ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistics in Society
Series A

J. R. Statist. Soc. A (2020)

Simple rules to guide expert classifications

Jongbin Jung,

Stanford University, USA

Connor Concannon,

John Jay College of Criminal Justice, New York, USA

Ravi Shroff,

New York University, USA

Sharad Goel

Stanford University, USA

and Daniel G. Goldstein

Microsoft Research, New York, USA

[Received February 2018. Final revision March 2020]

Summary. Judges, doctors and managers are among those decision makers who must often choose a course of action under limited time, with limited knowledge and without the aid of a computer. Because data-driven methods typically outperform unaided judgements, resource-constrained practitioners can benefit from simple, statistically derived rules that can be applied mentally. In this work, we formalize long-standing observations about the efficacy of improper linear models to construct accurate yet easily applied rules. To test the performance of this approach, we conduct a large-scale evaluation in 22 domains and focus in detail on one: judicial decisions to release or detain defendants while they await trial. In these domains, we find that simple rules rival the accuracy of complex prediction models that base decisions on considerably more information. Further, comparing with unaided judicial decisions, we find that simple rules substantially outperform the human experts. To conclude, we present an analytical framework that sheds light on why simple rules perform as well as they do.

Keywords: Heuristics; Judgement and decision making; Policy evaluation; Sensitivity analysis

1. Introduction

In field settings, decision makers often choose a course of action based on experience and intuition rather than on statistical analysis (Klein, 2017). This includes doctors classifying patients on the basis of their symptoms (McDonald, 1996), judges setting bail amounts (Dhami, 2003) or making parole decisions (Danziger *et al.*, 2011) and managers determining which ventures will succeed (Åstebro and Elhedhli, 2006) or which customers to target (Wübben and Wangenheim, 2008). Despite the prevalence of this approach, a large body of work shows that in many domains intuitive inferences are inferior to those based on statistical models (Meehl, 1954;

Address for correspondence: Sharad Goel, Department of Management Science and Engineering, Huang Engineering Center, Stanford University, 475 Via Ortega Avenue, Stanford, CA 94305-6015, USA.
E-mail: scgoel@stanford.edu

 0964–1998/20/183000

Dawes, 1979; Dawes *et al.*, 1989; Camerer and Johnson, 1997; Tetlock, 2005; Kleinberg *et al.*, 2015, 2017).

In this work, we generalize from research on improper linear models (Einhorn and Hogarth, 1975; Green, 1977; Dawes, 1979; Gigerenzer and Goldstein, 1996; Waller and Jones, 2011) to suggest a straightforward method for constructing simple yet accurate decision rules that experts can apply mentally. This *select, regress and round* method results in rules that are fast, frugal and clear: fast in that decisions can be made quickly in one's mind, without the aid of a computer; frugal in that they require very little information to reach a decision; and clear in that they expose the grounds on which classifications are made.

Decision rules satisfying these criteria have many benefits. Fast rules that can be applied mentally reduce transaction costs, encouraging persistent use. In medicine, frugal rules require fewer tests, which saves time, money and, in the case of triage situations, lives (Marewski and Gigerenzer, 2012). Frugal decision rules incorporating predictors that are broadly related to outcomes of interest are well suited for settings in which a model that is highly customized for one population may not generalize to other populations (Wyatt and Altman, 1995). The clarity of simple rules provides insight into how systems work and exposes where models may be improved (Gleicher, 2016; Sull and Eisenhardt, 2015), which may encourage adoption of such tools in clinical settings (Wyatt and Altman, 1995). Clarity can even become a legal requirement when society demands to know how algorithmic decisions are being made (Goodman and Flaxman, 2016; Corbett-Davies *et al.*, 2017).

After describing the select, regress and round method, we evaluate its efficacy on 21 data sets from the University of California, Irvine (UCI), Machine Learning Repository and show that in many cases simple rules are competitive with state of the art machine learning algorithms. To illustrate in detail the value of simple rules, we present a case-study of judicial decisions for pretrial release. On the basis of an analysis of over 100000 cases, we show that simple rules substantially improve on the efficiency and equity of unaided judicial decisions. In particular, we estimate that judges can detain a third fewer defendants while simultaneously increasing the number who appear at their court dates. In the judicial context, as in many policy settings, it is statistically challenging to evaluate decision rules based solely on historical data. The key difficulty is that we cannot observe what would have happened under an alternative course of action. What would have happened, for example, if one released a defendant who in reality was detained? We address this issue by first estimating the relevant counterfactual outcomes, and then assessing the sensitivity of our estimates to unobserved confounding, generalizing the technique of Rosenbaum and Rubin (1983a).

Our results add to a growing literature on *interpretable machine learning*. In addition to methods for better understanding complex machine learning models and data structures (Kim *et al.*, 2015; Ribeiro *et al.*, 2016), several methods have been introduced to construct interpretable decision rules, similar to the simple decision rules that we discuss here. For example, Van Belle *et al.* (2012) used convex optimization to build interval-coded scoring models for binary outcomes. More general methods for constructing interpretable decision rules have been recently proposed, including the supersparse linear integer model called 'SLIM' (Ustun and Rudin, 2016), Bayesian rule lists (Wang and Rudin, 2015) and interpretable decision sets (Lakkaraju *et al.*, 2016). These methods all produce rules that are easy to interpret and to apply but the methods differ considerably on the ease of rule creation. As an important practical consideration, the method that we investigate here can be carried out by a practitioner without extensive training in statistics, using popular open-source software—though it bears emphasis that appropriate application of all statistical methods requires both domain knowledge and familiarity with the relevant data.

The data that are analysed in the paper and the programs that were used to analyse them can be obtained from

`https://rss.onlinelibrary.wiley.com/hub/journal/1467985x/series-a-datasets`.

2. Select, regress and round: a simple method for creating simple rules

We begin by presenting a simple method—which we call *select, regress and round*—for constructing simple decision rules. This procedure generalizes ideas that appear throughout the judgement and decision-making literature on improper linear scoring rules and formalizes heuristics that are used by practitioners in creating decision aids.

The rules that we construct are designed to aid classification or ranking decisions by assigning each item in consideration a score z, computed as a linear combination of a subset S of the item features:

$$z = \sum_{j \in S} w_j x_j,$$

where the weights w_j are integers. Often, one seeks to make dichotomous decisions (e.g. whether to detain or to release an individual pending trial), which amounts to setting a threshold and then taking a particular course of action if and only if the score is above that threshold.

In the cases that we consider, the features themselves are typically 0–1 indicator variables (indicating, for example, whether a person is male, or whether an individual is 26–30 years old), and so the rule reduces to a weighted checklist, in which one simply sums up the (integer) weights of the applicable attributes. Although it is possible to apply select, regress and round to continuous features directly, in the spirit of simplicity and interpretability, we recommend discretizing continuous covariates, using, for example, three equal-sized bins, as proposed in Gelman and Park (2009). But in practice, as always, domain knowledge and technical considerations play an important role in determining appropriate transformations or discretization schemes. For example, rather than simply partitioning an age covariate into three bins, one might use 10-year buckets. Similarly, one might collapse categorical features with several levels into a smaller number of more semantically meaningful groups.

This class of rules has two natural dimensions of complexity: the number of features that are included in the subset S and the magnitude of the weights. Given integers $k \geqslant 1$ and $M \geqslant 1$, we apply the following three-step procedure to construct rules with at most k features and integer weights bounded by M (i.e. $|S| \leqslant k$ and $-M \leqslant w_j \leqslant M$).

Step 1: *select*—from the full set of features, select k features via forward stepwise selection. This is done by iteratively adding the feature that minimizes the Akaike information criterion AIC. For fixed k, we note that standard selection metrics (e.g. AIC or the Bayesian information criterion BIC) are theoretically guaranteed to yield the same set of features.
Step 2: *regress*—using only these k selected features, train an L^1-regularized (lasso) logistic regression model on the data, which yields (real-valued) fitted coefficients $\beta_1, \ldots, \beta_k$.
Step 3: *round*—rescale the coefficients to be in the range $[-M, M]$, and then round the rescaled coefficients to the nearest integer. Specifically, set

$$w_j = \text{round}\left(\frac{M \beta_j}{\max_i |\beta_i|} \right).$$

This select, regress and round method for rule construction extends research on unit-weighted

linear models by incorporating feature selection and by adopting more general integer weights to generate a richer family of rules, the accuracy of which we examine in the next section. In practice, we recommend that developers of such rules apply the procedure for a range of small values of k and M that are appropriate to their domain, and then pick the values that perform best on context-specific metrics, balancing simplicity with performance: an approach that we illustrate below.

We note that rules that are constructed in this way may have fewer than k features, since the lasso regression in step 2 may result in coefficients that are identically 0, and rescaling and rounding coefficients in step 3 may zero-out additional terms. For step 2, the regularization parameter λ is chosen via cross-validation. In our applications, following Friedman *et al.* (2010), we explore a regularization path with 1000 values of λ spaced evenly on the log-scale in the range $(\lambda_{\min}, \lambda_{\max})$, where $\lambda_{\min} = 10^{-4}$ and $\lambda_{\max}$ is selected as the minimum value such that all coefficients are regularized to 0. Unlike λ, the parameters k and M cannot be selected via an automated procedure unless we formally quantify the trade-off between performance and simplicity, since both performance and complexity increase with larger values of k and M. However, in practice, as we show below, we might achieve approximately the same performance as a traditional logistic regression model with relatively small values of k and M, meaning that the trade-off may be negligible.

3. Evaluating the efficacy of simple rules

We apply the select, regress and round procedure to 21 publicly available data sets to examine the trade-off between rule complexity and performance. These data sets all come from the UCI Machine Learning Repository (Table 1) and were selected according to four criteria:

(a) the data set involves binary classification (as opposed to a regression problem), where we set the plurality class as the target of prediction for those data sets whose outcome variable takes more than two values;
(b) the data set is provided in a standard and complete form;
(c) the data set involves more than 10 (binarized) features;
(d) the classification problem is a problem that a human could plausibly learn to solve with the given features.

For example, we included a data set in which the task was to determine whether cells were malignant or benign on the basis of various biological attributes of the cells, but we excluded image recognition tasks in which the features were represented as pixel values. This fourth requirement limits the scope of our analysis and conclusions to domains in which human decision makers typically act without the aid of a computer.

3.1. Benchmarking to complex prediction models

We benchmark the performance of our simple rules against three standard statistical models: logistic regression, L^1-regularized logistic regression and random forests. The random-forests method, in particular, is considered to be one of the best off-the-shelf classification algorithms in machine learning (Fernández-Delgado *et al.*, 2014; Kleinberg *et al.*, 2017). These models were fitted in R with the `glm`, `glmnet` and `randomForest` packages respectively. For the L^1-regularized logistic regression models, the `cv.glmnet` method was used to determine the best value of the regularization parameter λ with nested tenfold cross-validation and 1000 values of λ. We used 1000 trees for the random-forest models.

Across the 21 UCI data sets, variables are documented as *categorical* (discrete and

Table 1. Summary of UCI data sets†

Domain	*Instances*	*Features*	*Complete instances*	*Continuous features*	*Binarized features*	*Proportion positive*
1, adult	32561	14	30162	4	96	25
2, annealing	798	38	798	7	54	76
3, audiology-std	200	41	190	0	55	24
4, bank	41188	20	41188	9	62	11
5, bankruptcy	250	6	250	0	13	43
6, car	1728	6	1728	0	16	70
7, chess-krvk	28056	6	28056	0	35	10
8, chess-krvkp	3196	36	3196	0	37	52
9, congress-voting	435	16	232	0	17	53
10, contrac	1473	9	1473	2	20	43
11, credit-approval	690	15	653	6	44	45
12, ctg	2126	38	2126	33	67	78
13, cylinder-bands	541	39	279	19	65	65
14, dermatology	366	34	358	34	69	31
15, german_credit	1000	20	1000	7	56	70
16, heart-cleveland	303	13	299	6	26	46
17, ilpd	583	10	579	9	20	72
18, mammo	961	5	830	1	18	49
19, mushroom	8124	22	5644	0	76	38
20, wine	178	13	178	13	27	40
21, wine_qual	6497	12	6468	11	24	63

†For each domain, we report the name of the data set, the number of rows and features (columns excluding the class label) in the original data set, the number of complete rows with no missing data, the number of continuous features, as well as the number of features after discretizing continuous variables and expanding categorical variables to binary indicators, and the proportion of instances in the target class (proportion positive). The context of each domain is presented in detail in Appendix A.

unordered) *ordinal* (discrete and ordered) or *continuous*. For our simple rules, we represent discrete covariates—both categorical and ordinal—as a series of binary indicator variables, with one indicator per category. In particular, for simplicity, we ignore the category ranking in ordinal variables. Further, all continuous features are discretized into three approximately equal-sized bins representing (categorical) low, medium and high values of the feature, following Gelman and Park (2009). For the three complex models, we include the above feature representations, as well as the original (non-discretized) values of continuous variables. Also, for ordinal variables—in addition to their unordered categorical representation—we include a feature representation that preserves the order of categories. As is common, the categories of an ordinal variable are represented as sequential integers, with our complex models fitting orthogonal polynomials to these integer values (Chambers *et al.*, 1992).

