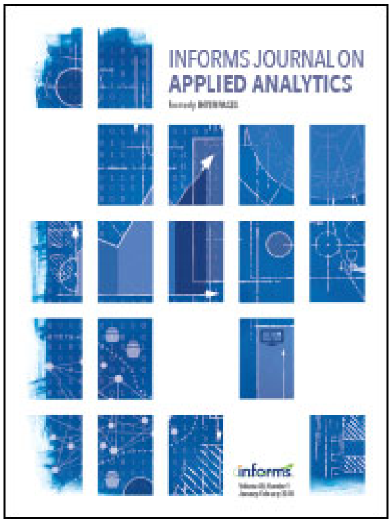


INFORMS Journal on Applied Analytics

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Optimized Scoring Systems: Toward Trust in Machine Learning for Healthcare and Criminal Justice

Cynthia Rudin, Berk Ustun

To cite this article:
Cynthia Rudin, Berk Ustun (2018) Optimized Scoring Systems: Toward Trust in Machine Learning for Healthcare and Criminal Justice. INFORMS Journal on Applied Analytics 48(5):449-466. https://doi.org/10.1287/inte.2018.0957

informs
http://pubsonline.informs.org/journal/inte/

INTERFACES
Vol. 48, No. 5, September–October 2018, pp. 449–466
ISSN 0092-2102 (print), ISSN 1526-551X (online)

Optimized Scoring Systems: Toward Trust in Machine Learning for Healthcare and Criminal Justice

Cynthia Rudin,[a] Berk Ustun[b]

[a] Departments of Computer Science, Electrical and Computer Engineering, and Statistical Science, Duke University, Durham, North Carolina 27708; [b] Center for Research in Computation for Society, Harvard John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts 02138

Contact: cynthia@cs.duke.edu, http://orcid.org/0000-0003-4283-2780 (CR); berk@seas.harvard.edu (BU)

https://doi.org/10.1287/inte.2018.0957

Abstract. Questions of trust in machine-learning models are becoming increasingly important as these tools are starting to be used widely for high-stakes decisions in medicine and criminal justice. Transparency of models is a key aspect affecting trust. This paper reveals that there is new technology to build transparent machine-learning models that are often as accurate as black-box machine-learning models. These methods have already had an impact in medicine and criminal justice. This work calls into question the overall need for black-box models in these applications.

Funding: Funding provided by the National Science Foundation, Philips, Siemens, and Wistron.

Keywords: machine learning • sparse linear models • scoring systems • trust • transparency • interpretability • healthcare • criminal justice • recidivism

There has been an increasing trend in healthcare and criminal justice to leverage machine learning for high-stakes prediction problems, such as detecting heart attacks (Weng et al. 2017), diagnosing Alzheimer's disease (Pekkala et al. 2017), and assessing recidivism risk (Berk and Bleich 2013, Tollenaar and van der Heijden 2013). In many of these problems, practitioners are deploying black-box machine-learning models that do not explain their predictions in a way that humans can understand. In some cases, model development is outsourced to private companies that build and sell proprietary predictive models using confidential data sets without regulatory oversight.

The lack of transparency and accountability of a predictive model can have severe consequences when it is used to make decisions that significantly affect human lives. In criminal justice, proprietary predictive models can lead to questions about due process or may discriminate based on race or poverty status (Wexler 2017b). In 2015, for example, Billy Ray Johnson was imprisoned based on evidence from software developed by a private company, TrueAllele, which refused to reveal how the software worked. This led to a landmark case (People v. Chubbs) in which the California appeals court ruled that such companies were not required to reveal how their software worked. As a different example, consider the controversy surrounding the COMPAS recidivism prediction model (Northpointe 2015), which is used for several applications in the U.S. criminal justice system but does not provide clear reasons for its predictions. COMPAS has been accused of discriminating on the basis of race (Angwin et al. 2016, Citron 2016) and possibly uses socioecononomic information, such as how often the individual is not paid above minimum wage. There has been some debate as to whether these claims are likely to be true (e.g., see Fisher et al. 2018).

A key problem with proprietary models is that they are prone to data-entry errors. There have been cases such as that of Glenn Rodríguez, a prisoner with a nearly perfect record, who was denied parole as a result of an incorrectly calculated COMPAS score (Wexler 2017a, b) with little recourse to argue or even to determine how his score was computed. There have been cases in which criminological risk scores (even simple ones) were miscalculated, allowing dangerous criminals to be released and, subsequently, to commit murders (Ho 2017) or other crimes. Issues such as those discussed herein have led to new regulations, such as the European Union's "right to explanation" (Goodman and Flaxman 2016), which requires explanations from any algorithmic decision-making tool that significantly affects humans.

Because mistakes in healthcare and criminal justice can be serious or even deadly, it can be beneficial for companies not to disclose their models. If the model is allowed to be hidden, the company never needs to fully justify why any particular prediction was made and could avoid liability when the model makes mistakes. This leads to misaligned incentives with which the users of the tools would strongly benefit from transparent predictive models, but this would equally undermine profits for selling predictive models. Because these industries have a strong disincentive from building transparent models, there has been little work done on determining the answers to the following questions:

1. *Are there interpretable predictive models that are as accurate as black-box models?* When we trust companies to build black-box models, we are implicitly assuming that their models are more accurate than transparent models. Is it possible that, for many given black-box models, an alternative model exists that is just as accurate but that is so simple that it can fit on an index card? We claim the answer is yes. A compelling argument of Breiman (2001), called the *Rashomon effect*, indicates that, for many applications, there may exist a large class of models that predict almost equally well. Among this large class of models are those from the various black-box machine-learning methods (e.g., support vector machines, random forests, boosted decision trees, neural networks). There is no inherent reason that this class would exclude interpretable models. This observation also helps to explain the 40 years of literature on the surprising performance of simple linear models (Dawes 1979, Holte 1993).

2. *What are the desired characteristics of an interpretable model if one exists?* The answer to this question changes for each audience and application (Kodratoff 1994, Pazzani 2000, Freitas 2014). We might desire accuracy in predictions, risks that are calibrated, and we might want the model to be calculated by a judge or a doctor without a calculator, which makes it easier to explain to a defendant or medical patient. Predictions from simpler models are much easier to verify, leading to fewer calculation errors and more robust decisions. A model with all of the characteristics listed here may not exist for any given problem, but if it does, it would be better to use than a black box.

3. *If an interpretable model does exist, is it possible to find it?* Interpretability, transparency, usability, and other desirable characteristics in predictive models lead to computationally hard optimization problems, such as mixed-integer nonlinear programs. It is much easier to find an accurate unintelligible model than an interpretable one.

The renaissance from proprietary predictive models back to interpretable predictive models can only be partially determined by regulations such as "right to explanation." Instead, the restoration to interpretable models should fundamentally be driven by technology. It must be demonstrated that interpretable models can achieve performance comparable with black-box models. That is what this work focuses on.

We present two machine-learning algorithms, called *supersparse linear integer models* (SLIM) and *risk-calibrated supersparse linear integer models* (RiskSLIM), which solve mixed-integer linear and nonlinear programs. They produce sparse linear models directly from data that are faithful to the century-old scoring-system model form, similar to the predictive models that have been used over the last century. SLIM produces scoring systems optimized for desired true-positive/false-positive trade-offs, whereas RiskSLIM produces risk scores. Both methods leverage modern optimization tools and avoid well-known pitfalls of rounding methods. The models come with optimality guarantees; that is, they allow one to test for the existence of interpretable models that are as accurate as black-box models. RiskSLIM's models are risk calibrated across the spectrum of true positives and false positives (or sensitivity and specificity), and both methods honor constraints imposed by the domain. Software for both methods is public and could be used to challenge the use of black-box models for high-stakes decisions.

SLIM and RiskSLIM are already challenging decision-making processes for applications in medicine and criminal justice. We focus on three of them in this work. (1) *Sleep apnea screening*: In joint work with Massachusetts General Hospital (Ustun et al. 2016), we determined that a scoring system built using a patient's medical history can be as accurate as one that relies on reported symptoms. This yields savings in the efficiency and effectiveness of medical care for sleep apnea patients. (2) *ICU seizure prediction*: In joint work with Massachusetts General Hospital (Struck et al. 2017), we created the first scoring system that uses continuous EEG measurements to predict seizures, called 2HELPS2B. The model provides concise reasons why a patient may be at risk. (3) *Recidivism prediction*: The recent public debate regarding recidivism prediction and whether COMPAS's proprietary predictions are racially biased (Angwin et al. 2016) leads to the question of whether interpretable models exist for recidivism prediction. In our studies of recidivism (Ustun and Rudin 2016a, 2017; Zeng et al. 2017), we used the largest publicly available data set on recidivism and showed that SLIM and RiskSLIM could produce small scoring systems that are as accurate as state-of-the-art machine-learning models. This calls into question the necessity of tools such as COMPAS and the reasons for government expenditures for predictions from proprietary models.

Scoring Systems: Applications and Prior Art

The use of predictive models is not new to society; only the use of black-box models is relatively new. Scoring systems, which are a widely used form of interpretable predictive model, date back at least to work on parole violation by Burgess (1928). One example of a scoring system is the CHADS$_2$ score (Gage et al. 2001)

Figure 1. The CHADS$_2$ Score to Assess Stroke Risk

1.	**C**ongestive Heart Failure	1 point	···
2.	**H**ypertension	1 point	+ ···
3.	**A**ge ≥ 75	1 point	+ ···
4.	**D**iabetes Mellitus	1 point	+ ···
5.	Prior **S**troke or Transient Ischemic Attack	**2** points	+
		SCORE	=

SCORE	0	1	2	3	4	5	6
RISK	1.9%	2.8%	4.0%	5.9%	8.5%	12.5%	18.2%

Source. Gage et al. (2001).
Notes. For each patient, the score is computed as the sum of the patients' points. The score is translated into the one-year stroke risk using the lower table.

(Figure 1), which predicts stroke in patients with atrial fibrillation and is arguably the most widely used predictive model in medicine. Scoring systems are sparse linear models with small integer coefficients. The coefficients are the point scores: for CHADS$_2$, the coefficients are 1, 1, 1, 1, and 2.

The vast majority of predictive models in the healthcare system and justice system are scoring systems. Other examples from healthcare include SAPS I, II, and III (Le Gall et al. 1993, Moreno et al. 2005); APACHE I, II, and III to assess ICU mortality risk (Knaus et al. 1981, 1985, 1991); TIMI to assess the risk of death and ischemic events (Antman et al. 2000); HEART (Six et al. 2008); EDACS (Than et al. 2014) for cardiac events; PCL to screen for PTSD (Weathers et al. 2013); and SIRS to detect system inflammatory response syndrome (Bone et al. 1992). Examples from criminal justice include the Ohio Risk Assessment System (Latessa et al. 2009), the Kentucky Pretrial Risk Assessment Instrument (Austin et al. 2010), the Salient Factor Score (Hoffman and Adelberg 1980, Hoffman 1994), and the Criminal History Category (U.S. Sentencing Commission 1987).

None of the scoring systems listed in the previous paragraphs were optimized for predictive performance on data. Each scoring system was created using a different method. Some were built using only domain expertise (i.e., no data), and some were created using logistic regression followed by rounding of coefficients to obtain integer-valued point scores.

