Supplementary materials

Judith Neve

Table S1: Allocated runtimes for each study.

Study	Number of datasets generated in one run	Number of runs	Original allocated time per run	Allocated runtime for rerunning of timed out runs	Number of timed out runs	Number of reruns taking longer than the original allocated runtime
1 (8 predictor scenarios)	1	3000	6 hours	8 hours	0	0
1 (16 predictor scenarios)	1	3000	30 hours	32 hours	188	35*
2	1	6000	3.5 hours	4 hours	1	0
3	10	600	6 hours	7 hours	7	0

*among these, 13 failed due to time-out. These were rerun with a maximum allocated time of 35 hours.

Timed out runs for Study 1 (16 predictor scenarios): 52, 64, 76, 82, 94, 100, 130, 142, 148, 154, 166, 184, 190, 202, 214, 220, 226, 232, 238, 244, 250, 256, 262, 268, 274, 280, 286, 292, 298, 304, 328, 340, 376, 400, 412, 424, 448, 454, 472, 478, 490, 508, 514, 550, 556, 562, 568, 580, 586, 610, 622, 628, 634, 646, 652, 658, 664, 670, 688, 694, 700, 706, 712, 730, 742, 760, 766, 772, 778, 784, 790, 796, 802, 808, 826, 838, 844, 850, 856, 862, 868, 886, 892, 934, 982, 988, 1030, 1060, 1066, 1072, 1078, 1150, 1192, 1198, 1204, 1330, 1336, 1456, 1480, 1486, 1492, 1498, 1594, 1606, 1654, 1750, 1756, 1768, 1774, 1822, 1846, 1852, 1882, 1888, 1894, 1900, 1912, 1918, 1930, 1936, 1942, 1948, 1954, 1972, 2026, 2092, 2098, 2104, 2146, 2158, 2194, 2200, 2224, 2230, 2236, 2248, 2272, 2296, 2302, 2320, 2326, 2374, 2380, 2386, 2410, 2416, 2422, 2428, 2446, 2452, 2458, 2464, 2470, 2482, 2494, 2512, 2524, 2530, 2536, 2566, 2590, 2608, 2638, 2644, 2710, 2716, 2746, 2758, 2764, 2776, 2788, 2806, 2830, 2836, 2848, 2866, 2878, 2884, 2896, 2908, 2914, 2932, 2938, 2944, 2956, 2962, 2986, 2992

Timed out reruns for Study 1 (16 predictor scenarios): 214, 274, 328, 376, 412, 448, 688, 700, 730, 742, 784, 802, 862

Timed out runs for Study 2: 788

Timed out runs for Study 3: 80, 296, 380, 476, 524, 560, 584

Table S2: Average performance of hyperparameter combinations for each scenario where $EF = 0.1$. Within each scenario, rows are sorted in ascending order of runtime.

Tuned hyperparameters	p	EF	n	AUC	Calibration slope	RMSD(slope)	Runtime (seconds)
				Mean (SD)	Median (IQR)		Mean (SD)
None	8	0.10	$0.5N$	0.70 (0.01)	0.66 (0.13)	0.44	1.00 (0000.50)
mtry + min.node.size	8	0.10	$0.5N$	0.70 (0.01)	0.83 (0.16)	0.27	63.10 (0009.70)
mtry + min.node.size + replace	8	0.10	$0.5N$	0.70 (0.01)	0.83 (0.17)	0.27	141.40 (0021.90)
mtry + min.node.size + splitrule	8	0.10	$0.5N$	0.71 (0.01)	1.00 (0.34)	0.26	189.70 (0032.00)
mtry + min.node.size + replace + splitrule	8	0.10	$0.5N$	0.71 (0.01)	0.97 (0.38)	0.30	425.00 (0078.20)
mtry + min.node.size + sample.fraction	8	0.10	$0.5N$	0.71 (0.01)	0.96 (0.36)	0.28	455.60 (0070.40)
mtry + min.node.size + sample.fraction + replace	8	0.10	$0.5N$	0.71 (0.01)	0.97 (0.39)	0.30	982.90 (0148.70)
mtry + min.node.size + sample.fraction + splitrule	8	0.10	$0.5N$	0.71 (0.01)	1.03 (0.37)	0.27	1370.40 (0231.30)
mtry + min.node.size + sample.fraction + replace + splitrule	8	0.10	$0.5N$	0.71 (0.02)	0.98 (0.38)	0.29	2948.60 (0483.60)
None	8	0.10	$1N$	0.71 (0.01)	0.74 (0.10)	0.32	1.70 (0000.40)
mtry + min.node.size	8	0.10	$1N$	0.72 (0.01)	0.90 (0.12)	0.17	138.50 (0019.40)
mtry + min.node.size + replace	8	0.10	$1N$	0.72 (0.01)	0.89 (0.12)	0.18	314.30 (0047.10)
mtry + min.node.size + splitrule	8	0.10	$1N$	0.73 (0.01)	1.07 (0.25)	0.21	396.30 (0067.70)
mtry + min.node.size + replace + splitrule	8	0.10	$1N$	0.73 (0.01)	1.04 (0.26)	0.20	900.20 (0168.10)
mtry + min.node.size + sample.fraction	8	0.10	$1N$	0.72 (0.01)	0.98 (0.22)	0.18	936.00 (0129.20)
mtry + min.node.size + sample.fraction + replace	8	0.10	$1N$	0.72 (0.01)	0.98 (0.23)	0.19	2034.90 (0285.70)
mtry + min.node.size + sample.fraction + splitrule	8	0.10	$1N$	0.73 (0.01)	1.07 (0.27)	0.20	2673.90 (0435.80)
mtry + min.node.size + sample.fraction + replace + splitrule	8	0.10	$1N$	0.73 (0.01)	1.06 (0.26)	0.20	5850.20 (0973.80)
None	16	0.10	$0.5N$	0.71 (0.01)	0.72 (0.11)	0.34	2.50 (0000.40)
mtry + min.node.size	16	0.10	$0.5N$	0.72 (0.01)	1.08 (0.16)	0.14	479.20 (0062.70)
mtry + min.node.size + replace	16	0.10	$0.5N$	0.71 (0.01)	1.07 (0.18)	0.15	1121.60 (0162.10)
mtry + min.node.size + splitrule	16	0.10	$0.5N$	0.72 (0.01)	1.14 (0.31)	0.24	1274.00 (0201.90)
mtry + min.node.size + sample.fraction	16	0.10	$0.5N$	0.72 (0.01)	1.12 (0.27)	0.21	2958.40 (0361.20)
mtry + min.node.size + replace + splitrule	16	0.10	$0.5N$	0.72 (0.01)	1.13 (0.31)	0.23	3025.00 (0523.20)
mtry + min.node.size + sample.fraction + replace	16	0.10	$0.5N$	0.71 (0.01)	1.09 (0.27)	0.21	6644.10 (0840.20)
mtry + min.node.size + sample.fraction + splitrule	16	0.10	$0.5N$	0.72 (0.01)	1.15 (0.32)	0.24	7997.80 (1145.30)
mtry + min.node.size + sample.fraction + replace + splitrule	16	0.10	$0.5N$	0.72 (0.01)	1.15 (0.31)	0.25	17940.90 (2767.00)
None	16	0.10	$1N$	0.72 (0.01)	0.80 (0.08)	0.23	5.30 (0000.80)
mtry + min.node.size	16	0.10	$1N$	0.72 (0.01)	1.12 (0.10)	0.14	1154.00 (0155.00)
mtry + min.node.size + replace	16	0.10	$1N$	0.72 (0.01)	1.11 (0.14)	0.13	2674.00 (0374.50)
mtry + min.node.size + splitrule	16	0.10	$1N$	0.73 (0.01)	1.16 (0.24)	0.21	2797.30 (0481.10)
mtry + min.node.size + replace + splitrule	16	0.10	$1N$	0.73 (0.01)	1.17 (0.24)	0.21	6561.30 (1212.50)
mtry + min.node.size + sample.fraction	16	0.10	$1N$	0.72 (0.01)	1.14 (0.14)	0.17	6998.70 (0906.30)
mtry + min.node.size + sample.fraction + replace	16	0.10	$1N$	0.72 (0.01)	1.12 (0.17)	0.16	15623.90 (2054.70)
mtry + min.node.size + sample.fraction + splitrule	16	0.10	$1N$	0.73 (0.01)	1.16 (0.23)	0.21	17164.90 (2802.10)
mtry + min.node.size + sample.fraction + replace + splitrule	16	0.10	$1N$	0.73 (0.01)	1.15 (0.21)	0.20	38521.20 (6617.70)

Table S3: Average performance of hyperparameter combinations for each scenario where $EF = 0.3$. Within each scenario, rows are sorted in ascending order of runtime.