On each of the UCI data sets that we analyse here, we construct a family of simple rules having $k \in \{1, \ldots, 10\}$ features, with feature weights bounded by $M \in \{1, 2, 3\}$. We count the number of features k before binarization. For example, a categorical covariate with five possible values—and hence converted to five binary variables—counts as one of the k features in the simple rule, not five. The head-to-head comparison with complex models provides a difficult test for the simple rules in part because the simple rules can only base their predictions on 1–10 features. The complex models, in contrast, can train and predict with all the features in a domain, which number between 5 and 41 with a mean of 20. We provide the complex models with an additional advantage over the simple rules by including continuous and ordinal features

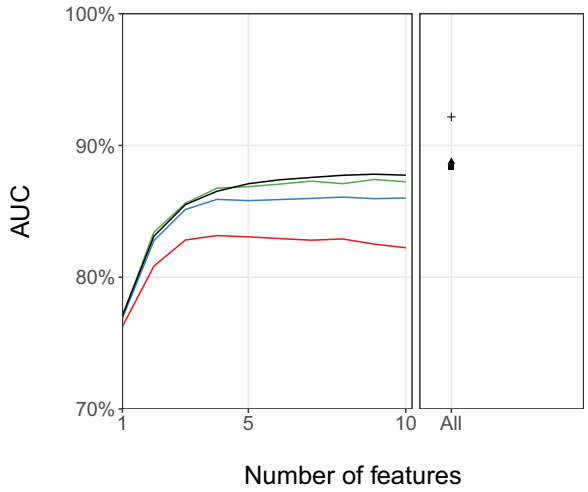

Fig. 1. Performance of simple and complex rules (performance is measured in terms of mean cross-validated AUC over all 21 data sets; the simple models can predict with up to 10 features; the number of 'all' features used by random forests (+), the lasso (▲) and logistic regression (■) varied by domain, with an average of 20): ——, simple models with no rounding; ——, ——, ——, simple models rounding coefficients to $[-3, 3]$, $[-2, 2]$ and $[-1, 1]$ respectively

in their native representation as well as their unordered, discretized versions. In contrast, the simple rules include only the unordered discretized versions. We note that although in theory the out-of-sample performance of logistic regression could be improved by incorporating a variable-selection step, we find no qualitative difference in performance when adding this step in our specific case.

Fig. 1 shows model performance, measured in terms of mean cross-validated area under the receiver operating characteristic curve, AUC, across the 21 data sets, as a function of model size and coefficient range. AUC for each model on each data set is computed via tenfold cross-validation. We find that simple rules with only five features and integer coefficients between -3 and 3 perform on a par with logistic regression and L^1-regularized logistic regression trained on the full set of features. For 1–10 features, the $[-3, 3]$ model (the green curve in Fig. 1) differs from the unrounded lasso model (the black curve) by less than 1 percentage point. The performance of the random-forest model—which is designed to capture non-linear structure—is somewhat better: trained on all features, the random-forest algorithm achieves a mean AUC of 92%; the mean AUC is 87% for simple rules with at most five features and integer coefficients between -3 and 3.

In Appendix B.2, we examine the performance of select, regress and round for each of the 21 UCI data sets separately. As Figs 11 and 12 there demonstrate, across almost all data sets, simple rules have AUC comparable with that of logistic regression (with or without regularization) and have slightly lower AUC than a random-forest model. As these results indicate, complex prediction methods certainly have their advantages, but the gap in performance between simple rules and fully optimized prediction methods is not as large as one might have thought.

3.2. Benchmarking to integer programming

The simple rules that we construct take the form of a linear scoring rule with integer weights. To produce such rules, mixed integer programming is a natural alternative to our select, regress

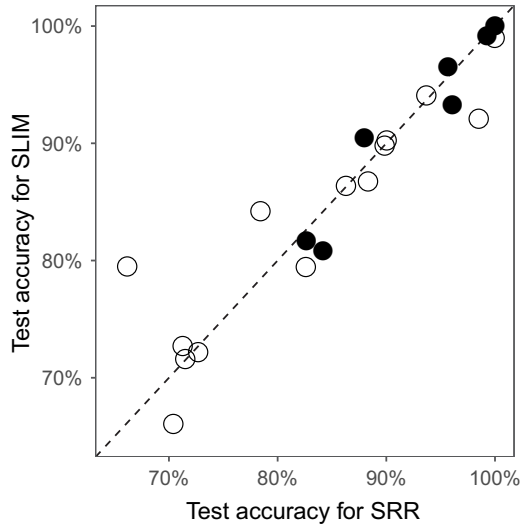

Fig. 2. Comparing binary classification accuracy for select, regress and round and SLIM on 21 UCI data sets: ●, cases in which SLIM successfully found an optimal integer solution; ○, cases in which the time limit of 6 h was exceeded

and round method, and SLIM (Ustun and Rudin, 2016) is the leading instantiation of that approach, with which we now compare. Integer programming is an 'NP-hard' problem, and so following Ustun and Rudin (2016) we set a time limit for SLIM: a 10-min limit was set in Ustun and Rudin (2016), but we allow up to 6 h of computation per model. For seven of the 21 data sets, SLIM found an integer-optimal solution within the time limit, and it returned approximate solutions in the remaining 14 cases.

Fig. 2 compares the binary classification accuracy of SLIM and select, regress and round on the 21 UCI data sets, where each point corresponds to a data set. Both methods are constrained to produce rules with at most five features and integer coefficients between −3 and 3. In comparing with SLIM, here we define the number of features k to be the number of *binarized* variables—for both SLIM and select, regress and round—since this method of accounting is what is used by SLIM. For example, whereas a single categorical variable with five possible values would have been considered as one feature in the previous section, each possible value is counted as a feature here, and hence including the entire categorical variable would result in a model with five features. We show 0–1 accuracy as opposed to AUC, since SLIM produces only optimized binary decisions, for which AUC is not applicable. In computing 0–1 accuracy for select, regress and round, we select a cut point that corresponds to approximately 0.5 on the probability scale. Accuracy is computed out of sample via tenfold cross-validation. Both methods for producing simple rules yield comparable results: averaged across all 21 data sets, SLIM and select, regress and round both achieve a mean accuracy of 86%. Even in the seven cases where SLIM found integer optimal solutions, the performance is nearly identical to that of the simple select, regress and round method.

In terms of classification accuracy, select, regress and round generates rules that are on a par with those obtained by solving mixed integer programs. We note, however, two advantages of our approach. First, whereas select, regress and round yields results almost instantaneously, integer programs can be computationally expensive to solve. Second, our approach is relatively simple, both conceptually and technically, accordingly easing adoption for practitioners.

4. Case-study: pretrial release decisions

To illustrate the value—and challenges—of applying simple decision rules in practice, we now turn to the domain of pretrial release determinations and present an extended case-study. In the USA, defendants are typically *arraigned* shortly after arrest in a court appearance where they are provided with written notice of the charges that are alleged by the prosecutor. At this time, a judge must decide whether a defendant, while awaiting trial, should be *released on one's own recognizance* or, alternatively, subject to monetary bail. In practice, if the judge rules that bail be set, defendants often await trial in jail since many of them do not have the financial resources to post bail. Moreover, when defendants can post bail, they often do so by contracting with a bail bondsman and in turn incur hefty fees. The judge, however, has a legal obligation to consider taking measures that are necessary to secure the defendant's appearance at required court proceedings. Pretrial release decisions must thus balance the risk of flight against the high burden that bail requirements place on defendants. In practice, judges may consider other factors—e.g. a defendant's threat to public safety or ability to afford bail—but risk of flight is the only legally relevant factor for the specific jurisdiction that we analyse below.

A key statistical challenge in this setting is that we cannot directly observe the effects of hypothetical decision rules. Unlike the class of prediction problems that was discussed in Section 3, outcomes in this domain are affected by a judge's decisions, and we observe only the outcomes that result from those decisions. For example, if a proposed policy recommends releasing some defendants who in reality were detained by the judge, we do not observe what would have happened if the rule had been followed. This counterfactual estimation problem—also known as offline policy evaluation (Dudík *et al.*, 2011)—is common in many domains. We address it here by adapting tools from causal inference to the policy setting, including the method of Rosenbaum and Rubin (1983a) for assessing the sensitivity of estimated causal effects to unobserved confounding.

Our analysis is based on 165000 adult cases involving non-violent offences charged by a large urban prosecutor's office and arraigned in criminal court between 2010 and 2015. This set was obtained by starting with a random sample of 200000 cases provided to us by the prosecutor's office, and then restricting to those cases involving non-violent offences and for which the records were complete and accurate. Our initial sample of 200000 cases does not include instances where defendants accepted a plea deal at arraignment, obviating the need for a pretrial release decision. For each case, we have a rich set of attributes: 49 features describe characteristics of the current charges (e.g. theft, gun related), and 15 describe characteristics of the defendant (e.g. gender, age or prior arrests). We also observe whether the defendant was released on recognizance (ROR) and whether the defendant had a failure to appear (FTA) at any of the subsequent court dates. We note that, even if bail is set, a defendant may still fail to appear since one can post bail and then miss a court date. Overall, 69% of defendants are ROR and 15% of ROR defendants fail to appear. Of the remaining 31% of defendants for whom bail is set, 45% are eventually released and 9% fail to appear. As a result, the overall FTA rate is 13%.

In our analysis below, we randomly divide the full set of 165000 cases into three approximately equal subsets; we use the first fold to construct decision rules (both simple and complex), and the second and third to evaluate these rules, as described next.

4.1. Rule construction

We start by constructing complex statistical decision rules for balancing the risk of flight against the burdens of bail. These rules serve as a benchmark for evaluating the simple rules that we create below. On the first fold of the data, we restrict to cases in which the defendant was ROR by

Table 2. A simple rule for estimating flight risk with five features: age, prior FTAs, major charge category, housing instability and defence attorney type†

Feature	*Score*	*Feature*	*Score*
$18 \leqslant \text{age} < 26$	2	1 prior FTA	2
$26 \leqslant \text{age} < 31$	1	2 prior FTAs	3
Major charge group A	−2	3 or more prior FTAs	3
Major charge group B	−1	Unstable housing	3
Major charge group C	1	Defence attorney type A	2
Major charge group D	2	Defence attorney type B	−1
Major charge group E	2	Defence attorney type C	−3

†A defendant's flight risk is obtained by summing the corresponding scores for the features that apply to the case.

the judge and then fit an L^1-regularized logistic (lasso) regression and random forest, using the procedures that were described in Section 3.1, to estimate the likelihood that an individual fails to appear at any of their subsequent court dates. We fit these models on all available information about the case and the defendant, excluding race because of legal and policy concerns with basing decisions on protected attributes (Corbett-Davies *et al.*, 2017; Corbett-Davies and Goel, 2018). We note, however, that including race does not significantly affect performance. The fitted models let us compute risk scores (i.e. estimated flight risk if ROR) for any defendant. These risk scores can in turn be converted into a binary decision rule by selecting a threshold for releasing individuals. For example, a defendant might be ROR if and only if their estimated risk of flight is below 20%.

We construct a family of simple rules by applying select, regress and round as described in Section 2, using all available features. The exact discretization scheme that was used for numerical features—such as age and a defendant's number of prior failures to appear—was determined in consultation with domain experts in the prosecutor's office with which we worked. The resulting rule using five features with integer coefficients between −3 and 3 is presented in Table 2. Unsurprisingly, missing court appearances in the past is a strong indicator of risk of flight and an individual's risk also declines with age, in line with conventional wisdom. The rule in Table 2, however, may be inappropriate for implementation given that some features and their associated scores could be challenged as undesirable. For example, defendants with unstable housing are rated a higher risk, which may be statistically true but which could lead to adverse outcomes for poorer defendants. Particularly in policy domains, feature selection often requires careful thought.

