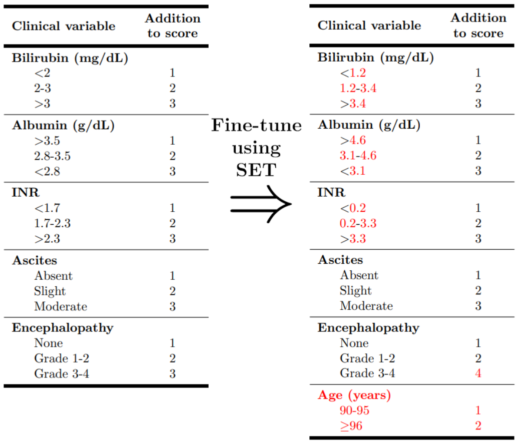
Multi-Objective Fine-Tuning of Clinical Scoring Tables: Adapting to Variations in Demography and Data

Kei Sen Fong FONGKEISEN@U.NUS.EDU
Department of Electrical and Computer Engineering, National University of Singapore

Mehul Motani MOTANI@NUS.EDU.SG
Department of Electrical and Computer Engineering, Institute of Data Science, N.1 Institute for Health, Institute for Digital Medicine (WisDM), National University of Singapore

Abstract

Clinical scoring tables (e.g., CURB-65 for pneumonia severity and mortality estimation) are widely used for estimating outcomes in healthcare, but their applicability is limited by i) demographic variations, ii) incomplete data availability of clinical variables, or iii) the need to incorporate data of new cohort-relevant clinical variables. We introduce a novel constrained multi-objective evolutionary machine learning (ML) optimization framework, **SET** (**S**coring-table **E**volutionary **T**uning), that fine-tunes established clinical scoring tables to enhance performance while maintaining familiarity. SET works by iteratively making small constrained changes to the original table to improve performance across multiple metrics, while maintaining a similar structure, ensuring that minimal adjustments are made. This is in contrast to ML-based proposals that replace scoring tables with entirely new models or tables, which may encounter barriers to clinical adoption. Extensive evaluations across 8 established scoring tables and cohorts demonstrate that SET allows existing clinically-trusted scoring tables to adapt to variations in demography, enhancing performance. We also show that in situations with incomplete data availability of key clinical variables, SET can still augment scoring tables and perform competitively. Additionally, SET can also augment existing tables to incorporate new cohort-relevant features.


Fine-tune
using
SET
⇒

Clinical variable	Addition to score
Bilirubin (mg/dL)	
<2	1
2-3	2
>3	3
Albumin (g/dL)	
>3.5	1
2.8-3.5	2
<2.8	3
INR	
<1.7	1
1.7-2.3	2
>2.3	3
Ascites	
Absent	1
Slight	2
Moderate	3
Encephalopathy	
None	1
Grade 1-2	2
Grade 3-4	3

Clinical variable	Addition to score
Bilirubin (mg/dL)	
<1.2	1
1.2-3.4	2
>3.4	3
Albumin (g/dL)	
>4.6	1
3.1-4.6	2
<3.1	3
INR	
<0.2	1
0.2-3.3	2
>3.3	3
Ascites	
Absent	1
Slight	2
Moderate	3
Encephalopathy	
None	1
Grade 1-2	2
Grade 3-4	4
Age (years)	
90-95	1
≥96	2

Figure 1: SET fine-tunes established clinical scoring tables, optimizing discrimination, calibration and reclassification objectives concurrently. See full-size examples in Appendix C, D, E for producing fine-tuned with the same variables, missing variables and additional variables, respectively.

Data and Code Availability Data and code used in this work are available to other researchers. This paper works on eight scores: Child-Pugh score (Child and Turcotte, 1964), COVID in-hospitality mortality score (CIMS) (Dueñas-Espín et al., 2023), CURB-65 score (Lim et al., 2003), GRACE score (Fox et al., 2006), HAS-BLED score (Pisters et al., 2010), LACE score (Van Walraven et al., 2010), NEWS (RCoP, 2012), and PSI score (Fine et al., 1997). Data used to evaluate and fine-tune each of these eight scores is available in Chang et al. (2020); Dueñas-Espín et al. (2023); Millman et al. (2017); Zhu et al. (2020); AlAmmari et al. (2021); Robinson and Hudali (2017); Mitsunaga et al. (2019); Chang et al. (2024), respectively. Code is available at https://github.com/kentridgeai/SET.

1. Introduction

Clinical scoring tables play a fundamental role in medical decision-making, functioning as transparent and interpretable tools for estimating outcomes.

Widely adopted scores such as CURB-65 (Lim et al., 2003) for pneumonia severity and mortality estimation and GRACE (Fox et al., 2006) for cardiovascular risk and mortality prediction in acute coronary syndrome, help guide clinicians in prognosticating outcomes. However, their applicability across diverse populations is often limited due to demographic variations (Pennells et al., 2024), incomplete data availability of key clinical variables (Le Gal et al., 2006), and the need to incorporate data of new cohort-relevant features (Raffa et al., 2022). While traditional machine learning (ML) models are alternatives to clinical scoring tables that can address the limitations above, they often lack familiarity, making them challenging to integrate into clinical workflows where interpretability and trust are paramount (Clark et al., 2023; Muralidharan et al., 2024).

In this work, we recommend utilizing fine-tuned versions of established, interpretable clinical scoring tables (see Figure 1). We mean fine-tune in the sense it is used in neural networks, where a fine-tuned network starts with numerical parameters derived from an initial task or dataset, followed by updates to those parameters for a subsequent task or dataset. We introduce **SET** (**S**coring-table **E**volutionary **T**uning), a widely-applicable, multi-purpose and robust framework that fine-tunes established clinical scoring tables to enhance performance while maintaining familiarity (see fine-tuned tables in Appendix C, D, E). However, a key challenge in fine-tuning these scoring tables is the high dimensionality of their components. Each table consists of multiple scoring parameters and clinical variable thresholds that influence the final score. Given a scoring table with n individual scoring components and m candidate scores per component, the total number of possible configurations is exponential, approximately $O(m^n)$, despite excluding optimizing the thresholds. A simple exhaustive search to optimize these parameters is computationally impractical, even with modern high-performance computing, necessitating an alternative approach to explore the search space.

Another key challenge is achieving an optimal balance across multiple performance dimensions, including discrimination, calibration and reclassification. Discrimination metrics evaluate a model's ability to distinguish between individuals with and without the outcome of interest, while calibration metrics assess the alignment between predicted probabilities and observed event rates. Specific to comparing between scores, reclassification metrics quantify the improvement in risk stratification compared to a baseline model. Ensuring a harmonious balance among these metrics is essential for the development of robust and clinically meaningful scoring tables.

To address these challenges, we leverage multi-objective evolutionary optimization, specifically using PyGAD (Gad, 2024), a genetic algorithm (GA)-based library. Evolutionary optimization is well-suited for this task because it efficiently searches high-dimensional, constrained solution spaces without requiring explicit gradient information (Zames, 1981; Mitchell, 1998), and supports optimizing of multiple objectives (i.e., discrimination, calibration and reclassification) concurrently.

We also add constraints, such as starting with an initial population filled with the original scoring table and include several penalty mechanisms in place to ensure that solutions remain clinically familiar while maximizing predictive performance. Since practicing clinicians trust and rely on established clinical scoring tables, the philosophy behind our framework is to introduce minimal but meaningful refinements rather than replacing them entirely, aligning with the core principle of familiarity-a critical factor in clinical adoption (Clark et al., 2023).

The **main contributions** of this paper are:

1. We propose a novel constrained multi-objective evolutionary optimization framework, SET, for fine-tuning clinical scoring tables, achieving higher performance by adapting to specific patient cohorts while preserving familiar structure and behavior.
2. We evaluate SET across multiple well-established clinical scoring tables and diverse patient cohorts, showing that the fine-tuned tables achieve improved performance from multiple perspectives (i.e., discrimination, calibration and reclassification metrics) while maintaining familiarity.
3. We demonstrate how SET enables dynamic feature adaptability, allowing the incorporation of cohort-relevant clinical variables and the removal of unavailable features, ensuring practical applicability across different healthcare settings without disrupting existing workflows.

2. Related Work

The refinement of existing clinical scoring tables has been studied independently in various isolated contexts and significant gaps remain in having a universal way to adapt these tables to specific cohorts. Our work builds upon studies in the following key areas.

2.1. Traditional Clinical Scoring Tables and Rescoring

Clinical scoring tables have been widely used for risk stratification and estimating outcomes. These tables provide interpretable decision-support tools that clinicians trust due to their simplicity and empirical validation. Each of these scoring tables are designed for a specific clinical purpose. 8 clinical scoring tables are explored in this paper and are available in Appendix C. Despite their widespread use, scoring tables are often developed in specific populations and may not generalize well to new cohorts due to demographic variability (Pennells et al., 2024), differences in clinical practice, and variations in data (Le Gal et al., 2006; Raffa et al., 2022).

Rescoring[1] of clinical scores have been done (Brudvik et al., 2019; Lee et al., 1999), but many require making an entirely new table, or require domain-specific judgment on the selection of parameters that are difficult to replicate. These methods can be both labor-intensive and suboptimal. Moreover, traditional rescoring approaches may not fully capture the complex relationships between clinical variables and outcomes, particularly when aspects are involved (i.e., discrimination, calibration, reclassification). In contrast, our framework employs constrained multi-objective evolutionary optimization to fine-tune scoring tables in a data-driven manner, ensuring a more flexible and computationally efficient adaptation process while preserving interpretability and clinical usability. Importantly, our framework, SET, has a universal applicability, as we show later in our experiments across a diverse range of scores and cohorts.

2.2. Machine Learning for Clinical Decision-Making

While typical ML models often achieve high predictive accuracy, they suffer from a lack of interpretability, making them less practical for clinical decision-making where explainability is crucial. Several studies have attempted to enhance ML interpretability through feature importance analysis and rule-based approximations, but these approaches still fall short of the inherent transparency offered by clinical scoring tables, evidenced by their low adoption rate (Clark et al., 2023; Muralidharan et al., 2024). Our framework attempts to bridge the gap between ML and adoption by incorporating data-driven optimization within the constraints of established scoring models, ensuring both improved predictive performance and interpretability. Furthermore, unlike building a score from scratch or using existing automated scoring frameworks (such as in Li et al. (2022); Xie et al. (2020)), our method attempts to maintain clinician trust by refining existing expert-crafted frameworks rather than replacing them entirely.

Another aspect to consider is that a fundamental requirement in clinical scoring is maintaining consistency in variable-output relationships, ensuring that predictions align with established medical knowledge. For example, for many domains, an increase in age should correspond to an increased or unchanged risk score for mortality, assuming all other factors remain the same. However, most black-box ML models are unable to strictly guarantee such behavior.

Beyond monotonic relationships, some clinical scoring tables exhibit more intricate variable-output patterns. For instance, in NEWS, temperature influences risk in a quadratic manner: extremely low and extremely high temperatures both correspond to a higher risk score, while mid-range temperatures are less concerning. Standard ML models lack explicit mechanisms to ensure a strictly quadratic behavior (i.e., a patient with an extremely low temperature may score higher than another patient with even lower temperature, all other variables being equal). In contrast, our approach adheres to the expert-crafted patterns developed by clinicians, ensuring that scoring refinements preserve expected clinical relationships while allowing data-driven improvements.

Finally, from different perspective, SET can also be viewed as an extension of Symbolic Regression (SR) that treats clinical scoring tables as equations. Some related work which uses SR for medical problems have been done (Fong and Motani, 2024b,a), but none of which address clinical scoring tables.

2.3. Multi-Objective Evolutionary Optimization

Evolutionary optimization has been widely used in various domains to solve complex search and optimization problems. GA provide an effective means of exploring high-dimensional parameter spaces while adhering to predefined constraints (Zames, 1981; Mitchell, 1998). Given the combinatorial explosion associated with optimizing clinical scoring tables, an exhaustive search approach is computationally prohibitive. GA is also extremely well-suited in this

[1] See Appendix A for clarification against similar terms

task because it i). exploits the multi-objective nature of finding improved clinical scoring tables and ii). allows for navigation on non-differentiable objectives (e.g. AUC). Our framework leverages PyGAD, a GA-based optimization library, to efficiently fine-tune clinical scoring tables while preserving their interpretability and adherence to medical guidelines (Gad, 2024). The optimization process in GAs is also effective in optimizing multiple objectives, in contrast to alternative methods like combining the various objective into a single value (e.g., linear combination of objectives) (Deb et al., 2002). Due to space constraints, we leave further details to Appendix K and Deb et al. (2002).

3. Evaluation Metrics

In this section, we detail the metrics used to assess the i) discrimination, ii) calibration and iii) reclassification ability of our predictive models. These will be essential in explaining our SET framework.

