NBER WORKING PAPER SERIES

MASTERING THE ART OF COOKBOOK MEDICINE: MACHINE LEARNING, RANDOMIZED TRIALS, AND MISALLOCATION

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

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

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

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

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

Mastering the Art of Cookbook Medicine: Machine Learning, Randomized Trials, and Misallocation
Jason Abaluck, Leila Agha, David C. Chan Jr, Daniel Singer, and Diana Zhu
NBER Working Paper No. 27467
July 2020
JEL No. I11,I18,O33

ABSTRACT

The application of machine learning (ML) to randomized controlled trials (RCTs) can quantify and improve misallocation in healthcare. We study the decision to prescribe anticoagulants for atrial fibrillation patients; anticoagulation reduces stroke risk but increases hemorrhage risk. We combine observational data on treatment choice and guideline use with ML estimates of heterogeneous treatment effects from eight RCTs. When physicians adopt a clinical guideline, treatment decisions shift towards the recommendation but adherence remains far from perfect. Improving guideline adherence would produce larger gains than informing physicians about guidelines. Adherence to an optimal rule would prevent 47% more strokes without increasing hemorrhages.

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

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

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

Daniel Singer
Harvard Medical School and
Massachusetts General
Hospital
DESINGER@mgh.harvard.edu

Diana Zhu
Yale University
dianayilunzhu@gmail.com

1 Introduction

In many settings, optimal decisions hinge on the relative magnitude of treatment effects. In medicine, this recognition spurs large investments in research, including randomized controlled trials (RCTs). Clinical guidelines are one of the principal ways that the resulting evidence is communicated to physicians, with approximately 250,000 peer-reviewed papers about clinical scoring systems published over the past 50 years (Challener et al., 2019). Despite these investments in developing evidence-based recommendations, research documents large inefficiencies in care allocation.[1]

One explanation for persistent misallocation is the low adherence rate to clinical guidelines. An alternative view is that the application of existing guidelines may not substantially improve care allocation or may even worsen health outcomes. Critics deride guidelines as "cookbook medicine," arguing that they distract physicians from using clinically relevant information not explicitly encoded in guidelines.[2] Modern machine learning (ML) techniques allow the estimation of heterogeneous treatment effects from randomized trials while imposing fewer ex ante restrictions on the variables used or functional forms. Decision rules based on granular ML-predicted treatment effects may mitigate concerns about cookbook medicine. ML techniques also open a window to evaluate these concerns in the context of existing recommendations.

In this paper, we study how physicians employ an existing clinical guideline and then evaluate their treatment choices using novel ML estimates of treatment effects that incorporate many patient characteristics absent from current guidelines. We also consider a range of policies that would vary the content and implementation of clinical guidelines. We distinguish between changes in *adoption*, the event when a physician becomes aware of a guideline and begins to use it in decision-making, and changes in *adherence*, the degree to which a physician follows guideline recommendations. This distinction separates the extensive margin of guideline take-up (adoption) from the intensive margin (adherence), echoing a distinction made in prior research on technology diffusion (Cohen and Dupas, 2010; Bensch and Peters, 2015). Relative to decision-making without guidelines, adoption may improve patient outcomes if guidelines provide new information without distracting physicians from other relevant factors. Increasing adherence even among adopters may be desirable if physicians continue to depart from guideline recommendations in ways that lead to worse patient outcomes.

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

[2] For example, see Costantini et al. (1999); Woolf et al. (1999); Basu et al. (2014).

We focus on the clinical setting of atrial fibrillation, a common condition afflicting more than 5 million people in the US (Colilla et al., 2013). The principal risk in atrial fibrillation is debilitating or deadly ischemic stroke (hereafter, stroke); untreated patients have a 5% risk of stroke per year (Atrial Fibrillation Investigators, 1995). Anticoagulation (blood thinning) has been shown in clinical trials to reduce stroke risk; however, anticoagulation also increases the risk of life-threatening hemorrhage (hereafter, bleed), including intracranial bleeding (Atrial Fibrillation Investigators, 1995). Therefore, the consequences of misallocating anticoagulation can be serious. In response to these stakes, researchers have developed the $CHADS_2$ score: a simple predictive score of stroke risk for patients with atrial fibrillation (Gage et al., 2001). Clinical guidelines recommend tailoring treatment decisions on the basis of patients' $CHADS_2$ scores (Hirsh et al., 2008). The $CHADS_2$ score is among the most well known and widely used risk scores used for any clinical condition.[3]

We study guideline adoption and treatment decisions of nearly 5,800 physicians treating 113,000 newly diagnosed atrial fibrillation patients in the Veterans Health Administration (VHA) from 2002-2013. For each physician, we measure the date of $CHADS_2$ adoption by the earliest mention of the $CHADS_2$ score in the physician's clinical notes. Following the publication of $CHADS_2$-based treatment guidelines in 2008, we see steady growth in guideline adoption. Prior to adoption, physicians treat roughly 50% of patients with atrial fibrillation, and treatment probability is largely invariant to the $CHADS_2$ score. After the first $CHADS_2$ mention, practice patterns pivot towards $CHADS_2$-based recommendations: prescriptions to patients with low risk scores fall by 4.9 percentage points, while prescriptions to patients with high risk scores increase by 1.6 percentage points. Despite these changes, adopters still fail to adhere to guidelines in more than 40% of cases.

Whether greater adherence is desirable depends on whether deviations from guideline recommendations are responding to information excluded from guidelines in ways that improve patient outcomes. To assess this, we generate novel ML estimates of heterogeneous treatment effects from RCT data, which we then link to chart-level data from the VHA to better understand physician behavior. To estimate heterogeneous treatment effects, we use detailed patient characteristics, clinical outcomes, and randomized treatment status of each patient from eight RCTs in the Atrial Fibrillation Investigators database (hereafter, AFI database) (van Walraven et al., 2009). Using a causal forest model (Wager and Athey, 2017), we obtain treatment effects on strokes and bleeds that vary both

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

with patient characteristics included in the $CHADS_2$ score and with other detailed characteristics. To establish the external validity of our causal forest estimates, we demonstrate that treatment effects estimated on subsets of our trial database predict treatment effects in "out-of-bag" trials not used in estimation. We find that the $CHADS_2$ score explains 25% of the variation in causal forest stroke treatment effects, indicating substantial residual variation.

With ML-estimated treatment effects in hand, we estimate a model of clinician treatment decisions in the VHA data. Our goal is to understand the extent to which treatment deviations from $CHADS_2$ recommendations may be motivated by residual variation in stroke and bleed treatment effects that are not codified in the guideline. Our first finding is that residual variation in stroke treatment effects (i.e., orthogonal to the $CHADS_2$ score) explains little of the variation in treatment decisions. Adoption of the $CHADS_2$ score (proxied by first mention of the score in a doctor's clinical notes) leads doctors to place greater decision weight on $CHADS_2$-related stroke treatment effects and, in some specifications, less weight on bleed treatment effects. We add to the model detailed patient characteristics that may relate to the benefits of treatment despite not being available in the trial data; these include variables that predict patient medication adherence, bleed risk, and fall risk. Conditional on age, these characteristics explain very little of the variance in treatment choices given estimated treatment effects and do not impact the other coefficients in the model.

We then use the model to simulate the impact of broader adoption or stricter adherence to existing and new guidelines. To do so, we consider the frontier of stroke and bleed outcomes which are possible following each decision rule and varying the fraction of patients treated from 0% to 100%. Our findings suggest that physicians allocate treatments to patients with atrial fibrillation only slightly better than a *random* decision rule. Some of this inefficiency is explained by physicians' aversion to treating older patients, as the rate of strokes prevented per bleed induced remains roughly constant as patients age, but doctors are much less likely to treat older patients. Extending $CHADS_2$ adoption (with observed levels of adherence) to non-adopters would modestly improve treatment allocation, preventing an additional 8 strokes per 10,000 patients annually, or a 4% increase in strokes prevented by warfarin.

Our main result is that the benefits of strict guideline adherence are much larger than the benefits of universal adoption. Reallocating the same number of treatments in the observed allocation to patients with the highest $CHADS_2$ scores would prevent 13% more strokes than the status quo for the same number of bleeds. Strict adherence to a guideline which incorporates all information about heterogeneity in stroke treatment effects would prevent 38% more strokes. Reallocating treatments to

patients with the greatest ratio of strokes prevented per bleed induced could prevent 47% more strokes without increasing the number of bleeds. These results suggest that policies that aim to increase adherence at the intensive margin may produce much larger improvements in patient outcomes than policies that only broaden adoption at the extensive margin.

Incorporating more variables can improve guidelines, but only if that information is properly weighted. A subsequent refinement to the $CHADS_2$ score called the CHA_2DS_2-VASc was introduced in 2010, changing the weighting on age and adding vascular disease as a component. While vascular disease does predict stroke treatment effects, the CHA_2DS_2-VASc score gave it too much weight and the new score performs *worse* than the prior version at minimizing strokes per induced bleed. Incorporating more variables that predict strokes need not improve guidelines if those variables are not weighted in proportion to their impact on treatment effects.

Our research relates to several strands of literature. First, we contribute to an active literature in economics studying the potential for machine-based algorithms to improve decision-making. Papers in this literature have typically relied on observational data and quasi-experimental assumptions to reach conclusions about misallocation.[4] A recent econometric literature has pointed out that many quasi-experiments that recover treatment effects averaged across subgroups of data may nevertheless fail to meet the strict assumptions required for identifying treatment effects within each subgroup of data (e.g., Kolesar et al., 2015; de Chaisemartin, 2017; Frandsen et al., 2019). Another challenge in observational data is miscoding in outcome variables. For example, data used for claims reimbursement may imperfectly capture whether a patient admitted to the hospital with a diagnosis of stroke had a new stroke or is being treated for complications of an existing stroke. Our approach exploits data from several large RCTs, where assumptions for internally valid heterogeneous treatment effects are more plausible and where special care is taken to measure critical outcomes of interest for a given clinical condition.

Our findings also relate to recent work highlighting challenges to using evidence from trials to model policy changes. Manski (2017) cautions that we should not privilege the internal validity of trials over external validity in the setting of interest. We address this concern by "clustering" our

[4]This literature spans allocations made by physicians (Abaluck et al., 2016), hospitals (Chandra and Staiger, 2020), judges (Kleinberg et al., 2018), and human resource managers (Hoffman et al., 2018). Many of these papers, cited earlier, compare treatments and outcomes across decision-makers, usually assuming a common ranking of cases across decision-makers (e.g., Abaluck et al., 2016; Kleinberg et al., 2018; Chandra and Staiger, 2020). Newer approaches in Chan et al. (2019) and Arnold et al. (2020) relax this assumption but restrict the direction of treatment effects. In another vein, recent papers have applied ML techniques to predict outcomes using observational data, again with necessary quasi-experimental assumptions because outcomes are selectively observed (e.g., Ribers and Ullrich, 2019; Mullainathan and Obermeyer, 2020).

treatment effect estimates so that the objective is to fit treatment effects in "out-of-bag" trials not used in estimation. We further validate predicted treatment effects in out-of-bag trials with a "best linear predictor" regression, following (Chernozhukov et al., 2018). Einav et al. (2019) and Oster (2020) both report that healthier patients are more likely to take up recommended health screenings and diets. This selection can bias observational estimates and may imply that those treated as a result of guidelines have smaller treatment effects than inframarginal patients. In contrast to these papers, we study guidelines that are primarily targeted at physician treatment decisions, rather than patient-driven health behaviors, and we find no evidence that marginal patients treated due to $CHADS_2$ adoption have smaller (or larger) treatment effects. These differences suggest that the degree of selection on treatment effects may depend on whether physicians or patients are adopting the guideline.

In the medical literature, research has focused on how treatment decisions relate to clinical guidelines. The literature has shown widespread lack of adherence, not only in the case of the $CHADS_2$ score but also across many clinical risk scores and guidelines.[5] Our paper builds on this literature by documenting that even adopting physicians who are aware of the guideline continue to deviate frequently. The crucial difficulty in interpreting non-adherence is the lack of evidence on whether counterfactual treatment decisions promoted by guidelines will improve health outcomes. For example, Mehta et al. (2015) report that physicians who adhere more closely to guidelines have better outcomes but do not separate the impact of guidelines from other differences across physicians. We address this problem by recovering granular heterogeneous treatment effects with ML techniques and RCT data and applying these treatment effects to assess observed and counterfactual treatment behavior.

The remainder of this paper is organized as follows. Section 2 provides clinical background. Section 3 describes our data. Section 4 provides reduced form evidence of the impact of $CHADS_2$ adoption on treatment. Section 5 presents our estimates of causal forest treatment effects. Section 6 models how guideline adoption impacts the relationship between treatment behavior and treatment effects. Section 7 considers counterfactual policies. Section 8 concludes with a discussion of policy implications.

[5] An older review article estimated that 40% of patients were not receiving guideline-recommended care for chronic conditions (Schuster et al., 1998). More recent research suggests non-adherence to guidelines continues to be widespread across a variety of clinical contexts (Lasser et al., 2016; Valle et al., 2015; Chen et al., 2015; Grimshaw and Russell, 1993; Prior et al., 2008; Rosenberg et al., 2015). For the $CHADS_2$ score specifically, Chapman et al. (2017) find evidence of substantial non-adherence to the guideline.

2 Atrial Fibrillation and the $CHADS_2$ Score

Atrial fibrillation is the most common cardiac arrhythmia. It afflicts over 5 million Americans; for adults older than 40 years, one in four will develop the condition (Hsu et al., 2016). Atrial fibrillation increases stroke risk by five-fold and is responsible for 40% of strokes among patients older than 80 years (Piccini and Fonarow, 2016). The main treatment to reduce stroke risk among patients with atrial fibrillation is anticoagulation by warfarin.[6] While anticoagulation is effective in reducing stroke risk, by 68% on average, it has also been shown to increase the risk of major bleeding by more than twofold (Atrial Fibrillation Investigators, 1995; Kearon et al., 2012). Given the large potential benefits and risks of anticoagulation, an important task for clinicians evaluating patients with atrial fibrillation is to decide which patients to treat with anticoagulation.

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

Designed to be easy to use, the $CHADS_2$ score is an index of five patient characteristics: "C" for congestive heart failure (1 point), "H" for hypertension (1 point), "A" for age $\geq$ 75 years (1 point), "D" for diabetes (1 point), and "S" for stroke (2 points) (Table 2). Since its introduction, the $CHADS_2$ score has become one of the most widely recognized risk scores in clinical practice.[7] However, despite its widespread recognition, studies in a variety of settings have shown that adherence to the

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

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

$CHADS_2$ score has been low, typically with only half of recommended patients being prescribed anticoagulation (Hsu et al., 2016; Piccini and Fonarow, 2016).

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

3 Data

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

3.1 Veterans Health Administration Cohort

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

We first identify potentially new diagnoses of atrial fibrillation (ICD9 code beginning with 427.3) among patients with no previous such diagnosis within three years, extending our data back to October 1999 to perform this exclusion. We also require an electrocardiogram (the primary means to diagnose atrial fibrillation) near the time of initial diagnosis and no anticoagulation prior to the initial diagnosis. After a diagnosis of atrial fibrillation, the anticoagulation decision is typically made by a physician who provides longitudinal care and makes prescription decisions for the patient. Therefore,

to attribute patients to physicians who are likely responsible for anticoagulation decisions, we require each patient to have a visit with a VHA cardiologist or primary care physician (PCP) within 90 days after the initial diagnosis (to decide on treatment). Further, the patient must have received at least one drug (other than warfarin) prescribed by the attributed physician within one year before or after the atrial fibrillation diagnosis. We also require each attributed physician to have at least 30 other patients with atrial fibrillation and to have prescribed warfarin for another patient. Our sample restrictions result in an analytic cohort of 113,270 patients (see Table 1 for details).

For each patient in this cohort, we capture a broad array of characteristics that may influence the anticoagulation decision following an initial diagnosis. These characteristics include demographic information, comorbidities, laboratory test results, body measurements, and blood pressure readings. We use these characteristics to construct the $CHADS_2$ score, match the other clinical characteristics recorded in the RCT data, and to proxy for other concerns not fully captured in the RCT data (e.g., bleed risk, fall risk, frailty, multi-morbidity).[8] To capture the anticoagulation decision, we rely on VHA records of prescriptions for warfarin or a newer oral anticoagulants (e.g., dabigatran, rivaroxaban, apixaban, edoxaban).[9]

Finally, we measure adoption of the $CHADS_2$ score at the physician level by searching physician visit notes for mention of the $CHADS_2$ score.[10] We consider physicians who have mentioned the $CHADS_2$ score in a note as "adopters" of the $CHADS_2$ score guideline, in the sense that they recognize the $CHADS_2$ score and have used it in their decision-making. We proxy a physician's time of $CHADS_2$ adoption by the physician's first note mentioning the $CHADS_2$ score. We categorize physicians without any note mentioning the $CHADS_2$ score as non-adopters.

Table 3 reports summary statistics for adopters and non-adopters. Overall, 50% of patients in our sample receive anticoagulation. Patient characteristics are similar for adopters and non-adopters. While roughly half of the physicians in our sample are adopters, adopters tend to have more patients

[8]For further details on the additional patient comorbidities and risk factors we extracted from the VHA sample, see Appendix Section A.4.

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

[10]Recall that physicians in our analytic sample are cardiologists and PCPs who have each treated at least 30 atrial fibrillation patients. We increase our detection of $CHADS_2$ mentions for these physicians by using visit notes within 6 months of initial diagnosis in our broad cohort of 844,312 atrial fibrillation patients, who may be patients with previously established diagnoses (see Table 1), and search for non-case-sensitive occurrences of the string `chads`. We settled on this string after spot-checking several variants for false positives. Consistent with our spot-checking results, we find no positive mentions of this string in the first two years of our data, prior to diffusion of the $CHADS_2$ score. This suggests that false positives are very rare.

in our sample than do non-adopters. Patients seen by adopters and non-adopters are similar in age and have similar prevalence rates of $CHADS_2$ risk factors.

