient in the AFI database, we observe treatment randomization status and subsequent stroke and bleeding events. The investigators harmonized several patient characteristics across trials,

(Joglar et al. 2024).

6. For a limited component of outcome ascertainment, we supplement VHA records with Medicare claims available for our cohort of veterans in the VHA data infrastructure; Appendix A.2 provides details.

7. We require no prior AF diagnosis within three years, an electrocardiogram near the initial diagnosis, no prior anticoagulation, attribution to a VHA primary care physician or cardiologist, and minimum physician-volume restrictions. Appendix Table A.1, Panel A, details the cohort construction steps.

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

9. The original AFI database contains ten trials. For our analysis, we define aspirin-only arms as untreated with anticoagulation and exclude observations receiving low-dose warfarin or low-dose warfarin plus aspirin. After these modifications, eight trials remain with both treatment and control arms. Appendix Table A.1, Panel B, details the sample construction steps.

including all variables underlying the $CHADS_2$ score and additional variables such as demographics, height, weight, blood pressure, hemoglobin, smoking status, comorbidities, and history of transient ischemic attack (TIA), stroke, anginal symptoms, and myocardial infarction. While these AFI covariates include plausible predictors of stroke and allow us to construct both the $CHADS_2$ and CHA_2DS_2-VASc scores, they omit several bleeding-relevant HAS-BLED components available in the VHA data, including abnormal renal or liver function, prior bleeding, and concomitant NSAID or alcohol use (Appendix Table A.2). Appendix Tables A.3 and A.4 summarize the trial dates, eligibility criteria, treatment arms, and sample sizes; the eligibility criteria are relevant for interpreting how well the randomized evidence represents patients treated in routine practice, an issue we return to in Section 3.

2.3 Guideline Use

We use the VHA electronic health record to document the salience of the $CHADS_2$ guideline in clinical practice, focusing on $CHADS_2$ rather than its later CHA_2DS_2-VASc variant because the former was much more widely recognized during our cohort period.[10] The $CHADS_2$ score also served as a benchmark for guideline-concordant care: formal quality measures recommended anticoagulation for high-risk AF patients defined by $CHADS_2$, while allowing documented medical or patient exceptions.[11]

Based on the 2006 ACC and 2008 ACCP guidelines, AF patients with a $CHADS_2$ score of 0 or 1 could forego warfarin, while those with a score of 2 or higher were recommended anticoagulation (Hirsh et al. 2008; Fuster et al. 2006). Abaluck et al. (2020) show that $CHADS_2$ became widely known in VHA practice but only modestly affected treatment patterns. By 2013, nearly 70% of physicians had mentioned the score at least once, yet anticoagulation remained well below guideline recommendations. Around a physician's first documented mention, anticoagulation declined modestly among low-score patients, with little change among higher-score patients. Thus, $CHADS_2$ was salient but did not determine treatment decisions.

3 The Evidence Base

Clinical guidelines seek to translate evidence on treatment benefits and harms into treatment recommendations. Doing so requires evidence on how treatment effects and underlying clinical risks vary across patients. In the case of anticoagulation for atrial fibrillation, treatment reduces the risk of stroke but may increase the risk of bleeding. This section defines the target objects relevant to this benefit-harm trade-off

10. Specifically, while 24% of patient encounters in our data are with physicians who have previously mentioned the $CHADS_2$ score, fewer than 2% of patient encounters are with physicians who have previously mentioned CHA_2DS_2-VASc in their notes.

11. Measure #326/NQF #1525 allowed exceptions for medical reasons, including bleeding risk, and for patient reasons such as refusal or noncompliance (Centers for Medicare & Medicaid Services 2017).

and assesses what randomized-trials and VHA data can tell us about them.

Let $Y_i^s(d) \in \{0,1\}$ denote whether patient i experiences a stroke within one year under treatment status $d \in \{0,1\}$, and let $Y_i^b(d) \in \{0,1\}$ denote whether patient i experiences a major bleeding event. Let X_i denote a vector of patient characteristics. The patient-specific stroke-reduction *benefit* of anticoagulation is

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

while the patient-specific treatment-induced bleeding *harm* is

$$H^*(x) \equiv E\left[Y_i^b(1) - Y_i^b(0) \,\middle|\, X_i = x\right]. \tag{2}$$

These causal objects describe how the benefits and harms of anticoagulation vary across patients.

We also consider untreated bleeding *risk*,

$$R^*(x) \equiv \Pr\Big(Y_i^b(0) = 1 \,\Big|\, X_i = x\Big). \tag{3}$$

Unlike $B^*(x)$ and $H^*(x)$, untreated bleeding risk is not a treatment effect; it captures baseline heterogeneity in patients' propensity to experience bleeding events. Such baseline risks are central to many clinical guidelines—the CHADS_2 score, for example, is based on predictors of untreated stroke risk rather than direct estimates of treatment-effect heterogeneity.[12]

3.1 The Randomized Trial Evidence

Randomized controlled trials provide the strongest causal evidence for learning about the treatment-effect objects defined above. By randomly assigning anticoagulation, trials make treatment status independent of potential outcomes, allowing estimation of the causal effects of treatment on stroke and bleeding. Consistent with random assignment, Appendix Table A.5 shows balance between treatment and control arms on patient demographics, medical history, smoking status, and other clinical measures, including blood pressure and hemoglobin. Because patient-level trial data include treatment assignment, baseline characteristics, and adjudicated clinical outcomes, they provide a natural starting point for assessing whether the effects of anticoagulation vary across patients. Trial outcome measurement is especially useful in this setting because administrative data may not distinguish new strokes or bleeding events from historical diagnoses.[13]

12. We also estimate baseline stroke risk in constructing the ML proxies below, but do not introduce separate notation for it because the policy framework depends on treatment-induced benefits and harms.

13. For example, events such as strokes are often recorded in administrative or electronic health record data using diagnosis codes, which may reflect either a new event or a historical diagnosis. In our audit of VHA data at the Palo Alto VA, approximately 40% of patients recorded as experiencing a current stroke in diagnosis codes were revealed on detailed chart review to have only a history of stroke without recurrence.

At the same time, randomized trials have important limitations for detecting treatment-effect heterogeneity. First, as of 2004, when the $CHADS_2$ score was first recommended for clinical practice, fewer than 10,000 AF patients had been enrolled in randomized trials studying anticoagulation, while the worldwide population of patients with AF likely exceeded 30 million individuals (Chugh et al. 2014). Second, even when trials identify average treatment effects, they may be underpowered to detect treatment-effect heterogeneity, and naive subgroup analyses can be prone to both false-positive and false-negative findings (Kent et al. 2010). Third, trial populations may differ from the patients encountered in clinical practice. Patients at higher risk of treatment-related harms, such as elderly or multimorbid patients, are often underrepresented in randomized trials, raising concerns about the external validity of trial-based estimates (Rothwell 2005; Abaluck, Agha, and Shah 2025).

These general limitations also apply to the AFI database. Appendix Tables A.3 and A.4 report the sample sizes and inclusion and exclusion criteria of the RCTs included in the AFI database. Many trials excluded patients with a prior indication or contraindication for oral anticoagulant or antiplatelet therapy, avoiding randomization of patients for whom treatment or non-treatment was already clinically compelled. The trials also commonly excluded patients with safety concerns such as uncontrolled hypertension, bleeding risk factors, predisposition to intracranial hemorrhage, hemostasis disorders, prior gastrointestinal or intracranial hemorrhage, alcoholism, and related conditions. In addition, few AFI covariates map to bleeding risk factors. This limited coverage of bleeding-risk factors restricts the dimensions of treatment-harm heterogeneity that AFI trials can study (see Appendix Table A.2). These limitations are especially important for evaluating heterogeneity in bleeding harms, which are relatively rare in the trials.

3.2 Treatment-Effect Estimation and Validation

Patient characteristics relevant to anticoagulation decisions can be highly multidimensional, and any treatment policy that varies across patients requires empirical inputs summarizing how expected benefits and harms vary with those characteristics. Because the true patient-specific treatment effects $B^*(x)$ and $H^*(x)$ are not directly observed, we estimate functions of patient characteristics that predict variation in these effects. We call these predicted treatment-effect signals machine-learning (ML) proxies. Let $Z_B(x)$ denote the ML proxy for the stroke-reduction benefit $B^*(x)$, and let $Z_H(x)$ denote the ML proxy for the treatment-induced bleeding harm $H^*(x)$. These proxies summarize the trial-based signal of treatment-effect heterogeneity and serve as empirical inputs for the treatment policies studied below.

We estimate $Z_B(x)$ and $Z_H(x)$ using patient-level data from the eight randomized trials in the AFI database. The prediction procedure combines causal forests (Wager and Athey 2018) and LASSO methods (Belloni, Chernozhukov, and Hansen 2014), allowing the proxies to capture both interactions among discrete patient characteristics and smoother relationships involving continuous predictors. Because the

proxies must be applied to the VHA population, we restrict the RCT prediction models to covariates observed in both the AFI and VHA data. To reduce sensitivity to finite-sample variation, we aggregate predictions across repeated 95% subsamples of the trial data, following the logic of bagging (Breiman 1996). Appendix A.1 provides additional details on the prediction procedure.

We validate the ML proxies using the AFI database's multi-trial structure. Following Chernozhukov et al. (2025), we implement a leave-one-trial-out best linear predictor (BLP) exercise. For each trial, we train the ML proxies on the remaining seven trials and then estimate the best linear predictor for the hold-out trial using OLS regression that interacts treatment with the proxy. The interaction coefficient tests whether the proxy (linearly) predicts treatment-effect variation in the hold-out trial. This cross-trial validation exercise asks whether the proxy captures heterogeneity that is stable across trial populations, rather than idiosyncratic variation within a particular trial. We use the BLP primarily as a validation tool. For descriptive analyses in the remainder of this section, let $B_{\mathrm{BLP}}(x)$ denote the fitted stroke-benefit value obtained by applying the estimated BLP projection to $Z_B(x)$. This measure captures the component of variation in predicted stroke benefit supported by randomized evidence. We use $B_{\mathrm{BLP}}(x)$ to describe how stroke benefits vary across patients and how they relate to untreated bleeding risk and the CHADS_2 score.

The validation results in Table 1 differ starkly between stroke and bleeding outcomes. The stroke-benefit proxy strongly validates across trials: patients predicted to benefit more from anticoagulation experience larger treatment-induced reductions in stroke risk in hold-out trials (Column 1). By contrast, the bleeding-harm proxy does not systematically predict treatment-induced bleeding in hold-out trials (Column 2), and even ML predictions of bleeding risk derived from the RCT data do not generalize across trials (Column 3). Figure 1 further shows that, unlike the stroke-benefit proxies, the bleeding-harm proxies exhibit almost no patient-level dispersion relative to the average bleeding treatment effect in the RCTs. Thus, the flexible ML procedure provides little evidence of usable, cross-trial-stable heterogeneity in causal bleeding harms. A simpler analysis nevertheless provides some evidence that treatment-induced bleeding harms may vary with untreated bleeding risk. Figure 2 relates treatment-induced bleeding harms in the RCTs to predicted untreated bleeding risk, $R_C(x)$. Within trials, patients with higher predicted bleeding risk experience larger treatment-induced bleeding harms. Across trials, however, the relationship is imprecise and opposite in sign. The randomized evidence therefore does not establish a quantitative relationship between baseline bleeding risk and treatment-induced harm that can be confidently transported to the VHA population.

3.3 Stroke Benefits and Bleeding Risks in the Target Population

We next estimate two measures of untreated one-year bleeding risk in the VHA data; both predict bleeding risk in the absence of anticoagulation and are therefore baseline-risk predictions rather than treatment effects. The common-covariate measure, $R_C(x)$, uses only characteristics recorded in both the VHA and

AFI data. We apply the same VHA-estimated model to patients in both samples, allowing us to compare bleeding-risk distributions and their relationships with trial evidence on a common covariate basis. The full-covariate measure, $R(x)$, uses the richer set of characteristics available in the VHA, including bleeding-relevant factors not recorded in the AFI trials. We use $R_C(x)$ for cross-sample and trial-based comparisons and $R(x)$ for the main target-population and policy analyses. Appendix A.2 describes their estimation and covariates.

Figure 3 compares the marginal distributions of stroke-reduction benefits and untreated bleeding risk in the VHA and RCT samples. Panel A shows substantial variation in predicted stroke benefits in the VHA population: patients in the bottom decile of $B_{\mathrm{BLP}}(x)$ have average predicted benefits of 1.4 percentage points, compared with 9.3 percentage points for patients in the top decile. At the same time, the VHA distribution of predicted stroke benefits lies largely within the range of trial-specific average stroke reductions in the RCT sample. Panel B compares $R_C(x)$ across the two samples. Its mean is 7.4% among VHA patients and 4.1% when the same fitted risk model is applied to RCT patients. Observed control-arm bleeding rates in the RCTs are lower still. This gap suggests that trial participants may have been selected on unmeasured characteristics associated with lower bleeding risk, although differences in bleeding measurement between the RCT and VHA data could also contribute.

Figure 4 shows that stroke benefits and bleeding risks are positively related in the VHA population. The Pearson correlation between $B_{\mathrm{BLP}}(x)$ and $R(x)$ is 0.41, and the Spearman rank correlation is 0.48. This relationship is also clinically meaningful: untreated bleeding risk rises from 3.5% in the bottom decile of predicted stroke benefit to 11.9% in the top decile. Thus, patients predicted to gain the most from anticoagulation also have higher baseline bleeding risk. Figure 5 relates these objects to the CHADS_2 score used in practice. Predicted stroke benefits increase with the CHADS_2 score, consistent with the clinical role of CHADS_2 as a stroke-risk score. Untreated bleeding risks also increase with CHADS_2, reflecting the fact that age and comorbidities predict both stroke and bleeding outcomes. However, the figure also shows substantial within-score variation in both outcomes. Patients with the same guideline score can differ meaningfully in predicted stroke benefits and bleeding risks, implying that coarse guideline categories may group together patients with different benefit–risk profiles.

These descriptive patterns create a central challenge for treatment prioritization. Under a scientific-null benchmark in which treatment-induced bleeding harms do not vary predictably across patients, prioritizing patients with the largest stroke benefits would be relatively straightforward. If, instead, treatment-induced bleeding harms increase with untreated bleeding risk, as suggested by the within-trial evidence in Figure 2, then the same patients who benefit most from treatment may also face the greatest bleeding harms. Because the randomized evidence does not establish a reliable relationship between treatment-induced bleeding harms and patient characteristics, the welfare implications of existing guidelines and alternative

treatment rules depend on how such harms vary across patients.

4 Evaluating Treatment Rules Under Scientific Uncertainty

Anticoagulation decisions require translating incomplete evidence into patient-level treatment recommendations. The evidence leaves two distinct forms of uncertainty relevant for policy. For stroke benefits, cross-trial validation supports a common heterogeneous treatment-effect signal, but finite trial samples generate sampling uncertainty in those estimates. For bleeding harms, the evidence does not establish a sufficiently stable conditional relationship to support a single model of harm heterogeneity, so we study policy under alternative harm structures.

This section develops a framework for evaluating treatment rules under these evidentiary limitations. We first translate validated benefit information into patient orderings and treatment thresholds using empirical welfare maximization. We then evaluate existing clinical guidelines under alternative harm structures, characterize the resulting assumption-specific optimal rules, and introduce treatment rules that are robust to this ambiguity. Finally, we incorporate empirical finite-sample uncertainty in estimated benefits and assess its implications for patient rankings and welfare.

4.1 From Evidence to Treatment Rules

Treatment rules require two ingredients: a patient ordering and a threshold. The ordering determines which patients are prioritized for treatment; the threshold determines how far down that ordering treatment extends. This view applies to existing guidelines, optimal rules, and robust rules. This distinction matters because the evidence summarized above primarily informs how patients differ in stroke-prevention benefit, whereas policy evaluation also requires deciding where to set treatment thresholds. Throughout this paper, we use *ordering* to denote the priority relation induced by a score or rule, *ranking* to denote the realized prioritization when that ordering is applied to a finite empirical sample, and *rank* to denote a patient's numerical position within that ranking.

Let $Z_B(x)$ denote the validated proxy for stroke-prevention benefit, oriented so that larger values correspond to larger expected benefit. Section 3 used a best linear predictor (BLP) projection to validate this proxy as a source of treatment-effect heterogeneity across trials. For policy evaluation, however, a validated ordering alone is not sufficient: benefits must be placed on a cardinal scale so that they can be compared with harms. A calibrated benefit $s(Z_B(x))$ defines threshold rules of the form $s(Z_B(x)) \geq t$, where different values of t correspond to different benefit cutoffs.[14]

14. After the harm structures are introduced below, a benefit cutoff t corresponds to the preference-weighted harm cutoff $t = \lambda\overline{H}$ under Assumption A; see Section 4.3. We use the same calibrated benefit object in all subsequent rule comparisons to put stroke-prevention benefits on a common scale. This lets the A-optimal, B-optimal, and robust rules discussed below differ only

We estimate this calibration using a criterion motivated by empirical welfare maximization (EWM) (Kitagawa and Tetenov 2018). Let ψ_i^B denote the cross-fitted augmented inverse-probability-weighted (AIPW) pseudo-outcome for the stroke-prevention benefit of treatment, constructed from the randomized trial data. Let $s \in \mathcal{S}$ denote a candidate monotone transformation of Z_B with values in $[0, 1]$. For any benefit threshold t, s induces the treatment rule $\mathbf{1}(s(Z_{B,i}) \geq t)$ and implies an empirical welfare gain $\mathbb{P}_n[(\psi_i^B - t)\mathbf{1}(s(Z_{B,i}) \geq t)]$, where $\mathbb{P}_n$ denotes the sample average. We choose the calibration that maximizes this empirical welfare gain averaged over benefit thresholds:

$$\widehat{s}_{\text{EWM}} = \arg\max_{s \in \mathcal{S}} \int_0^1 \mathbb{P}_n\left[(\psi_i^B - t)\mathbf{1}(s(Z_{B,i}) \geq t)\right] dt = \arg\max_{s \in \mathcal{S}} \mathbb{P}_n\left[\psi_i^B s(Z_{B,i}) - \frac{1}{2}s(Z_{B,i})^2\right]. \quad (4)$$

The integration range reflects the bounded, nonnegative scale of stroke-prevention benefits. The second equality follows from $s(Z_{B,i}) \in [0, 1]$; equivalently, the criterion fits a bounded calibrated benefit to the pseudo-outcome by least squares. We implement this calibration using $s(z) = \Lambda(g(z))$, where $\Lambda(\cdot)$ is the logistic link and $g(\cdot)$ is a monotone spline. This preserves the ordering in Z_B while placing calibrated benefits in $(0, 1)$. Appendix A.3 provides additional details and relates this calibration to BLP.