3.1. Notation and Common Symbols

For clarity and consistency throughout this paper, we adopt the following notation:

- N: Total number of subjects,
- i: Index for a subject, where $i = 1, \ldots, N$,
- $y_i \in \{0, 1\}$: True binary outcome for subject i, with $y_i = 1$ indicating the occurrence of the event,
- S_i: Continuous prediction score for subject i, as generated by the clinical model,
- $\hat{p}_i$: Calibrated prediction probability for subject i (in this work, this is computed by grouping subjects with identical scores and taking the observed frequency of the event),
- TP(c), TN(c), FP(c), FN(c): The counts of true positives, true negatives, false positives, and false negatives at a given score threshold c, respectively,
- PPV(c) and NPV(c): The positive and negative predictive values at score threshold c, respectively,
- G: The number of groups (bins) used in calibration analyses (in this work $G = 10$),
- baseline: Original clinical scoring table before fine-tuning.

3.2. Discrimination Metrics

Discrimination metrics quantify a model's ability to differentiate between subjects with and without the outcome.

Area Under the ROC Curve (AUC) (Metz et al., 1998): For a pair of subjects i and j with $y_i = 1$ and $y_j = 0$, the AUC is defined as

$$\mathrm{AUC} = \Pr\Big(S_i > S_j \mid y_i = 1,\ y_j = 0\Big) + \frac{1}{2}\Pr\Big(S_i = S_j \mid y_i = 1,\ y_j = 0\Big).$$

Youden Index (Youden, 1950): At a given score threshold c, let Sensitivity(c) $= \frac{\mathrm{TP}(c)}{\mathrm{TP}(c)+\mathrm{FN}(c)}$, Specificity($c$) $= \frac{\mathrm{TN}(c)}{\mathrm{TN}(c)+\mathrm{FP}(c)}$. Then the Youden index is defined as $J(c) = \mathrm{Sensitivity}(c) + \mathrm{Specificity}(c) - 1$. The **optimal score threshold** c^* is defined as the value that maximizes $J(c)$, consistent with literature (Hirschfeld and do Brasil, 2014; Lai et al., 2012).

3.3. Calibration Metrics

Calibration metrics assess the agreement between prediction probabilities and observed outcomes.

Brier Score (Brier, 1950): The Brier score is given by

$$\text{Brier Score} = \frac{1}{N}\sum_{i=1}^{N}(\hat{p}_i - y_i)^2.$$

Expected Calibration Error (ECE) (Naeini et al., 2015): Divide the prediction probability interval $[0, 1]$ into G bins. In this work $G = 10$. For bin g, let n_g be the number of subjects, $\overline{p}_g$ be the mean of all $\hat{p}_i$ in the group, and $\overline{y}_g$ be the observed rate of events in the group. The ECE is given by

$$\mathrm{ECE} = \sum_{g=1}^{G}\frac{n_g}{N}\left|\overline{p}_g - \overline{y}_g\right|.$$

Hosmer-Lemeshow Test (Hosmer and Lemesbow, 1980): Divide the prediction probability interval $[0, 1]$ into G groups. In this work $G = 10$. For group g, let O_g denote the observed number of events and E_g the predicted expected number of events, with group size n_g. The Hosmer-Lemeshow statistic is then defined as

$$C = \sum_{g=1}^{G}\frac{(O_g - E_g)^2}{E_g\left(1 - \frac{E_g}{n_g}\right)},$$

with degrees of freedom $df = G - 2$. The corresponding p-value is given by $p = 1 - F_{\chi^2}(C;\, df)$, where F_{χ^2} denotes the cumulative distribution function of the chi-square distribution.

3.4. Reclassification Metrics

Reclassification metrics compare the candidate model against a baseline model to assess improvements in risk classification.

Net Reclassification Improvement (NRI) (Pencina et al., 2008): Let $p_i^{(b)}$ and $p_i^{(c)}$ denote the baseline and candidate prediction probabilities, respectively. Define $\mathbb{I}\{\cdot\}$ as the indicator function, which equals 1 if the condition is true and 0 otherwise. Also, let $N_1 = \sum_{i=1}^{N} \mathbb{I}\{y_i = 1\}$ and $N_0 = \sum_{i=1}^{N} \mathbb{I}\{y_i = 0\}$. For subjects with events $(y_i = 1)$,

$$\mathrm{NRI}_1 = \frac{1}{N_1} \sum_{i:y_i=1} \left[\mathbb{I}\{p_i^{(c)} > p_i^{(b)}\} - \mathbb{I}\{p_i^{(c)} < p_i^{(b)}\} \right],$$

and for subjects without events $(y_i = 0)$,

$$\mathrm{NRI}_0 = \frac{1}{N_0} \sum_{i:y_i=0} \left[\mathbb{I}\{p_i^{(c)} < p_i^{(b)}\} - \mathbb{I}\{p_i^{(c)} > p_i^{(b)}\} \right].$$

The overall NRI is given by $\mathrm{NRI} = \mathrm{NRI}_1 + \mathrm{NRI}_0$.

Integrated Discrimination Improvement (IDI) (Pencina et al., 2008): Define the Yates discrimination slopes (Yates, 1982) for the baseline and candidate models as

$$\Delta^{(b)} = \overline{p}_1^{(b)} - \overline{p}_0^{(b)}, \quad \Delta^{(c)} = \overline{p}_1^{(c)} - \overline{p}_0^{(c)},$$

where $\overline{p}_1^{(b)}$ and $\overline{p}_0^{(b)}$ are the mean prediction probabilities for events and non-events in the baseline model, respectively, and $\overline{p}_1^{(c)}$ and $\overline{p}_0^{(c)}$ are defined similarly for the candidate model. Then, $\mathrm{IDI} = \Delta^{(c)} - \Delta^{(b)}$.

3.5. Summary of Metrics

Table 1 summarizes the evaluation metrics defined in Section 3.2 to Section 3.4 and includes theoretical ranges and preferred directions for ease-of-interpretation of results.

Metric	Range	Direction
Discrimination		
AUC	$[0, 1]$	Higher better
Youden Index	$[-1, 1]$	Higher better
Sensitivity	$[0, 1]$	Higher better
Specificity	$[0, 1]$	Higher better
PPV	$[0, 1]$	Higher better
NPV	$[0, 1]$	Higher better
Calibration		
Brier Score	$[0, 1]$	Lower better
ECE	$[0, 1]$	Lower better
Hosmer-Lemeshow p-value	$[0, 1]$	(> 0.05) preferred
Reclassification		
NRI	Depends	Positive preferred
IDI	Depends	Positive preferred

Table 1: Summary of evaluation metrics, including theoretical ranges and preferred directions.

4. SET: Scoring-table Evolutionary Tuning

In this section, we introduce **SET** (**S**coring-table **E**volutionary **T**uning), our proposed optimization framework that fine-tunes existing clinical scoring tables using constrained multi-objective evolutionary optimization. SET is designed to optimize the multiple performance metrics discussed in Section 3.2 to Section 3.4 while preserving the familiar structure of clinically established scoring tables.

Algorithm 1 provides a summary of the SET framework, with additional details discussed below. The optimization process begins by loading clinical data that has clinical variables (e.g., length of stay in LACE score) and an outcome (e.g., 30 days readmission in LACE score), which is then split into training and test sets while maintaining class stratification. This ensures that the dataset remains representative and that the model generalizes well to unseen data. We use a 60-40 train-test split in this work for all experiments. The next step is defining the score to tune, denoted as CST_W, where W are numerical parameters representing the score per component and clinical variable thresholds used. Note that yes/no in the scores are not modified to preserve the directionality with respect to the clinical variable, though the magnitude of the addition to score can still change. We also call W the baseline numerical parameters, in which we show a specific example in brown font in Table 2.

However, though possible, it would be naive to represent W in its raw form as shown in Table 2 consisting a vector of 14 numbers (i.e., $[1, 30, 2, 90, 60, 1, 1, 2, 1, 3, 5, 2, 6, 4]$ from traversing the Table 2 topmost, leftmost first). Notice that in "Number of ED visits", there are two interesting characteristics. First, the ranges 1-2, 3-5, ≥ 6 are contiguous. Second, the score for each range is strictly increasing, which is a behavior we want to preserve. Our strategy is to represent these sections as a concise summation of weights multiplied by a

Algorithm 1 SET Framework (Code available in Supplementary Materials)

```
Input: Training dataset $\mathcal{D}$, selected clinical scoring table to tune $\text{CST}_W(\cdot)$, baseline numerical parameters $W$ (i.e., score per component and clinical variable thresholds), GA parameters $\Theta$
Output: Fine-tuned parameters $\hat{W}$ (to get $\text{CST}_{\hat{W}}(\cdot)$)
Compute baseline scores by evaluating:
    $\text{Scores}_{base} \leftarrow \text{CST}_W(\mathcal{D})$
Evaluate baseline performance metrics $M_{\text{base}}$ (e.g., AUC, Brier score, NRI) from $\text{Scores}_{base}$
Define the fitness function $F(\mathbf{x})$ for candidate numerical parameters $\mathbf{x}$ which takes the steps:
    (a) Compute scores: $S_{\text{cand}} \leftarrow \text{CST}_{\mathbf{x}}(\mathcal{D})$
    (b) Compute Active AUC (see Section 3.2):
        $\Delta_{\text{AUC}} \leftarrow \text{AUC}(S_{\text{cand}}, \mathcal{D}) - \text{AUC}(S_{\text{base}}, \mathcal{D})$
    (c) Compute Active Brier (see Section 3.3):
        $\Delta_{\text{Brier}} \leftarrow \text{Brier}(S_{\text{base}}, \mathcal{D}) - \text{Brier}(S_{\text{cand}}, \mathcal{D})$
    (d) Compute NRI (see Section 3.4):
        $\text{NRI} \leftarrow \text{NRI}_1 + \text{NRI}_0$
    (e) Compute penalty $P(\mathbf{x})$ for constraint violations and deviation from $W$.
    (f) Return the penalized discrimination, calibration and reclassification objectives, where $\sigma$ is the ReLU function:
        $\left[\sigma(\Delta_{\text{AUC}} - P(\mathbf{x})), \sigma(\Delta_{\text{Brier}} - P(\mathbf{x})), \mathbb{I}(\text{NRI})\right]$
Initialize the GA population with only copies of $W$
Run the GA using PyGAD (following NSGA-II (Deb et al., 2002) selection policy) with fitness function $F(\cdot)$ and $\Theta$ (e.g.,, number of generations, population size)
return the best candidate solution $\hat{W}$ (to get $\text{CST}_{\hat{W}}(\cdot)$)
```

boolean condition with a single variable inequality. Thus, we structure the "Number of ED visits" section in Table 2 as $1{*}(ED \geq 1){+}1{*}(ED \geq 3){+}2{*}(ED \geq 6)$ instead, where ED is the variable denoting the number of ED visits. This way, instead of representing the numbers in the section as $[..., 1, 2, 1, 3, 5, 2, 6, 4]$, we can represent it as $[..., 1, 1, 1, 3, 2, 6]$, reducing the size of W from 8 to 6. More importantly, as long as the values chosen are greater than one, the clinical variable will always be treated contiguously, and the score for each range strictly increasing. There are many clinical variables that are included in clinical scoring tables in the way "Number of ED visits" is included in Table 2 (e.g., see "Bilirubin" in Table 11). Thus, it is important to employ the strategy of representing such sections as a concise summation of weights multiplied by a boolean condition with a single variable inequality.

Clinical variable	Addition to score
Confusion	
Yes	1
Respiratory rate (breaths/min)	
≥30	2
Blood pressure (mmHg)	
SBP <90 or DBP ≤60	1
Number of ED visits	
1-2	1
3-5	2
≥6	4

Table 2: Baseline numerical parameters shown in brown (this score is made-up for example).

After extracting W from the original chosen clinical scoring table, the next step is to initialize the GA with the right settings. The GA is initialized with a fitness function that evaluates candidate solutions based on three key objectives: discrimination, calibration, and reclassification. Discrimination is quantified as the improvement in AUC, denoted as Δ_{AUC}, which we also term as Active AUC. Calibration is measured by the improvement in the Brier score, Δ_{Brier}, which we also term as Active Brier. Reclassification is assessed through the Net Reclassification Index (NRI). These three metrics were selected to have a diversity of metrics, each looking at an orthogonal criterion.

To prevent over-optimization of any single objective at the cost of others, we apply the rectified linear unit (ReLU) function to Δ_{AUC} and Δ_{Brier} to ensure that improvements are only considered when they positively contribute to overall performance. This transformation is designed with the effects of crowding distance (Deb et al., 2002) in NSGA-II in mind, preventing solutions that provide negligible gains from being overly favored. We deemed reclassification to be the least important metric since it is least granular, thus, we binarized it via the indicator function, $\mathbb{I}$, which prevents NRI from dominating the selection pressure, but still allows it to identify candidate solutions with a net positive reclassification.

To ensure clinical validity, a penalty function $P(\mathbf{x})$ is introduced. Constraints enforce the logical ordering of thresholds to maintain clinical coherence, ensuring that risk scores increase or decrease in accordance with medical knowledge. For example in the earlier example of $1 * (ED \geq 1) + 1 * (ED \geq 3) + 2 * (ED \geq 6)$, the thresholds are $t_1 = 1, t_2 = 3, t_3 = 6$, in which after tuning, $t_1 < t_2 < t_3$ should still be preserved. Additionally, deviation constraints penalize excessive departures from the original scoring structure, preserving interpretability. The penalty function is structured as follows:

$$P(\mathbf{x}) = \lambda_1 \sum_k \max(0, t_k - t_{k+1} + 1) + \lambda_2 E_l \left[\left| \frac{\hat{W}_l - W_l}{W_l} \right| \right]$$

where λ_1 and λ_2 control the magnitude of penalties applied to threshold violations and parameter (i.e., thresholds, weights) deviations, respectively.

