Personalized Medicine and Patient Selection: Discretion vs. Guidelines

Jason Abaluck, Leila Agha and David Chan

PRELIMINARY AND INCOMPLETE [*]

December 7, 2016

Abstract

Efficient care requires that care be administered to patients who stand to benefit most. Clinical risk scores are an increasingly common tool to achieve this. However, scores typically summarize baseline risk absent treatment while clinicians would like to know treatment effects. We make use of a large database of detailed clinical records from the Veterans Health Administration to study anticoagulation treatment choices and outcomes for patients with atrial fibrillation. We first document that adoption of a popular guideline (the CHADS2 score) reshaped prescription patterns. Next, we estimate how anticoagulation affects stroke and hemorrhage across patients with different characteristics. We use variation generated by quasi-random assignment of patients to physicians with different propensities to prescribe anticoagulation to estimate a model which allows heterogeneous treatment effects to vary flexibly with observable characteristics and allows for the possibility that physicians are selecting patients into treatment based on unobservable treatment effect heterogeneity. We use the model to compare patient outcomes under the status quo, under strict adherence to standard risk scores and optimal risk scores, and under discretionary adherence to standard and optimal risk scores. Optimal guidelines have the potential to reduce stroke incidence by two thirds relative to current practice, without increasing the rate of adverse side effects.

[*] Abaluck: Yale University and NBER, jason.abaluck@yale.edu. Agha: Dartmouth College and NBER, leila.agha@dartmouth.edu. Chan: Stanford University and NBER, david.chan@stanford.edu. Thanks to Brian Abaluck, Joe Altonji, Joshua Aronson, Danielle Li, Adam Sacarny, Chris Walters as well as seminar participants at Wharton, the National Bureau of Economic Research, the National Tax Association annual meeting, and Yale. Funding for this work was provided by NIA Grant Number T32-AG0000186 to the NBER.

1 Introduction

Recent advances in machine learning and genetics, as well as the widespread adoption of electronic medical records, make possible more personalized assessments of the benefits of alternative treatments (Collins and Varmus 2015). In at least one setting, the benefits of better allocating a fixed number of diagnostic tests given patients' medical history were found to be five times as large as the benefits from ordering the right number of tests for the population in question (Abaluck et al. 2014). In other words, the traditional question of medical researchers - "which patients will benefit from treatment?" - may have greater welfare consequences than the traditional question of health economists - "do coinsurances or reimbursement lead physicians to treat too little or too much?" But efforts to personalize medicine face a fundamental challenge: existing randomized experiments are not powered to uncover heterogeneity in treatment effects and attempts to do so using observational data are confounded by selection into treatment based on unobservable determinants of outcomes.

We present a framework for estimating heterogeneous treatment effects given selection of patients into treatment by combining machine learning methods with an explicit selection model identified based on quasi-random assignment of patients to physicians. Current efforts at tailoring treatment plans center on the application of evidence-based care guidelines. We first show that popular existing guidelines for the use of anticoagulants to prevent stroke among atrial fibrillation patients impact treatment decisions. We then use the machine learning causal effects framework to determine how these existing guidelines impact health outcomes and to construct optimal guidelines. Finally, we conduct simulations comparing the status quo to strict adherence to existing and optimal guidelines as well as discretionary adherence in which physicians integrate information not observable in medical records into their treatment decisions.

Physicians typically construct guidelines by examining which of a small number of clinically relevant factors best predict risk among untreated patients (e.g. Gage et al. (2001)). There are three problems with this approach: first, treatment effects need not be proportional to risk, and the patients who stand to benefit the most from treatment are not necessarily those with the highest ex ante risk. If warfarin reduces one patient's stroke probability from 50% to 49% and a second patient's stroke probability from 40% to 10%, the second patient benefits more. Second, a large number of covariates are often observed with an exponentially larger number of possible interactions between covariates, and existing procedures to select a subset of covariates to include in guidelines are often ad hoc. Third, untreated patients are a selected sample in two respects. The relationship between patient characteristics and risk among untreated patients will not yield a consistent estimator of patient risk absent treatment if those same characteristics also impact the treatment decision. Additionally, treatment effects may themselves vary with the propensity to treat, e.g. the first 10% of patients whom doctors decide to treat with a given set of observables may benefit greatly while the last 10% may not benefit.

We attempt to solve all three of these problems. We identify treatment effects separately from risk by building on the jackknifed instrumental variables strategy used by Aizer and Doyle (2013) and Kling (2006) which exploit random assignment of defendents to judges to estimate the impact of sentencing. We argue that in the Veterans Health Administration, patients are as good as randomly assigned to physicians and present several balance tests which validate this assumption. Given random assignment, we can estimate marginal returns by comparing outcomes among physicians with different treatment intensities. Physician A treats 30/100 patients with a given set of comorbidities; physician B treats 20/100 patients. Assuming A and B would agree about which patients should definitely be tested (an assumption we further investigate), comparing outcomes between physicians A and B identifies the treatment effect among the marginal patients treated only by physician A.

Unlike Aizer and Doyle (2013) and Kling (2006), we want to estimate how the benefits of treatment vary flexibly with a large vector of observables. To do so, we extend the method developed in Athey and Imbens (2015) for estimating treatment effects using machine learning techniques under conditional random assignment to our setting with instrumental variables.

It may still be the case that untreated patients with a given set of observables differ in important ways from treated patients: perhaps some risk visible to the doctors (but not visible to the econometrician) makes the untreated patients ill-suited for treatment. To allow for this, we combine the above methods with a Roy model approach to estimating treatment effects developed by Heckman and Vytlacil (2005) and applied to the problem of estimating treatment effects in healthcare by Chandra and Staiger (2011) and Abaluck et al. (2014).

The approach we develop can also be contrasted with the approach used by operations researchers and computer scientists to mine observational data in order to determine which patients should be treated. A typical approach in those literatures is to use sophisticated machine learning methods to match patients on observable characteristics and then estimate treatment effects by comparing outcomes among treated and untreated patients (e.g. Bertsimas et al. (2016)). This approach will fail to the extent that physicians also select patients into treatment based on unobservables that are correlated with outcomes. Our model can be thought of as supplementing this approach with a selection correction. We continue to use the variation that comes from comparing treated and untreated patients for a given set of characteristics, but we explicitly adjust for the degree to which treated patients are unobservably different.

Our framework can also be applied to a wide variety of settings outside of medical care. The problems of an employer deciding which applicants to hire to maximize productivity, a bank deciding which consumers to loan to in order to minimize defaults or colleges deciding which applicants to admit to maximize a known objective all share the underlying features that we seek to address. One would like to estimate heterogeneous treatment effects by comparing treated and untreated beneficiaries, but to do so one must adjust properly for selection.

The paper proceeds as follows. Section 2 lays out the medical context and data; section 3 provides results of our reduced form analysis; section 4 describes our econometric model; section 5 reports our structural estimation results.

2 Clinical context and data

In this paper, we examine the decision to prescribe anticoagulation for patients with atrial fibrillation. Atrial fibrillation is the most common type of heart arrhythmia; the heart beats irregularly, causing palpitations, shortness of breath and weakness. Atrial fibrillation is a leading cause of stroke, associated with over 450,000 hospitalizations and 99,000 deaths each year (Ott et al. 1997). To treat patients with atrial fibrillation, physicians must weigh the benefit of reduced stroke risk from anticoagulation against the cost of increased hemorrhage risk. Anticoagulation guidelines recommend treatment for patients with the highest risk of stroke as estimated by popular risk scores (Camm et al. 2012; You et al. 2012; January et al. 2014), namely the CHADS2 score which was first published in 2001, and the CHA2DS2-VASc score which was first published in 2009 (Gage et al. 2001; Go et al. 2003; Lip et al. 2010). Table 1 describes the CHADS2 risk score and associated practice guidelines.

Recent evidence suggests that only 40% of atrial fibrillation patients at high risk of stroke are treated in accordance with the widely accepted risk score based guidelines for atrial fibrillation (Glazer et al. 2007). Yet, whether these deviations from guidelines lead to adverse patient outcomes is unclear for three key reasons.

First, the risk score guidelines may not place the optimal weights on each risk factor. The CHADS2 and CHA2DS2-VASc are formulated to predict which patients are at high risk for stroke; they do not necessarily predict how causal effects of anticoagulation vary across patients. If treatment effects are not in constant proportion to stroke risk, then these guidelines may not target the patients with the greatest benefit from anticoagulation. Differences between the stroke risk among a physicians specific patient population and the often specially selected study population may also lead the weights on popular risk scores to be an imperfect basis for care decisions.

Second, guidelines for atrial fibrillation are based solely on stroke risk and fail to consider the competing risk of hemorrhage, while clinical practice must consider both outcomes to optimize care.

Finally, simplistic risk scores omit myriad clinical and social factors that may affect a patients expected benefit from treatment; doctors may apply information not in risk scores better tailor their care. Some of these judgments could be codified with more detailed guidelines, whereas others may be difficult to consistently describe or extract and don't lend themselves to easy inclusion in standardized guidelines.

This study relies on electronic health records from the Veteran's Health Administration (VHA) to construct a detailed database of patients diagnosed with atrial fibrillation. We have collected data on each patients clinical risk factors for stroke and bleeding, anticoagulation choice (warfarin or alternative treatment), and clinical outcomes (including incidence of stroke and head bleed). For each physician in our sample, we will develop a case history of his experiences treating atrial fibrillation. We will also identify groups of doctors who practice within the same clinical location. These data are collected from the VA Corporate Data Warehouse, which includes patient scheduling, inpatient and outpatient visits, prescriptions, laboratory tests, diagnoses, and demographics. The data span the years 2000-2014.

To identify patients with a new diagnosis of atrial fibrillation, we require patients to have atrial fibrillation recorded on two separate visits at least 30 days but no more than 365 days apart. Using both the inpatient and outpatient encounters, we record each patients related health outcomes: stroke, intracranial hemorrhage, gastrointestinal hemorrhage, and falls (which increase the risk of intracranial hemorrhage). Our sample includes over 396,000 patients newly diagnosed with atrial fibrillation in the VHA over a fifteen year period. This large sample size will yield greater statistical power than is typically attainable in a randomized trial to investigate how treatment effects vary across patients, depending on their risk factors.

