Learning Optimal Predictive Checklists

Haoran Zhang
Massachusetts Institute of Technology
haoranz@mit.edu

Quaid Morris
Memorial Sloan Kettering Cancer Center
morrisq@mskcc.org

Berk Ustun*
UC San Diego
berk@ucsd.edu

Maryzeh Ghassemi*
Massachusetts Institute of Technology
mghassem@mit.edu

Abstract

Checklists are simple decision aids that are often used to promote safety and reliability in clinical applications. In this paper, we present a method to learn checklists for clinical decision support. We represent predictive checklists as discrete linear classifiers with binary features and unit weights. We then learn globally optimal predictive checklists from data by solving an integer programming problem. Our method allows users to customize checklists to obey complex constraints, including constraints to enforce group fairness and to binarize real-valued features at training time. In addition, it pairs models with an optimality gap that can inform model development and determine the feasibility of learning sufficiently accurate checklists on a given dataset. We pair our method with specialized techniques that speed up its ability to train a predictive checklist that performs well and has a small optimality gap. We benchmark the performance of our method on seven clinical classification problems, and demonstrate its practical benefits by training a short-form checklist for PTSD screening. Our results show that our method can fit simple predictive checklists that perform well and that can easily be customized to obey a rich class of custom constraints.

1 Introduction

Checklists are simple tools that are widely used to assist humans when carrying out important tasks or making important decisions [8, 13, 18, 35, 44, 51, 61, 62, 65, 66, 67]. These models are often used as predictive models in modern healthcare applications. In such settings, a checklist is a set of Boolean conditions that predict a condition of interest – e.g., a list of $N$ symptoms of which $M$ or more must be checked to flag a patient for a critical illness. These kinds of “predictive checklists” are often used for clinical decision support because they are easy to use and easy to understand [33, 55]. In contrast to other kinds of predictive models, clinicians can easily scrutinize a checklist, and make an informed decision regarding its adoption. Once they have decided to use a checklist, they can integrate the model into their clinical workflow without extensive training or technology (e.g., as a printed sheet [53]).

Considering these benefits, one of the key challenges in using predictive checklists in healthcare applications is finding a reliable way to create them [33]. Most predictive checklists in medicine are either hand-crafted by panels of experts [30, 41], or built by combining statistical techniques and heuristics [e.g., logistic regression, stepwise feature selection, and rounding 39]. These approaches

* Equal supervision.
make it difficult to develop checklists that are sufficiently accurate – as panel or pipeline will effectively need to specify a model that performs well under stringent assumptions on model form. Given the simplicity of the model class, it is entirely possible that some datasets may never admit a checklist that is sufficiently accurate to deploy in a clinical setting – as even checklists that are accurate at a population-level may perform poorly on a minority population [58, 71].

In this paper, we introduce a machine learning method to learn checklists from data. Our method is designed to streamline the creation of predictive checklists in a way that overcomes specific challenges of model development in modern healthcare applications. Our method solves an integer programming problem to return the most accurate checklist that obeys user-specified constraints on model form and/or model performance. This approach is computationally challenging, but provides specific functionality that simplifies and streamlines model development. First, it learns the most accurate checklist by optimizing exact measures of model performance (i.e., accuracy rather than a convex surrogate measure). Second, it seeks to improve the performance of checklists by adaptively binarizing features at training time. Third, it allows practitioners to train checklists that obey custom requirements on model form or on prediction, by allowing them to encode these requirements as constraints in the optimization problem. Finally, it provides practitioners with an optimality gap, which informs them when a sufficiently accurate checklist does not exist.

The main contributions of this paper are:

1. We present a machine learning method to learn checklists from data. Our method allows practitioners to customize models to obey a wide range of real-world constraints. In addition, it pairs checklists with an optimality gap that can inform practitioners in model development.
2. We develop specialized techniques that improve the ability of our approach to train a checklist that performs well, and to pair this model with a small optimally gap. One of these techniques can be used to train checklists heuristically.
3. We conduct a broad empirical study of predictive checklists on clinical classification tasks [24, 32, 37, 52, 59]. Our results show that predictive checklists can perform as well as state-of-the-art classifiers on some datasets, and that our method can provide practitioners with an optimality gap to flag when this is not the case. We highlight the ability of our method to handle real-world requirements through applications where we enforce group fairness constraints on a mortality prediction task and build a short-form checklist to screen for PTSD.
4. We provide a Python package to train customized checklists with open-source and commercial solvers, including CBC [29] and CPLEX [20] (see https://github.com/MLforHealth/predictive_checklists).

