DIAGNOSING PHYSICIAN ERROR - A MACHINE LEARNING APPROACH TO LOW-VALUE HEALTH CARE (QJE, 2022)

Sendhil Mullainathan, Ziad Obermeyer

Joseph Paul

Heriot-Watt University

November 3, 2023

INTRODUCTION: DIAGNOSING PHYSICIAN ERROR

- Testing for heart attacks is costly and invasive, yet misdiagnosis can be fatal.
- Machine learning as a diagnostic tool to assess *physician decision-making efficiency*.
- Examination of *246,265 emergency visits* over 2010-2015 at a leading hospital.
- A novel approach: Instead of direct comparison, the study uses algorithmic predictions to identify potential inefficiencies in testing decisions.

ALLOCATIVE INEFFICIENCY IN TESTING

- Identification of two core inefficiencies:
  1. Over-testing: 62% of tests costing over $150,000 per life-year.
  2. Under-testing: High-risk patients missing out on crucial tests.
- Traditional measures of efficiency would indicate that testing on average was efficient, hiding a lot of heterogeneity.

CONTRIBUTIONS

- Insights into **bounded rationality** and **representativeness heuristic** in medical decision-making.
- The challenge to moral hazard perspectives: balancing the risks of over- and under-testing.
- Policy implications: the necessity of considering systematic mistakes in health care models—**behavioral hazard**.

MEDICAL CONTEXT: ACUTE CORONARY SYNDROME (ACS)

- **Coronary Arteries:** Supply blood to the heart. Blockage kills heart muscle, leading to ACS.
  - *Consequences:* Immediate (arrhythmia, sudden death) and long-term (fatigue, heart failure).
- **Treatments:**
  - *Stenting:* Inserts a tube into the blocked artery to restore flow.
  - *Bypass:* For severe cases, open-heart surgery to bypass the blockage.

MEDICAL CONTEXT: ACUTE CORONARY SYNDROME (ACS)

- **Coronary Arteries:** Supply blood to the heart. Blockage kills heart muscle, leading to ACS.
  - *Consequences:* Immediate (arrhythmia, sudden death) and long-term (fatigue, heart failure).
- **Treatments:**
  - *Stenting:* Inserts a tube into the blocked artery to restore flow.
  - *Bypass:* For severe cases, open-heart surgery to bypass the blockage.

MEDICAL CONTEXT: ACUTE CORONARY SYNDROME (ACS) CONTINUED

- **Diagnosis Challenges:** Blockages often show subtle symptoms, e.g., mild chest squeezing, breath shortness.
- **ED Tests:**
  - *ECG:* Measures heart’s electrical activity.
  - *Troponin:* Blood test for dying heart muscle cells.
- **Definitive Test:**
  - *Cardiac Catheterization:* Invasive procedure to visualise blockages.
  - *Alternative:* Stress testing to detect blockage, then catheterization for treatment.
- **Cost and Risks:** Both tests are costly and have health risks, especially catheterization.
- **Decision Context:** Weigh treatment benefits against costs and risks.

FRAMEWORK OVERVIEW

- Patients are characterised by (X, Z). Z is visible only to the physician.
- Blockage B occurs with probability $b(X, Z)$.
- Test T yields a positive result with:

$$Pr(Y = 1 \mid B, X, Z) = Pr(y = 1 \mid B) = p + B(q - p),$$

  where p and q are the false and true positive rates.
- Stenting S treats the blockage, but is contingent on a positive test result ($Y = 1$).

POTENTIAL OUTCOMES FRAMEWORK

Patient's health, W^S, depends on stenting:

$$W^S = W - B(\eta - S\tau^K)$$

where

- $E[W \mid X, Z, B] = w(X, Z)$
- η is the negative health impact from blockage.
- $\tau^K = \tau - \theta K$: Benefit from stenting, adjusted by patient type K.
- Patients with $K = 1$ have known risks known to physicians.

BENEFIT OF STENTING

Benefit of stenting someone with a positive test:

$$\tilde{\tau}^K = E[W^1 - W^0 \mid Y = 1, K]$$

Average benefit of treating those with a blockage:

$$\tau^K = E[W^1 - W^0 \mid B = 1, K]$$

ADVERSE EVENTS AND COSTS

Adverse events *A* happen if a blockage is untreated:

$$A = \mu + B(\zeta - S\phi)$$

Socially optimal testing and treatment to maximise expected health:

$$\max_{S,T} w(X, Z) - b(X, Z)(\eta - S\tau^{K}) - c_T T - c_S S$$

Subject to: Only patients with positive test are stinted.