We assign patients to the prescribing doctor by studying the first visit with a primary care provider following diagnosis with atrial fibrillation. We assume that the doctor responsible for this visit is making the decision about whether to prescribe anticoagulants. To capture whether or not anticoagulation is prescribed, we use prescription drug files derived from billing data. Prescription drug files allow us to capture patients other drugs as well to account for the role of drug interactions in physician decisions and patient outcomes.

We extract a detailed accounting of patients medical histories and risk factors related to atrial fibrillation, stroke and hemorrhage. Using a three year history of all clinical encounters with the VHA system, we will capture patients active medical conditions by the ICD9 diagnosis codes associated with each visit. Key clinical considerations for prescribing anticoagulation for atrial fibrillation include conditions in the CHADS2 score: congestive heart failure, hypertension, age over 75 years, diabetes, and stroke. We augment these measures with additional clinical factors, including the Elixhauser comorbidity set and additional variables identified from the medical literature as relevant to stroke or bleed risk including fall risk, history of bleeds, and vision problems.

By using the rich electronic medical record data from the VA Corporate Data Warehouse, we can go beyond coding diagnosis history through ICD9 codes. In particular, we capture the results of common lab tests for INR and platelet levels which may be important determinants of a patients suitability for anticoagulation. We can also construct measures of patient compliance by measuring the consistency with which the patient keeps previous appointments and continues to refill prescriptions as appropriate.

A limitation of our data set is that while we observe a record of all care delivered within the VHA system, we do not capture care delivered outside the VHA. We address this limitation in a few ways. Finally, we link VHA records to two sources of claims data to capture care delivered outside the VHA system. We use Fee Basis data to identify care paid for by the VHA but delivered elsewhere. Finally, we link Medicare claims data to observe outcomes and diagnoses from care performed outside the VHA, for the subset of the patient population above age 65.

Summary statistics are reported in Table 2, by CHADS2 score. Patients with CHADS2 scores of 0 or 1 have the lowest average stroke rate and the lowest rates of warfarin prescription. Stroke rate is increasing monotonically in CHADS2 score. Only 1.6% of patients scoring 0 have a stroke within 6 months of diagnosis; strokes are ten times more frequent at the highest CHADS2 score, with 16.6% of patients experiencing a stroke within 6 months. Notably, while warfarin is prescribed less often to patients scoring 0 or 1, prescription rates do not increase monotonically with CHADS2 score. Very high scoring patients are less likely to be prescribed warfarin than lower scoring patients.

One possible explanation for this pattern can be seen in the average bleed rates also reported in this table. The bleed rates increase with CHADS2 score, ranging from 2.5%

of low score patients and increasing to 9.5% of patients with a maximum score of 6. While some of the increases in observed bleed rate might be explained by the higher rates of warfarin prescription (since warfarin increases the risk of bleeds), we see bleed rates increasing even over CHADS2 score ranges where the warfarin prescription rate is decreasing, e.g. from a score CHADS2 of 3 to 4 or from 5 to 6. It is possible that even though these high scoring patients are at higher risk of stroke, they are also at greater risk of this risky adverse event—bleeds—which limits prescription rates among the highest risk patients. We will explore these patterns in more detail after estimating heterogeneous causal treatment effects with our selection model.

3 Does the CHADS2 Score Impact Behavior?

Over our study period, growing awareness of the guidelines appears to have re-shaped physician's treatment of low risk patients. Popular guidelines recommend prescription anticoagulants such as warfarin for all patients with a CHADS2 score greater than two. Discretion is advised for patients with lower CHADS2 scores; the guidelines suggest that prescription anticoagulants are not required for many of these lower risk patients.

We begin by exploring trends in warfarin use. In Figure 1, we observe that for the first several years of our study, warfarin prescription rates were similar in levels and trending along the same path for patients with high and low CHADS2 scores. Beginning in 2008, prescription patterns for high and low risk patients diverge. There was a decline in warfarin use for patients with low CHADS2 scores (and so low estimated risk of stroke) relative to the trend for high score patients. These trends provide our first evidence that CHADS2 diffusion may have led to reduced use of anticoagulation for low risk patients.

Figure 2 displays trends in mentions of the CHADS2 score in physician notes. Concomitant with the decrease in warfarin use for low scoring patients documented in Figure 1, we also find an uptick in mentions of the CHADS2 score beginning around 2008. Although it was first published in 2001, the CHADS2 score appears to have been slow to diffuse within the VHA. In 2002, 27% of doctors had mentioned the CHADS2 score at least once in their clinical notes; by 2014 this rate had more than doubled, reaching 63%. Perhaps more tellingly, the share of primary care visits for atrial fibrillation patients mentioning the CHADS2 score increased from less than 1% to over 12%.

Figure 1 and Figure 2 provide preliminary evidence suggesting that the CHADS2 score impacted prescribing behavior. In Figure 1, we see that warfarin prescriptions for patients with a CHADS2 score of 0/1 fell dramatically relative to warfarin prescriptions for patients CHADS2 scores greater than 2 and that this fall precisely tracks the increase in CHADS2 score use recorded in Figure 2.

We can also explicitly examine how prescription behavior changes after a physician mentions the CHADS2 score in his notes for the first time, using a difference-in-differences framework. Specifically, we test whether prescription patterns change differentially for high risk patients relative to low risk patients, following the doctor's first CHADS2 mention. We assume that in the absence of a new awareness of the CHADS2 score, trends in prescription rates would be the same for high and low score patients, conditional on doctor fixed effects, flexible time fixed effects (for each year-month and day of the week) as well as our rich set of patient controls.

We construct a vector of relative year dummies indicating the year relative to the doctor's first mention of the CHADS2 score in their free-text physician notes for an atrial fibrilation patient. Year 0 is the calendar year that includes the doctor's first CHADS2 mention; negative years mark years prior to that mention; positive relative years mark years afterwards. We then model the warfarin prescription decision for patient i treated by doctor d at time t and relative year r as:

$$\begin{aligned} Warfarin_{idrt} = \sum_r \quad & \beta_r^1 \quad (CHADS2 >= 2)_i \\ + \quad & \sum_r \beta_r^2 + \alpha_t + \gamma_d + x_{id}\delta + \epsilon_{idt} \end{aligned}$$

The regression controls for year-month fixed effects, day of the week fixed effects, doctor fixed effects, and patient covariates (including all of the inputs into the CHADS2 score).

Results of this reduced form analysis are reported in Figure 3. For those patients for whom physicians note the CHADS2 score, Warfarin prescriptions are about 7 percentage points higher for patients with CHADS2 scores greater than 2 relative to patients with low CHADS2 scores. This provides further confirmation that physicians respond to new guidelines, incorporating them into their clinical practice and making consequential changes to their treatment choices.

4 Treatment Effects and Guidelines

In this section, we estimate a model of treatment effects which we will use to assess how strict adherence to existing and optimal guidelines would impact stroke rates for a given number of bleeds. We also allow physicians to select into treatment those patients with the highest expected return given unobservables. As a result, we can use the model to ask how patients would be impacted given discretionary adherence to guidelines in which physicians weight observables appropriately but continue to use information in unobservables to shape treatment decisions.

4.1 Model of Treatment Effects

Our goal is ultimately to estimate heterogeneous treatment effects accounting for selection. A naive approach would assume no selection on unobservables and compare treated and untreated individuals for a given set of observables. We can instead estimate the degree of selection on unobservables by observing how outcomes vary across doctors with different propensities to treat patients with a given set of observables. Estimating the model will require assumptions about how the degree of selection on unobservables varies across patients with different observable characteristics.

We begin by laying out a version of the general model of treatment effects developed by Heckman and Vytlacil (2005) - this model is without loss of generality. We then make explicit the assumptions we will use to avoid the dimensionality problems presented by the more general model. This is a generalization of the model used for treatment effect estimation in healthcare by Chandra and Staiger (2011) and Abaluck et al. (2014).

We start with a standard potential outcomes framework. Denote by $Y^o_{idc}(0)$ the outcome if patient i assigned to doctor d in clinic c is untreated and by $Y^o_{idc}(1)$ the outcome if patient i is treated. In our setting, the outcome is $o \in \{Bleed, Stroke\}$.

We start by writing:

$$\begin{aligned} Y^o_{idc}(0) &= f^o(x_{idc}) + \eta^o_{0idc} \qquad (1) \\ Y^o_{idc}(1) &= h^o(x_{idc}) + \eta^o_{1idc} \qquad (2) \end{aligned}$$

where $E(\eta^o_{0idc}|x_{idc}) = 0$ and $E(\eta^o_{1idc}|x_{idc}) = 0$. The average treatment effect is thus $g^o(x_{idc}) = h^o(x_{idc}) - f^o(x_{idc})$. This is the difference in outcomes if everyone with a given set of observables went from being untreated to being treated. Let $\Delta\eta^o_{idc} = \eta^o_{1idc} - \eta^o_{0idc}$.

Doctors treat if and only if:

$$B_{idc} = g^*(x_{idc}, z_{idc}) + \eta'_{idc} < 0 \qquad (3)$$

where $E(\eta'_{idc}|x_{idc}, z_{idc}) = 0)$ and z_{idc} are variables which impact the decision to treat but not outcomes, so $E(\eta^o_{0idc}|x_{idc}, z_{idc}) = E(\eta^o_{1idc}|x_{idc}, z_{idc}) = 0$.

Without further restrictions on $g^*(x_{idc}, z_{idc})$, this is a without loss of generality descriptive model of the treatment decision. If for example, $g^*(x_{idc}, z_{idc}) = g^s(x_{idc}) +$

$g^b(x_{idc})$ and $\eta'_{idc} = \Delta\eta^s_{idc} + \Delta\eta^b_{idc}$, then doctors would be minimizing the sum of strokes and bleeds. But we require no such assumption - $g^*(x_{idc}, z_{idc})$ and η'_{idc} can have an arbitrary relationship with the underlying treatment effects or be totally unrelated.