3.2 Atrial Fibrillation Investigators Database

To estimate heterogeneous treatment effects—which we use to assess physician decision-making and to evaluate counterfactual anticoagulation decisions on patient outcomes—we rely on the Atrial Fibrillation Investigators database (hereafter, AFI database). The AFI database contains patient-level observations from eight trials in which patients were randomized to anticoagulants versus a placebo or control.[11] Details of the AFI database have been documented elsewhere (e.g., van Walraven et al., 2009).

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

4 $CHADS_2$ Guideline Adoption and Adherence

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

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

[12] In the 2008 guidelines, patients with a $CHADS_2$ score of 0 were recommended aspirin, while patients with a $CHADS_2$ score of 1 could be treated with either aspirin or anticoagulation. Later guidelines suggested anticoagulation for patients with a $CHADS_2$ score of 1 and some patients with a $CHADS_2$ score of 0 but a CHA_2DS_2-VASc score of 1 (Lip et al.,

for groups of patients defined by their $CHADS_2$ score.

Figure 1 displays trends in anticoagulation rates for patients with low risk of stroke ($CHADS_2$ score of 0 or 1), patients with moderate risk of stroke ($CHADS_2$ score of 2 or 3), and patients with high risk of stroke ($CHADS_2$ score of 4 or greater). Prior to the 2008 ACCP guideline, patients with lower $CHADS_2$ scores were remarkably *more* likely to be treated with anticoagulation. This relationship held both between patients with low vs. moderate stroke risk and between patients with moderate vs. high stroke risk. After 2008, we find a reduction in anticoagulation rates for low-risk patients for whom the guideline allowed for management without anticoagulation. There appears to be a small increase in treatment rates for patients at moderate or high risk. However, even among groups where anticoagulation is recommended, prescription rates remain below 55% for our sample period. Patients with high stroke risk ($CHADS_2$ score of 4 or greater) remain slightly less likely to be treated than patients with moderate stroke risk ($CHADS_2$ of 2 or 3).

It is evident from Figure 1 that physicians had only weak adherence to recommendations to treat patients with high $CHADS_2$ risk scores and to leave patients with low $CHADS_2$ risk scores untreated. We next examine when physicians mention the $CHADS_2$ score in their clinical notes. We interpret these mentions of the risk score as a measure of adoption on the extensive margin, in the sense that physicians recognized the score as a tool to assist in the evaluation of stroke risk and the anticoagulation decision. In Figure 2, we show that almost no physician mentioned the $CHADS_2$ score in the period prior to the ACCP guideline in 2008, despite the fact that the $CHADS_2$ score was introduced in 2001 and validated in 2004. Following 2008, we find a steady rise in the proportion of physicians who have previously mentioned the $CHADS_2$ score at least once, approaching 70% near the end of our study period in 2013. We note that this represents a lower bound to awareness of the $CHADS_2$ score, as physicians may be aware of the score yet not mention it in their notes.

We proceed by estimating the causal impact of guideline adoption on prescription choice and investigating how guideline adherence changes after a physician incorporates the guideline into her clinical practice. In the Panel A of Figure 3, we first plot how anticoagulation rates among high- and low-score patients change following an adopting physician's first note mentioning the $CHADS_2$ score. Following the adoption event, the anticoagulation rate for low-scoring patients drops by several percentage points, while the anticoagulation rate for higher-scoring patients increases slightly.

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

2010). This clinical consensus applied mostly to after our study period.

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

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

Taken together, these results suggest that although the $CHADS_2$ score was becoming widely adopted, adherence to anticoagulation recommendations increased only modestly with adoption. Prior to the 2008 guideline, almost no physician appeared to be using the $CHADS_2$ score in documented clinical decision-making, yet by the end of 2013, the vast majority had explicitly mentioned it in their notes. While our event-study results suggest a clear behavioral shift in prescribing at the time of adoption, most of this response is from avoiding treatment for low-risk patients, not increasing treatment for high-risk patients. In Figure A.1, we further show that, while adherence varied substantively across physicians, few physicians reached an adherence rate of 80%.

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

These findings are consistent with a clinical literature that emphasizes the importance of reducing stroke risk by anticoagulation and that documents widespread awareness of the $CHADS_2$ score

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

among physicians, now a decade after the 2008 ACCP guideline (Ashburner et al., 2018; Amroze et al., 2019). Yet in numerous settings, only about half of patients with the highest stroke risks are treated with anticoagulation (Hsu et al., 2016). Our results further show that there is low adherence even among physicians who discuss the $CHADS_2$ score in their decision-making. Survey evidence suggests that physicians remain hesitant to anticoagulate patients who are elderly, frail, and multimorbid, out of concern that anticoagulation may result in severe bleeding for these patients (Fawzy et al., 2019). We find descriptive evidence to support this reason behind nonadherence: Appendix Figure A.2 shows that treatment rates actually *decline* with age after 75 years, despite age being a strong predictor of stroke risk.

5 The Effects of Anticoagulation

The results in the previous section show a gap between extensive-margin $CHADS_2$ guideline adoption and full adherence to the $CHADS_2$-based recommendations. Our interpretation of this gap, as well as policy recommendations for incorporating evidence from clinical trials into practice, will depend on the relative value of physician discretion versus guideline adherence for patient health outcomes.

In the case of atrial fibrillation, physicians may depart from $CHADS_2$-based recommendations for good reasons. First, the $CHADS_2$ score predicts stroke risk, while treatment decisions should be based on stroke treatment effects. Second, the decision to anticoagulate must weigh the reduction of stroke risk against the inducement of bleed risk. The $CHADS_2$ score is silent on the latter; while risk scores for bleeding have subsequently been developed (Pisters et al., 2010), their adoption has been much lower than the $CHADS_2$ score. Third, clinicians have access to more patient information than has been encoded in simple risk scores. Tailoring their treatment decisions to this additional information could lead to departures from $CHADS_2$-based guidelines.

5.1 Setup and Design

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

others that are omitted from it. Specifically, for one-year stroke and bleeding, respectively,

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

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

where X_i is a set characteristics belonging to patient i.

We set about estimating these objects by applying "causal forest" ML techniques, as described in Athey and Wager (2019), to RCT-generated data in the AFI database (van Walraven et al., 2002). The experimental design of RCTs is well-suited for this application, in contrast to many quasi-experimental designs commonly used in the economics literature. First, random assignment of treatment within each cell of patient characteristics x, a crucial requirement to estimate CATEs, is more plausible in RCTs. Second, many quasi-experiments involve monotonicity and exclusion-restriction violations within cells of the data, even if these violations "average out" in the entire sample (e.g., Kolesar et al., 2015; Frandsen et al., 2019). Third, RCTs carefully collect information on infrequent yet important outcomes (e.g., stroke and bleeding) for the disease and treatment being studied. Capturing these events, particularly their timing relative to initiation of treatment, may be challenging in observational data.

5.2 Causal Forest Implementation

We use the algorithm developed in Athey and Wager (2019) for estimating causal forests with conditional random assignment, supplemented by a LASSO variable selection step. We find that the latter improves performance, especially for bleeds (the rarer of the two outcomes in our data). The LASSO with a cross-validated penalty parameter selects the variables shown in Appendix Table A.2 (column 2 for stroke and column 4 for bleed) from the full set of variables in Appendix Table A.1 (as well as the $CHADS_2$ score itself, which is a function of these variables). For bleeds, this procedure leaves just two variables which strongly predict bleed risk: age and race. For strokes, 10 variables are selected. For the remainder of the algorithm, we proceed with the LASSO-selected variables.

We then use regression forests to adjust outcomes and treatment status with expectations based on patient characteristics x and the RCT trial identifier j:

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

$$W(x;j) = E[W_i|X_i = x, j(i) = j] \tag{5}$$

where $j(i)$ indicates the RCT trial for individual i, and where $Y^o(x;j)$ is formed using observations in the control arm of each trial. We compute "out of bag" estimates $\hat{Y}^o_{-i}(x;j)$ and $\hat{W}_{-i}(x;j)$ from data excluding i. In our causal forest estimation, we use re-centered outcome and treatment variables: $\tilde{Y}^s_i = Y_i - \hat{Y}^o_{-i}(X_i; j(i))$ and $\tilde{W}_i = W_i - \hat{W}_{-i}(X_i; j(i))$. The causal forest generates predictions of stroke and bleeding CATEs, $\tau^s(x)$ and $\tau^b(x)$ respectively, capturing heterogeneity in treatment effects along patient characteristics observable in the AFI database.[14] In addition to the variables selected by the LASSO, we include the associated risk score for each outcome, $\hat{Y}^o_{-i}(X_i; j(i))$, in the causal forest.

To improve the prediction of treatment effects out of sample (i.e., in the VHA data), we use trial identifiers in the AFI database to implement a "cluster-robust" causal forest with clusters at the trial level (Athey and Wager, 2019). This means that "out-of-bag" estimates used in training the forest are computed based on separate trials from the original observation. This procedure constructs estimates which are externally valid across trials within the AFI database.[15]

To provide insight into sources of treatment effect heterogeneity, we summarize the "variable importance" of patient characteristics in predicting treatment effect heterogeneity.[16] Appendix Table A.2 lists variables ranked by importance in the stroke and bleed causal forests, as well as regression forests estimated to predict risk in the control groups. We also report the sign of each variable in a linear regression of treatment effects on all included variables. The single most important predictor of treatment effects in both the stroke and bleed models is the corresponding regression-forest model of risk. The variables in the $CHADS_2$ score generally rank highly in both the causal forest and regression forest models, but several variables not in the $CHADS_2$ score also matter: e.g. lower hemoglobin and history of angina or myocardial infarction predict larger stroke treatment effects.

5.3 Validation and Best Linear Predictors

In this section, we validate the estimated heterogeneity to assess concerns about potential over-fitting both within and across trials. Following Athey and Wager (2019), we project outcomes onto leave-

[14] We estimate treatment effects as a function of x, exclusive of trial indicators, since we will be applying our predictions to an external sample (i.e., the VHA data). The 'centering' procedure described in the text is formally justified in Section 6.1.1 of Athey et al. (2019).

[15] We cross-validate all parameters in the causal forest except for sample fraction and minimum node size. The sample fraction can get too low for estimation to proceed properly when being tuned. Therefore we set the sample fraction to its default value 0.5. We set the minimum node size to 100 to avoid overfitting that occurs at node sizes that are too small. When selected via cross-validation, the causal forest prefers full-grown trees with arbitrarily small node sizes but this specification leads treatment effects to vary implausibly with small changes in continuous variables such as age. These irregularities disappear at a minimum node size of 100.

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

out-trial CATE predictions of treatment effects. This exercise also allows us to construct a "best linear predictor" (BLP) of treatment effects given the causal forest estimates (Chernozhukov et al., 2018).[17] Specifically, for observations in each trial in the AFI database, we make out-of-bag CATE predictions from causal forests grown exclusively from data in the *other* trials. Denote these leave-out-trial CATE predictions for stroke and bleeding as $\hat{\tau}^o_{-j(i)}(x)$, $o \in \{s, b\}$, for individual i in trial $j(i)$ with characteristics $X_i = x$ (we hereafter suppress the i in $j(i)$ to simplify notation). Using regression forests, we also construct and potentially control for predictions of $Y^o(x)$, based on other trials: $\hat{Y}^o_{-j,1}(x)$, using only control group data, and $\hat{Y}^o_{-j,2}(x)$, using all data but not conditioning on treatment status.

We assess external validity across trials by a regression of realized outcomes Y^o_i on out-of-bag predictions and trial fixed effects ζ^o_j:[18]

$$Y^o_i = \left(\delta^o_0 + \delta^o_1 \hat{\tau}^o_{-j}(X_i)\right) W_i + \gamma^o_1 \hat{Y}^o_{-j,1}(X_i) + \gamma^o_2 \hat{Y}^o_{-j,2}(X_i) + \zeta^o_j + \varepsilon_i. \tag{6}$$

The coefficient δ^o_1 quantifies the predictive power of heterogeneous CATEs that are estimated in other trials; a coefficient value of $\delta^o_1 = 1$ would suggest that the causal forest estimates are well calibrated and outcomes for a patient with characteristics x would increase one-for-one with treatment and the relevant $\hat{\tau}^o_{-j}(x)$. Using Equation (6), we can further construct, for later use, BLP-based adjustments to the CATEs estimated in Section 5.2:

$$\hat{\tau}^o_{BLP}(x) = \hat{\delta}^o_0 + \hat{\delta}^o_1 \hat{\tau}^o(x). \tag{7}$$

We will use this adjustment directly for CATEs on bleeding.

For strokes, we consider the exact analog of this regression as well as a modified procedure where we decompose treatment effects into two dimensions: stroke treatment effects that vary with the $CHADS_2$ score, and stroke treatment effects that are orthogonal to the $CHADS_2$ score. Specifically, we first use a regression to project leave-out-trial stroke CATEs, $\hat{\tau}^s_{-j}(x)$, onto indicators of the $CHADS_2$ score. We call the $CHADS_2$-projected component $\hat{\tau}^{s(c)}_{-j}(x)$ and the residual component

[17] The regression is an OLS version of the BLP procedure. Chernozhukov et al. (2018) suggest a GLS version that has more power in the general case when treatment probabilities vary across trials.

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

$\hat{\tau}^{s(r)}_{-j}(x)$, noting that $\hat{\tau}^{s}_{-j}(x) = \hat{\tau}^{s(c)}_{-j}(x) + \hat{\tau}^{s(r)}_{-j}(x)$. We next perform the following BLP projection:

$$\begin{aligned} Y^s_i &= \left(\delta^s_0 + \delta^{s(c)}_1 \hat{\tau}^{s(c)}_{-j}(X_i) + \delta^{s(r)}_1 \hat{\tau}^{s(r)}_{-j}(X_i)\right) W_i + \\ &\quad \gamma^o_1 \hat{Y}^o_{-j,1}(X_i) + \gamma^o_2 \hat{Y}^o_{-j,2}(X_i) + \zeta^o_j + \varepsilon_i. \end{aligned} \tag{8}$$

Using equation (8), we construct BLP-based adjustments to the stroke CATEs as follows:

$$\hat{\tau}^s_{BLP}(x) = \hat{\delta}^s_0 + \hat{\delta}^{s(c)}_1 \hat{\tau}^{s(c)}(x) + \hat{\delta}^{s(r)}_1 \hat{\tau}^{s(r)}(x). \tag{9}$$

The BLP estimation validates the performance of our causal forest procedure on held-out trials. Results of the BLP procedure are reported in Appendix Table A.4. All of our BLP coefficients on the recovered causal forest treatment effects are significantly different from 0, and we cannot reject the null that each coefficient equals 1. For stroke CATEs, we estimate $\hat{\delta}^s_1 = 1.027$ (standard error of 0.242). For bleed CATEs, we estimate $\hat{\delta}^b_1 = 1.149$ (standard error of 0.324). For the $CHADS_2$ component, we estimate $\hat{\delta}^{s(c)}_1 = 0.763$ (standard error of 0.340); for the residual component, we estimate $\hat{\delta}^{s(r)}_1 = 1.302$ (standard error of 0.348). In other words, both components predict treatment effect variation in held-out trials, although the evidence is even stronger for the residual component.

In our subsequent analyses, we will adjust CATEs predicted in Section 5.2 for stroke and bleeding by using Equation (9) for $\hat{\tau}^s_{BLP}$ and Equation (7) for $\hat{\tau}^b_{BLP}$. In Appendix Table A.9, we replicate our main results using unadjusted $\hat{\tau}^s$ and $\hat{\tau}^b$ as robustness checks and nothing changes qualitatively.

5.4 External Validity in the VHA Data

While clustering at the trial level should yield predictions that are externally valid across trials within the AFI database, we ultimately seek to use treatment effects estimated in the AFI database to evaluate counterfactuals in the VHA data. In order to assess the external validity of CATEs in the VHA data, we compare observable attributes of patients in the AFI database to those in the VHA data to assess whether, at least with respect to observable characteristics, the VHA data lies roughly within the support of the AFI trials.

Table 3 compares the VHA data with the AFI database on some key patient characteristics. The clearest difference in average patient characteristics is in the share of male patients. We estimate CATEs for both male and female patients from the AFI database, allowing the causal forest to use gender to predict treatment effect heterogeneity. Our analysis of variable importance, shown in Ap-

pendix Table A.2, finds that patient gender is unimportant in predicting CATEs. Among other patient characteristics, the AFI database has a larger share of the population over 65, a lower incidence of hypertension and diabetes, and a higher rate of congestive heart failure than the VHA data on average. However, in Appendix Table A.5, we can see that for all patient characteristics (including gender), the VHA data mean is within the range of trial means. This mitigates concerns that the types of patients seen in the VHA are not represented in the AFI database.

5.5 Treatment Effect Predictions for VHA Patients

We take causal forest prediction rules trained and validated in the AFI database, in the form of causal forest splits and leaf values, and apply these rules to patients in the VHA data. For the remainder of the paper, we use BLP-adjusted CATEs, with weights defined by our validation exercise in Section 5.3, Equations (7) and (9). Figure 4 shows variability in the distribution of estimated stroke and bleed (BLP-adjusted) CATEs. The 10th percentile stroke treatment effect (corresponding to the largest reductions in stroke risk) is -0.083, while the 90th percentile is -0.017. The impact of warfarin on bleeds is smaller but still variable, ranging from 0.010 at the 10th percentile (smallest increases in blled risk) to 0.038 at the 90th percentile.

We next assess the importance of characteristics omitted from the $CHADS_2$ score in predicting treatment effects in the VHA population. While stroke treatment effects and the $CHADS_2$ score are highly correlated, we find substantial residual variation in stroke treatment effects after conditioning on the $CHADS_2$ score. The R^2 from regressing BLP-adjusted stroke treatment effects, $\hat{\tau}^s_{BLP}$ on $CHADS_2$ score indicators is 0.25. Figure 5 shows that stroke treatment effects increase monotonically with the $CHADS_2$ score, but that wide variation also exists in the residual component of stroke treatment effects. While patients with higher $CHADS_2$ scores also have larger anticoagulation inducements in bleeds, there also is variation in bleed treatment effects within each $CHADS_2$ score. This suggests scope for using more detailed stroke and bleed CATEs in treatment decisions, which we will evaluate in the next section.

