# 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 | 0.98 (0000.50) |
| mtry + min.node.size | 8 | 0.10 | $0.5N$ | 0.70 (0.01) | 0.83 (0.16) | 0.27 | 63.05 (0009.69) |
| mtry + min.node.size + replace | 8 | 0.10 | $0.5N$ | 0.70 (0.01) | 0.83 (0.17) | 0.27 | 141.44 (0021.91) |
| mtry + min.node.size + splitrule | 8 | 0.10 | $0.5N$ | 0.71 (0.01) | 1.00 (0.34) | 0.26 | 189.66 (0032.02) |
| mtry + min.node.size + replace + splitrule | 8 | 0.10 | $0.5N$ | 0.71 (0.01) | 0.97 (0.38) | 0.30 | 425.03 (0078.16) |
| mtry + min.node.size + sample.fraction | 8 | 0.10 | $0.5N$ | 0.71 (0.01) | 0.96 (0.36) | 0.28 | 455.60 (0070.43) |
| mtry + min.node.size + sample.fraction + replace | 8 | 0.10 | $0.5N$ | 0.71 (0.01) | 0.97 (0.39) | 0.30 | 982.88 (0148.73) |
| mtry + min.node.size + sample.fraction + splitrule | 8 | 0.10 | $0.5N$ | 0.71 (0.01) | 1.03 (0.37) | 0.27 | 1370.38 (0231.33) |
| mtry + min.node.size + sample.fraction + replace + splitrule | 8 | 0.10 | $0.5N$ | 0.71 (0.02) | 0.98 (0.38) | 0.29 | 2948.55 (0483.62) |
| None | 8 | 0.10 | $1N$ | 0.71 (0.01) | 0.74 (0.10) | 0.32 | 1.71 (0000.42) |
| mtry + min.node.size | 8 | 0.10 | $1N$ | 0.72 (0.01) | 0.90 (0.12) | 0.17 | 138.54 (0019.44) |
| mtry + min.node.size + replace | 8 | 0.10 | $1N$ | 0.72 (0.01) | 0.89 (0.12) | 0.18 | 314.32 (0047.10) |
| mtry + min.node.size + splitrule | 8 | 0.10 | $1N$ | 0.73 (0.01) | 1.07 (0.25) | 0.21 | 396.28 (0067.72) |
| mtry + min.node.size + replace + splitrule | 8 | 0.10 | $1N$ | 0.73 (0.01) | 1.04 (0.26) | 0.20 | 900.20 (0168.08) |
| mtry + min.node.size + sample.fraction | 8 | 0.10 | $1N$ | 0.72 (0.01) | 0.98 (0.22) | 0.18 | 936.02 (0129.23) |
| mtry + min.node.size + sample.fraction + replace | 8 | 0.10 | $1N$ | 0.72 (0.01) | 0.98 (0.23) | 0.19 | 2034.94 (0285.71) |
| mtry + min.node.size + sample.fraction + splitrule | 8 | 0.10 | $1N$ | 0.73 (0.01) | 1.07 (0.27) | 0.20 | 2673.90 (0435.76) |
| mtry + min.node.size + sample.fraction + replace + splitrule | 8 | 0.10 | $1N$ | 0.73 (0.01) | 1.06 (0.26) | 0.20 | 5850.18 (0973.79) |
| None | 16 | 0.10 | $0.5N$ | 0.71 (0.01) | 0.72 (0.11) | 0.34 | 2.47 (0000.37) |
| mtry + min.node.size | 16 | 0.10 | $0.5N$ | 0.72 (0.01) | 1.08 (0.16) | 0.14 | 479.18 (0062.67) |
| mtry + min.node.size + replace | 16 | 0.10 | $0.5N$ | 0.71 (0.01) | 1.07 (0.18) | 0.15 | 1121.57 (0162.11) |
| mtry + min.node.size + splitrule | 16 | 0.10 | $0.5N$ | 0.72 (0.01) | 1.14 (0.31) | 0.24 | 1274.03 (0201.87) |
| mtry + min.node.size + sample.fraction | 16 | 0.10 | $0.5N$ | 0.72 (0.01) | 1.12 (0.27) | 0.21 | 2958.45 (0361.24) |
| mtry + min.node.size + replace + splitrule | 16 | 0.10 | $0.5N$ | 0.72 (0.01) | 1.13 (0.31) | 0.23 | 3024.95 (0523.16) |
| mtry + min.node.size + sample.fraction + replace | 16 | 0.10 | $0.5N$ | 0.71 (0.01) | 1.09 (0.27) | 0.21 | 6644.06 (0840.19) |
| mtry + min.node.size + sample.fraction + splitrule | 16 | 0.10 | $0.5N$ | 0.72 (0.01) | 1.15 (0.32) | 0.24 | 7997.78 (1145.27) |
| mtry + min.node.size + sample.fraction + replace + splitrule | 16 | 0.10 | $0.5N$ | 0.72 (0.01) | 1.15 (0.31) | 0.25 | 17940.93 (2766.98) |