During each iteration of the GA, candidate solutions are generated through mutation and crossover operations. Each candidate $\mathbf{x}$ is evaluated using:

$$F(\mathbf{x}) = [\sigma(\Delta_{\text{AUC}} - P(\mathbf{x})), \sigma(\Delta_{\text{Brier}} - P(\mathbf{x})), \mathbb{I}(\text{NRI})]$$

where σ is the ReLU function.

To guide the search for optimal solutions, the GA is configured with the parameters Θ that has a population size of 500 and runs for 100 generations. The NSGA-II selection rule is employed to maintain diversity among solutions and efficiently explore the Pareto front of optimal trade-offs among competing objectives. The initial population is also seeded with copies of the original scoring table, ensuring refinements remain clinically familiar rather than diverging into entirely new and potentially uninterpretable models. This way, the variable-outcome patterns that are hand-crafted by experts are preserved.

Once the algorithm converges, the best candidate solution is selected based on its overall performance across discrimination, calibration, and reclassification metrics while adhering to the imposed constraints and eventually based on the crowding distance selection rules in NSGA-II (alternatively, we note that a practitioner also has an option to return the Pareto front of tables from NSGA-II instead). The optimized (fine-tuned) scoring table is then evaluated on the test set to verify its performance against the original baseline.

Through its structured optimization approach, SET enables fine-tuning of clinical scoring tables for specific patient cohorts, ensuring that they remain effective in diverse healthcare settings. The combination of evolutionary search, penalty-based constraints, and multi-objective optimization makes SET a powerful tool for adapting clinical risk models while preserving the transparency required for adoption in real-world medical decision-making.

5. Results and Discussion

To evaluate the effectiveness of SET, we conduct experiments on eight established clinical scoring tables: Child-Pugh, CIMS, CURB-65, GRACE, HAS-BLED, LACE, NEWS, and PSI. For each score, we utilized a unique cohort dataset, as described in the "Data and Code Availability" paragraph. By leveraging a diverse set of scoring tables and datasets encompassing a broad range of clinical conditions and demographic variations, we aimed to ensure the robustness and generalizability of our findings. The objective of these experiments is to determine whether SET can enhance the predictive performance of these scoring tables while preserving their interpretability and clinical usability.

5.1. Experiment Details

Each dataset was preprocessed to retain only relevant clinical variables required for scoring. We partitioned each dataset into 60-40 training and test sets in a stratified manner to preserve the original class distributions. The training data was used for optimizing the scoring tables via SET, while the test data was used for performance evaluation.

The GA, using PyGAD's implementation, was configured with a population size of 500, a mutation probability of 0.2, 100 mating parents per generation, ran for 100 generations and used the NSGA-II selection strategy. The fitness function simultaneously optimized three objectives and the initial population was filled with copies of the original scoring tables as a starting point, as detailed in Section 4. The candidate numerical parameters were constrained as well. For simplicity and consistency, all candidate numerical parameters were constrained to take on an integer range from 1 to $\lceil 1.5W \rceil$, ensuring that they do not deviate too much from the original clinical scoring table's parameters W. For the penalty function, we set $\lambda_1 = 1$ to strongly penalize violations of the threshold order, and λ_2 was selected among $[0.01, 0.05, 0.1, 0.5, 1]$ to account for the difference in the magnitude of performance metrics across the different clinical scores and datasets.

5.2. SET with Same Clinical Variables

We start by first analyzing using SET for fine-tuning clinical scoring tables using the exact same variable set, meaning that the clinical variables and structure of the table remain the same (see examples in Tables

Table 3: Performance metrics of original score and SET fine-tuned score on the 40% test set. For all metrics except c^*, Brier and ECE, the higher the better. For Brier and ECE, the lower the better. Best performances are bolded. SET demonstrates improved performances across most metrics across all 8 clinical scores. See Tables 43 & 44 for confidence intervals and significance tests.

	Discrimination							Calibration			Reclassification	
	AUC	c^*	Youden Index	Sensitivity	Specificity	PPV	NPV	Brier	ECE	HL p-value	NRI	IDI
Child-Pugh												
Original	0.615	9	0.193	0.500	**0.693**	**0.127**	0.940	0.0750	**0.00628**	**1.00**	0.00%	0.00%
SET	**0.654**	10	**0.251**	**0.700**	0.551	0.122	**0.954**	**0.0741**	0.0154	0.997	**+33.3%**	**+0.830%**
CIMS												
Original	0.818	43	0.487	0.789	**0.697**	**0.438**	0.917	**0.139**	**0.0198**	0.544	0.00%	0.00%
SET	**0.822**	43	**0.501**	**0.833**	0.668	0.429	**0.930**	**0.139**	**0.0198**	**0.590**	**+14.8%**	**+0.660%**
CURB-65												
Original	0.708	1	0.362	**0.769**	0.592	0.174	0.958	0.0837	0.0268	0.0402	0.00%	0.00%
SET	**0.712**	1	**0.387**	**0.769**	**0.618**	**0.183**	**0.960**	**0.0836**	**0.0262**	**0.0415**	**+36.7%**	**+0.360%**
GRACE												
Original	0.786	167	0.294	0.500	**0.794**	0.222	0.931	0.122	0.112	**0.998**	0.00%	0.00%
SET	**0.841**	152	**0.529**	**0.750**	0.779	**0.286**	**0.964**	**0.0987**	**0.0921**	0.981	**+26.5%**	**+13.2%**
HAS-BLED												
Original	0.687	4	0.293	0.767	**0.526**	0.173	0.946	0.0997	**0.00692**	**0.992**	0.00%	0.00%
SET	**0.700**	4	**0.389**	**0.918**	0.471	**0.184**	**0.978**	**0.0988**	0.0179	0.963	**+4.74%**	**+2.07%**
LACE												
Original	0.676	10	0.221	**1.00**	0.221	0.0946	**1.00**	0.0730	0.0693	0.115	0.00%	0.00%
SET	**0.688**	9	**0.349**	**1.00**	**0.349**	**0.111**	**1.00**	**0.0690**	**0.0231**	**0.984**	**+46.7%**	**+3.21%**
NEWS												
Original	0.536	9	0.0151	0.187	**0.828**	0.446	0.579	0.248	0.0537	0.0510	0.00%	0.00%
SET	**0.540**	9	**0.0412**	**0.272**	0.769	**0.466**	**0.588**	**0.247**	**0.0335**	**0.169**	**+14.2%**	**+0.666%**
PSI												
Original	0.825	94	0.478	0.726	**0.752**	**0.358**	0.935	0.144	0.0892	**0.00**	0.00%	0.00%
SET	**0.828**	80	**0.482**	**0.877**	0.606	0.298	**0.963**	**0.136**	**0.0881**	**0.00**	**+7.44%**	**+6.12%**

4, 5 and Appendix C). Table 3 presents a comparison between the original and fine-tuned versions of each scoring table using the same clinical variables. Across all eight scoring tables, SET consistently improved key performance metrics on the test set. The observed improvements were particularly strong in discrimination, with AUC increasing across all scores. Notably, GRACE demonstrated a substantial AUC improvement from 0.786 to 0.841. In terms of calibration, SET fine-tuned version of scores had the best Brier values in all 8 scores and improved ECE values across multiple scores suggest that the fine-tuned models align more closely with observed event rates, making risk estimations more reliable. Finally, using the reclassification metrics, which helps verify if patients are being classified better compared to the original score, SET fine-tuned scoring tables always had a net positive improvement.

These results highlight SET's capability to fine-tune scoring tables to specific cohorts without disrupting their original structure, making it easier to integrate fine-tuned versions into existing clinical workflows. The improvement across multiple metrics at once also demonstrates the effectiveness in multi-objective optimization, which is difficult to obtain with single-objective approaches that tend to over-optimize or overfit to a single objective while sacrificing the performance on the remaining objectives. **SET simultaneously enhances discrimination, calibration, and reclassification of existing scores to a specific-cohort while preserving familiarity.**

5.3. SET for Simplification of Tables

Another advantage of SET is its ability to simplify clinical scoring tables while maintaining, or in some cases, improving performance. By following the same settings as earlier, but zero-ing out the parameters associated with an unavailable or removed clinical variable(s), SET is able to generate simplified clinical scoring tables. Examples of simplified tables are shown in Tables 6, 7 and Appendix D.

Table 7 illustrates how SET optimized CURB-65 by eliminating the urea component, resulting in a more streamlined version while still improving AUC from 0.584 to 0.664. The results in Table 8 further

Clinical variable	Addition to score
Length of stay (days)	
1	1
2	2
3	3
4-6	4
7-13	5
≥14	7
Acute (emergent) admission	
Yes	3
Charlson Comorbidity Index	
1	1
2	2
3	3
≥4	5
Number of ED visits within 6 months	
1	1
2	2
3	3
≥4	4

Table 4: **Before -** Original Baseline LACE Score.

Clinical variable	Addition to score
Length of stay (days)	
1	1
2	2
3-5	3
6-7	4
8-16	5
≥17	6
Acute (emergent) admission	
Yes	2
Charlson Comorbidity Index	
1	1
2-3	2
4-5	3
≥6	6
Number of ED visits within 6 months	
1-2	1
3	2
4-5	3
≥6	4

Table 5: **After -** Tuned LACE Score by SET.

demonstrate that SET-derived simplified tables generally performed on par with, or slightly better than, their original counterparts.

However, the extent of improvement was generally lower than when SET fine-tuned scores with the same clinical variables (see Section 5.2). This is likely because reducing input variables limits the model's capacity to leverage informative predictors, highlighting the trade-off between simplification and performance. Importantly, the impact of simplification is score-dependent. While the removal of urea from CURB-65 led to an improvement, reducing variables in HAS-BLED resulted in more modest gains, emphasizing that not all scoring tables can be simplified without potential loss of predictive power.

Clinical variable	Addition to score
Confusion	
Yes	1
Urea (mmol/L)	
>7	1
Respiratory rate (breaths/min)	
≥30	1
Blood pressure (mmHg)	
SBP <90 or DBP ≤60	1
Age (years)	
≥65	1

Table 6: **Before -** Original Baseline CURB-65 Score.

Clinical variable	Addition to score
Confusion	
Yes	1
Respiratory rate (breaths/min)	
≥27	1
Blood pressure (mmHg)	
SBP < 87 or DBP ≤37	1
Age (years)	
≥43	1

Table 7: **After -** Simplified CURB-65 Score without Urea by SET.

Despite these trade-offs, SET remains a powerful alternative to conventional missing-data handling approaches, such as discarding the score altogether or assigning arbitrary constant values to missing variables. In scenarios where clinical variables are unavailable due to cost, time constraints, or feasibility concerns, SET provides a structured optimization-driven approach that ensures the resulting scores remain both interpretable and clinically useful. **By adapting scoring tables to real-world constraints while safeguarding their predictive validity, SET can help bridge the gap between theoretical clinical models and the practical challenges of healthcare decision-making (e.g., unavailable clinical measurements).**

5.4. SET for Extension of Tables

SET also enables the extension of clinical scoring tables by incorporating additional variables, allowing for customization based on clinical intuition or emerging evidence. In practice, clinicians often identify new

Table 8: Performance metrics of original score and SET fine-tuned score on the 40% test set. For all metrics except c^*, Brier and ECE, the higher the better. For Brier and ECE, the lower the better. Best performances are bolded. See Tables 45 & 46 for confidence intervals and significance tests.

	Discrimination							Calibration			Reclassification	
	AUC	c^*	Youden Index	Sensitivity	Specificity	PPV	NPV	Brier	ECE	HL p-value	NRI	IDI
CIMS												
Original	0.809	42	**0.469**	0.751	**0.720**	**0.445**	0.906	**0.141**	0.0208	0.285	0.00%	0.00%
SET (Simplified)	**0.811**	38	**0.469**	**0.760**	0.710	0.439	**0.908**	**0.141**	**0.0166**	**0.565**	**+11.0%**	**+0.506%**
CURB-65												
Original	0.584	1	0.144	0.385	**0.760**	0.152	0.917	0.0882	0.0264	0.964	0.00%	0.00%
SET (Simplified)	**0.664**	1	**0.285**	**0.654**	0.631	**0.165**	**0.942**	**0.0876**	**0.0151**	**0.984**	**+29.3%**	**+1.62%**
GRACE												
Original	0.770	157	0.250	0.500	0.750	0.190	0.927	0.140	0.123	0.962	0.00%	0.00%
SET (Simplified)	**0.836**	147	**0.404**	**0.625**	**0.779**	**0.250**	**0.946**	**0.145**	**0.105**	**1.00**	**+2.94%**	**+0.490%**
HAS-BLED												
Original	0.620	5	**0.204**	0.384	**0.821**	**0.217**	0.911	**0.103**	**0.00462**	0.958	0.00%	0.00%
SET (Simplified)	**0.629**	3	0.160	**0.863**	0.297	0.137	**0.944**	**0.103**	0.00709	**1.00**	**+26.2%**	**+0.197%**
PSI												
Original	0.828	80	0.511	**0.863**	0.648	0.318	**0.961**	**0.138**	**0.0916**	**0.00**	**0.00%**	0.00%
SET (Simplified)	**0.833**	77	**0.514**	0.822	**0.692**	**0.337**	0.953	0.148	0.0983	**0.00**	-4.44%	**+0.254%**

potential risk factors through observations and experience, but integrating these variables into established scoring systems is challenging. SET provides a structured, data-driven approach to adjust the weight and threshold of newly introduced variables while maintaining the table's fundamental logic and usability. We do this by enforcing SET to have weights and thresholds that are non-zero for these new variables, even if performance decreases.