6 Assessing Physician Decisions in the VHA

With ML predictions of treatment effects in hand, we turn to assess how physician treatment decisions in the VHA relate to treatment effects. We introduce a model to characterize how treatment decisions respond to treatment effect variation, separately considering variation that is and is not captured by

guidelines. This model allows us consider the relative impact of adoption and adherence on patient outcomes. In particular, we will evaluate whether guidelines may lead physicians to neglect information relevant for health outcomes but not incorporated into a guideline. Finally, we use the model to simulate the counterfactual impact of adopting guidelines that incorporate more information about how treatment effects vary across patients.

6.1 Stylized Model of Treatment Decisions

We consider decisions made by physicians as a function of treatment effects and their guideline adoption status. Guideline adoption status is denoted $g \in \{\text{never}, \text{pre}, \text{post}\}$, corresponding to decisions made by physicians who never adopt the CHADS_2 score, decisions made by physicians before adopting the CHADS_2 score, and decisions made by physicians after adopting the CHADS_2 score. Anticoagulation decisions W_i in state g are made as follows:

$$W_i = \mathbf{1}\left\{\beta^s \tilde{\tau}^s_{i,g} + \beta^b \tilde{\tau}^b_{i,g} + f(X_i) + v_i > 0\right\}. \tag{10}$$

This model includes two components. First, physicians consider Bayesian posterior beliefs of patient-specific stroke and bleeding treatment effects, $\tilde{\tau}^s_{i,g}$ and $\tilde{\tau}^b_{i,g}$, with respective preference weights β^s and β^b. Guideline adoption may improve physicians' information about treatment effects, giving them more accurate posterior beliefs. The second component consists of other factors—including both observable characteristics (to the econometrician) $f(X_i)$ and otherwise unobservable factors v_i—which impact treatment decisions.

In Appendix A.1, we specify a Bayesian model in which posterior beliefs about treatment effects result from a prior and noisy signals. The model implies that posterior beliefs are a linear function of true treatment effects and prior beliefs:[19]

$$\tilde{\tau}^s_{i,g} = \lambda^{s(c)}_g \tau^{s(c)}_i + \lambda^{s(r)}_g \tau^{s(r)}_i + \mu^s_g + \upsilon^s_{i,g}; \tag{11}$$

$$\tilde{\tau}^b_{i,g} = \lambda^b_g \tau^b_i + \mu^b_g + \upsilon^b_{i,g}. \tag{12}$$

In the equations above, μ^s_g and μ^b_g are constants within g (which depend on physician's priors), and $\upsilon^s_{i,g}$ and $\upsilon^b_{i,g}$ are noise terms with variances that depend on the precision of the signals that physicians

[19] This linear projection can be exactly microfounded by a standard Bayesian model with normal true treatment effects and normal noise, which we detail in Appendix A.1. However, absent a joint-normal model of signals and noise, Equations (11) and (12) are linear approximations of Bayesian updating, common in empirical Bayes applications (e.g., Chetty et al., 2014).

receive. True stroke treatment effects, τ_i^s, are decomposed into a $CHADS_2$-related component $\tau_i^{s(c)}$ and a residual component $\tau_i^{s(r)}$, such that $\tau_i^s = \tau_i^{s(c)} + \tau_i^{s(r)}$. τ_i^b is the true treatment effect for bleeding. The parameters $\lambda^{\tilde{o}}$, $\tilde{o} \in \{s(c), s(r), b\}$, correspond to the signal-to-noise ratio of posterior beliefs $\tilde{\tau}_{i,g}^s$ and $\tilde{\tau}_{i,g}^b$ with respect to $\tau_i^{s(c)}$, $\tau_i^{s(r)}$, and τ_i^b. Physicians may have more precise signals for some treatment effects types $\tilde{o}$ than others, and the precision of their signals may change with guideline adoption status g.

In the model, $CHADS_2$ adoption increases the precision of the doctor's signal of $\tau_i^{s(c)}$. The model also allows for the possibility of "distraction effects," whereby guideline adoption leads physicians to neglect consideration of other decision-relevant factors (i.e. the "cookbook medicine" critique). In the language of the model, distraction effects correspond to physicians forming less precise beliefs about $\tau_i^{s(r)}$ and τ_i^b.

6.2 Estimating Equation

To estimate policy-relevant parameters from the data, we now map our behavioral model in Equation (10) to a heteroskedastic probit model:

$$W_i = \mathbf{1}\left\{\frac{1}{\sigma_{\varepsilon,g}}\left(\alpha_g^{s(c)}\tau^{s(c)}(X_i) + \alpha_g^{s(r)}\tau^{s(r)}(X_i) + \alpha_g^b\tau^b(X_i) + \mu_g + f(X_i) + \varepsilon_{i,g}\right) > 0\right\}. \tag{13}$$

where $\varepsilon_{i,g}$ is a normally distributed error with variance $\sigma_{\varepsilon,g}^2$.

From Equations (10) to (12), the coefficient $\alpha_g^{\tilde{o}}$, for treatment effect $\tilde{o} \in \{s(c), s(r), b\}$ measures the responsiveness of decisions to a treatment effect and can be interpreted as $\alpha_g^{\tilde{o}} = \beta^s \lambda_g^{\tilde{o}}$. $\varepsilon_{i,g}$ includes v_i from Equation (10) and the "noise" components of posterior beliefs, $\upsilon_{i,g}^s$ and $\upsilon_{i,g}^b$. In estimation, we normalize $\sigma_{\varepsilon,\text{pre}}^2 = 1$. In our baseline specification, we allow $f(X_i)$ to depend on age, to account for a bias against treating older patients, and on year fixed effects, to capture secular shifts in anticoagulation over time.

Guideline adoption may change the responsiveness to treatment effects in three ways. First, $CHADS_2$ adoption may give physicians more precise information about $CHADS_2$-related variation in stroke treatment effects; this effect increases the responsiveness to treatment effects $\alpha^{s(c)}$ by increasing $\lambda^{s(c)}$. Second, because noisy assessment of treatment effects are incorporated into $\varepsilon_{i,g}$, $\sigma_{\varepsilon,g}^2$ may change with guideline adoption.[20] Finally, $CHADS_2$ adoption may distract physicians from consid-

[20] If guideline adoption sufficiently increases the precision of physician signals, then $\sigma_{\varepsilon,g}^2$ will decrease. However, because $\varepsilon_{i,g}$ includes both noisy assessments and the weight that physicians' place on them, it is possible that $\varepsilon_{i,g}$ may increase if better information causes physicians to place more weight on their signals, including the noisy component of

ering information about other treatment effects, decreasing $\lambda^{s(r)}$ or λ^{b} and thereby leading to smaller $\alpha^{s(r)}$ or α^{b}, respectively.

To recover and interpret these objects, we require the following assumptions, which we will assess in our results below. First, to recover unbiased (and correctly signed) estimates of $\alpha_g^{\tilde{o}}$, we require that $\varepsilon_{i,g}$ is uncorrelated with $\tau^{\tilde{o}}(X_i)$. Second, in order to interpret deviations from treating according to treatment effects as worsening strokes or bleeds, we assume that v_i does not include treatment effects that we do not measure. Finally, in order to identify changes in σ_{ε_g} in our heteroskedastic probit model, we assume that $f(X_i)$ is invariant to guideline adoption. In our baseline specification, we will use age in $f(X_i)$ for this purpose, which assumes that CHADS_2 adoption changes treatment decisions with respect to age only through changes in sensitivity to treatment effects.

In estimating Equation (13), we use ML predictions of treatment effects, $\hat{\tau}_{BLP}^{s(c)}(x)$, $\hat{\tau}_{BLP}^{s(r)}(x)$, and $\hat{\tau}_{BLP}^{b}(x)$, in place of true treatment effects, $\tau^{s(c)}(x)$, $\tau^{s(r)}(x)$, and $\tau^{b}(x)$. These estimates are measured with error in the sense that they differ from the treatment effects we could obtain with infinite data. Our main concern is that differential measurement error of $\hat{\tau}^{s(c)}(x)$ and $\hat{\tau}^{s(r)}(x)$ may differentially attenuate $\alpha_g^{s(c)}$ and $\alpha_g^{s(r)}$. However, our BLP-adjustment in Section 5.3 of CHADS_2-related and residual stroke treatment effects provides a means to ensure that $\alpha_g^{s(c)}$ and $\alpha_g^{s(r)}$ can be interpreted on the same scale. This approach follows a "regression calibration" literature that addresses measurement error (George and Foster, 2000) and also resembles the first stage of "split-sample" instrumental variables approaches in economics (Angrist and Krueger, 1995).

6.3 Results

Table 4 reports estimates of $\alpha_g^{\tilde{o}}$, for treatment effect type $\tilde{o} \in \{s(c), s(r), b\}$ and for each group of physician decisions $g \in \{\text{never}, \text{pre}, \text{post}\}$, from Equation (13). To facilitate statistical comparisons of $\alpha_g^{\tilde{o}}$ across adoption states, the estimation takes $\alpha_{\text{pre}}^{\tilde{o}}$ as the baseline, and uses interaction terms to report differences between $\alpha_{\text{post}}^{\tilde{o}}$ or $\alpha_{\text{never}}^{\tilde{o}}$ and this baseline. Since both strokes and bleeding events are undesirable, we expect the sign of coefficients on both stroke and bleeding treatment effects to be negative. In other words, all else equal, physicians should be more likely to treat patients with larger reductions in stroke risk and less likely to treat patients with larger increases in bleeding risk. Our baseline specification controls for year fixed effects and cubic splines in patient age. Column 1 shows these results. Columns 2 and 3 show alternative specifications controlling for differential time trends in the sensitivity to treatment effects, in order to account for secular trends in overall anticoagulation

them. This point follows formally from the microfoundation in Appendix A.1.

and awareness of treatment effects.

For physicians who have not adopted the $CHADS_2$ score (i.e., $g \in \{\text{never}, \text{pre}\}$), anticoagulation decisions are already sensitive to the $CHADS_2$ component of stroke treatment effects and to bleed treatment effects, but they are not sensitive to the residual component of stroke treatment effects. In Appendix Figure A.3, we show that, conditional on $CHADS_2$ score, treated patients have similar stroke treatment effects before and after guideline adoption. This again suggests that physicians are not selecting on factors that predict stroke treatment effects conditional on $CHADS_2$ score.[21] Among decisions without $CHADS_2$ adoption, there are no substantive differences in the responsiveness to treatment effects between physicians who never adopt (i.e., $g = \text{never}$) and adopters who have not yet adopted (i.e., $g = \text{pre}$). This finding suggests that in the absence of the guideline, eventual adopters are neither lower nor higher skilled than non-adopters at matching patients to appropriate treatment.

Following adoption, physicians become at least twice as sensitive to $CHADS_2$-related stroke treatment effects, suggesting guideline adoption improves physicians' information about this variation in treatment benefits. Further, physicians remain insensitive to residual stroke treatment effects. We find some evidence that physicians become less sensitive to bleed treatment effects following adoption of the $CHADS_2$ score, although this effect is not statistically significant when controlling more richly for time trends in Column 3. In addition, $\sigma_{\varepsilon,g}$ changes little with adoption, despite the change in $\alpha_g^{s(c)}$ (see Appendix Table A.6).

We next consider the average marginal effects implied by our model estimates. As shown in Appendix Table A.7, a one-percentage-point reduction in $CHADS_2$-related yearly treatment effects leads to a 1.9 percentage points increase in treatment probability prior to $CHADS_2$ adoption; this effect grows to a 4.3 percentage points increase in treatment probability after $CHADS_2$ adoption (a 2.4 percentage point difference). On average, low $CHADS_2$ patients have stroke treatment effects that are 2 percentage points larger in magnitude. After $CHADS_2$ adoption, we thus expect relative treatment probabilities for low $CHADS_2$ patients to decline by 4.8 percentage points relative to high $CHADS_2$ patients. The magnitude of these marginal effects are consistent with our reduced-form results in Section 4.

[21] This result contrasts with Einav et al. (2019) and Oster (2020), who find that healthier patients select into treatment. In our setting, where physicians are the key decision-maker, patterns of selection might be different.

6.4 Interpretation and Robustness

Under the lens of our model, these findings suggest that the $CHADS_2$ score improves physicians' information on $CHADS_2$-related stroke treatment effects. Specifically, we find that $CHADS_2$ adoption doubles the precision of physicians' signal of $CHADS_2$-related treatment effects, from $\lambda^{s(c)}_{\text{pre}}$ to $\lambda^{s(c)}_{\text{post}}$. In contrast, physicians appear to have little systematic knowledge of residual stroke treatment effects, and adopting the $CHADS_2$ score has no impact on this. The precision of physicians' information about $CHADS_2$-related stroke treatment effects is much greater than the precision of their information about other sources of treatment effect variation, i.e. $\lambda^{s(c)}_g \gg \lambda^{s(r)}_g$, regardless of g. The negligible implied $\lambda^{s(r)}_g$ suggests that it is also unlikely that physicians are responding to variation in treatment effects that depends on characteristics unmeasured in the AFI database; doctors' information about stroke treatment effect variation appears to be limited to those factors directly modeled in the $CHADS_2$ score.

This evidence suggests that guideline adoption may provide important informational benefits. However, three factors highlight the potential limits of adoption. First, we find some evidence of distraction effects which may limit the degree to which $CHADS_2$ adoption reduces strokes per induced bleed. Second, although we see a large relative increase in physicians' responsiveness to $CHADS_2$-related treatment effects, we see no significant reduction in $\sigma_{\varepsilon,g}$ with $CHADS_2$ adoption. This result suggests that uncertainty about $CHADS_2$-related treatment effects is responsible for only a small share of variation in treatment decisions. Third, even with adoption of the $CHADS_2$ score, physicians continue to be less inclined to treat older patients (who also have higher $CHADS_2$ scores), and this behavior cannot be explained by treatment effects.

Before we take our results to counterfactual simulations, we can use the rich set of covariates in the VHA data to assess some of our key assumptions. In Appendix A.4, we describe several patient characteristics that may influence anticoagulation decisions via $f(X_i)$; these include additional comorbidities not in the AFI database, as well as variables that may be related to the HAS-BLED score, frailty and fall risk, patients' ability to adhere with dosage monitoring, and physician specialty codes. In Panel A of Figure A.4, we show that our estimates of $\lambda^{s(c)}_{\text{pre}}/\lambda^{s(c)}_{\text{post}}$ do not vary much as we progressively add these additional covariates to the model. We can also use these additional covariates to test whether departures from perfect adherence are well explained by variables that might impact treatment effects but are not included in the AFI database. In Panel B of Figure A.4, we show that these additional variables do not systematically explain a large fraction of treatment decisions and therefore

do not explain the high rates non-adherence.[22] Even as we control for detailed patient characteristics, we find little increase in the explained share of variance in treatment decisions. Combined with the important fact that physicians do not even respond to $\hat{\tau}^{s(r)}_{BLP}$, this suggests that departures from treating according to measured treatment effects is unlikely to be explained by unmeasured variation in treatment effects but could rather represent practice style variation across physicians or idiosyncratic decision-making within each physician.[23]

7 Counterfactual Adoption and Adherence

Based on our ML-predicted treatment effects in Section 5 and our analysis of physician decision-making in Section 6, we simulate outcomes under counterfactual scenarios of guideline adoption and adherence. When discussing counterfactual outcomes, it is useful to compare outcomes to a few benchmarks. Treating *all* patients with newly diagnosed atrial fibrillation in the VHA would prevent 400 strokes (hereafter, "preventable strokes") and induce 277 bleeding events (hereafter, "inducible bleeds") per 10,000 patients after one year.

In Figure 6 and Table 5, we show key results on prevented strokes and induced bleeds under counterfactual scenarios. We first show that status quo physician decisions are only slightly better than *random* anticoagulation of atrial fibrillation patients: physicians prescribe anticoagulation to 49.8% of patients and prevent 49.8% of preventable strokes but induce 48.4% of inducible bleeds. That is, status quo decisions prevent only $\frac{49.8\%}{48.4\%} - 1 \approx 3\%$ more strokes per bleed than random decisions with the same induced bleeds. While we found in Section 6.3 that physicians are sensitive to treatment effects, the benefits of this sensitivity are offset by the tendency not to treat older patients who often have desirable ratios of stroke to bleed treatment effects. We will later revisit this issue with analyses that reallocate anticoagulation only within age bins.

Next, we consider counterfactual outcomes under scenarios varying the extensive margin of guideline adoption. Adoption of the $CHADS_2$ score had relatively muted effects on outcomes. Under the counterfactual scenario of no $CHADS_2$ adoption, 49.7% of patients would be treated, preventing

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

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

49.2% of preventable strokes and inducing 48% of inducible bleeds (preventing 2.7% more strokes per bleed than a random allocation). Universal $CHADS_2$ adoption increases treatment to 50.2% of patients, preventing 51.3% of preventable strokes and inducing 49.5% of inducible bleeds (preventing 3.7% more strokes per bleed than random). We can also use the model to simulate adoption of a more comprehensive guideline revealing all information about stroke treatment effects. We assume that physicians who have adopted this guideline will have equal information on $\hat{\tau}^{s(c)}(x)$ and $\hat{\tau}^{s(r)}(x)$ (that is, in our counterfactual, $\lambda^{s(r)} = \lambda^{s(c)}_{\text{post}}$), so that the coefficient $\alpha^{s(r)} = \alpha^{s(c)}_{\text{post}}$.[24] In this scenario, at a treatment rate of 49.8%, physicians prevent 55.5% of preventable strokes and induce 48.4% of inducible bleeds (preventing 14.8% more strokes per bleed than random). A more comprehensive guideline substantially outperforms the $CHADS_2$ score.