2 Related Work

The use of checklist-style classifiers – e.g., classifiers that assign coefficients of ±1 to Boolean conditions – dates to the work of Burgess [9]. The practice is often referred to as “unit weighting,” and is motivated by observations that it often performs surprisingly well [see e.g., 7, 19, 23, 27].

Our work is strongly related to methods that learn sparse linear classifiers with small integer coefficients [see e.g., 4, 11, 31, 38, 69, 70, 76]. We learn checklists from data by solving an integer program. Our problem can be viewed as a special case of an IP proposed by Ustun and Rudin [69], Zeng et al. [78] to learn sparse linear classifiers with small integer coefficients. The problem that we consider is considerably easier to solve from a computational standpoint because it restricts coefficients to binary values and does not include \( \ell_0 \)-regularization. Here, we make use of this “slack” in computation to add functionality that improves the performance of checklists, namely: (1) constraints to binarize continuous and categorical features into items that can be included in a checklist [see also 10, 12, 72]; and (2) develop specialized techniques to reduce computation during the learning process.

Checklists are \( M \)-of-\( N \) rules — i.e., classifiers that predict \( y = +1 \) when \( M \) out of \( N \) conditions are true. Early work in machine learning used \( M \)-of-\( N \) rules in an auxiliary capacity – e.g., to explain the predictions of neural networks [68], or to use as features in decision trees [79]. More recent work has focused on learning \( M \)-of-\( N \) rules for standalone prediction – e.g., Chevaleyre et al. [16] describe a method to learn \( M \)-of-\( N \) rules by rounding the coefficients of linear SVMs to \( \{-1, 0, 1\} \) using a randomized rounding procedure [see also 80]. Methods to learn disjunctions and conjunctions are also relevant since they correspond to 1-of-\( N \) rules and \( N \)-of-\( N \) rules, respectively. These methods
learn models by solving a special case of the ERM problem in (1) where \( M = 1 \) or \( M = N \). Given that this problem is NP-hard, existing methods often reduce computation by using algorithms that return approximate solutions – e.g., simulated annealing [73], set cover [49, 50], or ERM with a convex surrogate loss [21, 22, 47, 48]. These approaches could be used to learn checklists by solving the ERM problem in (1). However, they would not be able to handle the discrete constraints needed for binarization and customization. More generally, they would not be guaranteed to recover the most accurate checklists for a given dataset, nor pair models with a certificate of optimality that can be used to inform model development.

3 Methodology

We start with a dataset of \( n \) examples \( (x_i, y_i)_{i=1}^n \) where \( x_i = [x_{i1}, \ldots, x_{id}] \in \{0, 1\}^d \) is a vector of \( d \) binary variables and \( y_i \in \{0, 1\} \) is a label. Here, \( x_i \) are Boolean variables that could be used as items in a checklist. We denote the indices of positive and negative examples as \( I^+ = \{i \mid y_i = +1\} \) and \( I^- = \{i \mid y_i = -1\} \) respectively, and let \( n^+ = |I^+| \) and \( n^- = |I^-| \). We denote the set of positive integers up to \( k \) as \([k] = \{1, \ldots, k\}\).

We use the dataset to learn a predictive checklist – i.e., a Boolean threshold rule that predicts \( \hat{y} = +1 \) when at least \( M \) of \( N \) items are checked. We represent a checklist as a linear classifier with the form:

\[
\hat{y}_i = \text{sign}(\lambda^T x_i \geq M). 
\]

Here \( \lambda = [\lambda_1, \ldots, \lambda_d] \in \{0, 1\}^d \) is a coefficient vector and \( \lambda_j = 1 \) iff the checklist contains item \( j \). We denote the number of items in a checklist with coefficients \( \lambda \) as \( N = \sum_{j=1}^d \lambda_j \), and denote the threshold number of items that must be checked to assign a positive prediction as \( M \). We learn checklists from data by solving an empirical risk minimization problem with the form:

\[
\min_{\lambda, M} \quad l(\lambda, M) + \epsilon_N N + \epsilon_M M \\
\text{s.t.} \quad N = \|\lambda\| \\
M \in [N] \\
\lambda \in \{0, 1\}^d 
\]