As demonstrated in Table 40 for the LACE score, we applied SET to extend existing scoring tables by injecting additional clinical variables, such as age and sex. Other extended tables for other scores are shown in Appendix E. In these tables, the added variables are chosen based on data availability rather than clinical evidence, which limits the performance results. The performance results in Table 10 show that while the extended scores did not always lead to substantial AUC improvements, reclassification and calibration metrics generally showed positive trends. This highlights SET's capability to integrate additional risk factors in a way that refines patient stratification without disrupting the performance and familiarity of the original scoring system.

These results underscore SET's potential as a clinician-guided optimization tool-allowing domain experts to introduce variables they deem important while ensuring that the final scoring system remains quantitatively validated. **By facilitating the seamless injection of new clinical variables into existing models, SET empowers clinicians to fine-tune and adapt scoring tables based on evolving medical knowledge and real-world observations.**

We also include multiple ablation studies in Appendix G and comparison with AdaBoost in Appendix H.

6. Conclusion

We introduce **SET** (**S**coring-table **E**volutionary **T**uning), a novel framework that fine-tunes clinical scoring tables to enhance predictive performance while preserving interpretability and clinical familiarity. Unlike black-box ML models, SET fine-tunes existing scores with minimal structural changes, to preserve expert-identified variable-outcome patterns and allow for seamless adoption.

Evaluations across 8 diverse clinical scoring tables and datasets demonstrate SET's ability to improve discrimination, calibration, and reclassification concurrently. Its versatility allows for i) fine-tuning existing scores, ii) simplifying tables when data is missing, and iii) integrating new risk factors, making it a powerful tool for adapting clinical models to real-world challenges.

SET bridges the gap between machine learning advancements and clinician trust, enabling scoring tables to remain relevant in evolving healthcare settings. Future work will focus on collaborating with domain-specific clinical experts to further refine SET with more expert-identified constraints to ensure that the optimized scoring tables better align with practical needs and enhance patient outcomes.

Institutional Review Board (IRB) This work does not require IRB approval.

Acknowledgments

This research/project is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG3-PhD-2023-08-052T), and A*STAR, CISCO Systems (USA) Pte. Ltd and National University of Singapore under its Cisco-NUS Accelerated Digital Economy Corporate Laboratory (Award I21001E0002).

References

Maha AlAmmari, Khizra Sultana, Abdulrahman Alturaiki, Abin Thomas, Monirah AlBabtain, Fakahr AlAyoubi, and Hanie Richi. The development and validation of a multivariable model to predict the bleeding risk score for patients with non-valvular atrial fibrillation using direct oral anticoagulants in the arab population. *Plos one*, 16(5):e0250502, 2021.

Glenn W Brier. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78 (1):1–3, 1950.

Kristoffer W Brudvik, Robert P Jones, Felice Giuliante, Junichi Shindoh, Guillaume Passot, Michael H Chung, Juhee Song, Liang Li, Vegar J Dagenborg, Åsmund A Fretland, et al. Ras mutation clinical risk score to predict survival after resection of colorectal liver metastases. *Annals of surgery*, 269(1):120–126, 2019.

Shu-Ching Chang, Gary L Grunkemeier, Jason D Goldman, Mansen Wang, Paul A McKelvey, Jennifer Hadlock, Qi Wei, and George A Diaz. A simplified pneumonia severity index (psi) for clinical outcome prediction in covid-19. *Plos one*, 19(5): e0303899, 2024.

Te-Sheng Chang, Ying-Huang Tsai, Yi-Heng Lin, Chun-Hsien Chen, Chung-Kuang Lu, Wen-Shih Huang, Yao-Hsu Yang, Wei-Ming Chen, Yung-Yu Hsieh, Yu-Chih Wu, et al. Limited effects of antibiotic prophylaxis in patients with child–pugh class a/b cirrhosis and upper gastrointestinal bleeding. *PloS one*, 15(2):e0229101, 2020.

CG Child and JG Turcotte. Surgery and portal hypertension. *Major Problems in Clinical Surgery*, 1: 1–85, 1964.

Phoebe Clark, Jayne Kim, and Yindalon Aphinyanaphongs. Marketing and us food and drug administration clearance of artificial intelligence and machine learning enabled software in and as medical devices: a systematic review. *JAMA Network Open*, 6(7):e2321792–e2321792, 2023.

Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and TAMT Meyarivan. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE transactions on evolutionary computation*, 6(2):182–197, 2002.

Iván Dueñas-Espín, María Echeverría-Mora, Camila Montenegro-Fárez, Manuel Baldeón, Luis Chantong Villacres, Hugo Espejo Cárdenas, Marco Fornasini, Miguel Ochoa Andrade, and Carlos Solís. Development and validation of a scoring system to predict mortality in patients hospitalized with covid-19: A retrospective cohort study in two large hospitals in ecuador. *Plos one*, 18(7):e0288106, 2023.

Michael J Fine, Thomas E Auble, Donald M Yealy, Barbara H Hanusa, Lisa A Weissfeld, Daniel E Singer, Christopher M Coley, Thomas J Marrie, and Wishwa N Kapoor. A prediction rule to identify low-risk patients with community-acquired pneumonia. *New England journal of medicine*, 336 (4):243–250, 1997.

Kei Sen Fong and Mehul Motani. Explainable and privacy-preserving machine learning via domain-aware symbolic regression. In *Conference on Health, Inference, and Learning*, pages 198–216. PMLR, 2024a.

Kei Sen Fong and Mehul Motani. Symbolic regression for discovery of medical equations: A case study on glomerular filtration rate estimation equations. In *2024 IEEE Conference on Artificial Intelligence (CAI)*, pages 1373–1379. IEEE, 2024b.

Keith AA Fox, Omar H Dabbous, Robert J Goldberg, Karen S Pieper, Kim A Eagle, Frans Van de Werf, Álvaro Avezum, Shaun G Goodman, Marcus D Flather, Frederick A Anderson, et al. Prediction of risk of death and myocardial infarction in the six months after presentation with acute coronary syndrome: prospective multinational observational study (grace). *bmj*, 333(7578):1091, 2006.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics*, 28(2):337–407, 2000.

Ahmed Fawzy Gad. Pygad: An intuitive genetic algorithm python library. *Multimedia tools and applications*, 83(20):58029–58042, 2024.

Gerrit Hirschfeld and Pedro Emmanuel Alvarenga Americano do Brasil. A simulation study into the performance of "optimal" diagnostic thresholds in the population:"large" effect sizes are not enough. *Journal of clinical epidemiology*, 67(4):449–453, 2014.

David W Hosmer and Stanley Lemesbow. Goodness of fit tests for the multiple logistic regression model. *Communications in statistics-Theory and Methods*, 9(10):1043–1069, 1980.

Chin-Ying Lai, Lili Tian, and Enrique F Schisterman. Exact confidence interval estimation for the youden index and its corresponding optimal cutpoint. *Computational statistics & data analysis*, 56(5):1103–1114, 2012.

Grégoire Le Gal, Marc Righini, Pierre-Marie Roy, Olivier Sanchez, Drahomir Aujesky, Henri Bounameaux, and Arnaud Perrier. Prediction of pulmonary embolism in the emergency department: the revised geneva score. *Annals of internal medicine*, 144(3):165–171, 2006.

Thomas H Lee, Edward R Marcantonio, Carol M Mangione, Eric J Thomas, Carisi A Polanczyk, E Francis Cook, David J Sugarbaker, Magruder C Donaldson, Robert Poss, Kalon KL Ho, et al. Derivation and prospective validation of a simple index for prediction of cardiac risk of major noncardiac surgery. *Circulation*, 100(10):1043–1049, 1999.

Anthony Li, Ming Lun Ong, Chien Wei Oei, Weixiang Lian, Hwee Pin Phua, Lin Htun Htet, and Wei Yen Lim. Unified auto clinical scoring (uni-acs) with interpretable ml models. In *Machine Learning for Healthcare Conference*, pages 26–53. PMLR, 2022.

Wei Shen Lim, Menno M Van der Eerden, R Laing, Wim G Boersma, Noel Karalus, George I Town, SA Lewis, and JT1746657 Macfarlane. Defining community acquired pneumonia severity on presentation to hospital: an international derivation and validation study. *Thorax*, 58(5):377–382, 2003.

Jiachang Liu, Chudi Zhong, Boxuan Li, Margo Seltzer, and Cynthia Rudin. Fasterrisk: fast and accurate interpretable risk scores. *Advances in Neural Information Processing Systems*, 35:17760–17773, 2022.

Charles E Metz, Benjamin A Herman, and Jong-Her Shen. Maximum likelihood estimation of receiver operating characteristic (roc) curves from continuously-distributed data. *Statistics in medicine*, 17(9):1033–1053, 1998.

Alexander J Millman, Adena Greenbaum, Sibongile Walaza, Adam L Cohen, Michelle J Groome, Carrie Reed, Meredith McMorrow, Stefano Tempia, Marietjie Venter, Florette K Treurnicht, et al. Development of a respiratory severity score for hospitalized adults in a high hiv-prevalence setting—south africa, 2010–2011. *BMC pulmonary medicine*, 17:1–8, 2017.

Melanie Mitchell. *An introduction to genetic algorithms*. MIT press, 1998.

Toshiya Mitsunaga, Izumu Hasegawa, Masahiko Uzura, Kenji Okuno, Kei Otani, Yuhei Ohtaki, Akihiro Sekine, and Satoshi Takeda. Comparison of the national early warning score (news) and the modified early warning score (mews) for predicting admission and in-hospital mortality in elderly patients in the pre-hospital setting and in the emergency department. *PeerJ*, 7:e6947, 2019.

Vijaytha Muralidharan, Boluwatife Adeleye Adewale, Caroline J Huang, Mfon Thelma Nta, Peter Oluwaduyilemi Ademiju, Pirunthan Pathmarajah, Man Kien Hang, Oluwafolajimi Adesanya, Ridwanullah Olamide Abdullateef, Abdulhammed Opeyemi Babatunde, et al. A scoping review of reporting gaps in fda-approved ai medical devices. *npj Digital Medicine*, 7(1):273, 2024.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

Michael J Pencina, Ralph B D'Agostino Sr, Ralph B D'Agostino Jr, and Ramachandran S Vasan. Evaluating the added predictive ability of a new marker: from area under the roc curve to reclassification and beyond. *Statistics in medicine*, 27(2):157–172, 2008.

Lisa Pennells, Stephen Kaptoge, and Emanuele Di Angelantonio. Adapting cardiovascular risk prediction models to different populations: the need for recalibration, 2024.

Ron Pisters, Deirdre A Lane, Robby Nieuwlaat, Cees B De Vos, Harry JGM Crijns, and Gregory YH Lip. A novel user-friendly score (has-bled) to assess 1-year risk of major bleeding in patients with atrial fibrillation: the euro heart survey. *Chest*, 138(5):1093–1100, 2010.

Jesse D Raffa, Alistair EW Johnson, Zach O'Brien, Tom J Pollard, Roger G Mark, Leo A Celi, David Pilcher, and Omar Badawi. The global open source severity of illness score (gossis). *Critical care medicine*, 50(7):1040–1050, 2022.

London RCoP. National early warning score (news): standardising the assessment of acute-illness severity in the nhs. *Report of working party. London: Royal College of Physicians*, 2012.

Robert Robinson and Tamer Hudali. The hospital score and lace index as predictors of 30 day readmission in a retrospective study at a university-affiliated community hospital. *PeerJ*, 5:e3137, 2017.

Berk Ustun and Cynthia Rudin. Learning optimized risk scores. *Journal of Machine Learning Research*, 20(150):1–75, 2019.

Carl Van Walraven, Irfan A Dhalla, Chaim Bell, Edward Etchells, Ian G Stiell, Kelly Zarnke, Peter C Austin, and Alan J Forster. Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community. *Cmaj*, 182(6):551–557, 2010.

Feng Xie, Bibhas Chakraborty, Marcus Eng Hock Ong, Benjamin Alan Goldstein, Nan Liu, et al. Autoscore: a machine learning–based automatic clinical score generator and its application to mortality prediction using electronic health records. *JMIR medical informatics*, 8(10):e21798, 2020.

J Frank Yates. External correspondence: Decompositions of the mean probability score. *Organizational Behavior and Human Performance*, 30(1):132–156, 1982.

William J Youden. Index for rating diagnostic tests. *Cancer*, 3(1):32–35, 1950.

