An Assessment of Risk Assessment in Cases of Domestic Abuse

Jeffrey Grogger

December 2024

Motivation

- Domestic abuse is a far-reaching problem with extensive consequences for victims and others
  - Roughly 1/3 of women around the world are affected by domestic abuse (WHO (2021); National Coalition Against Domestic Violence (2014), Her Majesty's Inspectorate of Constabulary (2014))
  - It has negative consequences for:
    - Employment, earnings, welfare dependency of victims (Bhuller, et al 2021, Adams, et al 2024)
    - Health of babies in utero at time of abuse (Aizer (2011))
    - Test scores, school attendance, etc of victims' children (Bhuller, et al (2021), Gutierrez and Molina (2020))

Preventing recidivism

- What can be done to prevent recidivism?
- In many jurisdictions, police carry out risk assessment
- In England and Wales, risk assessment is universal
  - Police administer quex ("DASH") at the scene
  - Predict risk of serious recidivism
  - Intervene in cases at high risk

Potential benefits of risk assessment

- May provide support, services in cases where incident is not classified as crime, charges are not pressed
- Potentially provides help to victims who need it, outside of constraints imposed by legal system

Questions

- Does it work?
  - Does it predict?
  - Does it protect?
- Can we explain officers' predictions?
- Discuss findings from three papers that bear on these questions
  - Based on work with Dan A. Black, Sean Gupta, Ria Ivandic, Andy Jordan, Tom Kirchmaier, and Koen Sanders
  - They are not to blame for conclusions below

Common features of studies

- Data from Greater Manchester Police
  - Population approx. 3 million
  - 3rd-largest force in England and Wales
- Data from 2 sources
  - Administrative records
  - DASH risk assessment protocol

Administrative records

- All DA-related calls, regardless of disposition
- Info on incident, participants' DA and criminal histories
- Observed outcome: $Y_i = 1$ if there is a repeat call for DA incident involving serious violence or sex offense within 12 months

Domestic Abuse, Stalking, and Harassment

Risk assessment protocol in use throughout England and Wales starting in 2009

- Procedure
  - Step 1: Officer administers standard 28-item quex
  - Step 2: Based on quex and own judgment, classifies case as standard-, medium-, or high-risk
- Definition of high risk
  - "There are identifiable indicators of risk of serious harm. The potential event could happen at any time and the impact would be serious."
- High-risk cases are provided with protective services

DASH questions and frequencies

Table 2: DASH Questions, Response Frequencies, and Distribution of Risk Assessments

(a) DASH questions and Response Frequencies

	Y	N	O
Current incident result injury?	0.136	0.737	0.127
Victim very frightened?	0.283	0.547	0.170
What victim frightened of?	0.299	0.519	0.182
Victim feel isolated?	0.126	0.689	0.185
Victim feel depressed?	0.174	0.637	0.190
Victim separate from the abuser?	0.451	0.369	0.180
Any conflict over child contact?	0.151	0.676	0.173
Abuser stalks the victim?	0.190	0.622	0.188
Victim pregnant recently?	0.162	0.665	0.172
Any children not the abuser's?	0.154	0.673	0.174
Abuser ever hurt children?	0.028	0.785	0.187
Abuser ever threatened children?	0.025	0.786	0.189
Abuse happening more often?	0.205	0.600	0.194
Abuse getting worse?	0.198	0.607	0.195
Abuser tried to control victim?	0.254	0.548	0.198
Abuser ever used weapons?	0.093	0.707	0.200
Abuser ever threatened to kill?	0.103	0.695	0.202
Abuser ever strangle/choke/etc.?	0.132	0.665	0.202
Abuse of a sexual nature?	0.074	0.721	0.206
Other person threatened the vic?	0.034	0.764	0.201
Abuser hurt anyone else?	0.116	0.683	0.202
Abuser ever mistreated animal?	0.036	0.763	0.201
Any financial issues?	0.166	0.637	0.197
Abuser had problems alcohol, etc.?	0.373	0.435	0.192
Abuser ever threatened suicide?	0.171	0.626	0.203
Abuser ever breached bail?	0.109	0.691	0.199
Abuser trouble with the police?	0.448	0.365	0.188
Other relevant information?	0.186	0.674	0.140

Sample

- Focus on male-female IPV cases
- Roughly 154,000 incidents over period 2014-2019

Question 1: Does DASH predict?