PHYSICIAN'S OBJECTIVES

Physicians may:

- Derive additional benefit $\nu > 0$ from testing.
- Misestimate the probability: $h(X, Z)$ vs $b(X, Z)$.

Physician's objective function:

$$\max_{S,T} w(X, Z) - h(X, Z)(\eta - S\tau^K) - (c_T - \nu)T - c_S S$$

Socially optimal

$$\max_{S,T} w(X, Z) - b(X, Z)(\eta - S\tau^K) - c_T T - c_S S$$

PHYSICIAN'S OBJECTIVES

Physicians may:

- Derive additional benefit $\nu > 0$ from testing.
- Misestimate the probability: $h(X, Z)$ vs $b(X, Z)$.

Physician's objective function:

$$\max_{S,T} w(X, Z) - h(X, Z)(\eta - S\tau^K) - (c_T - \nu)T - c_S S$$

Socially optimal

$$\max_{S,T} w(X, Z) - b(X, Z)(\eta - S\tau^K) - c_T T - c_S S$$

TESTING RULES

Socially Optimal

$$\text{Test iff } b(X, Z) > \frac{c_T + pc_S}{q\tau^K - c_S(q - p)}$$

Physician Testing

$$\text{Test iff } h(X, Z) > \frac{c_T - \nu + pc_S}{q\tau^K - c_S(q - p)}$$

Divergence from optimal due to: i) Private benefits or ii) Misestimation of risk.

Definition: Overtested V

- Tested patients in V have a lower than average yield.
- Conditions:
  - Among tested patients in set V, the average outcomes (yield) is below the global average: $$\mathbb{E}[Y|(X,Z) \in V, T = 1] < \mathbb{E}[Y|T = 1]$$
  - Positive probability of testing patients in V: $$\mathbb{E}[T|(X,Z) \in V] > 0$$
  - Expected yield from testing in V is below the threshold: $$\mathbb{E}[Y|(X,Z) \in V, T = 1] < \frac{c_T}{\tilde{\tau}^0 - c_S}$$

Definition: Overtested *V*

- Tested patients in *V* have a lower than average yield.
- Conditions:
  - Among tested patients in set *V*, the average outcomes (yield) is below the global average:
    $$\mathbb{E}[Y|(X,Z) \in V, T = 1] < \mathbb{E}[Y|T = 1]$$
  - Positive probability of testing patients in *V*:
    $$\mathbb{E}[T|(X,Z) \in V] > 0$$
  - Expected yield from testing in *V* is below the threshold:
    $$\mathbb{E}[Y|(X,Z) \in V, T = 1] < \frac{c_T}{\tilde{\tau}^0 - c_S}$$

Definition: Undertested V

Conditions:

- Among patients characterised by set V and who are of type $K = 0$, the probability they are tested is less than 1.

$$\mathbb{E}[T|(X, Z) \in V, K = 0] < 1$$

- For patients in set V of type $K = 0$ and who aren't tested, the expected adverse events are greater than the threshold defined by the constant adverse event rate μ plus a term that relates the costs of testing and stenting to the true positive rate and false positive rate of the test.

$$\mathbb{E}[A \mid (X, Z) \in V, K = 0, T = 0] > \mu + \eta \left(\frac{c_t + pc_S}{q\tau^0 - c_S(q - p)} \right)$$

Note: If physician judgments are erroneous, $h(X, Z) \neq b(X, Z)$, then both undertested and overtested patient subsets can exist. If accurate, overtested subsets happen only if $\nu > 0$.

Definition: Undertested V

Conditions:

- Among patients characterised by set V and who are of type $K = 0$, the probability they are tested is less than 1.
$$\mathbb{E}[T|(X, Z) \in V, K = 0] < 1$$

- For patients in set V of type $K = 0$ and who aren't tested, the expected adverse events are greater than the threshold defined by the constant adverse event rate μ plus a term that relates the costs of testing and stenting to the true positive rate and false positive rate of the test.
$$\mathbb{E}[A \mid (X, Z) \in V, K = 0, T = 0] > \mu + \eta \left(\frac{c_t + \ pc_S}{q\tau^0 - c_S(q - \ p)} \right)$$

Note: If physician judgments are erroneous, $h(X, Z) \neq b(X, Z)$, then both undertested and overtested patient subsets can exist. If accurate, overtested subsets happen only if $\nu > 0$.