G Zames. Genetic algorithms in search, optimization and machine learning. *Inf Tech J*, 3(1):301, 1981.

Houyong Zhu, Zhaodong Li, Xiaoqun Xu, Xiaojiang Fang, Tielong Chen, and Jinyu Huang. Predictive value of three inflammation-based glasgow prognostic scores for major cardiovascular adverse events in patients with acute myocardial infarction during hospitalization: a retrospective study. *PeerJ*, 8:e9068, 2020.

Appendix A. Differences between Recalibration, Reweighting and Rescoring

In this appendix, we clarify the distinctions between three similar terms: **recalibration**, **reweighting**, and **rescoring** in the context of clinical scoring tables.

Recalibration refers to updating the risk estimates associated with a given score, without modifying the structure, components, or weightings of the score itself. The primary goal of recalibration is to ensure that risk predictions align with clinical data. The numerical score remains unchanged but the estimated probability of an outcome per score group is adjusted.

Reweighting involves altering the contribution of individual variables within a scoring system. This is typically done when new evidence suggests that certain variables have more or less impact on the outcome than originally assumed. The components of the score remain the same but the assigned weight (point value or coefficient) of each component is modified.

Rescoring refers to modifying the overall scoring algorithm, which may include changes to the weightings, cutoff thresholds, or the inclusion or exclusion of specific variables.

Table 9 summarizes the key differences between **recalibration**, **reweighting**, and **rescoring**.

Aspect	**Recalibration**	**Reweighting**	**Rescoring**
Risk estimates updated	Yes	Yes	Yes
Weighting of components changed	No	Yes	Yes
Score cutoffs modified	No	No	Yes
Variables added or removed	No	No	Yes

Table 9: Comparison of recalibration, reweighting, and rescoring.

Appendix B. Performance of Extended Scoring Tables

Table 10: Discrimination, calibration and reclassification metrics performance between original score and SET fine-tuned score on the 40% test set. For all metrics except c^*, Brier and ECE, the higher the better. For Brier and ECE, the lower the better. Best performances are bolded.

	Discrimination							**Calibration**			**Reclassification**	
	AUC	c^*	Youden Index	Sensitivity	Specificity	PPV	NPV	Brier	ECE	HL p-value	NRI	IDI
Child-Pugh												
Original	0.615	9	0.193	0.500	**0.693**	**0.127**	0.940	**0.0750**	**0.00628**	**1.00**	**0.00%**	0.00%
SET (Extended)	**0.645**	10	**0.238**	**0.667**	0.571	0.122	**0.950**	**0.0750**	0.0112	0.998	-9.46%	**+0.0488%**
LACE												
Original	**0.676**	10	0.221	**1.00**	0.221	0.0946	**1.00**	0.0730	0.0693	0.115	0.00%	0.00%
SET (Extended)	0.654	16	**0.255**	0.714	**0.541**	**0.112**	0.959	**0.0716**	**0.0290**	**0.964**	**+27.0%**	**+2.38%**
NEWS												
Original	0.536	9	0.0151	0.187	**0.828**	0.446	0.579	0.248	0.0537	0.0510	0.00%	0.00%
SET (Extended)	**0.560**	10	**0.0635**	**0.312**	0.751	**0.481**	**0.596**	**0.242**	**0.00403**	**1.00**	**+15.3%**	**+1.56%**

Appendix C. Fine-tuned Tables

Clinical variable	Addition to score
Bilirubin (mg/dL)	
<2	1
2-3	2
>3	3
Albumin (g/dL)	
>3.5	1
2.8-3.5	2
<2.8	3
INR	
<1.7	1
1.7-2.3	2
>2.3	3
Ascites	
Absent	1
Slight	2
Moderate	3
Encephalopathy	
None	1
Grade 1-2	2
Grade 3-4	3

Table 11: **Before -** The Child-Pugh Score.

Clinical variable	Addition to score
Bilirubin (mg/dL)	
<0.5	1
0.5-3	2
>3	3
Albumin (g/dL)	
>5.1	1
3.0-5.1	2
<3.0	3
INR	
<1.1	1
1.1-2.5	2
>2.5	3
Ascites	
Absent	1
Slight	2
Moderate	3
Encephalopathy	
None	1
Grade 1-2	2
Grade 3-4	4

Table 12: **After -** Tuned Child-Pugh Score by SET.

Clinical variable	Addition to score
Sex	
Male	6
Age (years)	
45-57	9
58-68	20
69-102	24
Hypoxemia	
Yes	7
Glucose (mg/dL)	
<70	14
>140	5
AST to ALT ratio	
>1	9
C-reactive protein (mg/dL)	
>10	8
Arterial pH	
<7.35	7
>7.45	2
White blood cell count per μL	
>10 $\times 10^3$	9

Table 13: **Before -** CIMS.

Clinical variable	Addition to score
Sex	
Male	6
Age (years)	
45-57	9
58-68	20
69-102	24
Hypoxemia	
Yes	7
Glucose (mg/dL)	
<70	14
>140	6
AST to ALT ratio	
>1	9
C-reactive protein (mg/dL)	
>15	8
Arterial pH	
<7.35	7
>7.45	1
White blood cell count per μL	
>10 $\times 10^3$	13

Table 14: **After -** Tuned CIMS by SET.

Clinical variable	**Addition to score**
Confusion	
Yes	1
Urea (mmol/L)	
>7	1
Respiratory rate (breaths/min)	
≥30	1
Blood pressure (mmHg)	
SBP <90 or DBP ≤60	1
Age (years)	
≥65	1

Table 15: **Before -** CURB-65 Score.

Clinical variable	**Addition to score**
Confusion	
Yes	1
Urea (mmol/L)	
>7	1
Respiratory rate (breaths/min)	
≥30	1
Blood pressure (mmHg)	
SBP <90 or DBP ≤59	1
Age (years)	
≥65	1

Table 16: **After -** Tuned CURB-65 Score by SET

Clinical variable	Addition to score
Killip class	
II	20
III	39
IV	59
Systolic blood pressure (mmHg)	
≤80	58
80-99	53
100-119	43
120-139	34
140-159	24
160-199	10
Heart rate (beats/min)	
50-69	3
70-89	9
90-109	15
110-149	24
150-199	38
≥200	46
Age (years)	
30-39	8
40-49	25
50-59	41
60-69	58
70-79	75
80-89	91
≥90	100
Creatinine (mg/dL)	
0-0.39	1
0.40-0.79	4
0.80-1.19	7
1.20-1.59	10
1.60-1.99	13
2.00-3.99	21
>4.0	28
Other Risk Factors	
Cardiac Arrest at Admission	39
ST-Segment Deviation	28
Elevated Cardiac Enzyme Levels	14

Table 17: **Before -** GRACE Score.

Clinical variable	Addition to score
Killip class	
II	20
III	44
IV	59
Systolic blood pressure (mmHg)	
≤80	61
80-99	56
100-119	48
120-139	37
140-159	27
160-282	12
Heart rate (beats/min)	
50-69	3
70-89	9
90-109	15
110-149	24
150-199	38
≥200	46
Age (years)	
30-39	9
40-49	26
50-59	42
60-69	51
70-79	68
80-122	84
≥123	93
Creatinine (mg/dL)	
0-0.39	1
0.40-0.79	2
0.80-1.19	5
1.20-1.59	8
1.60-1.99	10
2.00-3.99	18
>4.0	25
Other Risk Factors	
Cardiac Arrest at Admission	18
ST-Segment Deviation	3
Elevated Cardiac Enzyme Levels	14

Table 18: **After -** Tuned GRACE Score by SET.

Clinical variable	**Addition to score**
Hypertension	
Yes	1
Renal disease	
Yes	1
Liver disease	
Yes	1
Stroke history	
Yes	1
Prior major bleeding or predisposition to bleeding	
Yes	1
Labile INR	
Yes	1
Age (years)	
>65	1
Medication usage predisposing to bleeding	
Yes	1
Alcohol use	
Yes	1

Table 19: **Before -** HAS-BLED Score.

Clinical variable	**Addition to score**
Hypertension	
Yes	1
Renal disease	
Yes	1
Liver disease	
Yes	1
Stroke history	
Yes	1
Prior major bleeding or predisposition to bleeding	
Yes	1
Labile INR	
Yes	1
Age (years)	
>43	1
Medication usage predisposing to bleeding	
Yes	1
Alcohol use	
Yes	1

Table 20: **After -** Tuned HAS-BLED Score by SET.

Clinical variable	**Addition to score**
Length of stay (days)	
1	1
2	2
3	3
4-6	4
7-13	5
≥14	7
Acute (emergent) admission	
Yes	3
Charlson Comorbidity Index	
1	1
2	2
3	3
≥4	5
Number of ED visits within 6 months	
1	1
2	2
3	3
≥4	4

Table 21: **Before -** LACE Score.

Clinical variable	**Addition to score**
Length of stay (days)	
1	1
2	2
3-5	3
6-7	4
8-16	5
≥17	6
Acute (emergent) admission	
Yes	2
Charlson Comorbidity Index	
1	1
2-3	2
4-5	3
≥6	6
Number of ED visits within 6 months	
1-2	1
3	2
4-5	3
≥6	4

Table 22: **After -** Tuned LACE Score by SET.

Clinical variable	Addition to score
Respiratory rate (breaths/min)	
≤8	3
9-11	1
21-24	2
≥25	3
Oxygen saturations (%)	
≤91	3
92-93	2
94-95	1
Any supplemental oxygen	
Yes	2
Temperature (°C)	
≤35.0	3
35.1-36.0	1
38.1-39.0	1
≥39.1	2
Systolic blood pressure (mmHg)	
≤90	3
91-100	2
101-110	1
≥220	3
Heart rate (beats/min)	
≤40	3
41-50	1
91-110	1
111-130	2
≥131	3
AVPU score	
Voice, Pain, or Unresponsive (V, P, U)	3

Table 23: **Before -** NEWS.

Clinical variable	Addition to score
Respiratory rate (breaths/min)	
≤4	3
5-11	1
21-24	2
≥25	3
Oxygen saturations (%)	
≤44	3
45-93	2
94-95	1
Any supplemental oxygen	
Yes	3
Temperature (°C)	
≤35.0	3
35.1-36.0	1
38.1-39.0	1
≥39.1	2
Systolic blood pressure (mmHg)	
≤90	3
91-100	2
101-110	1
≥220	1
Heart rate (beats/min)	
≤40	3
41-50	1
91-110	1
111-130	2
≥131	3
AVPU score	
Voice, Pain, or Unresponsive (V, P, U)	3

Table 24: **After -** Tuned NEWS by SET.

Clinical variable	Addition to score
Age (years)	Age Value
Sex	
Female	-10
Nursing home resident	
Yes	10
Neoplastic disease	
Yes	30
Liver disease history	
Yes	20
CHF history	
Yes	10
Cerebrovascular disease history	
Yes	10
Renal disease history	
Yes	10
Altered mental status	
Yes	20
Respiratory rate (breaths/min)	
$\geq$30	20
Systolic blood pressure (mmHg)	
<90	20
Temperature (°C)	
<35 or >39.9	15
Pulse (beats/min)	
$\geq$125	10
pH	
<7.35	30
BUN (mg/dL)	
$\geq$30	20
Sodium (mmol/L)	
<130	20
Glucose (mg/dL)	
$\geq$250	10
Hematocrit (%)	
<30	10
Partial pressure of oxygen (mmHg)	
<60	10
Pleural effusion on x-ray	
Yes	10

Table 25: **Before -** PSI Score.

Clinical variable	Addition to score
Age (years)	Age Value
Sex	
Female	-10
Nursing home resident	
Yes	10
Neoplastic disease	
Yes	30
Liver disease history	
Yes	20
CHF history	
Yes	10
Cerebrovascular disease history	
Yes	10
Renal disease history	
Yes	10
Altered mental status	
Yes	20
Respiratory rate (breaths/min)	
$\geq$30	20
Systolic blood pressure (mmHg)	
<90	20
Temperature (°C)	
<35 or >39.9	15
Pulse (beats/min)	
$\geq$125	11
pH	
<7.35	30
BUN (mg/dL)	
$\geq$30	16
Sodium (mmol/L)	
<130	20
Glucose (mg/dL)	
$\geq$250	10
Hematocrit (%)	
<30	10
Partial pressure of oxygen (mmHg)	
<60	9
Pleural effusion on x-ray	
Yes	10

Table 26: **After -** Tuned PSI Score by SET.

Appendix D. Simplified Tables

Clinical variable	Addition to score
Sex	
Male	6
Age (years)	
45-57	9
58-68	20
69-102	24
Hypoxemia	
Yes	7
Glucose (mg/dL)	
<70	14
>140	5
AST to ALT ratio	
>1	9
C-reactive protein (mg/dL)	
>10	8
Arterial pH	
<7.35	7
>7.45	2
White blood cell count per μL	
>10 $\times 10^3$	9

Table 27: **Before -** CIMS.

Clinical variable	Addition to score
Sex	
Male	8
Age (years)	
45-57	6
58-68	17
69-102	21
Hypoxemia	
Yes	7
Glucose (mg/dL)	
<70	14
>140	5
AST to ALT ratio	
>1	9
C-reactive protein (mg/dL)	
>5	8
White blood cell count per μL	
>10 $\times 10^3$	9

Table 28: **After -** Simplified CIMS without Arterial pH by SET.