Then the probability of treatment is given by: $P(W_{idc} = 1|x_{idc}, z_{idc}) = P(B_{idc} < 0) = H_x(g^*(x_{idc}, z_{idc}))$ where the function H is a monotonically increasing function which depends on the distribution of η'_{idc} (which in turn can vary arbitrarily with all observable characteristics, including interactions with doctor and clinic). Then we can write: $g^*(x_{idc}, z_{idc}) = H_x^{-1}(P(W_{idc} = 1|x_{idc}, z_{idc})$. Let $\lambda_x^{o,+}(P(W_{idc} = 1|x_{idc}, z_{idc}) = E(\eta^o_{1idc}|H_x^{-1}(P(W_{idc} = 1|x_{idc}, z_{idc}) + \eta'_{idc} < 0)$. The expected outcome among treated patients is given by:

$$E(Y^o_{idc}(1)|W_{idc} = 1) = h^o(x_{idc}) + \lambda_x^{o,+}\left(P(W_{idc} = 1|x_{idc}, z_{idc})\right) \tag{4}$$

$h^0(x_{id})$ tells us the average likelihood of a stroke if we treated all patients with a given set of observables ($\lambda_x^{o,+}(1) = 0$ since $E(\eta_{1idc}|x_{idc}, z_{idc}) = 0$. The slope of $\lambda_x^{o,+}(\cdot)$ is undetermined. When physicians treat more patients with a given set of observables, we might see fewer strokes if the first patients treated are highest ex ante risk or more strokes if patients have similar ex ante risks and the first patients treated benefit the most from treatment. Additionally, doctors may fail to order patients based on treatment effects (η'_{idc} may bear no relationship to $\Delta\eta_{idc}$).

Analogously, the expected outcome among untreated patients is given by:

$$E(Y^o_{idc}(0)|W_{idc} = 0) = f^o(x_{id}) + \lambda_x^{o,-}\left(P(W_{idc} = 1|x_{idc}, z_{idc})\right) \tag{5}$$

where $\lambda_x^{o,-}(P(W_{idc} = 1|x_{idc}, z_{idc}) = E(\eta^o_{0idc}|H_x^{-1}(P(W_{idc} = 1|x_{idc}, z_{idc})) + \eta'_{idc} \geq 0)$.

Now, $f^o(x_{id})$ gives us the average outcome if we left all patients untreated (formally, when no patients are treated, $\lambda^{o,-}(0) = 0$ because $E(\eta_{0idc}|x_{idc}, z_{idc}) = 0)$. The slope of $\lambda^{o,-}(\cdot)$ is once again indeterminate. Even if doctors appropriately order patients based on treatment effects, when more patients are treated, the remaining untreated patients might be lower risk, or they may be high risk but impervious to treatment.

Provided we observed sufficient variation in z_{idc} such that the probability of treatment ranged from 0 to 1 for every set of x_{idc}, we could estimate the five equations above (equations 3 and equations 4 and 5 for both strokes and bleeds) and this would recover all treatment effects of interest. We could use these equations to determine how many strokes and bleeds would occur for patients with any given set of observable characteristics x_{idc} as the probability of treatment for those patients went from 0 to 1 given the way that doctors currently select on unobservables.

To do so nonparametrically would require an unrealistic amount of data - we would need to observe enough variation in z_{idc} so that the probability of treatment ranged from 0 to 1 *for every set of patient characteristics for every doctor in our sample.* In our application, the instrument z_{idc} will be constructed at the physician level so we do not observe any within physician variation. To estimate the degree of selection on unobservables using variation across physicians, we will need to restrict the degree to which selection on unobservables varies across physicians (we also restrict how it varies across patients to avoid the dimensionality problem that arises if we need to estimate separate $\lambda^{0,+}(\cdot)$ and $\lambda^{o,-}(\cdot)$ for every set of patient characteristics).

We make the following assumption:

Assumption 1. *$\lambda^{o,+}_{x_{idc}} = \lambda^{o,+}_{A(x_i)}$ and $\lambda^{o,-}_{x_{idc}} = \lambda^{o,-}_{A(x_i)}$ where $A(x_i)$ is a known function of patient characteristics.*

In structural terms, this is an assumption about the distribution of η'_{idc}, the unobservable term in the decision to treat equation. We are assuming that η'_{idc} has the same distribution for any set of observables such that $A(x_i)$ takes the same value. In our baseline estimates, $A(x_i)$ gives quantiles of stroke risk conditional on x_i. Thus, we are assuming that patients with similar observable stroke risk have a similar distribution of unobservable characteristics.

Because we do not observe a large number of patients with every set of patient characteristics at every clinic let alone every physician, we assume that physician and clinic fixed effects are additively separable in the treatment effects as well as the decision to treat. Since patients are randomly assigned to physicians within clinics, the physician fixed effects can be omitted in the outcomes equation (the variables z_{id} allows the treatment decision to vary with the fact that different physicians have different propensities to treat identical patients). Thus, we assume:

Assumption 2. $f^o(x_{idc}) = f^o(x_i) + \theta^{fo}_c$
$g^0(x_{ic}) = g^o(x_i) + \theta^{go}_c$
$g^*(x_{idc}, z_{idc}) = g^*(x_i, z_{id}) + \theta^{g^*}_c$

Finally, to facilitate estimation, we assume that the error term in the treatment equation is i.i.d. and uniformly distributed so that we have:

Assumption 3. $P(W_{idc} = 1 | x_{idc}, z_{idc}) = \hat{g}^*(x_{id}, z_{id}) + \hat{\theta}_c$

where the hats reflect the fact that $\hat{g}$ and $\hat{\theta}$ are linear transformations of the underlying structural parameters. Were it not for the assumption that $\hat{\theta}_c$ is additively separable, this assumption would be without loss of generality - changing the distribution of η'_{idc} would be equivalent to changing the shape of the function $g^*(\cdot)$. Because of the additive separability assumption, the uniform assumption is substantive - the relationship between patient characteristics and treatment probabilities is assumed to be the same for all clinics and to vary across clinics only by an additive constant.

4.2 Identification

Identifying the functions $\lambda^{o,+}(\cdot)$ and $\lambda^{o,-}(\cdot)$ requires constructing instruments z_{idc} which impact the probability of treatment but have no other direct impact on outcomes. Intuitively, the perfect experiment is one in which we take patients with a given set of observables, sequence them based on physicians' unobservable assessment of suitability for treatment, and then examine how outcomes vary if we randomly treat different fractions of patients: how does the number of strokes vary among treated and untreated patients as we treat a greater fraction of patients?

The intuition for our instrumental variables strategy mirrors this thought experiment. We construct instruments using each physicians' propensity to administer warfarin among all other atrial fibrillation patients treated by that physician. Given the assumption of conditional random assignment of patients to physicians within clinics and assuming monotonicity (a common ranking of patients based on unobservables), we can identify treatment effects for each set of observables by comparing outcomes for physicians who treat more or less within a given clinic.

Let $d(i)$ denote all of the patients physician d considers other than patient i. A simple approach, following Aizer and Doyle (2013), would be to construct:

$$Z_{idc} = \frac{1}{n_{d(i)}} \sum_{k \in d(i)} \tilde{W}_{kdc} \tag{6}$$

where $\tilde{W}_{kdc} = W_{kdc} - \hat{g}(x_{idc})$, the residual variation in Warfarin administration after subtracting out the variation due to observables.

This instrument is inefficient because it does not appropriately account for the fact that this propensity is estimated much more precisely for some physicians than others because they have a greater number of patients, and it does not account for drift over time. In Appendix A, we describe a more involved procedure to construct an efficient estimator. This procedure constructs each physician's propensities to treat in each period (the jackknifed propensity in the current period) and then constructs a weighted average of those propensities based on the covariance between each alternative period and the current period as well as the number of observations used in each period. This alternative procedure increases the nR^2 of our instrument by a factor of 2.

The validity of our estimation strategy requires that patients do not sort differentially within a clinical site to doctors with different prescibing propensities. The institutional setting of the VHA supports this assumption; patients are assigned to physicians according to strict rotations and it is rare that care is transferred across doctors by patient request or due to differential expertise within a clinical specialty.

We can also probe this assumption empirically by testing for differential sorting on the basis of observable patient risk factors. To test for covariate balance, we regress the value of our jackknife prescription propensity on a vector of patient characteristics, including age, race, veteran status, and comorbidities included in the CHADS2 score. The regression controls for year dummies, and day of week by hour by clinic site fixed effects. Our identification argument is that patients are randomly assigned to doctors within a clinical site, conditional on their scheduling preferences. By controlling for day of week and hour of day at each clinical site, we are effectively removing variation in patient sorting to doctors that is related to patient and doctor schedule availability. Within a particular time slot, e.g. Mondays between 10-11 AM, and a clinical site, e.g. the primary care practice at the Palo Alto VA Center, we assume that patients are as good as randomly assigned. We exploit variation in the prescription propensities of doctors prescribing at that site and schedule slot to identify the effect of warfarin prescription on patient outcomes.

Table 2 presents balance tests for this instrument. We can see that observable characteristics are extremely well-balanced across quantiles of the instrument, consistent with our story that, conditional on clinic and date fixed effects, patients are randomly assigned to physicians.

4.3 Estimation

To estimate the model we proceed as follows. First, we estimate the selection into treatment equation, equation 3, which gives estimates of $P(W_{idc}|x_{idc}, z_{idc})$. Second, we estimate equations 4 and 5 given these estimates.

Consider first estimation of $P(W_{idc}|x_{idc}, z_{idc})$. To estimate this, we proceed as follows. First, we demean at the clinic level and estimate $g^*(\cdot)$ using the two-step LASSO procedure described in Belloni, Chernozhukov, and Hansen (2014) including all quadratic functions and interactions of the x_{idc} variables as well as z_{idc}. Finally, we estimate the clinic fixed effects fixing the remaining coefficients in the model.

The second step in our estimation process is to estimate equations 4 and 5 given

$\hat{P}(W_{idc}|x_{idc}, z_{idc})$. We do so using an analogous procedure. First, we demean at the clinic level to eliminate the θ_c variables. Next, we use the LASSO procedure in Belloni, Chernozhukov, and Hansen (2014) to estimate $f(x_{id}, \hat{P}(W_{idc}|x_{idc}, c, d))$. Finally, given estimates of $f(\cdot)$, we estimate the clinic fixed effects as the means of $Y^o_{idc} - \hat{f}(\cdot)$.