We define the EWM-calibrated benefit as $B_{\text{EWM}}(x) = \widehat{s}_{\text{EWM}}(Z_B(x))$. For the remainder of the main text, we write $B(x) \equiv B_{\text{EWM}}(x)$, since this is the calibrated stroke-benefit object used in policy comparisons unless otherwise noted. Below, we show that the substantive rule comparisons are similar when calibrated benefits are instead constructed from the linear BLP projection, B_{BLP}.

We next define the harm structures used to evaluate treatment rules. Let $\overline{H}$ denote the average treatment-induced bleeding harm and let $R(x)$ denote the predicted untreated bleeding risk, with population average $\overline{R}$. We consider two assumptions on the harm structure.

$$H_A(x) = \overline{H}; \quad (5)$$

$$H_B(x) = \frac{\overline{H}}{\overline{R}} R(x). \quad (6)$$

These assumptions describe the policy-relevant conditional mean of treatment-induced bleeding harms. Assumption A is a scientific-null benchmark: absent reliable evidence of predictable heterogeneity in treatment-induced bleeding harms, conditional mean harm is treated as constant, $H_A(x) = \overline{H}$. Assumption B instead allows conditional mean harms to rise in proportion to untreated bleeding risk, consistent with the within-trial relationship documented in Figure 2, and is normalized to preserve the same average harm. We use these as two empirically interpretable benchmark harm structures rather than as an exhaustive set of possible relationships between patient characteristics and treatment-induced harms.

For a harm structure $\omega \in \{A, B\}$ and preference weight $\lambda \geq 0$, define the expected net benefit from

because of changes in the harm structure, not because the benefit object is recalibrated under each harm structure.

treating a patient with covariates x and the welfare of treatment rule D as

$$U_\omega(x;\lambda) = B(x) - \lambda H_\omega(x), \tag{7}$$

$$W_\omega(D;\lambda) = E\left[D(X)U_\omega(X;\lambda)\right]. \tag{8}$$

The welfare-maximizing rule treats patients for whom $U_\omega(x;\lambda) \geq 0$.[15] Under mean-independent harms, treatment is optimal when $B(x) \geq \lambda\overline{H}$, so the generic benefit threshold t used in the calibration has the direct interpretation $t = \lambda\overline{H}$. Under proportional harms, the same calibrated benefit object is instead compared with the patient-specific threshold $\lambda(\overline{H}/\overline{R})R(x)$. The two harm assumptions therefore do not change the definition of stroke benefit; they change the harm term against which the common benefit object $B(x)$ is compared.[16]

Any ordering of patients can be combined with different thresholds to trace out the number of strokes prevented and bleeds induced as treatment expands. Our comparisons hold stroke prevention fixed and ask how many bleeding harms are induced by different rules to prevent the same number of strokes.[17] This fixed-benefit comparison separates the quality of the patient ordering from the overall aggressiveness of the treatment threshold and provides a common scale for comparing existing guidelines, assumption-specific optimal rules, and robust rules.

4.2 Existing Guidelines Under Competing Harm Assumptions

We begin by evaluating existing score-based guidelines before deriving new treatment rules. This provides a benchmark for the rest of the section and shows how a guideline's welfare implications depend on the structure of bleeding harms. As described above, anticoagulation guidelines have historically relied on stroke-risk scores to determine treatment eligibility. We focus on the two most prominent score-based rules in our setting, the CHADS_2 and CHA_2DS_2-VASc scores.

Each score induces an ordering of patients. Let Ω_G denote the ordering implied by guideline score G, with patients assigned higher priority for treatment when their score is higher. Combining this ordering

15. Our benchmark welfare criterion is additive in expected stroke prevention and treatment-induced bleeding harm. Consequently, policy evaluation requires the marginal treatment effects on stroke and bleeding, but not their joint distribution. More general welfare functions allowing interactions between stroke and bleeding outcomes would generally require additional information about their joint potential-outcome distribution, which we do not attempt to identify.

16. The framework abstracts from a separate non-bleeding burden of anticoagulation, such as drug-taking or monitoring costs. A patient-invariant burden would enter net benefit as an additive treatment cost. Under Assumption *A*, this simply raises the benefit threshold. Under Assumption *B*, the added constant burden is equivalent, after rescaling, to placing positive weight on the mean-independent harm component. Thus, a common non-bleeding treatment burden moves the proportional-harm case toward the mean-independent case rather than requiring a separate analysis. Patient-specific treatment burdens would add another source of heterogeneity.

17. Because the benefit–harm frontiers in our main comparisons do not cross over the relevant range, the qualitative conclusions are not driven by the particular stroke-prevention benchmark used for scalar summaries; a frontier-wide average over stroke-prevention benchmarks would preserve the same qualitative ordering.

with different thresholds traces out a benefit-harm frontier. Because the scores are discrete, moving the threshold through a score cell means treating some but not all patients in that cell. In the empirical implementation, we consider outcomes we would obtain if we randomly treated some but not all patients within a score cell, allowing us to compare guideline orderings over the full range of treatment intensities rather than only at a single recommended cutoff.

Figure 6 plots the benefit-harm frontiers induced by the $CHADS_2$ and CHA_2DS_2-VASc orderings under the two harm assumptions. Under mean-independent harms, both guidelines outperform a random ordering. At the benchmark that prevents 50% of preventable strokes, $CHADS_2$ and CHA_2DS_2-VASc induce 32.0% and 30.5% fewer bleeds than random treatment, respectively. Thus, when bleeding harms are unrelated to patient characteristics, the existing stroke-risk guidelines capture meaningful information about which patients have larger stroke-prevention benefits. Under proportional harms, their performance deteriorates sharply. At the same 50%-stroke-prevention benchmark, $CHADS_2$ induces 3.0% fewer bleeds than random treatment, while CHA_2DS_2-VASc induces 1.1% more bleeds. Under this assumption, the patients prioritized by stroke-risk scores also tend to have high untreated bleeding risk, so the guideline ordering trades off more sharply against bleeding harms.

Although CHA_2DS_2-VASc is newer and more granular than $CHADS_2$, it performs slightly worse under both harm assumptions, especially under Assumption B. This suggests that adding granularity to a guideline does not necessarily improve welfare when the additional detail continues to prioritize the same stroke-risk margin while leaving the structure of bleeding harms unresolved.

4.3 Assumption-Specific Optimal Treatment Rules

We next characterize the treatment rules optimal under each harm assumption. For a preference weight λ, let $D_A^*(\cdot;\lambda)$ denote the rule that maximizes welfare under Assumption A, and let $D_B^*(\cdot;\lambda)$ denote the rule that maximizes welfare under Assumption B. We call these the A-optimal and B-optimal treatment rules, respectively. Because they recommend treatment whenever the expected net benefit in Equation (7) is positive, these rules are

$$D_A^*(x;\lambda) = \mathbf{1}(B(x) \geq \lambda\overline{H}), \tag{9}$$

$$D_B^*(x;\lambda) = \mathbf{1}\left(B(x) \geq \lambda\frac{\overline{H}}{\overline{R}}R(x)\right) = \mathbf{1}\left(\frac{B(x)}{R(x)} \geq \lambda\frac{\overline{H}}{\overline{R}}\right). \tag{10}$$

The A-optimal rule therefore orders patients by calibrated treatment benefit $B(x)$. The B-optimal rule orders patients by the ratio of calibrated benefit to untreated bleeding risk, $B(x)/R(x)$. Both rules can be written compactly as $D_\omega^*(x;\lambda) = \mathbf{1}(B(x) \geq \lambda H_\omega(x))$.

Figure 7 illustrates the geometry of this tension in benefit-risk space for $\lambda = 0.8$. The horizontal

line corresponds to the threshold implied by the A-optimal rule: patients are treated if their calibrated benefit exceeds a constant threshold. The upward-sloping line corresponds to the threshold implied by the B-optimal rule: patients with higher bleeding risk require proportionally higher benefit to justify treatment. In the normalization used in the figure, the slope of this threshold is λ. The deepest-shaded and white regions are areas of agreement, where both rules either treat or do not treat, respectively. The lighter-shaded regions are areas of disagreement, where patients would be treated under one optimal rule but not the other. Because benefits and risks are positively correlated in the VHA sample (Figure 4), more patients lie in the disagreement regions than would be the case if benefits and risks were negatively correlated.

Disagreement varies with the preference parameter λ, which encodes the utility cost of a bleed harm relative to the utility benefit of stroke prevention. When $\lambda = 0$, treatment has no harm cost, and both rules treat all patients. As λ becomes arbitrarily large, both rules treat no patients. For intermediate values, however, the rules can differ substantially: the share of VHA patients for whom the A-optimal and B-optimal recommendations disagree reaches 41% of the VHA population. Thus, even after calibrating evidence on treatment benefits and agreeing on the utility cost of a bleed relative to a stroke, different assumptions about how bleeding harms vary across patients can imply different recommendations for nearly half of the target population.

4.4 Robust Treatment Rules

We have shown that the A-optimal and B-optimal rules perform best under the harm assumptions for which they are designed, but can lose much of their advantage when evaluated under the other assumption. We now ask how to choose treatment rules when the structure of bleeding harms is itself uncertain. Specifically, we consider rules that are robust over the benchmark ambiguity set spanned by these two harm structures.

Let $\theta \in [0, 1]$ denote the weight placed on Assumption A, so that $\theta = 1$ corresponds to mean-independent harms and $\theta = 0$ corresponds to proportional harms. For any treatment rule D, define welfare under this ambiguity set as

$$W(D; \lambda, \theta) = \theta W_A(D; \lambda) + (1 - \theta) W_B(D; \lambda). \tag{11}$$

We consider two robust criteria. The maximin rule maximizes welfare under the worst-case harm assumption:

$$D^*_{\text{MM}}(\cdot; \lambda) = \arg\max_D \min_{\theta \in [0,1]} W(D; \lambda, \theta). \tag{12}$$

The minimax-regret rule minimizes the largest welfare loss relative to the rule that would have been chosen

if the harm assumption were known:

$$D^*_{\mathrm{MR}}(\cdot;\lambda) = \arg\min_{D} \max_{\theta\in[0,1]} \left\{ \max_{D'} W(D';\lambda,\theta) - W(D;\lambda,\theta) \right\}. \tag{13}$$

Thus, maximin evaluates a rule by its lowest welfare across the ambiguity set, while minimax regret evaluates a rule by its largest loss (or regret) relative to the assumption-specific optimum.

Appendix A.4 shows that the robust rules agree with the A-optimal and B-optimal rules wherever those rules agree: patients treated under both assumption-specific rules are treated by the robust rules, and patients treated under neither are not treated. Robustness matters only in the disagreement regions, where treatment raises expected net benefit under one harm assumption and lowers it under the other.

The same appendix shows that, in these disagreement regions, robust rules can be characterized by a cutoff on the ratio of the welfare gain under the favorable assumption to the welfare loss under the unfavorable assumption. To see why this implies a straight-line threshold, consider the region where the A-optimal rule treats but the B-optimal rule does not. In this region, $U_A(x;\lambda) > 0$ and $U_B(x;\lambda) < 0$, so a gain-loss cutoff takes the form

$$U_A(x;\lambda) \geq \chi_{\mathrm{crit}}(\lambda)\{-U_B(x;\lambda)\},$$

where crit $\in \{\mathrm{MM}, \mathrm{MR}\}$ indexes the robust criterion. Substituting Equation (7) and rearranging shows that treatment is optimal when stroke benefit exceeds a weighted average of the two harm thresholds. Appendix A.4 shows that the same parameter $\chi_{\mathrm{crit}}(\lambda)$ governs the opposite disagreement region as well, with the inequality reversed. The robust rule can therefore be written as

$$D^*_{\mathrm{crit}}(x;\lambda) = \mathbf{1}\left(B(x) \geq \lambda \left[\frac{1}{1+\chi_{\mathrm{crit}}(\lambda)} H_A(x) + \frac{\chi_{\mathrm{crit}}(\lambda)}{1+\chi_{\mathrm{crit}}(\lambda)} H_B(x) \right] \right). \tag{14}$$

Thus, the robust threshold is a weighted average of the harm thresholds for Assumptions A and B, with relative weight governed by $\chi_{\mathrm{crit}}(\lambda)$. Since $H_A(x) = \overline{H}$ and $H_B(x) = (\overline{H}/\overline{R})R(x)$, Equation (14) is a threshold for $B(x)$ that is linear in $R(x)$. When $\chi_{\mathrm{crit}}(\lambda) = 0$, the rule coincides with the A-optimal rule; as $\chi_{\mathrm{crit}}(\lambda)$ approaches infinity, it approaches the B-optimal rule. Finite positive values place partial weight on untreated bleeding risk. The value of $\chi_{\mathrm{crit}}(\lambda)$ is determined by the robust criteria stated in Equations (12) and (13).

Figure 7 illustrates this geometry. Robust thresholds lie between the horizontal A-optimal threshold and the upward-sloping B-optimal threshold. For an interior maximin solution, Appendix A.4 shows that the rule equalizes welfare under the two endpoint harm assumptions:

$$W_A(D^*_{\mathrm{MM}};\lambda) = W_B(D^*_{\mathrm{MM}};\lambda). \tag{15}$$

Because the benefit term is the same under both assumptions, this equality reduces to an equality of average harms among treated patients. Under Assumption A, average bleeding harm among treated patients is proportional to the population mean $\overline{R}$. Under Assumption B, average bleeding harm among treated patients is proportional to the average untreated bleeding risk among treated patients. Thus, an interior maximin rule satisfies

$$E[R(X) \mid D^*_{\text{MM}}(X;\lambda) = 1] = \overline{R}. \tag{16}$$

The maximin rule therefore chooses a treated set whose average untreated bleeding risk matches the population average, thereby equalizing welfare across the two harm assumptions.

The maximin rule need not always be interior. In particular, if the A-optimal rule has weakly lower welfare under Assumption A than under Assumption B, so that

$$W_A(D^*_A;\lambda) \leq W_B(D^*_A;\lambda),$$

then D^*_A is itself maximin. The reason is that D^*_A already maximizes the worst-case welfare, corresponding to Assumption A. This case is intuitive when benefits and bleeding risks are negatively correlated such that $E[R(X) \mid D^*_A(X;\lambda) = 1] \leq \overline{R}$: prioritizing patients with the largest absolute benefits would yield even larger welfare under Assumption B. The need to move away from the A-optimal rule in our setting therefore stems from the positive empirical correlation between stroke-reduction treatment benefits and untreated bleeding risks.[18]

The minimax-regret rule has a parallel characterization. Rather than equalizing welfare across harm assumptions, it equalizes regret:

$$W_A(D^*_A;\lambda) - W_A(D^*_{\text{MR}};\lambda) = W_B(D^*_B;\lambda) - W_B(D^*_{\text{MR}};\lambda). \tag{17}$$

Maximin therefore balances welfare levels across assumptions, while minimax regret balances losses relative to the assumption-specific optima.

A final theoretical issue is whether a stable patient ordering summarizes the family of robust rules as λ varies. Such an ordering exists only if the treated sets are nested: increasing the harm weight should move the treatment cutoff without reordering patients. Appendix A.4 gives a sufficient condition for nestedness of the interior maximin rules based on the conditional quantile functions of benefits given bleeding risk. For minimax regret, analogous global nestedness requires additional restrictions. In the empirical implementation, both the maximin and minimax-regret treatment sets are nested over the relevant

18. Appendix A.4 also gives the symmetric corner condition for the B-optimal rule: if $W_B(D^*_B;\lambda) \leq W_A(D^*_B;\lambda)$, then $D^*_{\text{MM}} = D^*_B$. This condition is equivalent to the B-optimal treated set having an average untreated bleeding risk that is weakly above the population mean, or $E[R(X) \mid D^*_B(X;\lambda) = 1] \geq \overline{R}$. In that case, Assumption B is the worst-case endpoint for D^*_B; because D^*_B already maximizes welfare under Assumption B, no alternative rule can achieve a higher endpoint worst-case welfare.

range of λ.

Figure 8 plots the benefit-harm frontiers generated by the assumption-specific and robust treatment rules. At the same 50%-stroke-prevention benchmark used above, the maximin rule induces 36.5% fewer bleeds than random under both Assumption *A* and Assumption *B*. Thus, relative to the A-optimal rule, maximin sacrifices little performance under Assumption *A* while substantially improving performance under Assumption *B*. The minimax-regret rule induces 33.4% fewer bleeds than random under Assumption *A* and 40.5% fewer bleeds than random under Assumption *B*. Relative to maximin, minimax regret accepts a larger sacrifice under mean-independent harms in exchange for stronger protection under proportional harms.

4.5 Statistical Uncertainty and Welfare

We next incorporate finite-sample uncertainty in estimated stroke benefits. Holding the VHA target population and bleeding-risk predictions fixed, Appendix A.5 constructs 100 joint realizations of patient-level benefit estimates from the randomized evidence. We use the empirical distribution of these realizations to represent sampling uncertainty in benefits, while still treating Assumptions *A* and *B* as alternative harm structures under ambiguity.

Under the additive welfare function in Equation (8), expected welfare from any fixed treatment rule is linear in patient-level benefits. Averaging welfare across the benefit realizations therefore yields the same welfare as evaluating the rule at their mean. As a result, incorporating this sampling uncertainty leaves the A-optimal, B-optimal, and maximin rules unchanged. Minimax regret can change, however, because its benchmark is the optimal rule under each realized set of benefit estimates. Appendix A.5 derives this sampling-aware minimax expected-regret criterion and reports the corresponding welfare decomposition.

Sampling uncertainty nevertheless produces meaningful instability in fine patient rankings. Under Assumption *A*, 14.2% of patient pairs have unresolved treatment orderings; under Assumption *B*, the corresponding share is 20.2%. Yet the welfare consequences are small. At the 50%-stroke-prevention benchmark, the expected welfare loss from sampling uncertainty in benefit is 1.56 per 10,000 patients under Assumption *A* and 2.06 under Assumption *B*, compared with 17.58 and 17.09, respectively, from using the sampling-aware minimax-regret rule rather than the corresponding assumption-specific optimum. The sampling-aware minimax-regret rule differs from the rule based on the mean estimated benefits for only 0.12% of patients. Pairwise ranking reversals need not change treatment assignments—for example, when both patients remain inframarginal—and even assignment changes may have small welfare consequences when the welfare difference between treating and not treating the affected patients is small. Thus, relatively sparse experimental evidence can generate noticeable uncertainty about fine patient rankings while still supporting stable treatment rules and welfare outcomes.