Clinical variable	**Addition to score**
Confusion	
Yes	1
Urea (mmol/L)	
>7	1
Respiratory rate (breaths/min)	
≥30	1
Blood pressure (mmHg)	
SBP <90 or DBP ≤60	1
Age (years)	
≥65	1

Table 29: **Before -** CURB-65 Score.

Clinical variable	**Addition to score**
Confusion	
Yes	1
Respiratory rate (breaths/min)	
≥27	1
Blood pressure (mmHg)	
SBP < 87 or DBP ≤37	1
Age (years)	
≥43	1

Table 30: **After -** Simplified CURB-65 Score without Urea by SET.

Clinical variable	Addition to score
Killip class	
II	20
III	39
IV	59
Systolic blood pressure (mmHg)	
≤80	58
80-99	53
100-119	43
120-139	34
140-159	24
160-199	10
Heart rate (beats/min)	
50-69	3
70-89	9
90-109	15
110-149	24
150-199	38
≥200	46
Age (years)	
30-39	8
40-49	25
50-59	41
60-69	58
70-79	75
80-89	91
≥90	100
Creatinine (mg/dL)	
0-0.39	1
0.40-0.79	4
0.80-1.19	7
1.20-1.59	10
1.60-1.99	13
2.00-3.99	21
>4.0	28
Other Risk Factors	
Cardiac Arrest at Admission	39
ST-Segment Deviation	28
Elevated Cardiac Enzyme Levels	14

Table 31: **Before -** GRACE Score.

Clinical variable	Addition to score
Killip class	
II	20
III	44
IV	59
Systolic blood pressure (mmHg)	
≤80	61
80-99	56
100-119	48
120-139	37
140-159	27
160-282	12
Heart rate (beats/min)	
50-69	3
70-89	9
90-109	15
110-149	24
150-199	38
≥200	46
Age (years)	
30-39	9
40-49	26
50-59	42
60-69	51
70-79	68
80-122	84
≥123	93
Other Risk Factors	
Cardiac Arrest at Admission	18
ST-Segment Deviation	3
Elevated Cardiac Enzyme Levels	14

Table 32: **After -** Simplified GRACE Score without Creatinine level by SET.

Clinical variable	Addition to score
Hypertension	
Yes	1
Renal disease	
Yes	1
Liver disease	
Yes	1
Stroke history	
Yes	1
Prior major bleeding or predisposition to bleeding	
Yes	1
Labile INR	
Yes	1
Age (years)	
>65	1
Medication usage predisposing to bleeding	
Yes	1
Alcohol use	
Yes	1

Table 33: **Before -** HAS-BLED Score.

Clinical variable	Addition to score
Hypertension	
Yes	1
Renal disease	
Yes	1
Liver disease	
Yes	1
Stroke history	
Yes	1
Prior major bleeding or predisposition to bleeding	
Yes	1
Age (years)	
>59	1
Medication usage predisposing to bleeding	
Yes	1
Alcohol use	
Yes	1

Table 34: **After -** Simplified HAS-BLED Score without Labile INR by SET.

Clinical variable	**Addition to score**
Age (years)	Age Value
Sex	
Female	-10
Nursing home resident	
Yes	10
Neoplastic disease	
Yes	30
Liver disease history	
Yes	20
CHF history	
Yes	10
Cerebrovascular disease history	
Yes	10
Renal disease history	
Yes	10
Altered mental status	
Yes	20
Respiratory rate (breaths/min)	
≥30	20
Systolic blood pressure (mmHg)	
<90	20
Temperature (°C)	
<35 or >39.9	15
Pulse (beats/min)	
≥125	10
pH	
<7.35	30
BUN (mg/dL)	
≥30	20
Sodium (mmol/L)	
<130	20
Glucose (mg/dL)	
≥250	10
Hematocrit (%)	
<30	10
Partial pressure of oxygen (mmHg)	
<60	10
Pleural effusion on x-ray	
Yes	10

Table 35: **Before -** PSI Score.

Clinical variable	**Addition to score**
Age (years)	Age Value
Sex	
Female	-15
Nursing home resident	
Yes	8
Neoplastic disease	
Yes	9
Altered mental status	
Yes	11
Respiratory rate (breaths/min)	
≥30	23
Systolic blood pressure (mmHg)	
<90	2
Temperature (°C)	
<35 or >39.9	8
Pulse (beats/min)	
≥125	4
pH	
<7.35	32
BUN (mg/dL)	
≥30	8
Sodium (mmol/L)	
<130	8
Glucose (mg/dL)	
≥250	6
Hematocrit (%)	
<30	15
Partial pressure of oxygen (mmHg)	
<60	12
Pleural effusion on x-ray	
Yes	13

Table 36: **After -** Simplified PSI Score without patient history by SET.

Appendix E. Extended Scoring Tables

Clinical variable	Addition to score
Bilirubin (mg/dL)	
<2	1
2-3	2
>3	3
Albumin (g/dL)	
>3.5	1
2.8-3.5	2
<2.8	3
INR	
<1.7	1
1.7-2.3	2
>2.3	3
Ascites	
Absent	1
Slight	2
Moderate	3
Encephalopathy	
None	1
Grade 1-2	2
Grade 3-4	3

Table 37: **Before -** Child-Pugh Score.

Clinical variable	Addition to score
Bilirubin (mg/dL)	
<1.2	1
1.2-3.4	2
>3.4	3
Albumin (g/dL)	
>4.6	1
3.1-4.6	2
<3.1	3
INR	
<0.2	1
0.2-3.3	2
>3.3	3
Ascites	
Absent	1
Slight	2
Moderate	3
Encephalopathy	
None	1
Grade 1-2	2
Grade 3-4	4
Age (years)	
90-95	1
≥96	2

Table 38: **After -** Extended Child-Pugh Score with Age by SET.

Clinical variable	**Addition to score**
Length of stay (days)	
1	1
2	2
3	3
4-6	4
7-13	5
≥14	7
Acute (emergent) admission	
Yes	3
Charlson Comorbidity Index	
1	1
2	2
3	3
≥4	5
Number of ED visits within 6 months	
1	1
2	2
3	3
≥4	4

Table 39: **Before -** LACE Score.

Clinical variable	**Addition to score**
Length of stay (days)	
1	1
2	2
3	3
4-6	4
7-13	5
≥14	7
Acute (emergent) admission	
Yes	3
Charlson Comorbidity Index	
1	1
2	2
3	3
≥4	5
Number of ED visits within 6 months	
1	1
2	2
3	3
≥4	4
Age (years)	
20-39	1
40-59	2
≥60	3
Sex	
Male	1

Table 40: **After -** Extended LACE Score with Age and Sex by SET.

Clinical variable	**Addition to score**
Respiratory rate (breaths/min)	
≤8	3
9-11	1
21-24	2
≥25	3
Oxygen saturations (%)	
≤91	3
92-93	2
94-95	1
Any supplemental oxygen	
Yes	2
Temperature (°C)	
≤35.0	3
35.1-36.0	1
38.1-39.0	1
≥39.1	2
Systolic blood pressure (mmHg)	
≤90	3
91-100	2
101-110	1
≥220	3
Heart rate (beats/min)	
≤40	3
41-50	1
91-110	1
111-130	2
≥131	3
AVPU score	
Voice, Pain, or Unresponsive (V, P, U)	3

Table 41: **Before -** NEWS.

Clinical variable	**Addition to score**
Respiratory rate (breaths/min)	
≤8	4
9-10	1
21-27	2
≥28	3
Oxygen saturations (%)	
≤91	3
92-93	2
94-95	1
Any supplemental oxygen	
Yes	3
Temperature (°C)	
≤25.4	3
25.5-36.0	1
38.1-41.7	1
≥41.8	2
Systolic blood pressure (mmHg)	
≤90	3
91-100	2
101-110	1
≥220	1
Heart rate (beats/min)	
≤40	2
41-50	1
83-94	1
95-107	2
≥107	3
AVPU score	
Voice, Pain, or Unresponsive (V, P, U)	3
Age (years)	
38-80	1
≥81	3

Table 42: **After -** Extended NEWS with Age by SET.

Appendix F. Confidence Intervals and Statistical Significance Tests

Table 43: 95% confidence intervals via bootstrapping and bootstrap statistical significance test.

	Discrimination (95% CI)						Calibration (95% CI)		Reclassification (95% CI)	
	AUC	Youden Index	Sensitivity	Specificity	PPV	NPV	Brier	ECE	NRI	IDI
Child-Pugh										
Original	0.583–0.647	0.137–0.254	0.445–0.559	**0.678–0.708***	**0.109–0.147**	0.93–0.949	0.068–0.0826	**0.00156–0.0155***	0.00%	0.00%
SET	**0.624–0.684***	**0.195–0.306***	**0.649–0.753***	0.534–0.567	0.107–0.139	**0.945–0.963***	**0.0674–0.0817***	0.0069–0.0248	**+23.3%–+43.3%***	**+0.525%–+1.15%***
CIMS										
Original	0.812–0.826	0.473–0.503	0.778–0.803	**0.690–0.704***	**0.427–0.449***	0.912–0.923	**0.136–0.142**	0.016–0.0258	0.00%	0.00%
SET	**0.816–0.83***	**0.489–0.516***	**0.822–0.845***	0.661–0.676	0.419–0.44	**0.926–0.936***	**0.136–0.142**	**0.0149–0.0252**	**+11.4%–+18%***	**+0.357%–+0.864%***
CURB-65										
Original	0.677–0.741	0.304–0.418	**0.716–0.821**	0.573–0.612	0.153–0.195	0.947–0.969	0.0754–0.0921	0.0188–0.0389	0.00%	0.00%
SET	**0.683–0.744**	**0.331–0.443***	**0.716–0.821**	**0.599–0.638***	**0.162–0.207***	**0.95–0.971**	**0.0757–0.0919**	**0.0198–0.0381**	**+27.9%–+45.8%***	**-0.393%–+1.7%***
GRACE										
Original	0.742–0.829	0.175–0.413	0.386–0.617	**0.763–0.823***	0.16–0.283	0.911–0.951	0.1–0.144	0.0908–0.133	0.00%	0.00%
SET	**0.81–0.872***	**0.429–0.63***	**0.657–0.847***	0.748–0.809	**0.222–0.345***	**0.948–0.978***	**0.0793–0.118***	**0.073–0.111***	**+17.3%–+36.4%***	**+8.53%–+18.3%***
HAS-BLED										
Original	0.67–0.705	0.259–0.327	0.735–0.8	**0.513–0.538***	0.16–0.185	0.937–0.954	0.094–0.106	**0.00322–0.0146**	0.00%	0.00%
SET	**0.685–0.717***	**0.365–0.412***	**0.897–0.938***	0.457–0.483	**0.171–0.196***	**0.972–0.983***	**0.0933–0.104**	0.011–0.0249	**-3.35%–+12%***	**+1.58%–+2.5%***
LACE										
Original	0.645–0.708	0.201–0.241	**1–1**	0.201–0.241	0.0809–0.11	**1–1**	0.0636–0.0839	0.0583–0.0813	0.00%	0.00%
SET	**0.655–0.72***	**0.327–0.37***	**1–1**	**0.327–0.37***	**0.0955–0.129***	**1–1**	**0.0601–0.0793***	**0.0132–0.0353***	**+32.4%–+59%***	**+2.18%–+4.34%***
NEWS										
Original	0.524–0.548	-0.00168–0.0322	0.174–0.2	**0.818–0.839***	0.421–0.472	0.568–0.59	0.246–0.25	0.0435–0.064	0.00%	0.00%
SET	**0.528–0.552***	**0.0232–0.0598***	**0.258–0.287***	0.759–0.78	**0.444–0.487***	**0.577–0.6***	**0.245–0.249***	**0.0238–0.0435***	**+10.2%–+18.2%***	**+0.413%–+0.92%***
PSI										
Original	0.81–0.84	0.443–0.514	0.692–0.759	**0.739–0.766***	**0.333–0.382***	0.926–0.944	0.136–0.153	0.0804–0.0993	0.00%	0.00%
SET	**0.813–0.842***	**0.453–0.512**	**0.852–0.9***	0.591–0.621	0.278–0.317	**0.955–0.97***	**0.128–0.145***	**0.0785–0.0975**	**+1.25%–+13.7%***	**+3.9%–+8.48%***

* indicates $p < 0.05$ for the bootstrap test, with the null hypothesis being that there is no significant difference between SET and Original.

Table 44: Mean performance difference (and Wilcoxon signed-rank test) between SET and Original score (i.e., SET − Original) across different data splits. Improvement in metrics due to SET are bolded.

