Are Guidelines Worth Following?
Treatment Decisions Under Scientific Uncertainty

Jason Abaluck, Leila Agha, David Chan, Daniel Singer, Diana Zhu

February 2025

Motivation

- Information in health care is increasingly complex
  - Proliferation of technologies, evidence, and guidelines
- Yet information is still incomplete/unrepresentative (*scientific uncertainty*)
  - Evidence usually disease-specific; RCTs usually exclude higher-risk patients
  - Guidelines often silent on managing “complicated” patients, who are quite common in the real world (20% of Medicare beneficiaries have 5+ chronic conditions; 50% on 5+ medications
- Adherence to guidelines low

Motivating Questions

- How should we use existing information when it is incomplete and unrepresentative?
- How do we assess physician behavior in light of scientific uncertainty?

This Paper

- Study physician response to the introduction of an influential guideline for anticoagulation in atrial fibrillation (AF)
- Obtain causal estimates of heterogeneous TEs—benefits (reduced strokes) and harms (induced bleeds)—using machine learning (ML) methods in RCT data
- Develop framework to account for scientific uncertainty about harms
- Assess treatment rules: known guidelines, “optimal” rules (under specific assumptions), physician behavior

Preview of Findings

- Study physician response to the introduction of CHADS_2 score, an influential guideline for anticoagulation in atrial fibrillation (AF)
    - Widespread awareness but modest response and low adherence
- Obtain causal estimates of heterogeneous benefits (reduced strokes) and harms (induced bleeds) using machine learning (ML) methods in RCT data
    - RCT population unrepresentative
    - ML methods detect wide heterogeneity in benefits (stroke TEs); estimation limited for harms (bleed TEs)
    - However, *risks* (bleeds in absence of treatment) positively correlated with benefits

Preview of Findings

- Develop framework for evaluating performance under scientific uncertainty, considering two scenarios:
  - *A*: harms uncorrelated with benefits
  - *B*: harms proportional to underlying risk
- Develop theory for *maxmin* treatment rule: optimizing the worst outcome under uncertainty
  - Positive correlation between risks and benefits $\Rightarrow$ tradeoff between scenario-specific strategies for *A* and *B*.

Preview of Findings

- Assess performance of known and optimal treatment rules under uncertainty
  - Known guidelines could do worse than random treatment—driven by positive correlation between benefits and risks
  - Even scenario-specific optimal treatment rules perform relatively poorly
- Assess physician behavior
  - Providers appear to weigh benefits against risks, contrary to CHADS_2 score
  - $>$ 90% of physicians outperform CHADS_2 score under maxmin criterion and *B* (contrast with 0% under *A*)

1. Setting and Data

2. Provider Response

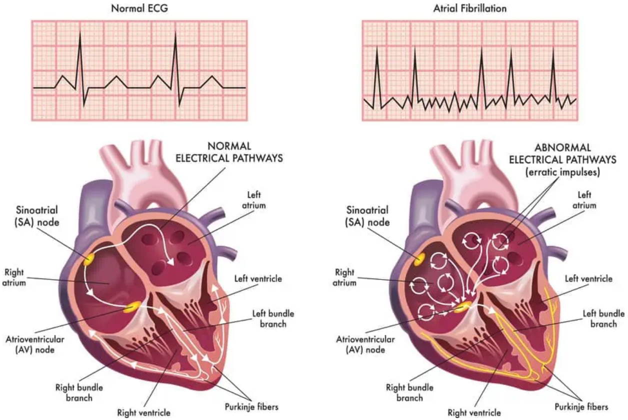
3. Heterogeneous Treatment Effects

4. Guideline Performance Under Scientific Uncertainty

5. Physician Behavior

Atrial Fibrillation (AF)

One in four adults over age 40 will develop AF; increases stroke risk by five-fold



Guidelines

- Primary treatment for stroke prevention is anticoagulation (warfarin)
    - Difficult tradeoff: prevent strokes but induce potentially life-threatening bleeds
- CHADS$_2$ score: predicts stroke *risk*; validated for clinical practice in 2004, first adopted as a guideline in 2006
    - **C**: congestive heart failure (1 point)
    - **H**: hypertension (1 point)
    - **A**: age $\geq$ 75 years (1 point)
    - **D**: diabetes (1 point)
    - **S**$_2$: stroke (2 points)
- Later variant in 2010: CHA$_2$DS$_2$-VASc score
- Provide no explicit guidance on how to calculate or use bleed risks/harms