Tuned hyperparameters	p	EF	n	AUC	Calibration slope	RMSD(slope)	Runtime (seconds)
				Mean (SD)	Median (IQR)		Mean (SD)
None	8	0.30	$0.5N$	0.68 (0.02)	0.81 (0.16)	0.26	0.80 (0001.20)
mtry + min.node.size	8	0.30	$0.5N$	0.68 (0.02)	0.93 (0.25)	0.25	38.30 (0005.70)
mtry + min.node.size + replace	8	0.30	$0.5N$	0.68 (0.02)	0.92 (0.29)	0.28	83.40 (0011.70)
mtry + min.node.size + splitrule	8	0.30	$0.5N$	0.69 (0.02)	1.07 (0.51)	0.36	123.90 (0018.40)
mtry + min.node.size + replace + splitrule	8	0.30	$0.5N$	0.69 (0.02)	1.01 (0.51)	0.38	273.50 (0042.30)
mtry + min.node.size + sample.fraction	8	0.30	$0.5N$	0.68 (0.01)	1.00 (0.34)	0.29	302.90 (0044.90)
mtry + min.node.size + sample.fraction + replace	8	0.30	$0.5N$	0.68 (0.02)	0.95 (0.37)	0.34	635.80 (0092.00)
mtry + min.node.size + sample.fraction + splitrule	8	0.30	$0.5N$	0.69 (0.02)	1.05 (0.44)	0.35	953.60 (0144.80)
mtry + min.node.size + sample.fraction + replace + splitrule	8	0.30	$0.5N$	0.68 (0.02)	1.01 (0.46)	0.35	2026.40 (0324.30)
None	8	0.30	$1N$	0.70 (0.01)	0.92 (0.14)	0.14	1.10 (0000.70)
mtry + min.node.size	8	0.30	$1N$	0.70 (0.01)	1.03 (0.25)	0.18	72.00 (0009.00)
mtry + min.node.size + replace	8	0.30	$1N$	0.70 (0.01)	1.01 (0.26)	0.18	159.80 (0020.20)
mtry + min.node.size + splitrule	8	0.30	$1N$	0.71 (0.01)	1.15 (0.28)	0.26	228.50 (0034.30)
mtry + min.node.size + replace + splitrule	8	0.30	$1N$	0.71 (0.01)	1.11 (0.33)	0.26	512.90 (0082.50)
mtry + min.node.size + sample.fraction	8	0.30	$1N$	0.70 (0.01)	1.04 (0.27)	0.20	513.90 (0065.60)
mtry + min.node.size + sample.fraction + replace	8	0.30	$1N$	0.70 (0.01)	1.01 (0.27)	0.20	1102.50 (0141.90)
mtry + min.node.size + sample.fraction + splitrule	8	0.30	$1N$	0.71 (0.01)	1.15 (0.31)	0.25	1619.40 (0236.60)
mtry + min.node.size + sample.fraction + replace + splitrule	8	0.30	$1N$	0.71 (0.01)	1.10 (0.33)	0.25	3508.40 (0521.30)
None	16	0.30	$0.5N$	0.68 (0.01)	0.96 (0.16)	0.13	1.30 (0000.30)
mtry + min.node.size	16	0.30	$0.5N$	0.69 (0.01)	1.16 (0.30)	0.24	202.20 (0022.70)
mtry + min.node.size + replace	16	0.30	$0.5N$	0.69 (0.01)	1.10 (0.32)	0.23	459.80 (0050.60)
mtry + min.node.size + splitrule	16	0.30	$0.5N$	0.69 (0.01)	1.21 (0.40)	0.32	649.60 (0094.30)
mtry + min.node.size + sample.fraction	16	0.30	$0.5N$	0.69 (0.01)	1.16 (0.33)	0.25	1345.60 (0142.30)
mtry + min.node.size + replace + splitrule	16	0.30	$0.5N$	0.69 (0.01)	1.15 (0.45)	0.31	1527.70 (0246.60)
mtry + min.node.size + sample.fraction + replace	16	0.30	$0.5N$	0.68 (0.01)	1.11 (0.36)	0.26	2935.90 (0310.70)
mtry + min.node.size + sample.fraction + splitrule	16	0.30	$0.5N$	0.69 (0.01)	1.16 (0.36)	0.29	4260.30 (0566.50)
mtry + min.node.size + sample.fraction + replace + splitrule	16	0.30	$0.5N$	0.69 (0.01)	1.14 (0.36)	0.29	9490.00 (1334.30)
None	16	0.30	$1N$	0.70 (0.01)	1.06 (0.14)	0.11	2.70 (0000.40)
mtry + min.node.size	16	0.30	$1N$	0.70 (0.01)	1.18 (0.24)	0.22	472.20 (0053.40)
mtry + min.node.size + replace	16	0.30	$1N$	0.70 (0.01)	1.17 (0.23)	0.21	1081.50 (0122.60)
mtry + min.node.size + splitrule	16	0.30	$1N$	0.70 (0.01)	1.26 (0.29)	0.28	1448.90 (0219.90)
mtry + min.node.size + sample.fraction	16	0.30	$1N$	0.70 (0.01)	1.18 (0.27)	0.22	3027.10 (0329.60)
mtry + min.node.size + replace + splitrule	16	0.30	$1N$	0.70 (0.01)	1.22 (0.30)	0.26	3422.30 (0574.20)
mtry + min.node.size + sample.fraction + replace	16	0.30	$1N$	0.70 (0.01)	1.15 (0.22)	0.20	6648.40 (0738.10)
mtry + min.node.size + sample.fraction + splitrule	16	0.30	$1N$	0.70 (0.01)	1.22 (0.28)	0.25	9222.00 (1335.20)
mtry + min.node.size + sample.fraction + replace + splitrule	16	0.30	$1N$	0.70 (0.01)	1.18 (0.28)	0.24	20665.90 (3156.00)

Table S4: Average performance of hyperparameter combinations for each scenario where $EF = 0.5$. Within each scenario, rows are sorted in ascending order of runtime.