	ΔDiscrimination						ΔCalibration		ΔReclassification	
	AUC	Youden Index	Sensitivity	Specificity	PPV	NPV	Brier	ECE	NRI	IDI
Child-Pugh										
Difference	**+0.0425**†	**+0.0481**†	**+0.173**†	-0.125†	-0.000781	**+0.00974**†	**-0.000543**†	+0.00143†	**+16.6%**†	**+0.589%**†
CIMS										
Difference	**+0.00494**†	**+0.0098**†	**+0.0407**†	-0.0309†	-0.0123†	**+0.0115**†	**-0.000665**†	**-0.00107**†	**+17.5%**†	**+0.571%**†
CURB-65										
Difference	**+0.00725**†	**+0.0198**†	**+0.000118**	**+0.0198**†	**+0.00777**†	**+0.000665**†	+0.000178	**-0.00052**	**+54.9%**†	-0.0338%
GRACE										
Difference	**+0.0314**†	**+0.121**†	**+0.111**†	**+0.00985**†	**+0.0321**†	**+0.0156**†	**-0.0171**†	**-0.0182**†	**+25.9%**†	**+14.1%**†
HAS-BLED										
Difference	**+0.0157**†	**+0.117**†	**+0.168**†	-0.0511†	**+0.0144**†	**+0.037**†	**-0.00269**†	+0.00274†	**+16.5%**†	**+3.16%**†
LACE										
Difference	**+0.0264**†	**+0.0649**†	-0.0871†	**+0.152**†	**+0.00915**†	-0.00157	**-0.00159**†	**-0.0107**†	**+19.2%**†	**+1.26%**†
NEWS										
Difference	**+0.00482**†	**+0.0218**†	**+0.0782**†	-0.0563†	**+0.0142**†	**+0.00761**†	**-0.00136**†	**-0.00219**	**+6.51%**†	**+0.705%**†
PSI										
Difference	**+0.00165**†	-0.00285	**+0.139**†	-0.142†	-0.0561†	**+0.0236**†	**-0.0056**†	**-0.00489**†	**+3.84%**†	**+4.49%**†

† indicates $p < 0.05$ for the Wilcoxon signed-rank test.

F.1. SET with Same Clinical Variables

We repeat the experiments performed in Table 3 via i). 1000 different bootstraps (see Table 43) and ii). 100 different data splits (see Table 44). Each table uses a different appropriate statistical significant test used in medical and machine learning literature, respectively. For all metrics the higher the better, except for Brier and ECE in which the lower the better. SET demonstrates statistically significant improved performances across most metrics across all 8 clinical scores.

Table 45: 95% confidence intervals via bootstrapping and bootstrap statistical significance test.

	Discrimination (95% CI)						Calibration (95% CI)		Reclassification (95% CI)	
	AUC	Youden Index	Sensitivity	Specificity	PPV	NPV	Brier	ECE	NRI	IDI
CIMS										
Original	0.803–0.816	0.454–0.484	0.737–0.763	**0.712–0.727***	**0.434–0.456***	0.9–0.911	**0.138–0.144**	0.0169–0.0263	0.00%	0.00%
SET (Simplified)	**0.804–0.818***	**0.455–0.485**	**0.746–0.772***	0.702–0.717	0.428–0.45	**0.902–0.913***	**0.138–0.144***	**0.0126–0.0223***	**+7.67%–+14.1%***	**+0.284%–+0.761%***
CURB-65										
Original	0.553–0.617	0.083–0.207	0.328–0.444	**0.743–0.777***	0.126–0.179	0.904–0.93	0.0788–0.0979	0.0145–0.0379	0.00%	0.00%
SET (Simplified)	**0.632–0.699***	**0.226–0.346***	**0.599–0.712***	0.61–0.651	**0.144–0.188**	**0.931–0.954***	**0.0785–0.097**	**0.00963–0.0263***	**+17%–+41.7%***	**+0.0235%–+3.25%***
GRACE										
Original	0.724–0.816	0.132–0.37	0.386–0.617	0.718–0.781	0.136–0.241	0.906–0.949	0.116–0.162	0.102–0.145	0.00%	0.00%
SET (Simplified)	**0.805–0.868***	**0.292–0.521***	**0.514–0.74***	**0.748–0.809***	**0.189–0.31***	**0.928–0.966***	**0.121–0.168**	**0.0842–0.125***	**+0.708%–+5.29%***	**-1.33%–+2.21%***
HAS-BLED										
Original	0.6–0.643	**0.169–0.243***	0.35–0.42	**0.811–0.831***	**0.195–0.24**	0.904–0.919	**0.0964–0.109**	**0.00216–0.0125**	0.00%	0.00%
SET (Simplified)	**0.609–0.65***	0.133–0.186	**0.838–0.887***	0.285–0.308	0.127–0.147	**0.933–0.954***	**0.0965–0.109***	0.00155–0.0148	**+18.4%–+33.5%***	**+0.0147%–+2.17%***
PSI										
Original	0.812–0.841	0.48–0.539	**0.838–0.887***	0.632–0.663	0.296–0.339	**0.953–0.968***	**0.131–0.146**	**0.953–0.968**	**0.00%**	0.00%
SET (Simplified)	**0.818–0.846***	**0.482–0.542***	0.796–0.848	**0.676–0.706***	**0.314–0.358***	0.946–0.96	0.14–0.157	0.0881–0.109	-11.1%–+2.31%	**-1.99–+2.22%**

* indicates $p < 0.05$ for the bootstrap test, with the null hypothesis being that there is no significant difference between SET and Original.

Table 46: Mean performance difference (and Wilcoxon signed-rank test) between SET and Original score (i.e., SET − Original) across different data splits. Improvement in metrics due to SET are bolded.

	ΔDiscrimination						ΔCalibration		ΔReclassification	
	AUC	Youden Index	Sensitivity	Specificity	PPV	NPV	Brier	ECE	NRI	IDI
CIMS										
Difference	**+0.0021**†	**+0.00202**†	**+0.0106**†	-0.0086†	-0.00427†	**+0.00253**†	**-0.000275**†	**-0.000871**	**+7.88%**†	**+0.457%**†
CURB-65										
Difference	**+0.0562**†	**+0.0907**†	**+0.234**†	-0.144†	-0.0108†	**+0.022**†	+0.000792†	+0.00104	**+22.2%**†	-0.01%
GRACE										
Difference	**+0.04**†	**+0.134**†	**+0.106**†	**+0.0285**†	**+0.0433**†	**+0.0155**†	**-0.000589**	**-0.00552**†	**+5.83%**†	**+0.896%**
HAS-BLED										
Difference	**+0.0101**†	-0.000738	**+0.515**†	-0.516†	-0.0541†	**+0.0393**†	**-0.00062**†	+0.00432†	**+29%**†	**+0.778%**†
PSI										
Difference	**+0.0152**†	**+0.0278**†	-0.0228†	**+0.0506**†	**+0.0272**†	-0.00247†	+0.00218†	+0.00449†	**+10.8%**†	**+1.44%**†

† indicates $p < 0.05$ for the Wilcoxon signed-rank test.

F.2. SET for Simplification of Tables

We repeat the experiments performed in Table 8 via i). 1000 different bootstraps (see Table 45) and ii). 100 different data splits (see Table 46). Each table uses a different appropriate statistical significant test used in medical and machine learning literature, respectively. For all metrics the higher the better, except for Brier and ECE in which the lower the better. SET demonstrates statistically significant improved performances across most metrics across simplification of all 5 clinical scores.

Appendix G. Abalation Studies

Clinical variable	Addition to score
Respiratory rate (breaths/min)	
≤8	3
9-11	1
21-24	2
≥25	3
Oxygen saturations (%)	
≤91	3
92-93	2
94-95	1
Any supplemental oxygen	
Yes	2
Temperature (°C)	
≤35.0	3
35.1-36.0	1
38.1-39.0	1
≥39.1	2
Systolic blood pressure (mmHg)	
≤90	3
91-100	2
101-110	1
≥220	3
Heart rate (beats/min)	
≤40	3
41-50	1
91-110	1
111-130	2
≥131	3
AVPU score	
Voice, Pain, or Unresponsive (V, P, U)	3

Table 47: **Before -** NEWS.

Clinical variable	Addition to score
Respiratory rate (breaths/min)	
≤3	4
4-9	1
12-34	2
≥35	3
Oxygen saturations (%)	
≤60	3
61-100	2
100-101	1
Any supplemental oxygen	
Yes	2
Temperature (°C)	
≤9.6	3
9.7-24.2	1
31.1-34.8	1
≥34.9	2
Systolic blood pressure (mmHg)	
≤24	3
25-46	2
47-116	1
≥274	3
Heart rate (beats/min)	
≤7	4
8-30	1
31-49	1
50-82	2
≥83	3
AVPU score	
Voice, Pain, or Unresponsive (V, P, U)	1

Table 48: **After -** Tuned NEWS by SET with penalty as separate objective. As observed, the penalty term is hardly enforced.

G.1. Ablation: Separate Penalty Objective

Separating penalty objective led to largely 'nonsensical' tables. We perform an ablation that treats the negative of the penalty term as the fourth objective. However this led to eventual fine-tuned clinical scoring tables with high penalties (see Table 48 where the numerical constants deviate strongly from the original table). As expected, these tables also do not perform well on the test set.

We found that this is because in every generation, many candidate scoring tables which incur high penalty still survive as long as they perform sufficiently well in the other 3 objectives, leading to a search process that did not respect the penalty term as intended, influencing even the first generation of evolution. Through experimentation, we found subtracting the penalty to be the more effective approach by far, which led to SET's final choice of having the penalty term subtracted from the objectives instead.

G.2. Ablation: Using Only One Objective

Table 49: Mean performance difference (and Wilcoxon signed-rank test) between SET (one-objective variant) and Original score across different data splits. Improvement in metrics due to SET (one-objective variant) are bolded. By only optimizing for one objective (i.e., Brier score), the other metrics demonstrate much poorer performance than multi-objective SET.

	ΔDiscrimination						**ΔCalibration**		**ΔReclassification**	
	AUC	Youden Index	Sensitivity	Specificity	PPV	NPV	Brier	ECE	NRI	IDI
Child-Pugh										
Difference	-0.108†	-0.0827†	-0.375†	**+0.292**†	**+0.222**†	-0.0106†	**-0.00225**†	+0.00162†	**+14%**†	**+3.67%**†
CIMS										
Difference	**+0.00283**†	**+0.0165**†	**+0.0306**†	-0.0142†	-0.00218†	**+0.00946**†	**-0.000425**†	+0.000109	**+3.81%**†	**+0.897%**†
CURB-65										
Difference	**+0.0125**†	-0.000738	-0.0816†	**+0.0809**†	**+0.0197**†	-0.00999†	**-0.00273**†	**-0.00299**†	-31.4%†	**+4.28%**†
GRACE										
Difference	-0.00515†	-0.0714†	-0.0657†	-0.00567	-0.0194†	-0.00911†	**-0.0187**†	**-0.014**†	**+36.3%**†	**+19.3%**†
HAS-BLED										
Difference	**+0.0144**†	**+0.117**†	**+0.168**†	-0.0518†	**+0.0142**†	**+0.037**†	**-0.00269**†	+0.00296†	**+15.6%**†	**+3.2%**†
LACE										
Difference	-0.0113†	-0.0275†	-0.621†	**+0.593**†	**+0.0164**†	-0.0278†	**-0.00381**†	+0.00657†	**+28.1%**†	**+9.69%**†
NEWS										
Difference	-0.049†	-0.00196	-0.0692†	**+0.0672**†	**+0.0123**†	-0.000998	**-0.00838**†	+0.0119†	**+35.2%**†	**+4.33%**†
PSI										
Difference	-0.000128	-0.00575	**+0.111**†	-0.116†	-0.0496†	**+0.0176**†	**-0.00229**†	+0.0113†	**+8.14%**†	**+6.75%**†

† indicates $p < 0.05$ for the Wilcoxon signed-rank test.

We also performed an ablation study that optimizes for just one of the 3 objectives (i.e., Brier score), with the results of SET (one-objective variant) in Table 49. Comparing these results against Table 44, it becomes clear that optimizing for Brier Score alone led to much poorer performance, in which discrimination performance such as **AUC decreases to be even worse than the original score itself** (i.e., difference of AUC < 0), and even Brier metric itself is lower on average on some problems. This exemplifies that finding good clinical scoring tables is inherently not a single-objective task, but rather a multi-objective task. In other words, optimizing for one objective alone will not automatically improve the performance on other objectives as well. Therefore, the multi-objective approach with SET uses is better and preferred.

G.3. Ablation: Replacing NSGA-II with Other Numerical Optimizers

In clinical scoring work, there exists alternative numerical solvers/optimizers to NSGA-II such as RiskSLIM-MINLP (Ustun and Rudin, 2019) and FasterRisk (Liu et al., 2022). In order to incorporate them into our work, we created 2 new variants of SET: SET-RS and SET-FR, which replaces the numerical solver in SET (i.e., NSGA-II) with RiskSLIMMINLP and FasterRisk respectively. We use the convention that if we specify SET without any dashes, it refers to using NSGA-II as the numerical solver by default. We report the results of SET-RS and SET-FR in Tables 50 and 51 respectively.