VHA Setting and Data

- Electronic medical records in the Veterans Health Administration (VHA) from 2002 to 2013
- Identify 112,000 potentially new diagnoses of atrial fibrillation (Turakhia et al. 2013; Perino et al. 2017)
  - No previous diagnosis within 3 years, EKG near initial diagnosis, no prior treatment
  - Visit with cardiologist or PCP within 90 days of diagnosis
  - Providers must have at least 30 AF patients, warfarin prescription history
- Leverage patient characteristics (demographics, comorbidities, laboratory tests, body measurements, vital signs), prescriptions, provider notes

1. Setting and Data

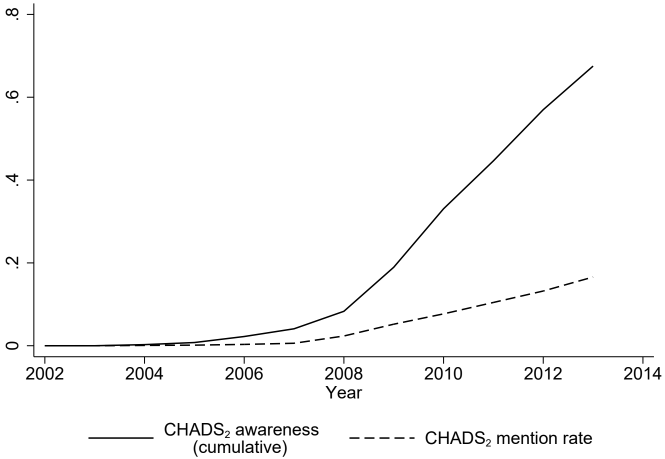
2. Provider Response

3. Heterogeneous Treatment Effects

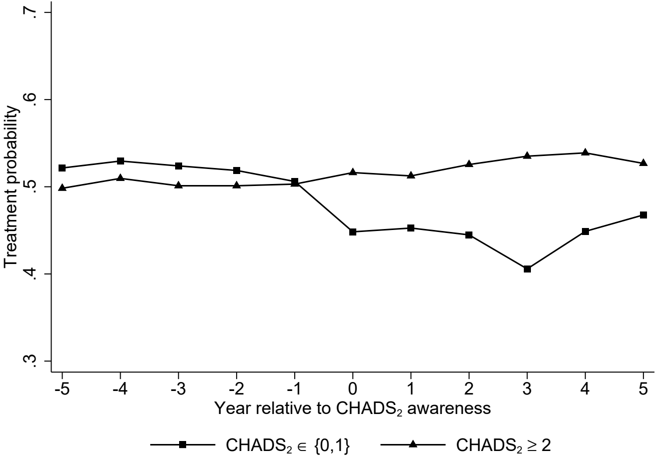
4. Guideline Performance Under Scientific Uncertainty

5. Physician Behavior

CHADS$_2$-Score Awareness



CHADS$_2$-Score Awareness



1. Setting and Data

2. Provider Response

3. Heterogeneous Treatment Effects

4. Guideline Performance Under Scientific Uncertainty

5. Physician Behavior

Heterogeneous Treatment Effects

- Foundation for guidelines: For whom will the benefits outweigh the harms of treatment?

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

  where $Y_i^s(D_i)$ and $Y_i^b(D_i)$ are potential stroke and bleed outcomes for patient i under treatment $D_i \in \{0, 1\}$.

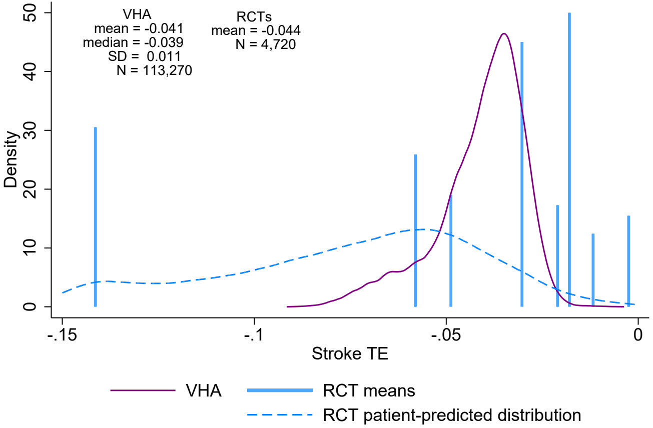
- To our knowledge, no existing estimates of TE heterogeneity for AF anticoagulation.
- Estimate using ML on RCT data