Approach

- Assess performance of officer's prediction (Step 2 of DASH protocol)
- Assess performance of different statistical models
  - Vary sets of predictors
  - Vary relative weights assigned to prediction errors

Assessing the DASH prediction

Table 4: Violent Recidivism by DASH Risk Assessment

	Lesser Risk	*High Risk*	*Row Share*	*Class Error*
Actual no	13,121	1,165	0.882	0.082
Actual yes	1,702	215	0.118	0.888
Col share	0.915	0.085	1	
Pred error	0.115	0.844		0.177

NOTE: AUC = 0.515.

Measures of performance

- Area under the curve (AUC)
- False negatives

Measures of performance

Area under the curve

- Based on signal detection theory
  - AUC $= 1$ implies perfect prediction
  - AUC $= .5$ implies random guessing
- Computed using FPR, $TPR = 1 - FNR$
  - AUC $= .515$
  - Pretty close to random guessing

Measures of performance

False negatives

- Reason to think that FNs are more costly than FPs
  - These are cases where cop predicts no recidivism, but recidivism occurs
  - Victim who could have benefited from protective services was not offered them
- FNR = .888
- Police miss the vast majority of cases where victim later suffers serious recidivism

Can we do better with a predictive algorithm?

Focus on random forests

- RF is adaptive non-linear regression estimator
- Good at learning functional form
- Can incorporate asymmetric cost of error
  - Believe cost(FN) $>$ cost(FP), but don't know actual numbers
  - For RF, only relative costs matter
  - Train models at 5:1, 10:1, and 15:1
  - Train on DASH only, admin only, both

Key takeaways

- RF based on DASH quex items performs better
  - AUC up to 0.60, FNR down to 0.28
- RF based on admin variables performs better still
  - AUC up to 0.64, FNR down to 0.24
- Adding DASH to admin variables doesn't change performance much

Conclusions

- Bad news: DASH protocol appears to predict poorly
  - Just better than chance
- Good news: Statistical models perform better

Question 2: Does DASH protect?

- Might expect not, since it appears poorly targeted
- Paradoxically, seemingly poor targeting could be result of effective interventions
  - Return to this below
- For now, analyze effect of intervention

Objectives

- Estimate ATT of DASH-based intervention on serious recidivism
- Estimate ATT of criminal charges, for comparison
- Estimate heterogeneity in treatment effects
  - Concerns about backfiring
  - Even absent backfiring, could help target resources

The evaluation problem

- Interventions may be correlated with characteristics of participants that predict recidivism
- If interventions are targeted toward cases which would be more likely to recidivate in the absence of the intervention, simple treatment-comparison group contrasts may understate effect

Methods

- Estimate the ATTs via IPW weighting
- Estimate heterogeneous treatment effects using causal forests

Assumptions

- Common support
  - Range of propensity scores similar for T, C groups
- Conditional independence
  - Conditional on predictors, treatment mean-independent of potential outcomes
  - Assignment to treatment effectively random, conditional on propensity score

Conditional independence

- Strong assumption
- Case for why it may hold
  - Many covariates that should (and do) strongly predict treatment status:
  - Charges
    - Variables implying assault, battery, stalking, and harassment (75% of charges)
    - Variable indicating high level of urgency
    - Measure of cop preferences
  - High-risk designation
    - Entire DASH quex
    - Measure of cop preferences

Conditional independence

- Case for why it may hold (cont.)
  - Four approaches to estimating propensity scores yield similar results
  - All propensity scores greatly reduce imbalance. Some achieve balance
  - All propensity scores greatly reduce imbalance on out-of-model covariates. Some achieve balance
  - Remaining unobservables may balance if officers base decisions on hunches, fuzzy recollections of prior cases

Results

Estimated ATTs

Table 3: Estimates of the average treatment effect on the treated (ATT)

A. Charges				
Weights from:	Logistic regression		Random Forest	CBPS
Predictors	Baseline	Expanded	Baseline	Baseline
Charges	-0.059	-0.048	-0.048	-0.05
	(0.005)	(0.004)	(0.003)	(0.004)
B. High-risk				
Weights from:	Logistic regression		Random Forest	CBPS
Predictors	Baseline	Expanded	Baseline	Baseline
High-risk	-0.012	-0.013	-0.006	-0.007
	(0.012)	(0.008)	(0.005)	(0.008)

Notes: ATT estimates are coefficients are from a weighted regression of the violent recidivism indicator on the treatment variable, using ATT weights from the indicated statistical model. Standard errors are clustered at the level of the dyad.