Tuned hyperparameters	p	EF	n	AUC	Calibration slope	RMSD(slope)	Runtime (seconds)
				Mean (SD)	Median (IQR)		Mean (SD)
None	8	0.50	$0.5N$	0.70 (0.01)	0.92 (0.16)	0.15	0.80 (0000.30)
mtry + min.node.size	8	0.50	$0.5N$	0.70 (0.01)	1.01 (0.31)	0.22	44.70 (0006.40)
mtry + min.node.size + replace	8	0.50	$0.5N$	0.69 (0.01)	0.99 (0.31)	0.22	97.40 (0014.10)
mtry + min.node.size + splitrule	8	0.50	$0.5N$	0.70 (0.01)	1.14 (0.38)	0.32	144.10 (0022.30)
mtry + min.node.size + replace + splitrule	8	0.50	$0.5N$	0.70 (0.01)	1.08 (0.42)	0.32	318.90 (0047.20)
mtry + min.node.size + sample.fraction	8	0.50	$0.5N$	0.70 (0.01)	1.05 (0.32)	0.25	339.70 (0047.40)
mtry + min.node.size + sample.fraction + replace	8	0.50	$0.5N$	0.69 (0.01)	1.03 (0.36)	0.28	718.80 (0106.80)
mtry + min.node.size + sample.fraction + splitrule	8	0.50	$0.5N$	0.70 (0.01)	1.11 (0.35)	0.30	1077.40 (0161.60)
mtry + min.node.size + sample.fraction + replace + splitrule	8	0.50	$0.5N$	0.70 (0.01)	1.08 (0.42)	0.31	2305.40 (0329.20)
None	8	0.50	$1N$	0.72 (0.01)	1.01 (0.15)	0.10	1.30 (0000.40)
mtry + min.node.size	8	0.50	$1N$	0.71 (0.01)	1.06 (0.23)	0.17	85.50 (0010.90)
mtry + min.node.size + replace	8	0.50	$1N$	0.71 (0.01)	1.05 (0.25)	0.17	189.70 (0023.70)
mtry + min.node.size + splitrule	8	0.50	$1N$	0.72 (0.01)	1.18 (0.26)	0.24	273.10 (0042.90)
mtry + min.node.size + sample.fraction	8	0.50	$1N$	0.72 (0.01)	1.07 (0.22)	0.18	602.20 (0077.50)
mtry + min.node.size + replace + splitrule	8	0.50	$1N$	0.72 (0.01)	1.15 (0.29)	0.23	610.60 (0100.70)
mtry + min.node.size + sample.fraction + replace	8	0.50	$1N$	0.71 (0.01)	1.06 (0.26)	0.19	1296.40 (0160.10)
mtry + min.node.size + sample.fraction + splitrule	8	0.50	$1N$	0.72 (0.01)	1.19 (0.29)	0.25	1894.80 (0274.20)
mtry + min.node.size + sample.fraction + replace + splitrule	8	0.50	$1N$	0.72 (0.01)	1.15 (0.29)	0.23	4123.40 (0613.40)
None	16	0.50	$0.5N$	0.69 (0.01)	1.04 (0.15)	0.11	1.30 (0000.70)
mtry + min.node.size	16	0.50	$0.5N$	0.69 (0.01)	1.16 (0.35)	0.24	177.60 (0019.50)
mtry + min.node.size + replace	16	0.50	$0.5N$	0.69 (0.01)	1.13 (0.31)	0.23	403.40 (0042.70)
mtry + min.node.size + splitrule	16	0.50	$0.5N$	0.70 (0.01)	1.23 (0.35)	0.30	581.00 (0080.20)
mtry + min.node.size + sample.fraction	16	0.50	$0.5N$	0.69 (0.01)	1.16 (0.31)	0.25	1195.20 (0126.80)
mtry + min.node.size + replace + splitrule	16	0.50	$0.5N$	0.70 (0.01)	1.18 (0.37)	0.29	1352.90 (0213.10)
mtry + min.node.size + sample.fraction + replace	16	0.50	$0.5N$	0.69 (0.01)	1.15 (0.32)	0.26	2604.40 (0271.80)
mtry + min.node.size + sample.fraction + splitrule	16	0.50	$0.5N$	0.70 (0.01)	1.23 (0.37)	0.29	3839.90 (0493.30)
mtry + min.node.size + sample.fraction + replace + splitrule	16	0.50	$0.5N$	0.70 (0.01)	1.16 (0.36)	0.28	8493.40 (1151.50)
None	16	0.50	$1N$	0.71 (0.00)	1.12 (0.12)	0.14	2.40 (0000.50)
mtry + min.node.size	16	0.50	$1N$	0.71 (0.00)	1.16 (0.23)	0.21	411.10 (0046.40)
mtry + min.node.size + replace	16	0.50	$1N$	0.71 (0.01)	1.16 (0.21)	0.20	941.80 (0107.00)
mtry + min.node.size + splitrule	16	0.50	$1N$	0.71 (0.01)	1.29 (0.24)	0.28	1283.50 (0196.90)
mtry + min.node.size + sample.fraction	16	0.50	$1N$	0.71 (0.01)	1.16 (0.21)	0.21	2637.80 (0300.30)
mtry + min.node.size + replace + splitrule	16	0.50	$1N$	0.71 (0.01)	1.23 (0.23)	0.25	3025.00 (0501.30)
mtry + min.node.size + sample.fraction + replace	16	0.50	$1N$	0.71 (0.01)	1.16 (0.22)	0.20	5798.00 (0657.90)
mtry + min.node.size + sample.fraction + splitrule	16	0.50	$1N$	0.71 (0.01)	1.27 (0.25)	0.27	8201.20 (1178.20)
mtry + min.node.size + sample.fraction + replace + splitrule	16	0.50	$1N$	0.71 (0.01)	1.22 (0.24)	0.25	18350.60 (2833.00)