In practice, we recommend that variable selection incorporates domain expertise. For example, starting from a list of predictive features, as in Table 2, one might exclude problematic variables. On the basis of discussions with experts in our partner prosecutor's office, we ultimately used only two features—age and prior history of failing to appear—which are generally viewed as acceptable considerations in pretrial decision making. In this case, we can think of the 'select' step in the select, regress and round strategy as incorporating both human and machine judgement. Specifically, we fit the following model:

$$\Pr(Y_i = 1) = \text{logit}^{-1}(\beta_0 + \beta_1^{\text{priors}} H_i^1 + \beta_2^{\text{priors}} H_i^2 + \beta_3^{\text{priors}} H_i^3 + \beta_{4+}^{\text{priors}} H_i^{4+} + \beta_{18-20}^{\text{age}} A_i^{18-20} + \ldots + \beta_{46-50}^{\text{age}} A_i^{46-50}),$$

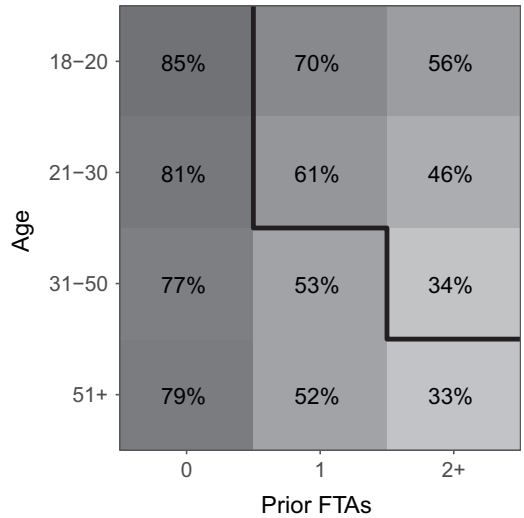

Fig. 3. Graphical representation of a simple rule based on the scores shown in Table 3 with a release threshold of 3.5; groups to the left of the black line in the grid are those that would be released under the rule; for comparison, the shading and numbers in the grid show the proportion of defendants who were actually ROR by judges in each group

Table 3. Simple rule for estimating the risk of flight, where a defendant's risk is obtained by summing the appropriate scores for age and prior history of FTA

Feature	*Score*	*Feature*	*Score*
$18 \leqslant \text{age} < 21$	3	No prior FTAs	0
$21 \leqslant \text{age} < 31$	2	1 prior FTA	2
$31 \leqslant \text{age} < 51$	1	2 or more prior FTAs	3
$51 \leqslant \text{age}$	0		

where $Y_i \in \{0, 1\}$ indicates whether the ith defendant failed to appear, $H_i^* \in \{0, 1\}$ indicates the defendant's number of prior failures to appear (exactly 1, 2, 3 or at least 4) and $A_i^* \in \{0, 1\}$ indicates the binned age of the defendant (18–20, 21–25, 26–30, 31–35, 36–40, 41–45 or 46–50 years). The parameters β_*^{priors} and β_*^{age} are the coefficients corresponding to each binary indicator variable. For identifiability, indicator variables for no prior FTAs and age 51 years and older have been omitted. As before, this model is fitted on the subset of cases in the first fold of data for which the judge released the defendant. Next, we rescale the age and prior FTA coefficients so that they lie in the interval $[-3, 3]$; specifically, we multiply each coefficient by the constant

$$\frac{3}{\max(|\beta_1^{\text{prior}}|, \ldots, |\beta_{4+}^{\text{priors}}|, |\beta_{18-20}^{\text{age}}|, \ldots, |\beta_{46-50}^{\text{age}}|)}.$$

Finally, we round the rescaled coefficients to the nearest integer.

Fig. 3 shows the result of this procedure. For any defendant, a risk score can be computed by summing the relevant terms in Table 3. These risk scores can be converted into a binary decision

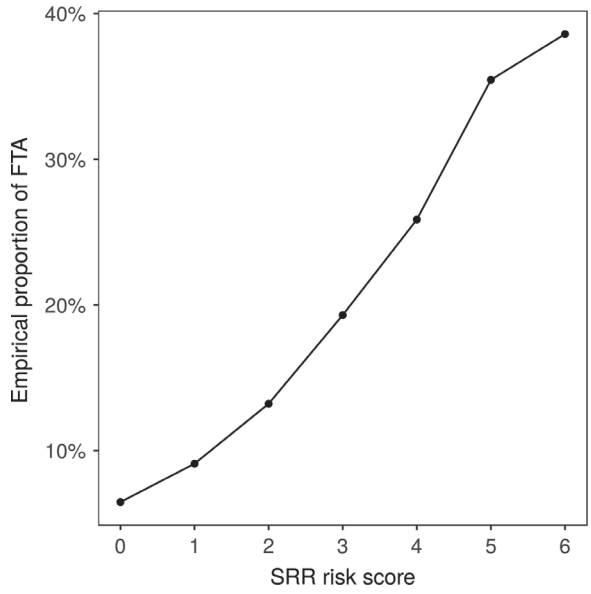

Fig. 4. Empirical frequency of FTA for each risk score, based on the simple rule shown in Fig. 3

rule by selecting a threshold for releasing individuals. For example, a defendant might be ROR if and only if their risk score is below 3.5; a graphical representation of such a binary decision rule is also shown in Fig. 3.

The application of a simple rule derived from the select, regress and round procedure yields an integer score for each defendant. However, in practice it may be useful also to have a probabilistic estimate of each defendant's risk (i.e. the probability that a defendant will fail to appear if released). A given integer score can be converted into a probabilistic risk estimate by considering all released defendants in the training set with that score, and then computing the empirical frequency that those defendants failed to appear. Fig. 4 shows the empirical frequency of FTA for each risk score based on the simple rule that is shown in Fig. 3. For example, a score of 3—the threshold value that was chosen in Fig. 3—corresponds to a risk estimate of 20%. These probability estimates characterize the risk among individuals who were in reality released. It is important, however, to note that those who were released may be qualitatively different from those who were not, and so these estimates provide only approximate risk in the full population of defendants, which is an issue that we consider in more detail in the following sections. We specifically examine the robustness of these probability estimates in Appendix B.1 and find that they are comparable with estimates from more complex prediction models.

4.2. Policy evaluation

AUC is a useful general measure of performance, and hence the metric that we consider when evaluating the 21 UCI data sets in Section 3. But in applied settings it is often necessary to measure the costs and benefits of any given rule directly. We do that here by assessing decision rules for pretrial release on two key dimensions:

(a) the proportion of defendants who are released under the rule and
(b) the resulting proportion who fail to appear at their court proceedings.

It is straightforward to estimate the former, since we need only to apply the rule to historical data

to see what actions would have been recommended. For example, if defendants are released if and only if their risk score is below 3.5, 79% would be ROR; under this rule, bail would be required for only two-thirds as many defendants relative to the *status quo*. Forecasting the proportion who would fail to appear, however, is generally much more difficult. The key problem is that, for any particular defendant, we observe only the outcome (i.e. whether or not the defendant failed to appear) conditionally on the action that the judge ultimately decided to take (i.e. release on recognizance or bail). Since the action that is taken by the judge may differ from that prescribed by the decision rule, we do not always observe what would have happened under the rule. This problem of *offline policy evaluation* (Dudík *et al.*, 2011) is a specific instance of the fundamental problem of causal inference.

To describe the estimation problem and our approach rigorously, we introduce some notation. For concreteness, we frame our methodology in terms of the pretrial release example, but the ideas that are presented here are common to many policy decisions. We denote the observed set of cases by $\Omega = \{(x_i, a_i, r_i)\}$, where x_i corresponds to the features of a case, $a_i \in \{\text{ROR}, \text{bail}\}$ is the action that was taken by the judge, and $r_i \in \{0, 1\}$ indicates whether the defendant failed to appear at a scheduled court date. We write $r_i(\text{ROR})$ and $r_i(\text{bail})$ to mean the *potential outcomes*: what would have happened under the two possible judicial actions. For any policy π, our goal is to estimate the FTA rate under the policy:

$$V^{\pi} = \frac{1}{|\Omega|} \sum_i r_i\{\pi(x_i)\},$$

where $\pi(x)$ denotes the action that is prescribed under the rule. The key statistical challenge is that only one of the two potential outcomes, $r_i = r_i(a_i)$, is observed. Policy evaluation is a generalization of estimating average treatment effects, namely, the average treatment effect can be expressed as $V^{\pi_{\text{ROR}}} - V^{\pi_{\text{bail}}}$, where π_{ROR} is the policy under which everyone is released and π_{bail} is defined analogously.

We investigated three approaches to estimating V^{π}—response surface modelling (Hill, 2012), inverse propensity weighting (Rosenbaum and Rubin, 1983b, 1984) and doubly robust estimation (Cassel *et al.*, 1976; Robins *et al.*, 1994; Robins and Rotnitzky, 1995; Kang and Schafer, 2007; Dudík *et al.*, 2011)—and found qualitatively similar results. Here we present the response surface modelling approach for its relative simplicity. With response surface modelling, the idea is to use a standard prediction model (e.g. logistic regression or random forests) to estimate the effect on each defendant of each potential judicial action. The model estimates of these potential outcomes are denoted by $\hat{r}_i(t)$, for $t \in \{\text{ROR}, \text{bail}\}$. Our estimate of V^{π} is then given by

$$\hat{V}^{\pi} = \frac{1}{|\Omega|} \sum_i [r_i I\{\pi(x_i) = a_i\} + \hat{r}_i\{\pi(x_i)\} I\{\pi(x_i) \neq a_i\}],$$

where $I(\cdot)$ is an indicator function evaluating to 1 if its argument is true and to 0 otherwise. If the prescribed action is in fact taken by the judge, then $r_i = r_i\{\pi(x_i)\}$ is directly observed and can be used; otherwise we approximate the potential outcome with $\hat{r}_i\{\pi(x_i)\}$. Table 4 illustrates this method for a hypothetical example.

Response surface modelling implicitly assumes that a judge's action is *ignorable* given the observed covariates (i.e. that, conditionally on the observed covariates, those who are ROR are similar to those who are not). Formally, ignorability means that

$$(r(\text{ROR}), r(\text{bail})) \perp\!\!\!\perp a | x.$$

This ignorability assumption is typically unavoidable, and it is similarly required for methods based on propensity scores (Rosenbaum and Rubin, 1983b, 1984; Cassel *et al.*, 1976; Robins

Table 4. Illustrative example of response surface modelling for offline policy evaluation†

Proposed action π	*Observed action* a	*Observed outcome* $r(a)$	$\hat{r}(ROR)$ *(%)*	$\hat{r}(bail)$ *(%)*
ROR	ROR	*0*	20	10
Bail	Bail	*1*	80	30
Bail	ROR	1	*90*	70
ROR	Bail	0	30	*25*
ROR	ROR	*0*	20	15

†For each defendant, $\hat{r}$(ROR) and $\hat{r}$(bail) are model-based estimates of the likelihood of FTA under each potential action. In cases where the observed action equals the proposed action, the observed outcome (FTA or not) is used to estimate the policy's effect; otherwise, the model-based estimates are used. Values in italics indicate which values are used in each instance. The overall FTA rate under the policy is estimated by averaging the italic values over all cases.

et al., 1994; Robins and Rotnitzky, 1995; Kang and Schafer, 2007; Dudík *et al.*, 2011). We examine this assumption in detail in Section 4.3 and find that our conclusions are robust under a common model of unobserved heterogeneity.

To carry out this approach, we derive estimates $\hat{r}_i(t)$ via an L^1-regularized logistic regression (lasso) model trained on the second fold of our data. For each individual, the model estimates the likelihood of FTA given all the observed features and the action that is taken by the judge. In contrast with the rule construction that was described above, this time we train the model on all cases (not just those for which the defendant ROR was by the judge) and include as a predictor the judge's action (ROR or bail); we also include the defendant's race. Although it is legally problematic to use race when *making* decisions, its use is acceptable—and indeed often required—when *evaluating* decisions. Then, on the third fold of the data, we use the observed and model-estimated outcomes to approximate the overall FTA rate for any decision rule. The model was fitted with the `glmnet` package in R. The `cv.glmnet` method was used to determine the best value for the regularization parameter λ with tenfold cross-validation and 1000 values of λ. The model includes all pairwise interactions between the judge's decision and defendant's features. We opt for the lasso instead of random forests for this prediction task because we empirically found that the lasso yielded better predictions in this case.

Fig. 5 shows estimated ROR and FTA rates for a variety of pretrial release rules. Points on the curves correspond to rules that were constructed via the lasso (the black curve) and random-forest (the grey curve) models that use all 64 available features, as described above, for various decision thresholds. The red points correspond to rules based on the simple scoring procedure in Fig. 3, using just age and prior FTA, again corresponding to various decision thresholds. For each rule, the horizontal axis shows the estimated proportion of defendants who were ROR under the rule, and the vertical axis shows the estimated proportion of defendants who would fail to appear at their court dates. The full black dot shows the *status quo*: 69% of defendants ROR and a 13% FTA rate. Finally, the open circles show the observed ROR and FTA rates for each of the 23 judges in our data who have presided over at least 1000 cases, sized in proportion to their case load.