Robustness

Issues

- Do interventions interact?
- Do the interventions exhibit dynamics?
- Is this all about short-term incapacitation while detained?
- Are we learning about changes in violence, or changes in reporting of violence?

Heterogeneous treatment effects

- Approach
  - Estimate causal forest to generate idiosyncratic CATEs
    - CATE is the expected treatment effect for the ith case, given observables for that case
  - Use exhaustive search to find low-dimensional decision rule identifying groups with different ATT's

Key takeaways

- Meaningful TE heterogeneity for charges
  - Group-specific ATTs range from -0.03 to -0.12
  - Results replicate to independent test sample
- No meaningful heterogeneity for intervention based on risk assessment
- No evidence of backfiring

Conclusions

ATT's

- Pressing charges has favorable effect on violent recidivism
  - Effect is large: -0.05 $\sim$ 40 percent of base recidivism rate
  - Meaningful heterogeneity
- DASH protocol has zero effect
  - Point estimate is -0.007
  - No meaningful heterogeneity

Question 3: What explains poor predictive performance?

Potential explanations

- Simple answer to question: officers don't get much training
  - "Very few police respondents in the three forces could recall receiving training relating specifically to risk assessment." (Robinson et al., 2016)
- If officers were Bayesian, they would learn from the data
  - Problem: they don't get much feedback
  - "Supervision and feedback that could reinforce any learning was also largely absent in the three study forces." (*ibid.*)
- In other words, the environment is ripe for heuristic reasoning, cognitive biases to affect predictions

Objectives

- Measure predictive skill among officers
  - Restrict attention to officers with at least 100 cases
- Analyze officer predictions, identify cognitive biases

Censoring problem

- First need to solve censoring problem
- Risk assessment, intervention may convert some latent TPs into observed FPs
  - This artificially reduces predictive performance
  - This is a type of censoring, also known as selective labels problem

