In Praise of Cookbook Medicine

Jason Abaluck, Leila Agha, David Chan, and Diana Zhu
EXTENDED ABSTRACT *

December 9, 2019

1 Background

The movement to promote evidence-based medical care has spurred the development of numerous clinical guidelines designed to help physicians incorporate insight from research into their daily clinical practice. Medical guidelines provide simple rules based on a small number of patient characteristics to determine which patients should be treated. They inform clinicians of underlying evidence and constrain biases in heuristic decision-making (Kahneman and Tversky 1972; Kahneman and Tversky 1973; McFadden et al. 1999). Yet critics worry that guidelines will lead to "cookbook medicine", where physicians adhere to guidelines even if it means neglecting otherwise important features (Costantini et al. 1999; Basu et al. 2014). The widespread adoption of electronic medical records makes possible the use of more complex guidelines, mitigating the concern that simple guidelines will mislead. However, whether guidelines can realize substantial improvements in clinical outcomes is unknown and will depend on the quality of physician decisions absent guideline adoption, as well as how physicians integrate guidelines into their decision-making.

In this paper, we study how guidelines impact physician behavior in the context of the CHADS2 score, a widely used guideline for whether patients with atrial fibrillation should receive anticoagulant (blood thinning) medication to prevent strokes (Gage et al. 2001). As shown in Table 2, the CHADS2 score was developed as an index of five variables that predict stroke risk, under the assumption that the benefits of anticoagulation are proportional to baseline stroke risk (Boode et al. 2004). Because anticoagulation increases the risk of dangerous bleeding, curbing prescriptions of anticoagulants to patients with low expected benefit is an important clinical goal.

This research builds on earlier work on clinical decision making. Glazer et al. (2007) find evidence of widespread non-adherence to the CHADS2 guideline, raising the question of whether the guideline is influencing clinical practice. Abaluck et al. (2016) find

*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. Zhu: Yale University, diana.zhu@yale.edu.

that diagnostic imaging tests fail to target patients with the greatest clinical benefit, suggesting that adoption of high quality clinical guidelines could substantially improve health outcomes. Einav et al. (2019) show that marginal women who receive mammograms at age 40 but not age 39 tend to be healthier and benefit less from screening than trials would otherwise suggest; in other words, the benefits of guidelines depend on which patients adopt. Building on this earlier work, our project will allow us to assess whether stronger incentives to adopt guidelines would be beneficial. Specifically, we need to understand both whether guidelines lead physicians to attend less to patient attributes not included in guidelines, as well as whether adopting physicians and non-adopting physicians differ in their ex ante behavior.

Our project uniquely combines data on physicians' guideline adoption and treatment decisions with randomized control trial evidence on treatment effect heterogeneity. This allows us to evaluate how physician decisions change when they start using guidelines – we can see how non-adopters and adopters compare in their treatment decisions before and after adoption, and we can use the trial data to evaluate the likely patient consequences. Importantly, the trial data also allows us to estimate treatment effects as a function of many more variables than existing guidelines (using a causal forest model). This allows us to evaluate whether guidelines lead physicians to neglect otherwise important variables not covered by the guideline.

2 Guideline diffusion and estimation of heterogeneous treatment effects

Using data scraped from physician notes in the Veterans Health Administration data (VHA), we can pinpoint when physicians first mention the CHADS2 score in their clinical notes. Although the CHADS2 guideline was first published in Gage et al. (2001), we find almost no evidence of guideline use prior to 2007. As shown in Figure 1, the CHADS2 guideline is mentioned in physician notes for fewer than 1% of patients with atrial fibrillation in 2002, climbing to 11% of patients in 2014. The fraction of physicians who have previously mentioned the guideline at least once also climbs steadily over the same period, rising from fewer than 1% to 60%.

For our cohort of 132,000 VHA patients newly diagnosed with atrial fibrillation, we use patients' clinical history to construct their CHADS2 score at the time of diagnosis and track their subsequent prescriptions for anticoagulants. Overall anticoagulant prescription rates are uncorrelated with clinical guidelines in the early years of the data, but prescription rates drop for patients with low CHADS2 scores in the later years of

our sample. These trends are shown in Figure 2.

Physicians become more guideline adherent following their first observed mention of the CHADS2 score: specifically, physicians become 8 percentage points (or 17%) less likely to prescribe anticoagulation for patients contraindicated by the CHADS2 score. For patients whose CHADS2 scores indicate they would benefit from anticoagulation, we find no change in prescription rates before and after the doctor's CHADS adoption. These results are shown graphically in Figure 3,

To evaluate whether these changes improve patient outcomes, we combine the VHA data with the Atrial Fibrillation Investigators database (hereafter, AFI data), a database containing patient level observations for 12 randomized clinical trials in which patients were randomized to anticoagulants versus a placebo or control.[1] This database includes patient's clinical characteristics at the time of randomization, patients' randomized treatment status, and subsequent stroke and bleed outcomes. Clinical characteristics include all variables underlying the CHADS2 score, as well as several additional variables (i.e. laboratory values from common blood tests, further detail on past stroke events, height, smoking status, and additional comorbidities and demographic variables).