Fig. 5. illustrates three key points. First, simple rules that consider only two features—age and prior FTAs—perform nearly identically to state of the art machine learning models (random

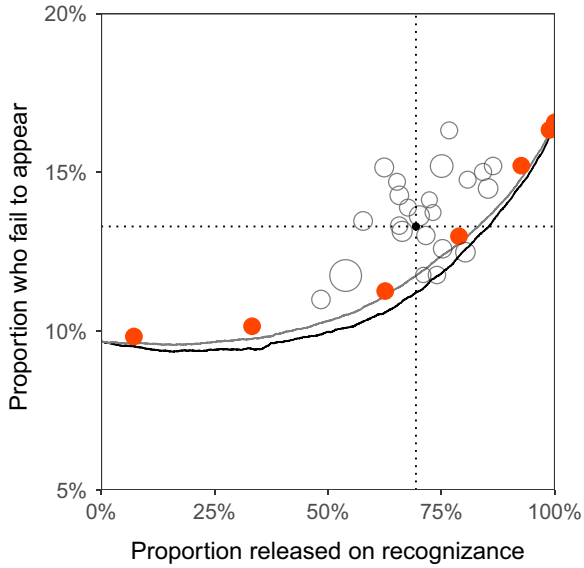

Fig. 5. Evaluation of simple and complex decision rules (each point on the curves corresponds to decision rules derived from a random forest (——) or lasso (——) risk model using all 64 features with varying thresholds for release): ●, policies based on the simple risk score using just two features for all possible release thresholds (the simple rules perform nearly identically to the random-forest models, and comparably with the lasso models) ○, observed ROR and FTA rates for each judge in our data who presided over at least 1000 cases, sized in proportion to their case load (in nearly every instance, the statistical decision rules outperform the human decision maker)

forests and lasso regression) that incorporate all 64 available features. Second, the statistically informed policies in the lower right quadrant all achieve higher rates of ROR and, simultaneously, lower rates of FTA than the *status quo*. In particular, by releasing defendants if and only if their risk score is below 3.5, we expect to release 79% of defendants while achieving an FTA rate of 13%. Relative to the existing policy, following this rule would result in detaining a third fewer defendants while also slightly decreasing the overall FTA rate—from 13.3% to 13.0%. Finally, for nearly every judge, there is a statistical decision rule that simultaneously yields both a higher rate of release and a lower rate of FTA than the judge currently achieves. The statistical decision rules consistently outperform the human decision makers.

Why do these statistical decision rules outperform the experts? Fig. 3 sheds light on this. Each cell in the grid corresponds to defendants binned by their age and prior number of FTAs. Under a rule that releases defendants if and only if their risk score is below 3.5, one would release everyone to the left of the black line, and set bail for everyone to the right of the line. The number in each cell shows the proportion of defendants in each bin who were actually released, and the cell shading graphically indicates this proportion. Aside from the lowest risk defendants, who have no prior FTAs, the likelihood of being released does not correlate strongly with the estimated risk of flight. For example, the high risk group of young defendants with two or more prior FTAs is released at about the same rate as the low risk group of older defendants with one prior FTA. This low correlation between risk of flight and release decision is in part attributable to extreme differences in release rates across judges, with some releasing more than 90% of defendants and others releasing just 50%. Whereas defendants experience dramatically different outcomes based on the judge whom they happened to appear in front of, statistical decision rules improve efficiency in part by ensuring consistency.

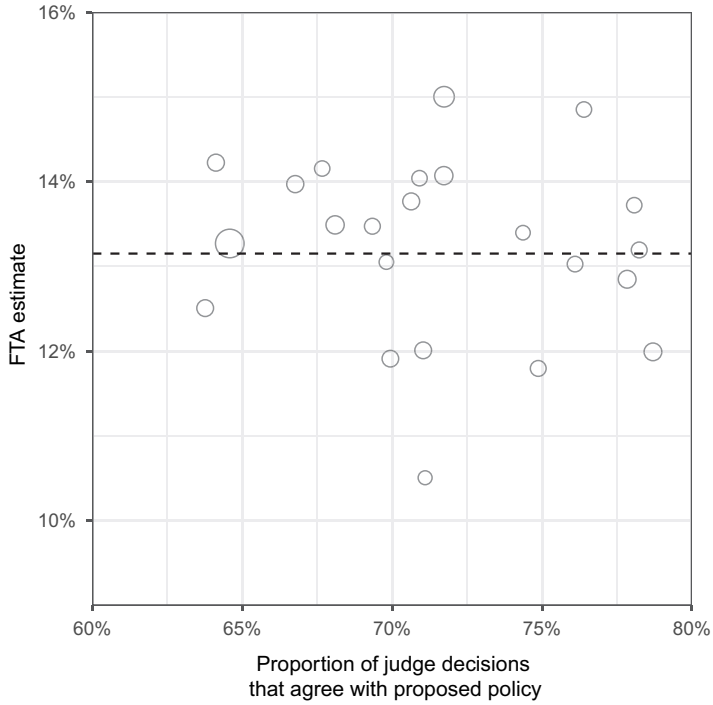

Fig. 6. Robustness of estimated FTA rate for the simple decision rule (the FTA rate is estimated by applying response surface modelling to each judge's cases, where each point corresponds to a judge; though judges have different criteria for releasing defendants—and the corresponding response models may thus differ—the FTA rate of the decision rule is consistently estimated to be approximately 12–14%): – – –, FTA rate of the decision rule estimated on the full set of cases

4.3. Sensitivity to unobserved heterogeneity

As noted above, our estimation strategy assumes that the judicial action that is taken is ignorable given the observed covariates. Under this ignorability assumption, we can accurately estimate the potential outcomes. Judges, however, might base their decisions in part on information that is not recorded in the data, which could in turn bias our estimates. For example, a judge, on meeting a defendant, might surmise that their risk of flight is higher than we would expect based on the recorded covariates alone and may accordingly require the defendant to post bail. In this case, since our estimates are based only on the recorded data, we may underestimate the defendant's counterfactual likelihood of failing to appear if released.

We take two approaches to gauge the robustness of our results to such hidden heterogeneity. First, on each subset of cases handled by a single judge, we use response surface modelling to estimate V^{π}. Each judge has idiosyncratic criteria for releasing defendants, as evidenced by the dramatically different release rates across judges; accordingly, the types and proportion of cases for which the policy π coincides with the observed action differ from judge to judge. This variation enables us to assess the sensitivity of our estimates to the observed actions $\{a_i\}$. In particular, if unobserved heterogeneity were significant, we would expect our estimates to vary systematically depending on the proportion of observed judicial actions that agree with the policy π. Fig. 6 shows the results of this analysis for the simple decision rule that is described in Fig. 3, where each point corresponds to a judge. We find that the FTA rate of the decision rule is consistently estimated to be approximately 12–14%. Moreover, some judges act in concordance with the de-

cision rule in nearly 80% of cases; for this subset of judges, where our estimates are largely based on directly observed outcomes, we again find that the FTA rate is estimated at around 12–14%.

As a second robustness check, we adapt the method of Rosenbaum and Rubin (1983a) for assessing the sensitivity of estimated causal effects to an unobserved binary covariate. We specifically tailor their approach to offline policy evaluation. At a high level, we assume that there is an unobserved covariate $u \in \{0, 1\}$ that affects both a judge's decision (ROR or bail) and also the outcome conditional on that action. For example, u might indicate that a defendant is sympathetic, and sympathetic defendants may be more likely to be ROR and also more likely to appear at their court proceedings. Our key assumption is that a judge's action is ignorable given the observed covariates x and the unobserved covariate u:

$$(r(\text{ROR}), r(\text{bail})) \perp\!\!\!\perp a | x, u. \tag{1}$$

There are four key parameters in this framework:

(a) the probability that $u = 1$;
(b) the effect of u on the judge's decision;
(c) the effect of u on the defendant's likelihood of FTA if ROR;
(d) the effect of u on the defendant's likelihood of FTA if bail is set.

Our goal is to quantify the extent to which our estimate of V^π changes as a function of these parameters.

Without loss of generality, we can write

$$\Pr(a = \text{ROR} | u, x) = \text{logit}^{-1}(\gamma_x + u\alpha_x) \tag{2}$$

for appropriately chosen parameters γ_x and α_x that depend on the observed covariates x. We note that randomness in judicial decisions may arise from a multitude of factors, including idiosyncrasies in how judges are assigned to cases. Here α_x is the change in log-odds of being ROR when $u = 0$ *versus* when $u = 1$. For $t \in \{\text{ROR}, \text{bail}\}$, we can similarly write

$$\Pr\{r(t) | u, x\} = \text{logit}^{-1}(\beta_x^t + u\delta_x^t) \tag{3}$$

for parameters β_x^t and δ_x^t. In this case, δ_x^{ROR} is the change in log-odds of failing to appear if ROR when $u = 0$ *versus* when $u = 1$, and δ_x^{bail} is the corresponding change if bail is set.

Now, for any posited values of $\Pr(u = 1|x)$, α_x, δ_x^{ROR} and δ_x^{bail}, we use the observed data to estimate γ_x, β_x^{ROR} and β_x^{bail}. We do this in three steps. First, by equation (2),

$$\Pr(a = \text{ROR} | x) = \Pr(u = 0|x)\ \text{logit}^{-1}(\gamma_x) + \Pr(u = 1|x)\ \text{logit}^{-1}(\gamma_x + \alpha_x).$$

The left-hand side of this equation can be estimated with a regression model fitted to the data. For fixed values of $\Pr(u = 1|x)$ and α_x, the right-hand side is a continuous increasing function of γ_x that takes values from 0 to 1 as γ_x goes from $-\infty$ to ∞. There is thus a unique value $\hat{\gamma}_x$ such that the right-hand side equals $\widehat{\Pr}(a = \text{ROR}|x)$. Rosenbaum and Rubin (1983a) derived a simple closed form solution for $\hat{\gamma}_x$, facilitating fast computation on large data sets, which we omit for brevity.

Second, we use the fitted values of γ_x to estimate the distribution of u given the observed covariates and judicial action. By Bayes's rule,

$$\begin{aligned}\Pr(u = 1 | a = t, x) &= \frac{\Pr(a = t | u = 1, x)\Pr(u = 1|x)}{\Pr(a = t|x)} \\ &= \frac{\Pr(a = t | u = 1, x)\Pr(u = 1|x)}{\Pr(a = t|u = 1, x)\Pr(u = 1|x) + \Pr(a = t|u = 0, x)\Pr(u = 0|x)}.\end{aligned}$$

With $\hat{\gamma}_x$, the $\Pr(a=t|u,x)$ terms on the right-hand side can be estimated from equation (2), and we can thus approximate the left-hand side.

Third, we have

$$\begin{aligned}\Pr\{r(t)=1|a=t,x\} &= \Pr(u=0|a=t,x)\Pr\{r(t)=1|a=t,x,u=0\} + \Pr(u=1|a=t,x)\\ &\quad\times \Pr\{r(t)=1|a=t,x,u=1\}\\ &= \Pr(u=0|a=t,x)\Pr\{r(t)=1|x,u=0\} + \Pr(u=1|a=t,x)\\ &\quad\times \Pr\{r(t)=1|x,u=1\}\\ &= \Pr(u=0|a=t,x)\,\text{logit}^{-1}(\beta_x^t) + \Pr(u=1|a=t,x)\,\text{logit}^{-1}(\beta_x^t+\delta_x^t).\end{aligned}$$

The second equality above follows from the ignorability assumption that is stated in equation (1), and the third equality follows from equation (3). The left-hand side can be approximated by the quantity $\hat{r}_x(t)$ that we obtain via response surface modelling. Importantly, $\hat{r}_x(t)$ is a reasonable estimate of $\Pr\{r(t)=1|a=t,x\}$ even though it may not be a good estimate of $r_x(t)$. This distinction is indeed the rationale of our sensitivity analysis. Given our above estimate of $\Pr(u=1|a=t,x)$ and our assumed value of δ_x^t, the only unknown on the right-hand side is β_x^t. As before, there is a unique value $\hat{\beta}_x^t$ that satisfies the constraint.