Problem

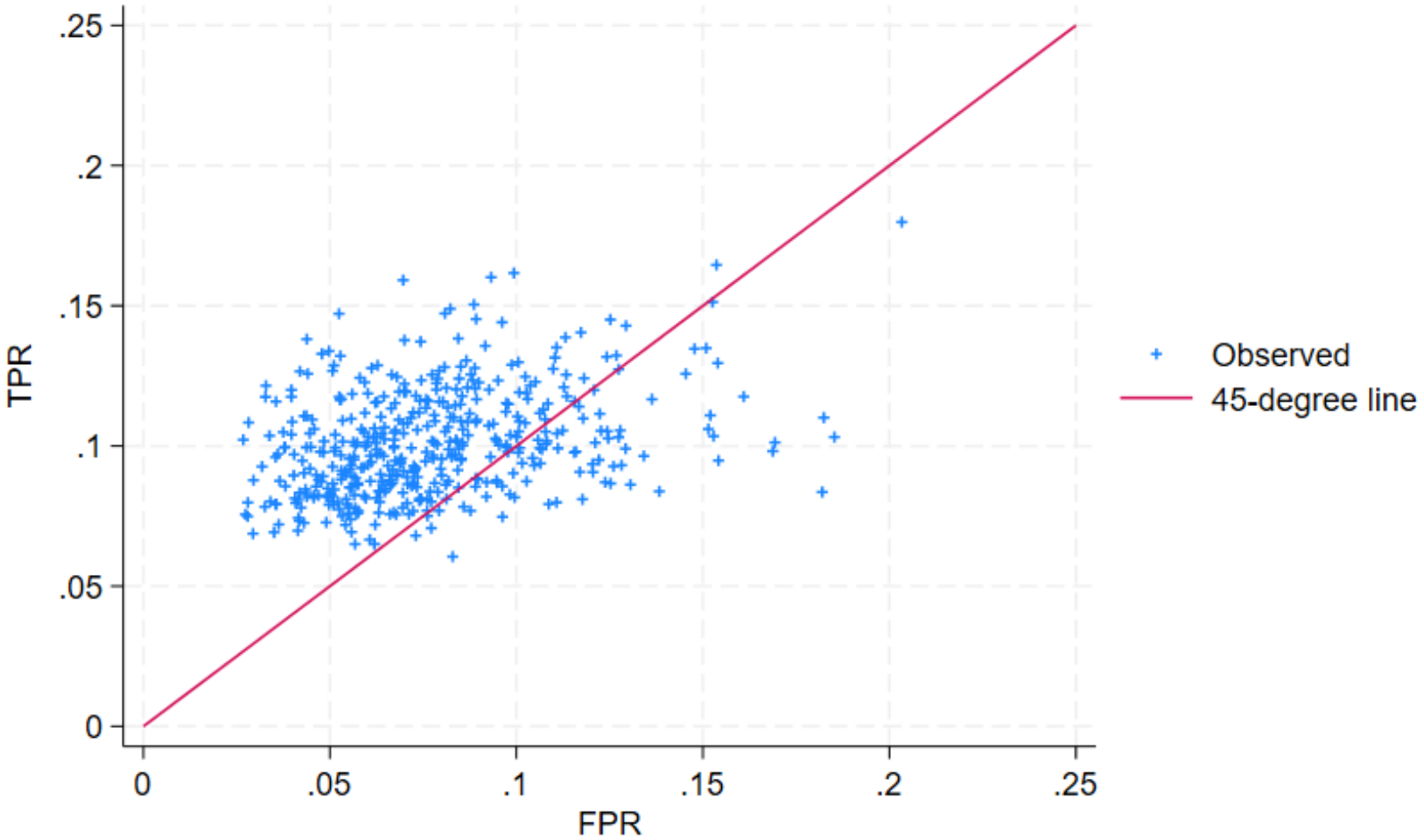
- Current solutions require random assignment of DM to cases
- Cops are not randomly assigned to incidents
  - Suspect this is not the only such case
- Need alternative solution to censoring problem
- We propose using causal forests

Solving the censoring problem via causal forests

- Causal forests estimate idiosyncratic treatment effects (CATEs) from observed outcome, treatment status, predictive covariates
- We reverse engineer the process, using observed outcome, treatment status, and estimated CATEs to impute latent outcome