Comparing these results to SET (Table 44), SET demonstrates robustly better performance compared to the 2 variants, SET-RS and SET-FR, on most metrics across the 8 clinical tasks. This is due to several differences:

1. The thresholds on the left column of clinical scoring tables are not adjustable via SET-RS and SET-FR. Rather, the thresholds have to be preselected before the optimization (see Definition 1 in (Ustun and Rudin, 2019) and Eq. 1 in (Liu et al., 2022)). Therefore, the algorithms and theorems obtained from RiskSLIMMINLP and FasterRisk do not apply to optimizing the thresholds, whereas NSGA-II can optimize thresholds, allowing SET greater flexibility.

Table 50: Mean performance difference (and Wilcoxon signed-rank test) between SET-RS and Original score across different data splits. Improvement in metrics due to SET-RS are bolded.

	ΔDiscrimination						**ΔCalibration**		**ΔReclassification**	
	AUC	Youden Index	Sensitivity	Specificity	PPV	NPV	Brier	ECE	NRI	IDI
Child-Pugh										
Difference	-0.0141†	-0.0351†	**+0.0012**	-0.0351†	-0.0106†	-0.00337†	+0.000323†	**-0.00652**†	-24.6%†	-0.606%†
CIMS										
Difference	-0.00504†	**+0.000381**	-0.00995†	**+0.0103**†	**+0.0058**†	-0.00231†	+0.00205†	**-0.00541**†	-16.1%†	-1.62%†
CURB-65										
Difference	**+0.0245**†	**+0.0692**†	**+0.116**†	-0.0468†	**+0.0467**†	-0.00348†	**-0.00373**†	**-0.000476**	**+36%**†	**+3.52%**†
GRACE										
Difference	**+0.0884**†	**+0.205**†	**+0.0714**†	**+0.133**†	**+0.159**†	**+0.0129**†	**-0.00634**†	**-0.0152**†	-2.63%	**+1.91%**
HAS-BLED										
Difference	**+0.0463**†	**+0.187**†	**+0.2**†	-0.0135†	**+0.0313**†	**+0.0463**†	**-0.00602**†	+0.00119	**+79.3%**†	**+6.48%**†
LACE										
Difference	**+0.0204**†	-0.0595†	**+0.114**†	-0.173†	-0.00589†	**+0.042**†	**-0.00142**†	**-0.0136**†	**+19.8%**†	**+0.738%**†
NEWS										
Difference	-0.0392†	-0.0203†	-0.192†	**+0.172**†	**+0.0342**	-0.00605†	**-0.000409**†	**-0.0214**†	**+2.43%**†	-0.171%†
PSI										
Difference	-0.0201†	**+0.0178**†	**+0.147**†	-0.129†	-0.0476†	**+0.0268**†	+0.00606†	**-0.00625**†	-27.8%†	-6.34%†

† indicates $p < 0.05$ for the Wilcoxon signed-rank test.

Table 51: Mean performance difference (and Wilcoxon signed-rank test) between SET-FR and Original score across different data splits. Improvement in metrics due to SET-FR are bolded.

	ΔDiscrimination						**ΔCalibration**		**ΔReclassification**	
	AUC	Youden Index	Sensitivity	Specificity	PPV	NPV	Brier	ECE	NRI	IDI
Child-Pugh										
Difference	-0.0123†	-0.0486†	-0.0277†	-0.0209†	-0.0127†	-0.00515†	+0.000242†	**-0.00182**†	-27.4%†	-0.482%†
CIMS										
Difference	**+0.00174**†	**+0.01**†	**+0.0481**†	-0.038†	-0.0155†	**+0.0135**†	**-5.95e-06**	**-0.00182**†	**+2.31%**†	-0.317%†
CURB-65										
Difference	**+0.0267**†	**+0.0692**†	**+0.116**†	-0.0468†	**+0.0467**†	-0.00348†	**-0.00376**†	+0.000528	**+39.9%**†	**+4.31%**†
GRACE										
Difference	**+0.0753**†	**+0.115**†	**+0.0714**†	**+0.0439**†	**+0.0502**†	**+0.011**†	+0.0107†	+0.00726†	-15.3%†	-11.2%†
HAS-BLED										
Difference	**+0.0116**†	**+0.114**†	**+0.168**†	-0.0545†	**+0.0135**†	**+0.0369**†	**-0.00274**†	+0.00278†	**+13.7%**†	**+3.18%**†
LACE										
Difference	**+0.0219**†	-0.0595†	**+0.114**†	-0.173†	-0.00589†	**+0.042**†	**-0.000843**†	**-0.0177**†	**+10.8%**†	**+0.354%**†
NEWS										
Difference	-0.0111†	-0.0203†	-0.192†	**+0.172**†	**+0.0275**	-0.00605†	**-0.0001**	**-0.0219**†	-6.68%†	-0.369%†
PSI										
Difference	-0.0185†	**+0.0198**†	**+0.172**†	-0.153†	-0.0538†	**+0.0328**†	+0.00112†	**-0.0169**†	-15%†	-3.43%†

† indicates $p < 0.05$ for the Wilcoxon signed-rank test.

2. While NSGA-II can optimize for discrimination, calibration and reclassification objectives concurrently, the works in RiskSLIMMINLP and FasterRisk can only optimize a single surrogate objective, the logistic loss, as specified in Definition 1 in (Ustun and Rudin, 2019) and Eq. 1 in (Liu et al., 2022). Thus, SET, which uses NSGA-II, is better poised to create fine-tuned scores that have increased performance in all 3 areas: discrimination, calibration and reclassification.

3. NSGA-II allows for flexibility in the objectives, allowing us to create novel objectives (i.e., Active AUC, Active Brier), which quantifies improvements relative to original baseline score, whereas RiskSLIMMINLP and FasterRisk have fixed objectives. Thus, NSGA-II is better aligned to the mission of this paper to improve existing clinical scoring tables.

Table 52: Mean performance difference (and Wilcoxon signed-rank test) between AdaBoost and Original score across different data splits. Improvement in metrics due to AdaBoost are bolded.

	Discrimination						**Calibration**		**Reclassification**	
	AUC	Youden Index	Sensitivity	Specificity	PPV	NPV	Brier	ECE	NRI	IDI
Child-Pugh										
Difference	**+0.02**†	**+0.0154**	-0.0523†	**+0.0676**†	**+0.0156**†	-0.000653	+0.0318†	+0.0714†	**+4.02%**†	-0.0935%
CIMS										
Difference	**+0.00977**†	**+0.0247**†	-0.0143†	**+0.039**†	**+0.0294**†	-0.000862	+0.0964†	+0.19†	-7.6%†	-19.1%†
CURB-65										
Difference	-0.0435†	-0.081†	-0.348†	**+0.267**†	**+0.0794**†	-0.0291†	+0.0298†	+0.0704†	-5.08%†	-12.9%†
GRACE										
Difference	**+0.00984**	-0.185†	-0.387†	**+0.202**†	**+0.236**	-0.0262†	**-0.0166**†	**-0.00763**†	**+3.16%**†	**+2.3%**†
HAS-BLED										
Difference	**+0.122**†	**+0.0823**†	-0.092†	**+0.174**†	**+0.0525**†	-0.00202	+0.0159†	+0.072†	**+39.2%**†	**+8.45%**†
LACE										
Difference	-0.351†	-0.233†	-0.799†	**+0.567**†	-0.0932†	-0.0919†	+0.0286†	+0.0466†	-29.8%†	-4.32%†
NEWS										
Difference	**+0.0674**†	**+0.107**†	**+0.336**†	-0.229†	**+0.0429**†	**+0.0505**†	+0.0857†	+0.167†	**+19.2%**†	**+5.66%**†
PSI										
Difference	-0.0367†	-0.0462†	**+0.0147**†	-0.0609†	-0.0455†	-0.00189	+0.0172†	+0.0191†	-17.7%†	-6.37%†

† indicates $p < 0.05$ for the Wilcoxon signed-rank test.

Appendix H. Comparison with AdaBoost

Tree-based machine learning models can be said to be related to clinical scoring table since both use thresholds. Typically, tree-based models utilize many thresholds sequentially which makes it incompatible and too different with scoring tables. However, **AdaBoost** (Friedman et al., 2000) in particular, is an ensemble of tree stumps, which allows it to be converted into a scoring table, albeit with scoring components that are non-integer values and with potentially many more thresholds than a typical scoring table.

Here, we fit AdaBoost to the same features as used in the original clinical scoring table, without any threshold values (e.g., Age is given as feature to AdaBoost instead of a binary feature such as Age$\geq$ 65), in the same way data is provided to SET. We compare the performance of AdaBoost against the original clinical scores in Table 52. By comparing these improvements with the improvements by SET (see Table 44), the results show that SET still yields better test performance (across multiple metrics and multiple clinical scores) even when compared to a more complex model such at AdaBoost (which tends to overfit), suggesting that SET's novel choice of building upon domain-experts-crafted existing clinical scoring tables help to identify high-quality generalizable components and thresholds that traditional machine learning algorithms do not incoporate.

Appendix I. More Dataset Details

This paper works on eight scores: Child-Pugh score (Child and Turcotte, 1964) using 913 samples with 0.0832 prevalence rate of outcome from Chang et al. (2020), COVID in-hospitality mortality score (CIMS) (Dueñas-Espín et al., 2023) using 4742 samples with 0.23 prevalence rate of outcome from Dueñas-Espín et al. (2023), CURB-65 score (Lim et al., 2003) using 646 samples with 0.0991 prevalence rate of outcome from Millman et al. (2017), GRACE score (Fox et al., 2006) using 188 samples with 0.101 prevalence rate of outcome from Zhu et al. (2020), HAS-BLED score (Pisters et al., 2010) using 1588 samples with 0.115 prevalence rate of outcome from AlAmmari et al. (2021), LACE score (Van Walraven et al., 2010) using 463 samples with 0.0778 prevalence rate of outcome from Robinson and Hudali (2017), NEWS (RCoP, 2012) using 2204 samples with 0.425 prevalence rate of outcome from Mitsunaga et al. (2019), and PSI score (Fine et al., 1997) using 1138 samples with 0.16 prevalence rate of outcome from Chang et al. (2024).

Table 53: Performance metrics of original score and SET fine-tuned score, an an alternative solution on the Pareto front of NSGA-II in SET. For all metrics except c^*, Brier and ECE, the higher the better. For Brier and ECE, the lower the better.

	Discrimination							**Calibration**			**Reclassification**	
	AUC	c^*	Youden Index	Sensitivity	Specificity	PPV	NPV	Brier	ECE	HL p-value	NRI	IDI
HAS-BLED												
Original	0.687	4	0.293	0.767	0.526	0.173	0.946	0.0997	0.00692	0.992	0.00%	0.00%
SET	0.700	4	0.389	0.918	0.471	0.184	0.978	0.0988	0.0179	0.963	+4.74%	+2.07%
SET (Another Solution)	0.699	4	0.387	0.918	0.469	0.183	0.978	0.0989	0.0186	0.953	+3.32%	+2.07%

Appendix J. Exploring Other Solutions on Pareto Set

Recall each candidate score $\mathbf{x}$ is evaluated using:

$$F(\mathbf{x}) = [\sigma(\Delta_{\text{AUC}} - P(\mathbf{x})), \sigma(\Delta_{\text{Brier}} - P(\mathbf{x})), \mathbb{I}(\text{NRI})]$$

where σ is the ReLU function.

Also recall that, once NSGA-II converges, the best candidate solution is selected based on its overall performance across discrimination, calibration, and reclassification metrics while adhering to the imposed constraints and eventually based on the crowding distance selection rules in NSGA-II. Suppose we want to select a score with lower $\sigma(\Delta_{\text{AUC}} - P(\mathbf{x}))$ and higher $\sigma(\Delta_{\text{Brier}} - P(\mathbf{x}))$ that does not follow the crowding distance selection rule, we can search for such a score on the Pareto front of the output of NSGA-II manually.

In HAS-BLED for instance, such a score exists on the Pareto front. In Table 53, we label this score as 'SET (Another Solution)'. However, do note that this Pareto front is assessed based on the objectives in NSGA-II which is on the train set, the performance on the test set, as seen in Table 53, may be different, in which the alternative SET solution performs worse on both AUC and Brier when compared to SET, although the Brier score objective was better on the train set during optimization.

Appendix K. Genetic Algorithms and NSGA-II

Genetic algorithms are a class of population-based optimization techniques inspired by the principles of natural selection and evolutionary biology. They work by evolving a population of candidate solutions over multiple generations. Each candidate solution, often represented as a vector or chromosome, is evaluated using a fitness function that quantifies its quality with respect to the objective. The algorithm then selects the fittest individuals to reproduce, generating new candidate solutions through recombination (crossover) and random mutation. Over time, the population converges toward better solutions, making GAs particularly effective for complex, nonlinear, or poorly understood search spaces.

NSGA-II is a widely used genetic algorithm designed specifically for multi-objective optimization problems, where two or more objectives may be in conflict. Rather than seeking a single optimal solution, NSGA-II identifies a set of Pareto-optimal solutions, each representing a trade-off between competing objectives. It improves upon earlier approaches by introducing a fast non-dominated sorting procedure, which ranks individuals based on levels of dominance in the population. It also introduces a crowding-distance metric to promote diversity along the Pareto front, ensuring that solutions are well spread out. Additionally, NSGA-II uses an elitist strategy, combining parent and offspring populations before selecting the next generation, which helps preserve the best solutions discovered so far.