5 Treatment Effect Estimates and Simulation Results

5.1 Heterogeneous Treatment Effects

In our first set of results, we explore the estimated treatment effects of our selection model outlined above and contrast them with two alternative estimation procedures. In the first set of alternative estimates, we assume warfarin causes a 60% relative risk reduction from the patient's baseline stroke risk absent treatment, since the randomized clinical trials find that warfarin causes a 60% reduction in stroke risk on average. This assumption of constant relative treatment effects underlies the current medical guidelines which recommend treatment to patients with high CHADS2 scores: the CHADS2 score was formulated to predict stroke risk among untreated patients and guidelines implicitly assume that warfarin treatment effects are proportional to stroke risk.

The second alternative procedure reports naive LASSO estimates of the treatment effect which do not correct for selection at all. These estimates apply machine learning to select variables that interact with the treatment variable and to select control variables. In the absence of selection on unobserved patient characteristics, these estimates should also recover estimates of heterogeneous causal treatment effects. This procedure mimics the approach of the operations research literature which applies machine learning to estimate treatment effects without accounting for selection on unobservables (e.g. Bertsimas et al. (2016)).

Figure 4 splits patients up into 10 deciles based on their predicted stroke risk absent treatment (as estimated by our selection model). The figure then displays the average treatment effect by decile of stroke risk as estimated by each of the three methodologies described above: the structural model of selection, assumption of constant 60% relative reduction in stroke, and the machine learning estimate without selection correction.

The figure reveals that the average treatment effect estimated with our selection model track closely to the estimates derived assuming a constant 60% reduction in stroke. This provides initial evidence that the basic premise of the CHADS2 score guideline, although previously untested to our knowledge, may have been a reasonable starting point. Stroke risk appears to predict the size of the average treatment effect of warfarin. However, it is important to note two caveats to interpreting this graph as validation of the CHADS2 guideline. The first caveat is that the figure obscures considerable heterogeneity in treatment effects within each bin of predicted stroke risk, which an optimal treatment rule would consider. The second caveat is that we have not yet considered bleeding treatment effects, which depending on their correlation with stroke treatment effects, could lead to a different optimal treatment rule.

Further, it is notable that the average treatment effect estimated by our selection model match closely to a simpler jackknife instrumental variables approach that uses equation 6 to define the instrumental variable, and estimates the local average treatment effect of warfarin within a standard instrumental variables framework. In the fourth quartile of predicted stroke risk, the basic instrumental variable model estimates that warfarin reduces stroke by 13 percentage points (standard error of 5.6 percentage points), which is very close to the average treatment effect estimated in the top two deciles of stroke risk by the machine learning selection model in Figure 4 (which predict stroke reduction by 10.7 p.p. and 12.5 p.p., respectively).

By contrast, the machine learning estimates that do not account for selection consistently underestimate the treatment effects at all deciles of stroke risk except the very lowest decile. The direction of bias is consistent with a basic selection story whereby patients at higher risk of stroke are more likely to be treated with warfarin. This comparison highlights the importance of accounting for selection within models that aim to estimate heterogeneous treatment effects. Even when conditioning on a rich vector of patient covariates drawn from a detailed electronic medical record, selection on unobservables remains an important confounder for estimating causal treatment effects. Applying estimates from the model without selection would lead to systematic errors in treatment decisions, in particular suggesting suboptimally low levels of anticoagulation.

Figure 5 reports results from a similar exercise, now comparing the predicted average treatment effect of warfarin on bleed risk across the three estimation strategies. The selection model and naive machine learning estimation procedures remain similar to the stroke case; the only difference is that bleed has been substituted as the outcome variable. To contrast with the perspective of the medical guidelines, we assume a 2.5% increase in bleed risk, taking the estimate from the randomized clinical trial and assuming it is constant by decile of bleed risk.

The selection model finds evidence of substantial heterogeneity in bleed risk, suggesting the importance of trading off heterogeneous effects of warfarin on bleed and stroke as they vary with patient characteristics. While warfarin is estimated to increase bleeds for patients at every decile of bleed risk, the average treatment effect is largest for patients in the lowest decile of bleed risk absent treatment. The naive machine learning estimates again consistently underestimate the size of the treatment effect, even suggesting that warfarin prescription reduces bleed risk for high risk patients. This runs contrary to medical teaching and the estimates of our selection model, and likely results from physicians selecting patients for treatment with warfarin who have lower bleed risk due to factors unobservable by the econometrician.

Again, we uncover a close match between the local average treatment effect implied by a basic instrumental variables estimate strategy and the average treatment effects estimated in this model. The basic instrumental variable estimation of the LATE finds that warfarin increases the risk of bleeds by 3 percentage points, which tracks closely to the average treatment effects of the machine learning selection model in the bottom two deciles.

5.2 Comparing estimated treatment effects to current guidelines

Next, we investigate how our estimated treatment effects correlate to the recommendations of CHADS2 score guidelines. Table 4 reports average treatment effects estimated with our machine learning model with selection correction by CHADS2 score. First, we note that warfarin is estimated to reduce stroke risk by a larger amount at higher CHADS2 scores. Again, this general pattern would at first seem to validate the CHADS2 approach to guideline construction–patients at higher stroke risk as predicted by CHADS2 score have the greatest benefits (in terms of stroke reduction) from warfarin treatment. However, considering the bleed average treatment effects tempers this conclusion considerably. While CHADS2 score does indeed predict stroke treatment effects, it also appears highly predictive of bleed treatment effect. Since the CHADS2 score jointly predicts both the benefit (reduced stroke) and the cost (increased bleed) of taking warfarin, its value as a treatment algorithm may be limited. Ideal treatment guidelines need to balance stroke reduction with bleed increases.

Finally, Table 5 compares the variables that are predictive of stroke treatment effects in our machine learning framework with the variables included in the CHADS2 and its successor, the CHADS2-VASc. The LASSO procedure retains the age and stroke history variables as predictive of stroke treatment effects, which were also in the original CHADS2 guideline. It also includes vascular disease, which was added to the guideline with the CHADS2-VASc amendment, suggesting this adjustment to the original guideline may have been helpful in targeting warfarin to patients with the largest stroke reduction. Our model also identifies four variables that are excluded from the CHADS2 and CHADS2-VASc but are predictive of stroke treatment effect–race, renal failure, fall risk, and neurological disorders. From this analysis alone, it is not clear if the variables identified by our model are appropriate for treatment guidelines, because we have not yet considered how they predict bleed effects.

To extend this exercise, we repeat now estimate our LASSO selection model with the outcome variable that calculates the combined stroke treatment effect plus the bleed treatment effect. A treatment guideline maximizing this outcome would minimize the combined incidence of strokes and bleeds, and would be optimal if the welfare cost of stroke and bleeds were equal. While this is a strong assumption which we will not impose in the subsequent welfare analysis, it allows some initial insight into the guideline construction process.

Variables that are the strongest predictors of treatment effect for stroke and bleed events include congestive heart failure and stroke history, both of which were included in the CHADS2 guidelines. Importantly, the sign of the congestive heart failure variable actually reverses in our model relative to the CHADS2 guideline. The CHADS2 guideline suggests that patients with congestive heart failure should be more likely to receive warfarin because they are likely to experience a larger decrease in stroke rate with treatment. However, our estimates suggest that because congestive heart failure is such a strong predictor of bleed treatment effects as well, these patients should be less likely to receive warfarin if the aim of treatment assignment is to minimize stroke and bleed events (with equal weight on each). The LASSO selection model also suggests four new variables which could improve treatment targeting further: bleed history, tumor, chronic pulmonary disease, which should all increase treatment likelihood and neurological disorders, which should reduce the likelihood of treatment.

5.3 Guideline simulations

In this section, we use our estimates of heterogeneous treatment effects from the machine learning selection model to consider a number of counterfactuals. Specifically, we want to compare the stroke and bleed rates that would be associated with status quo treatment decisions and contrast them with outcomes under strict adherence to the CHADS2 score and with treatment decisions that follow the optimal strict guideline based on our estimated treatment effects. In ongoing work, we are applying our selection model to analyze the benefits of discretionary adherence to guidelines, where physicians receive guidance of the sort, "half of patients with these comorbidities should be treated," and physicians use discretion to select which individual patients within that set to treat on the basis of unobserved characteristics.

To construct the optimal strict guideline, we consider minimizing the number of strokes subject to a constraint that holds the total number of bleeds constant at the rate currently observed in our sample. We call the guideline “strict” because it assigns warfarin to every patient or no patients within a given observable cell based on patient characteristics. It can be shown that the solution to this optimization problem is achieved by assigning warfarin to patients with the largest ratio of stroke average treatment effects divided by bleed average treatment effects, and continuing to assign warfarin to patients with a declining value of that ratio until the expected bleed rate equals the status quo bleed rate.

We apply cross validation techniques to avoid overfitting and thus overstating the benefits of adherence to the optimal strict guidelines. Specifically we identify the treatment rule given results estimated on half of our sample, which has been randomly assigned as the ”training” data set. We then evaluate the benefits of guideline adherence by predicting stroke and bleed outcomes using treatment effects estimating on the other ”test” half of our sample.

When considering strict CHADS2 adherence, we consider the counterfactual whereby physicians treat all patients starting with the highest CHADS2 score and moving to lower scores, stopping when the expected bleed rate equals the current observed bleed rate.

Finally, when considering status quo treatment decisions, we maintain treatment decisions observed in our sample, and compare actual stroke and bleed rates to the stroke and bleed rates predicted from applying our treatment effects estimated in the training data to predict outcomes in the test data.

Results of these simulations are reported in Table 6. We first note that simulated stroke and bleed rates based on estimated treatment effects match the observed rates in the test data set extremely closely, providing some validation of the model results.

Next, holding bleeds constant at the rate observed under status quo treatment decisions (as predicted by estimates in the training data), we investigate counterfactual stroke rates under strict CHADS2 adherence and optimal strict guideline adherence. We estimate that strict adherence to the CHADS2 score, treating all of the highest score patients with warfarin, could decrease the observed stroke rate from 4.4% to 2.7%.

Adherence to the optimal strict guideline could reduce strokes more dramatically to 1.6%, almost a third of the observed stroke rate. These estimates suggest potentially large gains to guideline improvements which optimally trade off stroke and bleed risk to better tailor treatment decisions.

6 Conclusion

This paper develops a new methodology for estimating heterogeneous returns to treatment, applying machine learning to a model that accounts for selection on unobservables. This methodology can be applied for development of new guidelines that perform substantially better than current guidelines or current observed treatment decisions in counterfactual simulations.