Serious problems with rounding heuristics are well documented in the optimization literature. When we solve a relaxed problem and round values to integers afterward, we know that (unless the problem has specific properties) either the solutions become infeasible or suboptimal. It is easy to find problems in discrete-optimization textbooks in which rounding leads to flawed solutions. In the case of linear regression or linear classification models, coefficients that are small are all rounded to zero; thus, an important part of the signal can easily be lost. We should not be using rounding heuristics if we want a reliable high-quality solution despite the government's recommendation (Gottfredson and Snyder 2005) to round logistic regression coefficients.

An additional set of challenges arises when models need to satisfy operational constraints, which are user-defined requirements for the model (e.g., false-positive rate below 20%). It is extremely difficult to design rounding heuristics that produce accurate models that also obey operational constraints. Heuristics for model design lead to suboptimal models, which, in turn, could lead to poor decision making for high-stakes applications.

Since its inception, the field of discrete optimization has been advancing although all of the scoring systems have been built without using discrete-optimization technology. Let us describe the optimization problems that we actually desire to solve when building scoring systems.

Optimization Problems and Methods

Here, we discuss two kinds of scoring systems:

1. *Decision rules* are scoring systems for decision making produced by SLIM. Here, predictions are based on whether the total score exceeds a threshold value (e.g., predict "yes" if total score >1). The choice of variables and points in the score function is optimized for accuracy at a specific decision point—a specific true-positive rate (TPR) or false-positive rate (FPR). The desired choice of TPR or FPR depends on the application. For medical screening, one might desire a larger false-positive rate so that the test is more likely to falsely identify someone as positive for a disease than to dismiss someone who has the disease by giving the person a negative test result. The user could specify the maximum false-positive rate the user is willing to tolerate, and SLIM will optimize the true-positive rate subject to that constraint.

2. *Risk scores* are scoring systems for risk assessment produced by RiskSLIM. These models use the score to generate a risk estimate. The choice of variables and points in the score function is optimized for risk calibration. A scoring system is risk calibrated when the predicted risk of the outcome (from the model) matches the risk of outcome in the data. These models do not optimize a specific TPR–FPR trade-off; rather they aim to achieve the highest true-positive rate for each false-positive rate.

We illustrate the difference between these two types of scoring systems in Figure 2, in which we show SLIM and RiskSLIM models for predicting whether a prisoner will be arrested within three years of being released from prison. Both models were built using the largest publicly available data set on recidivism and perform similarly to state-of-the-art machine-learning models (as we discuss in the *Applications and Insights*

Figure 2. Optimized Scoring Systems for Recidivism Predictions Built Using SLIM (Top) and RiskSLIM (Bottom)

SLIM scoring system

1.	Age at Release between 18 to 24	2 points	···
2.	Prior Arrests $\geq$ 5	2 points	+ ···
3.	Prior Arrest for Misdemeanor	1 point	+ ···
4.	No Prior Arrests	-1 point	+ ···
5.	Age at Release $\geq$ 40	-1 point	+ ···
		SCORE	= ···

PREDICT ARREST FOR ANY OFFENSE IF SCORE $>$ 1

RiskSLIM risk score

1.	Prior Arrests $\geq$ 2	1 point	···
2.	Prior Arrests $\geq$ 5	1 point	+ ···
3.	Prior Arrests for Local Ordinance	1 point	+ ···
4.	Age at Release between 18 to 24	1 point	+ ···
5.	Age at Release $\geq$ 40	-1 points	+ ···
		SCORE	= ···

SCORE	-1	0	1	2	3	4
RISK	11.9%	26.9%	50.0%	73.1%	88.1%	95.3%

Notes. The outcome variable for both models is whether a prisoner is arrested within three years of release from prison. The SLIM scoring system outputs a predicted outcome. It has a test TPR–FPR of 76.6%–44.5% and a mean fivefold cross-validation TPR–FPR of 78.3%–46.5%. The RiskSLIM scoring system outputs a risk estimate. It has a fivefold cross-validation mean test CAL–AUC of 1.7%–0.697% and training CAL–AUC of 2.6%–0.701%. We provide a definition of these performance metrics in the Evaluation Methodology for Machine-Learning Models section. See Zeng et al. (2017) and Ustun and Rudin (2016a) for more details.

section). The SLIM scoring system generates a decision rule (predict "yes" if the total score exceeds a threshold score), whereas the RiskSLIM scoring system outputs a table of risk estimates for each distinct score. In both cases, the choice of variables and the number of points are chosen to optimize the relevant performance metric by solving a discrete-optimization problem.

SLIM solves one constrained optimization problem to produce decision rules, and RiskSLIM solves a different problem to produce risk scores. Solving these optimization problems directly is principled, obviates the need for rounding and other manipulation, and directly encodes what we desire in a scoring system. The optimization problems are described mathematically in the appendix. In particular, we note the following:

- In both optimization problems (the decision-rule optimization and risk-score model optimization), hard constraints are used to force the coefficients to integer values.
- In both optimization problems, the objective we minimize includes a term that encourages the number of questions asked in the scoring system to be small (model sparsity).
- In the objective for SLIM, there is a term that encourages the point values to be small (e.g., it prefers value "one point" rather than value "seven points"). This also encourages the point values to be coprime; that is, they share no common prime factors. Thus, this formulation would never choose point scores 10, 10, 20, 10, 40; rather it would choose 1, 1, 2, 1, 4 to solve the same problem.
- In the formulation for RiskSLIM, the objective includes a term used in logistic regression (the *logistic loss*) that encourages the scores to be small and risk calibrated. As we define later, a model is risk calibrated when its predicted risks agree with risks calculated directly from the data.

Both optimization problems can accommodate constraints on the solution that are specific to the domain (operational constraints). Table 1 illustrates some types of constraints.

Both optimization problems are computationally hard, but theoretical results allow practical improvements

Table 1. Examples of Operational Constraints That Can Be Addressed

Constraint type	Example
Feature selection	Choose up to 10 features
Group sparsity	Include either *male* or *female*, not both
Optimal thresholding	Use at most three thresholds for *age*, for example, (age $\leq$ 30, age $\leq$ 50, age $\leq$ 75)
Logical structure	If *male* is in model, then also include *hypertension*
Probability	Predict $y = 1$ with at least 90% probability when *male* = true
Fairness	Ensure that the predicted outcome $\hat{y}$ is +1 an equal number of times for *male* and *female*

Notes. Both SLIM and RiskSLIM can handle constraints on model form. SLIM handles constraints related to error metrics (e.g., fairness constraints). RiskSLIM handles constraints on risk estimates (e.g., probability constraints as in the second to last row).

in speed. As a result, both the decision-rule optimization problem and the risk-score optimization problem can be solved for reasonably large data sets in minutes.

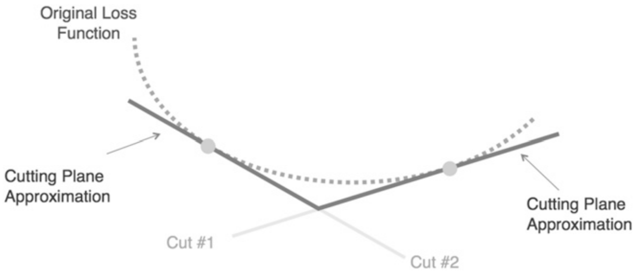
The risk-score problem is a mixed-integer nonlinear program because the logistic loss is nonlinear. However, because the logistic loss is convex, cutting planes would be a natural type of technique for this problem. Cutting-plane techniques produce piecewise linear approximations to the objective (cuts), which produce a surrogate lower bound, labeled "cutting-plane approximation" in the illustration in Figure 3. However, traditional cutting-plane methods fail badly for the risk-score problem. Because the feasible region is the integer lattice, a traditional cutting-plane method would need to solve a mixed-integer program (MIP) to optimality to develop each new cut. If this surrogate MIP is not solved to optimality, we have no way of knowing when we have reached the solution to the risk-score problem. After several iterations, enough cuts would accumulate that the MIP could not be solved to optimality in a reasonable amount of time, and the program would stall and fail to provide optimal scoring systems. This necessitates a new approach.

We developed a new branch-and-bound cutting-plane method used in RiskSLIM for solving the risk-score problem (Ustun and Rudin 2017). This method does not stall, involves solving linear programs rather than mixed-integer programs, and can be implemented using standard callback functions in CPLEX (ILOG 2007). The method gracefully handles arbitrarily large data sets (even millions of observations) because computation scales linearly with the number of observations. The RiskSLIM model in Figure 2 was fit on a data set with $N = 22,530$ observations in 20 minutes.

SLIM's decision-rule problem (unlike the risk-score problem we just described for RiskSLIM) is a mixed-integer linear program. It can be solved with optimization software, such as CPLEX, but the solver is made more efficient with a specialized bound that we constructed, which reduces the amount of data we use without changing the solution to the optimization problem (discussed in Ustun and Rudin 2016b).

In the appendix, we discuss the optimization problems solved by SLIM and RiskSLIM. Before we discuss applications, let us discuss means of evaluation.

Figure 3. A Convex Loss Function (Smooth Curve) and Its Surrogate Lower Bound (Lines)



Evaluation Methodology for Machine-Learning Models

The fields of machine learning and data mining use rigorous empirical evaluation techniques. *Cross-validation* is commonly used to provide a measure of uncertainty of prediction quality. To perform fivefold cross-validation, the data are divided into five equal-size folds. Four of the folds are used to train the algorithm, and predictions are made out-of-sample on the fifth "test" fold. The test fold rotates, and we report a mean and standard deviation (or range) across folds.

In this work, we are interested in the following evaluation measures for classification problems: the TPR is the fraction of positive test observations predicted to be positive. *Sensitivity* is also the true-positive rate. *Specificity* is the true-negative rate, the fraction of negative test observations predicted to be negative. The FPR is the fraction of negative test observations predicted to be positive, and FPR is equal to one minus the specificity. The receiver operator characteristic (ROC) curve is a plot of true-positive rate for each possible value of the false-positive rate. The *area under the ROC curve* (AUC) is important because, if the true-positive rate is high for each value of the false-positive rate, the algorithm has a high AUC and is performing well. An AUC value of 0.5 would be obtained for random guessing; an AUC of one is perfect; and for most of the problems we consider here, an AUC value of 0.8 would be considered excellent. AUC is a useful evaluation measure, particularly when the positive and negative classes are imbalanced; that is, only a small fraction of the data are positive (or negative). For example, for the seizure prediction problem we discuss later, only 13.5% of observations in the seizure prediction data correspond to true seizures, and the rest were nonseizures.

For risk-score prediction, we are also interested in *calibration* (CAL), which is a measure of how closely the predicted positive rate from the model matches the empirical positive rate in the data. We discuss CAL later.

In general, we find that, when the form or size of the model is not constrained, then, for the majority of applications, AUC values for all machine-learning algorithms tend to be similar. AUCs start to differ when operational constraints are imposed. We see this in more depth in the sleep apnea and seizure examples that follow.

Applications and Insights

Both SLIM and RiskSLIM have had an impact on several applications in healthcare and criminal justice. Here, we discuss three applications and provide insights gained by producing interpretable models.

