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)	
≥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 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)	
≥30	20
Systolic blood pressure (mmHg)	
<90	20
Temperature (°C)	
<35 or >39.9	15
Pulse (beats/min)	
≥125	11
pH	
<7.35	30
BUN (mg/dL)	
≥30	16
Sodium (mmol/L)	
<130	20
Glucose (mg/dL)	
≥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.