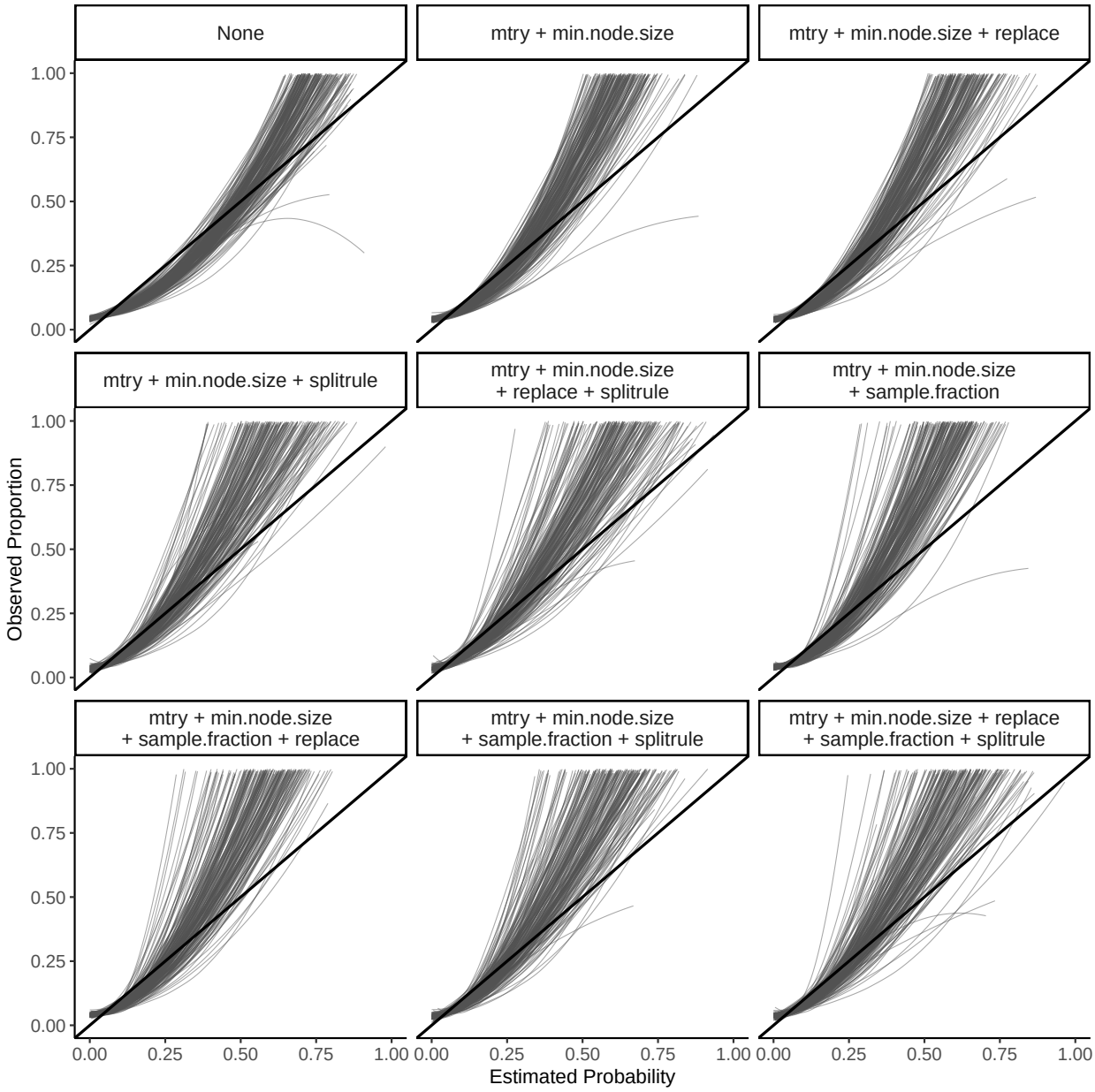

Figure S1: Calibration plots comparing hyperparameter combinations for every tenth dataset in scenarios where $EF = 0.1$.

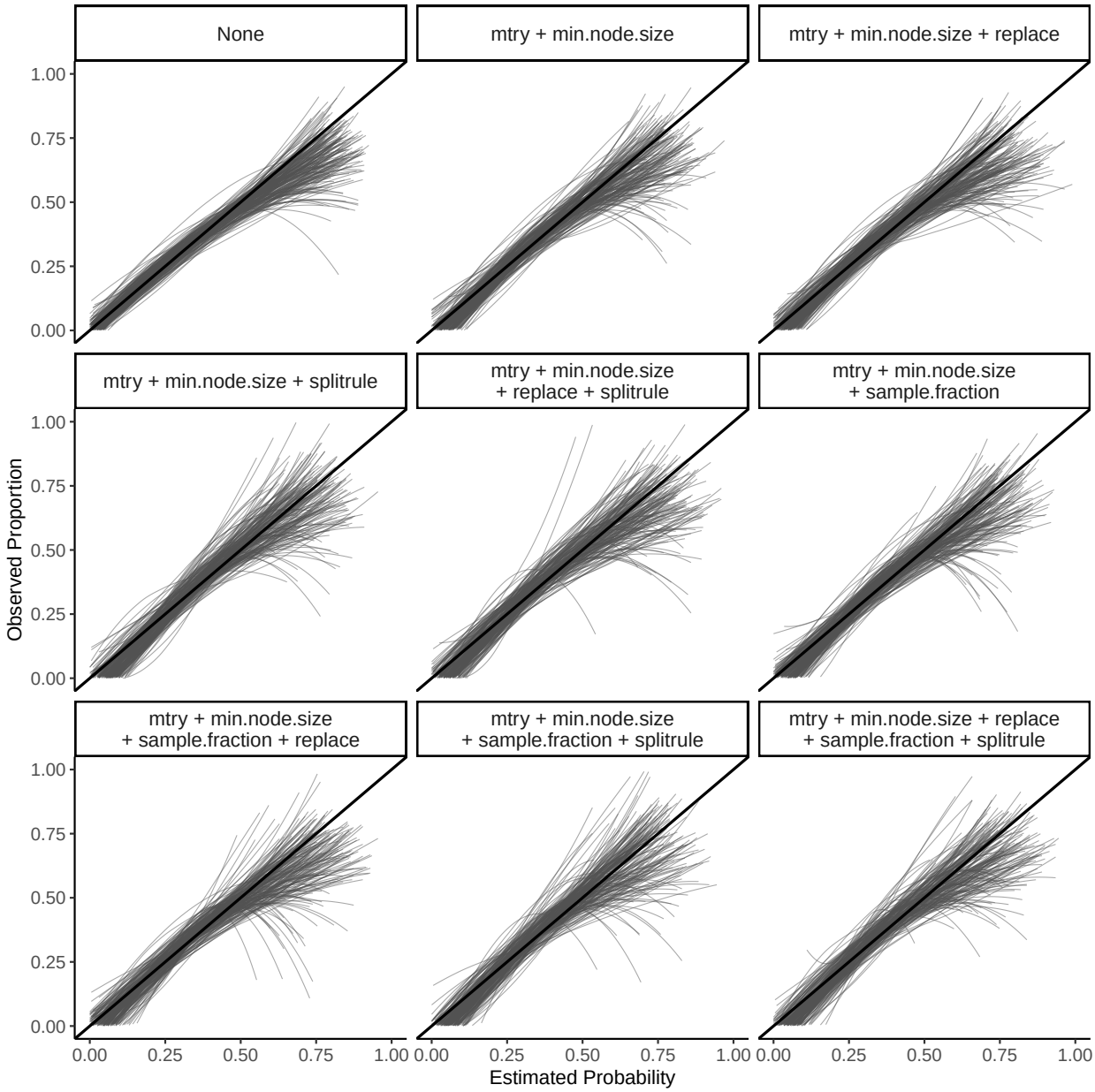

Figure S2: Calibration plots comparing hyperparameter combinations for every tenth dataset in scenarios where $EF = 0.3$.

Table S5: Average performance of optimisation criteria for each scenario where $EF = 0.1$. Within each scenario, rows are sorted in ascending order of RMSD(slope).

Optimisation criterion	p	EF	n	AUC	Calibration slope	RMSD(slope)	Runtime (seconds)
				Mean (SD)	Median (IQR)		Mean (SD)
Calibration intercept	8	0.10	$0.5N$	0.71 (0.01)	0.84 (0.15)	0.26	62.70 (010.70)
Logarithmic loss	8	0.10	$0.5N$	0.71 (0.01)	0.85 (0.17)	0.26	61.70 (010.40)
Brier score	8	0.10	$0.5N$	0.71 (0.02)	0.82 (0.19)	0.30	61.60 (010.20)
Calibration slope	8	0.10	$0.5N$	0.71 (0.02)	0.82 (0.19)	0.32	62.30 (010.20)
AUC	8	0.10	$0.5N$	0.70 (0.02)	0.74 (0.30)	0.49	61.90 (010.30)
Classification accuracy	8	0.10	$0.5N$	0.70 (0.02)	0.68 (0.26)	0.49	61.80 (010.40)
Cohen's Kappa	8	0.10	$0.5N$	0.69 (0.02)	0.49 (0.18)	0.76	61.90 (010.20)
Calibration intercept	8	0.10	$1N$	0.72 (0.01)	0.90 (0.11)	0.16	133.00 (019.20)
Logarithmic loss	8	0.10	$1N$	0.72 (0.01)	0.91 (0.13)	0.16	132.20 (018.90)
Calibration slope	8	0.10	$1N$	0.72 (0.01)	0.90 (0.14)	0.17	133.00 (019.30)
Brier score	8	0.10	$1N$	0.72 (0.01)	0.88 (0.18)	0.21	132.20 (018.90)
AUC	8	0.10	$1N$	0.72 (0.01)	0.86 (0.24)	0.32	132.60 (019.00)
Classification accuracy	8	0.10	$1N$	0.71 (0.01)	0.74 (0.21)	0.38	132.40 (019.10)
Cohen's Kappa	8	0.10	$1N$	0.71 (0.01)	0.58 (0.14)	0.59	132.50 (019.10)
Calibration intercept	16	0.10	$0.5N$	0.72 (0.01)	1.08 (0.13)	0.13	438.80 (069.00)
Logarithmic loss	16	0.10	$0.5N$	0.72 (0.01)	1.08 (0.17)	0.15	436.90 (069.40)
Brier score	16	0.10	$0.5N$	0.71 (0.01)	1.05 (0.24)	0.17	437.40 (069.90)
Calibration slope	16	0.10	$0.5N$	0.71 (0.01)	1.03 (0.26)	0.18	438.00 (069.80)
AUC	16	0.10	$0.5N$	0.71 (0.01)	1.05 (0.19)	0.22	437.20 (069.90)
Classification accuracy	16	0.10	$0.5N$	0.71 (0.01)	0.75 (0.23)	0.38	437.90 (070.20)
Cohen's Kappa	16	0.10	$0.5N$	0.70 (0.01)	0.55 (0.15)	0.62	437.50 (069.10)
Brier score	16	0.10	$1N$	0.72 (0.01)	1.08 (0.21)	0.14	1120.80 (163.60)
Calibration intercept	16	0.10	$1N$	0.72 (0.01)	1.13 (0.12)	0.14	1123.30 (161.80)
Calibration slope	16	0.10	$1N$	0.72 (0.01)	1.04 (0.22)	0.14	1125.30 (163.30)
Logarithmic loss	16	0.10	$1N$	0.72 (0.01)	1.13 (0.13)	0.15	1122.20 (166.30)
AUC	16	0.10	$1N$	0.72 (0.01)	1.11 (0.16)	0.18	1124.40 (161.60)
Classification accuracy	16	0.10	$1N$	0.72 (0.01)	0.80 (0.18)	0.28	1122.90 (161.90)
Cohen's Kappa	16	0.10	$1N$	0.71 (0.01)	0.65 (0.11)	0.45	1123.50 (162.50)