5 Physician Decisions

In this section, we turn to physician decisions. Although observed practice does not close the evidence gap on heterogeneous bleeding harms, it provides an empirical benchmark for understanding how treatment is allocated, how clinicians use information beyond existing guideline scores, and how to interpret guideline adherence amid scientific uncertainty.

We use physician decisions in two ways. First, we construct a physician-consensus rule, motivated by the "wisdom of the crowd" logic that aggregating expert decisions may reveal useful information even when no individual decision-maker is fully optimal (Gillen, Plott, and Shum 2017). Second, we estimate how physicians vary in treatment propensity and how closely their decisions align with alternative patient orderings. We then evaluate the treatment allocations implied by this estimated physician distribution using the benefit evidence from Section 3 and the harm assumptions from Section 4. This lets us compare the consensus rule with formal treatment rules and with the distribution of physician treatment rules under Assumptions *A* and *B*. These comparisons are benchmark-relative: observed physician choices do not identify the correct harm structure.

5.1 Observed Physician Behavior and the Consensus Rule

Section 2 showed that $CHADS_2$ became widely known in practice, but that changes in treatment after guideline awareness were modest. Existing guidelines therefore inform treatment behavior, but they do not fully determine anticoagulation decisions.

We summarize the common observable component of physician decision-making with the consensus rule. We estimate a treatment-prediction model using VHA patient characteristics and physician fixed effects, so we identify the patient-characteristic component from within-physician variation. The fitted patient component defines a predicted treatment propensity for each patient, and the consensus rule ranks patients by this propensity. This rule should be distinguished from realized physician performance: it evaluates an automated rule learned from the common observable component of physician behavior and applied mechanically to a common patient population, without physician discretion or unobserved bedside information.

5.2 A Skill-Threshold Model of Physician Treatment Decisions

To study physician heterogeneity, we use the skill-threshold framework of Chan, Gentzkow, and Yu (2022). This framework aligns with the treatment-rule representation in Section 4: a rule consists of an ordering of patients and a threshold along that ordering. Physicians may differ in thresholds, in how strongly their decisions follow a candidate ordering, or both.

For a harm assumption $\omega \in \{A, B\}$, let r_i^ω denote a normalized rank score for patient i under the ω-optimal ordering, oriented so that higher-priority patients receive larger r_i^ω. Equivalently, r_i^ω is a decreasing transformation of the patient's position in the ordering, where a lower position denotes higher treatment priority. Physician j treats patient i with probability

$$\Pr(D_{ij} = 1 \mid r_i^\omega, k_j^\omega, p_j^\omega) = \Phi\Big(k_j^\omega r_i^\omega - p_j^\omega\Big), \tag{18}$$

where $\Phi(\cdot)$ is the standard normal distribution function. The threshold parameter p_j^ω captures physician j's treatment aggressiveness: holding the ordering fixed, physicians with higher thresholds treat fewer patients. The skill parameter k_j^ω captures alignment with the ω-optimal ordering. Larger positive values imply stronger prioritization of high-priority patients; $k_j^\omega = 0$ implies no relationship with the ordering; and $k_j^\omega < 0$ implies treatment against the ordering.

Skill is therefore conditional on the ordering used to evaluate decisions, not an absolute measure of clinical quality. We estimate the model separately relative to the A- and B-optimal orderings, so skill is not an assumption-independent physician attribute. Under Assumption A, skill measures alignment with stroke-benefit prioritization. Under Assumption B, skill measures alignment with an ordering that discounts patients with higher untreated bleeding risk.

Because individual physicians treat too few patients to reliably estimate physician-specific skill and threshold parameters, we estimate a random-coefficients model for the physician population rather than estimating (k_j^ω, p_j^ω) separately for each physician. This population model makes it natural to evaluate simulated physicians drawn from the estimated distribution rather than analyzing actual physicians. We then evaluate each simulated physician on the same overall VHA patient population. Differences in simulated outcomes across physicians thus reflect only differences in their underlying treatment rules. Appendix A.6 provides the full model and estimation details.

5.3 Physician Performance and Guideline Adherence

We now evaluate the consensus rule and simulated physician decision rules under the same two harm assumptions used above. As in Section 4, performance is measured by bleeds induced when a rule prevents 50% of preventable strokes; lower values indicate better performance.

Figure 9 compares the consensus rule with formal treatment rules and with a summary of simulated physician performance.[19] To keep the physician benchmark comparable to the treatment-rule frontiers, we report the *median-skill physician*: the benefit-harm curve implied by the median of the marginal skill distribution, with the treatment threshold varied along that curve. Under Assumption A, the consensus rule

19. Appendix Figure A.1 reproduces Figure 9 using the BLP-projected stroke-benefit measure, $B_{\text{BLP}}(x)$, instead of the EWM-calibrated measure, $B_{\text{EWM}}(x)$; the main comparisons are unchanged.

performs worse than the median-skill physician, and both perform worse than random treatment. Under Assumption *B*, the consensus rule performs better than the median-skill physician, and both perform better than random treatment.

Appendix Figure A.2 examines the full distribution of simulated physician performance. Under Assumption *A*, the $CHADS_2$ guideline substantially outperforms the simulated physician distribution: no simulated physician outperforms $CHADS_2$. Under Assumption *B*, by contrast, 44.5% of simulated physicians outperform $CHADS_2$. Comparing the consensus rule with the physician distribution helps distinguish systematic from idiosyncratic components of treatment behavior. Aggregation removes physician-specific variation while preserving the component of treatment allocation that physicians share. Whether this improves welfare therefore depends on whether that common component is aligned with the relevant welfare benchmark. The common component performs poorly under Assumption *A* but is valuable under Assumption *B*.

Finally, we examine how guideline adherence relates to physician skill. Because most VHA patients have $CHADS_2$ scores warranting treatment, raw guideline adherence is closely related to treatment propensity. We therefore distinguish overall treatment aggressiveness from patient prioritization by examining the relationship between adherence and skill conditional on treatment propensity. The relationship reverses across harm assumptions: greater adherence is associated with higher skill under Assumption *A* but lower skill under Assumption *B* (Figure 10). These descriptive results reinforce that guideline adherence is not an assumption-independent measure of physician quality: its relationship with decision quality depends on the assumed structure of bleeding harms.

6 Conclusion

Clinical guidelines translate evidence into treatment decisions. Their value depends not only on the strength of the evidence but also on whether it identifies the welfare-relevant trade-offs guidelines implicitly make. In atrial fibrillation, anticoagulation prevents strokes but increases bleeding. Existing guidelines, especially $CHADS_2$ and CHA_2DS_2-VASc, summarize stroke risk and became widely known in practice. Yet the randomized evidence base identifies heterogeneous stroke-prevention benefits much more clearly than heterogeneous treatment-induced bleeding harms. This asymmetry creates scientific uncertainty: if treatment harms are mean-independent of stroke benefits, prioritizing high-benefit patients solves the relevant welfare problem; if harms are proportional to untreated bleeding risk, the same prioritization can expose high-risk patients to greater harms. Because stroke benefits and bleeding risks are positively related in the VHA population, these harm structures imply different patient rankings and different assessments of existing guidelines.

Our framework explicitly models uncertainty about the relationship between treatment benefits and harms. Existing guidelines perform well when high-stroke-risk patients are also the patients with the largest net benefits from treatment, but their performance can deteriorate when the patients prioritized for stroke prevention also face greater treatment-induced bleeding harms. Robust rules such as maximin and minimax regret preserve performance across plausible assumptions by accounting for what the evidence does not identify. These rules can also be made clinically legible. In this setting, they intuitively amount to prioritizing patients with large expected stroke-prevention benefits while avoiding recommendations that depend too heavily on unresolved assumptions about bleeding harms. Patients favored under both harm assumptions are natural candidates for treatment; patients disfavored under both are natural candidates for non-treatment; and patients whose recommendation changes across assumptions are cases where the evidence is least decisive. Common phenotypes in this disagreement set are natural targets for future experimentation.

We also use our framework to evaluate physician behavior under different assumptions about the structure of bleeding harms. The consensus rule and the skill-threshold model are consistent with physicians incorporating bleeding-risk information, which can be valuable when bleeding harms rise with untreated bleeding risk. But observed practice, or a machine-learning rule distilled from it, improves welfare only if the risk information physicians use corresponds to true treatment-induced harms. Existing guidelines raise the opposite concern: they prioritize patients using stroke-risk factors that are positively related to stroke benefits, so the welfare value of adherence to them depends on whether bleeding harms are sufficiently unrelated to the risks the guideline largely omits. Thus, practice-based decision rules and existing guidelines improve decisions under different assumptions about bleeding harms; neither provides a uniformly welfare-improving benchmark across plausible harm structures.

The broader lesson is that evidence-based guidelines are only as good as the benefit–harm tradeoffs they encode. The CHADS_2 family of scores is distinctive in its clinical importance, but the evidentiary problem it illustrates is more general. Guidelines often focus on disease-specific benefits that are relatively tractable to measure, while heterogeneous harms may depend on comorbidities, frailty, polypharmacy, or other patient complexity that randomized trials represent less well. When evidence identifies some welfare-relevant objects better than others, increasingly detailed decision rules may give a false sense of precision. A better response is to distinguish what the evidence supports from what it leaves ambiguous, evaluate policies under competing scientific assumptions, and design treatment rules that are robust to the remaining uncertainty.

References

Abaluck, Jason, Leila Agha, David C. Chan Jr, Daniel Singer, and Diana Zhu. 2020. "Fixing Misallocation with Guidelines: Awareness vs. Adherence." NBER Working Paper 27467.

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.

Abaluck, Jason, Leila Agha, and Sachin Shah. 2025. "Trials Avoid High Risk Patients and Underestimate Drug Harms." NBER Working Paper 34534.

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

Angelova, Victoria, Will Dobbie, and Crystal S Yang. 2026. "Algorithmic Recommendations and Human Discretion." *The Review of Economic Studies* 93 (4).

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

Athey, Susan, and Stefan Wager. 2021. "Policy Learning With Observational Data." *Econometrica* 89 (1): 133–161.

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.

Boone, Claire. 2025. "Discretion in Clinical Decision Making: Evidence from Bunching." *Review of Economics and Statistics,* 1–47.

Boyd, Cynthia M., Jonathan Darer, Chad Boult, Linda P. Fried, Lisa Boult, and Albert W. Wu. 2005. "Clinical Practice Guidelines and Quality of Care for Older Patients with Multiple Comorbid Diseases: Implications for Pay for Performance." *JAMA* 294 (6): 716–724.

Breiman, Leo. 1996. "Bagging Predictors." *Machine Learning* 24 (2): 123–140.

Centers for Medicare & Medicaid Services. 2017. *Measure #326 (NQF 1525): Atrial Fibrillation and Atrial Flutter: Chronic Anticoagulation Therapy.* Quality Payment Program quality measure specifications, Claims only.

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.

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.

Chernozhukov, Victor, Mert Demirer, Esther Duflo, and Iván Fernández-Val. 2025. "Fisher-Schultz Lecture: Generic Machine Learning Inference on Heterogeneous Treatment Effects in Randomized Experiments, With an Application to Immunization in India." *Econometrica* 93 (4): 1121–1164.

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." *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. 2026. "Rules versus Discretion: Treatment of Mental Illness in US Adolescents." *Journal of Political Economy* 134 (1): 478–522.

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

Deaton, Angus. 2010. "Instruments, Randomization, and Learning about Development." *Journal of Economic Literature* 48 (2): 424–455.

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.

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. "Ambiguity 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.

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. 2015. *Discretion in Hiring*. Working Paper 21709. National Bureau of Economic Research.

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.

Joglar, José A., Mina K. Chung, Anastasia L. Armbruster, Emelia J. Benjamin, Janice Y. Chyou, Edmond M. Cronin, Anita Deswal, et al. 2024. "2023 ACC/AHA/ACCP/HRS Guideline for the Diagnosis and Management of Atrial Fibrillation: A Report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines." *Circulation* 149 (1): e1–e156.

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.

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.

Kitagawa, Toru, and Aleksey Tetenov. 2018. "Who Should Be Treated? Empirical Welfare Maximization Methods for Treatment Choice." *Econometrica* 86 (2): 591–616.

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.

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.

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. 2004. "Statistical Treatment Rules for Heterogeneous Populations." *Econometrica* 72 (4): 1221–1246.

———. 2011. "Choosing Treatment Policies Under Ambiguity." *Annual Review of Economics* 3:25–49.

———. 2013. "Diagnostic Testing and Treatment Under Ambiguity: Using Decision Analysis to Inform Clinical Practice." *Proceedings of the National Academy of Sciences of the United States of America* 110 (6): 2064–2069.

Mullahy, John, Atheendar S. Venkataramani, Daniel L. Millimet, and Charles F. Manski. 2021. "Embracing Uncertainty: The Value of Partial Identification in Public Health and Clinical Research." *American Journal of Preventive Medicine* 61 (2): e103–e108.

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.

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.

Rothwell, Peter M. 2005. "External validity of randomised controlled trials: 'To whom do the results of this trial apply?" *The Lancet* 365 (9453): 82–93.

Savage, L. J. 1951. "The Theory of Statistical Decision." *Journal of the American Statistical Association* 46 (253): 55–67.

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.

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.

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.

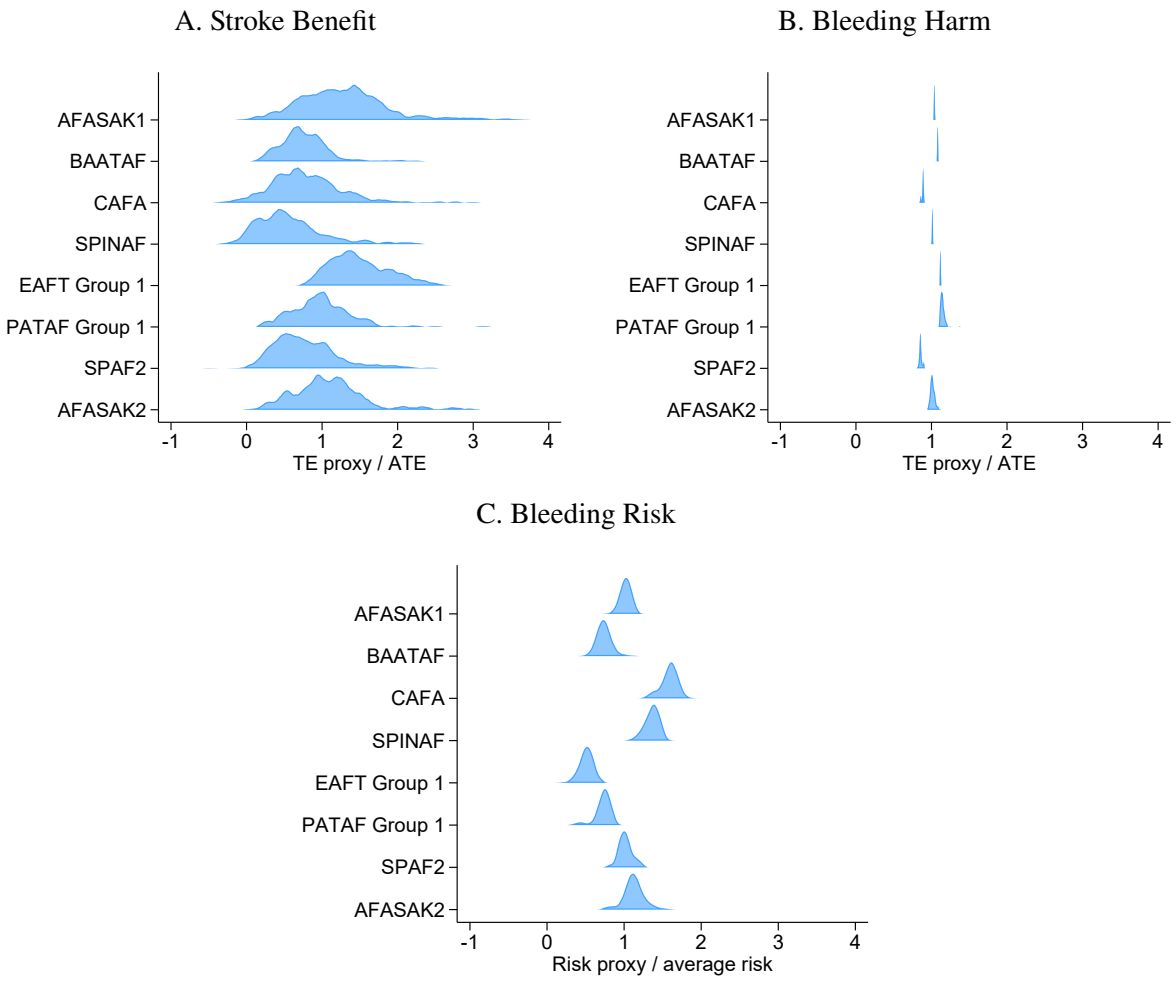
Turakhia, Mintu P., Donald D. Hoang, Xiangyan Xu, Susan Frayne, Susan Schmitt, Felix Yang, Ciaran S. Phibbs, Claire T. Than, Paul J. Wang, and Paul A. Heidenreich. 2013. "Differences and Trends in Stroke Prevention Anticoagulation in Primary Care vs Cardiology Specialty Management of New Atrial Fibrillation: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

van Walraven, Carl, 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.

Wager, Stefan, and Susan Athey. 2018. "Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests." *Journal of the American Statistical Association* 113 (523): 1228–1242.

Wald, Abraham. 1950. *Statistical Decision Functions.* New York: John Wiley & Sons.