Here, \( l(\lambda, M) = \sum_{i=1}^n \mathbb{I}[y_i \neq \hat{y}_i] \) counts the number of mistakes of a checklist with parameters \( \lambda \) and \( M \). The parameters \( \epsilon_N > 0 \) and \( \epsilon_M > 0 \) are small penalties used to specify lexicographic preferences. We set \( \epsilon_N < \frac{\epsilon_M}{M} \) and \( \epsilon_M < \frac{\epsilon_N}{N} \). These choices ensure that optimization will return the most accurate checklist, breaking ties between checklists that are equally accurate to favor smaller \( N \), and breaking ties between checklists that are equally accurate and sparse to favor smaller \( M \). Smaller values of \( N \) and \( M \) are preferable, as checklists with smaller \( N \) users check fewer items, and checklists with smaller \( M \) let users stop checking as soon as \( M \) items are checked (see Figure 3).

We recover a globally optimal solution to (1) by solving the integer program (IP):

\[
\min_{\lambda, z, M} \quad l^+ + W^- l^- + \epsilon_N N + \epsilon_M M \\
\text{s.t.} \quad B_i z_i \geq M - \sum_{j=1}^d \lambda_j x_{i,j} \quad i \in I^+ \quad (2a) \\
B_i z_i \geq \sum_{j=1}^d \lambda_j x_{i,j} - M + 1 \quad i \in I^- \quad (2b) \\
l^+ = \sum_{i \in I^+} z_i \\
l^- = \sum_{i \in I^-} z_i \\
N = \sum_{j=1}^d \lambda_j \\
M \in [N] \\
z_i \in \{0, 1\} \quad i \in [n] \\
\lambda_j \in \{0, 1\} \quad j \in [d] 
\]

Here, \( l^+ \) and \( l^- \) are variables that count the number of mistakes on positive and negative examples. The values of \( l^+ \) and \( l^- \) are computed using the mistake indicators \( z_i = \mathbb{I}[\hat{y}_i \neq y_i] \), which are set to
1 through “Big-M” Constraints in (2a) and (2a). These constraints depend on the “Big-M” parameters $B_i$, which can be set to its tightest possible value $\max \lambda_i M - \lambda^\top x_i$. $W$ is a user-defined parameter that reflects the relative cost of misclassifying a negative example. By default, we set $W = 1$, so that the objective minimizes training error.

Customization  The IP in (2) can be customized to fit checklists that obey a wide range of real-world requirements on performance and model form. We list examples of constraints in Table 1. Our method provides two ways of handling classification problems with different costs of False Negatives (FNs) and False Positives (FPs). First, the user can specify the relative misclassification costs for FPs and FNs in the objective function by setting $W$. Here, optimizing (2) corresponds directly to minimizing the weighted error. Second, the user can instead specify limits on FPR and FNR - e.g. minimize the training FPR (by setting $W^- = n^-$) subject to training FNR $\leq 20\%$.

<table>
<thead>
<tr>
<th>Model Requirement</th>
<th>Example</th>
<th>Constraint</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model Size</td>
<td>Use $\leq N_{max}$ items</td>
<td>$N \leq N_{max}$</td>
</tr>
<tr>
<td>Procedural</td>
<td>If checklist includes item about coughing, then include item about fever $\lambda_{fever} \geq \lambda_{cough}$</td>
<td></td>
</tr>
<tr>
<td>Prediction</td>
<td>Predict $y_i = +1$ for all patients with both fever and coughing $\lambda_{fever} \cdot y_i + \lambda_{cough} \cdot y_i \geq M \forall i \in [n]$</td>
<td></td>
</tr>
<tr>
<td>Class-Based Accuracy</td>
<td>Max FPR $\leq \gamma$</td>
<td>$f^- \leq</td>
</tr>
<tr>
<td>Group Fairness</td>
<td>Max FPR disparity of $\gamma%$ between males and females $</td>
<td>\frac{f^-_m}{\gamma} - \frac{f^-_f}{\gamma}</td>
</tr>
<tr>
<td>Minimax Fairness</td>
<td>No group with FNR worse than $\delta$</td>
<td>$\lambda_{\delta} \leq \delta \cdot n^\top \gamma \in G$</td>
</tr>
</tbody>
</table>

Table 1: Model requirements that can be addressed by our method. Each requirement can be directly encoded into IP (2). The IP can then be solved using the same MIP solver to obtain the most accurate checklist that obeys these constraints.