Table S6: Average performance of optimisation criteria for each scenario where $EF = 0.3$. Within each scenario, rows are sorted in ascending order of RMSD(slope).

Optimisation criterion	p	EF	n	AUC Mean (SD)	Calibration slope Median (IQR)	RMSD(slope)	Runtime (seconds) Mean (SD)
Calibration intercept	8	0.30	$0.5N$	0.68 (0.02)	0.99 (0.19)	0.17	39.10 (06.90)
Logarithmic loss	8	0.30	$0.5N$	0.68 (0.02)	0.95 (0.24)	0.22	38.40 (06.70)
Brier score	8	0.30	$0.5N$	0.68 (0.02)	0.94 (0.25)	0.24	38.30 (06.70)
Calibration slope	8	0.30	$0.5N$	0.68 (0.02)	0.92 (0.27)	0.25	39.10 (06.90)
AUC	8	0.30	$0.5N$	0.68 (0.02)	0.87 (0.29)	0.32	38.60 (06.80)
Classification accuracy	8	0.30	$0.5N$	0.68 (0.02)	0.84 (0.31)	0.33	38.50 (06.70)
Cohen's Kappa	8	0.30	$0.5N$	0.67 (0.02)	0.68 (0.22)	0.46	38.50 (06.60)
Calibration intercept	8	0.30	$1N$	0.70 (0.01)	1.12 (0.19)	0.17	70.10 (10.90)
Logarithmic loss	8	0.30	$1N$	0.70 (0.01)	1.06 (0.24)	0.17	69.20 (10.60)
AUC	8	0.30	$1N$	0.70 (0.01)	1.01 (0.25)	0.18	69.60 (10.60)
Brier score	8	0.30	$1N$	0.70 (0.01)	1.04 (0.26)	0.19	69.20 (10.70)
Calibration slope	8	0.30	$1N$	0.70 (0.01)	1.03 (0.28)	0.19	70.00 (10.80)
Classification accuracy	8	0.30	$1N$	0.70 (0.01)	0.92 (0.28)	0.23	69.50 (10.90)
Cohen's Kappa	8	0.30	$1N$	0.70 (0.01)	0.79 (0.18)	0.30	69.70 (11.30)
Calibration slope	16	0.30	$0.5N$	0.68 (0.01)	1.06 (0.34)	0.22	190.00 (26.60)
Classification accuracy	16	0.30	$0.5N$	0.68 (0.01)	0.99 (0.33)	0.23	188.90 (26.00)
Brier score	16	0.30	$0.5N$	0.69 (0.01)	1.16 (0.31)	0.23	188.30 (25.60)
Logarithmic loss	16	0.30	$0.5N$	0.69 (0.01)	1.16 (0.34)	0.23	188.40 (25.90)
AUC	16	0.30	$0.5N$	0.69 (0.01)	1.18 (0.29)	0.24	188.90 (25.90)
Calibration intercept	16	0.30	$0.5N$	0.69 (0.01)	1.27 (0.21)	0.27	190.30 (26.10)
Cohen's Kappa	16	0.30	$0.5N$	0.67 (0.01)	0.81 (0.17)	0.28	188.70 (25.80)
Calibration slope	16	0.30	$1N$	0.70 (0.01)	1.04 (0.23)	0.17	433.20 (56.60)
Cohen's Kappa	16	0.30	$1N$	0.69 (0.01)	0.90 (0.17)	0.18	431.00 (55.20)
Classification accuracy	16	0.30	$1N$	0.70 (0.01)	1.02 (0.29)	0.20	432.30 (60.70)
Brier score	16	0.30	$1N$	0.70 (0.01)	1.17 (0.27)	0.22	429.30 (54.20)
Logarithmic loss	16	0.30	$1N$	0.70 (0.01)	1.15 (0.25)	0.22	431.70 (58.00)
AUC	16	0.30	$1N$	0.70 (0.01)	1.26 (0.23)	0.26	431.00 (55.40)
Calibration intercept	16	0.30	$1N$	0.70 (0.01)	1.35 (0.18)	0.31	432.10 (55.50)

Table S7: Average performance of optimisation criteria for each scenario where $EF = 0.5$. Within each scenario, rows are sorted in ascending order of RMSD(slope).