We then turn to scenarios involving strict adherence to a guideline. Treatment decisions under these scenarios strictly follow an ordering according to guideline recommendations, and we remove any tendency not to treat older patients and the substantial idiosyncratic variation in treatment decisions. Each guideline implies a "score" that we use to order patients; patients with the same score (e.g., patients with the same $CHADS_2$ score for the $CHADS_2$ guideline) are randomly ordered. We evaluate the performance of adhering to each guideline-implied ordering by a *set* of counterfactual outcomes, moving from no patients treated to all patients treated. Under the assumption that the costs of treatment and monitoring are negligible relative to the clinical benefits,[25] two guideline orderings can be welfare-ranked if one guideline prevents more strokes than the other guideline, for any fixed number of induced bleeds.

Compared to expanding adoption, policies that achieve strict adherence to guidelines produce much better outcomes. Holding bleeds fixed at the status quo level, strict adherence to the $CHADS_2$ score prevents 56% of preventable strokes. Adherence to a score based on full stroke treatment effects performs better still, preventing 68.5% of preventable strokes for the same number of bleeds. Finally, we consider adherence to an optimal guideline that recommends treatment according to the ratio of full stroke and bleed treatment effects, or $-\hat{\tau}^s_{BLP}(x)/\hat{\tau}^b_{BLP}(x)$. Treatment orderings based on this ratio should optimize any linear objective function depending on strokes and bleeding events.[26] Ordering based on the complete ratio can prevent 73% of preventable strokes, or 47% more strokes than under

[24] Our model also does not identify how much a more comprehensive guideline would reduce noise in treatment choices and thus reduce $\sigma_{\varepsilon,\text{post}}$, but we can bound this effect. Our results are not sensitive to the bounds. We discuss this detail, as well as assumptions about distraction effects, in Appendix A.2.

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

[26] For an objective function $u_i = \beta^s \tau^s_i + \beta^b \tau^b_i$, the optimal rule is to treat any patient such that $-\tau^s_i/\tau^b_i > \beta^b/\beta^s$. Note the similarity with Equation (10). In that model, an agent will adhere fully to a ratio rule with observed treatment effects if $f(X_i) + v_i = 0$ for all i. We give details in Appendix A.3.

the status quo, while holding induced bleeds fixed at the status quo level.

In additional counterfactuals, we investigate the role of weighting vs. the number of variables considered. In Appendix Figure 6, we show that adherence to the CHA_2DS_2-VASc guideline prevents fewer strokes at any given number of bleeds than strict adherence to the $CHADS_2$ score (i.e. it performs worse). Our model suggests that vascular disease is an important predictor of stroke treatment effects, but the CHA_2DS_2-VASc guideline gives vascular disease too much weight relative to other variables.[27] Thus, more comprehensive guidelines are not always better. More comprehensive guidelines can improve outcomes, but only if the relevant factors are weighted appropriately given treatment effects. To understand the benefits of more comprehensive guidelines relative to the $CHADS_2$ score, we also consider an "adjusted CHADS" score that assigns integers to approximate the ratio $-\hat{\tau}^s_{BLP}(x)/\hat{\tau}^b_{BLP}(x)$ with each of the five conditions in the original $CHADS_2$ score. Holding induced bleeds fixed at the status quo level, adherence to this guideline prevents 62% of preventable strokes, about one-third of the difference between the $CHADS_2$ and the optimal guideline. In this sense, about one-third of the benefits of our optimal guideline come from better weighting and two-thirds come from including more variables.

The tendency not to treat old people cannot be rationalized by a lower value of a statistical life, since that would bear equally on strokes and bleeds, but it could be rationalized if doctors believed that the relative badness of strokes and bleeds differed for older patients or if the costs of monitoring were more substantial at older ages. In Appendix Figure A.6 and Appendix Table A.8, we explore outcomes in counterfactual scenarios that hold fixed the age distribution (in five-year bins) of treated patients. Observed physician decisions are better than random, but only by a small amount. The benefits of strict guideline adherence are slightly attenuated when constraining the age distribution of treated patients but comparable in magnitude to what we find in the unconstrained scenario. For example, adherence to the optimal ratio guideline can still prevent 44% more strokes than the status quo holding bleeds fixed and maintaining the age distribution of treated patients.

8 Conclusion

Our findings suggest that evidence-based clinical guidelines have the potential to improve patient health outcomes. The $CHADS_2$ score shifted physician behavior and likely prevented a small number

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

of additional strokes while inducing an even smaller number of additional bleeds. Adoption of more comprehensive guidelines that incorporates all of the variables that predict stroke treatment effects would have larger benefits. Stricter adherence to existing or novel treatment rules produces much larger gains than adoption with discretionary adherence. Strict adherence to an optimal treatment rule that minimizes strokes per bleed can prevent 47% more strokes without increasing the number of induced bleeds.

Our results suggest three important lessons for the use of guidelines in clinical care. First, the extensive margin of guideline *adoption*, in which physicians are aware of the guideline but exercise discretion in how to use it, achieves only a fraction of the benefits of greater intensive margin adherence. Second, incorporating more variables into guidelines can improve outcomes, but only if they are weighted properly. We find that optimally weighting the $CHADS_2$ variables achieves about one-third of the benefits of an optimal guideline with all available information. The CHA_2DS_2-VASc score provides a cautionary tale of a more complex score that underperforms relative to the simpler $CHADS_2$ score due to misweighting. Third, improving the prediction of treatment effects for a single outcome may underperform relative to treatment rules that balance a trade-off between different outcomes. An optimal rule combining information on stroke and bleed treatment effects outperforms one based on strokes alone. The best algorithms are designed to match a social objective function (Obermeyer et al., 2019).

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

Our results incorporate more information to estimate treatment effects than has been previously considered, but they only scratch the surface of what is possible. Data from additional trials could

be added and combined optimally with observational data in order to create more powerful predictors of treatment effects. While there remain logistical challenges to the widespread integration of machine-based algorithms into health IT systems (Kawamoto and McDonald, 2020), these are likely to be lessened as data integration and methods of validation in healthcare becomes more commonplace. As such algorithms are implemented, healthcare markets will be reshaped as physicians take on fundamentally different roles (Autor et al., 2003).

References

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

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

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

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

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

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

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

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

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

Autor, David H, Frank Levy, and Richard J Murnane, "The Skill Content of Recent Technological Change: An Empirical Exploration," *Quarterly Journal of Economics*, 2003, *118* (4), 1279–1333.

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

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

Bensch, Gunther and Jörg Peters, "The Intensive Margin of Technology Adoption - Experimental Evidence on Improved Cooking Stoves in Rural Senegal," *Journal of Health Economics*, July 2015, *42*, 44–63.

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

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

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

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

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

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

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

Cohen, Jessica and Pascaline Dupas, "Free Distribution or Cost-Sharing? Evidence from a Randomized Malaria Prevention Experiment," *Quarterly Journal of Economics*, February 2010, *125* (1), 1–45.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Mullainathan, Sendhil and Ziad Obermeyer, "A Machine Learning Approach to Low-Value Health Care: Wasted Tests, Missed Heart Attacks and Mis-Predictions," Working Paper 26168, National Bureau of Economic Research February 2020.

Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan, "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations," *Science*, 2019, *366* (6464), 447–453.

Oster, Emily, "Unobservable selection and coefficient stability: Theory and evidence," *Journal of Business & Economic Statistics*, 2019, *37* (2), 187–204.

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

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

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

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

Pisters, Ron, Deirdre A. Lane, Robby Nieuwlaat, Cees B. De Vos, Harry J.G.M. Crijns, and Gregory Y.H. Lip, "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*, 2010, *138* (5), 1093–1100.

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

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

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

Sacarny, Adam, Michael L Barnett, Jackson Le, Frank Tetkoski, David Yokum, and Shantanu Agrawal, "Effect of peer comparison letters for high-volume primary care prescribers of quetiapine in older and disabled adults: a randomized clinical trial," *JAMA psychiatry*, 2018, *75* (10), 1003–1011.

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

Singer, Daniel E, Yuchiao Chang, Margaret C Fang, Leila H Borowsky, Niela K Pomernacki, Natalia Udaltsova, and Alan S Go, "The net clinical benefit of warfarin anticoagulation in atrial fibrillation," *Annals of internal medicine*, 2009, *151* (5), 297–305.

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

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

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

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

_ , Robert G. Hart, Stuart Connolly, Peter C. Austin, Jonathan Mant, F.D. Richard Hobbs, Peter J. Koudstaal, Palle Petersen, Francisco Perez-Gomez, J. Andre Knottnerus, Beppie Boode, Michael D. Ezekowitz, and Daniel E. Singer, "Effect of Age on Stroke Prevention Therapy in Patients with Atrial Fibrillation: The Atrial Fibrillation Investigators," *Stroke*, 2009, *40* (4), 1410–1416.

Vytlacil, Edward, "Independence, Monotonicity, and Latent Index Models: An Equivalence Result," *Econometrica*, 2002, *70* (1), 331–341.

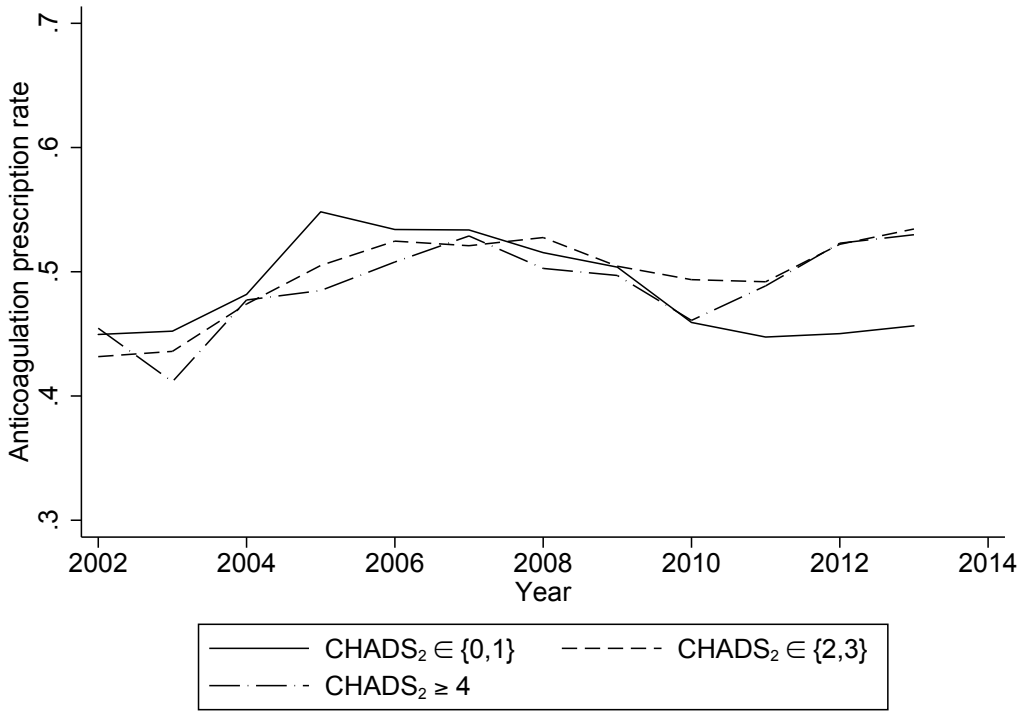
Wager, Stefan and Susan Athey, "Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests," *Journal of the American Statistical Association*, 2017, *0* (ja), 0–0.

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

Werner, Rachel M, Jonathan T Kolstad, Elizabeth A Stuart, and Daniel Polsky, "The effect of pay-for-performance in hospitals: lessons for quality improvement," *Health Affairs*, 2011, *30* (4), 690–698.

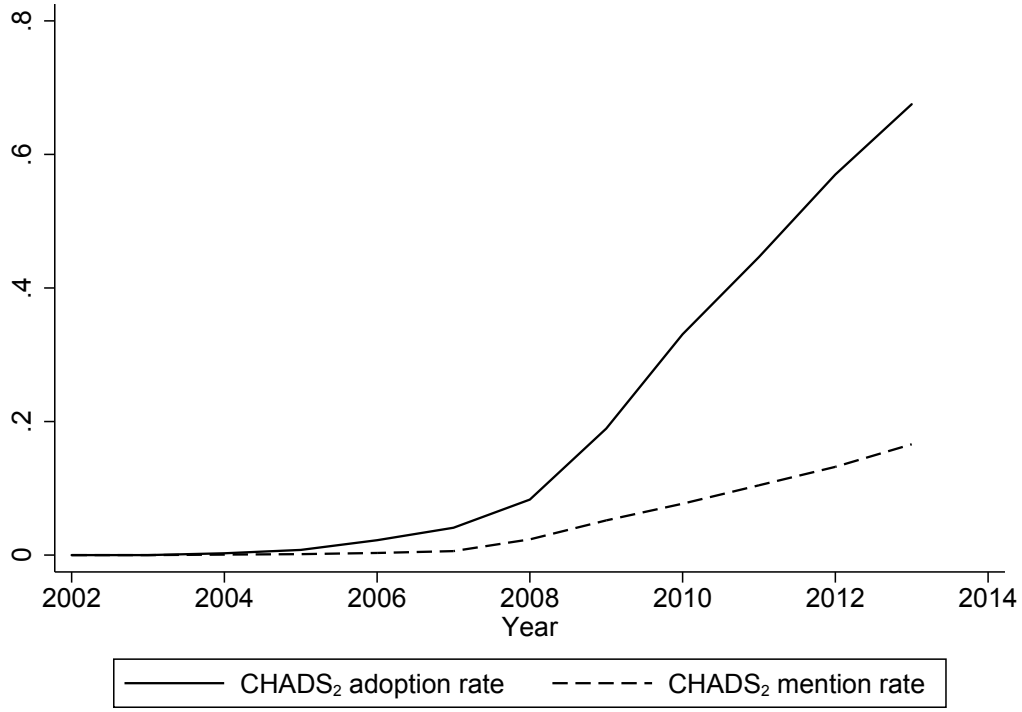
Woolf, Steven H., Richard Grol, Allen Hutchinson, Martin Eccles, and Jeremy Grimshaw, "Potential Benefits, Limitations, and Harms of Clinical Guidelines," *BMJ*, 1999, *318* (7182), 527–530.

Figure 1: Anticoagulation Trends by $CHADS_2$ Score



Notes: This figure shows the fraction of atrial fibrillation patients treated with anticoagulation over time, for three groups of patients by $CHADS_2$ score. The sample reflect patients with newly diagnosed atrial fibrillation in the VHA, and anticoagulation treatments are defined as prescriptions within 90 days of initial diagnosis. Table 1 provides further details about the sample selection.

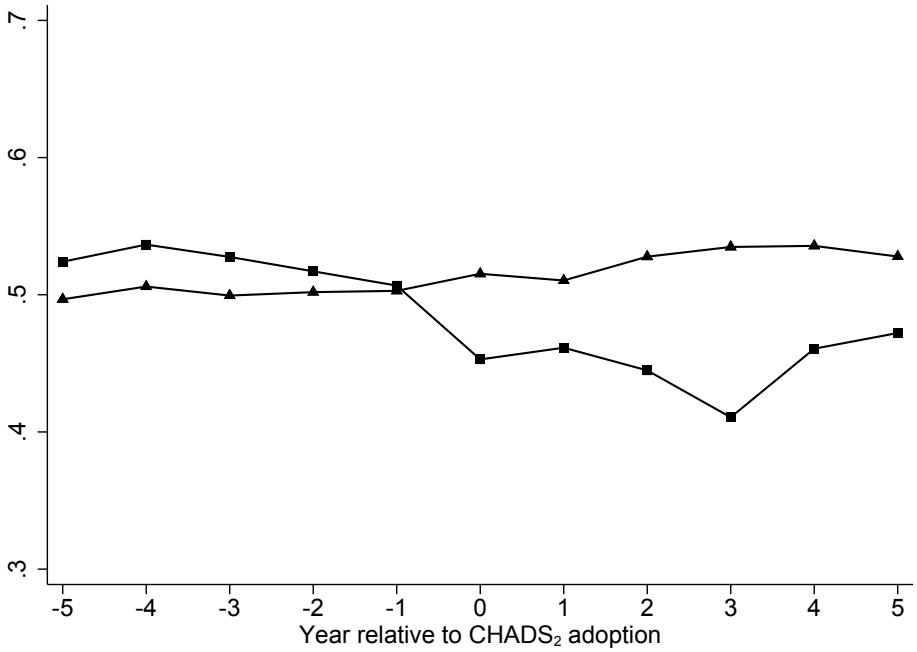
Figure 2: Diffusion of the $CHADS_2$ Score



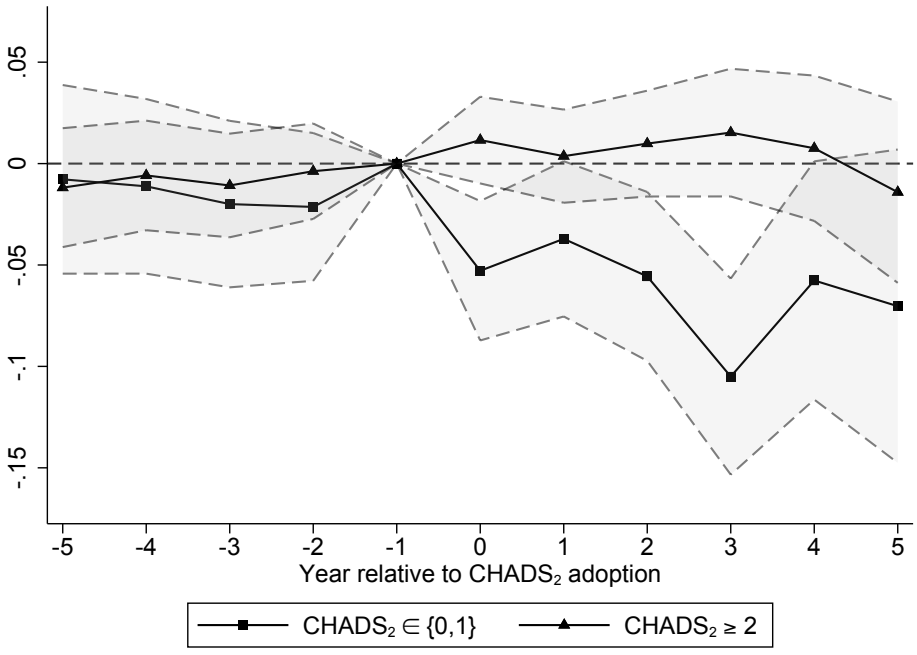
Notes: This figure shows the fraction of patients a given year with physicians who either mention the $CHADS_2$ score in the note for the index patient or have mentioned the $CHADS_2$ score in either the note for the index patient or in a previous note. We identify mentions by searching the note text for the phrase `chads` (not case-sensitive). We consider any physician who has mentioned the the $CHADS_2$ score in the current note or in a previous note as having *adopted* the $CHADS_2$ guideline, shown in the solid line. The dashed line reflects the rate of mentions in the index patient's note. The sample reflect patients with newly diagnosed atrial fibrillation in the VHA. Table 1 provides further details about the sample selection.