Adaptive Binarization  Practitioners often work with datasets that contain real-valued and categorical features. In such settings, rule-learning methods often require practitioners to binarize such features before training. Our approach allows practitioners to binarize features into items during training. This allows all items to be binarized in a way that maximizes accuracy. This approach requires practitioners to first define $T_j$ candidate items for each non-binary feature $u_{i,j}$. For example, given a real-valued feature $u_{i,j} \in \mathbb{R}$, the set of candidate items could take the form $\{x_{i,j,1}, x_{i,j,2}, \ldots, x_{i,j,T_j}\}$ where $x_{i,j,t} = \mathbb{I} \left[u_{i,j} \geq v_{i,j,t}\right]$ and $v_{i,j,t}$ is a threshold. We would add the following constraint to IP (2) to ensure that the checklist only uses one of the $T_j$ items: $\sum_{t \in [T_j]} \lambda_{i,j} \leq 1$. In this way, the IP would then choose a binary item that is aligned with the goal of maximizing accuracy. In practice, the set of candidate items can be produced automatically [e.g., using binning or information gain $54$] or specified on the basis of clinical standards (as real-valued features like blood pressure and BMI have established thresholds). This approach is general enough to allow practitioners to optimize over all possible thresholds, though this may lead to overfitting. Our experiments in Section 5.2 show that adaptive binarization produces a meaningful improvement in performance for our model class.

Optimality Gap  We solve IP (2) with a MIP solver, which finds a globally optimal solution through exhaustive search algorithms like branch-and-bound $74$. Given an instance of IP (2) and a time limit for when to stop the search, a solver returns: (i) the best feasible checklist found within the time limit – i.e., values of $\lambda$ and $M$ that achieve the best objective value $V^{max} = l(\lambda, M)$; and (ii) a lower bound on the objective value – i.e., $V^{min}$, the minimal objective value for any feasible solution. These quantities are used to compute the optimality gap $\varepsilon := 1 - \frac{V^{max}}{V^{min}}$. When the upper bound $V^{max}$ matches the lower bound $V^{min}$, the solver returns a checklist with an optimality gap of $\varepsilon = 0$ – i.e., one that is certifiably optimal.

In our setting, the optimality gap reflects the worst-case difference in training error between the checklist that a solver returns and the most accurate checklist that we could hope to obtain by solving IP (2). Given a solution to IP (2) with an objective value of $L$ and an optimality gap of $\varepsilon$, any feasible checklist has a training error of at least $\left(1 + \varepsilon\right)L$. When $\varepsilon$ is small, the most accurate checklist has a training error $\approx L$. Thus, if we are not satisfied with the performance of our model, we know that no checklist can perform better than $L$, and can make an informed decision to relax constraints, fit a classifier from a more complex hypothesis class, or not fit a classifier at all. When $\varepsilon$ is large, our problem may admit a checklist with training error far smaller than $L$. If we are not satisfied with the
performance of the checklist, we cannot determine whether this is because the dataset does not admit a sufficiently accurate model or because the solver has not been able to find it.

4 Algorithmic Improvements

In this section, we present specialized techniques to speed up the time required to find a feasible solution, and/or to produce a solution with a smaller optimality gap.