The new estimates of treatment effects can outperform existing guidelines for two reasons. The first is that these new estimates better identify variables that predict treatment effect heterogeneity rather than simply predicting baseline risk (and assuming treatment effects are proportional to risk). The second is that they can trade off bleed and stroke treatment effects and target treatment to patients with the largest benefits in terms of stroke reduction and smallest relative risk in terms of bleed increase.

Current efforts at applying machine learning to medical applications frequently fail to account for selection into treatment on the basis of unobserved factors. Even with a rich set of covariates observed from the patient's medical record, we demonstrate that failing to account for selection leads to systematic understatement of treatment effects in our setting.

We apply our model to understand the tradeoffs between allowing physician discretion and requiring strict guideline adherence. We observe that strict adherence to either the CHADS2 or the optimal strict guideline could reduce stroke rates without increasing the rate of bleed events. Of course, discretionary adherence to an optimal guideline could outperform both of these approaches. In ongoing work, we explore the value of discretion in the context of optimal guidelines.

Our selection model is identified by a jackknife instrumental variable approach which relies on monotonicity assumptions for identification. Specifically, we assume that physicians agree on the sorting of patients by suitability for treatment and differ only in their treatment intensity. In ongoing work, we are exploring the validity of this assumption by testing for monotonicity in subgroups identified by patient covariates. We can also explore relaxations of this assumption by allowing the selection function to vary across physician groups.

Our estimation also makes use of functional form restrictions for tractability of estimation, and these could also contribute to identification. While in principle, nonparametric identification is possible, even in our large sample of patients we do not have sufficient power for a completely nonparametric approach. However, in ongoing work, we are testing the impact of relaxing these functional form assumptions including additive separability of the instrument and uniformity of the error term.

References

Abaluck, J., L. Agha, C. Kabrhel, A. Raja, and A. Venkatesh (2014). Negative tests and the efficiency of medical care: What determines heterogeneity in imaging behavior? Technical report, National Bureau of Economic Research.

Aizer, A. and J. J. Doyle (2013). Juvenile incarceration, human capital and future crime: Evidence from randomly-assigned judges. *The Quarterly Journal of Economics 130*(2), 759–803.

Athey, S. and G. Imbens (2015). Machine learning methods for estimating heterogeneous causal effects. *arXiv preprint arXiv:1504.01132*.

Belloni, A., V. Chernozhukov, and C. Hansen (2014). High-dimensional methods and inference on structural and treatment effects. *The Journal of Economic Perspectives 28*(2), 29–50.

Bertsimas, D., N. Kallus, A. Weinstein, and D. Zhuo (2016). Personalized diabetes management using electronic medical records.

Camm, A. J., G. Y. Lip, R. De Caterina, I. Savelieva, D. Atar, S. H. Hohnloser, G. Hindricks, P. Kirchhof, J. J. Bax, H. Baumgartner, et al. (2012). 2012 focused update of the esc guidelines for the management of atrial fibrillation. *European heart journal*, ehs253.

Chandra, A. and D. Staiger (2011). Expertise, Overuse and Underuse in Healthcare. *Working Paper*.

Collins, F. S. and H. Varmus (2015). A new initiative on precision medicine. *New England Journal of Medicine 372*(9), 793–795.

Gage, B. F., A. D. Waterman, W. Shannon, M. Boechler, M. W. Rich, and M. J. Radford (2001). Validation of clinical classification schemes for predicting stroke: results from the national registry of atrial fibrillation. *Jama 285*(22), 2864–2870.

Glazer, N. L., S. Dublin, N. L. Smith, B. French, L. A. Jackson, J. B. Hrachovec, D. S. Siscovick, B. M. Psaty, and S. R. Heckbert (2007). Newly detected atrial fibrillation and compliance with antithrombotic guidelines. *Archives of Internal Medicine 167*(3), 246–252.

Go, A. S., E. M. Hylek, Y. Chang, K. A. Phillips, L. E. Henault, A. M. Capra, N. G. Jensvold, J. V. Selby, and D. E. Singer (2003). Anticoagulation therapy for stroke prevention in atrial fibrillation: how well do randomized trials translate into clinical practice? *Jama 290*(20), 2685–2692.

Heckman, J. J. and E. Vytlacil (2005). Structural equations, treatment effects, and econometric policy evaluation1. *Econometrica 73*(3), 669–738.

January, C. T., L. S. Wann, J. S. Alpert, H. Calkins, J. E. Cigarroa, J. B. Conti, P. T. Ellinor, M. D. Ezekowitz, M. E. Field, K. T. Murray, et al. (2014). 2014 aha/acc/hrs guideline for the management of patients with atrial fibrillation: a report of the american college of cardiology/american heart association task force on practice guidelines and the heart rhythm society. *Journal of the American College of Cardiology 64*(21), e1–e76.

Kling, J. R. (2006). Incarceration length, employment, and earnings. *The American Economic Review 96*(3), pp. 863–876.

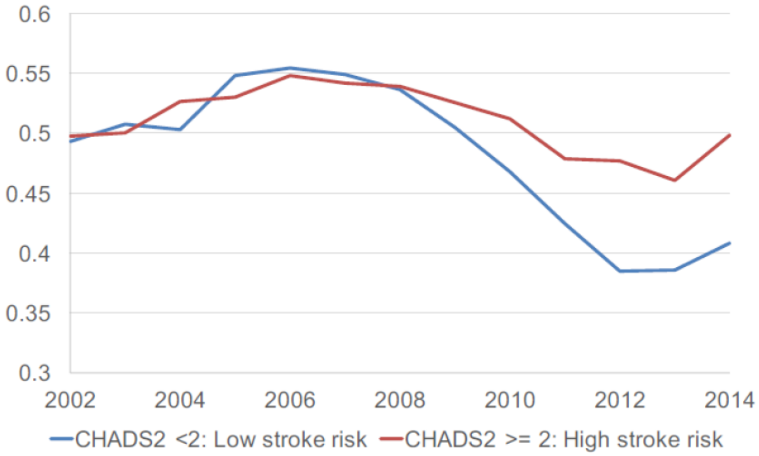
Lip, G. Y., R. Nieuwlaat, R. Pisters, D. A. Lane, and H. J. Crijns (2010). Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the euro heart survey on atrial fibrillation. *Chest Journal 137*(2), 263–272.

Ott, A., M. M. Breteler, M. C. de Bruyne, F. van Harskamp, D. E. Grobbee, and A. Hofman (1997). Atrial fibrillation and dementia in a population-based study the rotterdam study. *Stroke 28*(2), 316–321.

You, J. J., D. E. Singer, P. A. Howard, D. A. Lane, M. H. Eckman, M. C. Fang, E. M. Hylek, S. Schulman, A. S. Go, M. Hughes, et al. (2012). Antithrombotic therapy for atrial fibrillation: antithrombotic therapy and prevention of thrombosis: American college of chest physicians evidence-based clinical practice guidelines. *CHEST Journal 141*(2_suppl), e531S–e575S.

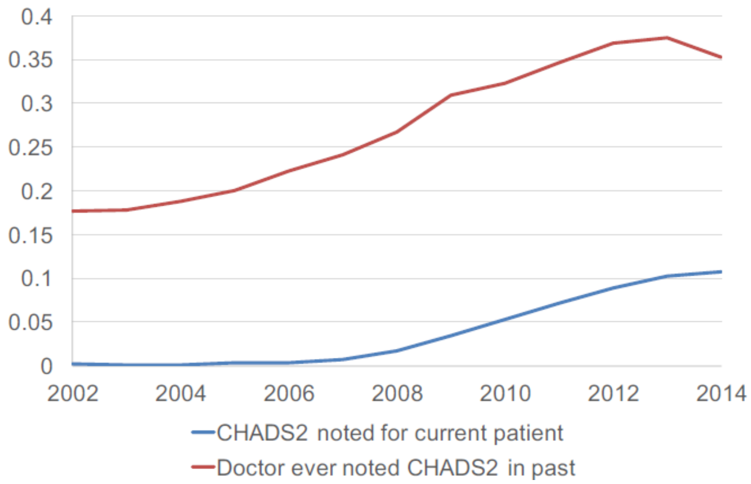
7 Tables and Figures

Figure 1: Trends in warfarin prescription rates over time by CHADS2 score



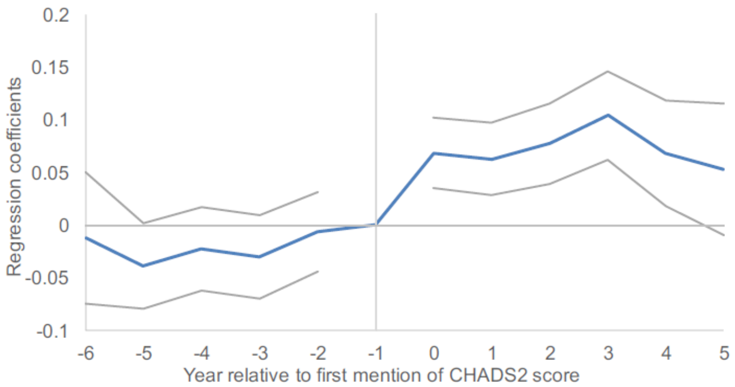
Notes: this figure plots rates of warfarin prescription by CHADS2 score over time. CHADS2 scores below 2 predict low stroke risk, and CHADS2 scores of 2 or greater predicted elevated risk of stroke. Data is on 400,000 patients with atrial fibrillation treated by the Veterans Health Administration.