Pros and Cons of RCTs as Evidence

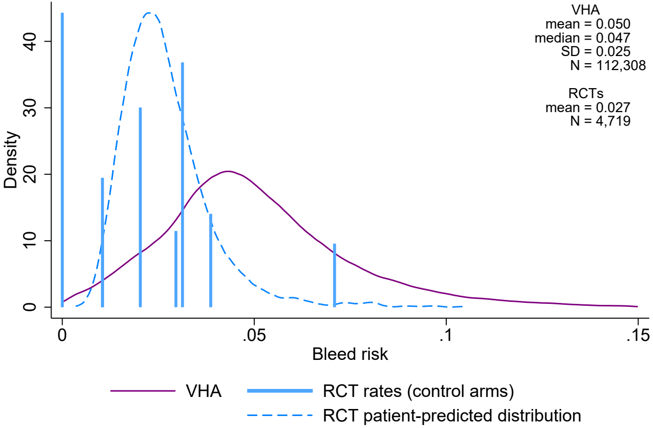
- Rationale:
  - The gold standard for clinical evidence
  - Rigorous outcome measurement
  - Assumptions for recovering (heterogeneous) CATEs within cells likely valid
- Limitations:
  - Sample may be unrepresentative (selected to show benefit)
  - Under-powered to detect heterogeneity in TEs
  - Focused on short-term outcomes

Decisions for many/most patients may be (RCT) evidence-free

Distribution of Stroke TEs



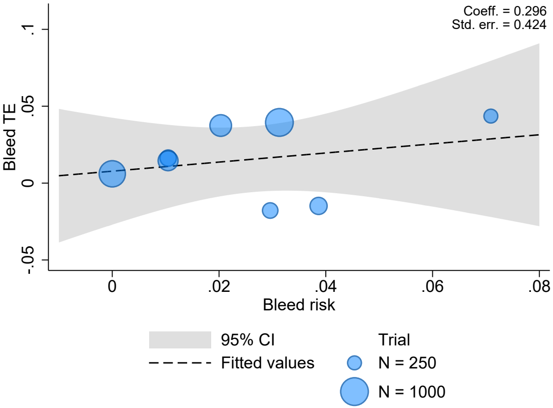
Distribution of Bleed Risk



▸ Deconvolved Across RCTs

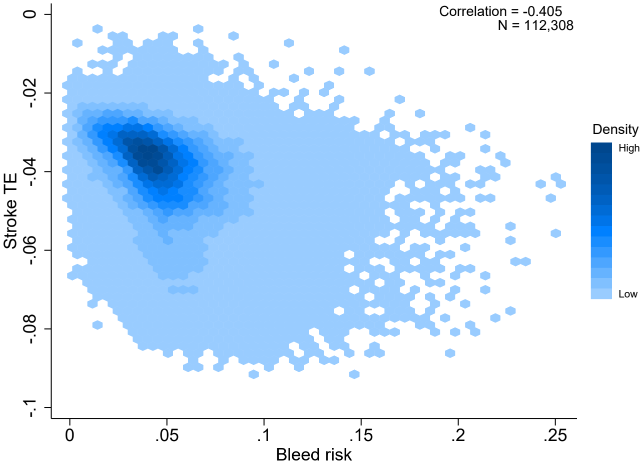
Challenge: Estimating Bleed TEs

Regression Across RCTs



Notes: Bleed risk = bleeds in control arms; mixed results when predicting bleed risk at patient level in hold-out data ▸ Regression Within RCTs

Joint Distribution



1. Setting and Data

2. Provider Response

3. Heterogeneous Treatment Effects

4. Guideline Performance Under Scientific Uncertainty

5. Physician Behavior

Framework for Evaluation

- Evaluate stroke and bleed outcomes when a subset of patients are treated
- Outcomes depend on share of treated patients and ranking of patients
- Given uncertainty about harms, consider two plausible scenarios:
  1. *A*: bleed TEs orthogonal to stroke TEs
  2. *B*: bleed TEs proportional to bleed risks
- Aversion to ambiguity/uncertainty (incalculable risks) → consider maxmin criterion: maximize worse-case utility