4.1 Submodular Heuristic

Our first technique is a submodular optimization technique that we use to generate initial feasible solutions for IP (2). We consider a knapsack cover problem. This problem can be solved to produce a checklist of \( N_{\text{max}} \) or fewer items that: (1) maximizes coverage of positive examples; (2) has a “budget” of negative examples \( B \); and (3) uses at most one item from each of the feature groups \( R_1, \ldots, R_T \) where each \( R_t \) denote a set of items derived from a particular real-valued feature and \( \bigcup_{t=1}^{T} R_t = [d] \).

\[
\begin{align*}
\max_{A} & \quad f^+_M(A) \\
\text{s.t.} & \quad |A| \leq N_{\text{max}} \\
& \quad |A \cap R_t| \leq 1 \quad t \in [T] \\
& \quad \sum_{j \in A} c(j) \leq B
\end{align*}
\]

Here, \( f^+_M := \sum_{i \in I^+} \min(\sum_{j \in A} x_{i,j}, M) \) counts the positive examples that are covered by at least \( M \) items in \( A \). Constraint (3b) limits the number of items in the checklist. Constraint (3c) requires the checklist to use at most one item from each feature group \( R_1, \ldots, R_T \). Constraint (3d) controls the number of negative examples that are covered by at least \( M \) items in \( A \). This is a knapsack constraint that assigns a cost to each item as \( c(j) = \sum_{i \in I^-} x_{i,j} \) and limits the overall cost of \( A \) using the budget parameter \( B > 0 \).

The optimization problem in (3) can be solved using a submodular minimization algorithm given that \( f^+_M \) is monotone submodular (see Appendix A.1 for a proof), and that constraints (3b) to (3d) are special kinds of matroid constraints. We solve (3) with a submodular minimization algorithm adapted from Badanidiyuru and Vondrák [2] (Appendix A.2). The procedure takes as input \( M, N_{\text{max}} \) and \( B \) and outputs a set of checklists that use up to \( N_{\text{max}} \) items. This procedure can run within 1 second. Given a problem where we would want to train a checklist that has at most \( N_{\text{max}} \) items, we would solve this problem \(|B|N_{\text{max}}\) times, varying \( M \in [N_{\text{max}}] \) and \( B \in B \).

**Algorithm 1: Sequential Training over \( N \) and \( M \)**

**Input:** maximum # of items \( N_{\text{max}} \)

**Input:** objective value function \( V(\cdot) \)

\[ \text{pool} \leftarrow \emptyset \]

for \( i \in [N_{\text{max}}], \ j \in [i] \) do

\[
S \leftarrow \{ C | C \in \text{pool}, N(C) \leq i, M(C) \leq j \} \\
C_{\text{init}} \leftarrow \arg\min_{C \in S} f(C) \\
C_{\text{new}} \leftarrow \text{solve IP (2) with initial solution } C_{\text{init}} \\
\quad \text{subject to } N \leq i \text{ and } M \leq j \\
\quad \text{Add } \{ C_{\text{new}} \} \text{ to pool}
\]

end

return pool

(a)

In Figure 1a, we show the effect of initializing the IP with checklists produced using the submodular heuristic. As shown, these solutions allow us to initialize the search with a good upper bound, allowing us to ultimately recover a certifiably optimal solution faster. Our empirical results in Section

(b)

Figure 1: (a) Performance profile when solving IP (2) with and without initialization. We show the optimality gap vs. the # of branch-and-bound nodes when solving IP (2) for the heart dataset where \( N \leq 6 \) and \( M \leq 2 \). (b) algorithm for sequential training over \( N \) and \( M \).
5.2 show that checklists from our submodular heuristic outperform checklists fit using existing baselines. This highlights the value of this approach as a heuristic to fit checklists (in that it can produce checklists that perform well), and as a technique for initialization (in that it can produce checklists that can be used to initialize IP (2) with a small upper bound).

4.2 Path Algorithms

Our second technique is a path algorithm for sequential training. Given an instance of the checklist IP (2), this algorithm sequentially solves smaller instances of IP (2). The algorithm exploits the fact that small instances can be solved to optimality quickly and thus “ruled out” of the feasible region of subsequent instances. The resulting approach can reduce the size of the branch-and-bound tree and improve the ability to recover a checklist with a small optimality gap. These algorithms can wrap around any procedure to solve IP (2).

In Algorithm 1, we present a path algorithm to train checklists over the full range of model sizes from 1 to \( N_{\text{max}} \). The algorithm recovers a solution to a problem with a hard model size limit by solving smaller instances with progressively larger limits of \( N_{\text{max}} \) and \( M_{\text{max}} \). We store the best feasible solution from each iteration in a pool, which is then used to initialize future iterations. If the best feasible solution is certifiably optimal, we add a constraint to constrict the search space for subsequent instances. This allows us to reduce the size of the branch-and-bound tree, resulting in a faster solution in each iteration.