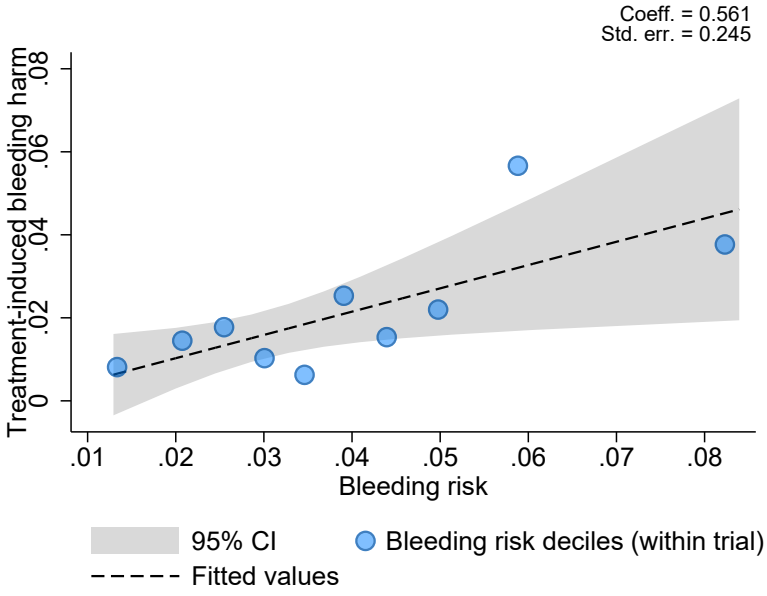
Figure 1: Distribution of Proxy Predictions for Validation



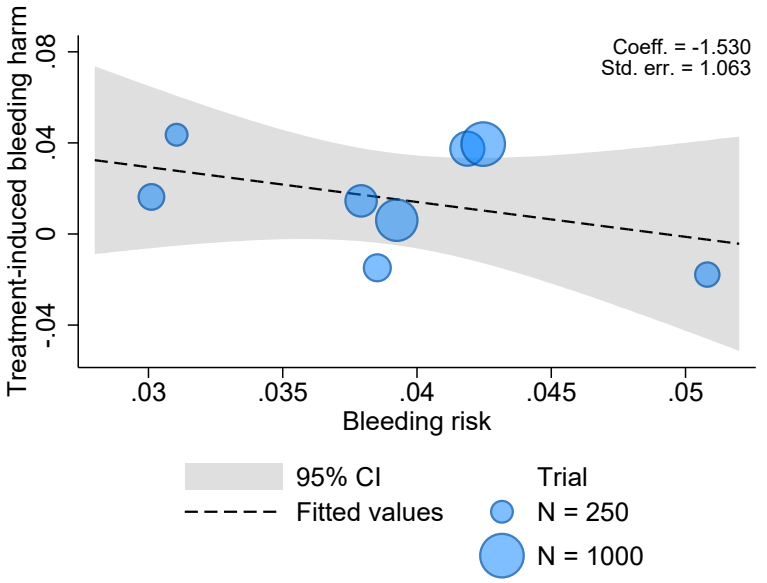
Notes: This figure shows the distributions of the prediction objects used in the cross-trial validation exercise across hold-out trials in Table 1. For patients in hold-out trial j, Panels A–C show the stroke-benefit proxy, $Z_{B,-j}(x)$; the bleeding-harm proxy, $Z_{H,-j}(x)$; and the RCT-estimated bleeding-risk proxy, $Z_{R,-j}(x)$. We estimate each proxy using data from trials other than trial j and then apply it to patients in the hold-out trial. The stacked kernel density plots show the resulting distribution in each hold-out trial. For comparability across outcomes, we normalize each proxy by subtracting and dividing by its corresponding population mean: the average treatment effect for each treatment-effect proxy and the average bleeding risk for the bleeding-risk proxy.

Figure 2: Trial Bleeding Risk and Harms

A. Within-Trial Design



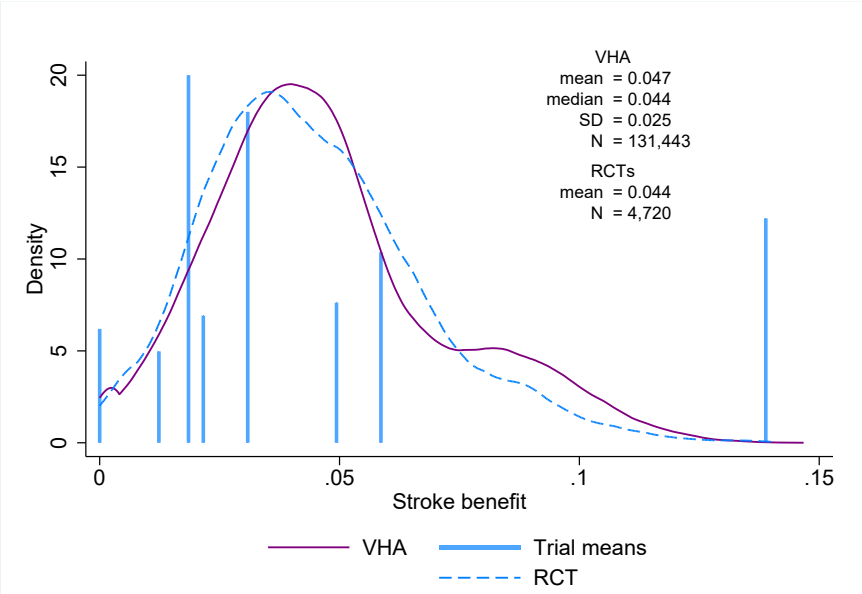
B. Across-Trial Design



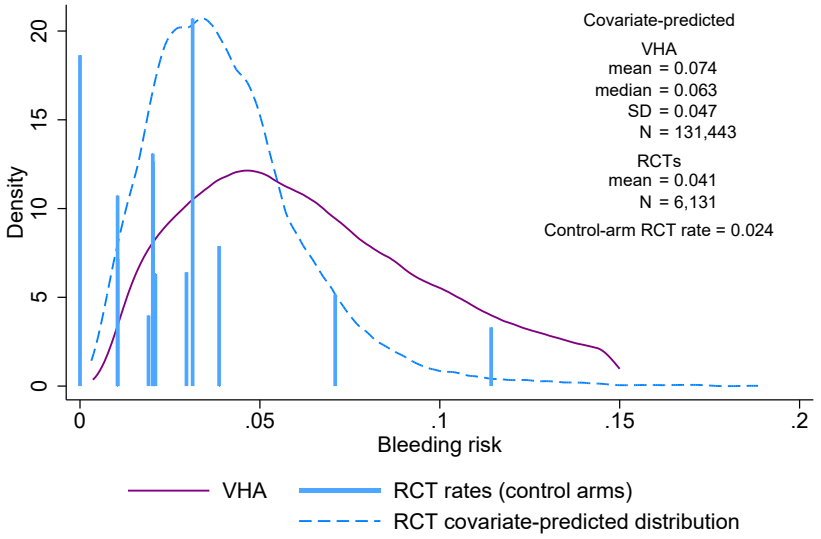
Notes: This figure relates untreated bleeding risk to treatment-induced bleeding harm in the RCT data. Bleeding risk is the VHA-estimated common-covariate prediction, $R_C(x)$, applied to RCT patients; bleeding harm is the treatment-control difference in bleeding-event probabilities. Panel A groups patients into risk deciles within each trial and plots average risk against the treatment-control difference within each trial–decile cell. Panel B plots each trial's average risk against its overall treatment-control difference. Dashed lines show linear fits, with 95% confidence bands. Confidence bands in Panel A are computed by bootstrap.

Figure 3: Distributions of Stroke Benefits and Bleeding Risk

A. Stroke Benefits

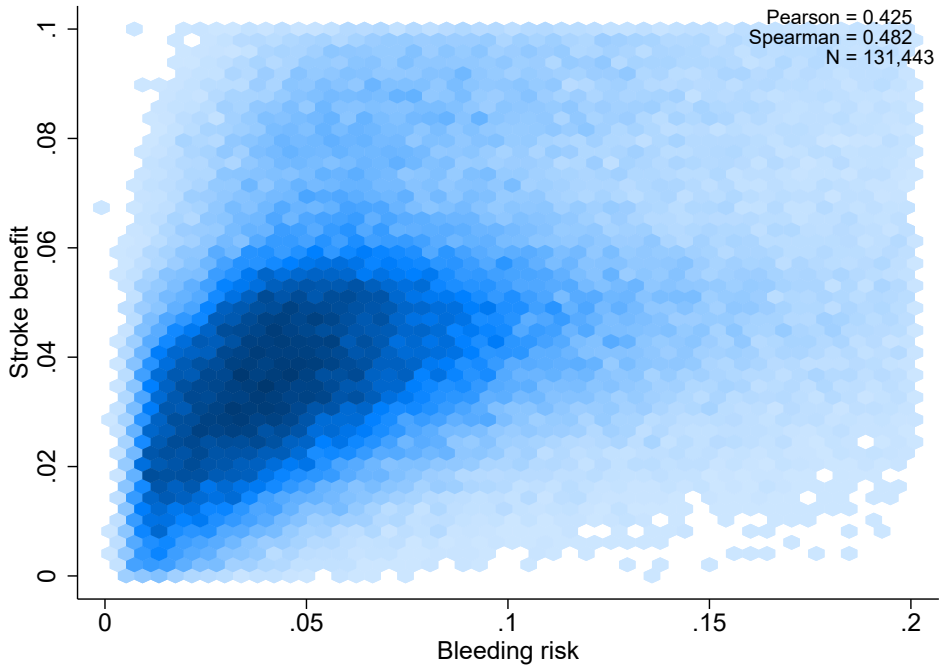


B. Bleeding Risk



Notes: This figure compares the distributions of predicted stroke benefits and untreated bleeding risk in the VHA and RCT samples. Panel A shows BLP-projected stroke benefits, $B_{\text{BLP}}(x)$. Panel B shows the VHA-estimated common-covariate prediction, $R_C(x)$, for VHA and RCT patients. Solid and dashed curves show the VHA and RCT distributions, respectively. Vertical lines show trial-specific estimates, with heights proportional to trial size: mean stroke treatment effects for the eight trials with treatment and control arms in Panel A, and observed control-arm bleeding rates for the twelve trials with control arms in Panel B. The RCT sample in Panel B is the 6,131 patients in these twelve trials: the Step 5 sample in Panel B of Appendix Table A.1, less the single trial group without a control arm.

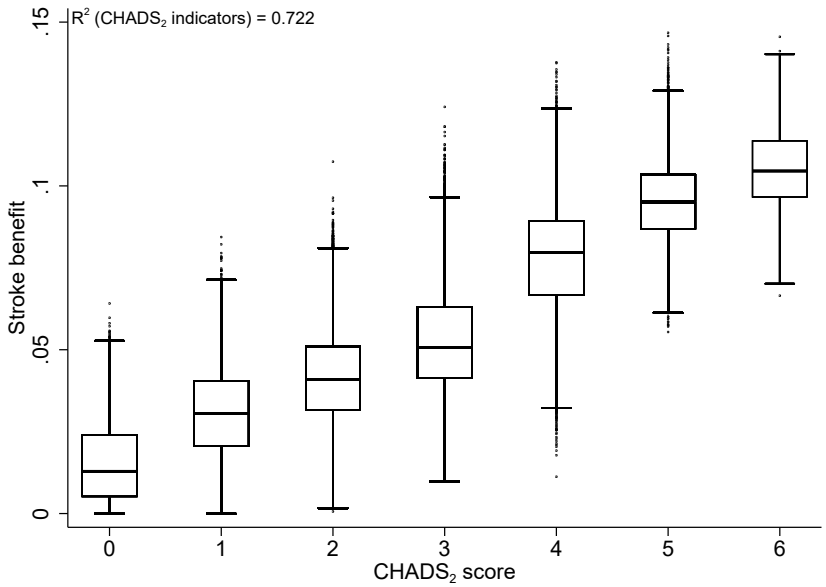
Figure 4: Joint Distribution of Stroke Benefits and Bleeding Risk



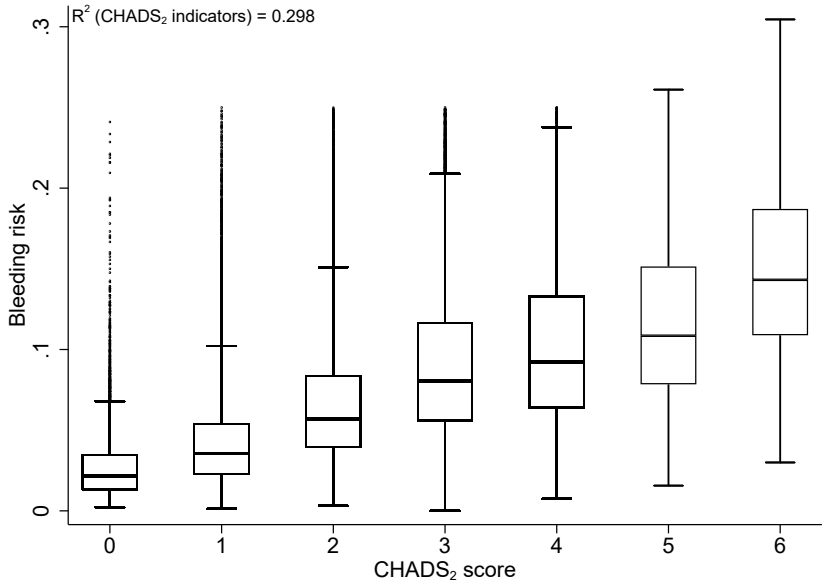
Notes: This figure displays the joint distribution of BLP-projected stroke benefits, $B_{\text{BLP}}(x)$, and predicted untreated bleeding risk, $R(x)$, among patients in the VHA sample. Stroke benefits are estimated from the AFI randomized trials using the ML proxy validation procedure described in Section 3. Untreated bleeding risk is predicted in the VHA data under no anticoagulation and is described further in Appendix A.2. The heat map shows the density of VHA patients in benefit–risk space.

Figure 5: Stroke Benefits and Bleeding Risk by $CHADS_2$ Score

A. Stroke Benefits



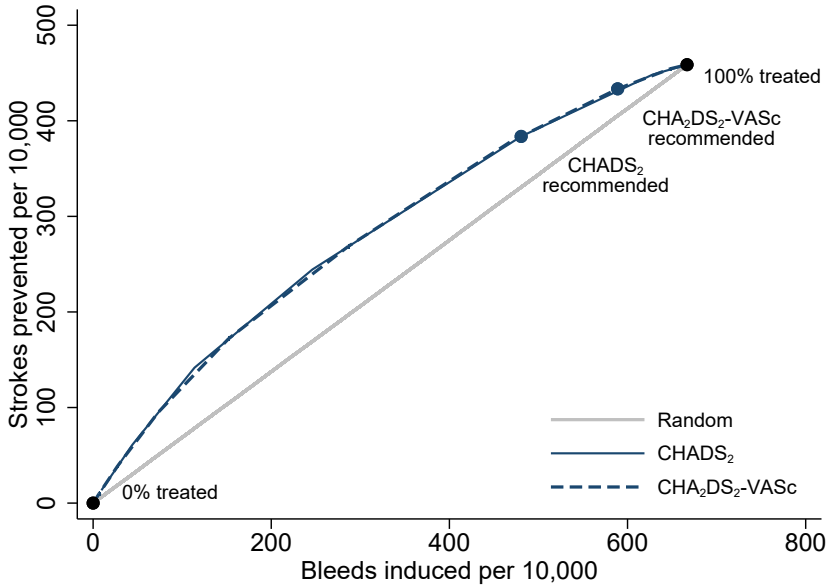
B. Bleeding Risk



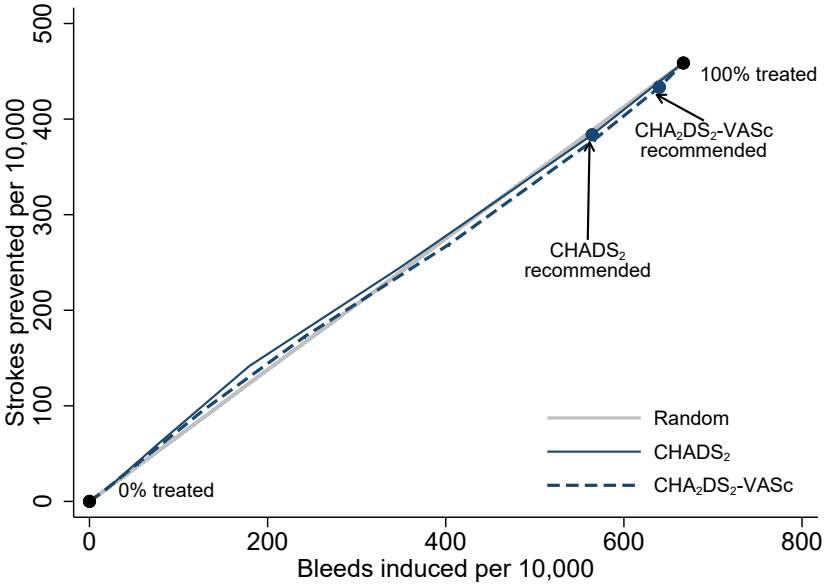
Notes: This figure shows box plots for the distribution of BLP-projected stroke benefits, $B_{\text{BLP}}(x)$, and predicted untreated bleeding risk, $R(x)$, by $CHADS_2$ score among patients in the VHA sample. Panel A shows stroke benefits, and Panel B shows untreated bleeding risk. Box bounds are the 25th and 75th percentiles, the horizontal line marks the median, and whiskers extend to the 5th and 95th percentiles.