Optimal Guidelines

- Consider expected utility of treating patient $i \in \mathcal{I}$, with stroke and bleed TEs $\tau_i^s \in [-1, 0]$ and $\tau_i^b \in [0, 1]$:

  $$u(i) = -\tau_i^s - \beta\tau_i^b,$$

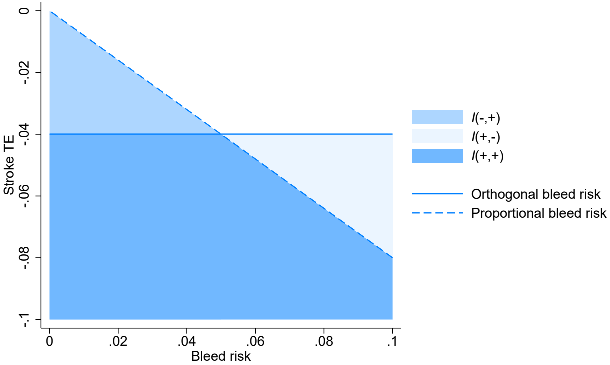
  where β is the utility cost of a bleed relative to a stroke
- Issue: Cannot observe τ_i^b, but can observe bleed risk, $\alpha_i^b \equiv E[Y_i^b(0)|X_i = x]$
- Scenario-specific optimal guideline depends on assumption $\omega \in \{A, B\}$:

  $$\begin{aligned} D_A^*(i) &= \mathbf{1}(\tau_i^s < \beta E[\tau_i^b]); \\ D_B^*(i) &= \mathbf{1}(\tau_i^s < \beta\kappa\alpha_i^b), \end{aligned}$$

  where $\kappa = E[\tau_i^b]/E[\alpha_i^b]$

Graphical Representation

Share of patients over which D_A^* and D_B^* disagree depends on $\Delta(\alpha_i^b, \tau_i^s)$



Notes: $\mathcal{I}(+,+)$: treat in both *A* and *B*; $\mathcal{I}(+,-)$: treat only in *A*; $\mathcal{I}(-,+)$: treat only in *B*.

Maxmin Solution

- Consider welfare $W_\omega(D)$ for treatment rule D under $\omega \in \{A, B\}$:

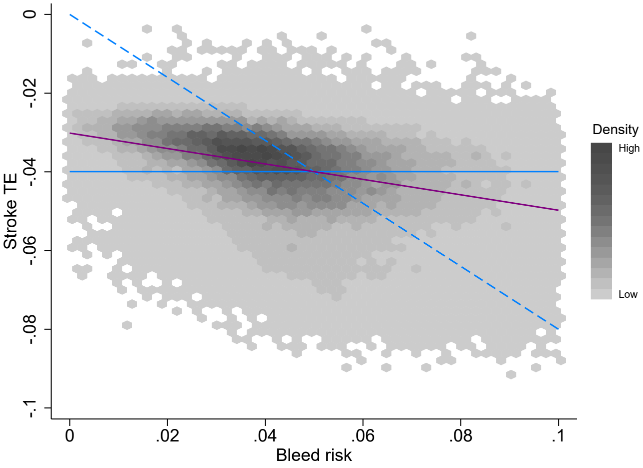
$$W_\omega(D) = \int_0^{P(D)} u_\omega(i) d\Omega(i; D)$$

- Solve for treatment rule that maximizes the worse welfare:

$$D^* = \arg\max_D (\min(W_A(D), W_B(D)))$$

- D^* is unique and satisfies a moment condition: $E[\alpha_i^b | D^*(i) = 1] = E[\alpha_i^b]$
  - Intuition: effectively orthogonalizes bleed risks relative to stroke TEs
  - If $\text{Corr}(\alpha_i^b, \tau_i^s) \geq 0$, $D^* = D_A^*$ (i.e., can rank by benefits and ignore harms)

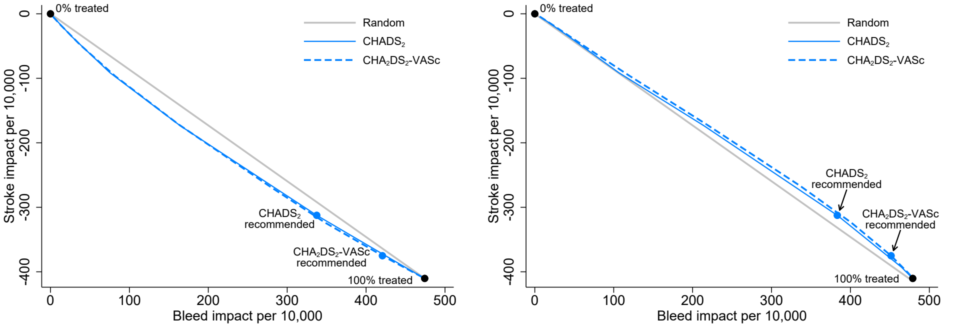
Maxmin Solution



Empirically Evaluating Treatment Rules

- Without knowing β, can compare sets of stroke-bleed outcomes for a given ranking implied by treatment rules
  - Between two rankings, a *dominating* ranking reduces more strokes for any level of bleeds (or induces fewer bleeds for any level of strokes)
- Can compare ranking performance between scenarios *A* and *B*