Sleep Apnea Screening

Obstructive sleep apnea (OSA) is a serious medical condition that can lead to morbidity and mortality and can severely affect quality of life. A major goal of every sleep clinic is to correctly screen patients for this disease. Testing for OSA is problematic. Preliminary screening is based mainly on patient-reported symptoms and scoring systems. Surprisingly, however, patient-reported symptoms are not particularly reliable as reported, nor are they very useful for determining whether a patient has OSA. In particular, to screen for OSA, doctors often use the Epworth sleepiness scale (Johns 1991) or other scoring systems that are based on typical reported OSA symptoms, such as snoring, nocturnal gasping, witnessed apneas, sleepiness, and other daytime complaints. Each of these predictive factors alone is weak; the comorbidities provided in medical records are much stronger. Hypertension, for example, is a good predictor of OSA. Thus, it is reasonable that the staff of the Massachusetts General Hospital hypothesized that an accurate scoring system could be created using information from only routinely available medical records—without reported symptoms—that could be just as accurate as the widely used scoring systems.

The data provided for this study were records from all patients at the Massachusetts General Hospital sleep lab who were over 18 years old and who underwent a definitive test for OSA called *polysomnography* (1,922 patients) between 2009 and 2013. Polysomnography is an expensive test for OSA; patients stay at a hospital overnight to allow medical technicians to collect information about their brain activity, blood oxygen levels, heart rate, breathing patterns, eye movements, and leg movements. Our goal was to predict OSA using only information that was available prior to the patient's polysomnography. Such information included standard medical information, such as gender, age, body mass index, past heart problems, hypertension, diabetes, and smoking as well as self-reported information on sleep patterns, such as caffeine consumption, insomnia, snoring, gasping, dry mouth in morning, leg jerks, and slowness in falling back to sleep. A full list of the features is provided in table 1 of Ustun et al. (2016).

The domain experts also required several operational constraints on the form of the model, such as constraints on the size of the model, and the signs of the coefficients. The domain experts considered these constraints vital to their trust in the model.

If a scoring system could be developed that accurately screens patients for sleep apnea, using only the patient's medical records, without using the patient-reported symptoms, it would create an actionable tool that could allow automatic screening (as opposed to manual screening, which involves a doctor). This type of automated scoring would allow wise usage of the limited resources available for direct patient encounters.

To summarize, our domain experts (Brandon Westover and Matt Bianchi at Massachusetts General Hospital) had two important goals: (1) create an accurate and transparent model for obstructive sleep apnea that obeyed operational constraints and (2) determine the value of the patient-reported symptoms (e.g., gasping, insomnia, caffeine consumption) as compared with information that is already in the patient's medical record.

Prior to our work, the best scoring system for sleep apnea screening was arguably the STOP-BANG score (Chung et al. 2008). STOP-BANG relies on eight features, including self-reported snoring, tiredness, and breathing problems, in addition to medical-record information. Its sensitivity is 83.6% and specificity is 56.4%, which precludes it from being used as a screening tool. The specificity is the percentage of negatives identified correctly; that is, the false-positive rate is 43.6% (100% − 56.4%), much higher than the FPR goal of 20% that our domain experts were seeking.

SLIM Model for Sleep Apnea Screening

One of the models that our collaboration produced has sensitivity 61.4% and specificity 79.1% so that the FPR was 20.9%. The scoring system was produced by SLIM and is in Figure 4.

Note that the model in Figure 4 does not contain patient-reported symptoms. After finding models like this, we wondered whether patient-reported symptoms were needed at all to achieve good prediction performance.

Patient-Reported Symptoms vs. Medical-Record Information

Using any machine-learning algorithm, it was easy to answer the second question of domain experts—that of

Figure 4. A SLIM Scoring System for Sleep Apnea Screening

1.	Age $\geq$ 60	4 points	···
2.	Hypertension	4 points	+ ···
3.	BMI $\geq$ 30	2 points	+ ···
4.	BMI $\geq$ 40	2 points	+ ···
5.	Female	-6 points	+ ···
		SCORE	= ···

PREDICT OBSTRUCTIVE SLEEP APNEA IF SCORE > 1

Notes. This model achieves a 10-fold cross-validation mean test TPR–FPR of 61.4%–20.9% and obeys all operational constraints. The model predicts OSA if the score exceeds one. There are no common prime factors because the threshold one is included in the set of factors; the coefficients are 1, 4, 4, 2, 2, −6, which are coprime. See Ustun et al. (2016) for more details.

measuring the importance of patient-reported symptoms. Patient-related symptoms are not nearly as important as medical-history information. Across every machine-learning method we tried, the models that used only patient-reported symptoms performed poorly, whereas models that used only medical-record information performed almost as well (and often as well) as the models that used both sets of information (see table S2 in the supplementary material of Ustun et al. 2016 for the AUC values of all machine-learning methods we tried). To illustrate this, Figure 5 shows the ROC curves for models built using all features (dashed curve), patient-reported symptoms only (lower solid curve), and features that were extracted from an electronic health record (gray curve, overlapping the dashed curve). This figure shows that performance does not degrade when omitting all the patient-reported variables.

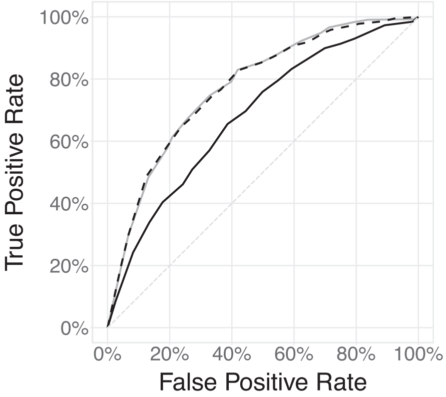
To summarize, SLIM was able to find a model using medical-record information only—without the patient-reported symptoms—with prediction quality that is essentially identical to the models that use both types of information.

An Insight from the Apnea Study: Operational Constraints Are Challenging for Nonmathematical-Programming-Based Machine-Learning Algorithms

The experiment for the sleep apnea project revealed severe shortcomings for nonmathematical-programming-based machine-learning methods, in that they are almost incapable of handling operational constraints.

Without considering operational constraints, our experiments indicated that SLIM models perform similarly to other machine-learning methods, such as support-vector machines with radial basis function kernels (Ustun et al. 2016). The differences between methods arise when operational constraints are considered.

Figure 5. Decision Points of SLIM Models over the Full ROC Curve



Notes. We show decision points of SLIM models for (1) all features (gray, overlapping with dashed curve), (2) features that can be extracted from an electronic health record (dashed), and (3) features related to patient-reported symptoms (black). See Ustun et al. (2016) for more details.

Our collaborators at Massachusetts General Hospital wanted a model fulfilling three simple operational constraints:

- *Max FPR*: Less than 20% false-positive rate. Our goal was to correctly detect as many cases of OSA as possible, limiting the falsely detected cases to 20%.
- *Model size*: Less than five terms in the model and small integer coefficients.
- *Sign constraints*: Some point values needed to be constrained to be either positive or negative. For example, it would not make sense to subtract points (i.e., predict lower risk of OSA) for patients who have hypertension than for those who do not because hypertension alone provides a significant risk for sleep apnea.

How would one obtain a model obeying these constraints with a standard machine-learning algorithm that does not use mathematical programming? As it turns out, this is not trivial. For standard methods, the only degrees of freedom given to the experimenter are parameters that govern the shape of the model. These parameters can be tuned until the constraints are obeyed; however, this proved to be challenging in practice. In particular, our results showed that for the standard machine-learning methods, even if we searched extensively through parameter values, we could rarely find feasible models (models that satisfy all constraints). Table 2 shows the number of parameter values we chose using a grid search, which is recorded in the "Total instances trained" column, and the parameter values we chose are in the "Values for free parameters" column. For example, we ran 975,000 instances of the standard machine-learning algorithm called "Elastic Net." Despite the large number of instances we trained, Table 2 indicates that the grid search rarely produced models that satisfied the constraints. The decision tree methods we tried (CART, C5.0 rules, C5.0 trees) had the worst problems; despite tuning, they were unable to produce any models with FPR < 20%. This can be seen in the column under "Percent of total instances satisfying" labeled "MaxFPR." Support vector machines (SVMs) with linear kernels were unable to produce models with simultaneously less than five terms and FPR < 20%, and ridge regression had the same problem. An SVM with radial basis function (RBF) kernels is nonparametric (i.e., it adapts dynamically to the data) and is highly nonlinear and, thus, not interpretable. The only algorithms that could be tuned to accommodate the constraints were Elastic Net, Lasso, and SLIM. For SLIM, the constraints are directly incorporated into the solver, and every solution it produces is feasible.

Table 2. Classification Methods Used for Sleep Apnea Screening Experiments

			Percentage of total instances satisfying		
Algorithm	Values for free parameters	Total instances trained	Maximum FPR, %	Maximum FPR and model size, %	Maximum FPR, model size, and signs, %
CART	39 values of $w^+ \in \{0.025, 0.05, \ldots, 0.975\}$	39	0.0	0.0	0.0
C5.0R	39 values of $w^+ \in \{0.025, 0.05, \ldots, 0.975\}$	39	0.0	0.0	0.0
C5.0T	39 values of $w^+ \in \{0.025, 0.05, \ldots, 0.975\}$	39	0.0	0.0	0.0
Lasso	39 values of $w^+ \in \{0.025, 0.05, \ldots, 0.975\}$ × 1,000 values of λ chosen by glmnet	39,000	19.6	4.8	4.8
Ridge	39 values of $w^+ \in \{0.025, 0.05, \ldots, 0.975\}$ × 1,000 values of α chosen by glmnet	39,000	20.9	0.0	0.0
Elastic Net	39 values of $w^+ \in \{0.025, 0.05, \ldots, 0.975\}$ × 1,000 values of λ chosen by glmnet × 19 values of $\alpha \in \{0.05, 0.10, \ldots, 0.95\}$	975,000	18.3	1.0	1.0
SVM linear	39 values of $w^+ \in \{0.025, 0.05, \ldots, 0.975\}$ × 25 values of $C \in \{10^{-3}, 10^{-2.75}, \ldots, 10^3\}$	975	18.7	0.0	0.0
SVM RBF	39 values of $w^+ \in \{0.025, 0.05, \ldots, 0.975\}$ × 25 values of $C \in \{10^{-3}, 10^{-2.75}, \ldots, 10^3\}$	975	15.8	0.0	0.0
SLIM	$w^+ = n^-/(1+n^-)$, $C_0 = 0.9w^-/nd$, $\lambda_0 \in \{-100, \ldots, 100\}$, $\lambda_j \in \{-10, \ldots, 10\}$	1	100.0	100.0	100.0

Notes. For each algorithm, we show the parameter settings, total number of instances trained, and the percentage of instances that fulfilled various combinations of operational constraints. Each instance is a unique combination of free parameters for a given method. The w^+ parameter is a unit misclassification cost for positive points. See Ustun et al. (2016) and Ustun and Rudin (2016b) for more details.

Of the feasible models found from the standard machine-learning methods, almost none are accurate predictive models. Figure 6 shows how Elastic Net, Lasso, and SLIM perform as we vary the model size. Here, both Lasso and Elastic Net would need eight variables to attain the accuracy of the five-variable SLIM model.