Figure 6: Benefit-Harm Frontiers for Existing Guidelines

A. Mean-Independent Harms

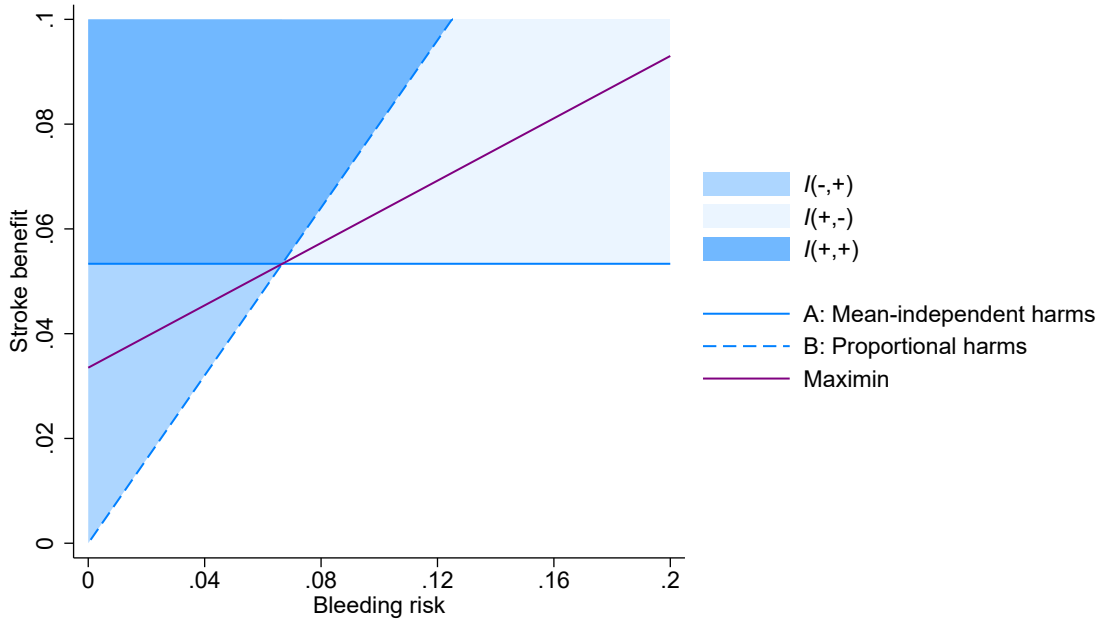


B. Proportional Harms



Notes: This figure plots benefit–harm frontiers for the $CHADS_2$ and CHA_2DS_2-VASc guideline orderings. Each ordering prioritizes patients for anticoagulation, and moving along a frontier varies the treatment threshold from treating 0% to 100% of patients. The vertical axis reports strokes prevented per 10,000 patients, and the horizontal axis reports bleeds induced per 10,000 patients. Panel A evaluates harms under Assumption *A* (mean-independent harms). Panel B evaluates harms under Assumption *B* (proportional harms). The gray line shows random treatment.

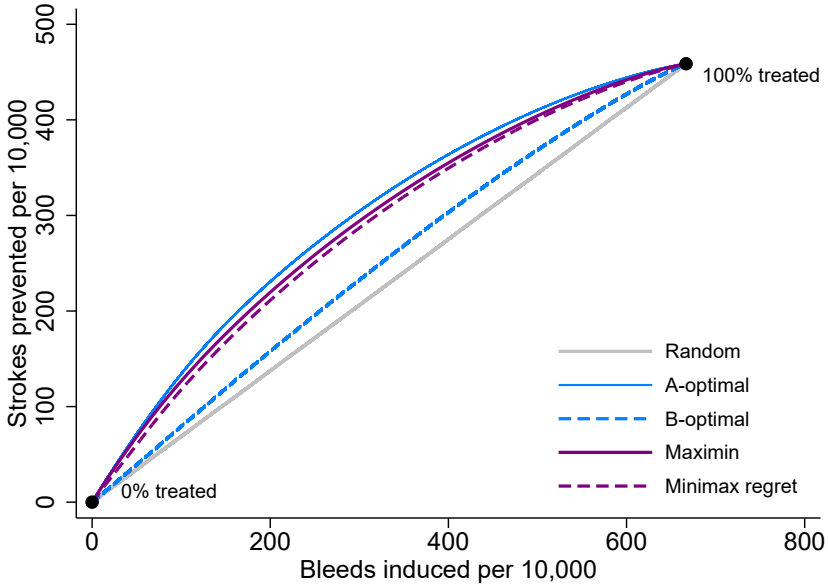
Figure 7: Optimal Treatment Rules in Benefit-Risk Space



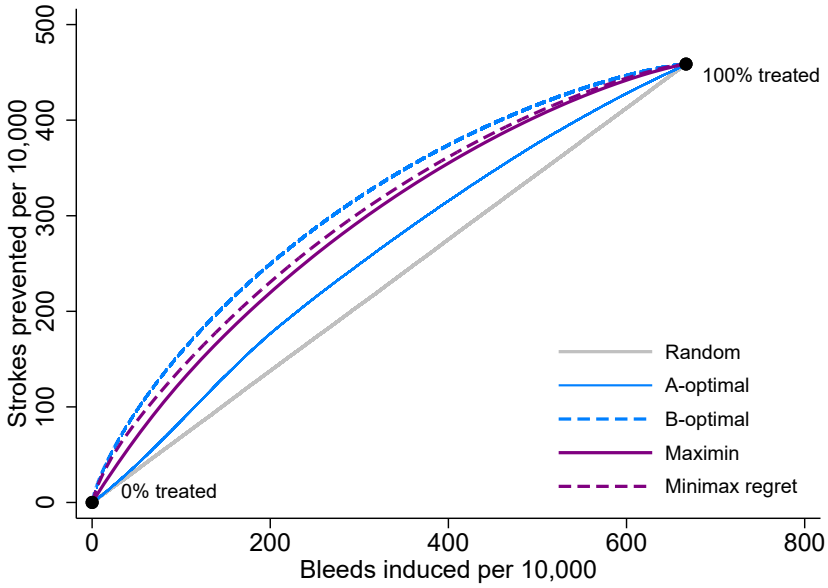
Notes: This figure illustrates treatment thresholds in benefit–risk space for $\lambda = 0.8$. The vertical axis is the calibrated stroke benefit, $B(x)$, and the horizontal axis is predicted untreated bleeding risk, $R(x)$. The horizontal line is the A-optimal threshold under Assumption A (mean-independent harms), and the upward-sloping line is the B-optimal threshold under Assumption B (proportional harms), where $H_B(x) = (\overline{H}/\overline{R})R(x)$. Region $\mathcal{I}(+,+)$ contains patients treated under both rules; $\mathcal{I}(+,-)$ contains patients treated only under Assumption A; and $\mathcal{I}(-,+)$ contains patients treated only under Assumption B. The purple line shows the maximin threshold, which lies between the two assumption-specific thresholds. For an interior solution, the maximin rule selects a treated set whose average untreated bleeding risk equals the population mean, as characterized in Equation (16). Figure 4 shows the empirical VHA distribution of stroke benefits and untreated bleeding risks underlying these comparisons.

Figure 8: Benefit-Harm Frontiers for Optimal and Robust Rules

A. Mean-Independent Harms

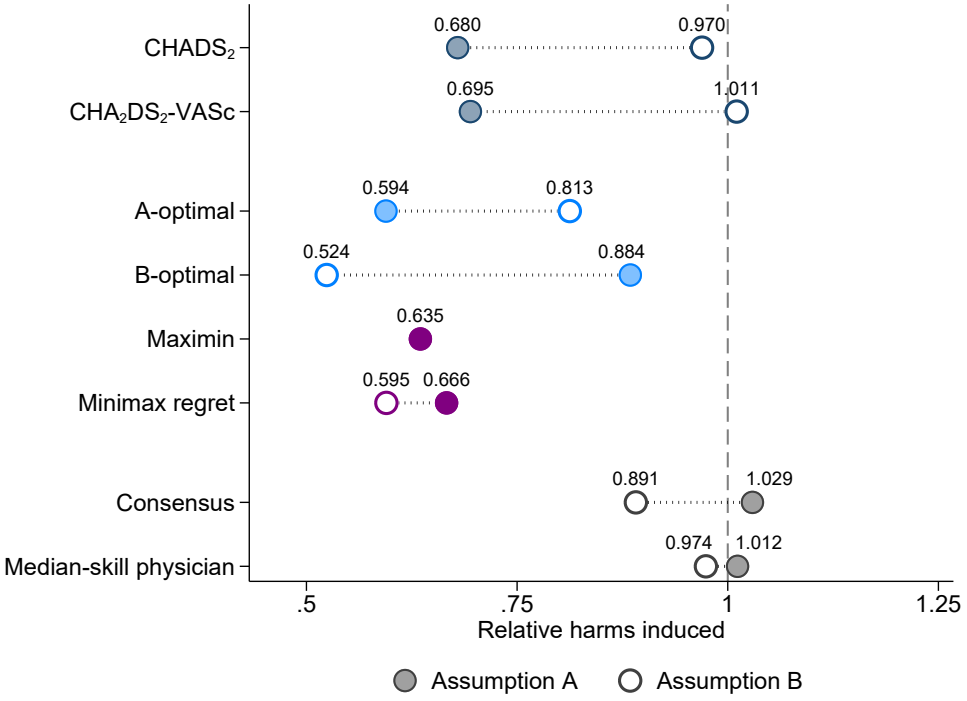


B. Proportional Harms



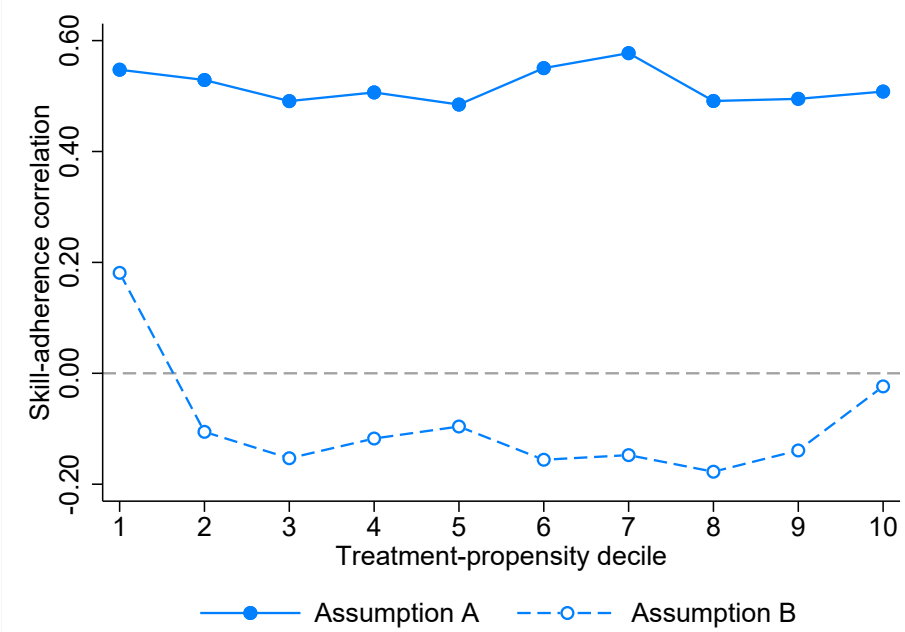
Notes: This figure plots benefit–harm frontiers for the A-optimal, B-optimal, maximin, and minimax-regret treatment orderings. Each ordering prioritizes patients for anticoagulation, and moving along a frontier varies the treatment threshold from treating 0% to 100% of patients. The vertical axis reports strokes prevented per 10,000 patients, and the horizontal axis reports bleeds induced per 10,000 patients. Panel A evaluates harms under Assumption *A* (mean-independent harms). Panel B evaluates harms under Assumption *B* (proportional harms). The gray line shows random treatment.

Figure 9: Performance Relative to Random Treatment



Notes: This figure compares random treatment, existing guideline orderings, optimal and robust treatment rules, the consensus rule, and the physician benchmark at a fixed stroke-prevention benchmark. All comparisons hold strokes prevented fixed at the level obtained by treating 50% of patients randomly. The horizontal axis reports bleeds induced relative to bleeds induced under 50% random treatment, so values below one indicate fewer bleeds than random treatment. Shaded circles evaluate harms under Assumption *A* (mean-independent harms), and hollow circles evaluate harms under Assumption *B* (proportional harms).

Figure 10: Adherence-Skill Correlation by Treatment Propensity



Notes: This figure shows the correlation between physician $CHADS_2$ adherence and empirical-Bayes posterior skill estimates within deciles of physician treatment propensity. Skill is defined relative to the A-optimal ordering under Assumption *A* (mean-independent harms) and the B-optimal ordering under Assumption *B* (proportional harms). The dashed line marks zero correlation.

Table 1: Cross-Trial Validation of ML Proxies

	Stroke	Bleeding	
	(1)	(2)	(3)
Treatment	-0.0441***	0.0186***	
	(0.00680)	(0.00518)	
Treatment × demeaned proxy	0.776**	-3.562	
	(0.315)	(3.570)	
Bleeding risk			-0.377
			(0.529)
Predicted outcome controls	Yes	Yes	
Trial-probability interaction	Yes	Yes	
Trial fixed effects	Yes	Yes	Yes
Outcome mean	0.065	0.030	0.030
Observations	4,720	4,720	4,720

Notes: This table reports hold-out-trial best linear predictor (BLP) validation regressions for the ML proxies. Column 1 validates the stroke-benefit proxy $Z_{B,-j}(x)$ by interacting treatment with the demeaned full stroke-benefit proxy. Column 2 validates the bleeding-harm proxy $Z_{H,-j}(x)$ using the same specification applied to the bleeding outcome. Column 3 reports a separate validation of the bleeding-risk proxy $Z_{R,-j}(x)$, regressing realized bleed outcomes on $Z_{R,-j}(x)$ with trial-by-arm fixed effects. "Predicted outcome controls" denotes LASSO and regression-forest predictions of the outcome estimated in the control groups of hold-out trials; "Trial-probability interaction" denotes the treatment probability in each trial interacted with the corresponding demeaned proxy. The validation sample corresponds to Appendix Table A.1, Panel B, Step 6. Standard errors are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

A.1 Trial-Based ML Proxies and Cross-Trial Validation

We use the AFI trial data to construct machine-learning (ML) proxies for heterogeneity in the stroke-prevention benefits and treatment-induced bleeding harms of anticoagulation. The two proxies are constructed using the same ensemble procedure and are treated *ex ante* symmetrically. We then use the multi-trial structure of the AFI database to test whether each proxy predicts treatment-effect heterogeneity in hold-out trials. Only the stroke-benefit proxy validates across trials, so the downstream policy analysis retains that proxy and does not use the bleeding-harm proxy as an empirical measure of heterogeneous harms. Appendix A.2 separately describes the VHA-estimated predictions of untreated bleeding risk used in the policy analysis.

A.1.1 Sample and Estimation Design

For cross-trial validation, we construct leave-one-trial-out predictions for each of the eight trials containing both treatment and control arms. For hold-out trial j, we fit the prediction procedure using the other trials and apply the fitted models to patients in trial j, producing $Z_{B,-j}(x)$ and $Z_{H,-j}(x)$. Outcome-risk components use control-group observations from all available training trials (Appendix Table A.1, Panel B, Step 5); treatment-effect components use trials containing both treatment and control arms (Panel B, Step 6). To obtain predictions for the VHA population, we also fit all-trials versions, denoted $Z_B(x)$ and $Z_H(x)$. To reduce sensitivity to sampling variation, we bag each leave-one-trial-out or all-trials model across ten estimates, each fit on a 95% subsample stratified by trial and treatment status. We average the ten predictions (Breiman 1996).

A.1.2 Construction of the ML Proxies

We apply the same prediction procedure to stroke and bleeding outcomes. Let Y_i^o denote outcome $o \in \{s, b\}$, where s denotes stroke and b denotes bleeding. The ensemble combines a post-LASSO component, used primarily for continuous predictors, with a causal-forest component that captures remaining heterogeneity using discrete predictors.

Post-LASSO component. Before estimation, we subtract trial-specific means from the outcome and predictors, which is equivalent to including unpenalized trial indicators. Using control-group observations, we first fit a LASSO outcome-risk model on patient characteristics observed in both the AFI and VHA data. We then estimate treatment-effect heterogeneity using both treatment arms by including the patient characteristics, the predicted outcome risk, the VHA-estimated common-covariate bleeding-risk prediction $R_C(X_i)$ described in Appendix A.2, and their interactions with treatment. We scale penalties by predictor standard deviations and select them via ten-fold cross-validation. We refit the selected continuous patient characteristics and risk predictions by OLS, following the post-LASSO approach of Belloni, Chernozhukov, and Hansen (2014).

Causal-forest component. We next subtract the fitted post-LASSO contribution and use a causal forest to estimate residual treatment-effect heterogeneity. Following Athey, Tibshirani, and Wager (2019), we locally center the residual outcome and treatment using out-of-bag nuisance estimates. The treatment model contains trial indicators, reflecting randomization within trials and differing treatment probabilities across trials. The forest uses eleven discrete characteristics—sex, white race, smoking status, and histories of angina, congestive heart failure, diabetes, hypertension, myocardial infarction, peripheral vascular disease, transient ischemic attack, and stroke—together with a regression-forest prediction of the outcome. We estimate the regression and causal forests using `grf`, with honest forests of 2,000 trees and tuning parameters selected by the package's cross-validation procedure (Tibshirani, Athey, and Wager 2020).

For outcome o, the ensemble treatment-effect prediction is

$$\widehat{\tau}^{o}(X_i) = \widehat{\tau}^{o}_{\text{post-lasso}}(X_i) + \widehat{\tau}^{o}_{\text{cf}}(X_i),$$

where the causal forest is fit to variation remaining after the post-LASSO component. We orient the proxies so that larger values indicate larger treatment consequences:

$$Z_B(x) \equiv -\widehat{\tau}^{s}(x), \qquad Z_H(x) \equiv \widehat{\tau}^{b}(x),$$

with leave-one-trial-out versions $Z_{B,-j}(x)$ and $Z_{H,-j}(x)$ defined analogously. Thus, larger Z_B indicates a larger predicted stroke reduction and larger Z_H a larger predicted increase in bleeding.

A.1.3 Cross-Trial Validation

We test whether the proxies predict treatment-effect heterogeneity in trials excluded from their estimation. For patient i in hold-out trial j, let $Z_{o,-j}(X_i)$ denote the relevant proxy, oriented so that larger values correspond to a larger treatment consequence, and define $q_s = -1$ and $q_b = 1$ to map that orientation back to the outcome scale. We also construct leave-one-trial-out LASSO and regression-forest predictions of each outcome using control observations from the other trials.

Following Chernozhukov et al. (2025), we estimate

$$\begin{aligned} Y_i^o = {} & \left[a_0^o + q_o a_1^o \{Z_{o,-j}(X_i) - \overline{Z}_o\}\right] D_i + q_o b_1^o \{Z_{o,-j}(X_i) - \overline{Z}_o\} p_j \\ & + b_2^o \widehat{m}^{o,\text{lasso}}_{-j}(X_i) + b_3^o \widehat{m}^{o,\text{rf}}_{-j}(X_i) + \zeta_j^o + \varepsilon_i^o, \end{aligned} \tag{A.1}$$

where p_j is the treatment probability in trial j, ζ_j^o denotes trial fixed effects, and $\overline{Z}_o$ is the mean leave-one-trial-out proxy. The interaction with p_j accommodates differences in randomization probabilities across trials. Trial fixed effects also absorb trial-level differences that could otherwise induce a mechanical relationship between a leave-one-trial-out proxy and the hold-out trial's realized average treatment effect.