Figure 3: Treatment Decisions and $CHADS_2$ Adoption

A. Trends Relative to $CHADS_2$ Adoption

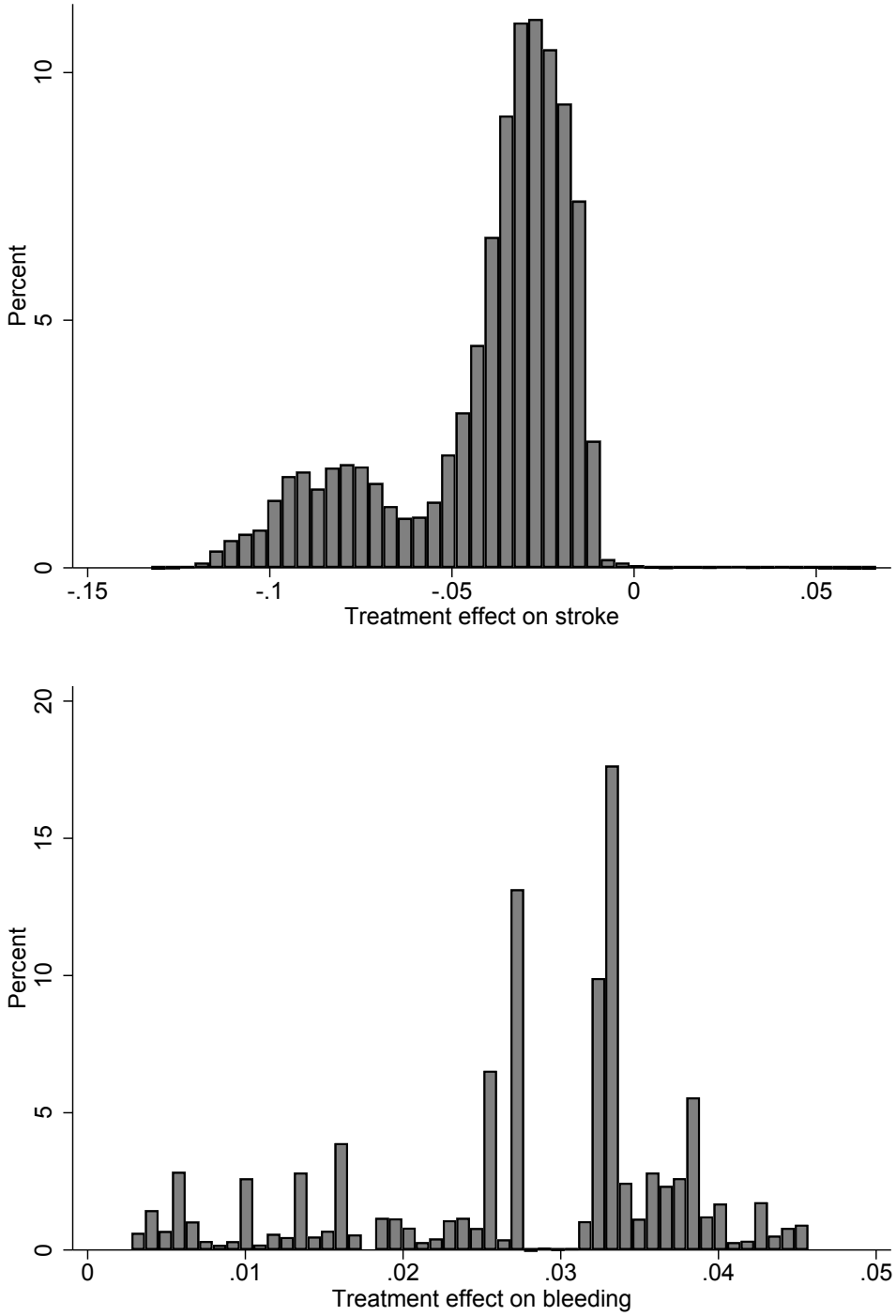


B. Event Study Estimates



Notes: Panel A displays the fraction of atrial fibrillation patients treated with anticoagulation in each year relative to $CHADS_2$ adoption for physicians who eventually adopt the $CHADS_2$ score. Panel B shows regression coefficients and 95% confidence intervals from Equation (1), run separately for patients with $CHADS_2 \in \{0, 1\}$ and for patients with $CHADS_2 \geq 2$. The 12-month period prior to the physician's first $CHADS_2$ mention is normalized to 0. The regression sample includes 104,585 VHA patients who either are treated within 5 years of their physician's observed $CHADS_2$ adoption or are treated by a non-adopting physician.

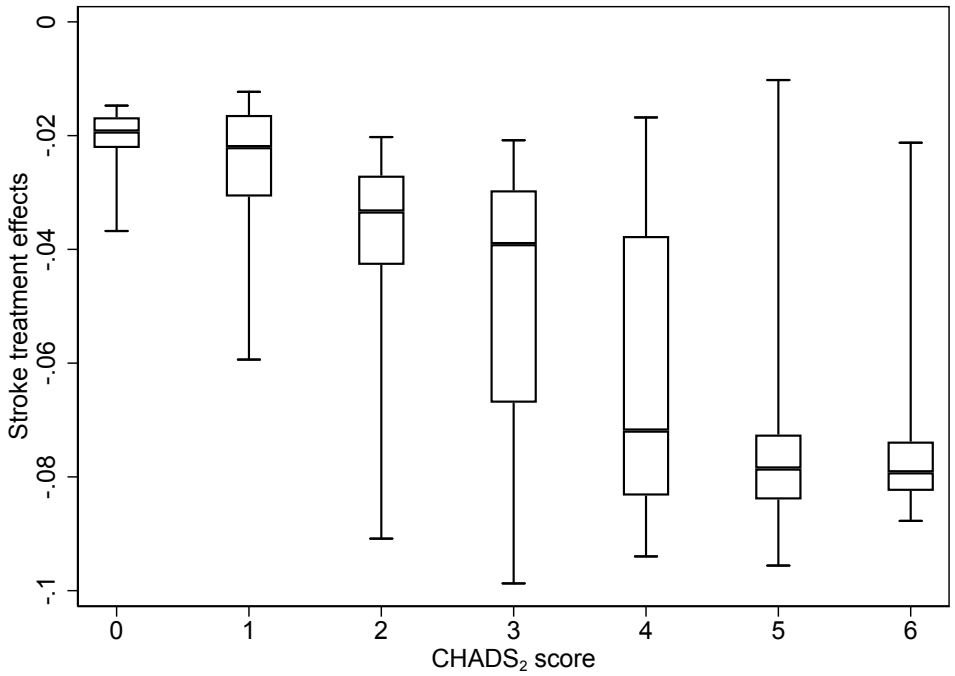
Figure 4: Distribution of Treatment Effects Across VHA Patients



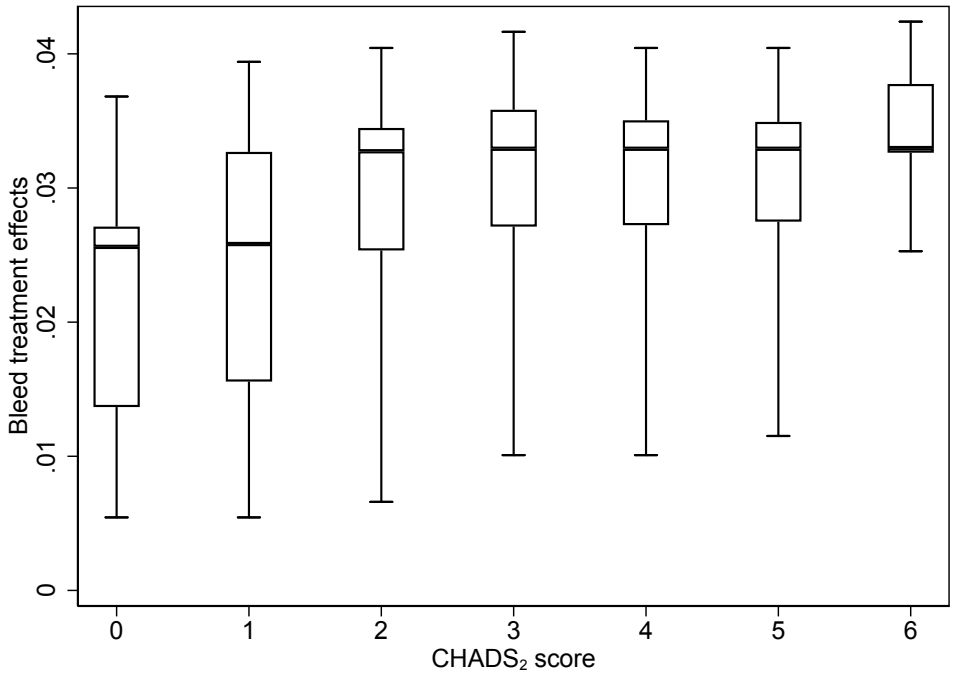
Notes: This figure displays histograms of stroke and bleed treatment effects in the VHA sample. Conditional average treatment effect (CATE) predictions are trained and validated by using causal forest methods, described in Section 5, applied to RCT data in the AFI database. We use the causal forest rules to calculate CATEs as a function of patient characteristics for each patient in the VHA data.

Figure 5: Treatment Effects by $CHADS_2$ Score

A. Stroke Treatment Effects



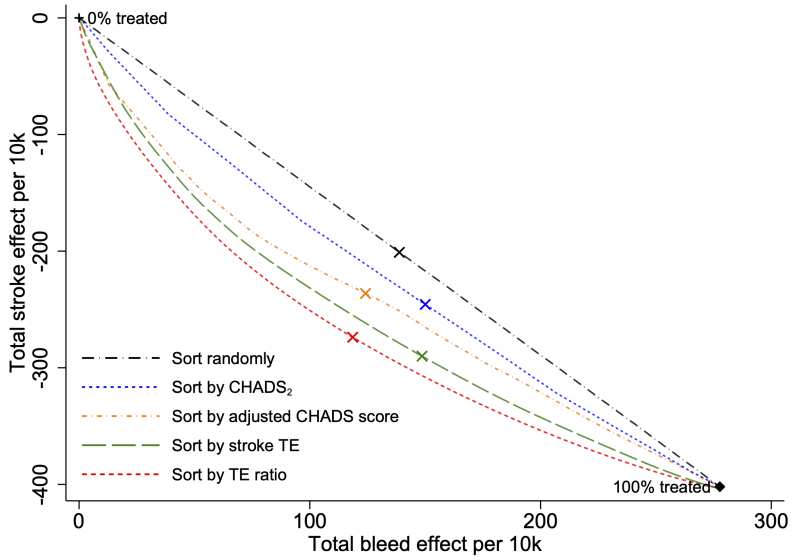
B. Bleed Treatment Effects



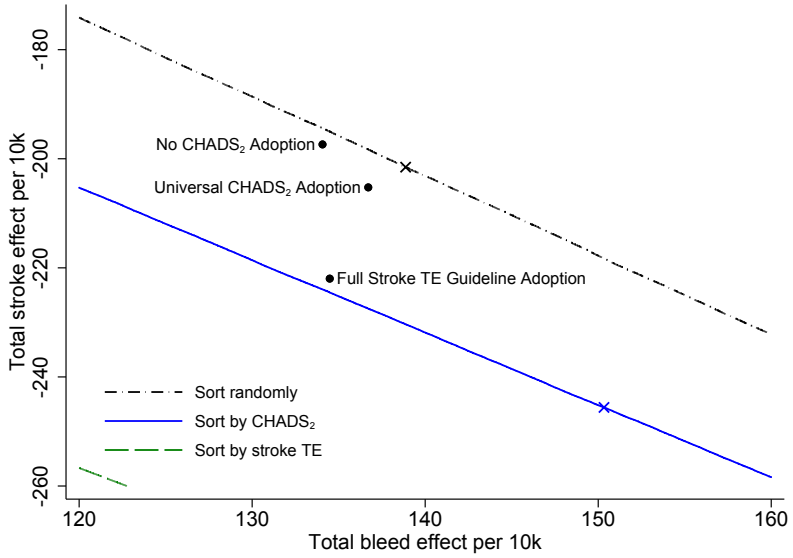
Notes: This figure shows box plots for the distribution of treatment effects by $CHADS_2$ score in the VHA data. Stroke treatment effects are shown in Panel A; bleed treatment effects are shown in Panel B. Bounds on the box plot are at the 25th and 75th percentile, with the median marked with a horizontal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

Figure 6: Counterfactual Outcomes

A. Strict Guideline Adherence



B. Guideline Adoption (Inset)



Notes: This figure shows strokes prevented and bleeds induced by anticoagulation in counterfactual scenarios. Strokes prevented per 10,000 patients are shown on the *y*-axis, and bleeds induced per 10,000 patients are shown on the *x*-axis. Both panels display outcomes under a random treatment allocation, ranging from 0% of patients treated (top-left corner of Panel A) to 100% of patients treated (bottom-right corner of Panel A). Panel A shows outcomes under counterfactual strict adherence to various guideline rules. Each rule implies a patient sorting, and the curves indicate counterfactual outcomes ranging from treating 0% to 100% of patients. "Sort by adjusted CHADS score" involves sorting by an integer score based on the $CHADS_2$ component conditions (see Table 2) but predicting the ratio of stroke treatment effect to bleed treatment effect. "Sort by TE ratio" sorts by the ratio of stroke treatment effect to bleed treatment effect. Crosses indicate the points where 50% of patients are treated. Panel B shows an inset area and plots outcomes for counterfactual adoption (i.e., imperfect adherence) scenarios.

Table 1: VHA Sample Selection

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

Notes: This table describes sample selection rules used to define a cohort of newly diagnosed atrial fibrillation patients in the VHA.

Table 2: $CHADS_2$ Score and Treatment Recommendations

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

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

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

Table 3: Summary Statistics

	VHA Data		AFI Database
	Adopting Physician	Non-Adopting Physician	
Characteristic	(1)	(2)	(3)
Treated with anticoagulation	0.50	0.50	0.39
Male	0.99	0.99	0.65
Age	74.0	74.2	70.8
	(10.0)	(9.7)	(9.5)
Stroke treatment effect	−0.041	−0.040	−0.043
	(0.025)	(0.024)	(0.026)
Bleed treatment effect	0.028	0.028	0.024
	(0.010)	(0.024)	(0.012)
$CHADS_2$ components:			
Congestive heart failure	0.16	0.15	0.29
Hypertension	0.84	0.83	0.48
Age ≥ 65	0.51	0.52	0.76
Diabetes	0.36	0.35	0.15
Previous stroke	0.15	0.14	0.14
Number of physicians	2,950	2,802	
Number of patients	64,592	48,678	6,707

Notes: This table reports mean and standard deviations of characteristics of patients in the VHA data and in the AFI database. Standard deviations are shown in parentheses. Columns 1 and 2 show characteristics of patients in the VHA data, with sample selection steps described in Table 1, for patients with physicians who have adopted the $CHADS_2$ guideline and for those with physicians who have adopted the guideline, respectively. Column 3 shows characteristics of patients from the AFI database.