5 Experimental Results

In this section, we benchmark the performance of our methods on seven clinical prediction tasks.

5.1 Setup

Data. We consider seven clinical classification tasks shown in Table 2. For each task, we create a classification dataset by using each of the following techniques to binarize features:

**Fixed:** We convert continuous and ordinal features into threshold indicator variables using the median, and convert categorical feature into an indicator for the most common category.

**Adaptive:** We convert continuous and ordinal features into 4 threshold indicator variables using quintiles as thresholds, and convert each categorical feature with a one-hot encoding.

**Optbinning:** We use the method proposed by Navas-Palencia [54] to binarize all features.

We process each dataset to oversample the minority class to equalize the number of positive and negative examples due to class imbalance.

Methods. We compare the performance of the following methods for creating checklists:

**MIP:** We fit a predictive checklist by solving an instance of IP (2) with the appropriate constraints. We solve this problem using CPLEX 12.10 [20] paired with the computational improvements in Section 4 on a 2.4 GHz CPU with 16 GB RAM for \( \leq 60 \) minutes.

**Cover:** We fit a predictive checklist using the submodular heuristic in Section 4.1. We set \( N_{\text{max}} = 8 \), and vary \( M \in [N_{\text{max}}] \) and \( B \in \{kn^+|k \in \{\frac{1}{3}, \frac{1}{2}, 1, 2, 3\}\} \).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>( d_T )</th>
<th>( d_{\text{pre}} )</th>
<th>( n )</th>
<th>Prediction Task</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>adhd</td>
<td>5</td>
<td>20</td>
<td>594</td>
<td>Patient diagnosed with ADHD</td>
<td>[52]</td>
</tr>
<tr>
<td>cardio</td>
<td>40</td>
<td>52</td>
<td>8,815</td>
<td>Patient with cardiogenic shock died in hospital</td>
<td>[59]</td>
</tr>
<tr>
<td>kidney</td>
<td>17</td>
<td>80</td>
<td>1,722</td>
<td>Patient with kidney failure died after renal replacement therapy</td>
<td>[37]</td>
</tr>
<tr>
<td>mortality</td>
<td>50</td>
<td>484</td>
<td>21,139</td>
<td>Patient died in hospital</td>
<td>[37]</td>
</tr>
<tr>
<td>ptsd</td>
<td>20</td>
<td>80</td>
<td>873</td>
<td>Patient has a PCL-5 PTSD diagnosis</td>
<td>[52]</td>
</tr>
<tr>
<td>readmit</td>
<td>42</td>
<td>94</td>
<td>9,766</td>
<td>Patient re-admitted to hospital within 30 days</td>
<td>[32]</td>
</tr>
<tr>
<td>heart</td>
<td>13</td>
<td>82</td>
<td>303</td>
<td>Patient has heart disease</td>
<td>[24]</td>
</tr>
</tbody>
</table>

Table 2: Clinical prediction tasks used in Section 5. Here, \( n \) is the number of samples prior to oversampling, \( d_T \) is the number of feature groups, and \( d_{\text{pre}} \) is the number of features after binarization using the adaptive method.
We report results for MIP and XGBoost (Table 3: Mptsd MCV test error). As shown in Figure 2, this improvement typically holds across all values of N. Lastly, we show predictive checklists trained using our method in Figure 3 and Appendix D.3. In what follows, we discuss these results.

### 5.2 Results

In Table 3, we report performance on each dataset when using Fixed binarization and Adaptive binarization. We report the corresponding results for Optbinning in Appendix D.2 due to space limitations. In Figure 2, we show the performance effects of varying N. Lastly, we show predictive checklists trained using our method in Figure 3 and Appendix D.3. In what follows, we discuss these results.

### On Performance

Our results in Table 3 show that our method MIP outperforms existing techniques to fit checklists (Unit) on 7/7 datasets in terms of training error, and on 5/7 datasets in terms of 5-fold CV test error. As shown in Figure 2, this improvement typically holds across all values of N.