Figure 2: Diffusion of CHADS2 score over time



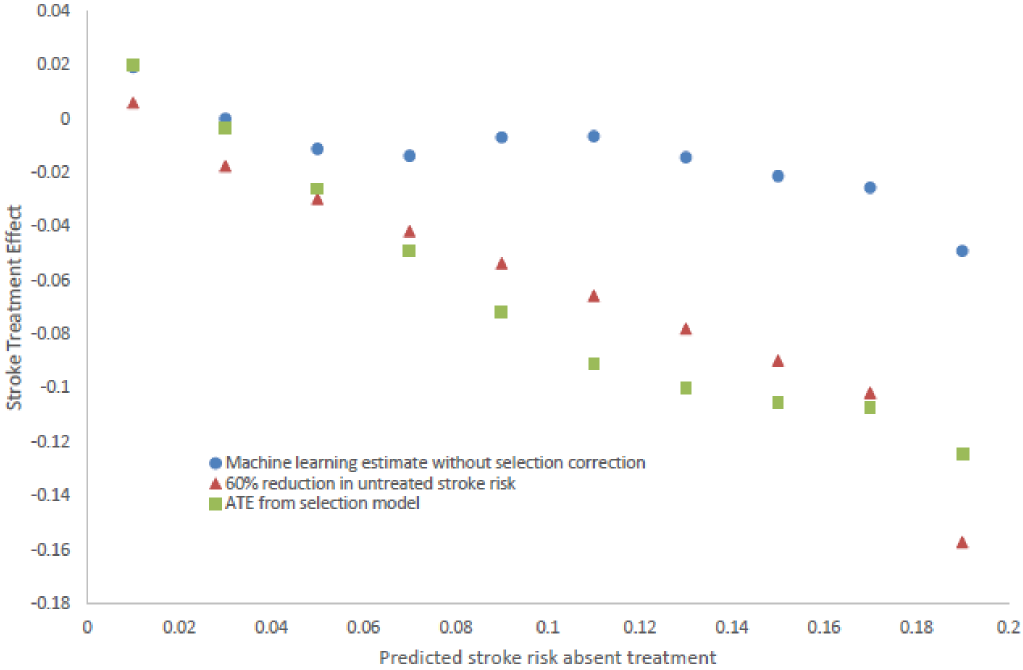
Notes: This figure plots rates of CHADS2 adoption over time. The red series displays the fraction of patients treated by a physician who has ever mentioned the CHADS2 score in his note. The blue series displays the fraction of patients for whom the CHADS2 score is mentioned in the note associated with his own visit for atrial fibrillation.

Figure 3: Impact of CHADS2 adoption on warfarin prescription rates:
Comparison of high score to low score patients



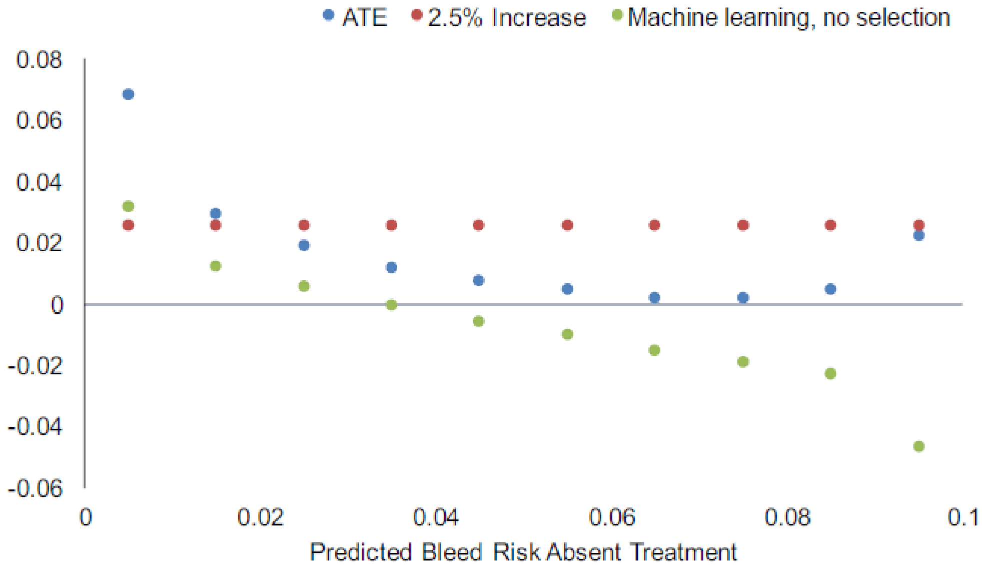
Notes: This figure plots coefficient estimates from a modified version of equation 1. The regression controls for year-month fixed effects, doctor fixed effects, patient covariates, and year relative to first mention of CHADS2 adoption. The plotted coefficients are the interaction of relative year fixed effects and a dummy variable for patient CHADS2 score of 2 or greater. The plot shows how prescription patterns change for high CHADS2 score patients relative to low score patients following their doctor's first mention of the CHADS2 guidelines. Year 0 is the year of first CHADS2 mention. 95% confidence interval is shown in grey. Standard errors are clustered at the physician level.

Figure 4: Estimates of stroke average treatment effect by decile of stroke risk



Notes: This figure plots estimates of average stroke treatment effects (ATE) by decile of predicted stroke risk absent treatment. We report three different estimation strategies for stroke ATE. First, we report estimates from our selection model outline in the paper using green boxes. Next, we plot estimates assuming warfarin causes a 60% reduction in stroke risk following the logic of the medical CHADS2 guidelines using red triangles. Finally, we report estimates of stroke treatment effects estimated using LASSO results that do not account for selection on unobservables using blue circles.

Figure 5: Estimates of bleed average treatment effect by decile of bleed risk



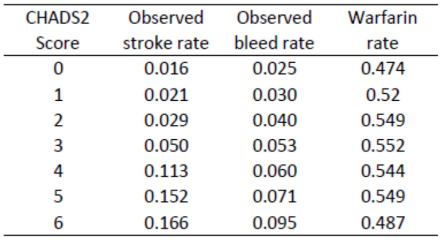
Notes: This figure plots estimates of average bleed treatment effects (ATE) by decile of predicted bleed risk absent treatment. We report three different estimation strategies for bleed ATE. First, we report estimates from our selection model outline in the paper using blue. Next, we plot estimates assuming warfarin causes a 60% reduction in stroke risk following the logic of the medical CHADS2 guidelines using red. Finally, we report estimates of stroke treatment effects estimated using LASSO results that do not account for selection on unobservables using green.

Table 1: CHADS2 score guidelines

CHADS2 Stroke Risk Score	
Congestive heart failure history	+1
Hypertension history	+1
Age ≥ 75 years	+1
Diabetes mellitus history	+1
Stroke or TIA symptoms previously	+2
Clinical recommendations	
Score of 2 or greater: high risk of stroke, oral anticoagulation recommended	
Score of 1: moderate risk of stroke, oral anticoagulation considered	
Score of 0: low risk of stroke, no anticoagulation recommended	

Notes: This table describes the CHADS2 score used to assess stroke risk among patients with atrial fibrillation. It also describes anticoagulation guidelines based on this risk score.

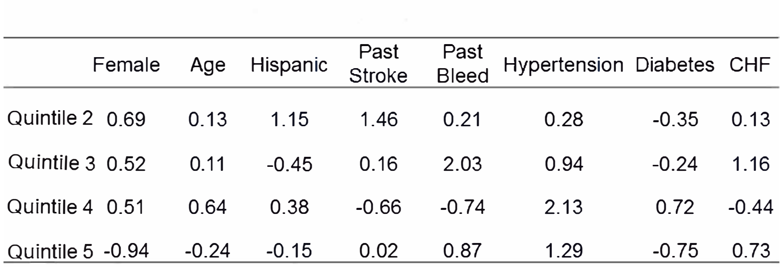
Table 2: Summary statistics by CHADS2 score



CHADS2 Score	Observed stroke rate	Observed bleed rate	Warfarin rate
0	0.016	0.025	0.474
1	0.021	0.030	0.52
2	0.029	0.040	0.549
3	0.050	0.053	0.552
4	0.113	0.060	0.544
5	0.152	0.071	0.549
6	0.166	0.095	0.487

Notes: This table summarizes the average stroke rate, bleed rate, and warfarin prescription rate by CHADS2 score in our sample.

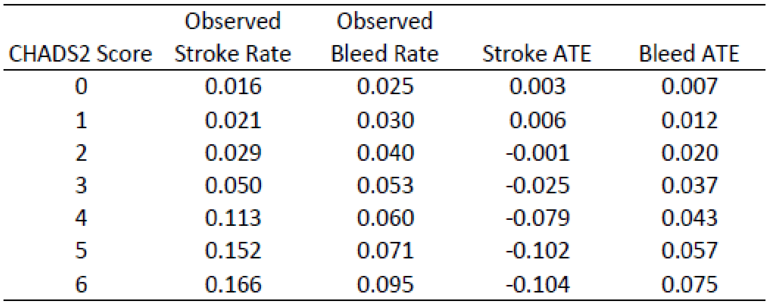
Table 3: Balance test: Regressions of patient characteristics on quintiles of IV



	Female	Age	Hispanic	Past Stroke	Past Bleed	Hypertension	Diabetes	CHF
Quintile 2	0.69	0.13	1.15	1.46	0.21	0.28	-0.35	0.13
Quintile 3	0.52	0.11	-0.45	0.16	2.03	0.94	-0.24	1.16
Quintile 4	0.51	0.64	0.38	-0.66	-0.74	2.13	0.72	-0.44
Quintile 5	-0.94	-0.24	-0.15	0.02	0.87	1.29	-0.75	0.73

Notes: This table reports regression results and standard errors (in parentheses) from eight separate regressions. The independent variable of interest is residual variation in patient characteristics (e.g. female, age, hispanic, etc.) after partialling out the clinic by year fixed effects and day of week by hour of day fixed effects. Independent variables are a series of dummy variables for each quintile of physician prescribing propensity.

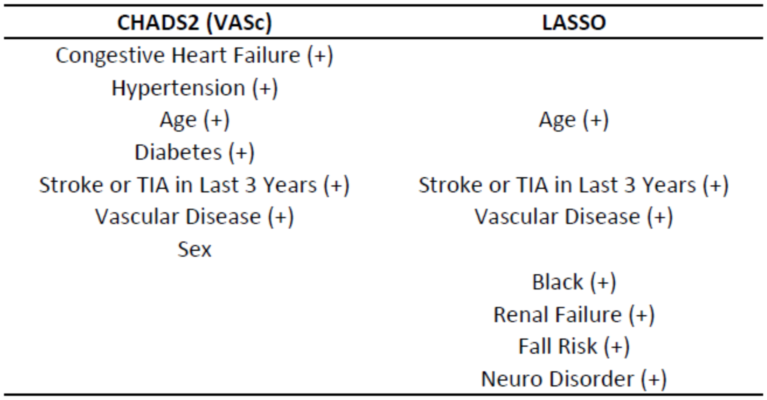
Table 4: Estimated average treatment effects by patient CHADS2 score



CHADS2 Score	Observed Stroke Rate	Observed Bleed Rate	Stroke ATE	Bleed ATE
0	0.016	0.025	0.003	0.007
1	0.021	0.030	0.006	0.012
2	0.029	0.040	-0.001	0.020
3	0.050	0.053	-0.025	0.037
4	0.113	0.060	-0.079	0.043
5	0.152	0.071	-0.102	0.057
6	0.166	0.095	-0.104	0.075

Notes: This table reports the average treatment effects from our machine learning model with selection correction, for in sample patients with each CHADS2 score.