Table 4: Probit Estimates

	Dependent Variable: Anticoagulant Prescription		
	(1)	(2)	(3)
$CHADS_2$-related stroke treatment effect, $\hat{\tau}^{s(c)}_{BLP}(x)$			
Pre-adoption baseline, $\alpha^{s(c)}_{\text{pre}}$	−4.705***	−4.885***	−6.707***
	(0.698)	(1.003)	(1.233)
Post-adoption difference, $\alpha^{s(c)}_{\text{post}} - \alpha^{s(c)}_{\text{pre}}$	−6.078***	−6.042***	−5.252***
	(1.097)	(1.138)	(1.474)
Never-adopter difference, $\alpha^{s(c)}_{\text{never}} - \alpha^{s(c)}_{\text{pre}}$	0.941	0.906	1.214
	(0.711)	(0.750)	(1.397)
Residual stroke treatment effect, $\hat{\tau}^{s(r)}_{BLP}(x)$			
Pre-adoption baseline, $\alpha^{s(r)}_{\text{pre}}$	0.609**	0.450	0.423
	(0.308)	(0.327)	(0.668)
Post-adoption difference, $\alpha^{s(r)}_{\text{post}} - \alpha^{s(r)}_{\text{pre}}$	−0.925*	−0.778	−0.794
	(0.481)	(0.495)	(0.785)
Never-adopter difference, $\alpha^{s(r)}_{\text{never}} - \alpha^{s(r)}_{\text{pre}}$	0.0430	0.144	0.392
	(0.404)	(0.422)	(0.961)
Bleed treatment effect, $\hat{\tau}^{b}_{BLP}(x)$			
Pre-adoption baseline α^{b}_{pre}	−3.272***	−3.128***	−3.066**
	(0.793)	(0.864)	(1.551)
Post-adoption difference, $\alpha^{b}_{\text{post}} - \alpha^{b}_{\text{pre}}$	2.462*	2.232*	1.933
	(1.280)	(1.345)	(1.896)
Never-adopter difference, $\alpha^{b}_{\text{never}} - \alpha^{b}_{\text{pre}}$	1.065	0.911	0.311
	(1.037)	(1.097)	(2.036)
Year fixed effects, age spline controls	Yes	Yes	Yes
Differential trends on treatment effects	No	Yes	No
Treatment effects interacted with year identities	No	No	Yes
Number of observations	113,270	113,270	113,270

Notes: This table reports estimates from probit regressions of anticoagulation treatment decisions, as specified in Equation (13). Key regressors of interest are causal forest predictions of CATEs: $CHADS_2$-related stroke treatment effects, or $\hat{\tau}^{s(c)}_{BLP}(x)$; residual stroke treatment effects, , or $\hat{\tau}^{s(r)}_{BLP}(x)$; and bleed treatment effects, or $\hat{\tau}^{b}_{BLP}(x)$. Column 2 includes linear trends interacted with each of these treatment effects. Column 3 includes three-way interactions of (i) each of the treatment effects, (ii) indicators for two-year periods, and (iii) adoption status. These probit models allow that $\sigma_{\varepsilon,g}$ may vary with adoption status and year fixed effects, reported in Appendix Table A.6. Standard errors are clustered at the physician level. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 5: Counterfactual Treatment Decisions and Outcomes

	Treated Patients	Strokes Prevented	Bleeds Induced
A: Benchmarks			
Observed treatment choices	49.8%	199	134
Randomly assigned treatment	49.8%	199	138
B: Guideline Adoption			
No $CHADS_2$ adoption	49.7%	197	133
Universal $CHADS_2$ adoption	50.2%	205	137
Universal adoption of stroke TE guideline	49.8%	222	134
C: Strict Guideline Adherence, Fixed Bleeds Induced			
Strict adherence to $CHADS_2$	44.6%	224	134
Strict adherence to adjusted CHADS score	58.2%	247	134
Strict adherence to stroke TE guideline	45.1%	274	134
Strict adherence to TE-ratio guideline	55.1%	292	134

Notes: This table reports treatment rates and patient outcomes in counterfactual adoption and adherence scenarios for patients in the VHA data. Outcomes of strokes prevented and bleeds induced are reported per 10,000 patients. Panel A reports treatment rates and outcomes under the status quo, which we observe in our data, and in a counterfactual assignment of the same number of treatments to random patients. Panel B reports treatment rates and outcomes under counterfactual adoption scenarios, assuming adherence implied by our structural model in Equation (13). "Universal adoption of stroke TE guideline" refers to guideline adoption in which physicians become equally informed of $CHADS_2$-related and residual stroke treament effects (i.e., $\hat{\tau}^{s(c)}_{BLP}(x)$ and $\hat{\tau}^{s(r)}_{BLP}(x)$). Panel C reports treatment rates and outcomes under patient orderings according to scores implied by counterfactual strict adherence to different guidelines. Patients with the same score are randomly ranked. "Strict $CHADS_2$ adherence" orders patients by their $CHADS_2$ score. "Strict adherence to adjusted CHADS score" orders patients by a point system approximating $\hat{\tau}^{s}_{BLP}(x)/\hat{\tau}^{b}_{BLP}(x)$ by $CHADS_2$ indicators in Table 2. "Strict adherence to stroke TE guideline" orders patients by $\hat{\tau}^{s}_{BLP}(x)$. "Strict adherence to TE-ratio guideline" orders patients by $\hat{\tau}^{s}_{BLP}(x)/\hat{\tau}^{b}_{BLP}(x)$.

Online Appendix

Mastering the Art of Cookbook Medicine:
Machine Learning, Randomized Trials, and Misallocation
Jason Abaluck, Leila Agha, David Chan, Daniel Singer, Diana Zhu

A.1 Bayesian Model of Decision-Making

In this appendix, we describe in greater detail the Bayesian model of decision-making specified in Equation (10), which we restate here:

$$W_i = \mathbf{1}\left\{\beta^s \tilde{\tau}^s_{i,g} + \beta^b \tilde{\tau}^b_{i,g} + f(X_i) + v_i > 0\right\},$$

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

A.1.1 Component Treatment Effects, Signals, and Beliefs

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

$$\tau^o_{i,k} \sim N\left(\overline{\tau}^o_k, \sigma^2_{\tau(o),k}\right), \tag{A.1}$$

for outcome o and each $k \in \mathcal{K}$. We assume that physicians know the moments of each component treatment effect $\left(\overline{\tau}^o_k, \sigma^2_{\tau(o),k}\right)_{k=1}^K$.[28]

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

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

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

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

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

[28] Our model in Equation (10) allows for potentially non-Bayesian beliefs that can shift decision-making via $f(X_i)$ and v_i. In order to study the effect of information in a Bayesian framework, we compartmentalize the two components of the model and consider the first component, described in this appendix, as Bayesian.

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

A.1.2 Regression Interpretation

The relationship between posterior beliefs and signals in Equation (A.3) can be interpreted as a regression of posterior beliefs on signals. This relationship may also be interpreted as a regression of posterior beliefs on true treatment effects, since the noise component of signals is orthogonal to treatment effects:

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

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

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

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

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

$$\tilde{\tau}^o_{i,g} = \lambda^o_g \dot{\tau}^o_{i,g} + (1-\lambda^o_g)\overline{\tau}^o. \tag{A.5}$$

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

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

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

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

A.1.3 CHADS$_2$ and Residual Treatment Effects

We are now in a position to state posterior beliefs as in Equations (11) and (12). For strokes, we can separate the set of components $\mathcal{K}_c$ that predict CHADS$_2$-related treatment effects and $\mathcal{K} \setminus \mathcal{K}_c$ components that predict residual treatment effects. We expect that the component posterior beliefs related to the CHADS$_2$ score should increase in informativeness. That is, we expect that $\lambda^s_{g,k}$ should increase with g = post, for $k \in \mathcal{K}_c$. We first define the two components of stroke treatment effects: $\tau_i^{s(c)} \equiv \sum_{k\in\mathcal{K}_c} \tau^s_{i,k}$, and $\tau_i^{s(r)} \equiv \sum_{k\notin\mathcal{K}_c} \tau^s_{i,k}$. Restating Equation (11) as

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

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

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

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

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

Unlike $\lambda_g^{s(c)}$ and $\lambda_g^{s(r)}$, μ^s_g and $\text{Var}\left(\upsilon^s_{i,g}\right)$ are not exactly invariant to the level of aggregation with which posterior beliefs are formed.[29] Nevertheless, regardless of this level of aggregation, qualitative interpretations are unchanged: μ^s_g is a function of the signal-to-noise ratio and prior beliefs, and $\upsilon^s_{i,g}$ is a function of signal-to-noise ratio and noise. If $\lambda^s_{g,k} = 1$ for all $k \in \mathcal{K}$, there is no noise, and $\upsilon^s_{i,g} = 0$. At the other extreme, if $\lambda^s_{g,k} = 0$ for all $k \in \mathcal{K}$, there is no meaningful signal. In this case, physicians will ignore all $\dot{\tau}^s_{g,k}$, and we will also have $\upsilon^s_{i,g} = 0$.

For bleeding events, we are only interested in overall treatment effects: $\tau_i^b \equiv \sum_{k\in\mathcal{K}} \tau^b_{i,k}$. Restating Equation (12) as

$$\tilde{\tau}^b_{i,g} = \lambda^b_g \tau^b_i + \mu^b_g + \upsilon^b_{i,g},$$

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

we interpret the signal-to-noise coefficient as

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

If we conceptualize the posterior belief as directly formed from $\dot{\tau}_i^b \equiv \tau_i^b + \sum_{k\in\mathcal{K}_c} \epsilon^b_{i,g,k}$, then we can similarly interpret the constant, μ_g^b, as a function of λ_g^b and prior beliefs, or $\mu_g^b = \left(1-\lambda_g^b\right)\sum_{k\in\mathcal{K}} \overline{\tau}^s_k$. The error term, $\upsilon^b_{i,g}$, is similarly a function of λ_g^b and noise, or $\upsilon^b_{i,g} = \lambda_g^b \sum_{k\in\mathcal{K}} \epsilon^s_{i,g,k}$.

A.2 Adoption of Guideline Revealing $\left(\tau^{s(c)}(x), \tau^{s(r)}(x)\right)$

In this appendix, we discuss how we evaluate a counterfactual scenario in which physicians adopt a guideline that contains information on both $\tau^{s(c)}(x)$ and $\tau^{s(r)}(x)$. We assume that physicians who have adopted this guideline will have equal information on $\hat{\tau}^{s(c)}(x)$ and $\hat{\tau}^{s(r)}(x)$, which implies that $\lambda^{s(r)}_{\text{post}} = \lambda^{s(c)}_{\text{post}}$. We assume the same distraction effects as the CHADS_2 score, so that α^b_{post} remains unchanged. To evaluate counterfactual outcomes, we need to know the effect of this policy on $\sigma_{\varepsilon,\text{post}}$. Although this object is not identified in the data, we proceed with a bounding exercise.

We first note that adoption of the CHADS_2 score did little to change $\sigma_{\varepsilon,\text{post}}$. This suggest that, although the CHADS_2 score had a significant impact on physicians' information on $\tau^{s(c)}(x)$ (i.e., doubling $\lambda^{s(c)}_{\text{post}}$ relative to $\lambda^{s(c)}_{\text{pre}}$), uncertainty about $\tau^{s(c)}(x)$ accounts for very little in the variation in treatment decisions across patients with the same treatment effects. We therefore expect very little impact of the comprehensive guideline that also informs physicians of $\tau^{s(r)}(x)$.

As a baseline assumption, we consider no additional effect of the comprehensive guideline on $\sigma_{\varepsilon,\text{post}}$. We also consider a lower bound on $\sigma_{\varepsilon,\text{post}}$, which corresponds to an upper bound on the welfare improvement from the comprehensive guideline. In this lower bound, we consider the effect of the comprehensive guideline reducing the variance of $\beta^s \lambda_g^{s(r)} \sum_{k\notin\mathcal{K}_c} \epsilon^s_{i,g,k}$ to 0. We note that

$$\begin{aligned}
\text{Var}\left(\beta^s \lambda_g^{s(r)} \sum_{k\notin\mathcal{K}_c} \epsilon^s_{i,g,k}\right) &= \left(\beta^s \lambda_g^{s(r)}\right)^2 \sum_{k\notin\mathcal{K}_c} \sigma^2_{\epsilon(s),g,k} \\
&= \left(\frac{\alpha_g^{s(r)}}{\lambda_g^{s(r)}}\right)^2 \left(\lambda_g^{s(r)}\right)^2 \sum_{k\notin\mathcal{K}_c} \sigma^2_{\epsilon(s),g,k} \\
&\leq \left(2\alpha_g^{s(r)}\right)^2 \frac{1}{4} \sum_{k\notin\mathcal{K}_c} \sigma^2_{\tau(s),g,k} \\
&= \left(\alpha_g^{s(r)}\right)^2 \text{Var}\left(\tau^{s(r)}(x)\right).
\end{aligned}$$

The inequality comes from the fact that $\left(\lambda_g^{s(r)}\right)^2 \sum_{k\notin\mathcal{K}_c} \sigma^2_{\epsilon(s),g,k}$ is largest when $\lambda_g^{s(r)} = \frac{1}{2}$. This occurs when $\sum_{k\notin\mathcal{K}_c} \sigma^2_{\epsilon(s),g,k} = \sum_{k\notin\mathcal{K}_c} \sigma^2_{\tau(s),g,k}$. To evaluate this bound, we use the empirical variance of $\hat{\tau}^{s(r)}_{BLP}(x) \approx 0.00013$ for $\text{Var}\left(\tau^{s(r)}(x)\right)$.

A.3 Optimal Treatment as Function of $\left(\tau^s(x), \tau^b(x)\right)$

In this appendix, we will examine the optimal ranking of patients as a function of $\left(\tau^s(x), \tau^b(x)\right)$. We will use this ranking to characterize the optimal guideline under strict adherence, in which physicians follow the ranking of patients specified by the guideline. Intuitively, we find that the optimal guideline should rank patients in terms of the ratio of stroke treatment effects per bleed treatment effect. In other words, patients should be ranked by $|\tau^s(x)|/\tau^b(x)$, and patients with the highest value of $|\tau^s(x)|/\tau^b(x)$.

In order to see this intuition, we may first consider a social welfare function of the form

$$U_i = \overline{\beta}^s Y_i^s + \overline{\beta}^b Y_i^b, \tag{A.8}$$

where $\overline{\beta}^s$ and $\overline{\beta}^b$ are social preference weights with respect to stroke and bleeds, and Y_i^s and Y_i^b are realize stroke and bleed outcomes. Using notation from Section 5.1, we note that $Y_i^o = W_i Y_i^s(1) + (1 - W_i)Y_i^s(0)$, for $o \in \{s, b\}$. If the planner knows conditional treatment effects, $\left(\tau^s(x), \tau^b(x)\right)$, defined in Equations (2) and (3) as

$$\begin{aligned} \tau^s(x) &\equiv E\left[Y_i^s(1) - Y_i^s(0) \middle| X_i = x\right]; \\ \tau^b(x) &\equiv E\left[Y_i^b(1) - Y_i^b(0) \middle| X_i = x\right], \end{aligned}$$

she may maximize $E[U_i]$ with the following decision rule:

$$W_i = \mathbf{1}\{\overline{\beta}^s \tau^s(X_i) + \overline{\beta}^b \tau^b(X_i) > 0\}. \tag{A.9}$$

This decision rule is equivalent to a rule that treats if and only if $|\tau^s(x)|/\tau^b(x) > \overline{\beta}^b/\overline{\beta}^s$.

Note the similarity between the decision rule implied by Equation (A.9) and our model of physician decision-making in Equation (10). Physician decisions W_i will maximize U_i if and only if $\overline{\beta}^b/\overline{\beta}^s = \beta^b/\beta^s$ and $f(X_i) + \nu_i = 0$ for all i. The first condition reflects alignment in physician preferences and social preference weights. The second condition implies that physicians perfectly observe $\left(\tau^s(x), \tau^b(x)\right)$ and have no other decision-making considerations.

The decision rule in Equation (A.9) can be generalized to any arbitrary social preference ratio $\pi = \overline{\beta}^b/\overline{\beta}^s$. Suppose we consider optimal decisions under two preference ratios, π and $\pi' < \pi$. All patients who should be treated under the preference ratio π should also be treated under the preference ratio π'. No patients who are not treated under the preference ratio π' should be treated under the preference ratio π. Therefore, there are only incremental patients who should be treated under π' and not under π, and not vice versa. These patients have a ratio of treatment effects $|\tau^s(x)|/\tau^b(x) \in (\pi', \pi]$. This equivalence between patient rankings and utility maximization is noted in Vytlacil (2002) and Chan et al. (2019).

A.4 Predicting Physician Treatment Decisions

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

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

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

3. *Variables related to frailty and fall risk:* neurologic disorder (including Parkinson's Disease), fall risk (neuropathy, muscle weakness, dizziness), vision problems, arthritis, head injury, fracture. Frailty and fall risk are frequently cited clinical explanations for not prescribing warfarin to patients with high $CHADS_2$ scores. Patients with high fall risk may be more likely to suffer intracranial bleeds if they are taking warfarin.

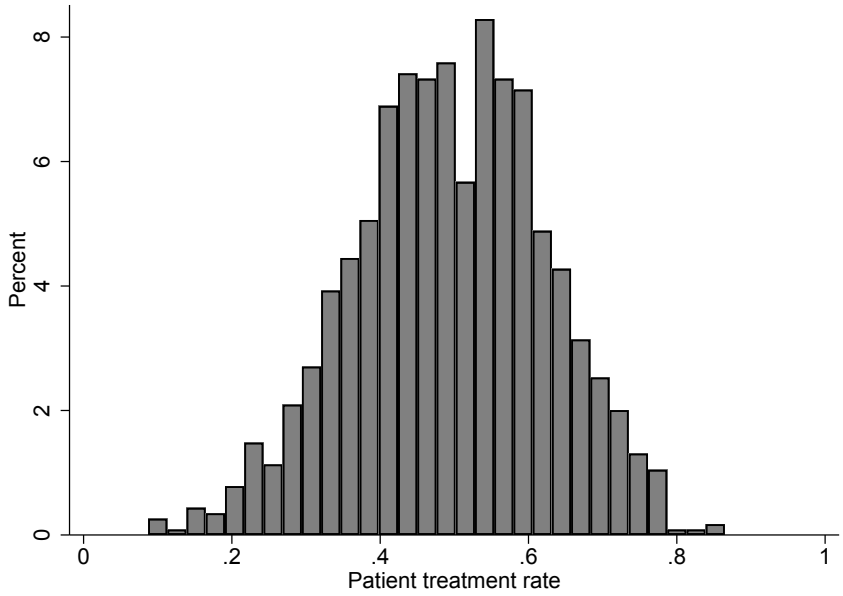
4. *Elixhauser comorbidities that are not in the AFI database:* HIV/AIDS, deficiency anemia, hypothyroidism, tumor, metastasis, lymphoma, obesity, weight loss, paralysis, pulmonary circulation disorders, ulcer, valvular disease. These are additional patient characteristics that have been shown to predict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

5. *Physician characteristics:* doctor specialty code (cardiology, internal medicine, primary care). This specialty coding variable indicates the doctor's training and role at the VHA.

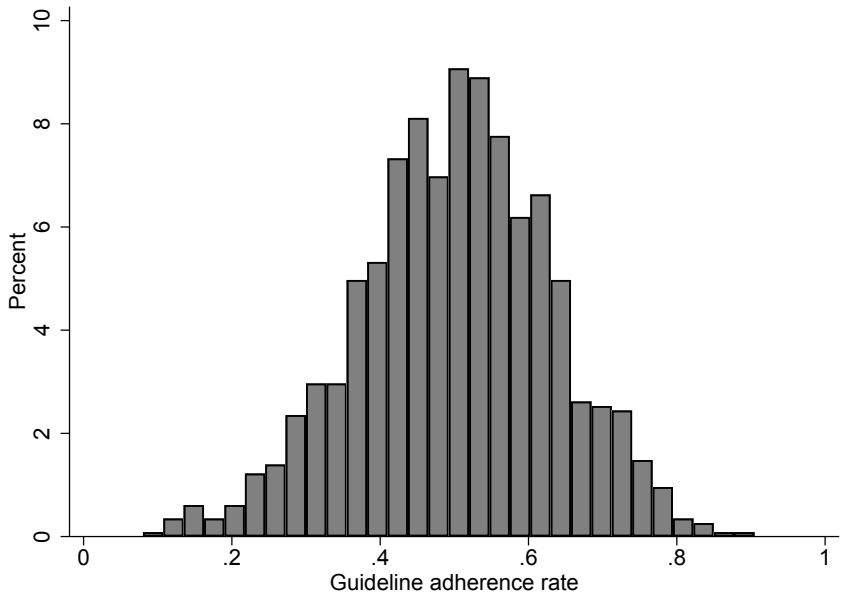
Controlling for these variables in our model estimation does not materially change the conclusions of our analysis. Figure A.4 reports the results of regressions that permute the control variable sets to cover every possible combination of the above list. In Panel A, we find a similar increase in sensitivity to the $CHADS_2$-component of stroke treatment effects after guideline adoption in each model, regardless of the set of included controls. In Panel B, we show that the unexplained variance in treatment propensity does not change substantially, even after we control for these detailed patient and physician characteristics.