Optimisation criterion	p	EF	n	AUC	Calibration slope	RMSD(slope)	Runtime (seconds)
				Mean (SD)	Median (IQR)		Mean (SD)
Calibration intercept	8	0.50	$0.5N$	0.69 (0.01)	1.10 (0.22)	0.18	44.50 (07.10)
Brier score	8	0.50	$0.5N$	0.69 (0.01)	1.01 (0.32)	0.21	43.60 (07.00)
Logarithmic loss	8	0.50	$0.5N$	0.69 (0.01)	1.01 (0.30)	0.21	43.60 (07.10)
AUC	8	0.50	$0.5N$	0.69 (0.01)	0.97 (0.27)	0.22	43.90 (07.00)
Calibration slope	8	0.50	$0.5N$	0.69 (0.01)	0.98 (0.29)	0.23	44.50 (07.10)
Classification accuracy	8	0.50	$0.5N$	0.69 (0.01)	0.90 (0.27)	0.24	43.70 (06.90)
Cohen's Kappa	8	0.50	$0.5N$	0.69 (0.01)	0.91 (0.30)	0.25	43.70 (07.00)
AUC	8	0.50	$1N$	0.71 (0.01)	1.05 (0.24)	0.16	82.50 (15.80)
Logarithmic loss	8	0.50	$1N$	0.71 (0.01)	1.03 (0.24)	0.16	81.70 (10.70)
Classification accuracy	8	0.50	$1N$	0.71 (0.01)	1.02 (0.25)	0.17	81.70 (10.50)
Brier score	8	0.50	$1N$	0.71 (0.01)	1.03 (0.23)	0.17	81.80 (10.70)
Calibration slope	8	0.50	$1N$	0.71 (0.01)	1.03 (0.27)	0.18	82.60 (10.90)
Cohen's Kappa	8	0.50	$1N$	0.71 (0.01)	0.97 (0.25)	0.18	82.00 (10.60)
Calibration intercept	8	0.50	$1N$	0.71 (0.01)	1.19 (0.17)	0.21	82.50 (10.70)
Calibration slope	16	0.50	$0.5N$	0.69 (0.01)	1.03 (0.32)	0.22	165.70 (22.20)
Cohen's Kappa	16	0.50	$0.5N$	0.69 (0.01)	1.08 (0.34)	0.22	164.50 (21.50)
Classification accuracy	16	0.50	$0.5N$	0.69 (0.01)	1.10 (0.37)	0.23	164.20 (21.40)
AUC	16	0.50	$0.5N$	0.69 (0.01)	1.19 (0.32)	0.24	164.80 (22.10)
Logarithmic loss	16	0.50	$0.5N$	0.69 (0.01)	1.15 (0.31)	0.24	164.20 (21.80)
Brier score	16	0.50	$0.5N$	0.69 (0.01)	1.17 (0.33)	0.25	164.20 (21.70)
Calibration intercept	16	0.50	$0.5N$	0.70 (0.01)	1.32 (0.23)	0.31	165.60 (22.50)
Calibration slope	16	0.50	$1N$	0.71 (0.01)	1.05 (0.22)	0.17	372.50 (46.60)
Classification accuracy	16	0.50	$1N$	0.71 (0.01)	1.16 (0.29)	0.23	370.60 (46.50)
Brier score	16	0.50	$1N$	0.71 (0.01)	1.19 (0.26)	0.23	370.00 (46.00)
Cohen's Kappa	16	0.50	$1N$	0.71 (0.01)	1.17 (0.30)	0.23	371.30 (48.20)
Logarithmic loss	16	0.50	$1N$	0.71 (0.01)	1.19 (0.25)	0.23	370.50 (46.10)
AUC	16	0.50	$1N$	0.71 (0.01)	1.25 (0.28)	0.26	371.50 (47.30)
Calibration intercept	16	0.50	$1N$	0.71 (0.01)	1.40 (0.23)	0.34	372.40 (46.30)

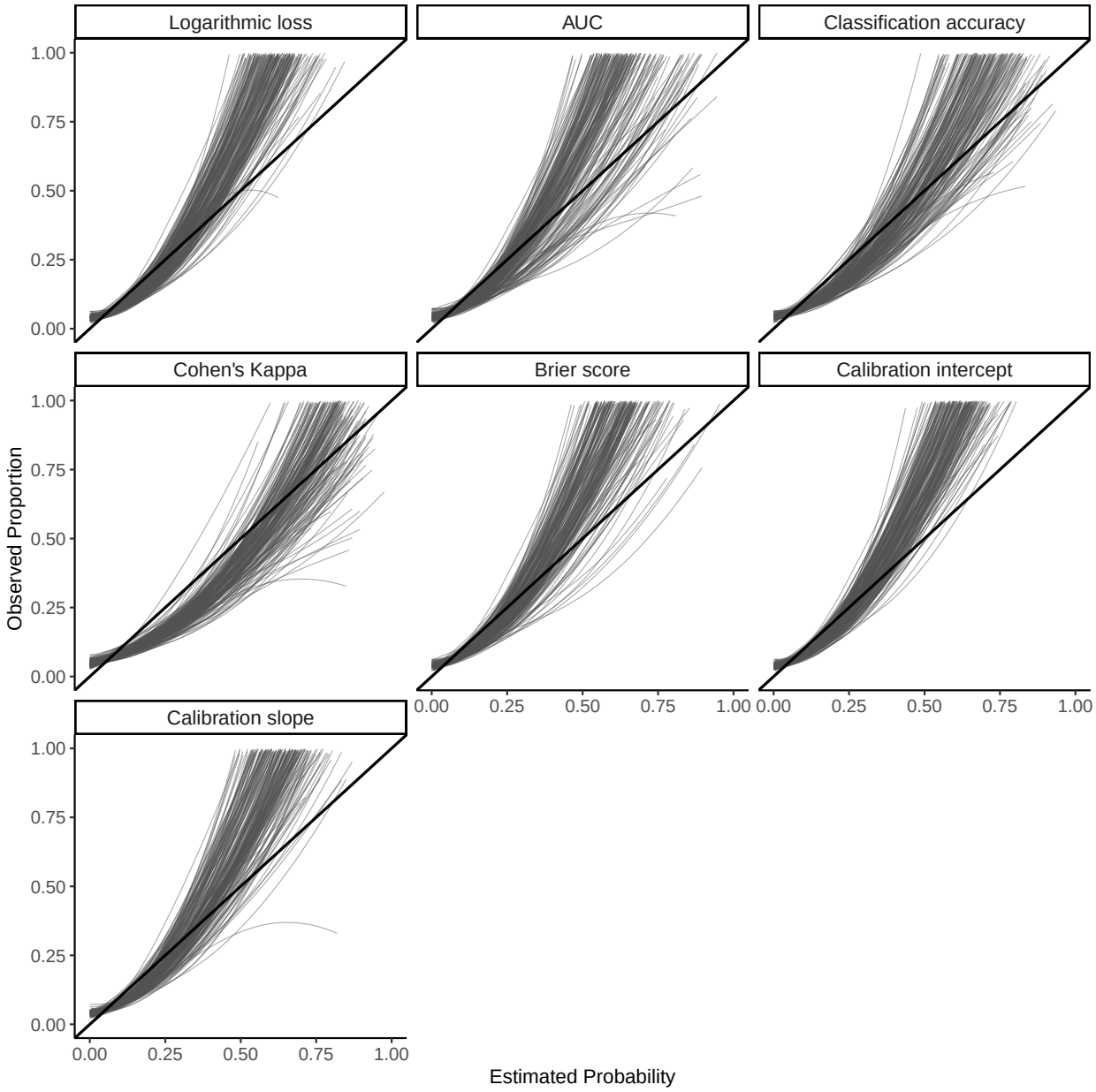

Figure S3: Calibration plots comparing optimisation criteria for every tenth dataset in scenarios where $EF = 0.1$.