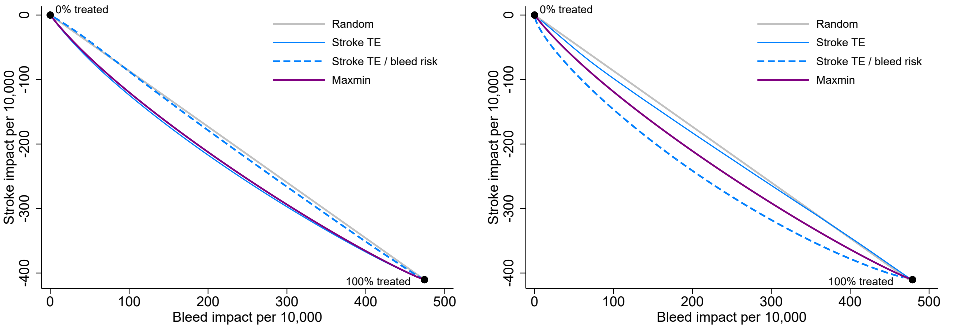
Outcomes: Known Guidelines



(a) Orthogonal Bleed Effects

(b) Proportional Bleed Effects

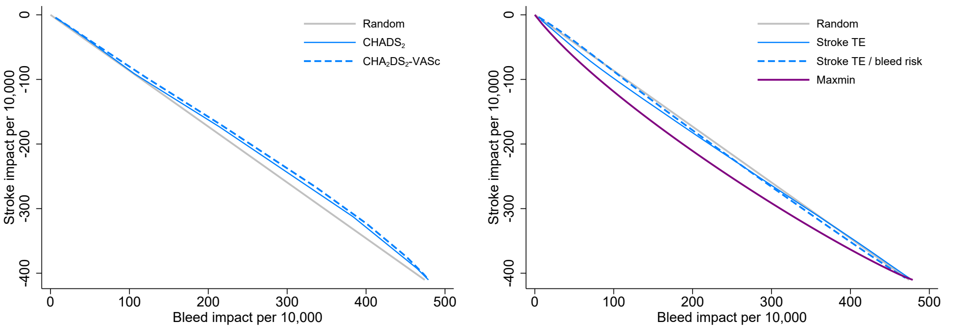
Outcomes: Optimal Guidelines



(a) Orthogonal Bleed Effects

(b) Proportional Bleed Effects

Outcomes: Worse Case



(a) Known Guidelines

(b) Optimal Guidelines

Implications of Scientific Uncertainty

- Guideline performance depends on (unknown) joint distribution of (τ_i^b, τ_i^s)
- Known guidelines (CHADS$_2$ and CHA$_2$DS$_2$-VASc scores) can perform worse than random treatment
- Optimal guidelines D_A^* and D_B^* can also perform worse than random treatment under the opposite assumption
  - In our case, they perform only slightly better than random
- Worse performance when (i) benefits and harms are positively correlated and (ii) potential variation in harms is larger

1. Setting and Data

2. Provider Response

3. Heterogeneous Treatment Effects

4. Guideline Performance Under Scientific Uncertainty

5. Physician Behavior

Physician Behavior

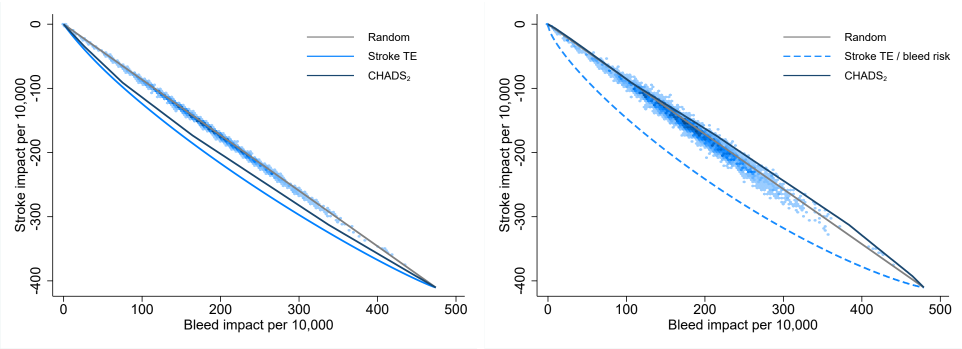
- Estimate random coefficients model of physician behavior based on τ_i^s and α_i^b:

  $$D_j(i) = \mathbf{1}\left(\tau_i^s < a_j - b_j(\alpha_i^b - E[\alpha_i^b]) + p_j + \varepsilon_{ij}/k_j\right),$$

  where $a_j \in (-1, 0)$; b_j, p_j, k_j normally distributed; $\varepsilon_{ij} \sim N(0, 1)$

  - Physicians generally respond to both τ_i^s and α_i^b, wide variation in $\Pr(D_j(i) = 1)$ across physicians
- Based on hyperparameters of (a_j, b_j, p_j, k_j), simulate population of physicians and their stroke/bleed outcomes under A and B
- Compare population performance with known and optimal guidelines

Physician-Level Outcomes

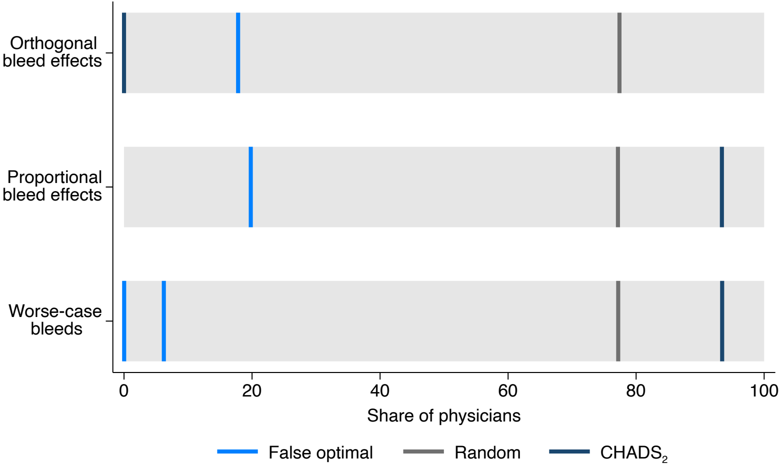


(a) Orthogonal Bleed Effects

(b) Proportional Bleed Effects

▸ Worse-Case Heatmap

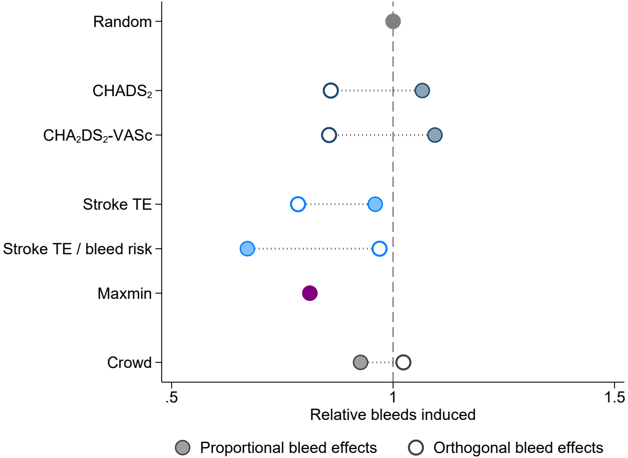
Physician Performance Relative to Benchmarks



Wisdom of the Crowd

- Revisit physician behavior by estimating implicit *aggregate ranking* $\Omega(X_i)$
- Focus on within-physician ranking, pooled across physicians
  - I.e., abstract from practice variation across physicians
  - Akin to “wisdom of the crowd” approach in AI (e.g., Agarwal et al. 2023); as yet unproven whether it will improve outcomes
- Estimate for physicians in general and subgroups (e.g., cardiologists vs. PCPs)