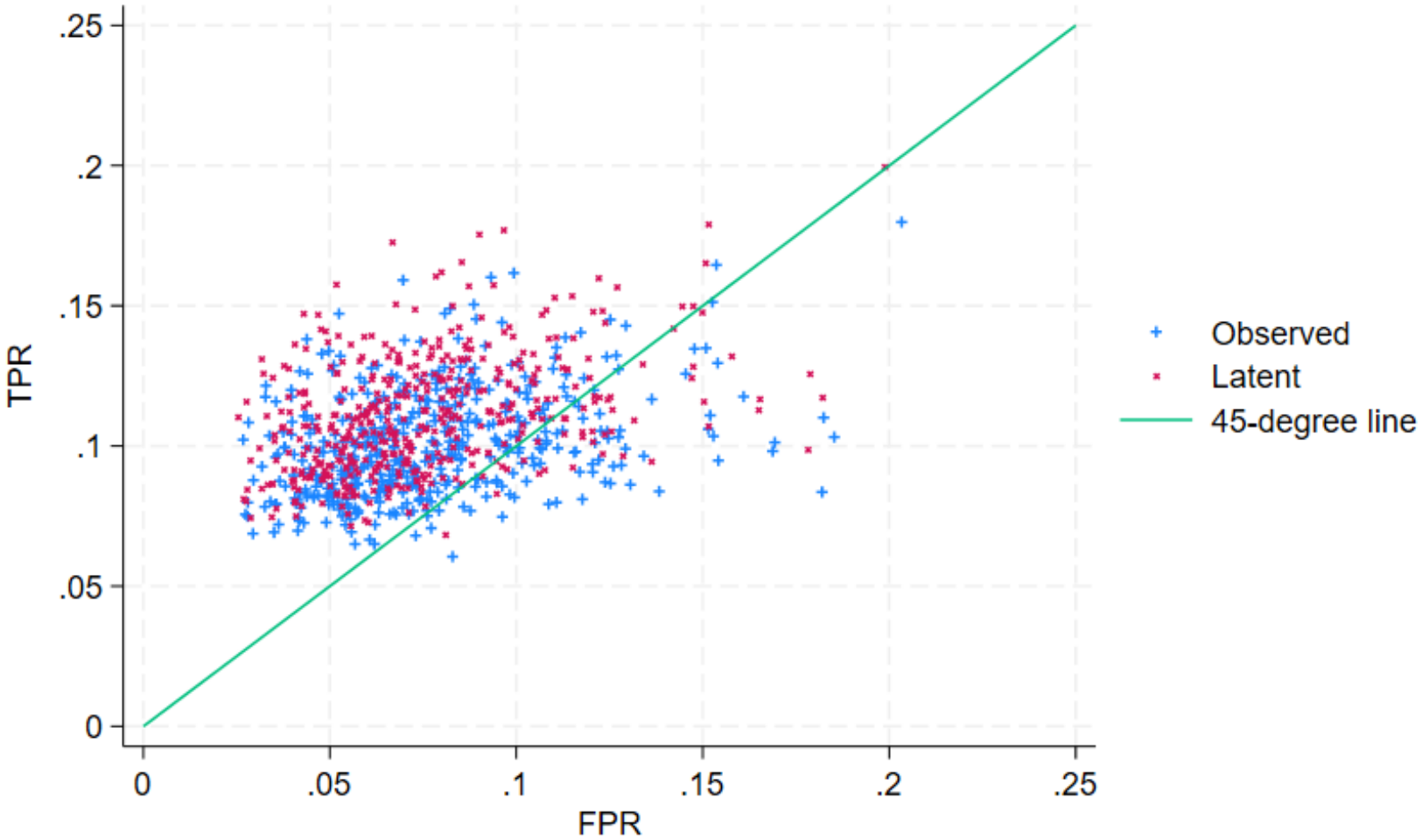
Distribution of officer performance

Observed data



Distribution of officer performance

Observed and imputed data



Distribution of officer performance, cont.

Differences between observed and imputed data

- Imputation works as expected
  - Concern was that interventions converted latent TPs into observed FPs
  - Plot based on imputed data has higher TPs, lower FPs than that based on observed data
- Still conclude that officer skill is fairly low

Heuristic reasoning

Regression models

- We estimate:

$$D_i = Q_i\alpha + Z_i\phi + \varepsilon_i \tag{1}$$

$$Y_i^* = Q_i\beta + Z_i\psi + u_i \tag{2}$$

- D_i denotes cop's prediction (=1 if high-risk, =0 otherwise)
- Y_i^* is latent recidivism
- Q_i is vector of dummies characterizing DASH traits
- Z_i is vector of covariates from admin records