While the AFI data has previously been analyzed (van Walraven et al. 2002; van Walraven et al. 2009), variable selection has been ad hoc, and machine learning methods have not previously been used to determine the best model of treatment effects given all available variables. We estimate a causal forest model to determine the mean-squared error minimizing predictor of stroke and bleed treatment effects for each patient as a function of all available patient characteristics. We find that the CHADS2 score is highly correlated with the causal forest model, although there is substantial residual variation (an R^2 of 0.36). For example, the causal forest suggests that the single most important variable in predicting stroke treatment effects is the recency of the last stroke, which is not included in the CHADS2 score.

3 Evaluating the benefits of guideline use

We then use the estimated causal forest model (from the AFI data) to assess whether the CHADS2 score improves treatment allocation (in the VHA data). Specifically, we estimate a model where physicians make treatment decisions based on three factors: the CHADS2 score; residual variation in causal forest treatment effects (orthogonal to the CHADS2 score); and noise. Although patient age directly enters the CHADS2 score and is predictive of treatment effects—older patients have larger stroke reduction

[1] Anticoagulants includes both oral anticoagulants and antiplatelet therapy.

benefits—we find that doctors are less inclined to treat older patients, contravening the guideline recommendation. Considerations unrelated treatment effects may discourage physicians from prescribing anticoagulant to older patients. Because warfarin (the most commonly prescribed anticoagulant in this population) interacts with patients' diet and other drugs, it requires regular blood work (typically 1-4 times per month) and possible dosing adjustment. Adherence to this regime may be especially challenging for older patients. Because of these possible clinical justifications for avoiding the treatment of old patients, we control for age in this analysis and consider only variation in treatment effects and CHADS2 score that is orthogonal to patient age.

We find that prior to adopting the CHADS2 score, physicians' treatment decisions are already somewhat sensitive to the component of treatment effects predicted by the CHADS2 score. Even before the codification and popularization of the CHADS2 score, consensus expert recommendations emphasized the importance of the same clinical risk factors in anticoagulant treatment decisions, though they did not explicitly apply them in an algorithmic scoring tool. Following their first mention of the CHADS2 score, physicians become more than twice as sensitive to the component of treatment effects predicted by the CHADS2 score. Both before and after CHADS2 adoption, physicians' treatment decisions are insensitive to the residual variation in treatment effects predicted by the causal forest but orthogonal to the CHADS2 score. Further, we find no evidence that physicians who eventually adopt the CHADS2 score were making better (or worse) treatment choices prior to their first CHADS2 mention than non-adopting physicians were at that time.

There are two potential ways in which CHADS2 score adoption could worsen patient outcomes. First, it may lead physicians to treat patients in a way that induces more bleeds. Stroke risk is highly correlated with bleed risk in both the VA data and the experimental data. By guiding physicians to treat patients with the greatest stroke risk, the CHADS2 score may inadvertently lead to treatment of patients who suffer the most potential harm from anticoagulation due to bleeding risk. Second, by emphasizing a handful of variables, the CHADS2 score may discourage physicians from considering other patient attributes that predict important variation in stroke reduction from anticoagulants.

We find no evidence that guidelines induce harmful effects along either dimension. In the trial data, we find that patients with larger stroke *treatment effects* from anticoagulants (as opposed to baseline risk) have *fewer* induced bleeds. And in fact, the CHADS2 score succeeds in targeting these patients who are at low risk of induced bleeds. Additionally, we find that physicians are largely insensitive to non-CHADS2 predictors

of stroke treatment effects both before and after adopting the guideline, suggesting the guideline did *not* discourage physicians from considering important variables that they were otherwise weighting in their decision.

These findings raise an important question: could better guidelines improve physician behavior further? To investigate, we simulate the effects of more comprehensive guidelines which utilize all of the information in our causal forest estimates of the heterogeneous impacts of strokes and bleeds. We consider scenarios in which physicians are as responsive to this new guideline as they are to existing guidelines, as well as scenarios where physicians reallocate treatment within age categories by strictly ordering patients according to the new guideline.