| None | 16 | 0.10 | $1N$ | 0.72 (0.01) | 0.80 (0.08) | 0.23 | 5.31 (0000.75) |
| mtry + min.node.size | 16 | 0.10 | $1N$ | 0.72 (0.01) | 1.12 (0.10) | 0.14 | 1153.99 (0155.01) |
| mtry + min.node.size + replace | 16 | 0.10 | $1N$ | 0.72 (0.01) | 1.11 (0.14) | 0.13 | 2674.04 (0374.50) |
| mtry + min.node.size + splitrule | 16 | 0.10 | $1N$ | 0.73 (0.01) | 1.16 (0.24) | 0.21 | 2797.31 (0481.07) |
| mtry + min.node.size + replace + splitrule | 16 | 0.10 | $1N$ | 0.73 (0.01) | 1.17 (0.24) | 0.21 | 6561.26 (1212.50) |
| mtry + min.node.size + sample.fraction | 16 | 0.10 | $1N$ | 0.72 (0.01) | 1.14 (0.14) | 0.17 | 6998.69 (0906.25) |
| mtry + min.node.size + sample.fraction + replace | 16 | 0.10 | $1N$ | 0.72 (0.01) | 1.12 (0.17) | 0.16 | 15623.88 (2054.65) |
| mtry + min.node.size + sample.fraction + splitrule | 16 | 0.10 | $1N$ | 0.73 (0.01) | 1.16 (0.23) | 0.21 | 17164.91 (2802.06) |
| mtry + min.node.size + sample.fraction + replace + splitrule | 16 | 0.10 | $1N$ | 0.73 (0.01) | 1.15 (0.21) | 0.20 | 38521.23 (6617.74) |

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 | 0.79 (0001.22) |
| mtry + min.node.size | 8 | 0.30 | $0.5N$ | 0.68 (0.02) | 0.93 (0.25) | 0.25 | 38.33 (0005.68) |
| mtry + min.node.size + replace | 8 | 0.30 | $0.5N$ | 0.68 (0.02) | 0.92 (0.29) | 0.28 | 83.42 (0011.75) |
| mtry + min.node.size + splitrule | 8 | 0.30 | $0.5N$ | 0.69 (0.02) | 1.07 (0.51) | 0.36 | 123.86 (0018.38) |
| mtry + min.node.size + replace + splitrule | 8 | 0.30 | $0.5N$ | 0.69 (0.02) | 1.01 (0.51) | 0.38 | 273.48 (0042.34) |
| mtry + min.node.size + sample.fraction | 8 | 0.30 | $0.5N$ | 0.68 (0.01) | 1.00 (0.34) | 0.29 | 302.85 (0044.93) |
| mtry + min.node.size + sample.fraction + replace | 8 | 0.30 | $0.5N$ | 0.68 (0.02) | 0.95 (0.37) | 0.34 | 635.75 (0091.97) |
| mtry + min.node.size + sample.fraction + splitrule | 8 | 0.30 | $0.5N$ | 0.69 (0.02) | 1.05 (0.44) | 0.35 | 953.61 (0144.84) |
| mtry + min.node.size + sample.fraction + replace + splitrule | 8 | 0.30 | $0.5N$ | 0.68 (0.02) | 1.01 (0.46) | 0.35 | 2026.45 (0324.34) |
| None | 8 | 0.30 | $1N$ | 0.70 (0.01) | 0.92 (0.14) | 0.14 | 1.14 (0000.68) |
| mtry + min.node.size | 8 | 0.30 | $1N$ | 0.70 (0.01) | 1.03 (0.25) | 0.18 | 72.02 (0009.02) |
| mtry + min.node.size + replace | 8 | 0.30 | $1N$ | 0.70 (0.01) | 1.01 (0.26) | 0.18 | 159.78 (0020.23) |
| mtry + min.node.size + splitrule | 8 | 0.30 | $1N$ | 0.71 (0.01) | 1.15 (0.28) | 0.26 | 228.50 (0034.28) |
| mtry + min.node.size + replace + splitrule | 8 | 0.30 | $1N$ | 0.71 (0.01) | 1.11 (0.33) | 0.26 | 512.87 (0082.53) |
| mtry + min.node.size + sample.fraction | 8 | 0.30 | $1N$ | 0.70 (0.01) | 1.04 (0.27) | 0.20 | 513.89 (0065.61) |
| mtry + min.node.size + sample.fraction + replace | 8 | 0.30 | $1N$ | 0.70 (0.01) | 1.01 (0.27) | 0.20 | 1102.52 (0141.94) |
| mtry + min.node.size + sample.fraction + splitrule | 8 | 0.30 | $1N$ | 0.71 (0.01) | 1.15 (0.31) | 0.25 | 1619.35 (0236.59) |
| mtry + min.node.size + sample.fraction + replace + splitrule | 8 | 0.30 | $1N$ | 0.71 (0.01) | 1.10 (0.33) | 0.25 | 3508.36 (0521.32) |