Definition: Undertested V

Conditions:

- Among patients characterised by set V and who are of type $K = 0$, the probability they are tested is less than 1.

$$\mathbb{E}[T|(X, Z) \in V, K = 0] < 1$$

- For patients in set V of type $K = 0$ and who aren't tested, the expected adverse events are greater than the threshold defined by the constant adverse event rate μ plus a term that relates the costs of testing and stenting to the true positive rate and false positive rate of the test.

$$\mathbb{E}[A \mid (X, Z) \in V, K = 0, T = 0] > \mu + \eta \left(\frac{c_t + \ pc_S}{q\tau^0 - c_S(q - \ p)} \right)$$

Note: If physician judgments are erroneous, $h(X, Z) \neq b(X, Z)$, then both undertested and overtested patient subsets can exist. If accurate, overtested subsets happen only if $\nu > 0$.

DATA SOURCE

- Primary data from the EHRs of a top urban academic medical center.
- Time period: January 2010 to May 2015.

DATA SELECTION AND DEFINITIONS

- Initial dataset: All Emergency Department (ED) visits within the period.
- Exclusions:
  - Age > 80 years, poor prognosis conditions.
  - Known recent blockage, immediate ED deaths.
- Sample: 246,265 ED visits by 129,859 distinct patients.
- Key Variables:
  - T_{ij}: Testing flag for patient i at visit j.
  - S_{ij}: Treatment flag.
  - Y_{ij}: Test yield.
  - K_{ij}: Contraindication flag.
  - A_{ij}: Major adverse cardiac event within 30 days.

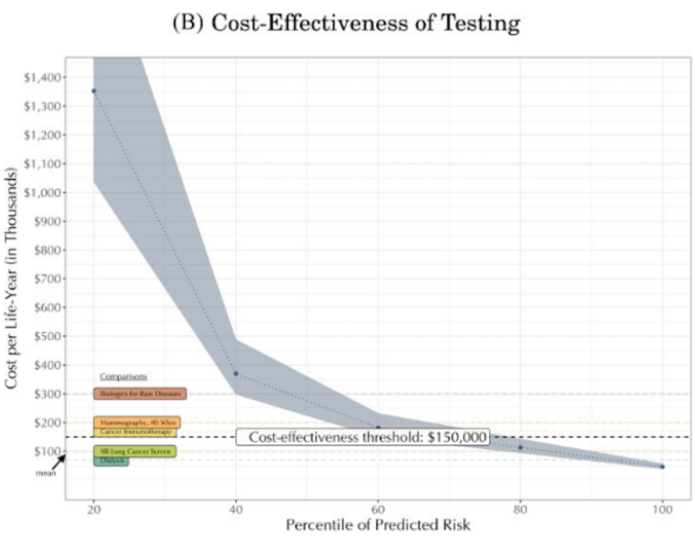
TRAINING PROCEDURE

- Objective: Develop estimator $\hat{m}(\cdot)$ using observed covariates X_{ij} to predict proxies for an (unobserved) blockage.
- Data proxies for blockage:
  - Positive test result leading to treatment S_{ij} when $T_{ij} = 1$.
  - Adverse event A_{ij} when $T_{ij} = 0$.
- Data splitting:
  - Training: 70%
  - Ensembling: 5%
  - Hold-out: 25%
- Four individual estimators:
  - Tested patients: Gradient boosted trees & LASSO to predict $S_{ij} = 1$.
  - Untested patients: Gradient boosted trees & LASSO to predict $A_{ij} = 1$.

ENSEMBLING AND FINAL MODEL

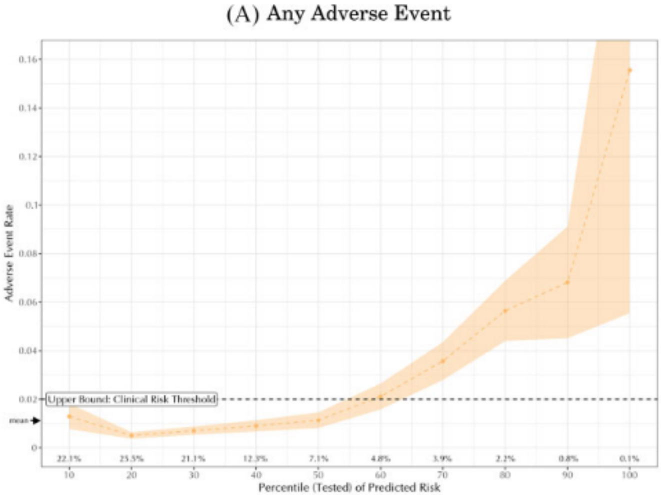
- Predictions generated in the 5% ensembling set.
- Logistic regression used to ensemble model predictions.
- Results yield final weights for LASSO and boosted trees predictions.
- Final model: Weighted ensemble $\hat{m}(X_{ij})$.