In general, we find that relaxing the value of M yield performance gains across all datasets (see e.g., results for ptsd where the difference in test error between MIP OR and MIP is 7.8%), though smaller values of M can produce checklists that are easier to use (see checklists in Appendix D.3).
Figure 2: Performance of checklists with $\leq N$ items on mortality fit using our approach and unit weighting.

Figure 3: Predictive checklist fit for readmit using MIP. The checklist has a train error of 33.1%, and a test error of 33.9%. We show sample checklists for all other datasets in Appendix D.3.

The results in Table 3 also highlight the potential of adaptive binarization to improve performance by binarizing during training time, as MIP with adaptive binarization achieves a better training error than fixed binarization for 6/7 datasets and a better test error on 5/7 datasets.

The performance loss of checklists relative to more complex model classes (LR and XGB) varies greatly between datasets. Checklists trained on heart and readmit perform nearly as well as more complex classifiers. On mortality and kidney, however, the loss may exceed 5%.

On Generalization Our results show that checklist performance on the training set tends to generalize well to the test set. This holds for checklists trained using all methods (i.e., Unit, Cover, MIP OR, and MIP), which is expected given that checklists belong to a simple model class. Generalization can help us customize checklists in an informed way as we can assume that any changes to the training loss (e.g. with the addition of a constraint) lead to similar changes in the test set. In this regime, practitioners can evaluate the impact of model requirements on predictive performance – since they can measure differences in training error between checklists that obey different requirements, and expect these differences in training error as a reliable proxy for differences in test error.

On the Value of an Optimality Gap One of the benefits of our approach is that it pairs each checklist with an optimality gap, which can inform model development by identifying settings where we cannot train a checklist that is sufficiently accurate (see Section 3). On heart, we train a checklist with a training error of 13.6% and an optimality gap of 54.5%, which suggests that there may exist a checklist with a training error of 6.2%. This optimality gap is not available for checklists trained with a heuristic (like unit weighting or domain expertise), so we would not be able to attribute poor model performance to a difficult problem, or to the heuristic being ineffective.

On Fairness We analyze the fairness of MIP checklists trained on the kidney dataset. The task is to predict in-hospital mortality for ICU patients with kidney failure after receiving Continuous Renal Replacement Therapy (CRRT). In this setting, a false negative is often worse than a false positive, so we train checklists that minimize training FPR under a constraint to limit training FNR to 20%, using an 80%/20% train/test split.

We find that our checklist exhibit performance disparities across sex and race, which is consistent with the tendency of simpler models to exhibit performance disparities [40]. To address this issue, we train a new checklist by solving a version of IP (2) with constraints that limit disparities in training FPR and FNR over 6 intersectional groups $g \in G = \{\text{Male,Female}\} \times \{\text{White,Black,Other}\}$. Specifically, 30 constraints that limit the FPR disparity across groups to 15%, and 6 constraints to cap the max FNR for each intersectional group to 20%. This is unique compared to prior methods in fair classification work [77] because it enforces group fairness using exact measures[i.e., without convex relaxations 46] and over intersectional groups.

We find that we can limit performance disparities across all values of $N$ (see Figures 4 and 5, and Appendix D.4), observing that MIP with group fairness constraints has larger FPR overall but exhibits lower FPR disparities over groups (e.g., reducing the test set FPR gap between Black Females and White Males from 54.5% to 30.6%).
Figure 4: FPR (left), maximum FPR disparity (center) and worst-case FNR (right) versus number of items for checklists for mortality prediction given CRRT using kidney. Group fairness constraints effectively limit the FPR disparity and worst-case FNR during training, at the cost of overall performance.

Figure 5: Checklists trained to predict mortality given CRRT (a) without group fairness constraints and (b) with group fairness constraints. Checklist (a) has better overall FPR, but has worse FNR for the worst-case group, and exhibits disparities in FPR across groups. Checklist (b) obeys group fairness constraints that attenuate the training FPR gap to 15% and the worst-case FNR to 20%.

6 Learning a Short-Form Checklist for PTSD Screening

In this section, we demonstrate the practical benefits of our approach by learning a short-form checklist for Post-Traumatic Stress Disorder (PTSD) diagnosis.