4 Discussion and lessons for guideline development

Finally, we consider lessons from our data for optimal guideline construction. In the medical literature, it is commonplace to construct guidelines which recommend treatment based on *risk among untreated patients*. Indeed, the CHADS2 score was developed based on such evidence (Gage et al. 2004). While risk among untreated patients is correlated with treatment effects, the relationship is not perfect. We compare the performance of guidelines developed with the risk score approach to guidelines developed using estimates of treatment effect heterogeneity. In addition, we consider the returns to complexity in guideline construction, contrasting the performance of the best five-factor scores to the full causal forest estimate.

This research suggests that evidence-based clinical guidelines have tremendous potential to improve patient health outcomes. There are two primary pathways to increasing the positive impact of clinical guidelines. First, increasing the adoption of existing guidelines would improve treatment allocation in the setting we study. We find no evidence that non-adopting physicians are higher skill or would otherwise experience lower benefits of guideline use, and yet 40% of physicians have still not adopted guidelines seven years after the first signs of guideline adoption among VHA physicians. Second, new causal machine learning approaches to guideline development could markedly improve guideline quality, further increasing the benefits of guideline adoption.

References

Abaluck, J., L. Agha, C. Kabrhel, A. Raja, and A. Venkatesh (2016, December). The determinants of productivity in medical testing: Intensity and allocation of care.

American Economic Review 106(12), 3730–64.

Basu, A., A. B. Jena, D. P. Goldman, T. J. Philipson, and R. Dubois (2014). Heterogeneity in action: the role of passive personalization in comparative effectiveness research. *Health economics 23*(3), 359–373.

Boode, B., P. Petersen, B. F. Gage, C. van Walraven, L. Pearce, R. G. Hart, and P. J. Koudstaal (2004). Selecting patients with atrial fibrillation for anticoagulation: Stroke risk.

Costantini, O., K. K. Papp, J. Como, J. Aucott, M. D. Carlson, and D. C. Aron (1999). Attitudes of faculty, housestaff, and medical students toward clinical practice guidelines. *Academic medicine: journal of the Association of American Medical Colleges 74*(10), 1138–1143.

Einav, L., A. Finkelstein, T. Oostrom, A. J. Ostriker, and H. L. Williams (2019). Screening and selection: The case of mammograms. Technical report, National Bureau of Economic Research.

Gage, B. F., C. van Walraven, L. Pearce, R. G. Hart, P. J. Koudstaal, B. Boode, and P. Petersen (2004). Selecting patients with atrial fibrillation for anticoagulation: stroke risk stratification in patients taking aspirin. *Circulation 110*(16), 2287–2292.

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.

Kahneman, D. and A. Tversky (1972). Subjective probability: A judgment of representativeness. *Cognitive psychology 3*(3), 430–454.

Kahneman, D. and A. Tversky (1973). On the psychology of prediction. *Psychological review 80*(4), 237.

McFadden, D., M. J. Machina, and J. Baron (1999). Rationality for economists? In *Elicitation of preferences*, pp. 73–110. Springer.

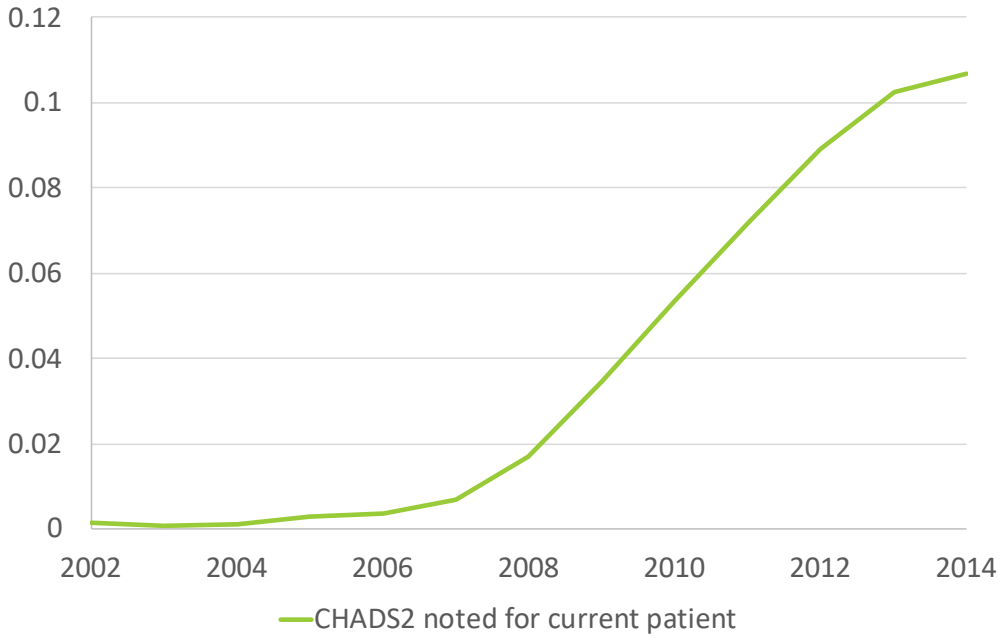
van Walraven, C., R. G. Hart, S. Connolly, P. C. Austin, J. Mant, F. R. Hobbs, P. J. Koudstaal, P. Petersen, F. Perez-Gomez, J. A. Knottnerus, et al. (2009). Effect

of age on stroke prevention therapy in patients with atrial fibrillation: the atrial fibrillation investigators. *Stroke 40*(4), 1410–1416.