RESULTS: OVERTESTING



Cost effectiveness of testing among those tested. The y-axis shows the implied cost-effectiveness of testing patients in a bin, in units of thousands of dollars per life-year.

RESULTS: UNDERTESTING



Adverse event rates within thirty days among untested individuals (y-axis) categorised by decile bins of predicted risk (x-axis). The horizontal line marks the 2% threshold, above which clinical guidelines advise testing. Notably, the top 14% (six bins) have rates that significantly exceed 2%.

WHY DO PHYSICIANS MAKE TESTING ERRORS

Boundedness in Physician Judgements

- True risk model is complex.
- Bounded rationality could lead physicians to a simpler model analogous to regularisation.

LASSO Regularisation

- They use LASSO to model bounded rationality, choosing the penalty so that the number of variables ranges from 0 to $1,500$.

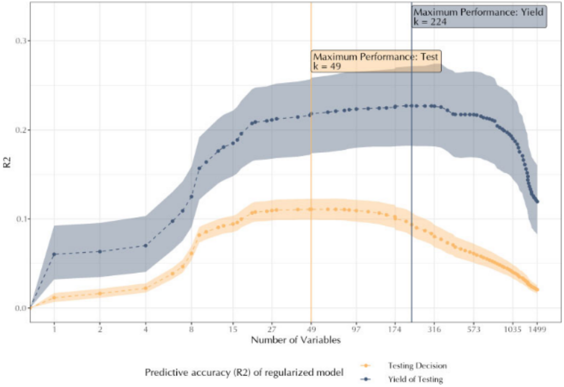


FIGURE V

Explanatory Power of Simple versus Complex Models of Risk

- Grey line: Predicts out-of-sample risk R^2.
- Yellow line: Predicts physician testing.
- Optimal for physicians: 49 variables. For risk: 224.

STATISTICAL TEST

This motivates the following statistical test. Let $\hat{m}_{simple}(X_{ij})$, which uses the 49 variables selected by LASSO, and $\hat{m}_{complex}(X_{ij})$, which uses the 224. Focusing on $[\hat{m}_{complex}(X_{ij}) - \hat{m}_{simple}(X_{ij})]$, which we can interpret as the additional risk information provided by the complex model, which we can call complex risk.

Consider models:

$$T_{ij} = \beta_0 + \beta_1 \hat{m}_{simple}(X_{ij}) + \beta_2[\hat{m}_{complex}(X_{ij}) - \hat{m}_{simple}(X_{ij})] + e_{ij},$$

$$Y_{ij} = \gamma 0 + \gamma_1 \hat{m}_{simple}(X_{ij}) + \gamma_2[\hat{m}_{complex}(X_{ij}) - \hat{m}_{simple}(X_{ij})] + e_{ij}.$$

PHYSICIANS' BOUNDED ATTENTION

If relying solely on a simple model:

1. Expect $\beta_2 = 0$.
2. Expect $\gamma_2 > 0$.

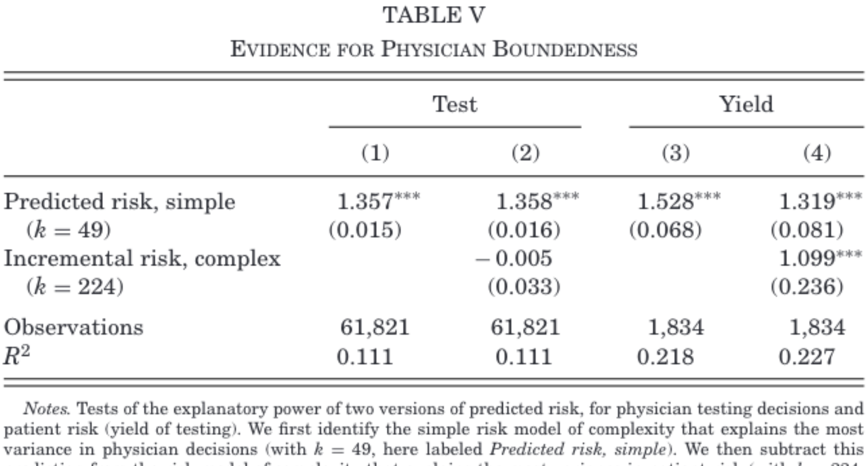


TABLE V

EVIDENCE FOR PHYSICIAN BOUNDEDNESS

	Test		Yield	
	(1)	(2)	(3)	(4)
Predicted risk, simple	1.357***	1.358***	1.528***	1.319***
($k = 49$)	(0.015)	(0.016)	(0.068)	(0.081)
Incremental risk, complex		– 0.005		1.099***
($k = 224$)		(0.033)		(0.236)
Observations	61,821	61,821	1,834	1,834
R^2	0.111	0.111	0.218	0.227