With $\hat{\beta}_x^t$ in hand, we can now approximate the potential outcome for the action that is *not* taken:

$$\Pr\{r(\bar{t})=1|a=t,x\}$$

where $\bar{t}\equiv$ ROR if $t\equiv$ bail, and vice versa. Specifically, we have

$$\widehat{\Pr}\{r(\bar{t})=1|a=t,x\} = \widehat{\Pr}(u=0|a=t,x)\,\text{logit}^{-1}(\hat{\beta}_x^{\bar{t}}) + \widehat{\Pr}(u=1|a=t,x)\,\text{logit}^{-1}(\hat{\beta}_x^{\bar{t}}+\delta_x^{\bar{t}}). \qquad (4)$$

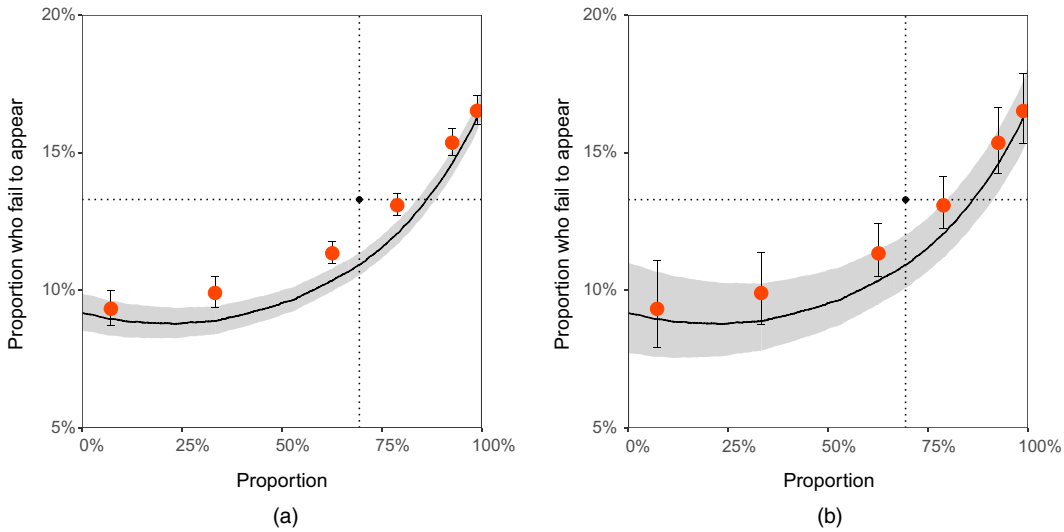

Fig. 7. Sensitivity of FTA estimates to unobserved heterogeneity (the grey bands (for the complex rules using the lasso) and the error bars (for the simple rules) indicate minimum and maximum FTA estimates for a variety of parameter settings): in (a) we assume that $\alpha = \log(2)$ and consider all combinations of $p(u=1) \in \{0.1, 0.2, \ldots, 0.9\}$, $\delta^{\text{ROR}} \in \{-\log(2), 0, \log(2)\}$, and $\delta^{\text{bail}} \in \{-\log(2), 0, \log(2)\}$, where all parameters are constant independent of x; in (b) we consider a more extreme situation, with $\alpha = \log(3)$, $\delta^{\text{ROR}} \in \{-\log(3), 0, \log(3)\}$ and $\delta^{\text{bail}} \in \{-\log(3), 0, \log(3)\}$; the results are relatively stable in these parameter regimes

Finally, the Rosenbaum and Rubin estimator adapted to policy evaluation is

$$\hat{V}^{\pi}_{\mathrm{RR}} = \frac{1}{|\Omega|} \sum_i [r_i \mathbf{I}\{\pi(x_i) = a_i\} + \hat{r}_i(\bar{a}_i)\mathbf{I}\{\pi(x_i) \neq a_i\}],$$

where $\hat{r}_i(\bar{a}_i) = \widehat{\Pr}\{r(\bar{a}_i) = 1 | a_i, x_i\}$ is computed via equation (4).

Fig. 7 shows the results of computing $\hat{V}^{\pi}_{\mathrm{RR}}$ on our data in two parameter regimes. In the first (Fig. 7(a)), we assume that $\alpha = \log(2)$ and consider all combinations of $p(u = 1) \in \{0.1, 0.2, \ldots, 0.9\}$, $\delta^{\mathrm{ROR}} \in \{-\log(2), 0, \log(2)\}$ and $\delta^{\mathrm{bail}} \in \{-\log(2), 0, \log(2)\}$. All parameters are constant independent of x. We thus assume that, holding the observed covariates fixed, a defendant with $u = 1$ has twice the odds of being ROR as a defendant with $u = 0$, and that u can double or halve the odds that a defendant will fail to appear. For each complex policy (i.e. one based on the lasso), the grey bands show the minimum and maximum value of $\hat{V}^{\pi}_{\mathrm{RR}}$ across all parameters in this set; the error bars on the red points show the analogous quantity for the simple rules. In Fig. 7(b) we consider a more extreme situation, with $\alpha = \log(3)$, $\delta^{\mathrm{ROR}} \in \{-\log(3), 0, \log(3)\}$ and $\delta^{\mathrm{bail}} \in \{-\log(3), 0, \log(3)\}$. We find that our estimates are relatively stable in these parameter regimes. In the first case ($\alpha = \log(2)$) the estimated FTA rate for a given policy typically varies by only half a percentage point. Even in the more extreme setting ($\alpha = \log(3)$), policies are typically stable to about 1 percentage point. It thus seems that our conclusions are robust to potentially unobserved heterogeneity across defendants.

5. The robustness of binary classification

Why is it that simple rules often perform as well as the most sophisticated statistical methods? In part, it is because binary classification accuracy is relatively robust to error in the underlying predictive model: an observation that we formalize in proposition 1 below.

To establish this result, we start by considering the prediction scores that are generated via a standard statistical method—such as logistic regression trained on the full set of available features—which we call the 'true' scores. As in linear discriminant analysis, we assume that the true scores for positive and negative instances are normally distributed with equal variance: $N(\mu_p, \sigma^2)$ and $N(\mu_n, \sigma^2)$ respectively. The homoscedasticity assumption guarantees that the Bayes optimal classifier is a threshold rule on the scores. For scores that are estimated via logistic regression, the normality assumption is reasonable if we consider the scores on the logit scale rather than on the probability scale. Fig. 8(a) shows such scores for one of the UCI data sets: heart-cleveland. We further assume that the process of generating simple rules—both limiting the number of features and also restricting the possible values of the weights—can be viewed as adding normal, mean 0 noise $N(0, \sigma^2_\epsilon)$ to the true scores; Fig. 8(b) plots the distribution of this noise for the same heart-cleveland data set as considered in Fig. 8(a). We estimate the noise distribution by taking the difference between the simple and true scores. Thus, with simple rules, instead of making classification decisions based on the true scores, we assume that decisions are made in terms of a noisy approximation. Under this analytic framework, proposition 1 shows that the drop in classification performance (as measured by AUC) can be expressed in terms of the 'true AUC' (i.e. the AUC under the true scores) and $\gamma = \sigma^2_\epsilon / \sigma^2$: the ratio of the noise to the within-class variance of the true scores. In particular, we find that when the magnitude of the noise is on a par with (or smaller than) the score variance (i.e. $\gamma \lesssim 1$), then the AUC of the noisy approximation is comparable with the true AUC.

Proposition 1. For a binary classification task, let Y be a continuous random variable that denotes the prediction score of a random instance, and let Y_p and Y_n denote the conditional

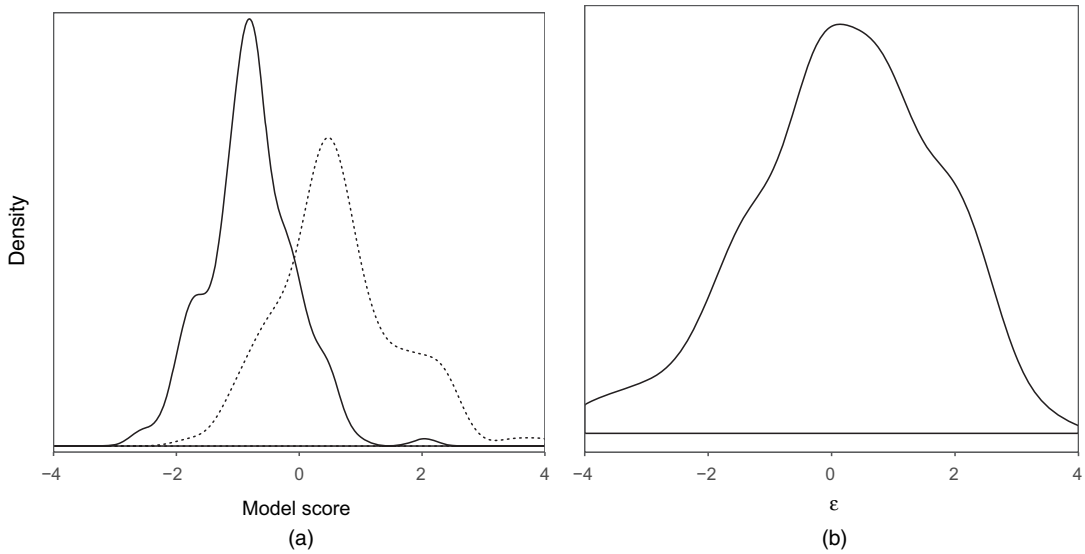

Fig. 8. Empirical estimation of noise added by simple rules: (a) empirical distribution of prediction scores, on the logit scale, for positive (· · · · · ·) and negative (——) instances of one UCI data set (heart-cleveland), generated via an L^1-regularized logistic regression model; (b) empirical distribution of ϵ for select, regress and round applied to the same data set

distributions of Y for positive and negative instances respectively. Suppose that $Y_p \sim N(\mu_p, \sigma^2)$ and $Y_n \sim N(\mu_n, \sigma^2)$. Then, for $\epsilon \sim N(0, \sigma_\epsilon^2)$ and $\hat{Y} = Y + \epsilon$,

$$\mathrm{AUC}_{\hat{Y}} = \Phi\left\{ \frac{\Phi^{-1}(\mathrm{AUC}_Y)}{\sqrt{(1+\gamma)}} \right\}, \tag{5}$$

where $\gamma = \sigma_\epsilon^2/\sigma^2$, and Φ is the cumulative distribution function for the standard normal distribution.

Proof. In general, AUC is equal to the probability that a randomly selected positive instance has a higher prediction score than a randomly selected negative instance, and so $\mathrm{AUC}_Y = \Pr(Y_p - Y_n > 0)$ (Su and Liu, 1993). Since $Y_p - Y_n$ is normally distributed with mean $\mu_p - \mu_n$ and variance $2\sigma^2$,

$$\frac{Y_p - Y_n - (\mu_p - \mu_n)}{\sqrt{2}\sigma} \sim N(0, 1).$$

Hence,

$$\begin{aligned} \mathrm{AUC}_Y &= \Pr\left\{ \frac{Y_p - Y_n - (\mu_p - \mu_n)}{\sqrt{2}\sigma} > -\frac{\mu_p - \mu_n}{\sqrt{2}\sigma} \right\} \\ &= \Phi\left(\frac{\mu_p - \mu_n}{\sqrt{2}\sigma} \right), \end{aligned}$$

where the last equality follows from symmetry of the normal distribution.

Now define $\hat{Y}_p = Y_p + \epsilon$, so $\hat{Y}_p \sim N(\mu_p, \sigma^2 + \sigma_\epsilon^2)$, with $\hat{Y}_n$ defined similarly. A short computation shows that

$$\mathrm{AUC}_{\hat{Y}} = \Pr(\hat{Y}_p > \hat{Y}_n)$$

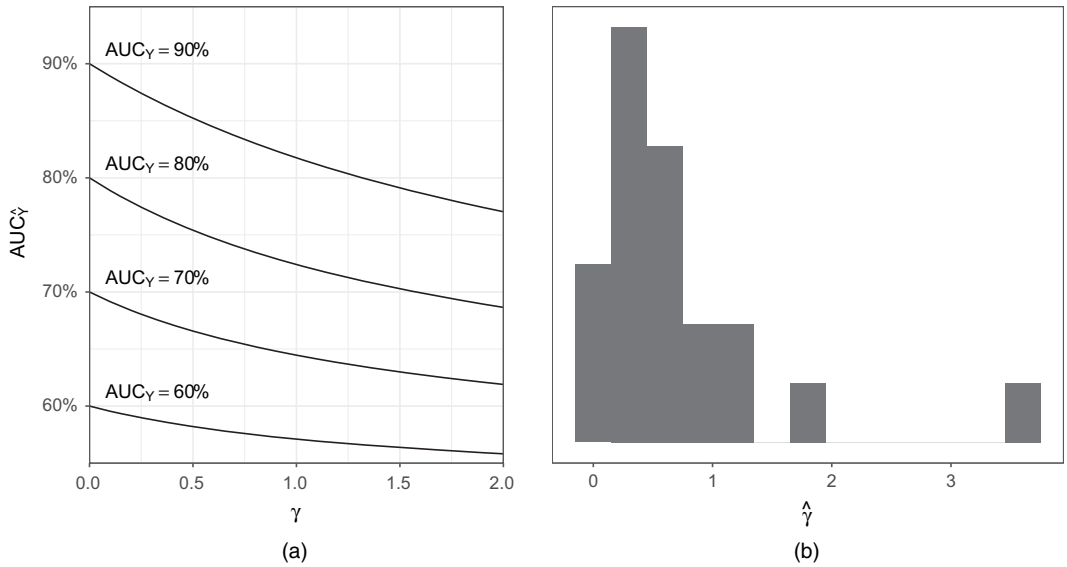

Fig. 9. Theoretical analysis of simple rules: (a) theoretical change in AUC, as a function of γ; (b) distribution of $\hat{\gamma}$, estimated across all simple rules for 21 data sets with $k = 5$ and $M = 3$

$$= \Phi\left\{\frac{\mu_p - \mu_n}{\sqrt{(2\sigma^2 + 2\sigma_\epsilon^2)}}\right\}$$
$$= \Phi\left\{\frac{\Phi^{-1}(\mathrm{AUC}_Y)}{\sqrt{(1+\gamma)}}\right\}. \qquad \square$$

Proposition 1 establishes a direct theoretical link between performance and noise in model specification. To give a better sense of how the analytic expression for $\mathrm{AUC}_{\hat{Y}}$ varies with AUC_Y and γ, Fig. 9(a) shows this expression for various parameter values. For example, Fig. 9(a) shows that, for $\mathrm{AUC}_Y = 90\%$ and $\gamma = 0.5$, we have $\mathrm{AUC}_{\hat{Y}} = 85\%$, i.e. if the amount of noise is equal to half the within-class variance of the true scores, then the drop in performance is relatively small.