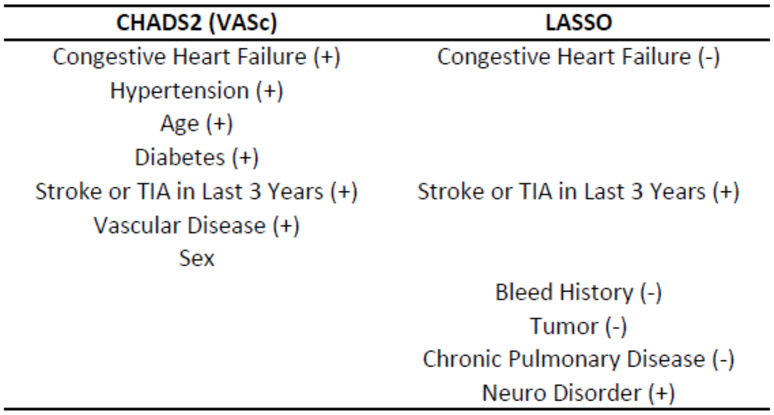
Table 5: LASSO variables that predict stroke treatment effects



CHADS2 (VASc)	LASSO
Congestive Heart Failure (+)	
Hypertension (+)	
Age (+)	Age (+)
Diabetes (+)	
Stroke or TIA in Last 3 Years (+)	Stroke or TIA in Last 3 Years (+)
Vascular Disease (+)	Vascular Disease (+)
Sex	
	Black (+)
	Renal Failure (+)
	Fall Risk (+)
	Neuro Disorder (+)

Notes: This table reports the variables that predict treatment effects for stroke outcomes in our LASSO estimation with selection correction.

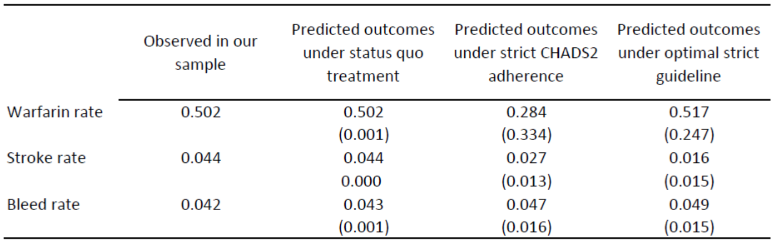
Table 6: LASSO variables that predict $|\ stroke\ | - bleed$ treatment effects



CHADS2 (VASc)	LASSO
Congestive Heart Failure (+)	Congestive Heart Failure (-)
Hypertension (+)	
Age (+)	
Diabetes (+)	
Stroke or TIA in Last 3 Years (+)	Stroke or TIA in Last 3 Years (+)
Vascular Disease (+)	
Sex	
	Bleed History (-)
	Tumor (-)
	Chronic Pulmonary Disease (-)
	Neuro Disorder (+)

Notes: This table reports the variables that predict treatment effects for stroke plus bleed outcomes in our LASSO estimation with selection correction.

Table 7: Simulated counterfactual treatment decisions



	Observed in our sample	Predicted outcomes under status quo treatment	Predicted outcomes under strict CHADS2 adherence	Predicted outcomes under optimal strict guideline
Warfarin rate	0.502	0.502	0.284	0.517
		(0.001)	(0.334)	(0.247)
Stroke rate	0.044	0.044	0.027	0.016
		0.000	(0.013)	(0.015)
Bleed rate	0.042	0.043	0.047	0.049
		(0.001)	(0.016)	(0.015)

Notes: Column 1 reports the observed rates of warfarin prescription, stroke and bleed in our testing sample. Column 2 uses the observed prescription decisions to estimate the stroke and bleed rate in the testing sample, based on the model estimated in the training sample. Column 3 simulates outcomes if warfarin were assigned to highest CHADS2 score patients first, until the estimated bleed rate equals the observed bleed rate. Column 4 simulates outcomes if warfarin were assigned to patients according to the optimal strict guideline that minimizes strokes while holding the bleed rate at the current levels.

A Empirical Bayes Jackknife Instrument

Suppose $Y_{idt} \sim_{i.i.d} Bernoulli(x_{id}\beta + \delta_c + \theta\mu_d \cdot x_{id} + p_{dt})$. Note that we can also rewrite this assumption as $Y_{idt} = x_{id}\beta + \delta_c + \theta\mu_d \cdot x_{id} + p_{dt} + e_{idt}$ where $E(e_{idt}|x_{dt}, \delta_c, p_{dt}) = 0$ and $e_{idt} = 1 - (x_{id}\beta + \delta_c + p_{dt})$ with probability $x_{id}\beta + \delta_c + p_{dt}$ and $-(x_{id}\beta + \delta_c + p_{dt})$ with probability $1 - (x_{id}\beta + \delta_c + p_{dt})$. Note that μ_d are doctor-level observables. When we regress outcomes on Warfarin use, we include as instruments our estimates of p_{dt} and $\theta\mu_d \cdot x_{id}$ and we include as controls clinic fixed effects as well as all the x_{id}. We assume further that e_{idt} are uncorrelated across time or across different physicians (in other words, conditional on the stroke probability at a given doctor time, and conditional on covariates, the fact that a given patient happens to have a stroke at time s has no further bearing on whether some other patient has a stroke at time t).

Assume that $Cov_d(p_{d(t+s)}, p_{dt}) = \sigma_{ps}$ where the subscript d denotes the empirical covariance of these parameters across doctors. We can consistently estimate β in a fixed effects regression of Y_{id} on x_{id} and fixed effects for p_{dt}. I assume below that β is precisely estimated so we can ignore any noise in $\hat{\beta}$. I will however allow for the possibility of imprecisely estimated clinic fixed effects. Define: $\tilde{Y}_{idt} \equiv Y_{idt} - x_{id}\beta - \theta\mu_d \cdot x_{id} = \delta_c + p_{dt} + e_{idt}$. I assume without loss of generality that the mean of p_{dt} within each clinic is 0.

If we estimate this regression by OLS, then our estimates of p_{dt} will be given by: $\hat{p}_{dt} = \frac{1}{N_{dt}} \sum_{i \in d(t)} (\tilde{Y}_{idt} - \bar{Y}_c)$ where $d(t)$ denotes the set of doctor d's patients at time t and $\bar{Y}_c = \frac{1}{N_c} \sum_{i \in c} \tilde{Y}_{idt}$. Note that we can write this as:

$$\begin{aligned}
\hat{p}_{dt} &= \frac{1}{N_{dt}} \sum_{i \in d(t)} (\tilde{Y}_{idt} - \bar{Y}_c) \\
&= \frac{1}{N_{dt}} \sum_{i \in d(t)} (\delta_c + p_{dt} + e_{idt} - (\delta_c + \frac{1}{N_c} \sum_{i \in c} e_{idt})) \\
&= p_{dt} + \frac{1}{N_{dt}} \sum_{i \in d(t)} (e_{idt} - \bar{e}_c) \qquad (7)
\end{aligned}$$

where $\bar{e}_c = \frac{1}{N_c} \sum_{i \in c} e_{idt}$.

In this model, we want to construct the best linear unbiased estimator of p_{dt} given $\hat{p}_{d(-t)}$ where $\hat{p}_{d(-t)} = (\hat{p}_{d1}, ..., \hat{p}_{d(t-1)})$. We write this estimator as:

$$\hat{p}_{dt} = \sum_{s=1}^{t-1} \psi_s(N_1, ..., N_{t-1}) \hat{p}_{ds} \qquad (8)$$

The weights ψ_s in each period are allowed to vary flexibly with the number of observations observed in all periods (you might give less weight to a prior period because you know that other periods are estimated with greater precision).

We will find ψ_s as the solution to the problem of minimizing:

$$\psi = \arg \min_{\{\psi_1,...,\psi_{(t-1)}\}} \sum_d \left(p_{dt} - \psi_0(N_{d1},...,N_{d(t-1)}) - \sum_{s=1}^{t-1} \psi_s(N_{d1},...,N_{d(t-1)})\hat{p}_{ds} \right)^2 \quad (9)$$

As in Chetty et. al., the resulting coefficients ψ are equivalent to those obtained from an OLS regression of $\hat{p}_{dt}$ on $\hat{p}_{d(-t)}$ identified using across doctor variation. If we observed a large number of doctors for every possible set of $N_d = \{N_{d1},...,N_{d(t-1)}\}$, then we could run this regression to recover the relevant coefficients. In lieu of that, we can make use of the underlying structural assumptions in the model to compute the regression coefficients. These are given by $\psi = \Sigma_p^{-1}\gamma$ where $\gamma = (Cov_d(p_{dt}, \hat{p}_{d1}|N_d),...,Cov_d(p_{dt}, \hat{p}_{d(t-1)}|N_d))'$ and Σ_p is the across-doctor covariance matrix of $\hat{p}_{d(-t)}$ (again conditional on N_d). The subscript d is used to indicate that this covariance is empirical covariance across doctors as opposed to the covariance of the random variables. This distinction will be important below for the variance terms (for example, the across doctor variance of p_{dt} is positive, whereas the variance of the "random variable" p_{dt} is 0 since p_{dt} is a constant).