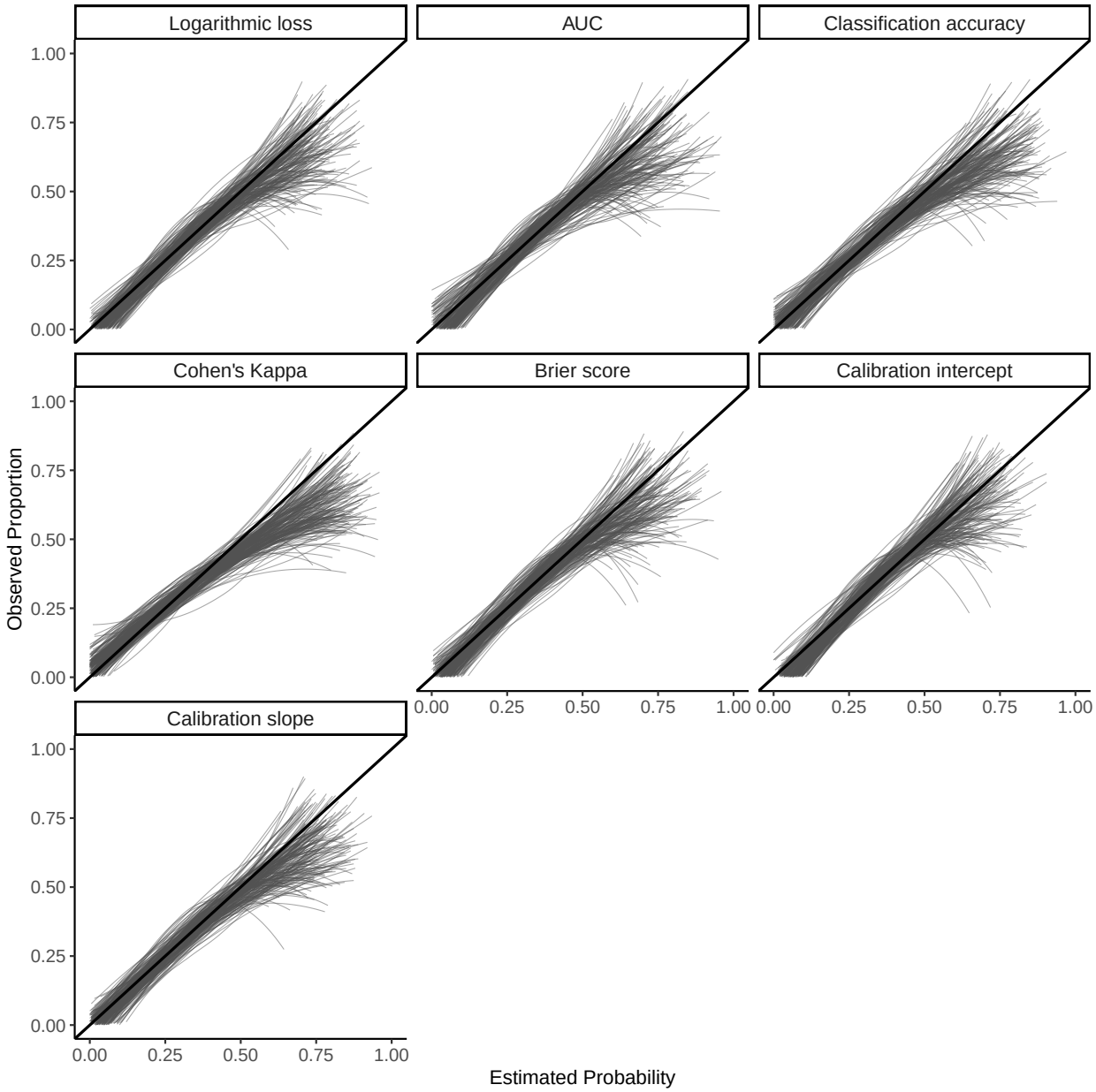

Figure S4: Calibration plots comparing optimisation criteria for every tenth dataset in scenarios where $EF = 0.3$.

Table S8: Average performance of hyperparameter search algorithms for each scenario. Within each scenario, rows are sorted in ascending order of runtime.

Search algorithm	p	EF	n	AUC	Calibration slope	RMSD(slope)	Runtime (seconds)
				Mean (SD)	Median (IQR)		Mean (SD)
Random search	8	0.10	$0.5N$	0.70 (0.01)	0.88 (0.19)	0.26	29.43 (005.12)
Model-based optimisation	8	0.10	$0.5N$	0.70 (0.01)	0.89 (0.22)	0.23	53.21 (011.51)
Grid search	8	0.10	$0.5N$	0.70 (0.01)	0.82 (0.16)	0.27	63.80 (011.09)
Random search	8	0.10	$1N$	0.72 (0.01)	0.96 (0.19)	0.17	58.48 (009.21)
Model-based optimisation	8	0.10	$1N$	0.72 (0.01)	0.96 (0.17)	0.16	66.44 (012.70)
Grid search	8	0.10	$1N$	0.72 (0.01)	0.91 (0.14)	0.16	136.30 (019.75)
Random search	8	0.30	$0.5N$	0.68 (0.02)	0.95 (0.34)	0.28	16.48 (003.76)
Grid search	8	0.30	$0.5N$	0.68 (0.02)	0.93 (0.25)	0.25	38.01 (005.97)
Model-based optimisation	8	0.30	$0.5N$	0.68 (0.02)	0.97 (0.32)	0.26	44.82 (008.17)
Random search	8	0.30	$1N$	0.70 (0.01)	1.04 (0.30)	0.21	24.63 (003.60)
Model-based optimisation	8	0.30	$1N$	0.70 (0.01)	1.04 (0.26)	0.20	54.92 (016.50)
Grid search	8	0.30	$1N$	0.70 (0.01)	1.03 (0.24)	0.17	71.64 (009.49)
Random search	8	0.50	$0.5N$	0.69 (0.01)	1.00 (0.33)	0.26	15.39 (002.08)
Grid search	8	0.50	$0.5N$	0.69 (0.01)	1.03 (0.30)	0.21	43.26 (005.19)
Model-based optimisation	8	0.50	$0.5N$	0.69 (0.01)	1.02 (0.33)	0.24	47.22 (009.69)
Random search	8	0.50	$1N$	0.71 (0.01)	1.05 (0.26)	0.19	23.97 (010.40)
Model-based optimisation	8	0.50	$1N$	0.71 (0.01)	1.05 (0.26)	0.17	58.76 (016.73)
Grid search	8	0.50	$1N$	0.71 (0.01)	1.07 (0.25)	0.17	86.61 (011.38)
Model-based optimisation	16	0.10	$0.5N$	0.71 (0.01)	1.12 (0.17)	0.17	86.04 (023.26)
Random search	16	0.10	$0.5N$	0.71 (0.01)	1.12 (0.22)	0.18	198.13 (031.91)
Grid search	16	0.10	$0.5N$	0.71 (0.01)	1.08 (0.17)	0.14	448.33 (066.24)
Model-based optimisation	16	0.10	$1N$	0.72 (0.01)	1.14 (0.13)	0.16	142.73 (022.72)
Random search	16	0.10	$1N$	0.72 (0.01)	1.14 (0.19)	0.17	476.86 (073.43)
Grid search	16	0.10	$1N$	0.72 (0.01)	1.12 (0.12)	0.14	1137.30 (150.00)
Model-based optimisation	16	0.30	$0.5N$	0.68 (0.01)	1.14 (0.36)	0.26	53.88 (009.08)
Random search	16	0.30	$0.5N$	0.68 (0.01)	1.13 (0.37)	0.26	65.38 (008.06)
Grid search	16	0.30	$0.5N$	0.69 (0.01)	1.13 (0.32)	0.23	186.15 (021.96)
Model-based optimisation	16	0.30	$1N$	0.70 (0.01)	1.16 (0.29)	0.22	93.41 (028.58)
Random search	16	0.30	$1N$	0.70 (0.01)	1.16 (0.26)	0.22	125.42 (018.80)
Grid search	16	0.30	$1N$	0.70 (0.01)	1.19 (0.27)	0.23	451.93 (057.52)
Random search	16	0.50	$0.5N$	0.69 (0.01)	1.11 (0.32)	0.25	49.84 (006.66)
Model-based optimisation	16	0.50	$0.5N$	0.69 (0.01)	1.14 (0.35)	0.25	56.82 (016.66)
Grid search	16	0.50	$0.5N$	0.69 (0.01)	1.14 (0.34)	0.24	167.45 (019.93)
Random search	16	0.50	$1N$	0.71 (0.01)	1.13 (0.21)	0.20	85.02 (013.68)
Model-based optimisation	16	0.50	$1N$	0.71 (0.01)	1.11 (0.24)	0.20	87.69 (021.43)
Grid search	16	0.50	$1N$	0.71 (0.01)	1.15 (0.22)	0.20	388.60 (053.45)