Although connecting model performance to model noise, proposition 1 leaves unanswered how much noise simple rules add to the underlying scores. This question seems difficult to answer theoretically. We can, however, empirically estimate how much noise simple rules add in the data sets that we analyse. To estimate $\gamma = \sigma_\epsilon^2/\sigma^2$ for a specific simple rule on a given data set, we first compute the average within-class variance of the true scores, where these scores are generated via an L^1-regularized logistic regression model. We estimate σ_ϵ^2 by taking the variance of the measured noise. Fig. 9(b) shows the distribution of $\hat{\gamma}$ across the 21 UCI data sets that we consider, when using rules with five features and a coefficient range from -3 to 3, with an average value of $\hat{\gamma} = 0.22$. This low empirically observed noise is in line with our finding that such simple rules perform well on these data sets. In Appendix B.1, we further test the empirical robustness of probabilistic risk predictions, in addition to binary classification, using simple rules. We find that probabilistic estimates from our simple rules are comparable with those from more complex statistical models.

6. Conclusion

Our work extends past research on improper linear models by formalizing and evaluating a

simple method for constructing simple rules—rules that experts can apply mentally to guide classification decisions. These simple rules take the form of a short checklist whose factors have small integer weights. In 22 domains of varying size and complexity, the rules that are produced by the select, regress and round method rivalled the accuracy of regularized logistic regression models, although using only a fraction of the information. In a detailed analysis of pretrial release decisions, the simple rules outperformed human judges and matched machine learning models that incorporated 64 features. (In Appendix C, we provide another detailed demonstration of select, regress and round to assess credit risk and we reach similar conclusions.)

Although our focus has been on the comparison between simple, statistically informed decision rules and more complex machine learning methods, our results are also in accordance with an extensive literature comparing predictions by human experts with those based on statistical models. Over 60 years ago, Meehl contrasted *clinical* methods for predicting behaviour, which rely on professional judgement, with *actuarial* methods, which rely on statistically derived patterns in data (Meehl, 1954). Subsequently, large meta-analyses have consistently demonstrated that actuarial methods outperform clinical approaches, including in the context of predicting criminal activity (Ægisdóttir *et al.*, 2006), and even for judgements by the most experienced professionals (see Goel *et al.* (2020) for a review). These results hold in the judicial context as well, where clinical assessments of risk by judges are generally worse at predicting recidivism than actuarial formulae (Gottfredson, 1999). Our analysis of a large data set of judicial decisions provides further evidence that simple statistical models can outperform experts in a high stakes domain.

Statistically informed rules, and simple checklists in particular, may result in improved accuracy and consistency compared with unaided human decisions, but some open questions remain. First, in many contexts, allowing human overrides of algorithmic decision aids may be legally mandated, but such overrides can reduce accuracy (Krauss, 2004). In the criminal justice setting, past work indeed suggests that judges may not apply the recommendations of risk assessments in a consistent manner (Christin, 2017; DeMichele *et al.*, 2018; Stevenson, 2018). It is important to strike an appropriate balance, allowing for human overrides in exceptional instances while not degrading overall performance. Second, in contexts where an outcome variable has a non-linear relationship with a set of predictors, the simple rules that are produced by select, regress and round may not be sufficiently flexible to make useful predictions (for example, in Fig. 12 in Appendix B.2 all linear models show poor performance on the chess-krvk data set). One solution may be to allow additional model flexibility in select, regress and round, though that approach could be at odds with the goals of transparency and interpretability. Finally, it is unclear how well simple rules would work in domains with little training data, but we also note that prediction tasks using small sample sizes remain challenging for more complex methods.

Our results complement a growing body of work in statistics and computer science on interpretable machine learning, in which sophisticated algorithms are used to create simple scoring systems and rule sets (Ustun and Rudin, 2016; Wang and Rudin, 2015; Lakkaraju *et al.*, 2016). Although many of these rule construction methods offer great flexibility, they in turn require considerable computational resources and expertise to carry out. In contrast, the method that we propose can easily be carried out by ordinary practitioners using popular open-source software. It has long been noted that statistical models tend to outperform unaided human judgement (Einhorn and Hogarth, 1975; Green, 1977; Dawes, 1979; Gigerenzer and Goldstein, 1996; Waller and Jones, 2011). We hope that providing practitioners with models that are both easy to apply and easy to construct will increase their adoption and, ultimately, the quality of decisions.

Acknowledgements

We thank Avi Feller, Andrew Gelman, Gerd Gigerenzer, Art Owen and Berk Ustun for helpful conversations. Code to replicate our results is available on line from `https://github.com/stanford-policylab/simple-rules`.

Appendix A: Description of University of California, Irvine, data sets

In Table 5 we provide a short description of the classification task that is associated with each of the 21 UCI data sets that we consider in Section 3.

Table 5. Description of each UCI data set

Data set	*Classification task*
adult	Predict whether income exceeds $50000 per year on the basis of census data: also known as the 'census income' data set
annealing	Classify steel types on the basis of various annealing properties
audiology-std	Standardized version of the original audiology database first presented in Bareiss *et al.* (1988)
bank	The data are related to direct marketing campaigns (phone calls) of a Portuguese banking institution: the classification goal is to predict whether the client will purchase a term deposit
bankruptcy	Predict bankruptcy on the basis of qualitative parameters measured by experts
car	Determine whether a car is 'acceptable' or not, on the basis of quantitative attributes: originally presented in Bohanec and Rajkovic (1988)
chess-krvk	Chess end game data for white king and rook against black king: the classification task is to determine whether white can win or not
chess-krvkp	Chess end game data for king and rook *versus* king and pawn on square A7 (usually abbreviated KRKPA7): the pawn on square A7 means that it is one square away from queening; it is the king and rook's turn (white) to move; the goal is to classify whether white can win or not
congress-voting	1984 US congressional voting records: the task is to classify votes as Republican or Democrat
contrac	A subset of the 1987 National Indonesia Contraceptive Prevalence Survey: the samples are married women who were either not pregnant or do not know whether they were at the time of interview; the problem is to predict the current contraceptive method choice (no use, long-term methods or short-term methods) of a woman on the basis of her demographic and socio-economic characteristics
credit-approval	A collection of credit card applications: the task is to determine whether the application was approved or not
ctg	Measurements of fetal heart rate FHR and uterine contraction UC features on cardiotocograms classified by expert obstetricians: the task is to classify the fetal state as normal, suspect or pathologic
cylinder-bands	Predict process delays known as 'cylinder bands' in rotogravure printing
dermatology	The aim of this data set is to determine the type of eryhaemato- squamous disease
german_credit	This data set classifies people described by a set of attributes as good or bad credit risks
heart-cleveland	The goal is to determine the presence of heart disease in the patients: the outcome is integer valued from 0 (no presence) to 4; experiments with the Cleveland database have concentrated on simply attempting to distinguish presence (values 1, 2, 3 and 4) from absence (value 0)
ilpd	This data set contains 416 liver patient records and 167 non-liver-patient records: the data were collected from the north-east of Andhra Pradesh, India

(*continued*)

Table 5 (*continued*)

Data set	*Classification task*
mammo	Discrimination of benign and malignant mammographic masses based on BI-RADS attributes and the patient's age
mushroom	From the Audobon Society field guide: mushrooms are described in terms of physical characteristics; the task is to classify them as either poisonous or edible
wine	Using chemical analysis, determine the origin of wines
wine_qual	Two data sets are included, related to red and white wine samples, from the north of Portugal: the goal is to model wine quality on the basis of physicochemical tests

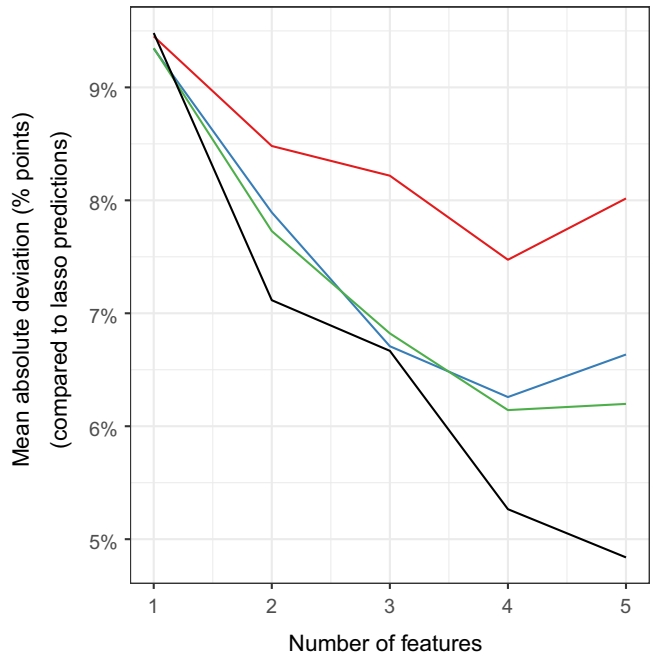

Fig. 10. Mean absolute deviation of probability estimates based on simple rules compared with those from a complex lasso model, averaged over all the UCI data sets: ——, $[-1, 1]$; ——, $[-2, 2]$; ——, $[-3, 3]$; ——, no rounding

Appendix B: Additional results for University of California, Irvine, data

B.1. Robustness of probability estimates with simple rules

To gauge the robustness of probability estimates that are derived from simple rules, we compare the mean absolute deviation of those estimates with the predictions from a lasso model that uses all available features. For each integer score that is produced by a select, regress and round model, we compute the corresponding probability estimate for the simple rule by considering all cases in the training set with that score, and then computing the empirical frequency of the outcome of interest, as detailed in Section 4.1. As shown in Fig. 10, using five features and rounding coefficients to the interval $[-3, 3]$, probability estimates by using select, regress and round deviate from the lasso predictions by about 6 percentage points on average across the UCI data sets.