What we have illustrated is a serious concern regarding the use of machine-learning methods for practical problems: in almost all machine-learning algorithms, user-defined constraints are not accommodated. Mathematical programming tools solve this issue.

Our work on sleep apnea was published in the *Journal of Clinical Sleep Medicine* (Ustun et al. 2016), which is the official journal of the American Academy of Sleep Medicine. More details can be found in SLIM's paper (Ustun and Rudin 2016b).

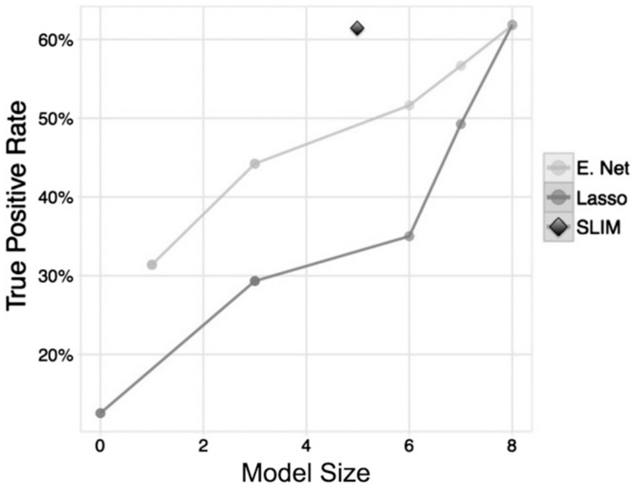
Seizure Prediction in the ICU

Patients in the intensive care unit of a hospital who may be at risk for dangerous seizures are monitored using continuous electroencephalography (cEEG), with which electrodes monitor electrical signals in the brain. A clinician monitors the patient and identifies features in the cEEG signal that may be predictive of seizure. The clinician may determine that the patient requires a potentially dangerous intervention to prevent seizures or (expensive) continued monitoring. Rather than having clinicians manually estimate seizure risk from cEEG signals, Massachusetts General Hospital staff aimed to assist clinicians by estimating this risk in a transparent way. We worked with a data set from the Critical Care EEG Monitoring Research Consortium, collected at several hospitals (e.g., Emory University Hospital, Brigham and Women's Hospital, and Yale University Hospital) over the course of three years. The database contains 5,427 cEEG recordings with 87 variables, and each patient had at least six hours of uninterrupted cEEG monitoring. The variables from the cEEGs included important pattern types: lateralized periodic discharges (LPDs), lateralized rhythmic delta (LRDA), generalized periodic discharges (GPDs), generalized rhythmic delta (GRDA), and bilateral periodic discharges (BiPDs). Additionally, we had medical history and secondary symptoms for each patient. The outcome we aimed to predict was whether the patient would have a seizure within 24 hours. A transparent automated tool to help with seizure-risk prediction would be particularly helpful in preventing false negatives: situations in which clinicians mistakenly label the patient as not being at risk.

RiskSLIM Model for Seizure Prediction

In Figure 7, we show a model that we built using RiskSLIM. This model has a mean AUC over five cross-validation folds of 0.819 (with a range of 0.776–0.849 over the five folds). It is similar to other medical scoring systems in that it can be memorized by its acronym: the "2H" stands for "GRDAs, LRDAs, BiPDs, LPDs, or GPDs with a frequency > 2 Hz" (one point); "E" stands for epileptiform discharges (one point); "L" stands for LPD or LRDA or BiPD (one point); "P" stands for GRDAs, LRDAs, BiPDs, LPDs, or GPDs with plus

Figure 6. Sensitivity and Model Size of Lasso and Elastic Net Models that Satisfy the Sign and FPR Constraints



Notes. For each method, we plot the instance that attains the highest 10-fold cross-validation mean test TPR at model sizes between zero and eight. Lasso and Elastic Net need at least eight coefficients to produce a model with the same sensitivity as SLIM. See Ustun and Rudin (2016b) for details.

features (superimposed rhythmic, fast, or sharp activity) (one point); "S" is any history of seizures (one point); and "2B" is brief potentially ictal rhythmic discharges (two points).

The 2HELPS2B score has no predecessors. It is the first scoring system to be developed for cEEG monitoring for seizure prediction and can be directly integrated into clinical workflow.

More details are in the neurology paper (Struck et al. 2017) and the RiskSLIM methodology papers (Ustun and Rudin 2016a, 2017).

Calibration was an important concern for our collaborators—models were deemed unacceptable if they were poorly calibrated. While constructing the 2HELPS2B score, it became apparent that the typical methods one might use to construct scoring systems had systematic problems with calibration. This is our second insight, which we now discuss.

Figure 7. The 2HELPS2B Scoring System Produced by RiskSLIM

1.	Any cEEG Pattern with Frequency **2 H**z	1 point	···
2.	**E**pileptiform Discharges	1 point	+ ···
3.	Patterns include [**L**PD, LRDA, BIPD]	1 point	+ ···
4.	**P**atterns Superimposed with Fast or Sharp Activity	1 point	+ ···
5.	Prior **S**eizure	1 point	+ ···
6.	**B**rief Rhythmic Discharges	**2** points	+ ···
		SCORE	= ···

SCORE	0	1	2	3	4	5	6+
RISK	<5%	11.9%	26.9%	50.0%	73.1%	88.1%	95.3%

Source. Struck et al. (2017).

An Insight from the Seizure Study: Risk Calibration Suffers When We Use Rounding to Compute Risk Scores

Risk calibration (CAL) measures how closely the estimated risks from the model match risks in the data. Risk calibration is essential for practical use in risk-scoring applications (e.g., Shah et al. 2018).

Let us define CAL precisely. The estimated risks for each individual i are calculated using the scoring system (e.g., from 2HELPS2B), and the risk for patient i from the model is denoted by p_i. Separately, for each possible value of the score s, we estimate the probability of the outcome $y = 1$ given s from the data; that is, $p(s) = P(y = 1|s)$. Then we compute the Euclidean distance between p_i and $p(s_i)$ across all patients i; this is precisely CAL. A calibration plot is a plot of $p(s_i)$ versus p_i. If the plot is a diagonal line, the model is nicely calibrated.

RiskSLIM minimizes the logistic loss that is used for logistic regression. Logistic regression produces risk-calibrated models (Zadrozny and Elkan 2002, Caruana and Niculescu-Mizil 2004), but when rounding or other postprocessing steps are done to a logistic regression model, it can drastically alter calibration. As we discussed earlier, rounding sends all small coefficients to zero (which eliminates part of the signal), and rounding coefficients upward makes variables more important than they should be in a calibrated model. An extensive set of experiments in Ustun and Rudin (2016a, 2017) considered several types of rounding techniques. In particular, it considered naïve rounding (denoted RD), which simply rounds coefficients to the nearest integer within the range $\{-5, -4, .., 0, \ldots, 4, 5\}$, and rescaled rounding (denoted RSRD), which scales all coefficients so that the largest one is ± 5 and then rounds to the nearest integer. Rescaled rounding tends to mitigate the problem of too many coefficients being rounded to zero.

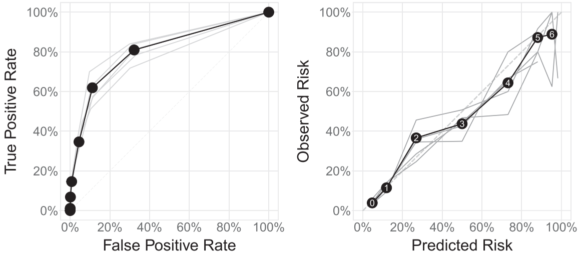
Calibration curves should always go upward: as the score increases, the risk should always increase. However, this does not hold for either RD or RSRD. Our collaborators determined that this was problematic because it is unreasonable that (for example) a patient with a score of three has a higher risk of seizure than a patient with a score of four. Figure 8 shows results from a controlled cross-validation experiment, including ROC curves and calibration curves for RiskSLIM and also for the RD and RSRD methods. The black curves in the figures are from a model computed across the five cross-validation folds, and models in gray are from each of the five folds. The problems with calibration are apparent: the curves simply do not always increase. Here, RiskSLIM's fivefold mean CAL was 2.5% (the best is 0%), whereas RD's was 5.3% and RSRD's was 12.2%. 2HELPS2B was determined separately from the controlled experiment; its ROC and calibration curves are shown in Figure 9. It has a mean CAL over the five folds of 2.7%.

Figure 8. Risk Scores, ROC Curves, and Reliability Diagrams for RiskSLIM and Heuristic Rounding Techniques

Optimized Risk Score (RiskSLIM)

1.	AnyBriefRhythmicDischarge	2 points	· · ·
2.	PatternsInclude LPD	2 points	+ · · ·
3.	AnyPriorSeizure	1 point	+ · · ·
4.	EpiDischarge	1 point	+
		SCORE	=

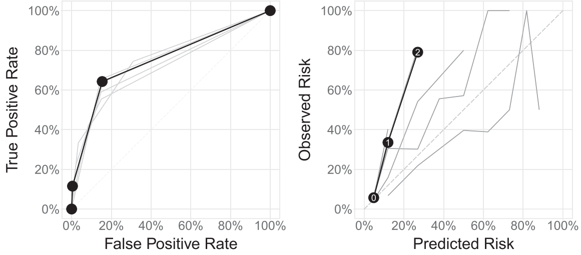
SCORE	0	1	2	3	4	5	6
RISK	4.7%	11.9%	26.9%	50.0%	73.1%	88.1%	95.3%