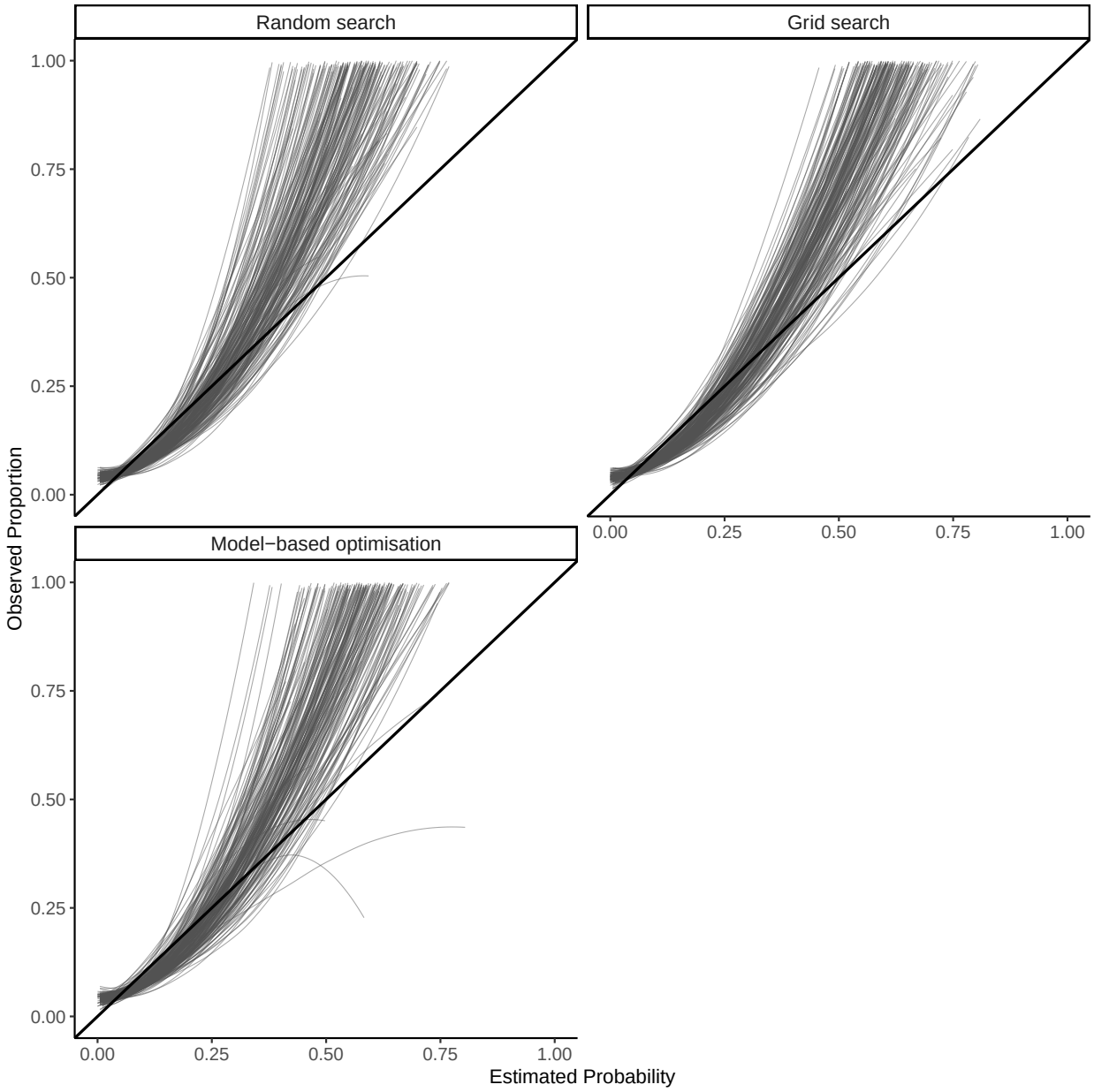

Figure S5: Calibration plots comparing hyperparameter search algorithms for every tenth dataset in scenarios where $EF = 0.1$.

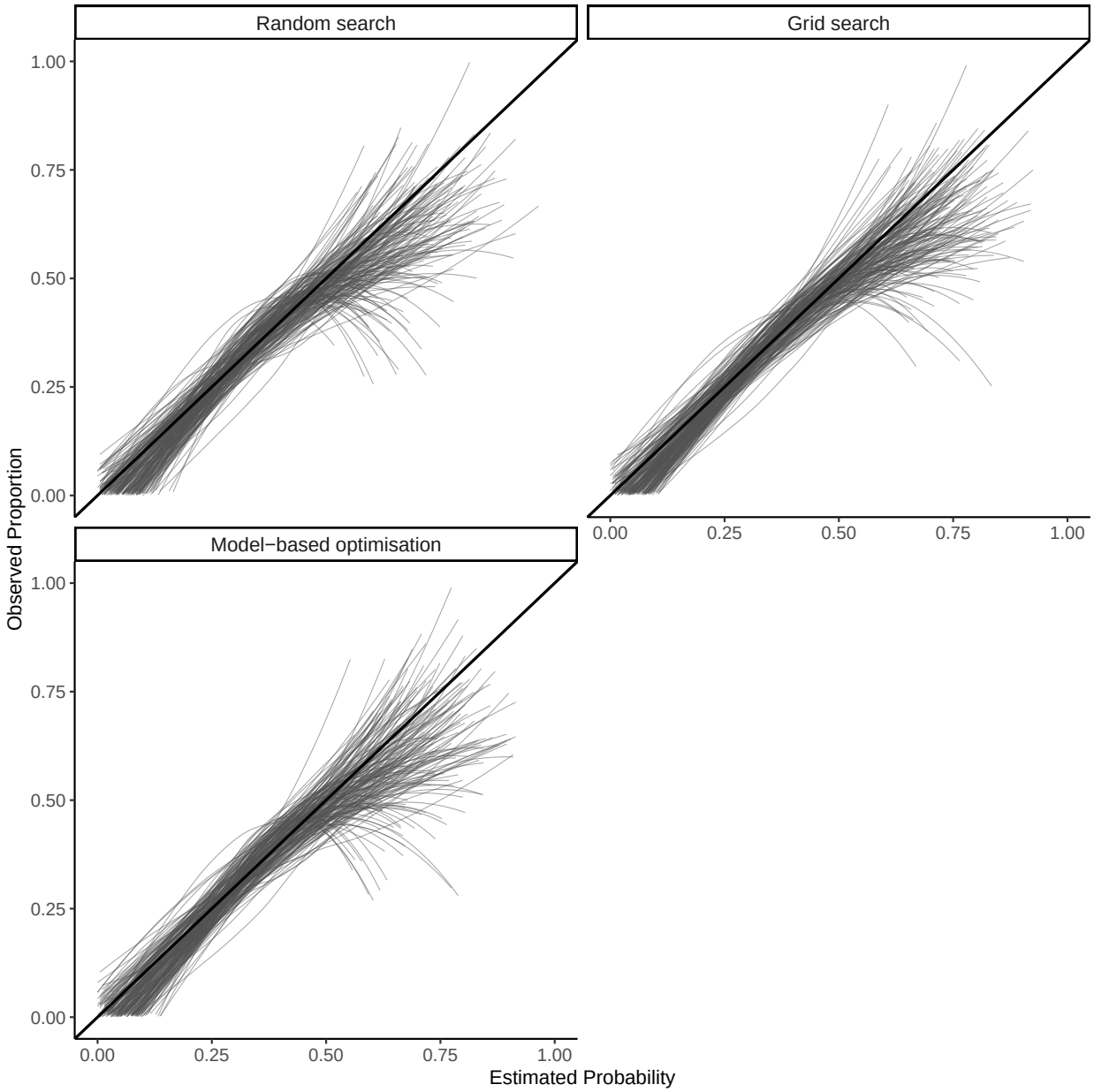

Figure S6: Calibration plots comparing hyperparameter search algorithms for every tenth dataset in scenarios where $EF = 0.3$.