Predictions, recidivism as function of DASH Qs

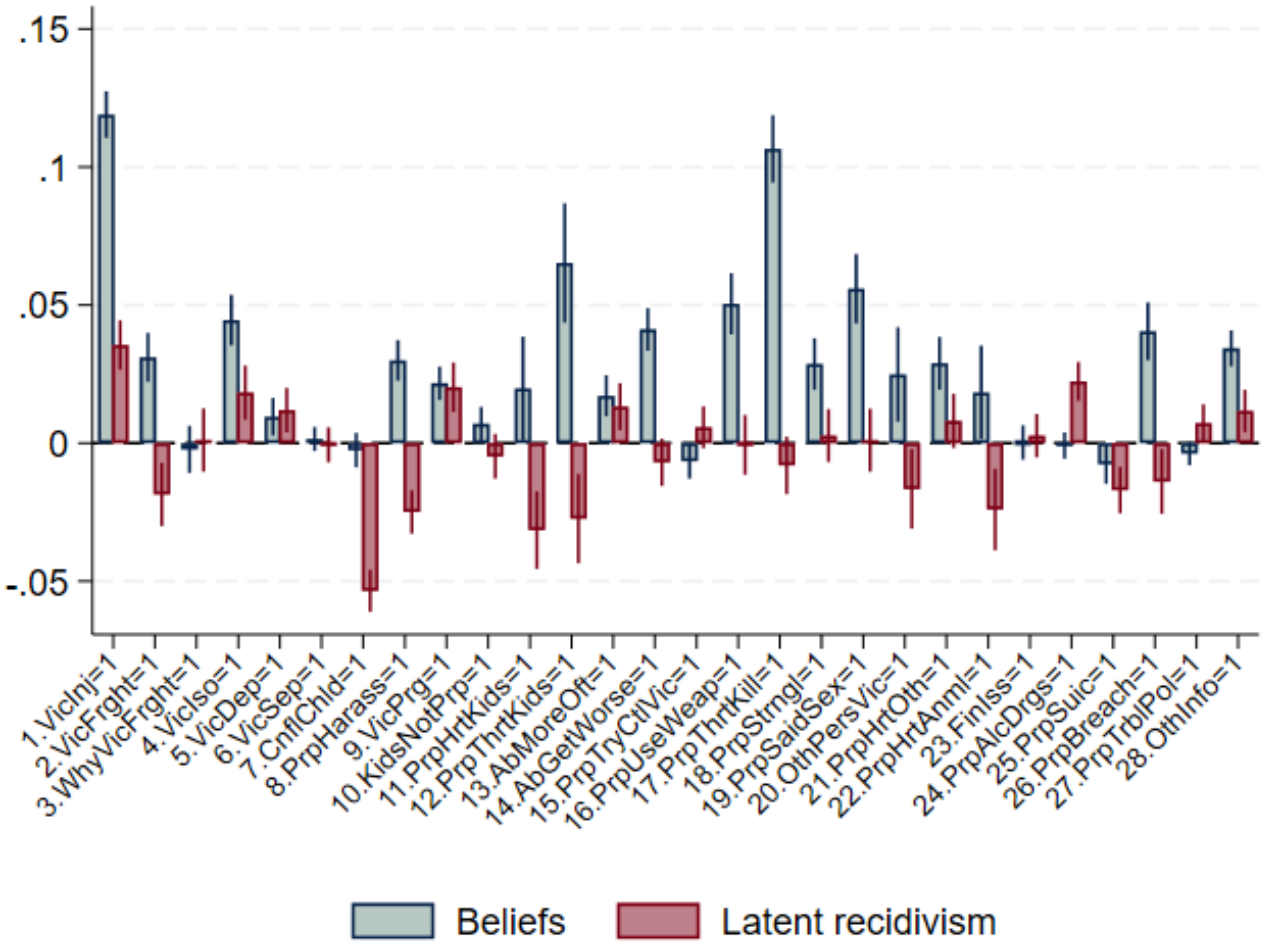


Figure: Coefficients on DASH traits from regressions of high-risk indicator and latent recidivism on DASH traits and Z

Three main patterns

- Poor questionnaire design/bad learning environment
- Over-reaction/salience
- Representativeness bias

Poor questionnaire design/bad learning environment

- Almost all coefficients from beliefs regression are positive, as face-value reading of quex suggests they should be
- 13 coefficients from outcome regression are *negative*
- Points to either bad questionnaire design or consequence of poor learning environment

Overreaction

- Quex design aside, cops generally overreact to incident traits
- Prominent examples:
  - Victim injury (Question 1): $\hat{\alpha} - \hat{\beta} = 0.084(0.005)$
  - Perp. threatens kids (Question 12): $\hat{\alpha} - \hat{\beta} = 0.092(0.012)$
  - Perp. has threatened to kill victim (Question 17): $\hat{\alpha} - \hat{\beta} = 0.115(0.007)$

Salience

- Tversky, Kahneman (1974) and Bordalo etal (2012) argue that DMs may form beliefs on the basis of traits that come most readily to mind
- Seems to fit 3 three traits involving greatest overreaction
  - English police declare they will "cause the peace to be kept ... and prevent all offences against people and property." (HMIC 2014)
  - These traits all imply desire to offend against a person
  - Victim injury demonstrates capability to bring such desires to fruition