Firstly, note that for $t \neq s$, $Cov_d(p_{dt}, \hat{p}_{ds}|N_{dt}, t) = E_d(p_{dt}\hat{p}_{ds}|N_{dt}, t) - E_d(p_{dt}|N_{dt}, t)E_d(\hat{p}_{ds}|N_{dt}, t) = \sigma_{p(t-s)}$. This quantity does not depend on N_{dt}, and it only depends on $t-s$, not t or s individually. We estimate this using the covariance of all $\tilde{Y}_{idt}$ and $\hat{p}_{ds}$ for all t and s that are the appropriate number of years apart and for which doctor d has at least 2 observations in year t. I denote this covariance by $Cov_i(\tilde{Y}_{idt}, \hat{p}_{ds})$. Let N_{t-s} denote the number of patients whose doctors have at least 2 observations in the current year and at least 1 observation $t-s$ years earlier. Then we have:

$$\begin{aligned}
Cov_i(\tilde{Y}_{idt}, \hat{p}_{ds}) &= E_i(\tilde{Y}_{idt}\hat{p}_{ds}) - E_i(\tilde{Y}_{idt})E_i(\hat{p}_{ds}) \\
&= E_i((\delta_c + p_{dt} + e_{idt})(p_{ds} - \bar{e}_c + \frac{1}{N_{ds}} \sum_{j \in d(s)} e_{jds})) - E_i(\tilde{Y}_{idt})E_i(\hat{p}_{ds}) \\
&= E_i((\delta_c + p_{dt} + e_{idt})(p_{ds} - \bar{e}_c + \frac{1}{N_{ds}} \sum_{j \in d(s)} e_{jds})) - E_i(\tilde{Y}_{idt})E_i(\hat{p}_{ds}) \\
&= Cov_i(p_{dt}, p_{ds}) - E_i(e_{idt}, \bar{e}_c) \\
&= Cov_i(p_{dt}, p_{ds}) - \frac{1}{N_{t-s}} \sum_i \frac{1}{N_{c(i)}} Var(e_{jdt}) \qquad (10)
\end{aligned}$$

where the fourth line follows since e_{idt} and e_{ids} are independent by assumption and so $E_i(e_{idt}e_{ids}) = 0$. Note further that $Cov_i(p_{dt}, p_{ds}) \to_p \sigma_{p(t-s)}$. Let $\hat{e}_{idt}$ denote the residuals from a regression of $\tilde{Y}_{idt}$ on doctor-time fixed effects. We can estimate the second term using:

$$
\begin{aligned}
&\frac{1}{N_{t-s}}\sum_i \frac{1}{N_{c(i)}}\frac{N_{dt(j)}}{N_{dt(j)}-1}\hat{e}^2_{jdt} \to_p \\
&\frac{1}{N_{t-s}}\sum_i \frac{1}{N_{c(i)}}\frac{N_{dt(j)}}{N_{dt(j)}-1}Var(\hat{e}_{jdt}) = \\
&\frac{1}{N_{t-s}}\sum_i \frac{1}{N_{c(i)}}\frac{N_{dt(j)}}{N_{dt(j)}-1}\left(\left(\frac{N_{dt(j)}-2}{N_{dt(j)}}\right)Var(e_{jdt}) + \frac{1}{N^2_{dt(j)}}\sum_{k\in d(t)}Var(e_{kdt})\right) = \\
&\frac{1}{N_{t-s}}\sum_i \frac{1}{N_{c(i)}}\left(\left(\frac{N_{dt(j)}-2}{N_{dt(j)}-1}\right)Var(e_{jdt}) + \frac{1}{N_{dt(j)}(N_{dt(j)}-1)}\sum_{k\in d(t)}Var(e_{kdt})\right) = \\
&\frac{1}{N_{t-s}}\sum_i \frac{1}{N_{c(i)}}\left(\left(\frac{N_{dt(j)}-2}{N_{dt(j)}-1}\right)Var(e_{jdt}) + \frac{1}{N_{dt(j)}-1}Var(e_{jdt})\right) = \\
&\frac{1}{N_{t-s}}\sum_i \frac{1}{N_{c(i)}}Var(e_{jdt}) \qquad (11)
\end{aligned}
$$

The case of the current year requires special consideration. We want to know $Cov(p_{dt}, \hat{p}^{JK}_{idt})$ where $\hat{p}^{JK}_{idt}$ is the jackknife estimate which we compute either using only prior observations or as a leave one out estimate. Let $t(i)$ denote the specific date on which individual i was treated and let N_{idt} denote the number of patients tested prior to patient i by doctor d in period t. In that case:

$$
\begin{aligned}
Cov_d(p_{dt}, \hat{p}^{JK}_{idt}) &= Cov_d(p_{dt}, \frac{1}{N_{idt}}\sum_{t(j)<t(i)}\tilde{Y}_{jdt}) \\
&= Cov_d(p_{dt}, \frac{1}{N_{idt}}\sum_{t(j)<t(i)}p_{dt} + \delta_c + e_{jdt}) \\
&= Var_d(p_{dt}) \qquad (12)
\end{aligned}
$$

We estimate this term below.

The above derivation tells us the off-diagonal terms of Σ_p as well as the vector γ. The diagonal terms are given by $Var_d(\hat{p}_{dt}|N_d) = Var_d(\hat{p}_{dt}|N_{dt})$ since the variance of the estimated $\hat{p}_{dt}$ depends only on the number of observations in that period. If we observed enough doctors in every period with each possible number of observations, we could compute this variance empirically. Since we do not, we again rely on structure to relate the variance for each N_d to a single underlying across patient variance. Specifically, we impose that the mean and variance variance of true p_{dt} across doctors is the same for all N- $Var_{d,N}(p_{dt}) = Var_d(p_{dt})$ and $E_{d,N}(p_{dt}) = E_d(p_{dt})$.

For each value of $N_{dt} = N$, let D_{Nt} denote the set of patients belonging to doctors with exactly N patients at time t and let $|D_{Nt}|$ denote the number of such patients. Further, let $N_{c(i),D_{Nt}}$ denote the number of patients in clinic c belonging to D_{Nt}. Then we have:

$$\begin{aligned}
& Var_i(\hat{p}_{dt}|N_{dt},t) =_p \\
& Var_d(p_{dt}|N_{dt},t) + Var_i\left(-\bar{e}_c + \frac{1}{N_{dt}}\sum_{i\in d(t)} e_{idt}|N_{dt},t\right) \\
= \quad & Var_d(p_{dt}|N_{dt},t) + \frac{1}{|D_{Nt}|}\sum_{i\in D_{Nt}}\left(-\bar{e}_{c(i)} + \frac{1}{N_{dt(i)}}\sum_{j\in dt(i)} e_{jdt}\right)^2 \\
\to_p \quad & Var_d(p_{dt}|N_{dt},t) \\
& +\frac{1}{|D_{Nt}|}\sum_{i\in D_{Nt}}\left(\frac{1}{N^2_{c(i)}}\sum_{j\in c(i)} e^2_{jtd} + \frac{1}{N^2_{dt(i)}}\sum_{j\in dt(i)} e^2_{jdt} - \frac{2}{N_{c(i)}N_{dt(i)}}\sum_{j\in dt(i)} e^2_{jdt}\right) \\
= \quad & Var_d(p_{dt}|N_{dt},t) + \frac{1}{|D_{Nt}|}\sum_{i\in D_{Nt}}\left(\frac{N_{c(i),D_{Nt}}}{N^2_{c(i)}}e^2_{itd} + \frac{1}{N_{dt(i)}}e^2_{idt} - \frac{2}{N_{c(i)}}e^2_{idt}\right) \\
= \quad & Var_d(p_{dt}|N_{dt},t) \\
& +\frac{1}{|D_{Nt}|}\sum_{i\in D_{Nt}}\left(\frac{N_{c(i),D_{Nt}}N_{dt(i)} + N^2_{c(i)} - 2N_{c(i)}N_{dt(i)}}{N^2_{c(i)}N_{dt(i)}}e^2_{idt}\right)
\end{aligned} \tag{13}$$

Analogous reasoning to equation 11 shows that we can compute the parenthetical term by substituting: $\frac{N_{dt(i)}}{N_{dtij)}-1}\hat{e}^2_{idt}$ for e^2_{idt} when $N_{dt} > 1$.

With sufficient data, we could estimate $Var_d(\hat{p}_{dt}|N_{dt},t)$ for every (N_{dt},t) pair and use this to determine $Var_d(p_{dt}|N_{dt},t)$. Note however that this requires we observe enough physicians at that pair that the law of large numbers applies in the above derivation. Because this is not true for many (N_{dt},t) pairs, we instead assume that $Var_d(p_{dt}|N_{dt},t) = Var_d(p_{dt})$, a constant. We additionally assume that $Var_d(\hat{p}_{dt}|N_{dt},t) = Var_d(\hat{p}_{dt}|N_{dt})$ which follows if, for example, we assume that the distribution of e_{idt} is the same in each period.

We then proceed as follows. First, we estimate $Var_d(\hat{p}_{dt}|N_{dt},t)$ for all (N_{dt},t) pairs where we observe at least 20 doctors with 2 observations. Second, we estimate $Var_d(p_{dt})$ as the physician weighted average of these estimates. For each such pair, our estimate is given by:

$$\begin{aligned} Var_d(p_{dt}) &= Var_d(\hat{p}_{dt}|N_{dt},t) - \\ & \frac{1}{|D_{Nt}|}\sum_{i\in D_{Nt}}\left(\frac{N_{c(i),D_{Nt}}N_{dt(i)}+N_{c(i)}^2-2N_{c(i)}N_{dt(i)}}{N_{c(i)}^2(N_{dt(i)}-1)}\hat{e}_{idt}^2\right) \qquad (14) \end{aligned}$$

Finally, we compute $Var_d(\hat{p}_{dt}|N_{dt})$ using the empirical distribution - we bin cases of $N_{dt} > 5$ (3% of all doctors - 2/3 of which have 6 or 7 observations) into a single bin - for the very small number of doctor-times with more than 5 doctors, this will tend to overstate the noise in the variance and thus understate the weight they should receive.

Finally, consider the constant term $\psi_0(N_{d1},...,N_{d(t-1)})$. This term is given by:

$$\begin{aligned} \psi_0(N_{d1},...,N_{d(t-1)}) &= E_d(p_{dt} - \sum_{s=1}^{t-1}\psi_s(N_{d1},...,N_{d(t-1)})\hat{p}_{ds}) \\ &= \bar{p}_t \qquad (15) \end{aligned}$$

since $E_d(\hat{p}_{ds}) = 0$ (by construction, since we partial out time fixed effects when we define $\tilde{Y}_{idt}$). To deal with this term we can just include time fixed effects.