$\ell_1 + \ell_2$ Penalized Logistic Regression + Rounding

1.	AnyBriefRhythmicDischarge	1 point	· · ·
2.	PatternsIncludeBiPD or LRDA or LPD	1 point	+
		SCORE	=

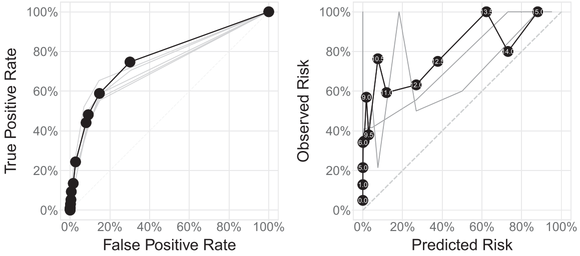
SCORE	0	1	2
RISK	4.7%	11.9%	26.9%



$\ell_1 + \ell_2$ Penalized Logistic Regression + Scaling + Rounding

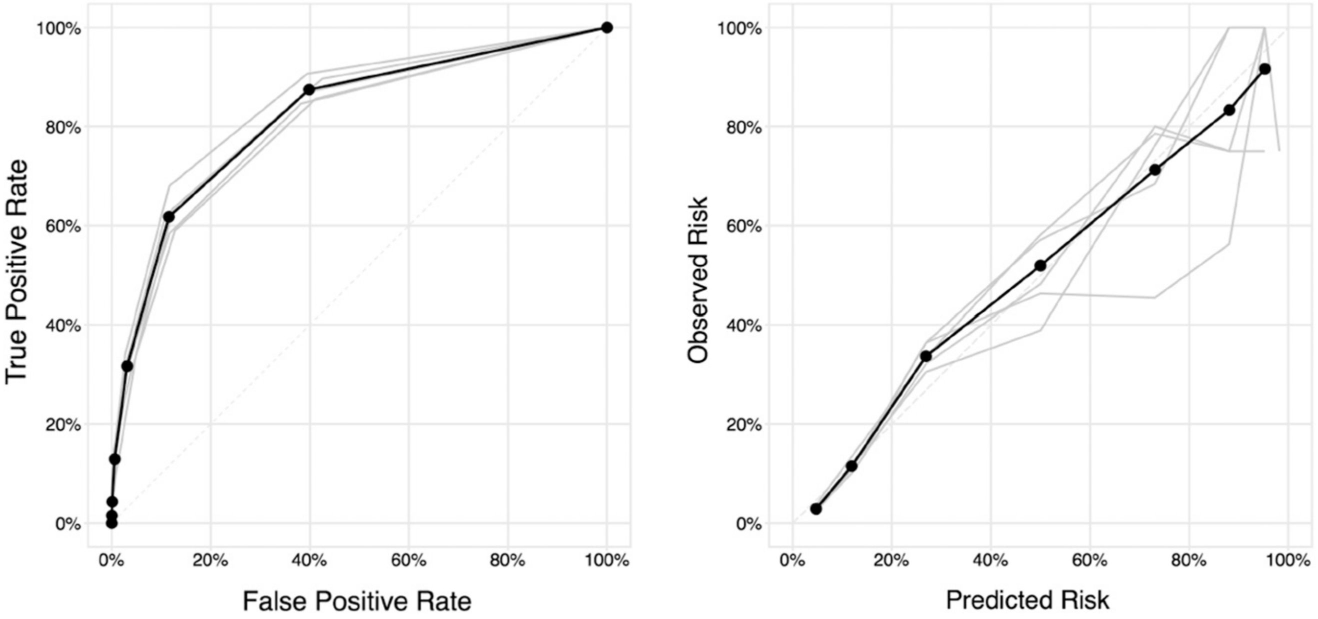
1.	PatternsIncludeBiPD or LRDA or LPD	5 points	· · ·
2.	MaxFreqFactorLPD	3 points	+ · · ·
3.	AnyPriorSeizure	1 point	+
		SCORE	=

SCORE	0.0 to 9.5	10.5	11	12	12.5	13.5	14	15
RISK	< 5.0%	7.6%	11.9%	26.9%	37.8%	62.2%	73.1%	88.1%



Source. Ustun and Rudin (2016a).
Note. We show the final model on training data in black and fold-based models on test data in gray.

Figure 9. ROC Curves and Calibration Curves for the 2HELPS2B Score Produced by RiskSLIM



Source. Struck et al. (2017).

These experiments with rounding are not surprising; when we move in an arbitrary direction in a high-dimensional space, we know from integer programming textbooks (Wolsey 1998) that we will find problems with solution quality. Further, by using rounding, all guarantees of optimality are lost. This becomes problematic for applications such as recidivism prediction as we discuss next.

Recidivism Prediction

In the United States, criminal sentencing is done according to a mandated federal guideline (e.g., the Criminal History Category; U.S. Sentencing Commission 2004). One of the latest public guidelines for recidivism risk prediction in the United States is the Pennsylvania Commission on Sentencing (2012), and other methods are used in Canada (Hanson and Thornton 2003), the Netherlands (Tollenaar and van der Heijden 2013), and the United Kingdom (Howard et al. 2009). There are a very large number of different risk scores for various applications, including sentencing, parole, and prison administration (see Zeng et al. 2017 for a longer list). These scores can be helpful: it is possible for a data-driven calculation to mitigate irregularities in decisions made by people. No human can keep a database in its head and accurately calculate recidivism risks. The decision-making process of judges can have high variance and rely on arbitrary factors. For example, there is (debated) evidence that judges are much less likely to make a favorable ruling immediately before a lunch break (Danziger et al. 2011, Kahneman 2013). A larger problem is that judges are not generally provided with feedback on the quality of their recidivism predictions; that is, they cannot learn from past mistakes.

Over the past few years, there has been an ongoing debate in the statistical community of criminologists. Some of them have claimed that traditional statistical methods are as accurate for predicting recidivism as modern machine-learning tools when the proper preprocessing has been done to create features (e.g., Berk and Bleich 2013, Bushway 2013, Tollenaar and van der Heijden 2013). As we have shown, however, traditional statistical tools have serious flaws when paired with rounding methods in terms of risk calibration and an inability to incorporate operational constraints.

As this debate is happening, companies such as Northpointe (now called Equivant) are selling predictions to the U.S. government, which uses them widely. These risk scores have the potential to be racially biased as argued by ProPublica (Angwin et al. 2016) although determining whether they are actually biased is difficult (see Fisher et al. 2018, who debate this using variable importance arguments). In 2016, in the case State v. Loomis, the Wisconsin Supreme Court ruled that judges can use black-box risk scores such as Northpointe's COMPAS but minimized the role that such scoring systems could play as evidence. An appeal was filed at the U.S. Supreme Court; however, the Court declined to hear the case in June 2017.

The goal of our project was to determine whether such black-box scoring systems were needed at all for recidivism prediction. If we could find a transparent model with the same accuracy as the best black-box model, we would no longer require the black-box model.

We used the largest publicly available data set on recidivism, which is the "Recidivism of Prisoners Released in 1994" data set collected by the U.S. Department of Justice, Bureau of Justice Statistics (U.S. Department of Justice, Bureau of Justice Statistics 2014). This data set contains information that we used from 33,796 prisoners, including criminal history from record-of-arrest-and-prosecution sheets along with demographic factors, such as gender and age. We omitted socioeconomic factors, such as race, for the main study but conducted experiments using race afterward (see Zeng et al. 2017). The outcomes we aimed to predict within three years of release were (1) arrest for any crime, (2) arrest for drug-related crime, (3) arrest for violent crime, (4) arrest for domestic violence crime, (5) arrest for sexual violence crime, and (6) arrest for crime involving fatal violence.

Results for Recidivism Prediction

Our results were consistent with those from other applications in that most machine-learning algorithms performed almost identically across the full ROC curve for all of the six prediction problems as shown in Figure 10. The decision tree methods (CART, C5.0T, C5.0R; Figure 10) sometimes performed poorly, particularly for imbalanced problems. This could potentially illustrate the reason why people often believe that an interpretable modeling algorithm does not perform as accurately as a black-box method; methods such as Cart that produce interpretable models are indeed not as accurate as other methods. CART (Breiman et al. 1984) is not based on optimization and was designed to operate within the limits of older computers (e.g., from 1984). CART's poor performance is not a convincing reason for why all interpretable modeling methods might perform poorly.

Figure 2 shows two of the models we produced using SLIM and RiskSLIM.

The basic findings (that interpretable models are as accurate as black-box models) were confirmed by Angelino et al. (2017, 2018) using another publicly available data set, that is, ProPublica Broward County data; in this case, small logical models were shown to be as accurate as the COMPAS score for predicting two-year recidivism.

Insight for Recidivism Prediction: Importance of Certifiable Optimality

Methods such as SLIM and RiskSLIM produce certificates of optimality, or they provide distance to optimality (optimality gaps) in the case in which the problems are not fully solved to optimality. These types of guarantees are useful for answering questions such as "Does there exist an interpretable model (of a given form) that achieves a particular value for predictive performance on the data set?"

Although it is true that optimizing performance on the training set does not correspond exactly to performance on the test set, training and test performances are guaranteed to be similar by statistical learning theory. If a method cannot achieve high-quality in-sample performance, it is difficult for it to achieve high-quality out-of-sample performance.

This work provides tools that can determine whether an interpretable model exists that performs well on a given data set. If an accurate, interpretable model does exist (which it does in many cases), we should use it rather than resorting to a black box, particularly for high-stakes decisions, such as bail, parole, and sentencing.

Other Applications

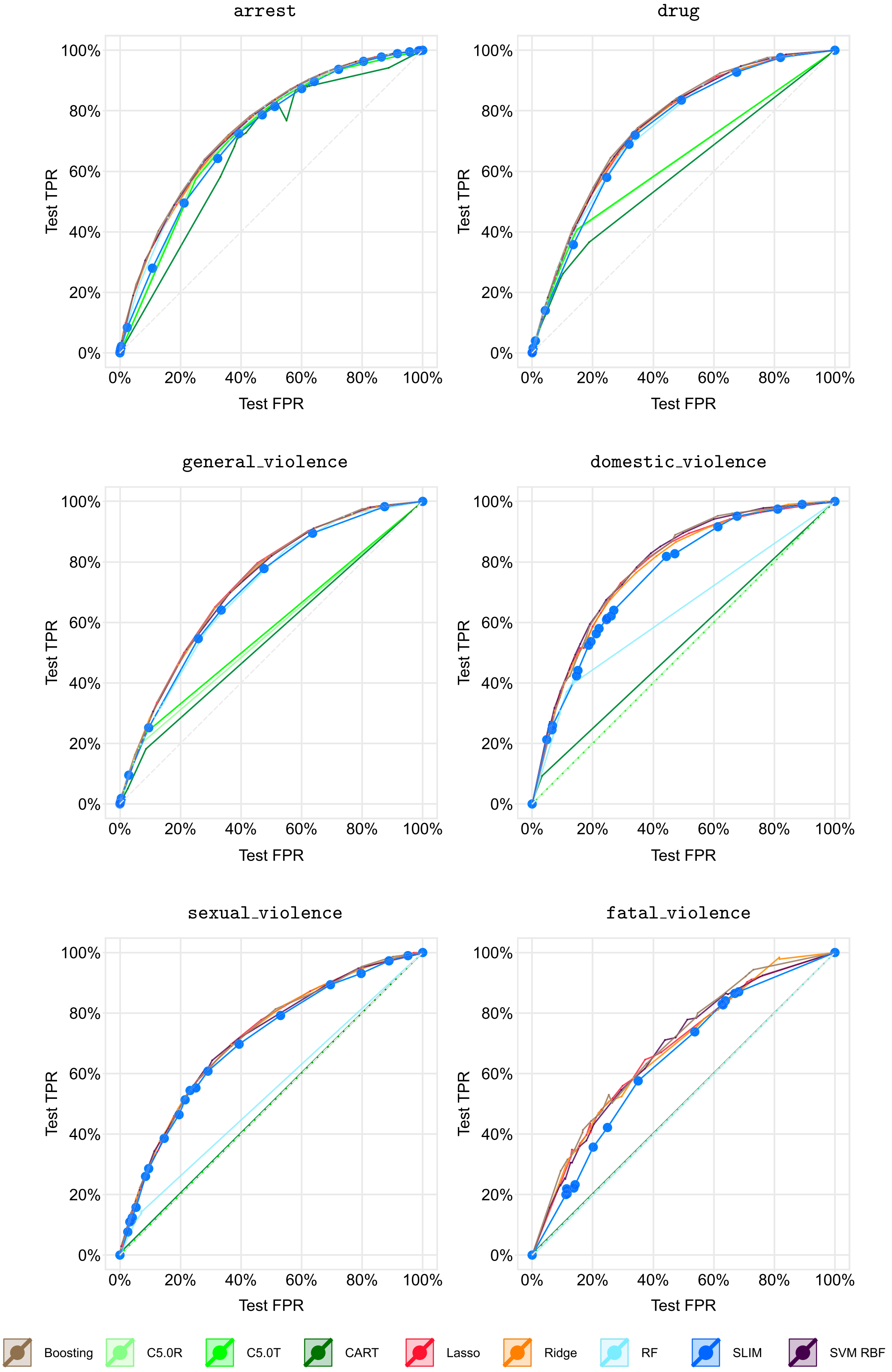
SLIM and RiskSLIM have been used for purposes besides those discussed herein. SLIM has been used to detect cognitive impairment, such as Alzheimer's disease, dementia, and Parkinson's disease. In particular, the clock-drawing test, which is a pen-and-paper test that has been used for a century to diagnose these disorders, has been updated to be digitized. Patients draw clocks with a digital pen, and this digitized test is automatically scored with a SLIM-based system. The new scoring system far surpasses the accuracy of all previously published scoring systems for the clock-drawing test and is a promising noninvasive technique for early identification of cognitive impairment. Our work on this project, in conjunction with several collaborators (Souillard-Mandar et al. 2016), won the 2016 INFORMS Innovative Applications in Analytics Award.