Representativeness bias/stereotyping

- Tversky, Kahneman (1974) propose that DMs may form beliefs based not on true distribution of traits, but on distribution that overweights traits that are representative of event that DM is trying to predict
- A trait is representative if it is more prevalent among recidivist cases than among non-recidivist cases

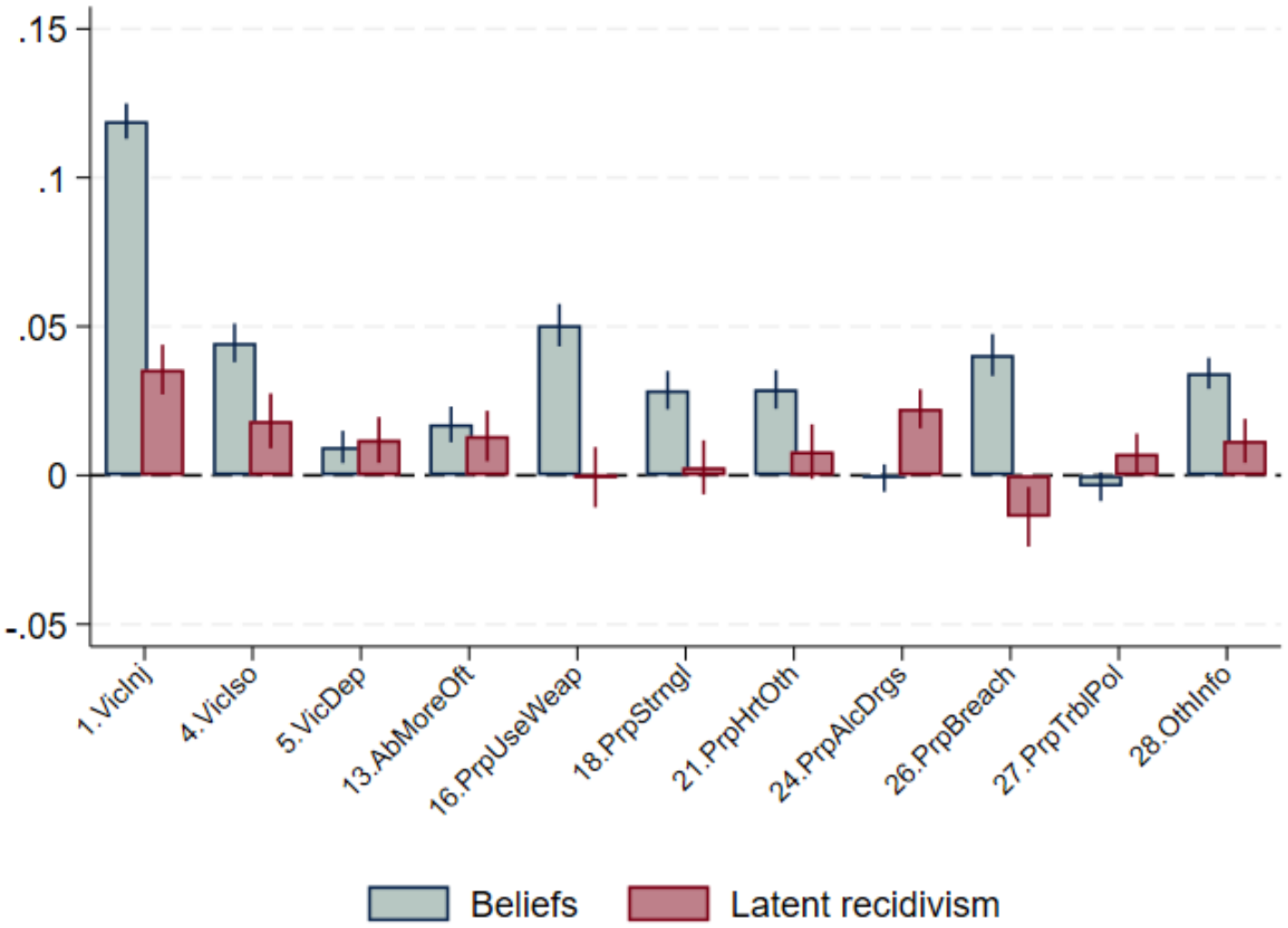
Representativeness bias/stereotyping

- Bordalo et al (2016) refine this hypothesis
  - Traits which are highly prevalent should not distort beliefs much
  - Those which are rare may distort them a great deal
  - Adding weight to prevalent trait may not change distribution of traits much; adding weight to rare trait may change it a lot