Figure A.1: Distribution of Physician Treatment Decisions

A. Treatment

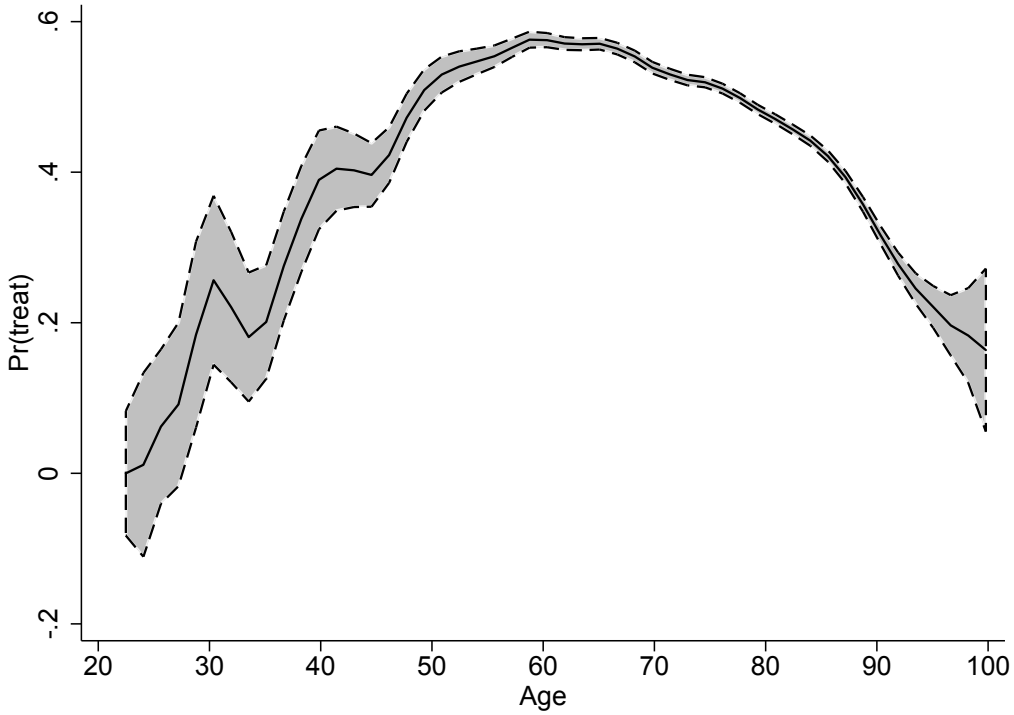


B. $CHADS_2$ Adherence



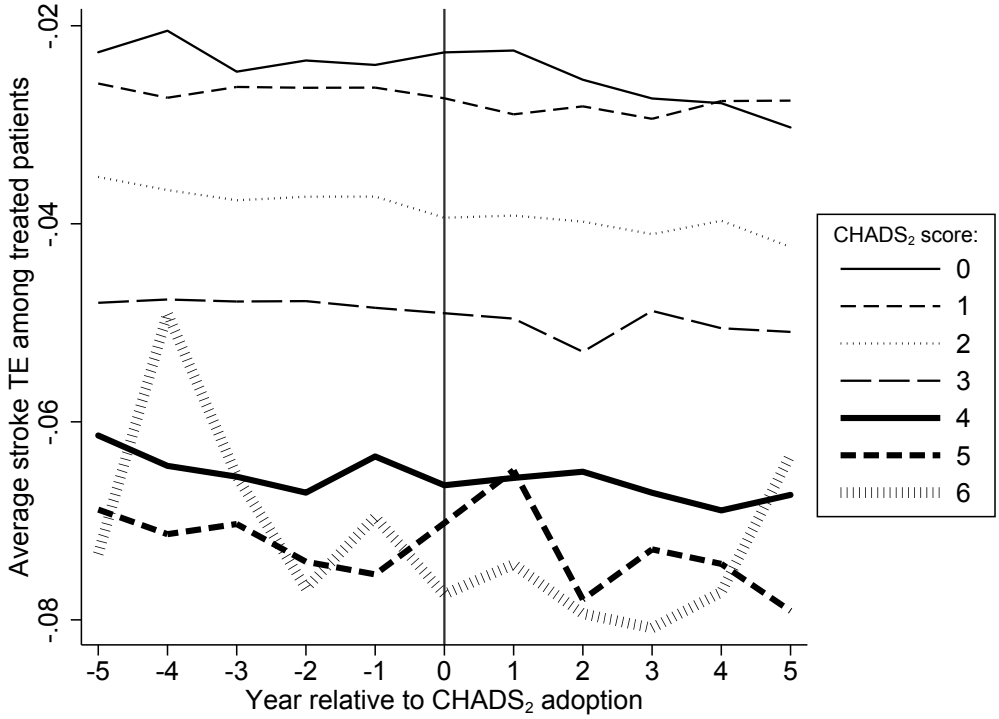
Notes: These figures show the distribution of treatment rates and $CHADS_2$ adherence rates across physicians. They cover the subsample of 1,146 physicians treating at least 30 patients in our final analysis sample. This covers 50,426 patients treated by the higher volume doctors, or a little less than half of the VHA sample defined in Table 1. Panel A shows the distribution of treatment rates. Panel B shows the distribution of $CHADS_2$ adherence rates. We define $CHADS_2$-adherent anticoagulation decisions as follows: No anticoagulation for patients with a $CHADS_2$ of 0 and anticoagulation for patients with a $CHADS_2$ score greater than or equal 2; we omit patients with a $CHADS_2$ score of 1 from this calculation, since the 2008 ACCP guideline allowed for either anticoagulation or aspirin for these patients.

Figure A.2: Treatment Probability by Patient Age



Notes: This figure shows the probability of anticoagulation as a function of patient age in the VHA sample. The curve fits the observed data with a kernel weighted local polynomial; the shaded area represents the 95% confidence interval.

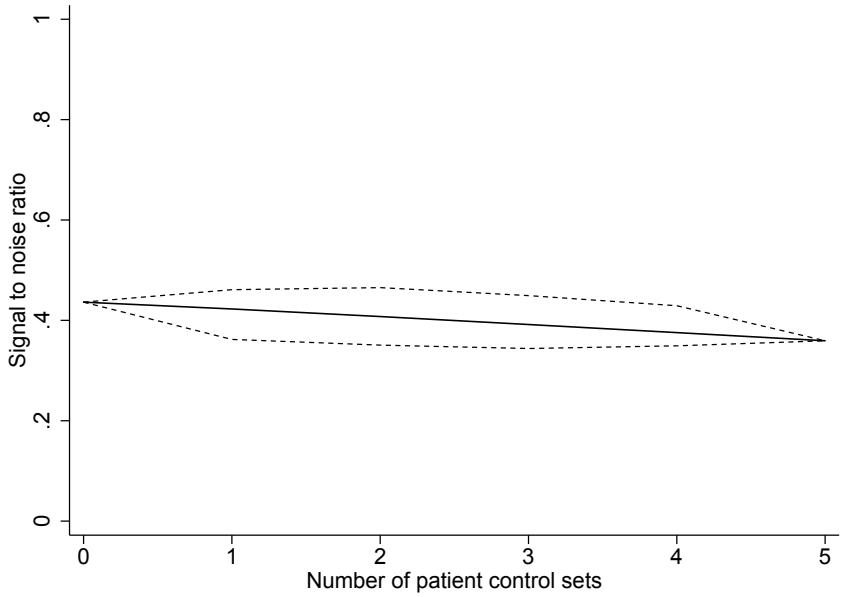
Figure A.3: Stroke Treatment Effects Among Treated Patients



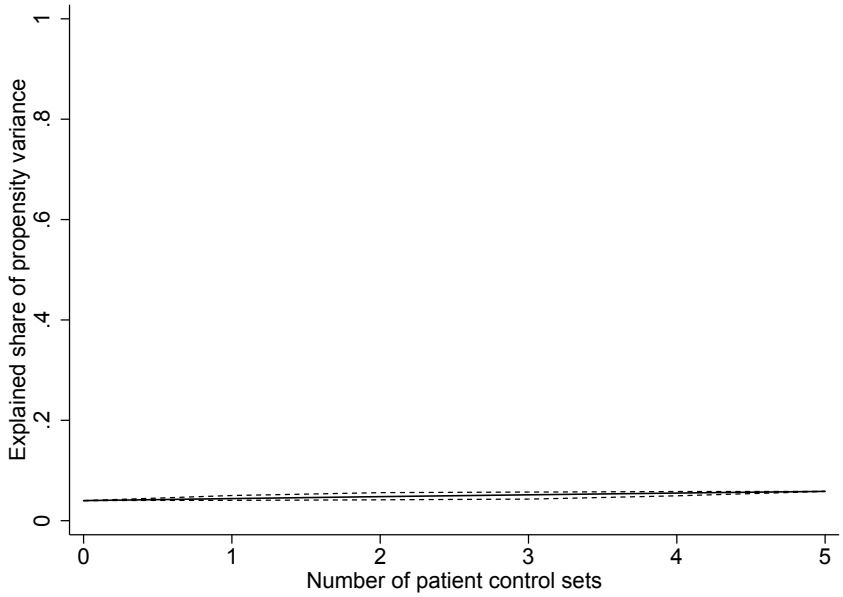
Notes: This figure displays time trends in the average stroke treatment effect among treated patients, separately by $CHADS_2$ score. The physician's first $CHADS_2$ mention occurs on the first day of year 0. Stroke treatment effects are predicted using causal forest rules trained and validated in the AFI database; the causal forest rules are then applied to patient characteristics in the VHA data. Adoption status and treatment decisions are measured in the VHA data.

Figure A.4: Stability of the Structural Results

A. Increase in Signal-to-Noise Ratio, $\lambda^{s(c)}_{\text{pre}}/\lambda^{s(c)}_{\text{post}}$



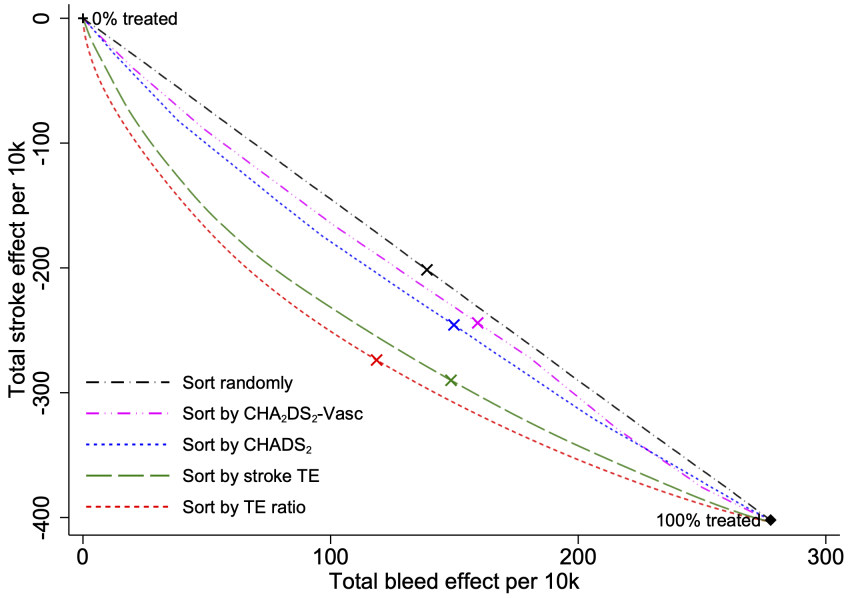
B. Explained Share of Latent Variable



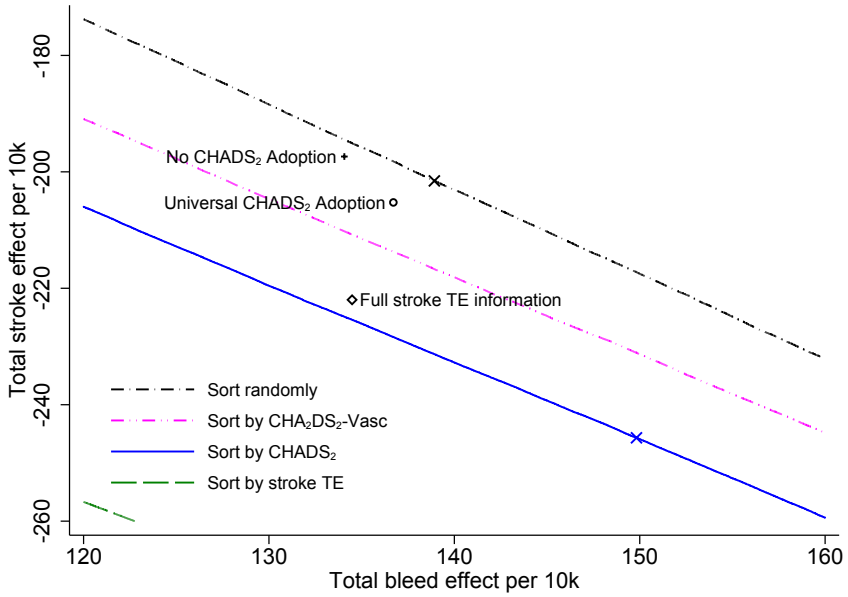
Notes: These graphs illustrate how key results of our structural model vary as we include various sets of control variables in its estimation. Panel A examines the increase in the informativeness of physicians' beliefs about CHADS_2-related stroke treatment effects with CHADS_2 adoption, or $\lambda^{s(c)}_{\text{pre}}/\lambda^{s(c)}_{\text{post}}$. Panel B examines the proportion of variance in the latent variable that we can explain with observable characteristics (i.e., the complement of the share explained by $\sigma^2_{\varepsilon,g}$). In each panel, we include varying sets of patient characteristics in $f(X_i)$ in our structural model stated in Equation (13). We estimate the baseline specification, shown in Column 1 of Table 4. The solid line shows the mean value of the statistic among specifications with the indicated number of control sets; the top (bottom) dashed line shows the maximum (minimum) of the statistic. The control variables are detailed in Appendix Section A.4.

Figure A.5: Counterfactual Outcomes with CHA_2DS_2-VASc Guideline

A. Strict Guideline Adherence



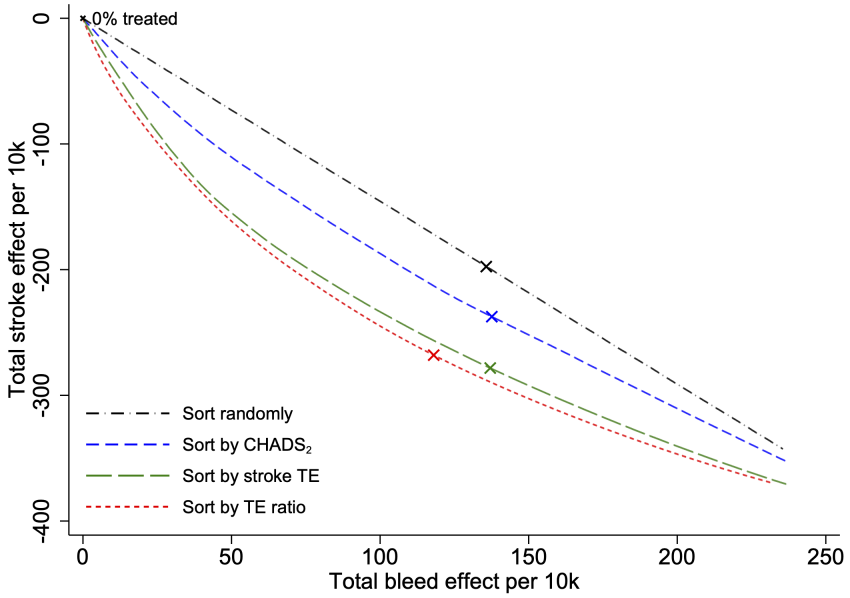
B. Guideline Adoption (Inset)



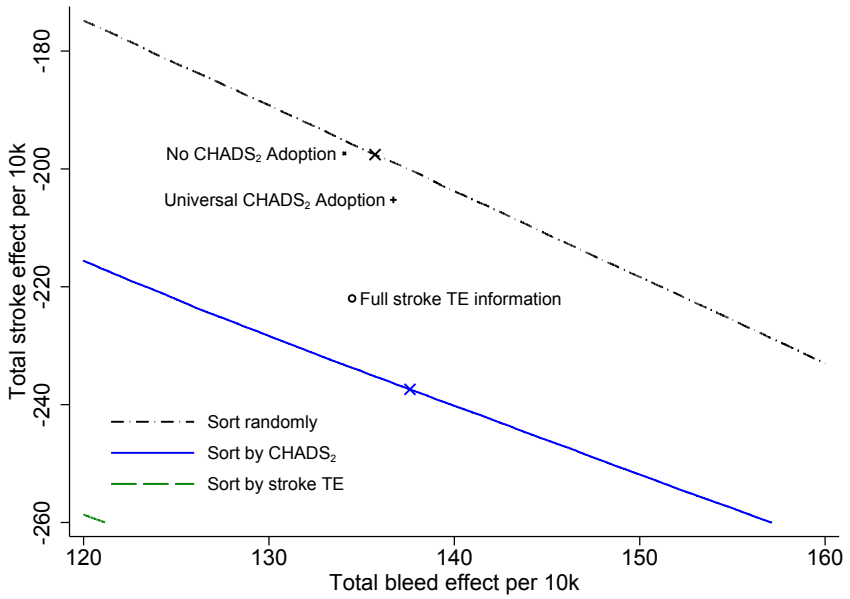
Notes: Relative to Figure 6, this figure includes an additional set of counterfactual outcomes under strict adherence to the CHA_2DS_2-VASc guideline. Like other counterfactuals of strict adherence, strict adherence to this guideline implies ranking patients by their CHA_2DS_2-VASc score. The CHA_2DS_2-VASc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