In a separate project using RiskSLIM, we created a screening scale for adult attention deficit hyperactivity disorder (ADHD) in collaboration with a team of psychiatrists. The test allows for a quick, risk-calibrated diagnosis based on the answers to six questions on a self-reported questionnaire. The questions include "How often do you have difficulty concentrating on what people say to you even when they are speaking to you directly?" and "How often do you leave your seat in meetings and other situations in which you are expected to remain seated?" The prediction performance was optimized based on clinical diagnoses using *Diagnostic and Statistical Manual of Mental Disorders* (DSM-5) criteria (American Psychiatric Association 2013), which is the new standard for adult ADHD diagnosis. According to the work's publisher, JAMA Psychiatry (Ustun et al. 2017), it has accumulated 13,840 views within its first year of publication (May 2017 through May 2018).

SLIM and RiskSLIM are optimization-based approaches. Ertekin and Rudin (2015) provide a Bayesian approach to forming scoring systems.

It is important to note that scoring systems are not the only forms of interpretable models. Logical models, such as decision trees and decision lists, have existed

Figure 10. (Color online) ROC Curves for Recidivism Prediction Problems



Source. Zeng et al. (2017).
Notes. TPR–FPR for SLIM models are plotted using large dots. All models perform similarly except those from C5.0R, C5.0T, and CART.

since the beginning of our use of artificial intelligence. Recent work on those models has been useful for recidivism prediction (Angelino et al. 2017, Lakkaraju and Rudin 2017, Yang et al. 2017), credit scoring (Chen and Rudin 2018), hospital readmission (Wang and Rudin 2015), and stroke prediction in atrial fibrillation patients (Letham et al. 2015). Logical models (in particular "or's of and's") are useful for modeling consideration sets used in marketing (Goh and Rudin 2014; Wang et al. 2016, 2017). Logical models with operational constraints can also be constructed with specialized optimization techniques (e.g., varieties of monotonicity constraints—see Wang and Rudin 2015 and Chen and Rudin 2018). For computer vision and other domains in which modeling in terms individual features are not meaningful, defining interpretability is more challenging (e.g., Li et al. 2018).

Comments from Practitioners

Although further research is needed to adopt any of our (very recent) models into routine medical practice, medical practitioners have already found the methods useful and have developed further avenues to investigate medical practice based on the models discovered by SLIM and RiskSLIM. In this section, we include comments from some of these practitioners about our methods and models, and we let them speak for themselves by quoting their words.

Sleep Apnea

Combs et al. (2016, p. 160) wrote a commentary, "Big-Data or Slim-Data: Predictive Analytics Will Rule with World," about our article on sleep apnea screening (Ustun et al. 2016). In their article, they note, "Potentially, the tool described by Ustun et al. could be integrated into the electronic medical record, flagging high-risk patients and prompting physicians to further screen for OSA. These high-risk patients could then be referred for diagnostic testing for OSA. The approach to use a screening tool that is not dependent upon patient-reported sleep symptoms sidesteps the barriers for detection of OSA in the busy clinic setting. While management of OSA by sleep-certified physicians may confer an advantage over providers with no prior experience in managing patients with OSA, such automated electronic medical record based systems could assist with case-finding and conceivably be comparable between providers who are not experienced, nor received training, in managing patients with as yet undiagnosed OSA versus those managed by sleep-certified physicians. Ustun and colleagues should be commended for bringing both big and slim data to our doorsteps."

ICU Seizure Risk

Dr. Brandon Westover, a collaborator on the faculties at Massachusetts General Hospital and Harvard Medical school, has written about our joint work in a letter of support for the Wagner Prize: "The model developed using RiskSLIM not only produced reliable risk estimates, but also obeyed an extensive set of constraints designed to ensure that it would be used and accepted by physicians who interpret cEEGs. The ability to produce calibrated risk estimates under these constraints is especially remarkable, as our early attempts with using current machine learning methods produced tools that either had poor risk calibration or were unusable (uninterpretable by physicians) in this setting. …The model that we built will help physicians quickly assess seizure risk and improve the way that we make crucial decisions in the ICU. I expect that it will help reduce costs associated with cEEG monitoring, and ultimately lead to better health outcomes for critically ill patients."

An editorial by Czeisler and Claassen (2017) about our work (Struck et al. 2017) states, "The authors should be commended for their robust evaluation of a complex issue. Many attempts at logistic regression models for such complex phenomena yield models that are mathematically useful but too complex for routine clinical use. By using the risk-calibrated supersparse linear integer model process for development, the authors were able to design a simple scale with good accuracy, which can be easily used by clinicians to estimate seizure risk in their patients."

Later on in their editorial, Czeisler and Claassen (2017, p. 1396) begin to ask questions of exactly the kind that would allow them to fine-tune models using SLIM and RiskSLIM. For example, in their discussion "What do these findings mean for clinical utility?" they consider whether one could create a score that would provide fine-grained risk scores for low-risk patients; this is something that SLIM and RiskSLIM can potentially construct given the right patient data. They also state, "Another potential use for this score would be for guidance as to what patterns to treat more aggressively than others. For example, patients with a high 2HELPS2B score might benefit from additional antiepileptic medication, while those with low scores might not. We do not know the answer to these questions yet, but appropriate risk stratification using 2HELPS2B may allow us to answer these questions adequately in the near future."

ADHD Screening

Ronald Kessler, a collaborator of ours who developed the main prior scoring system for ADHD, wrote about his experience with RiskSLIM in a support letter for the Wagner prize: "A standard screening scale for ADHD developed by my team has been used throughout the world for over a decade to screen for adult ADHD in primary care samples, workplace health-risk appraisal surveys, and specialty mental health intake assessments. This scale is longer than desired, though, and has a lower sensitivity than we would want to be the case.

Based on these problems, we recently used the Risk-SLIM technique developed by Rudin and Ustun to see if we could improve on our original screening scale in a reanalysis of our benchmark validity studies. The results were stunning. Cross-validated operating characteristics were improved dramatically with a short (six-item) scale that had excellent out of sample performance when applied to external validation data. The algorithm was very easy to use: The results were intuitive. And the ease of scoring makes implementation simple." Kessler also writes, "The optimization-based techniques used in RiskSLIM fill a crucial methodological gap in the emerging field of precision medicine. Although patient self-report screening scales are widely used, the methods used to create these scales vary dramatically across studies and are all suboptimal compared with RiskSLIM. I'm not exaggerating when I say that this algorithm is soon going to become the mainstay of scale development in the patient-reported outcome research area. RiskSLIM makes it far easier for us to build these scales than in the past and the scales will perform much better than in the past. The ease with which we were able to develop the ASRS has attracted significant interest in our community, with a number of researchers from around the world contacting me with interest in using it to build scales for other constructs."

Shaw et al. (2017, p. 527) wrote an editorial about our work (Ustun et al. 2017), stating, "To our knowledge, the article by Ustun et al. presents the first screening scale using DSM-5 criteria. …These fascinating findings will not only stimulate further research but could also result in less insistence on a childhood history of symptoms, perhaps even further increasing diagnostic rates. In short, as public awareness of adult ADHD increases, so too does the need for psychometrically robust screening tools both for research and to help identify those most likely to benefit from further expert assessment and treatment."

Indeed, this need for robust and interpretable screening and diagnosis tools extends to many areas of healthcare.

Looking Forward

Within the foreseeable future, there will be a business need to keep the details of machine-learning models as a trade secret. This may not be problematic in some domains, particularly when decisions have a minor effect on the lives of people. In other domains, such as healthcare and criminal justice, decisions are serious, and actions need to be defensible. The machine-learning algorithms presented here represent a fundamental change to the way transparent models are constructed, leveraging modern discrete-optimization techniques (cutting planes, data-reduction bounds, mixed-integer programming) and capabilities (callback functions, modern solvers). Code for SLIM and RiskSLIM is publicly available, at http://github.com/ustunb/slim-python and http://github.com/ustunb/risk-slim.

Appendix. Optimization Problems

We start with a data set of N independent and identically distributed training examples $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$, where $\boldsymbol{x}_i \in \mathcal{X} \subseteq \mathbb{R}^{d+1}$ denotes a vector of features $[1, x_{i,1}, \ldots, x_{i,d}]^\top$ and $y_i \in \mathcal{Y} = \{-1, 1\}$ denotes a class label. We consider linear classification models of the form $\hat{y} = \text{sign}(\langle \boldsymbol{\lambda}, \boldsymbol{x} \rangle)$, where $\boldsymbol{\lambda} = [\lambda_0, \lambda_1, \ldots, \lambda_d]^\top$ represents a vector of coefficients and λ_0 represents an intercept.

In this setup, the coefficient vector $\boldsymbol{\lambda}$ determines all parameters of a scoring system. In particular, the coefficient λ_j represents the *points* for feature j for $j = 1, \ldots, d$. Given an example with features $\boldsymbol{x}_i$, users first tally the points for all features such that $\lambda_j \neq 0$ to obtain a total *score* $\sum_{j=1}^d \lambda_j x_{i,j}$ then use the total score to obtain a predicted label (i.e., for decision making) or an estimate of predicted risk (i.e., for risk assessment).