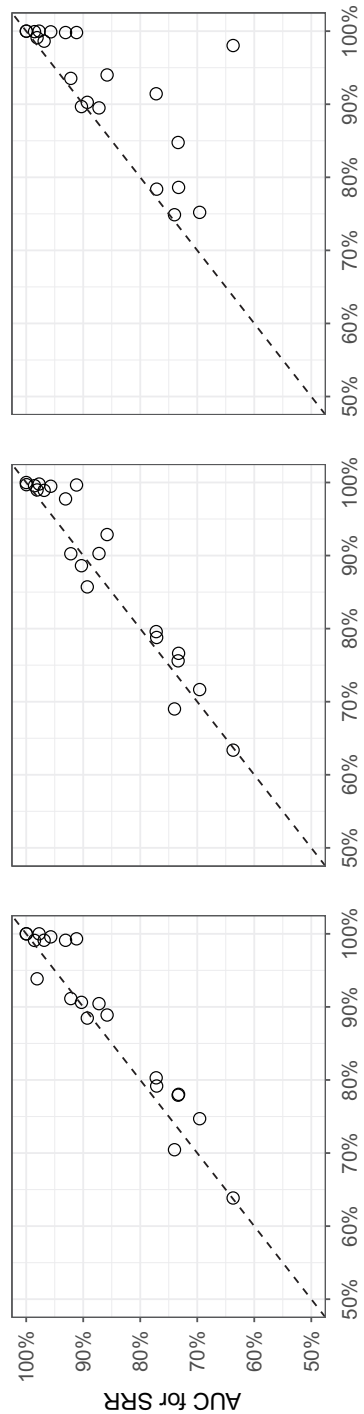

Fig. 11. Comparison of performance across each of the 21 UCI data sets, for simple rules using up to five features, and rounding to the nearest integer within the range $[-3, 3]$ (i.e. $k = 5$ and $M = 3$) (whereas random forests generally outperform select, regress and round, simple rules are comparable in performance with logistic and lasso regression across most data sets): (a) logistic; (b) lasso; (c) random forests

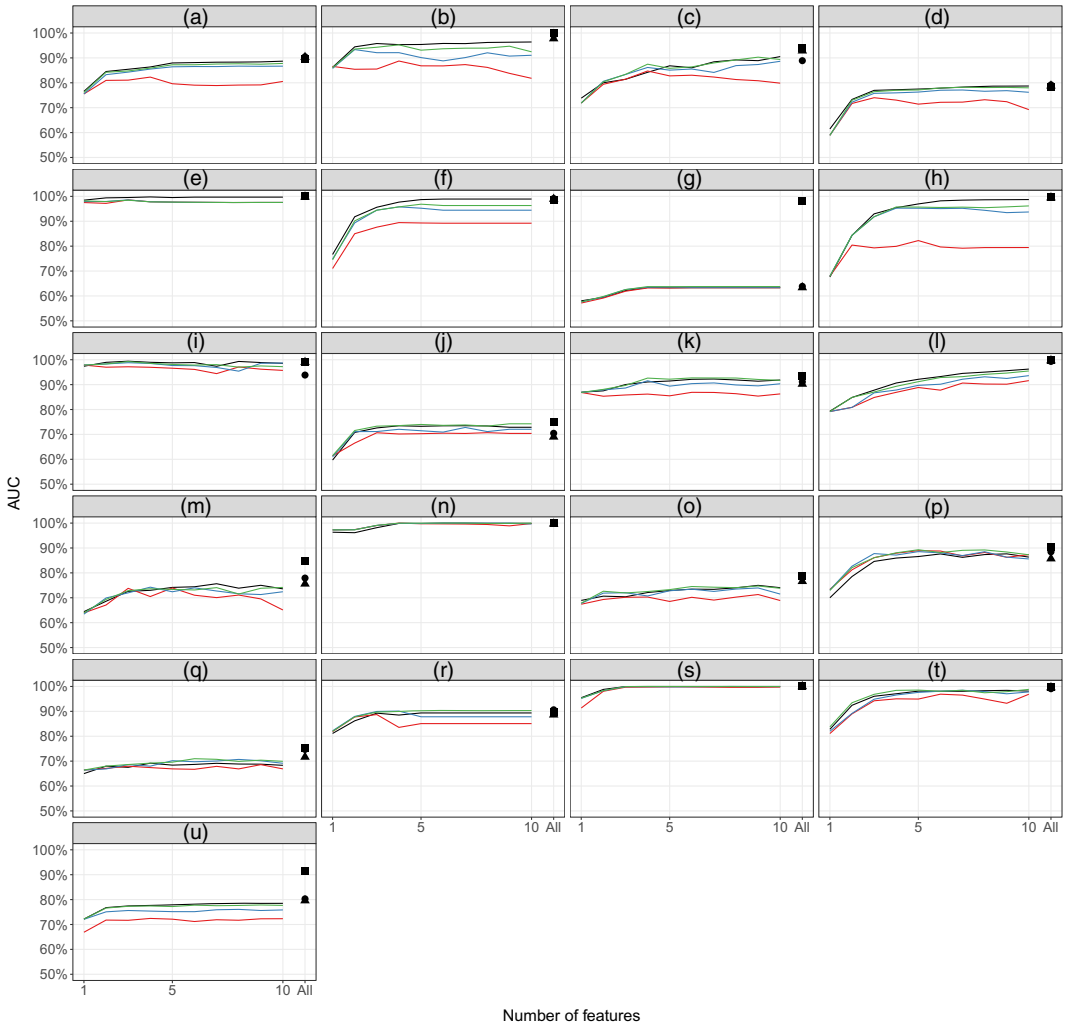

Fig. 12. Comparison of performance of AUC for each of the 21 UCI data sets, for simple rules with various k and M parameters (●, logistic; ▲, lasso; ■, random forests; ——, coefficient range $[-1, 1]$; ——, coefficient range $[-2, 2]$; ——, coefficient range $[-3, 3]$; ——, no rounding): (a) adult; (b) annealing; (c) audiology-std; (d) bank; (e) bankruptcy; (f) car; (g) chess-krvk; (h) chess-krvkp; (i) congress-voting; (j) contrac; (k) credit-approval; (l) ctg; (m) cylinder-bands; (n) dermatology; (o) german_credit; (p) heart-cleveland; (q) ilpd; (r) mammo; (s) mushroom; (t) wine; (u) wine_qual

B.2. Detailed results for individual data sets

Here, we disaggregate the results in Section 3.1 to compare simple rules and complex models on each of the 21 UCI data sets. First, Fig. 11 compares the performance of select, regress and round by using up to five features and rounding to the nearest integer in the range $[-3, 3]$ (i.e. $k = 5$ and $M = 3$) against each of the three benchmark models for each individual data set. In Fig. 11, each point represents a data set, and the corresponding horizontal and vertical positions show the cross-validated AUC of the complex models and simple rules respectively. For the logistic regression and lasso comparisons, all points are very close to the diagonal, indicating that select, regress and round performs on a par with these complex models for each individual data set. In contrast, we see that a random-forest model outperforms simple rules in many situations. Next, Fig. 12 provides a more detailed comparison by replicating Fig. 1 for each individual data set. We observe that, in general, model comparisons on individual data sets are similar to those that average

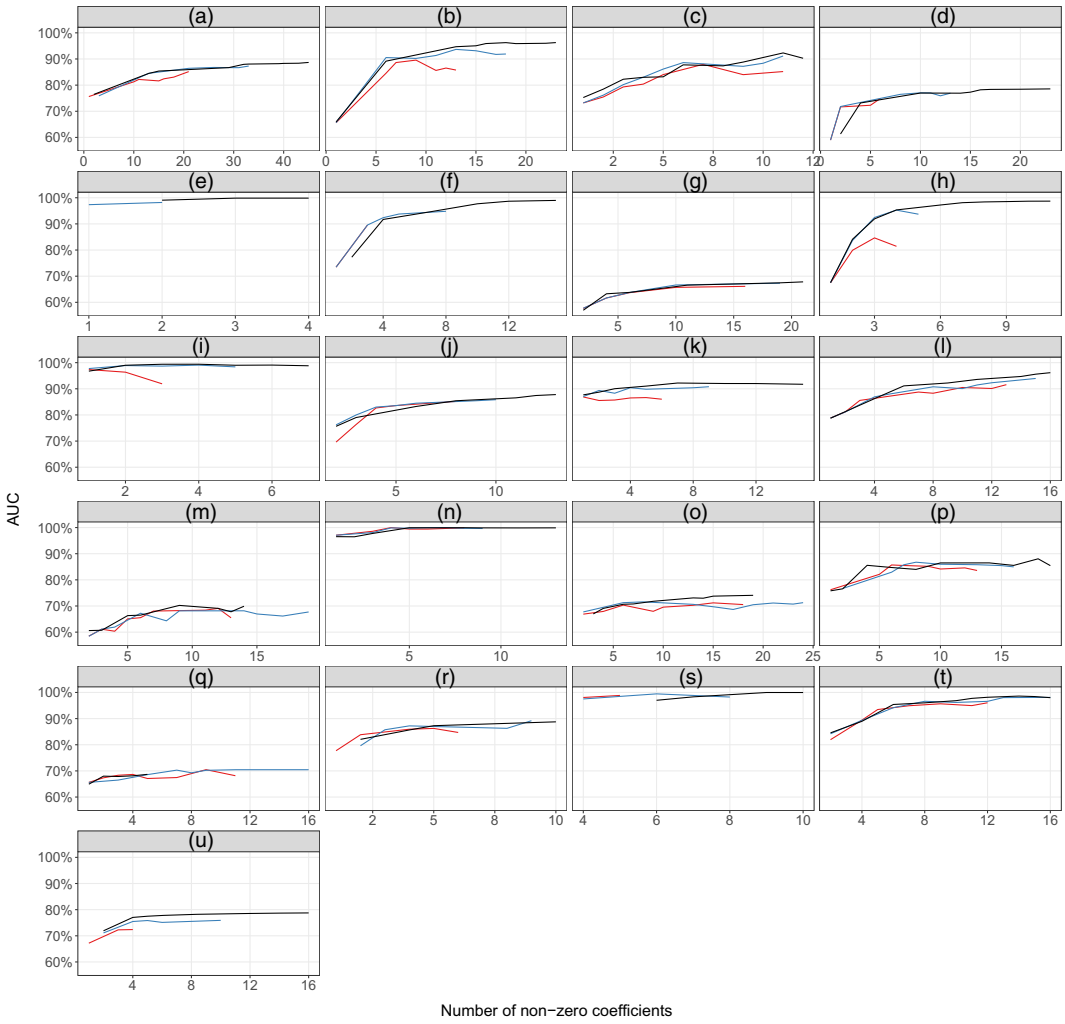

Fig. 13. Comparison of performance of AUC for each of the 21 UCI data sets, for simple rules with various k and M parameters (compared with Fig. 12, this figure shows the number of binarized features with non-zero coefficients on the horizontal axis) (—, coefficient range $[-1, 1]$; —, coefficient range $[-2, 2]$; —, coefficient range $[-3, 3]$; —, no rounding): (a) adult; (b) annealing; (c) audiology-std; (d) bank; (e) bankruptcy; (f) car; (g) chess-krvk; (h) chess-krvkp; (i) congress-voting; (j) contrac; (k) credit-approval; (l) ctg; (m) cylinder-bands; (n) dermatology; (o) german_credit; (p) heart-cleveland; (q) ilpd; (r) mammo; (s) mushroom; (t) wine; (u) wine_qual

over all data sets. Finally, in Fig. 13 we similarly plot performance but replace the horizontal axis with the number of non-zero coefficients instead of features. A model can have more non-zero coefficients than features, because a categorical variable with more than two categories will yield more than two non-zero coefficients after each category has been binarized. For example, the rule that is presented in Fig. 3 has two features, age and prior FTAs, but five non-zero coefficients.

Appendix C: Alternative case-study with german_credit data

We illustrate select, regress and round on the german_credit data. This data set consists of 1000 cases labelled as having either good or bad credit. Each row is described by 20 features: seven continuous and

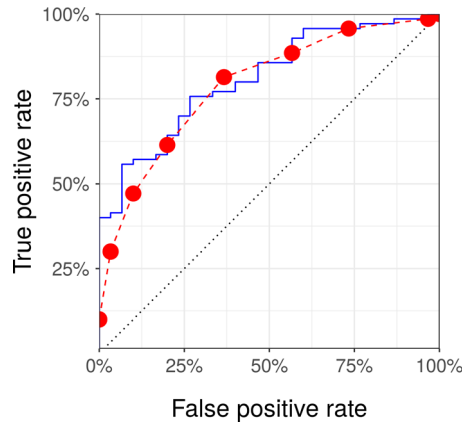

Fig. 14. Receiver operating characteristic curve comparing performance of a complex (——, random forest) model that uses all 20 available features to predict whether an individual has 'good' or 'bad' credit, *versus* the simple rule derived by using select, regress and round (——) as shown in Table 6: the curve for simple rules is shown as points, since a simple rule results in discrete cut-offs; the complex model and simple rule achieve AUCs of 0.80 and 0.78 respectively; whereas a decision maker would typically select a threshold based on various costs given a risk score, our results show that a simple rule achieves almost identical performance compared with a complex model for all possible threshold values, with the additional benefit of being transparent and interpretable

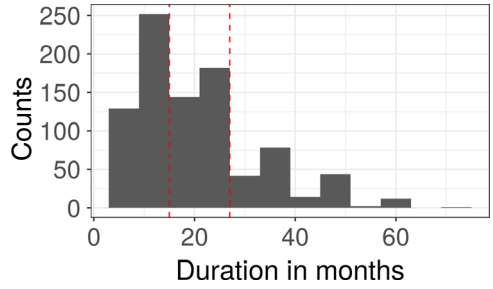

Fig. 15. Distribution of the continuous feature duration in months in the training data: ⋮, values at the 33rd and 67th percentiles, which are used to discretize the continuous feature into three bins of approximately equal size

13 categorical; a full description of the data set can be found at the UCI repository. The goal is to estimate the risk of default (labelled as 'bad' credit) for each case. We split the data randomly into 900 cases for training and 100 cases for evaluation. We use a single 9:1 split for simplicity here, but note that our main results that are reported in Fig. 1 were obtained via tenfold cross-validation.