Background The PTSD Checklist for DSM-5 (PCL-5) is a self-report screening tool for PTSD [5, 75]. The PCL-5 consists of 20 questions that assess the severity of symptoms of PTSD DSM-5. Patients respond to each question with answers of Not at all, A little bit, Moderately, Quite a bit, or Extremely. Given these responses, the PCL assigns a provisional diagnosis of PTSD by counting the number of responses of Moderately or more frequent across four clusters of questions. Patients with a provisional diagnosis are then assessed for PTSD by a specialist. The PCL-5 can take over 10 minutes to complete, which limits its use in studies that include the PCL-5 in a battery of tests (i.e., using the provisional diagnosis to evaluate the prevalence of PTSD and its effect as a potential confounder), and has led to the development of short-form models [see e.g., 6, 43, 81].

Problem Formulation Our goal is to create a short-form version of the PCL-5 — i.e., a checklist that assigns the same provisional diagnoses as PCL-5 using only a subset of the 20 questions. We train our model using data from the AURORA study [52], which contains PCL-5 responses of U.S. patients who have visited the emergency department following a traumatic event. Here, \( y_i = +1 \) if the patient is assigned a provisional PTSD diagnosis. We encode the response for each question into four binary variables \( x_{i,k,l} = 1 \) if \( q_k \geq l \) where \( q_k \) is the response to each question and \( l \in [4] \) is its response. Our final dataset contains \( d = 80 \) binary variables for \( n = 873 \) patients, which we split into an 80% training set and 20% test set.

Since this checklist would be used to screen patients, a false negative diagnosis is less desirable than a false positive diagnosis. We train a checklist that minimizes FPR and that obeys the following operational constraints: (1) limit FNR to 5%; (2) pick one threshold for each PCL-5 question; (3) use at most 8 questions. We make use of the same methods and techniques as in Section 5.2.
Results Our results show that short-form checklists fit using our method outperform other checklists (see Figure 6). In Figure 7, we display a sample $N = 8, M = 4$ checklist trained with our method which reproduces the PCL-5 diagnosis with high accuracy, offering an alternative to the PCL-5 with simpler binary items and less than half the number of questions.

Figure 6: FPR versus number of items for short-forms of the PCL-5. MIP outperforms Unit across all values of $N$.

Figure 7: Short-form of the PCL-5 trained with MIP. The checklist has a train/test FPR of 6.6%/10.8%, and a train/test FNR of 4.7%/6.3%. It reproduces the PCL-5 diagnosis on the training/test sets with 94.1/90.9% accuracy, and has a 70.8% optimality gap. We show condensed PCL-5 questions due to space constraint.

7 Concluding Remarks

In this work, we presented a machine learning method to learn predictive checklists from data by solving an integer program. Our method illustrates a promising approach to fit models to obey constraints related to qualities like safety [1] and fairness [3, 14]. Using our approach, practitioners can potentially co-design checklists alongside clinicians – by encoding requirements into the model and reporting their effects on predictive performance [17, 36].

We emphasize that effective adoption of the checklist models we create does not end at deployment. The proper integration of our models in a clinical setting will require careful negotiation at the personal and institutional level, and should be considered in the context of care delivery [28, 63, 64] and clinical workflows [26, 56].

We caution against the use of items in a checklist as causal factors for outcomes (i.e., questions in the short-form PCL are not “drivers” of PTSD), and against the use of our method solely due to its interpretability. Recent work has argued against the use of models simply because they are “interpretable” [25, 45] and highlights the importance of validating claims surrounding interpretability by user studies [42, 60]. Interesting directions for future work include: learning checklists for risk prediction [70]; optimizing the order of items to minimize response time; and studying interventions that further improve predictive performance in settings where checklists fail to perform well.
Acknowledgments and Disclosure of Funding

We acknowledge a Wellcome Trust Data for Science and Health (222914/Z/21/Z) and NSERC (PDF-516984) grant. Resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute. Dr. Marzyeh Ghassemi is funded in part by Microsoft Research, and a Canadian CIFAR AI Chair held at the Vector Institute. We would like to thank Taylor Killian, Bret Nestor, Aparna Balagopalan, and four anonymous reviewers for their valuable feedback.

References


[45] Zachary C Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018.