With this sign normalization, $a_1^o > 0$ means that larger proxy values predict larger treatment effects in the direction indicated by the proxy. For stroke, $\widehat{a}_1^s$ is positive and statistically significant (Table 1, Column 1), so larger $Z_{B,-j}$ predicts larger treatment-induced reductions in stroke in hold-out trials. For bleeding, $\widehat{a}_1^b$ is wrong-signed and statistically insignificant (Column 2), providing no evidence that $Z_{H,-j}$ captures

stable heterogeneity in treatment-induced bleeding harms.

We separately test whether untreated bleeding risk estimated within the RCT data transfers across trials. Let $Z_{R,-j}(x)$ denote a bleeding-risk prediction estimated on control observations outside trial j. We estimate

$$Y_i^b = d_R Z_{R,-j}(X_i) + \xi_{j,D_i} + \varepsilon_i^R,$$

where ξ_{j,D_i} denotes trial-by-treatment-arm fixed effects. The estimate of d_R is wrong-signed and statistically insignificant (Table 1, Column 3). Figure 1 shows that the bleeding-harm proxy also has very little dispersion relative to the stroke-benefit proxy. Together, these results motivate using trial-based heterogeneity in stroke benefits while treating heterogeneity in treatment-induced bleeding harms as unresolved.

A.1.4 All-Trials Predictions and Application to the VHA

We fit all-trials versions of the same ensemble and apply them to VHA patient characteristics, averaging predictions across the ten bagged iterations. Although this produces both $Z_B(x)$ and $Z_H(x)$ in the VHA sample, the policy analysis retains only $Z_B(x)$ because only the benefit proxy validates across trials.

For descriptive analyses, we place $Z_B(x)$ on the stroke-benefit scale using the cross-trial BLP:

$$B_{\text{BLP}}(x) = -\widehat{a}_0^s + \widehat{a}_1^s\{Z_B(x) - \overline{Z}_B\}.$$

Here, $\overline{Z}_B$ is the AFI-sample mean of the proxy and is used to center both AFI and VHA predictions; the negative intercept converts the average treatment effect on the stroke-outcome scale into a positive stroke-prevention benefit. The main policy analysis instead uses the EWM calibration $B_{\text{EWM}}(x)$ described in Appendix A.3.

A.2 VHA Predictions of Untreated Bleeding Risk

Untreated bleeding risk is a baseline outcome prediction, not a treatment effect. We estimate two measures in the VHA data:

$$R(x) \equiv R_{\text{full}}^{\text{VHA}}(x), \qquad R_C(x) \equiv R_{\text{common}}^{\text{VHA}}(x).$$

The full-covariate prediction $R(x)$ is the main policy object and uses the richer set of characteristics available in the VHA. The common-covariate prediction $R_C(x)$ restricts the model to characteristics observed in both the VHA and AFI data. We estimate $R_C(x)$ in the VHA and apply the same fitted model to both samples, allowing comparisons of predicted bleeding risk on a common covariate basis. We also use it in the trial-based ML procedure described in Appendix A.1.2. After these definitions, we suppress the VHA superscript.

For both measures, the outcome is a gastrointestinal or intracranial hemorrhage within one year of the atrial fibrillation diagnosis, identified from VHA records supplemented by Medicare inpatient claims for Medicare-covered patients. We first estimate a LASSO-penalized Poisson model and then refit the selected variables using post-LASSO Poisson regression. We include anticoagulant treatment without

penalization; to predict untreated bleeding risk, we set the treatment indicator to zero. Candidate predictors include HAS-BLED components, frailty and fall-risk measures, demographics, Elixhauser comorbidities (Elixhauser et al. 1998), other medical history, vital signs, laboratory values, body measurements, the $CHADS_2$ score, health-care utilization, and eligible pairwise interactions. We omit labile INR because the analysis sample is restricted to patients who have not yet initiated warfarin. Appendix Table A.2 summarizes the availability of HAS-BLED components in the VHA and AFI data.

The common-covariate model uses the same estimation procedure but restricts predictors to characteristics observed in both data sources. Because $R_C(x)$ is estimated using the same VHA model in both populations, differences between its VHA and RCT distributions reflect differences in observed common covariates rather than differences in the fitted risk model. By contrast, $R(x)$ incorporates the richer VHA information and is therefore used for within-VHA descriptive and policy analyses.

Neither $R(x)$ nor $R_C(x)$ is interpreted as a treatment-induced bleeding harm. Under Assumption B, $R(x)$ indexes the relative distribution of treatment-induced harms, while their population average remains anchored to the randomized treatment effect. Because the bleeding-risk models are estimated in a much larger VHA sample than the trial-based treatment-effect models, our sampling-uncertainty analysis focuses on uncertainty in estimated stroke benefits rather than in $R(x)$.

A.3 Empirical Welfare Maximization Calibration of Stroke Benefits

A.3.1 Relationship Between Integrated EWM and the BLP

Let Y_i^s denote patient i's observed stroke outcome, D_i the anticoagulation indicator, and X_i the vector of covariates. Define the treatment-arm outcome regressions as

$$\mu_d(x) = E[Y_i^s \mid D_i = d, X_i = x], \qquad d \in \{0, 1\}.$$

Let $\widehat{\mu}_{d,-k(i)}$ denote the corresponding regression estimated without observations in the fold $k(i)$ containing patient i. Because treatment is randomized within trial, we use the trial-specific randomization probability $p_{j(i)} = \Pr(D_i = 1 \mid j(i))$.

The cross-fitted augmented inverse-probability-weighted pseudo-outcome for stroke-prevention benefit is

$$\psi_i^B = \widehat{\mu}_{0,-k(i)}(X_i) - \widehat{\mu}_{1,-k(i)}(X_i) + \frac{(1 - D_i)\left[Y_i^s - \widehat{\mu}_{0,-k(i)}(X_i)\right]}{1 - p_{j(i)}} - \frac{D_i\left[Y_i^s - \widehat{\mu}_{1,-k(i)}(X_i)\right]}{p_{j(i)}}. \tag{A.2}$$

This doubly robust transformation targets the conditional stroke-prevention benefit $E[Y_i^s(0) - Y_i^s(1) \mid X_i]$, so larger values indicate larger benefit (Robins, Rotnitzky, and Zhao 1994; Chernozhukov et al. 2018).

Let $m(z) \equiv E[\psi^B \mid Z_B = z]$ denote the conditional mean of the pseudo-outcome given the validated proxy. A calibration s places the proxy on the benefit scale so that the threshold rule at target benefit t treats patients for whom $s(Z_{B,i}) \geq t$.

Consider the population analog of the integrated criterion in Equation (4), with target thresholds

weighted by a nonnegative function $\nu(t)$:

$$W_\nu(s) = \int \nu(t)\, E\big[(\psi^B - t)\, \mathbf{1}(s(Z_B) \geq t)\big]\, dt. \tag{A.3}$$

Using iterated expectations, and defining $A_\nu(t) \equiv \int_{-\infty}^{t} \nu(u)du$ and $M_\nu(t) \equiv \int_{-\infty}^{t} u\nu(u)du$, this becomes $W_\nu(s) = E[m(Z_B)A_\nu(s(Z_B)) - M_\nu(s(Z_B))]$. For a linear score $s(z) = a_0 + a_1 z$, the first-order conditions are

$$E[(m(Z_B) - s(Z_B))\nu(s(Z_B))] = 0, \qquad E[Z_B(m(Z_B) - s(Z_B))\nu(s(Z_B))] = 0. \tag{A.4}$$

These are weighted normal equations, with residuals weighted by the threshold weights ν evaluated at the fitted score. Because these weights generally depend on the candidate score, this is not an ordinary weighted least-squares problem with fixed observation weights. If ν is constant over the range of the linear score, however, the weights drop out and the equations reduce exactly to the ordinary BLP normal equations. Thus, linear BLP is the special case of integrated EWM with a linear score and constant threshold weights.

In our setting, target benefits outside $[0, 1]$ have no decision-theoretic interpretation because stroke-prevention benefits are bounded probability reductions. Restricting ν to be constant on $[0, 1]$ and zero elsewhere gives the integrated criterion in Equation (4) and breaks the literal equivalence with the unrestricted BLP. For any bounded score $s(Z_B) \in [0, 1]$, the inner integral equals

$$\int_0^1 (\psi^B - t)\mathbf{1}(s(Z_B) \geq t)dt = \psi^B s(Z_B) - \frac{1}{2}s(Z_B)^2.$$

Therefore, the sample criterion is

$$W_n(s) = \mathbb{P}_n\left[\psi_i^B s(Z_{B,i}) - \frac{1}{2}s(Z_{B,i})^2\right]. \tag{A.5}$$

Since

$$\psi^B s - \frac{1}{2}s^2 = \frac{1}{2}(\psi^B)^2 - \frac{1}{2}(\psi^B - s)^2,$$

maximizing W_n is equivalent to minimizing $\mathbb{P}_n[(\psi_i^B - s(Z_{B,i}))^2]$ over the chosen class of bounded scores. Thus EWM calibration is a least-squares projection of the pseudo-outcome onto a bounded monotone score class. It differs from the linear BLP because the admissible score class produces welfare-relevant benefit predictions in $[0, 1]$, and in our implementation because we impose monotonicity and a smooth logistic-spline functional form.

A.3.2 Implementation

We implement $\widehat{s}_{\text{EWM}}$ as the composition of three monotone transformations: recentering the proxy using the RCT sample mean, a monotone cubic spline, and the logistic link,

$$\widehat{s}_{\text{EWM}}(z) = \Lambda\Big(g_\beta\Big(z - \overline{Z}_B\Big)\Big), \tag{A.6}$$

where $\overline{Z}_B$ is the mean proxy in the RCT sample and g_β is a monotone cubic spline. Because each component is monotone, the composition preserves the patient ordering in the validated proxy.

The recentered proxy is

$$\widetilde{Z}_{B,i} = Z_{B,i} - \overline{Z}_B,$$

where $\overline{Z}_B$ is calculated once in the RCT sample and applied to both RCT and VHA. The spline and its knot vector are estimated in RCT and reused for VHA prediction. For VHA prediction, recentered proxy values below or above the RCT boundary knots are set to the lower or upper boundary knot, respectively, before using the fitted spline to calculate the calibrated benefit.

The spline index $g_\beta(\widetilde{z}) = \sum_{k=1}^{5} \beta_k b_k(\widetilde{z})$ is a cubic B-spline with one interior knot and five basis functions. We impose monotonicity through nondecreasing coefficients, parameterized as $\beta_k = \beta_1 + \sum_{j=2}^{k} \exp(d_j)$. The logistic link $\Lambda(\cdot)$ restricts calibrated benefits to the open unit interval. We estimate $(\beta_1, d_2, \ldots, d_5)$ by nonlinear least squares,

$$\min_{\beta_1, d_2, \ldots, d_5} \mathbb{P}_n\left[\left(\psi_i^B - \Lambda\Big(g_\beta(\widetilde{Z}_{B,i})\Big)\right)^2\right], \tag{A.7}$$

which implements the integrated criterion for this bounded monotone score class. The calibrated benefit for a patient with covariates x is $B_{\mathrm{EWM}}(x) = \widehat{s}_{\mathrm{EWM}}(Z_B(x))$.

A.3.3 Specification Choice and Diagnostics

We choose the functional form using cross-fit welfare. For a calibration $\widehat{s}$, the integrated welfare contribution is $\psi_i^B \widehat{s}_i - \widehat{s}_i^2/2$, so $W_n(\widehat{s})$ can be evaluated from out-of-fold predictions. We use a nested design: an inner five-fold, within-trial loop cross-fits the outcome models entering ψ_i^B, and an outer five-fold loop refits the calibration outside each fold and predicts $\widehat{s}_i$ on the hold-out fold.

We also inspect two diagnostics. A binned scatter plot compares the average pseudo-outcome and average fitted score across proxy bins, checking whether the calibration tracks $m(\cdot)$ over the proxy support. The calibrated-benefit distribution is inspected for artifacts such as mass points, boundary pileup, or abrupt endpoints. These diagnostics are not the welfare objective, but they are relevant because fixed target thresholds are applied in calibrated-benefit units and therefore depend on the distribution of $s(Z_B)$.

A.4 Optimal Treatment Rules

This section derives the assumption-specific and ambiguity-robust treatment rules used in the main text. We first state the two endpoint harm models and the maximin and minimax-regret criteria. We then show that both robust criteria select threshold rules that are linear in untreated bleeding risk, characterize the balancing conditions that distinguish maximin from minimax regret, and discuss when the resulting families of rules are nested as the weight on bleeding harms varies.

A.4.1 Setup and Robust Decision Criteria

Consider a continuum of patients $i \in \mathcal{I}$. Let $B_i \in [0,1]$ denote the stroke-prevention benefit of treatment and $R_i \in [0, R_{\max}]$ untreated bleeding risk. Let $\overline{H}$ denote average treatment-induced bleeding harm and $\overline{R} = E[R_i]$ average untreated bleeding risk. We consider the two endpoint harm structures from the main text:

Assumption *A* (Mean-Independent Harms). *Conditional mean bleeding harm is constant:*

$$H_A(i) = \overline{H}.$$

Assumption *B* (Proportional Harms). *Conditional mean bleeding harm is proportional to untreated bleeding risk:*

$$H_B(i) = \frac{\overline{H}}{\overline{R}} R_i.$$

For a utility cost of a bleed relative to a stroke of $\lambda \geq 0$, the net expected utilities of treatment are

$$U_A(i;\lambda) = B_i - \lambda \overline{H}, \tag{A.8}$$

$$U_B(i;\lambda) = B_i - \lambda \frac{\overline{H}}{\overline{R}} R_i. \tag{A.9}$$

For a treatment rule $D : \mathcal{I} \to \{0,1\}$, endpoint welfare is

$$W_A(D;\lambda) = E[U_A(i;\lambda) D(i)], \tag{A.10}$$

$$W_B(D;\lambda) = E[U_B(i;\lambda) D(i)]. \tag{A.11}$$

For $\theta \in [0,1]$, define the convex combination

$$W(D;\lambda,\theta) = \theta W_A(D;\lambda) + (1-\theta) W_B(D;\lambda). \tag{A.12}$$

The endpoint-optimal rules are therefore

$$D_A^*(i;\lambda) = \mathbf{1}\left(B_i \geq \lambda \overline{H}\right), \tag{A.13}$$

$$D_B^*(i;\lambda) = \mathbf{1}\left(B_i \geq \lambda \frac{\overline{H}}{\overline{R}} R_i\right). \tag{A.14}$$

The maximin and minimax-regret criteria are

$$D_{\mathrm{MM}}^*(\cdot;\lambda) \in \arg\max_{D} \min_{\theta \in [0,1]} W(D;\lambda,\theta), \tag{A.15}$$

$$D_{\mathrm{MR}}^*(\cdot;\lambda) \in \arg\min_{D} \max_{\theta \in [0,1]} \left\{ \max_{D'} W(D';\lambda,\theta) - W(D;\lambda,\theta) \right\}. \tag{A.16}$$

The maximin criterion follows Wald (1950). The minimax-regret criterion follows Savage (1951); see

also Manski (2004) and Manski (2011) for treatment-choice applications. Because $W(D;\lambda,\theta)$ is linear in θ, the worst-case welfare for any fixed rule occurs at Assumption A or B. For minimax regret, the best attainable welfare $\max_{D'} W(D';\lambda,\theta)$ is the pointwise maximum of linear functions of θ and is therefore convex. Regret is consequently convex in θ, so its maximum over $[0,1]$ also occurs at an endpoint. Hence

$$D^*_{\text{MM}} \in \arg\max_D \min\{W_A(D), W_B(D)\},$$
$$D^*_{\text{MR}} \in \arg\min_D \max\{L_A(D), L_B(D)\},$$

where $L_\omega(D) \equiv W_\omega(D^*_\omega) - W_\omega(D)$ for $\omega \in \{A,B\}$ and dependence on λ is suppressed when convenient.

A.4.2 Characterization of Robust Treatment Rules

Define four regions according to whether the endpoint-optimal rules treat each patient:

$$\mathcal{I}_{+,+} = \{i : D^*_A(i) = 1,\ D^*_B(i) = 1\}, \qquad \mathcal{I}_{+,-} = \{i : D^*_A(i) = 1,\ D^*_B(i) = 0\},$$
$$\mathcal{I}_{-,+} = \{i : D^*_A(i) = 0,\ D^*_B(i) = 1\}, \qquad \mathcal{I}_{-,-} = \{i : D^*_A(i) = 0,\ D^*_B(i) = 0\}.$$

The first and last are agreement regions; the middle two are disagreement regions.

Remark 1. Any maximin or minimax-regret rule treats all patients in $\mathcal{I}_{+,+}$ and no patients in $\mathcal{I}_{-,-}$.

The result follows immediately because treatment raises welfare under both endpoint assumptions in $\mathcal{I}_{+,+}$ and lowers welfare under both in $\mathcal{I}_{-,-}$. Robustness therefore matters only in the two disagreement regions.

For the results below, assume that (R,B) has an absolutely continuous joint distribution. This ensures that a nontrivial linear treatment index is exactly zero only with probability zero. In the proofs, it is convenient to temporarily allow treatment probabilities $D(i) \in [0,1]$. For maximin, let t denote the welfare level guaranteed under both endpoint assumptions; for minimax regret, let q denote the largest regret across the two endpoints. These auxiliary quantities let us write each criterion using two simple endpoint constraints. Solving either problem amounts to placing nonnegative weights that sum to one on the two endpoint criteria. Conditional on those weights, a patient is treated exactly when the corresponding weighted average of U_A and U_B is positive. The proposition below makes this argument precise. Absolute continuity then implies that exact indifference occurs only with probability zero, so allowing treatment probabilities does not change the optimal rule almost surely.