Notes. Tests of the explanatory power of two versions of predicted risk, for physician testing decisions and patient risk (yield of testing). We first identify the simple risk model of complexity that explains the most variance in physician decisions (with $k = 49$, here labeled *Predicted risk, simple*). We then subtract this

ALTERNATIVE MECHANISMS OF UNDER-/OVER-WEIGHTING

- Salience: A variable grabs attention.
- Representativeness: An event reflects salient features of its generation process.
- Example: Chest pain is seen as representative of blockage.

CONSTRUCTING NEW RISK PREDICTOR

- Reduce training data to subset of variables W.
- Train $\hat{m}_W(X_{ij})$.
- Regress test decision T_{ij} on $\hat{m}(X_{ij})$ and $\hat{m}_W(X_{ij})$.
- Positive coefficient of $\hat{m}_W$ indicates overweighting of variables in W.

Symptom Salience

- Symptoms are most salient piece of information to physicians. Often stressed in medical education.
- W restricted to variables indicating symptoms.
- Doctors overweight symptoms in decisions.

TABLE VI

Symptom Salience and Representativeness

	Test				
	(1)	(2)	(3)	(4)	(5)
Predicted risk, full	0.872***	0.715***	0.756***	0.619***	0.755***
	(0.053)	(0.049)	(0.061)	(0.045)	(0.066)
Predicted risk, subsets					
All symptoms		0.888***	0.860***	0.273***	
		(0.052)	(0.057)	(0.061)	
Representative symptoms				1.283***	
				(0.121)	
Demographics			0.139***		
			(0.031)		
Prior diagnoses			0.046**		
			(0.021)		
Prior procedures			– 0.053*		
			(0.030)		
Prior lab results and vital signs			– 0.209***		
			(0.019)		
Physician experience					
Experience (years)					– 0.0005**
					(<0.001)
Experience × risk					0.011***
					(0.005)
Observations	61,938	61,938	61,938	61,938	55,777
R^2	0.084	0.106	0.113	0.118	0.082

REPRESENTATIVENESS

$$Pr(B = 1 \mid M = 1) \times g\left(\frac{Pr(M = 1 \mid B = 1)}{Pr(M = 1 \mid B = 0)}\right),$$

- Symptoms common in patients with blockages are weighted more.
- Hypothesis: For patients with the same predict risk: patients with more (less) representative symptoms are more (less) likely to be tested.

TABLE VI

SYMPTOM SALIENCE AND REPRESENTATIVENESS

	Test				
	(1)	(2)	(3)	(4)	(5)
Predicted risk, full	0.872***	0.715***	0.756***	0.619***	0.755***
	(0.053)	(0.049)	(0.061)	(0.045)	(0.066)
Predicted risk, subsets					
All symptoms		0.888***	0.860***	0.273***	
		(0.052)	(0.057)	(0.061)	
Representative symptoms				1.283***	
				(0.121)	
Demographics			0.139***		
			(0.031)		
Prior diagnoses			0.046**		
			(0.021)		
Prior procedures			– 0.053*		
			(0.030)		
Prior lab results and vital signs			– 0.209***		
			(0.019)		
Physician experience					
Experience (years)					– 0.0005**
					(<0.001)
Experience × risk					0.011***
					(0.005)
Observations	61,938	61,938	61,938	61,938	55,777
R^2	0.084	0.106	0.113	0.118	0.082

REPRESENTATIVENESS

$$Pr(B = 1 \mid M = 1) \times g\left(\frac{Pr(M = 1 \mid B = 1)}{Pr(M = 1 \mid B = 0)}\right),$$

- Symptoms common in patients with blockages are weighted more.
- Hypothesis: For patients with the same predict risk: patients with more (less) representative symptoms are more (less) likely to be tested.

Conclusion: Symptoms as a whole may be salient, but representative symptoms drive physicians to test far more.

IMPLICATIONS AND CONCLUSIONS

- Simple views of less is more or more is less lack nuance.
- High-testing providers viewed as wasteful.
- Policy drives: incentives to test less.
- Biases by physicians challenge this policy.