van Walraven, C., R. G. Hart, D. E. Singer, A. Laupacis, S. Connolly, P. Petersen, P. J. Koudstaal, Y. Chang, and B. Hellemons (2002). Oral anticoagulants vs aspirin in nonvalvular atrial fibrillation: an individual patient meta-analysis. *Jama 288*(19), 2441–2448.

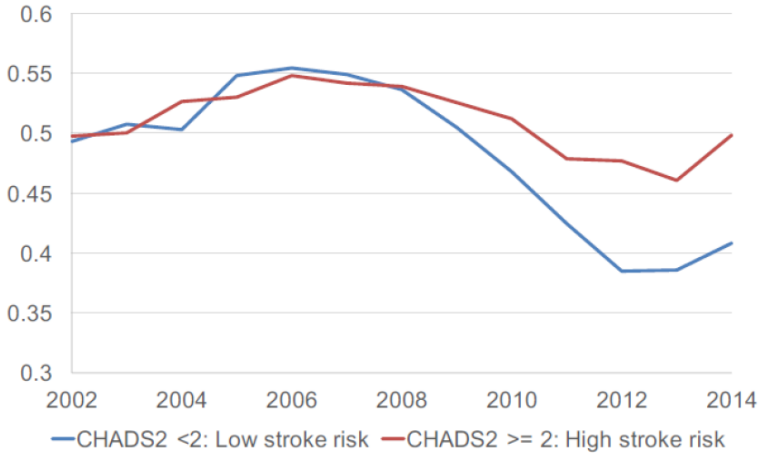
5 Tables and Figures

Figure 1: Diffusion of the CHADS2 Score over Time



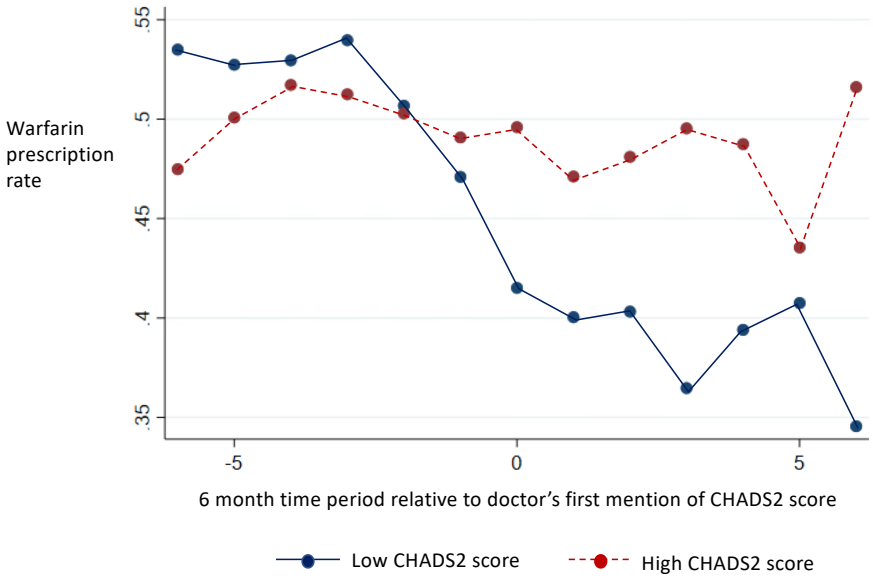
Notes: This figure shows the fraction of patients whose physician notes mention the CHADS2 guideline, among the cohort of patients newly diagnosed with atrial fibrillation. Data is from the Veterans Health Administration.

Figure 2: 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 127,000 patients with atrial fibrillation treated by the Veterans Health Administration.

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



Notes: This figure plots the average rate of warfarin prescription separately for high CHADS2 (score of 2 or above) and low CHADS2 score patients (score of 0 or 1). Data are averaged over 6-month periods, for three years before and three years after the doctor's first mention of the CHADS2 score in a note for any patient. Time period 0 marks the time of first CHADS2 score mention by the patient's physician. The plot shows how warfarin prescription rates for low score patients decline around the time of their doctor's first mention of the CHADS2 guidelines.

Table 2: 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.