Proposition 1 (Linear Robust Treatment Rules). *For either criterion* $\text{crit} \in \{\text{MM}, \text{MR}\}$ *and any* λ*, there exists a weight* $a_{\text{crit}}(\lambda) \in [0,1]$ *such that a robust rule can be written*

$$D^*_{\text{crit}}(i;\lambda) = \mathbf{1}\left(B_i \geq \lambda\left[a_{\text{crit}}(\lambda)\overline{H} + \{1 - a_{\text{crit}}(\lambda)\}\frac{\overline{H}}{\overline{R}}R_i\right]\right). \tag{A.17}$$

Equivalently, defining $\chi_{\text{crit}} = (1 - a_{\text{crit}})/a_{\text{crit}}$ *with the usual conventions at the endpoints,*

$$D^*_{\text{crit}}(i;\lambda) = \mathbf{1}\left(B_i \geq \lambda \left[\frac{1}{1+\chi_{\text{crit}}}\overline{H} + \frac{\chi_{\text{crit}}}{1+\chi_{\text{crit}}}\frac{\overline{H}}{\overline{R}}R_i\right]\right). \tag{A.18}$$

Thus every robust boundary is a straight line in (R, B) *space passing through* $(\overline{R}, \lambda\overline{H})$*, with slope between the A-optimal slope* 0 *and the B-optimal slope* $\lambda\overline{H}/\overline{R}$*.*

Proof. For maximin, write the relaxed problem as

$$\max_{D,t} t \quad \text{s.t.} \quad 0 \leq D(i) \leq 1, \qquad t \leq W_A(D), \qquad t \leq W_B(D).$$

Let α_A and α_B denote the nonnegative Lagrange multipliers on the two welfare constraints. The first-order condition for t gives $\alpha_A + \alpha_B = 1$. Writing $a = \alpha_A$, the part of the Lagrangian that depends on the treatment rule is therefore

$$aW_A(D) + (1-a)W_B(D).$$

Because welfare is additive across patients, this weighted welfare is maximized by setting $D(i) = 1$ when

$$aU_A(i) + (1-a)U_B(i) > 0$$

and $D(i) = 0$ when the index is negative. Absolute continuity makes equality a probability-zero event. Substituting Equations (A.8) and (A.9) yields Equation (A.17). The endpoint cases $a = 1$ and $a = 0$ reproduce the A- and B-optimal rules, respectively.

For minimax regret, write the relaxed problem as

$$\min_{D,q} q \quad \text{s.t.} \quad 0 \leq D(i) \leq 1, \qquad q \geq L_A(D), \qquad q \geq L_B(D).$$

Let α_A and α_B now denote the nonnegative multipliers on the two regret constraints. The first-order condition for q again gives $\alpha_A + \alpha_B = 1$. Since $W_A(D^*_A)$ and $W_B(D^*_B)$ do not depend on D, minimizing the weighted sum of endpoint regrets is equivalent to maximizing

$$\alpha_A W_A(D) + \alpha_B W_B(D).$$

The same patient-by-patient treatment rule therefore follows. Reparameterizing $a = \alpha_A$ by $\chi = (1-a)/a$ gives Equation (A.18). □

The weight a_{crit} has a direct interpretation. At $a_{\text{crit}} = 1$, only welfare under Assumption A matters and the rule is A-optimal; at $a_{\text{crit}} = 0$, only Assumption B matters and the rule is B-optimal. Interior values interpolate between these two harm structures.

Equation (A.18) also yields the gain–loss interpretation used in the main text. In $\mathcal{I}_{+,-}$, where $U_A > 0 > U_B$, treatment occurs when

$$U_A(i;\lambda) \geq \chi_{\text{crit}}(\lambda)\{-U_B(i;\lambda)\}.$$

In $\mathcal{I}_{-,+}$, where $U_A < 0 < U_B$, the same parameter satisfies

$$-U_A(i;\lambda) \leq \chi_{\text{crit}}(\lambda) U_B(i;\lambda).$$

Thus the same χ_{crit} governs both disagreement regions. In $\mathcal{I}_{+,-}$, the gain under Assumption A must be large enough relative to the loss under Assumption B; in $\mathcal{I}_{-,+}$, the loss under Assumption A must be small enough relative to the gain under Assumption B. Equivalently, χ_{crit} defines a common cutoff on $|U_A/U_B|$, with the inequality reversed across the two regions.

A.4.3 Balancing Conditions and Corner Solutions

The maximin and minimax-regret criteria select different values of the common weight in Proposition 1. Their distinguishing conditions are especially simple for interior solutions.

Proposition 2 (Maximin Equality of Welfare). *If the maximin solution is interior, so that it differs from both endpoint-optimal rules, then*

$$W_A(D^*_{\text{MM}};\lambda) = W_B(D^*_{\text{MM}};\lambda). \tag{A.19}$$

Proof. In the constrained formulation used above, a zero multiplier on one endpoint would reduce the weighted-welfare rule to the optimum for the other endpoint. An interior solution therefore places positive weight on both endpoint constraints. Complementary slackness then implies that both welfare constraints bind, which gives Equation (A.19). □

The maximin solution can instead be an endpoint rule.

Remark 2. If

$$W_A(D^*_A;\lambda) \leq W_B(D^*_A;\lambda),$$

then D^*_A is maximin. Symmetrically, if

$$W_B(D^*_B;\lambda) \leq W_A(D^*_B;\lambda),$$

then D^*_B is maximin.

For example, the first condition says that Assumption A is already the worst-case endpoint for the A-optimal rule. Since D^*_A maximizes welfare under A, no other rule can improve its worst-case welfare.

Proposition 3 (Minimax-Regret Equality of Regret). *When the endpoint-optimal rules differ on a set of positive measure, the minimax-regret solution equalizes endpoint regret:*

$$W_A(D^*_A;\lambda) - W_A(D^*_{\text{MR}};\lambda) = W_B(D^*_B;\lambda) - W_B(D^*_{\text{MR}};\lambda). \tag{A.20}$$

Proof. Use the constrained minimax-regret formulation in the proof of Proposition 1. Suppose, for example, that only the Assumption A regret constraint received positive weight. The weighted-welfare problem would

then select D_A^*. But at D_A^*, regret under Assumption A is zero while regret under Assumption B is strictly positive whenever the endpoint-optimal rules differ on a set of positive measure. The Assumption A constraint would therefore be slack rather than binding, contradicting its positive multiplier. The same argument rules out placing weight only on Assumption B. Hence both regret constraints receive positive weight, and complementary slackness implies that both bind, yielding Equation (A.20). □

For maximin, equality of endpoint welfare has an especially useful implication. For any rule D with positive treatment probability,

$$W_A(D;\lambda) - W_B(D;\lambda) = \lambda \frac{\overline{H}}{\overline{R}} \Pr(D(i)=1) \left\{ E[R_i \mid D(i)=1] - \overline{R} \right\}. \tag{A.21}$$

Combining this identity with Proposition 2 gives the following characterization.

Remark 3. Any interior maximin rule satisfies

$$E[R_i \mid D^*_{\mathrm{MM}}(i;\lambda)=1] = \overline{R}. \tag{A.22}$$

Thus the maximin treated set has the same average untreated bleeding risk as the population.

Equal regret also implies an analogous condition on average untreated bleeding risk among treated patients. Unlike maximin, however, the target need not equal the population mean $\overline{R}$; it depends on the gap between the best attainable welfare under Assumptions A and B and on the treated share. We do not use this auxiliary condition to compute the minimax-regret rule.

A.4.4 Nestedness as λ Varies

A single patient ordering summarizes a family of treatment rules only when the treated sets are nested as λ changes. We say that a family $\{D(\cdot;\lambda)\}_{\lambda \geq 0}$ is *nested* if, whenever $\lambda' < \lambda$,

$$D(i;\lambda) \leq D(i;\lambda') \qquad \text{for all } i.$$

The A- and B-optimal families are nested immediately from Equations (A.13) and (A.14). For robust rules, the weight in Equation (A.17) may itself vary with λ, so linearity of each individual boundary does not by itself guarantee nestedness.

The maximin moment condition gives a useful sufficient condition for the interior maximin family. Let $Q_{B|R}(q \mid r)$ denote the qth conditional quantile of stroke benefit B among patients with untreated bleeding risk $R = r$.

Remark 4 (A sufficient condition for interior maximin nestedness). Consider a range of λ for which the maximin solution is interior. For each λ, let $q(\lambda)$ satisfy

$$Q_{B|R}(q(\lambda) \mid \overline{R}) = \lambda \overline{H}.$$

Suppose that $r \mapsto Q_{B|R}(q(\lambda) \mid r)$ is linear and has slope between the A- and B-optimal slopes. Then the interior maximin treatment sets over this range are nested.

Proof. Fix λ. The rule that treats patients above the boundary $Q_{B|R}(q(\lambda) \mid R)$ treats the same share $1 - q(\lambda)$ at every value of R. Its treated patients therefore have average untreated bleeding risk $\overline{R}$, so Equation (A.21) implies that welfare is the same under Assumptions A and B. The assumed slope also places this boundary between the A- and B-optimal boundaries, so Proposition 1 implies that it maximizes a weighted average $aW_A + (1-a)W_B$ for some $a \in [0,1]$. A rule with strictly higher worst-case welfare would raise both endpoint welfares and hence this weighted average, which is impossible. The boundary is therefore maximin.

Now consider $\lambda' < \lambda$. At $R = \overline{R}$, the corresponding boundary levels are $\lambda'\overline{H} < \lambda\overline{H}$, so $q(\lambda') \leq q(\lambda)$. Conditional quantiles are ordered in their quantile index, implying

$$Q_{B|R}(q(\lambda') \mid r) \leq Q_{B|R}(q(\lambda) \mid r)$$

for every r. The treatment region can therefore only shrink as λ increases, establishing nestedness. □

For minimax regret, the same argument does not automatically apply because equal regret does not pin the treated group to a fixed average-risk target as λ changes. Rather than impose additional restrictions, the empirical implementation checks nestedness directly for the estimated minimax-regret family. We also check the full maximin family directly, including transitions between interior and endpoint solutions. These checks determine whether each robust family can be summarized by a single patient ordering over the relevant range of λ.

A.5 Sampling Uncertainty in Estimated Benefits

This appendix examines how finite-sample uncertainty in estimated stroke benefits affects patient rankings, treatment assignments, and welfare. We represent sampling uncertainty using repeated realizations of patient-level benefit estimates, while holding the VHA target population, bleeding-risk predictions, and the two benchmark harm structures fixed.

A.5.1 Benefit Realizations and Ranking Stability

Let $B^{(m)}(x)$ denote the estimated stroke-reduction benefit for patient characteristics x in realization m, for $m = 1, \ldots, M$ and $M = 100$. We begin with 100 re-estimates of the benefit model, each based on a random 95% subsample of RCT observations within trial-by-treatment cells. For each realization, we randomly select 10 re-estimates, average their RCT predictions and re-estimate the EWM calibration, and apply that calibration to the VHA predictions averaged over the same 10 re-estimates. Each realization is constructed jointly for all VHA patients, preserving the cross-patient dependence induced by the common resampled

trial data and estimated model. Define the mean benefit estimate as

$$\overline{B}(X_i) = \frac{1}{M}\sum_{m=1}^{M} B^{(m)}(X_i).$$

For each realization, we recompute the patient ordering under Assumptions A and B. Under Assumption A, patients are ordered by $B^{(m)}(X_i)$; under Assumption B, they are ordered by $B^{(m)}(X_i)/R(X_i)$. Let $\Omega_\omega^{(m)}(i)$ denote patient i's position under harm structure $\omega \in \{A, B\}$, with lower values indicating higher treatment priority. For an ordered patient pair (i, j), define

$$p_{ij}^{\omega} = \frac{1}{M}\sum_{m=1}^{M} \mathbf{1}\left(\Omega_\omega^{(m)}(i) < \Omega_\omega^{(m)}(j)\right).$$

At the stability standard $c = 0.9$, the pair is unresolved if

$$1 - c < p_{ij}^{\omega} < c.$$

Thus, a pair is resolved only if the same patient is prioritized first in at least 90% of benefit realizations. This is an empirical ranking-stability standard over the 100 realizations, not a conventional confidence interval or significance level. We evaluate these probabilities using Monte Carlo samples of ordered patient pairs across all realizations.

At $c = 0.9$, 14.2% of patient pairs are unresolved under Assumption A, and 20.2% are unresolved under Assumption B. These statistics measure instability in fine patient rankings; they do not by themselves imply changes in treatment assignments or welfare.

A.5.2 Sampling-Aware Welfare and Regret

We next incorporate the same benefit realizations into the welfare analysis. Let G denote the empirical distribution that places equal weight on the $M = 100$ joint benefit realizations. For harm structure $\omega \in \{A, B\}$, write $W_\omega(D; \lambda; B^{(m)})$ for welfare under treatment rule D and benefit realization m. Because welfare in Equation (8) is additive and linear in patient-level benefits,

$$\frac{1}{M}\sum_{m=1}^{M} W_\omega(D; \lambda; B^{(m)}) = W_\omega(D; \lambda; \overline{B})$$

for any fixed rule D. It follows that averaging over sampling uncertainty leaves the Assumption-A-optimal, Assumption-B-optimal, and maximin rules unchanged relative to evaluation at the mean benefit estimates $\overline{B}$.

Minimax regret can change because its benchmark is the optimal rule under each realized benefit realization. Define expected realization-specific optimal welfare under harm structure ω as

$$W_\omega^{*,G}(\lambda) = \frac{1}{M}\sum_{m=1}^{M} \max_{D'} W_\omega(D'; \lambda; B^{(m)}),$$

and define sampling-aware regret as

$$L_{\omega,G}(D;\lambda) = W^{*,G}_{\omega}(\lambda) - W_{\omega}(D;\lambda;\overline{B}).$$

If D^{*}_{ω} is the assumption-specific optimum under $\overline{B}$, define the Jensen gap

$$J^{\omega}(\lambda) = W^{*,G}_{\omega}(\lambda) - W_{\omega}(D^{*}_{\omega};\lambda;\overline{B}).$$

Then

$$L_{\omega,G}(D;\lambda) = L_{\omega}(D;\lambda) + J^{\omega}(\lambda),$$

where $L_{\omega}(D;\lambda)$ is regret relative to D^{*}_{ω} under the mean benefit estimates $\overline{B}$. Thus, $J^{\omega}(\lambda)$ is the expected welfare gain from knowing the realized patient-specific benefits before choosing the assumption-specific optimal rule. The sampling-aware minimax-regret rule, $D^{*,G}_{\mathrm{MR}}$, minimizes the maximum of $L_{A,G}$ and $L_{B,G}$.

Appendix Table A.6 reports the welfare and assignment consequences at the common 50%-stroke-prevention benchmark. The Jensen gap is 1.56 per 10,000 patients under Assumption A and 2.06 under Assumption B. By comparison, regret from using the sampling-aware minimax-regret rule rather than the corresponding assumption-specific optimum under the mean benefit estimates is 17.58 and 17.09, respectively. The sampling-aware minimax-regret assignment itself differs from the minimax-regret assignment based on the mean benefit estimates for only 0.12% of patients. Hence sampling uncertainty creates noticeable instability in fine rankings but relatively little instability in the resulting robust treatment assignment or welfare. Table A.6 additionally reports regret in bleed-equivalent units and assignment disagreement relative to the assumption-specific and realization-specific optima.

A.6 A Skill-Threshold Model for Physician Treatment Rules

Following Chan, Gentzkow, and Yu (2022), we estimate a physician-level skill-threshold model of anticoagulation decisions. The model takes the patient-level stroke-prevention benefit and bleeding risk as given and asks how strongly physicians' treatment decisions follow the patient ordering that is optimal under each harm assumption. We restrict the estimation sample to physicians with at least 20 patients.

A.6.1 Patient Rankings and Treatment Decisions

We estimate the model separately relative to the A-optimal and B-optimal patient orderings. For each specification $\omega \in \{A, B\}$, let $r_i^{\prime,\omega}$ denote the corresponding raw sort score, oriented so that a lower value denotes higher treatment priority. We rank patients from lowest to highest $r_i^{\prime,\omega}$, assigning average ranks to ties. If m_i^{ω} is patient i's rank among N patients, we transform the rank to a standard-normal score,

$$r_i^{\omega} = \Phi^{-1}\left(1 - \frac{m_i^{\omega}}{N+1}\right), \tag{A.23}$$

so that larger r_i^{ω} denotes higher treatment priority and the scale is comparable across the two specifications.

Physician j has a skill parameter k_j^ω and a threshold parameter p_j^ω. Treatment follows the latent-index model

$$d_{ij} = \mathbf{1}\left(k_j^\omega r_i^\omega + \varepsilon_{ij} > p_j^\omega\right), \qquad \varepsilon_{ij} \sim N(0,1), \tag{A.24}$$

or equivalently,

$$\Pr(d_{ij} = 1 \mid k_j^\omega, p_j^\omega, r_i^\omega) = \Phi\left(k_j^\omega r_i^\omega - p_j^\omega\right). \tag{A.25}$$

The threshold p_j^ω captures overall treatment aggressiveness: holding skill fixed, physicians with higher thresholds treat fewer patients. The slope k_j^ω captures alignment with the ω-optimal ordering. Larger positive values imply stronger prioritization of high-priority patients; $k_j^\omega = 0$ implies no relationship between treatment and the ordering; and negative values imply treatment against the ordering. Skill is therefore specific to the ordering used to evaluate decisions rather than an assumption-independent measure of clinical quality.

A.6.2 Population Distribution and Estimation

Because individual physicians treat too few patients to estimate (k_j^ω, p_j^ω) precisely in isolation, we model these parameters as random coefficients. For each ordering specification,

$$\begin{pmatrix} k_j^\omega \\ p_j^\omega \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_k^\omega \\ \mu_p^\omega \end{pmatrix}, \begin{pmatrix} (\sigma_k^\omega)^2 & \rho_{k,p}^\omega \sigma_k^\omega \sigma_p^\omega \\ \rho_{k,p}^\omega \sigma_k^\omega \sigma_p^\omega & (\sigma_p^\omega)^2 \end{pmatrix}\right). \tag{A.26}$$

The correlation $\rho_{k,p}^\omega$ allows physicians with different skill slopes to have systematically different treatment thresholds.

Let I_j denote the patients assigned to physician j, let $\vartheta_j^\omega = (k_j^\omega, p_j^\omega)$, and define the physician-level likelihood

$$\mathcal{L}_j^\omega(d_j \mid \vartheta_j^\omega) = \prod_{i \in I_j} \left[\Phi(k_j^\omega r_i^\omega - p_j^\omega)\right]^{d_i} \left[1 - \Phi(k_j^\omega r_i^\omega - p_j^\omega)\right]^{1-d_i}.$$

We estimate the population parameters $\eta^\omega = (\boldsymbol{\mu}^\omega, \boldsymbol{\Sigma}^\omega)$ by simulated maximum likelihood,

$$\widehat{\eta}^\omega = \arg\max_{\eta^\omega} \sum_j \log \int \mathcal{L}_j^\omega(d_j \mid \vartheta)\, f(\vartheta \mid \eta^\omega)\, d\vartheta. \tag{A.27}$$

The integral over physician random effects has no closed form, so we approximate it by simulation from the candidate bivariate-normal population distribution (Train 2003). We estimate the *A*- and *B*-ordering specifications separately, producing distinct population distributions of skill and thresholds under the two harm assumptions.