As a benchmark, we first fit a random-forest model with 1000 trees on the training data, using all 20 available features. The result is a complex model that achieves 0.80 AUC on the test set. A full receiver operating characteristic curve for the complex model is presented as a blue curve in Fig. 14. According to the steps that were presented in Section 2, we build a simple rule to score the risk of default for a given case. As described previously, we discretize continuous features into three bins of approximately equal sizes to prioritize simplicity. This is achieved by discretizing each continuous feature at the 33rd and 67th percentile in the training data, and applying the same cut-offs to the test data. For example, Fig. 15 shows the distribution of the duration in months feature in the training data, which represents the number of months that an applicant has lived at their current address. The 33rd and 67th percentile of this feature are 12 and 24 respectively: hence the feature is discretized at these points for both the training and the test data. In detail, we perform the following steps.

Table 6. Simple rule for determining the risk of default, derived by using select, regress and round on the german_credit data set with $k = 5$ and $M = 3$†

Selected feature	*Lasso coefficient*	*Select, regress and round score*
Checking account status		
Less than 200 DM	0.3	1
200 DM or above	0.92	2
No checking account	1.56	3
Months lived at current address		
Between 12 and 24 months	−0.39	−1
24 months or more	−1.02	−3
Credit history		
All credits at this bank paid back	−0.36	−1
Delayed payments in the past	0.56	1
Unpaid credits existing (not at this bank)	1.23	3
Savings account or bonds		
$100 \leqslant$ value < 500 DM	0.1	0
$500 \leqslant$ value < 1000 DM	0.67	2
1000 DM $\leqslant$ value	1.002	2
No known savings account	0.99	2
Guarantors		
Coapplicant	−0.34	−1
Guarantor	1.35	3

†Variables with a zero lasso coefficient have been omitted.

Step 1: *select*—from the full set of 20 features, we select $k = 5$ features via forward stepwise selection. The features that are selected, in order, are checking account status, months lived at current address, credit history, savings account/bonds, and guarantors.
Step 2: *regress*—using the five selected features, we train an L^1-regularized (lasso) logistic regression model to predict whether the credit is good (0) or bad (1) for each case. The regularization parameter λ is chosen via tenfold cross-validation. Following Friedman *et al.* (2010), we explore a regularization path with 1000 values of λ spaced evenly on a log-scale in the range $(\lambda_{\min}, \lambda_{\max})$, where $\lambda_{\min} = 10^{-4}$ and $\lambda_{\max}$ is set to 0.141, the minimum value such that all coefficients are regularized to 0. We find that λ^*, the value of λ that maximizes cross-validated performance, is 0.004. The second column of Table 6 shows the fitted lasso model coefficients.
Step 3: *round*—we rescale the coefficients of the model from step 2 to be in the range $[-3, 3]$ (e.g. $M = 3$), and then round the rescaled coefficients to the nearest integer. The final scores corresponding to each variable are listed in the third column of Table 6.

Fig. 14 shows a comparison of receiver operating characteristic curve performance on the held-out test set, between the random-forest model (blue) and our simple rule in Table 5 (red). In practice, a decision maker would typically select a threshold based on various costs to determine which loan applications to approve or reject. However, our results demonstrate that, across all threshold values, a simple, transparent rule achieves nearly identical performance when compared with a considerably more complex model.

This case-study and all the results can be replicated by running the `case_study.R` script that is provided in our public code repository: `https://github.com/stanford-policylab/simple-rules`. In addition, we have made it easy to generate simple rules for any combination of parameters for each of the 21 UCI data sets, by providing an R markdown file that can be run by using freely available software.

References

Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., Nichols, C. N., Lampropoulos, G. K., Walker, B. S., Cohen, G. and Rush, J. D. (2006) The meta-analysis of clinical judgment project: fifty-six years of accumulated research on clinical versus statistical prediction. *Counslng Psychol.*, **34**, 341–382.

Åstebro, T. and Elhedhli, S. (2006) The effectiveness of simple decision heuristics: forecasting commercial success for early-stage ventures. *Mangmnt Sci.*, **52**, 395–409.

Bareiss, E. R., Porter, B. W. and Wier, C. C. (1988) Protos: an exemplar-based learning apprentice. *Int. J. Man-Mach. Stud.*, **29**, 549–561.

Bohanec, M. and Rajkovic, V. (1988) Knowledge acquisition and explanation for multi-attribute decision making. In *Proc. 8th Int. Wrkshp Expert Systems and Their Applications, Avignon*, pp. 59–78.

Camerer, C. F. and Johnson, E. J. (1997) The process-performance paradox in expert judgment. In *Research on Judgment and Decision Making: Currents, Connections, and Controversies* (eds W. M. Goldstein and R. M. Hogarth). New York: Cambridge University Press.

Cassel, C. M., Särndal, C. E. and Wretman, J. H. (1976) Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, **63**, 615–620.

Chambers, J. M., Hastie, T. J. and Pregibon, D. (1992) *Statistical Models in S*. Pacific Grove: Wadsworth and Brooks–Cole.

Christin, A. (2017) Algorithms in practice: comparing web journalism and criminal justice. *Big Data Soc.*, **4**, 1–14.

Corbett-Davies, S. and Goel, S. (2018) The measure and mismeasure of fairness: a critical review of fair machine learning. *Preprint arXiv:1808.00023*. Stanford University, Stanford.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S. and Huq, A. (2017) Algorithmic decision making and the cost of fairness. In *Proc. 23rd Int. Conf. Knowledge Discovery and Data Mining* (eds S. Matwin, S. Yu and F. Farooq), pp. 797–806. New York: Association for Computing Machinery.

Danziger, S., Levav, J. and Avnaim-Pesso, L. (2011) Extraneous factors in judicial decisions. *Proc. Natn. Acad. Sci. USA*, **108**, 6889–6892.

Dawes, R. M. (1979) The robust beauty of improper linear models in decision making. *Am. Psychol.*, **34**, 571–582.

Dawes, R. M., Faust, D. and Meehl, P. E. (1989) Clinical versus actuarial judgment. *Science*, **243**, 1668–1674.

DeMichele, M., Baumgartner, P., Barrick, K., Comfort, M., Scaggs, S. and Misra, S. (2018) What do criminal justice professionals think about risk assessment at pretrial? *Preprint*. RTI International. (Available from `https://ssrn.com/abstract=3168490`.)

Dhami, M. K. (2003) Psychological models of professional decision making. *Psychol. Sci.*, **14**, 175–180.

Dudík, M., Langford, J. and Li, L. (2011) Doubly robust policy evaluation and learning. Microsoft. (Available from `http://arxiv.org/abs/1103.4601`.)

Einhorn, H. J. and Hogarth, R. M. (1975) Unit weighting schemes for decision making. *Organznl Behav. Hum. Perform.*, **13**, 171–192.

Fernández-Delgado, M., Cernadas, E., Barro, S. and Amorim, D. (2014) Do we need hundreds of classifiers to solve real world classification problems? *J. Mach. Learn. Res.*, **15**, 3133–3181.

Friedman, J., Hastie, T. and Tibshirani, R. (2010) Regularization paths for generalized linear models via coordinate descent. *J. Statist. Softwr.*, **33**, 1–22.

Gelman, A. and Park, D. K. (2009) Splitting a predictor at the upper quarter or third and the lower quarter or third. *Am. Statistn*, **63**, 1–8.

Gigerenzer, G. and Goldstein, D. G. (1996) Reasoning the fast and frugal way: models of bounded rationality. *Psychol. Rev.*, **103**, 650–659.

Gleicher, M. (2016) A framework for considering comprehensibility in modeling. *Big Data*, **4**, 75–88.

Goel, S., Shroff, R., Skeem, J. L. and Slobogin, C. (2020) The accuracy, equity, and jurisprudence of criminal risk assessment. Stanford University, Stanford. (Available from `http://dx.doi.org/10.2139/ssrn.3306723`.)

Goodman, B. and Flaxman, S. (2016) EU regulations on algorithmic decision-making and a "right to explanation". *Wrkshp Human Interpretability in Machine Learning, New York*. (Available from `http://arxiv.org/abs/1606.08813v1`.)

Gottfredson, D. M. (1999) Effects of judges' sentencing decisions on criminal careers. Office of Justice Programs, US Department of Justice, National Institute of Justice, Washington DC.

Green, B. F. (1977) Parameter sensitivity in multivariate methods. *Multiv. Behav. Res.*, **12**, 263–287.

Hill, J. L. (2012) Bayesian nonparametric modeling for causal inference. *J. Computnl Graph. Statist.*, **20**, 217–240.

Kang, J. D. and Schafer, J. L. (2007) Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Statist. Sci.*, **22**, 523–539.

Kim, B., Shah, J. A. and Doshi-Velez, F. (2015) Mind the gap: a generative approach to interpretable feature selection and extraction. In *Advances in Neural Information Processing Systems* (eds C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama and R. Garnett), pp. 2260–2268.

Klein, G. A. (2017) *Sources of Power: how People Make Decisions*. Cambridge: MIT Press.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J. and Mullainathan, S. (2017) Human decisions and machine predictions. *Q. J. Econ.*, **133**, 237–293.

Kleinberg, J., Ludwig, J., Mullainathan, S. and Obermeyer, Z. (2015) Prediction policy problems. *Am. Econ. Rev.*, **105**, 491–495.
Krauss, D. A. (2004) Adjusting risk of recidivism: do judicial departures worsen or improve recidivism prediction under the federal sentencing guidelines? *Behav. Sci. Law*, **22**, 731–750.
Lakkaraju, H., Bach, S. H. and Leskovec, J. (2016) Interpretable decision sets: a joint framework for description and prediction. In *Proc. 22nd Int. Conf. Knowledge Discovery and Data Mining*. New York: Association for Computing Machinery.
Marewski, J. N. and Gigerenzer, G. (2012) Heuristic decision making in medicine. *Dial. Clin. Neursci.*, **14**, 77–89.
McDonald, C. J. (1996) Medical heuristics: the silent adjudicators of clinical practice. *Ann. Intern. Med.*, **124**, 56–62.
Meehl, P. E. (1954) *Clinical vs. Statistical Prediction*. Minneapolis: University of Minnesota Press.
Ribeiro, M. T., Singh, S. and Guestrin, C. (2016) "Why should I trust you?": Explaining the predictions of any classifier. In *Proc. 22nd Int. Conf. Knowledge Discovery and Data Mining*, pp. 1135–1144. New York: Association for Computing Machinery.
Robins, J. M. and Rotnitzky, A. (1995) Semiparametric efficiency in multivariate regression models with missing data. *J. Am. Statist. Ass.*, **90**, 122–129.
Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994) Estimation of regression coefficients when some regressors are not always observed. *J. Am. Statist. Ass.*, **89**, 846–866.
Rosenbaum, P. R. and Rubin, D. B. (1983a) Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *J. R. Statist. Soc.* B, **45**, 212–218.
Rosenbaum, P. R. and Rubin, D. B. (1983b) The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70**, 41–55.
Rosenbaum, P. R. and Rubin, D. B. (1984) Reducing bias in observational studies using subclassification on the propensity score. *J. Am. Statist. Ass.*, **79**, 516–524.
Stevenson, M. (2018) Assessing risk assessment in action. *Minn. Law Rev.*, **103**, 303–384.
Su, J. Q. and Liu, J. S. (1993) Linear combinations of multiple diagnostic markers. *J. Am. Statist. Ass.*, **88**, 1350–1355.
Sull, D. and Eisenhardt, K. M. (2015) *Simple Rules: how to Thrive in a Complex World*. New York: Houghton Mifflin Harcourt.
Tetlock, P. (2005) *Expert Political Judgment: how Good is It?: how Can We Know*? Princeton: Princeton University Press.
Ustun, B. and Rudin, C. (2016) Supersparse linear integer models for optimized medical scoring systems. *Mach. Learn.*, **102**, 349–391.
Van Belle, V. M., Van Calster, B., Timmerman, D., Bourne, T., Bottomley, C., Valentin, L., Neven, P., Van Huffel, S., Suykens, J. A. and Boyd, S. (2012) A mathematical model for interpretable clinical decision support with applications in gynecology. *PLOS One*, **7**, article e34312.
Waller, N. and Jones, J. (2011) Investigating the performance of alternate regression weights by studying all possible criteria in regression models with a fixed set of predictors. *Psychometrika*, **76**, 410–439.
Wang, F. and Rudin, C. (2015) Falling rule lists. In *Artificial Intelligence and Statistics* (eds G. Lebanon and S. Vishwanathan), pp. 1013–1022. San Diego: Society for Artificial Intelligence and Statistics.
Wübben, M. and Wangenheim, F. V. (2008) Instant customer base analysis: managerial heuristics often get it right. *J. Marktng*, **72**, 82–93.
Wyatt, J. C. and Altman, D. G. (1995) Prognostic models: clinically useful or quickly forgotten? *Br. Med. J.*, **311**, 1539–1541.