Performance Relative to Random Treatment



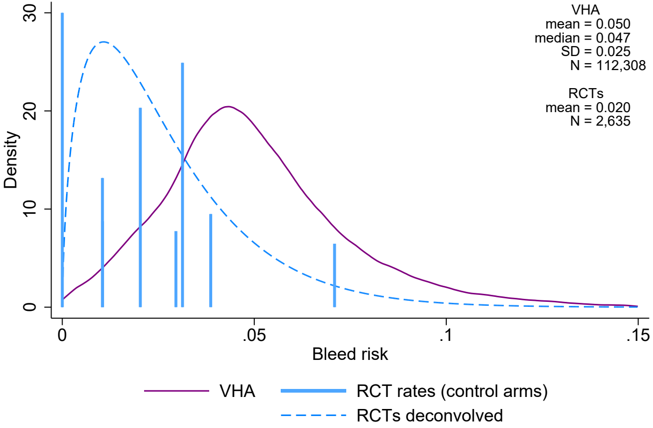
Notes: Strokes fixed at 50% random treatment ▸ Subsets of Crowd

Conclusion

- We study a well-known guideline for AF treatment
  - Despite expert recommendations, strict adherence may worsen outcomes relative to random treatment
  - Danger in focusing on where we have information
- Difficult question: how to incorporate information yet account for its gaps?
  - Following the crowd is imperfect but may be better than following existing guidelines
  - Potential of developing treatment rules robust to gaps in information

Appendix

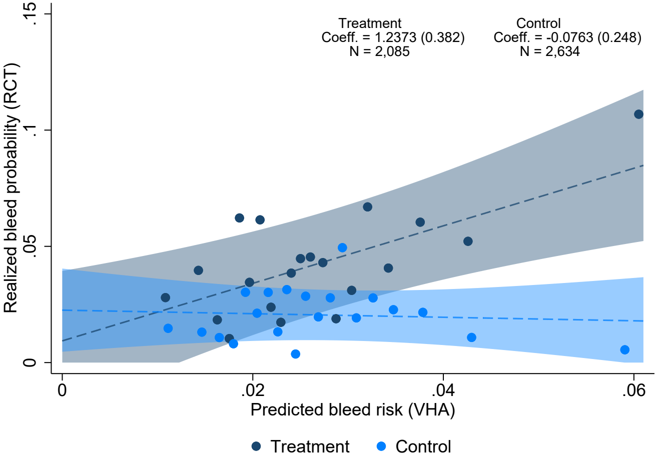
Distribution of Bleed Risk (Deconvolved Across RCTs)



Back

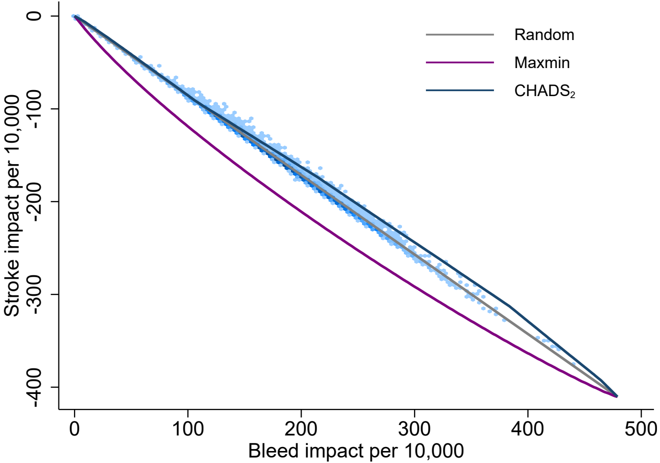
Challenge: Estimating Bleed TEs

Regression Within RCTs



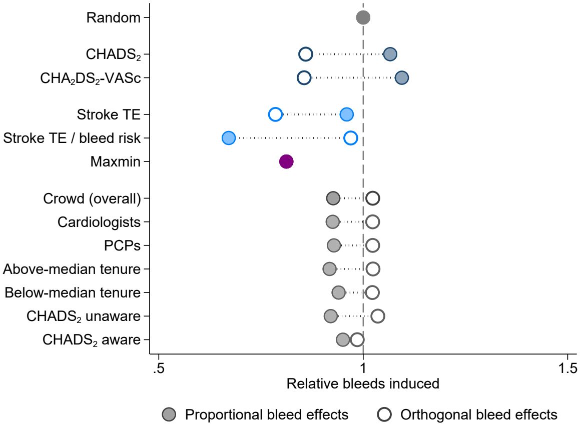
Back

Physician-Level Outcomes: Worse Case



Back

Performance Relative to Random Treatment



Notes: Strokes fixed at 50% random treatment

Back