Representative traits, by rare/prevalent

Rare (prevalence $<$ .17)	Prevalent (prevalence $>$ .37)
1. VicInj	24. PrpAlcDrgs
4. VicIso	26. PrpTrblPol
5. VicDep	
13. AbMoreOft	
16. PrpUseWeap	
18. PrpStrngl	
21. PrpHrtOth	
26. PrpBreach	
28. OthInfo	

Results


.15
.1
.05
0
-.05
1.VicInj
4.VicIso
5.VicDep
13.AbMoreOft
16.PrpUseWeap
18.PrpStrngl
21.PrpHrtOth
24.PrpAlcDrgs
26.PrpBreach
27.PrpTrblPol
28.OthInfo
Beliefs
Latent recidivism

Results, cont.

- All representative/rare traits move predictions positively, significantly
- Neither representative/prevalent trait moves predictions at all

Results, cont.

- Alcohol and drug issues positively predict recidivism, but police miss it
- They are also quite commonplace, involved in nearly 40 percent of DA incidents
- Result is substantial number of victims at risk who are not provided with protective services

Robustness

- Estimates are consistent if omitted variables are uncorrelated with DASH questions
- Consider three checks
  - Cop fixed effects
  - Predictive observables
  - Use Q28 to infer unobservables

Conclusions

- Findings
  - Predictive performance is generally low
  - Seem to rely on heuristics: salience effects, representativeness bias

Policy Issues

- Main question for policy: what to do about risk assessment?
  - Abandon it
  - Provide better training to officers
  - Replace human judgment with algorithmic prediction

Should we abandon risk assessment?

- Key benefit is ability to provide resources to victims whose cases aren't prosecuted
- Seems way too valuable to give up on

Should we provide better training to officers?

- Key question is how much training for how much improvement?
- Consider the case of diagnostic radiologists
  - Read diagnostic images, predict presence of disease
  - Training
    - 4 years med school
    - 4 years as resident
    - Optional 1 or 2 years as fellow
  - In predicting pathology from chest x-rays, algorithms do better than almost 2/3 of radiologists (Agarwal et al 2023)

Should we provide better training to officers? cont.

- Now consider the case of police
  - Training
    - Chicago PD recruits spend 6 weeks at academy
    - London Metropolitan Police spend 16 weeks in the classroom
    - Training covers all aspects of their job
  - Hard to imagine providing enough training for cops to do as well as radiologists

Should we replace human judgment with algortihmic predictions?

- Pluses and minuses of human judgment
  - Plus: use information not available to algorithm
  - Minus: use irrelevant information, use relevant information inconsistently
- Pluses and minuses of algorithms
  - Plus: use information consistently, downweight irrelevant information
  - Minus: only use information made available to it
- Conceptually, not clear who should do better

Should we replace human judgment with algortihmic predictions? cont.

- Empirically, evidence is very clear
  - Grove et al (2000) analyzed 136 studies, showed that algorithmic predictions at least weakly outperform human judgment 3/4 of the time
  - More recent studies show that human DMs usually do worse when they overturn algorithmic recommendations (Agarwal et al 2023; Angelova et al 2023; Stevenson and Doleac 2022)
  - Analyses above show that police risk assessment in cases of domestic abuse is no exception to rule

Objections

The algorithm will make mistakes

- All systems, human or algorithmic, will make mistakes
- Mistakes will cause suffering, and each mistake will be tragic
- Goal should be to make as few mistakes as possible
- All evidence shows that algorithmic predictions have the edge

Robinson Amanda L., Myhill Andy, Wire Julia, Roberts Jo, Tilley Nick. Risk-led policing of domestic abuse and the DASH risk model. 2016. 1–46.

Tversky A., Kahneman D. Judgment under Uncertainty : Heuristics and Biases Linked references are available on JSTOR for this article : Judgment under Uncertainty : Heuristics and Biases // Science. 1974. 185, 4157. 1124–1131.