SLIM's Optimization Framework for Decision Making

In decision-making applications, we use the score to output a predicted label $\hat{y} \in \{-1, 1\}$ through a decision rule of the form

$$\hat{y}_i = \begin{cases} +1 & \text{if } \displaystyle\sum_{j=1}^{d} \lambda_j x_{i,j} + \lambda_0 > 0, \\ -1 & \text{if } \displaystyle\sum_{j=1}^{d} \lambda_j x_{i,j} + \lambda_0 \leq 0. \end{cases} \tag{A.1}$$

In this setting, we learn the values of coefficients by solving a discrete-optimization problem that we refer to as the *decision-rule problem*. The optimal solution to the decision-rule problem is a supersparse linear integer model. The decision-rule problem is a discrete-optimization problem of the form

$$\begin{aligned} \min_{\boldsymbol{\lambda}} \quad & l_{01}(\boldsymbol{\lambda}) + C_0 \|\boldsymbol{\lambda}\|_0 \\ \text{s.t.} \quad & \boldsymbol{\lambda} \in \mathcal{L}, \\ & \gcd(\boldsymbol{\lambda}) = 1, \end{aligned} \tag{A.2}$$

where

- $l_{01}(\boldsymbol{\lambda}) = \frac{1}{N}\sum_{i=1}^N \mathbf{1}[\hat{y}_i \neq y_i]$ is the fraction of misclassified observations (notation $\mathbf{1}$[statement] evaluates to one if the statement is true and zero otherwise);
- $\|\boldsymbol{\lambda}\|_0 = \sum_{j=1}^d \mathbf{1}[\lambda_j \neq 0]$ is the count of nonzero coefficients, ℓ_0-seminorm;
- $\mathcal{L} \subset \mathbb{Z}^{d+1}$ is a finite user-provided set of feasible coefficient vectors, usually chosen to be small integers, $\mathcal{L} = \{-10, -9, \ldots, 9, 10\}^{d+1}$;
- $C_0 > 0$ is a user-chosen trade-off parameter to balance accuracy and sparsity (if we choose C_0 as 0.01, we make a 1% sacrifice in training accuracy to reduce the model by one term in size);
- $\gcd(\boldsymbol{\lambda}) = 1$ is a symmetry-breaking constraint to ensure coefficients are coprime. Here "gcd" stands for greatest common divisor.

Here, the objective minimizes the empirical probability of misclassification and penalizes the number of nonzero terms to encourage the model to be sparse. The feasible region can be customized to include additional operational constraints (see Table 1).

To implement the decision-rule problem as a mathematical program, there is a simple trick for encoding the constraint that the gcd of the coefficients is one. In particular, if we add a term to the objective that is the sum of the absolute coefficients multiplied by a very small number (ϵ in the formulation that follows), it forces the gcd to be one without influencing either accuracy or sparsity. The reason this trick works is because the loss and sparsity terms take on only discrete values. Among all models that are equally accurate and equally sparse, the formulation will choose the one with the smallest absolute sum of terms, $\sum_j |\lambda_j|$, also written $\|\boldsymbol{\lambda}\|_1$. Because the values of the λ_j are also integers, they must be coprime.

In practice, the fraction of misclassifications in the objective is replaced with a weighted sum of false positives and false negatives for applications with which the user has determined that one of these is more important to reduce than the other.

Incorporating the separate weights for false positives and false negatives (w^- and w^+) and using the additional term in the objective to force the gcd to one, the optimization problem is as follows:

$$\begin{aligned}\min_{\boldsymbol{\lambda}} \quad & \frac{w^+}{N_+}\sum_{i:\,y_i=1}\mathbf{1}[\hat{y}_i \neq 1] + \frac{w^-}{N_-}\sum_{i:\,y_i=-1}\mathbf{1}[\hat{y}_i \neq -1] \\ & + C_0\|\boldsymbol{\lambda}\|_0 + \epsilon\|\boldsymbol{\lambda}\|_1 \\ \text{s.t.} \quad & \boldsymbol{\lambda} \in \mathcal{L},\end{aligned}$$

where $\hat{y}$ depends on $\boldsymbol{\lambda}$ through Equation (A.1) and N_+ and N_- are the number of positive observations and negative observations, respectively. The value ϵ needs to be sufficiently small that the gcd term only makes the coefficients in $\boldsymbol{\lambda}$ coprime and does not effect the solution in any other way.

The relative importance of false positives and false negatives, w^+ and w^-, should generally be chosen by the user, depending on how much a false positive is worth relative to a false negative in the application. Often, we try many possible values of w^+ and w^- to create several models that are optimized for specific points on the ROC curve. If w^+ is two and w^- is one, it means that each positive is worth twice that of a negative in this calculation.

This optimization problem is amenable to mixed-integer linear programming, which we discuss in depth in Ustun and Rudin (2016b).

We have finished discussing the optimization problem solved by SLIM; now we move on to RiskSLIM.

RiskSLIM's Optimization Framework for Risk Assessment

In risk assessment applications, we use the score to estimate predicted risk. Specifically, we estimate the predicted risk that example *i* belongs to the positive class using the logistic link function as

$$\Pr(y_i = +1 \mid \boldsymbol{x}_i) = \frac{1}{1+\exp(-\boldsymbol{\lambda}^T\boldsymbol{x}_i)}.$$

We learn the values of the coefficients from data by solving the following mixed integer nonlinear program (MINLP), which we refer to as the *risk score problem* or RiskSlimMINLP:

$$\begin{aligned}\min_{\boldsymbol{\lambda}} \quad & l(\boldsymbol{\lambda}) + C_0\|\boldsymbol{\lambda}\|_0 \\ \text{s.t.} \quad & \boldsymbol{\lambda} \in \mathcal{L},\end{aligned} \qquad \text{(A.3)}$$

where

- $l(\boldsymbol{\lambda}) = \frac{1}{N}\sum_{i=1}^N \log(1+\exp(-\boldsymbol{\lambda}^T y_i \boldsymbol{x}_i))$ is the logistic loss function;
- $\|\boldsymbol{\lambda}\|_0 = \sum_{j=1}^d \mathbf{1}[\lambda_j \neq 0]$ is the ℓ_0-seminorm;
- $\mathcal{L} \subset \mathbb{Z}^{d+1}$ is a set of feasible coefficient vectors (user-provided);
- $C_0 > 0$ is a trade-off parameter to balance fit and sparsity (user provided).

The optimal solution to the risk score problem is a scoring system that we refer to as a risk-calibrated supersparse linear integer model.

Here, the objective minimizes the logistic loss from logistic regression to achieve high values of the AUC and to achieve risk calibration. The objective penalizes the ℓ_0-seminorm for sparsity. The trade-off parameter C_0 controls the balance between these competing objectives and represents the maximum log-likelihood that is sacrificed to remove a feature from the optimal model. The feasible region restricts coefficients to a small set of bounded integers, such as $\mathcal{L} = \{-10, \dots, 10\}^{d+1}$, and may be further customized to include operational constraints, such as those in Table 1.

To fit a RiskSLIM scoring system, we need to solve the MINLP. This MINLP is difficult to solve using any commercial solver. Cutting-plane algorithms are a natural choice for this problem because the objective is continuous and convex, but we were not able to use a traditional cutting-plane algorithm because of the discrete domain of the optimization problem. Instead, we designed a specialized cutting-plane technique that creates a series of branches, and we compute cutting planes on each branch. This allows us to solve very large problems and parallelize easily. This algorithm is called the "lattice cutting plane method"; more details can be found in the work of Ustun and Rudin (2016a).

References

American Psychiatric Association (2013) *Diagnostic and Statistical Manual of Mental Disorders (DSM-5)* (American Psychiatric Association Publishing, Washington, DC).

Angelino E, Larus-Stone N, Alabi D, Seltzer M, Rudin C (2017) Learning certifiably optimal rule lists for categorical data. *Proc. 23rd ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining* (ACM, New York), 35–44.

Angelino E, Larus-Stone N, Alabi D, Seltzer M, Rudin C (2018) Certifiably optimal rule lists for categorical data. *J. Machine Learn. Res.* 18:1–78.

Angwin J, Larson J, Mattu S, Kirchner L (2016) Machine bias. Accessed January 1, 2018, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Antman EM, Cohen M, Bernink PJ, McCabe CH, Horacek T, Papuchis G, Mautner B, Corbalan R, Radley D, Braunwald E (2000) The TIMI risk score for unstable angina/non–ST elevation MI. *J. Amer. Medical Assoc.* 284(7):835–842.

Austin J, Ocker R, Bhati A (2010) Kentucky pretrial risk assessment instrument validation. *Bureau of Justice Statistics*. (October), https://www.ncjrs.gov/App/Publications/abstract.aspx?ID=267494.

Berk RA, Bleich J (2013) Statistical procedures for forecasting criminal behavior. *Criminol. Public Policy* 12(3):513–544.

Bone R, Balk R, Cerra F, Dellinger R, Fein A, Knaus W, Schein R, Sibbald W, Abrams J, Bernard G, et al. (1992) American College of Chest Physicians/Society of Critical Care Medicine consensus conference: Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. *Critical Care Medicine* 20(6):864–874.

Breiman L (2001) Statistical modeling: The two cultures. *Statist. Sci.* 16(3):199–231.

Breiman L, Friedman J, Stone CJ, Olshen RA (1984) *Classification and Regression Trees* (CRC Press, Boca Raton, FL).

Burgess EW (1928) Factors determining success or failure on parole. Bruce AA, Harno AJ, Landesco J, Burgess EW, eds. *Parole and the Indeterminate Sentence: A Report to the Chairman of the Parole Board of Illinois on "The Workings of the Indeterminate Sentence Law and the Parole System in Illinois"* (Committee on the Study of the Workings of the Indeterminate Sentence Law and Parole, Springfield, IL), 205–249.

Bushway SD (2013) Is there any logic to using logit. *Criminology Public Policy* 12(3):563–567.

Caruana R, Niculescu-Mizil A (2004) Data mining in metric space: An empirical analysis of supervised learning performance criteria. *Proc. 10th ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining* (ACM, New York), 69–78.

Chen C, Rudin C (2018) An optimization approach to learning falling rule lists. Storkey A, Perez-Cruz F, eds. *Proc. Artificial Intelligence Statistics (AISTATS)* (PMLR, Fort Lauderdale, FL), 604–612.

Chung F, Yegneswaran B, Liao P, Chung SA, Vairavanathan S, Islam S, Khajehdehi A, Shapiro CM (2008) Stop questionnaire: A tool to screen patients for obstructive sleep apnea. *Anesthesiology* 108(5): 812–821.

Citron D (2016) (Un)fairness of risk scores in criminal sentencing. *Forbes* (January 13), https://www.forbes.com/sites/daniellecitron/2016/07/13/unfairness-of-risk-scores-in-criminal-sentencing/#10d06e974ad2.

Combs D, Shetty S, Parthasarathy S (2016) Big-data or slim-data: Predictive analytics will rule with world. *J. Clinical Sleep Medicine* 12(2):159–160.

Czeisler BM, Claassen J (2017) A novel clinical score to assess seizure risk. *JAMA Neurology* 74(12):1395–1396.

Danziger S, Levav J, Avnaim-Pesso L (2011) Extraneous factors in judicial decisions. *Proc. Natl. Acad. Sci. USA* 108(17): 6889–6892.

Dawes RM (1979) The robust beauty of improper linear models in decision making. *Amer. Psychol.* 34(7):571–582.

Ertekin Ş, Rudin C (2015) A Bayesian approach to learning scoring systems. *Big Data* 3(4):267–276.

Fisher A, Rudin C, Dominici F (2018) Model class reliance: Variable importance measures for any machine learning model class, from the "Rashomon" perspective. Working paper, Cornell University, Ithaca, New York.

Freitas AA (2014) Comprehensible classification models: A position paper. *ACM SIGKDD Explorations Newsletter* 15(1):1–10.

Gage BF, Waterman AD, Shannon W, Boechler M, Rich MW, Radford MJ (2001) Validation of clinical classification schemes for predicting stroke. *J. Amer. Medical Assoc.* 285(22):2864–2870.

Goh ST, Rudin C (2014) Box drawings for learning with imbalanced data. *Proc. 20th ACM SIGKDD Conf. Knowledge Discovery Data Mining (KDD)* (ACM, New York), 333–342.