A.6.3 Empirical Bayes Physician Skill

For the physician-level adherence analysis, we summarize each physician's skill using the empirical Bayes posterior mean of k_j^{ω} (Morris 1983). Conditional on the estimated population distribution,

$$\widehat{k}_j^{\mathrm{EB},\omega} = \frac{\int k\, \mathcal{L}_j^{\omega}(d_j \mid k, p)\, f(k, p \mid \widehat{\eta}^{\omega})\, dk\, dp}{\int \mathcal{L}_j^{\omega}(d_j \mid k, p)\, f(k, p \mid \widehat{\eta}^{\omega})\, dk\, dp}. \tag{A.28}$$

The posterior mean shrinks physician-specific skill toward the population mean, with greater shrinkage when a physician's treatment decisions contain less information about skill.

We evaluate the posterior mean by Monte Carlo integration using 2,000 draws from the fitted bivariate-normal population distribution for each specification. Each draw is weighted by the likelihood of physician j's observed treatment decisions, and the posterior skill estimate is the likelihood-weighted average of the simulated skill draws. Because the draws come directly from the fitted population distribution, no additional prior-density weighting is required.

References

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

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.

Breiman, Leo. 1996. "Bagging Predictors." *Machine Learning* 24 (2): 123–140.

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.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. "Double/Debiased Machine Learning for Treatment and Structural Parameters." *The Econometrics Journal* 21 (1): C1–C68.

Chernozhukov, Victor, Mert Demirer, Esther Duflo, and Iván Fernández-Val. 2025. "Fisher-Schultz Lecture: Generic Machine Learning Inference on Heterogeneous Treatment Effects in Randomized Experiments, With an Application to Immunization in India." *Econometrica* 93 (4): 1121–1164.

Elixhauser, Anne, Claudia Steiner, D. Robert Harris, and Rosanna M. Coffey. 1998. "Comorbidity Measures for Use with Administrative Data." *Medical Care* 36 (1): 8–27.

Manski, Charles F. 2004. "Statistical Treatment Rules for Heterogeneous Populations." *Econometrica* 72 (4): 1221–1246.

———. 2011. "Choosing Treatment Policies Under Ambiguity." *Annual Review of Economics* 3:25–49.

Morris, Carl N. 1983. "Parametric Empirical Bayes Inference: Theory and Applications." *Journal of the American Statistical Association* 78 (381): 47–55.

Robins, James M., Andrea Rotnitzky, and Lue Ping Zhao. 1994. "Estimation of Regression Coefficients When Some Regressors are Not Always Observed." *Journal of the American Statistical Association* 89 (427): 846–866.

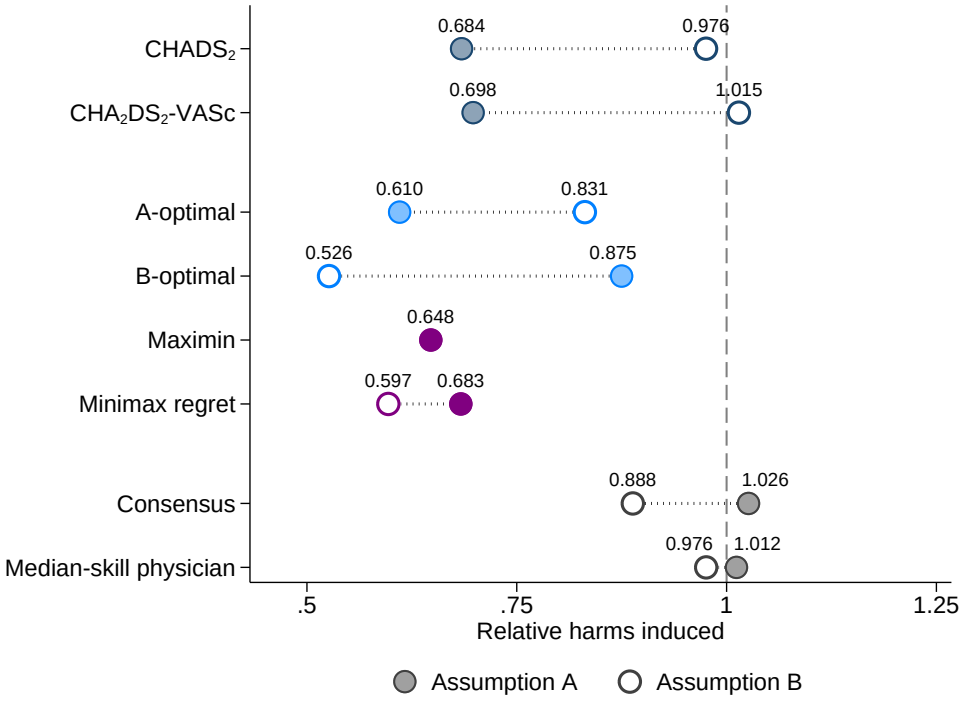
Savage, L. J. 1951. "The Theory of Statistical Decision." *Journal of the American Statistical Association* 46 (253): 55–67.

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

Train, Kenneth. 2003. *Discrete Choice Methods with Simulation.* University of California, Berkeley: Cambridge University Press.

Wald, Abraham. 1950. *Statistical Decision Functions.* New York: John Wiley & Sons.

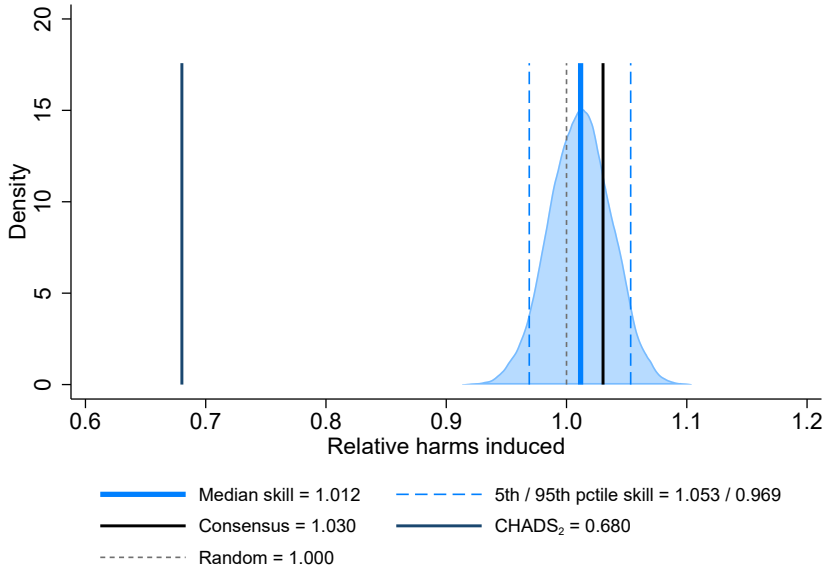
Figure A.1: Performance Using BLP-Projected Stroke Benefits



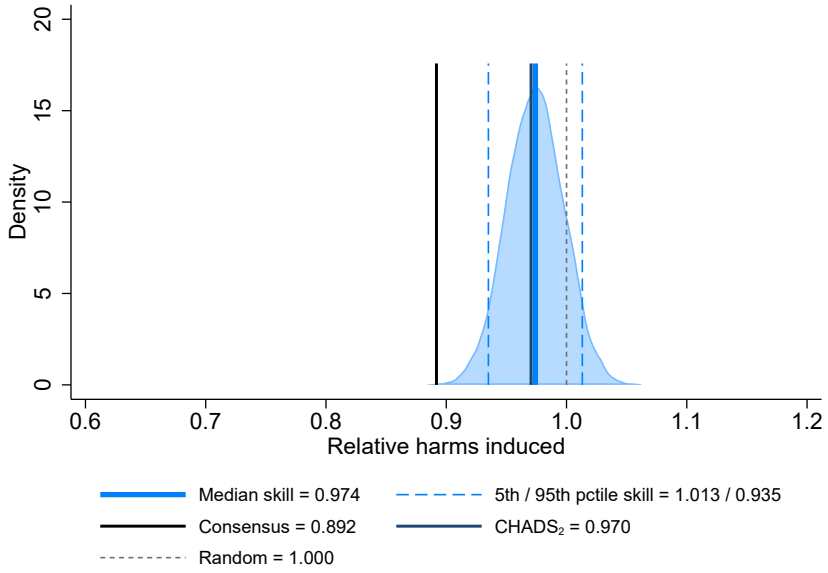
Notes: This figure reproduces Figure 9 using BLP-projected stroke benefits, $B_{\text{BLP}}(x)$, instead of EWM-calibrated stroke benefits, $B_{\text{EWM}}(x)$. It compares random treatment, existing guideline orderings, optimal and robust treatment rules, the consensus rule, and the physician benchmark at a fixed stroke-prevention benchmark. All comparisons hold strokes prevented fixed at the level obtained by treating 50% of patients randomly. The horizontal axis reports bleeds induced relative to those induced by 50% random treatment, so values below one indicate fewer bleeds. Shaded circles evaluate harms under Assumption A (mean-independent harms), and hollow circles evaluate harms under Assumption B (proportional harms).

Figure A.2: Simulated Physician Performance at the Fixed-Stroke Benchmark

A. Mean-Independent Harms



B. Proportional Harms



Notes: This figure displays the distribution of simulated physician performance at the fixed stroke-prevention benchmark. Simulated physicians are drawn from the estimated skill-threshold model described in Section 5.2 and Appendix A.6. For each simulated physician, we vary the treatment threshold along the physician's implied treatment ordering and evaluate the number of bleeds induced when the rule prevents the same number of strokes as treating 50% of patients randomly. The horizontal axis reports bleeds induced relative to bleeds induced under 50% random treatment, so values below one indicate fewer bleeds than random treatment. Panel A evaluates harms under Assumption *A* (mean-independent harms), and Panel B evaluates harms under Assumption *B* (proportional harms). Vertical lines mark selected percentiles of the simulated physician distribution and selected benchmark rules. No simulated physician outperforms the $CHADS_2$ guideline under Assumption *A*, while 44.5% do so under Assumption *B*.

Table A.1: Sample Selection

Sample step	Description	Observations Dropped	Observations Remaining
Panel A: VHA sample			
1. Identify potentially new AF patients	AF diagnosis with no AF diagnosis recorded in the prior three years.		844,845
2. Prescription restriction	VA prescription fill in the prior year and no prior anticoagulation prescription.	300,222	544,623
3. PCP visit before index diagnosis	VA PCP visit within six months before the index visit.	102,948	441,675
4. Confirmed AF diagnosis	EKG within 30 days before or after the index diagnosis and a second AF diagnosis 30–365 days later.	183,433	258,242
5. PCP or cardiologist visit	PCP or cardiology visit within 90 days; the earliest visit identifies the attributed physician, who must have prescribed at least one non-warfarin drug within one year before or after the visit.	94,425	163,817
6. Physicians with sufficient sample	Attributed physician has at least 30 AF patients and at least one warfarin prescription in the unrestricted step-1 sample.	32,374	131,443
Panel B: RCT sample			
1. Raw AFI meta-analysis data	All patients across the 10 trials in the AFI database.		9,155
2. Restrict to eligible patients	Match the 2009 AFI Stroke cohort and apply each trial's original randomization eligibility criteria.	1,754	7,401
3. Drop observations with missing variables	Require nonmissing histories of TIA or stroke, diabetes, hypertension, and congestive heart failure.	2	7,399
4. Drop unusual observations	Drop patients with negative computed age.	54	7,345
5. Drop low-dose warfarin arms	Exclude low-dose warfarin and low-dose warfarin plus aspirin arms.	691	6,654
6. Drop one-armed trial groups	Drop SPAF1, SPAF3, EAFT group 2, and PATAF group 2, which lack a treatment or control arm.	1,934	4,720

Notes: Panel A describes key VHA sample-selection steps. Panel B describes key RCT sample-selection steps.

Table A.2: HAS-BLED Component Availability by Data Source

HAS-BLED component	VHA definition	VHA	AFI
Hypertension	Systolic blood pressure above 160, using the reading closest to AF diagnosis	Yes	Yes
Abnormal renal function	Elixhauser renal failure indicator	Yes	No
Abnormal liver function	Elixhauser liver disease indicator	Yes	No
Stroke history	Prior stroke or TIA diagnosis	Yes	Yes
Prior bleeding	Any prior hemorrhage diagnosis	Yes	No
Labile INR	Excluded	Excluded	Excluded
Elderly	Age above 65 at AF diagnosis	Yes	Yes
Drugs (NSAID)	NSAID prescription fill in the year before AF diagnosis	Yes	No
Alcohol use	Elixhauser alcohol abuse indicator	Yes	No

Notes: This table reports the availability of each HAS-BLED component in the VHA and AFI trial data and its definition in the VHA data. Only hypertension, stroke history, and age are recorded in the AFI trials. The labile-INR component is excluded from both models because the sample is restricted to patients not yet prescribed warfarin.

Table A.3: 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

Notes: This table summarizes the first four randomized trials in the AFI database, including trial dates, eligibility criteria, exclusions, treatment arms, and sample sizes. Full-dose warfarin or oral-anticoagulant arms are treated as anticoagulation; placebo, control, and aspirin-only arms are treated as untreated with anticoagulation. Low-dose warfarin arms are excluded. Abbreviations: AF = atrial fibrillation; AP = antiplatelet therapy; BP = blood pressure; DBP = diastolic blood pressure; ECG = electrocardiogram; INR = international normalized ratio; OAC = oral anticoagulation; SBP = systolic blood pressure; TIA = transient ischemic attack.

Table A.4: 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

Notes: This table summarizes the remaining randomized trials in the AFI database, including trial dates, eligibility criteria, exclusions, treatment arms, and sample sizes. Full-dose warfarin, coumarin, or oral-anticoagulant arms are treated as anticoagulation; placebo, control, and aspirin-only arms are treated as untreated with anticoagulation. Low-dose anticoagulation arms are excluded. Abbreviations: AF = atrial fibrillation; AP = antiplatelet therapy; ECG = electrocardiogram; EF = ejection fraction; GI = gastrointestinal; INR = international normalized ratio; OAC = oral anticoagulation; TIA = transient ischemic attack.

Table A.5: Balance in Patient Characteristics

Patient characteristic	Means		Missing (%)	Within-trial difference			
	Treatment	Control		Non-imputed		Imputed	
				Est.	SE	Est.	SE
Panel A: Demographics							
Age	70.3	70.4	0.00%	0.180	0.247	0.180	0.248
Age $\geq$ 65	0.76	0.77	0.00%	0.003	0.012	0.003	0.013
Male	0.68	0.67	0.00%	-0.016	0.013	-0.016	0.013
White	0.94	0.95	23.01%	0.003	0.007	0.003	0.006
Panel B: Patient medical history							
Congestive heart failure	0.288	0.305	0.00%	-0.006	0.012	-0.006	0.012
Hypertension	0.459	0.447	0.00%	-0.008	0.015	-0.008	0.015
Diabetes	0.138	0.136	0.00%	-0.006	0.010	-0.006	0.010
Stroke	0.125	0.174	0.00%	-0.006	0.008	-0.006	0.008
Myocardial infarction	0.106	0.110	0.00%	-0.011	0.009	-0.011	0.009
Angina	0.168	0.166	0.00%	0.003	0.011	0.003	0.011
Transient ischemic attack	0.032	0.040	0.00%	0.006	0.005	0.006	0.005
Peripheral vascular disease	0.085	0.070	29.75%	0.011	0.009	0.007	0.007
Panel C: Lifestyle factor							
Smoker	0.516	0.534	40.04%	-0.004	0.017	-0.004	0.011
Panel D: Clinical measures							
Weight	175.9	173.3	16.44%	-1.198	1.240	-1.118	1.039
Systolic blood pressure	142.4	144.4	0.06%	-0.405	0.565	-0.379	0.563
Diastolic blood pressure	82.8	83.6	0.08%	-0.267	0.287	-0.243	0.287
Hemoglobin	14.8	14.7	30.28%	-0.008	0.049	-0.018	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" columns show the difference when using the original data, dropping observations with missing values. The "imputed" columns show 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 reported in the SE columns.

Table A.6: Regret from Sampling Uncertainty and Harm Ambiguity

Measure	Assumption A	Assumption B
Panel A: Welfare regret per 10,000 patients		
Jensen gap	1.56	2.06
Regret of $D^{*,G}_{\text{MR}}$ relative to D^{*}_{ω}	17.58	17.09
Expected regret of $D^{*,G}_{\text{MR}}$ relative to $D^{*,(m)}_{\omega}$	19.15	19.15
Panel B: Welfare regret in bleed-equivalent units per 10,000 patients		
Jensen gap	1.98	2.61
Regret of $D^{*,G}_{\text{MR}}$ relative to D^{*}_{ω}	22.29	21.66
Expected regret of $D^{*,G}_{\text{MR}}$ relative to $D^{*,(m)}_{\omega}$	24.28	24.28
Panel C: Treatment-assignment disagreement		
$D^{*,G}_{\text{MR}}$ versus D^{*}_{MR}	0.12%	
$D^{*,G}_{\text{MR}}$ versus D^{*}_{ω}	14.46%	20.69%
$D^{*,G}_{\text{MR}}$ versus $D^{*,(m)}_{\omega}$, averaged across benefit realizations	15.03%	21.08%

Notes: For Assumption $\omega \in \{A, B\}$, D^{*}_{ω} denotes the assumption-specific optimal rule under $\overline{B}$, $D^{*,(m)}_{\omega}$ denotes the assumption-specific optimal rule under benefit realization $B^{(m)}$, and D^{*}_{MR} denotes the minimax-regret rule under $\overline{B}$. The Jensen gap measures the expected welfare gain from knowing the benefit realization before choosing the assumption-specific optimal rule. The second row of Panel A measures regret from accommodating harm ambiguity when benefits are fixed at $\overline{B}$. The third row combines the two sources of uncertainty and equals, up to rounding, the sum of the preceding two rows within each assumption. Panel B converts the Panel A quantities to bleed-equivalent units by dividing by λ^{G}_{50}. Panel C reports the corresponding treatment-assignment disagreements.