Figure A.6: Counterfactual Outcomes, Fixed Treated Age Distribution

A. Strict Guideline Adherence



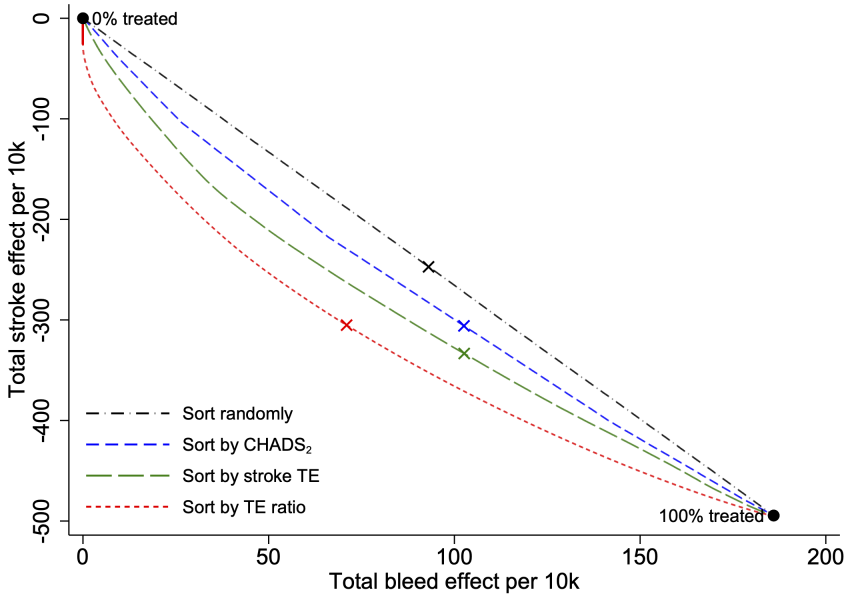
B. Guideline Adoption (Inset)



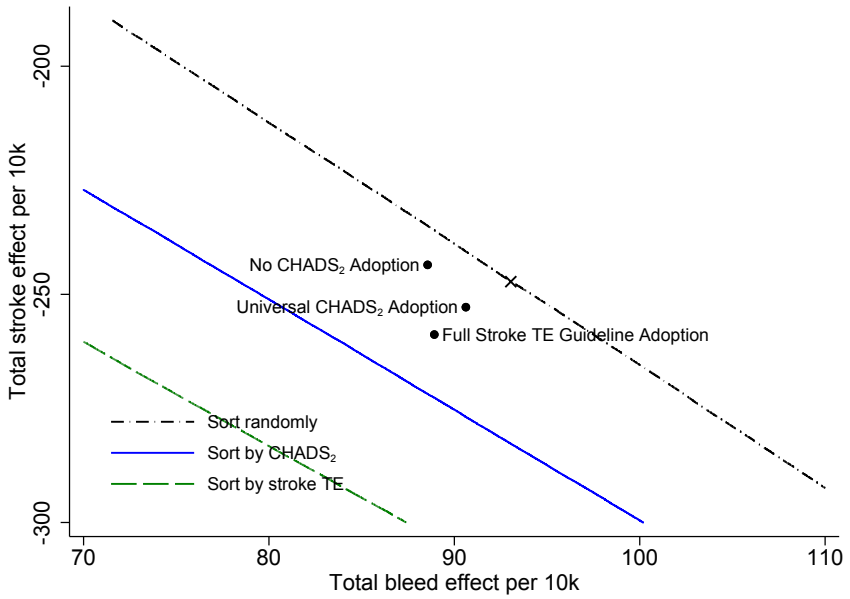
Notes: Relative to Figure 6, this figure shows counterfactual outcomes for patient rankings that hold fixed the age distribution of treated patients at every point along the curve. The fraction of treated patients in each 5-year age bin matches the fraction observed treated in our sample. Within each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

Figure A.7: Counterfactual Outcomes, Unadjusted Treatment Effects

A. Strict Guideline Adherence



B. Guideline Adoption (Inset)



Notes: Relative to Figure 6, this figure shows counterfactual outcomes that apply the raw stroke and bleed treatment effects from the causal forest instead of the best linear predictor (BLP) adjusted estimates. For this exercise, we bound stroke treatment effects above by 0, trimming the treatment effects for 0.06% of the sample with wrong-signed predictions of stroke treatment effects. We bound bleed treatment effects from below by 0, trimming the treatment effects for 6% of the sample with wrong-signed bleed treatment effects. While wrong-signed bleed treatment effects are more common when we do not apply the BLP adjustment, the point estimates are very close to zero, with the largest magnitude wrong-signed bleed treatment effect being −0.003. For more details, see notes to Figure 6.

Table A.1: AFI Database Covariates

Age
Sex
Height
Weight
Race
Smoker
Systolic blood pressure
Diastolic blood pressure
History of angina
History of congestive heart failure
History of diabetes
History of hypertension
History of myocardial infarction
History of TIA or stroke
Worst prior event: TIA or stroke
Time from last stroke or TIA

Notes: This table lists all the covariates from the AFI database that are used in our causal forest implementation. We set the treatment variable to be 1 for patients treated with warfarin and 0 for patients patients on control or ASA therapy (aspirin). Observations are dropped for those on low warfarin and low warfarin plus ASA therapy. Patients with especially relevant disease histories missing are also excluded from the sample. These disease histories include Transient Ischemic Attack (TIA), stroke, diabetes and hypertension. We regrouped the variable "Race" so that it equals 1 when the patient is white and 0 otherwise. We constructed two binary variables related to smoking history based on the variable "Smoker" in the AFI database, which takes 3 values. The new variables, current smoker and never smoker indicates whether the patient is a current smoker and whether he or she has ever smoked. Otherwise, missing variables are replaced by their mean and mode.

Table A.2: Variable Importance

Stroke Causal Forest	Stroke Regression Forest	Bleed Causal Forest	Bleed Regression Forest
Stroke Risk (−)	Past Stroke or TIA (+)	Bleed Risk (+)	Age (+)
Age (−)	$CHADS_2$ (+)	Age (+)	Race (−)
Systolic Blood Pressure (−)	White (−)	Race (+)	
Hemoglobin (+)	Systolic Blood Pressure (+)		
Past Stroke or TIA (−)	Age (+)		
Height (+)	Hemoglobin (−)		
$CHADS_2$ (−)	Height (−)		
White (−)	Past Diabetes (+)		
Past Angina (−)	Past Angina (+)		
Past Diabetes (−)	Past MI (+)		
Past MI (−)			

Notes: Each column of this table shows the important variables for each forest in descending order of importance. Only variables selected with the LASSO procedure are included for the forests. For bleeds, LASSO only selected age and race. Risks computed from regression forest using only the control sample are then used as an input into causal forests. The +/− signs following each variable indicates the sign of its coefficient in a bivariate linear regression with the causal forest output as the dependent variable.

Table A.3: Balance Table

Patient Characteristics	Control Group Mean	Treatment Group Mean	Coefficient
Age	70.9	70.7	0.134
			(0.269)
Congestive Heart Failure	0.27	0.32	-0.004
			(0.012)
Age above 65	0.76	0.77	0.003
			(0.012)
History of Hypertension	0.47	0.50	-0.007
			(0.014)
History of Stroke	0.19	0.15	-0.007
			(0.008)
History of Diabetes	0.15	0.15	-0.007
			(0.010)
Male	0.64	0.66	-0.015
			(0.013)

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

Table A.4: Causal Forest BLP Validation Regressions

	Dependent Variable		
	Stroke		Bleed
	(1)	(2)	(3)
Treatment, W_i	−0.043	−0.041	0.024
	(0.007)	(0.007)	(0.005)
Treatment effect interactions			
$W_i \times \hat{\tau}^o_{-j}(X_i)$	1.027		1.149
	(0.242)		(0.324)
$W_i \times \hat{\tau}^{s(c)}_{-j}(X_i)$		0.763	
		(0.340)	
$W_i \times \hat{\tau}^{s(r)}_{-j}(X_i)$		1.301	
		(0.348)	
Outcome mean	0.071	0.071	0.030
Observations	6,707	6,707	6,707
Trial fixed effects	Yes	Yes	Yes
Predicted outcome controls			
$\hat{Y}^o_{-j,1}(X_i)$, control group	Yes	Yes	Yes
$\hat{Y}^o_{-j,2}(X_i)$, control and treatment groups	Yes	Yes	Yes

Notes: This table reports the coefficients of best linear predictor (BLP) validation regressions of stroke and bleed outcomes on treatment W_i and interactions with treatment effects. Treatment effects are demeaned, so that the coefficient on the treatment indicator W_i reflects the average treatment effect. All specifications control for trial fixed effects and for regression forest predictions of the outcome, $\hat{Y}^o_{-j,1}(X_i)$, that are estimated in the control groups of leave-out trials as and analogous predictors $\hat{Y}^o_{-j,2}(X_i)$, that are estimated in both control and treatment groups of leave-out trials. Columns 1 and 3 corresponds to Equation (6), which interacts treatment with the full treatment effect, or $\hat{\tau}^o_{-j}(X_i)$; Column 2 corresponds to Equation (8), which interacts treatment with the $CHADS_2$ and the residual components of stroke treatment effects, or $\hat{\tau}^{s(c)}_{-j}(X_i)$ and $\hat{\tau}^{s(r)}_{-j}(X_i)$. Standard errors are shown in parentheses.

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

	VHA Data	AFI Database Trial							
		AFASAK1	BAATAF	CAFA	SPINAF	SPAF2	AFASAK2	EAFT Group 1	PATAF Group 1
Age	74.0	73.0	67.9	68.0	67.9	70.3	73.7	70.7	70.5
Stroke treatment effect, $\hat{\tau}^s_{BLP}$	−0.05	−0.05	−0.03	−0.03	−0.03	−0.04	−0.04	−0.08	−0.03
Bleed treatment effect, $\hat{\tau}^b_{BLP}$	0.03	0.03	0.02	0.02	0.02	0.02	0.03	0.02	0.02
Congestive heart failure	0.16	0.52	0.26	0.22	0.31	0.21	0.70	0.09	0.00
Hypertension	0.51	0.32	0.51	0.39	0.59	0.53	0.44	0.44	0.36
Age ≥ 65	0.84	0.84	0.67	0.68	0.69	0.73	0.90	0.81	0.79
Diabetes	0.15	0.08	0.15	0.12	0.19	0.16	0.12	0.13	0.17
Stroke history	0.18	0.06	0.03	0.03	0.08	0.06	0.05	0.76	0.00
Male	0.99	0.54	0.72	0.75	1.00	0.70	0.61	0.59	0.46
Number of patients	113,270	1,007	420	378	571	1,100	339	668	272

Note: This table shows the mean of each listed patient characteristic in the VHA data, as well as each of the eight trials with both control and treatment arms in the AFI database. There are a total of 10 trials in the AFI database. In three of the trials, patients are divided into eligible versus ineligible groups for anticoagulation and then randomized within each group. In the estimation of causal forest, we treat these groups as separate trials. AFASAK1: Atrial Fibrillation, Aspirin, and Anticoagulation Study 1; BAATAF: Boston Area Anticoagulation Trial for Atrial Fibrillation; CAFA: Canadian Atrial Fibrillation Anticoagulation; SPINAF: Stroke Prevention in Non-rheumatic Atrial Fibrillation; SPAF2: Stroke Prevention in atrial Fibrillation; AFASAK2: Atrial Fibrillation, Aspirin, and Anticoagulation Study 1; EAFT Group 1: European Atrial Fibrillation Trial; PAATAF Group 1: Primary Prevention of Arterial Thromboembolism in Atrial Fibrillation.

Table A.6: Probit Estimates: Additional Results

	Dependent Variable: Anticoagulant Prescription		
	(1)	(2)	(3)
Adoption status intercepts			
Post-adoption intercept, μ_{post}	−0.296***	−0.274***	−0.232***
	(0.046)	(0.048)	(0.052)
Never-adopter intercept, μ_{never}	0.015	0.016	0.023
	(0.038)	(0.038)	(0.041)
Standard deviation of error term			
Post-adoption, $\ln(\sigma_{\varepsilon,\text{post}})$	0.015	0.032	0.115
	(0.077)	(0.080)	(0.0868)
Never-adopter, $\ln(\sigma_{\varepsilon,\text{never}})$	0.012	0.013	0.0232
	(0.055)	(0.055)	(0.0576)
Year fixed effects, age spline controls	Yes	Yes	Yes
Differential trends on treatment effects	No	Yes	No
Treatment effects interacted with year identities	No	No	Yes
Number of observations	113,270	113,270	113,270

Notes: This table shows additional results from the specifications reported in Table 4. The probit models allow that $\sigma_{\varepsilon,g}$ may vary with adoption status (post-adoption and non-adopter) and year fixed effects. See notes for Table 4 for additional details. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table A.7: Probit Estimates: Average Marginal Effects

	Dependent Variable: Anticoagulant Prescription		
	(1)	(2)	(3)
$CHADS_2$-related stroke treatment effect, $\hat{\tau}^{s(c)}_{BLP}(x)$			
Prior to adoption, $\alpha^{s(c)}_{\text{pre}}$	−1.861***	−1.916***	−2.429***
	(0.210)	(0.228)	(0.387)
Post-adoption difference, $\alpha^{s(c)}_{\text{post}} - \alpha^{s(c)}_{\text{pre}}$	−2.404***	−2.369***	−1.903***
	(0.373)	(0.381)	(0.496)
Never-adopter difference, $\alpha^{s(c)}_{\text{never}} - \alpha^{s(c)}_{\text{pre}}$	0.372	0.355	0.440
	(0.282)	(0.294)	(0.506)
Residual stroke treatment effect, $\hat{\tau}^{s(r)}_{BLP}(x)$			
Prior to adoption, $\alpha^{s(r)}_{\text{pre}}$	0.241	0.177	0.153
	(0.121)	(0.128)	(0.698)
Post-adoption difference, $\alpha^{s(r)}_{\text{post}} - \alpha^{s(r)}_{\text{pre}}$	−0.366*	−0.305	−0.288
	(0.188)	(0.191)	(0.283)
Never-adopter difference, $\alpha^{s(r)}_{\text{never}} - \alpha^{s(r)}_{\text{pre}}$	0.017	0.056	0.142
	(0.160)	(0.165)	(0.347)
Bleed treatment effect, $\hat{\tau}^{b}_{BLP}(x)$			
Prior to adoption, α^{b}_{pre}	−1.294***	−1.226***	−1.111**
	(0.321)	(0.353)	(0.568)
Post-adoption difference, $\alpha^{s(r)}_{\text{post}} - \alpha^{s(r)}_{\text{pre}}$	0.974*	0.875	0.700
	(0.517)	(0.541)	(0.695)
Never-adopter difference, $\alpha^{b}_{\text{never}} - \alpha^{b}_{\text{pre}}$	0.421	0.357	0.113
	(0.416)	(0.436)	(0.739)
Year fixed effects, age spline controls	Yes	Yes	Yes
Differential trends on treatment effects	No	Yes	No
Treatment effects interacted with year identities	No	No	Yes
Number of observations	113,270	113,270	113,270

Notes: This table reports average marginal effects from the specifications reported in Table 4. See notes for Table 4 for additional details. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table A.8: Counterfactual Outcomes, Fixed Treated Age Distribution

	Treated Patients	Strokes Prevented	Bleeds Induced
A: Benchmarks (repeated from Table 5)			
Status quo	49.8%	199	134
Randomly assigned treatment	49.8%	196	135
B: Strict Guideline Adherence, Fixed Treated Age Distribution			
Strict $CHADS_2$ adherence	48.9%	234	134
Strict adherence to stroke TE guideline	49.1%	275	134
Strict adherence to TE-ratio guideline	55.6%	287	134

Notes: This table reports counterfactual outcomes that hold the fixed the age distribution of treated patients. The fraction of treated patients in each 5-year age bin matches the fraction observed treated in our sample. We also hold the overall percentage of treated patients fixed at 49.8%. Within each age group, patients are treated according to rankings implied by the noted guideline. Figure A.6 shows counterfactual outcomes varying the overall percentage of treated patients. For more details, see notes to Table 5 and Figure A.6.

Table A.9: Counterfactual Outcomes, Unadjusted Treatment Effects

	Treated Patients	Strokes Prevented	Bleeds Induced
A: Benchmarks			
Observed treatment choices	49.8%	245	89
Randomly assigned treatment	49.8%	246	93
B: Guideline Adoption			
No $CHADS_2$ adoption	49.7%	244	89
Universal $CHADS_2$ adoption	50.2%	252	91
Universal adoption of stroke TE guideline	49.8%	256	89
C: Strict Guideline Adherence, Fixed Bleeds Induced			
Strict $CHADS_2$ adherence	43.1%	273	89
Strict adherence to stroke TE guideline	43.4%	304	89
Strict adherence to TE-ratio guideline	58.5%	344	89

Notes: This table reports counterfactual outcomes that apply the raw stroke and bleed treatment effects from the causal forest rather than the best linear predictor (BLP) adjusted estimates. For this exercise, we bound stroke treatment effects above by 0, trimming the treatment effects for 0.06% of the sample with wrong-signed predictions of stroke treatment effects. We bound bleed treatment effects from below by 0, trimming the treatment effects for 6% of the sample with wrong-signed bleed treatment effects. While wrong-signed bleed treatment effects are more common when we do not apply the BLP adjustment, the point estimates are very close to zero, with the largest magnitude wrong-signed bleed treatment effect being -0.003. For more details, see notes to Table 5 and Figure A.7.