Goodman B, Flaxman S (2016) European Union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine* 38(3):arXiv:1606.08813 [stat.ML].

Gottfredson DM, Snyder HN (2005) *The Mathematics of Risk Classification: Changing Data into Valid Instruments for Juvenile Courts* (Department of Justice, Office of Juvenile Justice and Delinquency Prevention, Washington, DC).

Hanson R, Thornton D (2003) *Notes on the Development of Static-2002* (Department of the Solicitor General of Canada, Ottawa, Ontario).

Ho V (2017) Miscalculated score said to be behind release of alleged twin peaks killer. SFGate, *San Francisco Chronicle* (August 14), https://www.sfgate.com/crime/article/Miscalculated-score-said-to-be-behind-11818814.php.

Hoffman PB (1994) Twenty years of operational use of a risk prediction instrument: The United States parole commission's salient factor score. *J. Criminal Justice* 22(6):477–494.

Hoffman PB, Adelberg S (1980) The salient factor score: A nontechnical overview. *Federal Probation* 44(1):44–52.

Holte RC (1993) Very simple classification rules perform well on most commonly used datasets. *Machine Learn.* 11(1):63–90.

Howard P, Francis B, Soothill K, Humphreys L (2009) OGRS 3: The revised offender group reconviction scale, Technical Report (Ministry of Justice, London).

ILOG (2007) *CPLEX 11.0 User's Manual* (IBM, New York).

Johns MW (1991) A new method for measuring daytime sleepiness: The Epworth sleepiness scale. *Sleep* 14(6):540–545.

Kahneman D (2013) *Thinking, Fast and Slow* (Farrar, Straus and Giroux, New York).

Knaus WA, Draper EA, Wagner DP, Zimmerman JE (1985) Apache II: A severity of disease classification system. *Critical Care Medicine* 13(10):818–829.

Knaus WA, Zimmerman JE, Wagner DP, Draper EA, Lawrence DE (1981) Apache-acute physiology and chronic health evaluation: A physiologically based classification system. *Critical Care Medicine* 9(8):591–597.

Knaus WA, Wagner D, Draper E, Zimmerman J, Bergner M, Bastos P, Sirio C, Murphy D, Lotring T, Damiano A (1991) The Apache III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults. *Chest J.* 100(6):1619–1636.

Kodratoff Y (1994) The comprehensibility manifesto. *KDD Nugget Newsletter* (IOS Press, Amsterdam, Netherlands), 83–85.

Lakkaraju H, Rudin C (2017) Learning cost-effective and interpretable treatment regimes. *Proc. 20th Internat. Conf. Artificial Intelligence Statistics* (PMLR, Fort Lauderdale, FL), 166–175.

Latessa E, Smith P, Lemke R, Makarios M, Lowenkamp C (2009) Creation and validation of the Ohio risk assessment system: Final report. Center for criminal justice research, school of criminal justice (University of Cincinnati, Cincinnati, OH), http://www.ocjs.ohio.gov/ORAS_FinalReport.pdf.

Le Gall JR, Lemeshow S, Saulnier F (1993) A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study. *J. Amer. Medical Assoc.* 270(24):2957–2963.

Letham B, Rudin C, McCormick TH, Madigan D (2015) Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model. *Ann. Appl. Statist.* 9(3):1350–1371.

Li O, Liu H, Chen C, Rudin C (2018) Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. *Proc. 32nd AAAI Conf. Artificial Intelligence* (AAAI Press, Palo Alto, CA), 1–8.

Moreno RP, Metnitz PG, Almeida E, Jordan B, Bauer P, Campos RA, Iapichino G, Edbrooke D, Capuzzo M, Le Gall JR (2005) SAPS 3-from evaluation of the patient to evaluation of the intensive care unit. Part 2: Development of a prognostic model for hospital mortality at ICU admission. *Intensive Care Medicine* 31(10): 1345–1355.

Northpointe (2015) Correctional offender management profiling for alternative sanctions (COMPAS). Accessed January 1, 2018, http://www.northpointeinc.com/files/technical_documents/FieldGuide2_081412.pdf.

Pazzani MJ (2000) Knowledge discovery from data? Intelligent systems and their applications. *IEEE* 15(2):10–12.

Pekkala T, Hall A, Lötjönen J, Mattila J, Soininen H, Ngandu T, Laatikainen T, Kivipelto M, Solomon A (2017) Development of a late-life dementia prediction index with supervised machine learning in the population-based CAIDE study. *J. Alzheimer's Disease* 55(3):1055–1067.

Pennsylvania Commission on Sentencing (2012) *Risk/Needs Assessment Project Interim Report 4: Development of Risk Assessment Scale* (Pennsylvania Commission on Sentencing, State College, PA).

Shah N, Steyerberg E, Kent D (2018) Big data and predictive analytics: Recalibrating expectations. *J. Amer. Medical Assoc.* 320(1): 27–28.

Shaw P, Ahn K, Rapoport JL (2017) Good news for screening for adult attention-deficit/hyperactivity disorder. *JAMA Psychiatry* 74(5): 527.

Six A, Backus B, Kelder J (2008) Chest pain in the emergency room: Value of the heart score. *Netherlands Heart J.* 16(6):191–196.

Souillard-Mandar W, Davis R, Rudin C, Au R, Libon DJ, Swenson R, Price CC, Lamar M, Penney DL (2016) Learning classification models of cognitive conditions from subtle behaviors in the digital clock drawing test. *Machine Learn.* 102(3):393–441.

Struck AF, Ustun B, Rodriguez Ruiz A, Lee JW, LaRoche S, Hirsch LJ, Gilmore EJ, Rudin C, Westover BM (2017) A practical risk score for EEG seizures in hospitalized patients. *JAMA Neurology* 74(12):1419–1424.

Than M, Flaws D, Sanders S, Doust J, Glasziou P, Kline J, Aldous S, Troughton R, Reid C, Parsonage WA, et al. (2014) Development and validation of the emergency department assessment of chest pain score and 2 h accelerated diagnostic protocol. *Emergency Medicine Australasia* 26(1):34–44.

Tollenaar N, van der Heijden P (2013) Which method predicts recidivism best? A comparison of statistical, machine learning and data mining predictive models. *J. Royal Statist. Soc. Ser. A* 176(2): 565–584.

U.S. Department of Justice, Bureau of Justice Statistics (2014) Recidivism of prisoners released in 1994. Accessed January 1, 2018, http://doi.org/10.3886/ICPSR03355.v8.

U.S. Sentencing Commission (1987) 2012 guidelines manual: Chapter four - criminal history and criminal livelihood. Accessed January 1, 2018, http://www.ussc.gov/guidelines-_manual/2012/2012-_4a11.

U.S. Sentencing Commission (2004) Measuring recidivism: The criminal history computation of the federal sentencing guidelines. Accessed January 1, 2018, https://www.ussc.gov/sites/default/files/pdf/research-and-publications/research-publications/2004/200405_Recidivism_Criminal_History.pdf.

Ustun B, Rudin C (2016a) Learning optimized risk scores for large-scale datasets. arXiv:1610.00168.

Ustun B, Rudin C (2016b) Supersparse linear integer models for optimized medical scoring systems. *Machine Learn.* 102(3):349–391.

Ustun B, Rudin C (2017) Optimized risk scores. *Proc. 23rd ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining* (ACM, New York), 1125–1134.

Ustun B, Westover MB, Rudin C, Bianchi MT (2016) Clinical prediction models for sleep apnea: The importance of medical history over symptoms. *J. Clinical Sleep Medicine* 12(2):161–168.

Ustun B, Adler LA, Rudin C, Faraone SV, Spencer TJ, Berglund P, Gruber MJ, Kessler RC (2017) The World Health Organization adult attention-deficit/hyperactivity disorder self-report screening scale for DSM-5. *JAMA Psychiatry* 74(5):520–526.

Wang F, Rudin C (2015) Falling rule lists. *Proc. 18th Internat. Conf. Artificial Intelligence Statistics (AISTATS)*, May 9–12, San Diego, CA.

Wang T, Rudin C, Doshi F, Liu Y, Klampfl E, MacNeille P (2016) Bayesian or's of and's for interpretable classification with application to context aware recommender systems. Lebanon G, Vishwanathan SVN, eds. *Internat. Conf. Data Mining (ICDM)* (PMLR, Fort Lauderdale, FL), arXiv:1504.07614 [cs.LG].

Wang T, Rudin C, Doshi-Velez F, Liu Y, Klampfl E, MacNeille P (2017) A Bayesian framework for learning rule sets for interpretable classification. *J. Machine Learn. Res.* 18(70):1–37.

Weathers FW, Litz BT, Keane TM, Palmieri PA, Marx BP, Schnurr PP (2013) The PTSD checklist for DSM-5 (pcl-5). National Center for PTSD, http://www.ptsd.va.gov.

Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N (2017) Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS One*. Accessed June 1, 2017, http://journals.plos.org/plosone/article/authors?id=10.1371/journal.pone.0174944.

Wexler R (2017a) Code of silence: How private companies hide flaws in the software that governments use to decide who goes to prison and who gets out. *Washington Monthly*, https://washingtonmonthly.com/magazine/junejulyaugust-2017/code-of-silence/.

Wexler R (2017b) When a computer program keeps you in jail: How computers are harming criminal justice. *New York Times* (June 13), https://www.nytimes.com/2017/06/13/opinion/how-computers-are-harming-criminal-justice.html.

Wolsey LA (1998) *Integer Programming*, Vol. 42 (Wiley, New York).

Yang H, Rudin C, Seltzer M (2017) Scalable Bayesian rule lists. Precup D, Teh YW, eds. *Proc. 34th Internat. Conf. Machine Learn. (ICML)* (PMLR, Fort Lauderdale, FL), 3921–3930.

Zadrozny B, Elkan C (2002) Transforming classifier scores into accurate multiclass probability estimates. *Proc. 8th ACM SIGKDD Internat. Conf. on Knowledge Discovery Data Mining* (ACM, New York), 694–699.

Zeng J, Ustun B, Rudin C (2017) Interpretable classification models for recidivism prediction. *J. Royal Statist. Soc. Ser. A* 180(3): 689–722.

Cynthia Rudin is an associate professor of computer science, electrical and computer engineering, and statistics at Duke University. Previously, Prof. Rudin held positions at MIT, Columbia, and NYU. She holds a BS and BA from the University at Buffalo, and a PhD from Princeton. She is the recipient of the 2013 and 2016 INFORMS Innovative Applications in Analytics Awards and an NSF CAREER award, and was named one of the "Top 40 Under 40" by Poets and Quants in 2015 and by Businessinsider.com as one of the 12 most impressive professors at MIT in 2015. She has served on committees for DARPA, the NIJ, the NASEM, AAAI, INFORMS, and ASA.

Berk Ustun is a postdoctoral fellow at the Harvard University Center for Research in Computation for Society. His research focuses on fairness and interpretability in machine learning and causal inference. He is the recipient of the 2016 INFORMS Innovative Applications in Analytics award and the 2017 INFORMS Computing Society Best Student Paper award. He holds a PhD in electrical engineering and computer science from MIT and BS degrees in operations research and economics from UC Berkeley.