| None | 16 | 0.30 | $0.5N$ | 0.68 (0.01) | 0.96 (0.16) | 0.13 | 1.34 (0000.26) |
| mtry + min.node.size | 16 | 0.30 | $0.5N$ | 0.69 (0.01) | 1.16 (0.30) | 0.24 | 202.17 (0022.69) |
| mtry + min.node.size + replace | 16 | 0.30 | $0.5N$ | 0.69 (0.01) | 1.10 (0.32) | 0.23 | 459.78 (0050.63) |
| mtry + min.node.size + splitrule | 16 | 0.30 | $0.5N$ | 0.69 (0.01) | 1.21 (0.40) | 0.32 | 649.56 (0094.33) |
| mtry + min.node.size + sample.fraction | 16 | 0.30 | $0.5N$ | 0.69 (0.01) | 1.16 (0.33) | 0.25 | 1345.60 (0142.27) |
| mtry + min.node.size + replace + splitrule | 16 | 0.30 | $0.5N$ | 0.69 (0.01) | 1.15 (0.45) | 0.31 | 1527.65 (0246.56) |
| mtry + min.node.size + sample.fraction + replace | 16 | 0.30 | $0.5N$ | 0.68 (0.01) | 1.11 (0.36) | 0.26 | 2935.87 (0310.67) |
| mtry + min.node.size + sample.fraction + splitrule | 16 | 0.30 | $0.5N$ | 0.69 (0.01) | 1.16 (0.36) | 0.29 | 4260.33 (0566.51) |
| mtry + min.node.size + sample.fraction + replace + splitrule | 16 | 0.30 | $0.5N$ | 0.69 (0.01) | 1.14 (0.36) | 0.29 | 9490.01 (1334.26) |
| None | 16 | 0.30 | $1N$ | 0.70 (0.01) | 1.06 (0.14) | 0.11 | 2.66 (0000.41) |
| mtry + min.node.size | 16 | 0.30 | $1N$ | 0.70 (0.01) | 1.18 (0.24) | 0.22 | 472.18 (0053.36) |
| mtry + min.node.size + replace | 16 | 0.30 | $1N$ | 0.70 (0.01) | 1.17 (0.23) | 0.21 | 1081.51 (0122.58) |
| mtry + min.node.size + splitrule | 16 | 0.30 | $1N$ | 0.70 (0.01) | 1.26 (0.29) | 0.28 | 1448.88 (0219.87) |
| mtry + min.node.size + sample.fraction | 16 | 0.30 | $1N$ | 0.70 (0.01) | 1.18 (0.27) | 0.22 | 3027.06 (0329.56) |
| mtry + min.node.size + replace + splitrule | 16 | 0.30 | $1N$ | 0.70 (0.01) | 1.22 (0.30) | 0.26 | 3422.28 (0574.22) |
| mtry + min.node.size + sample.fraction + replace | 16 | 0.30 | $1N$ | 0.70 (0.01) | 1.15 (0.22) | 0.20 | 6648.42 (0738.12) |
| mtry + min.node.size + sample.fraction + splitrule | 16 | 0.30 | $1N$ | 0.70 (0.01) | 1.22 (0.28) | 0.25 | 9221.98 (1335.15) |
| mtry + min.node.size + sample.fraction + replace + splitrule | 16 | 0.30 | $1N$ | 0.70 (0.01) | 1.18 (0.28) | 0.24 | 20665.92 (3156.03) |

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 | 44.68 (0006.42) |
| mtry + min.node.size + replace | 8 | 0.50 | $0.5N$ | 0.69 (0.01) | 0.99 (0.31) | 0.22 | 97.42 (0014.14) |
| mtry + min.node.size + splitrule | 8 | 0.50 | $0.5N$ | 0.70 (0.01) | 1.14 (0.38) | 0.32 | 144.09 (0022.26) |
| mtry + min.node.size + replace + splitrule | 8 | 0.50 | $0.5N$ | 0.70 (0.01) | 1.08 (0.42) | 0.32 | 318.90 (0047.24) |
| mtry + min.node.size + sample.fraction | 8 | 0.50 | $0.5N$ | 0.70 (0.01) | 1.05 (0.32) | 0.25 | 339.68 (0047.35) |
| mtry + min.node.size + sample.fraction + replace | 8 | 0.50 | $0.5N$ | 0.69 (0.01) | 1.03 (0.36) | 0.28 | 718.84 (0106.82) |
| mtry + min.node.size + sample.fraction + splitrule | 8 | 0.50 | $0.5N$ | 0.70 (0.01) | 1.11 (0.35) | 0.30 | 1077.38 (0161.58) |
| mtry + min.node.size + sample.fraction + replace + splitrule | 8 | 0.50 | $0.5N$ | 0.70 (0.01) | 1.08 (0.42) | 0.31 | 2305.38 (0329.17) |
| None | 8 | 0.50 | $1N$ | 0.72 (0.01) | 1.01 (0.15) | 0.10 | 1.27 (0000.37) |
| mtry + min.node.size | 8 | 0.50 | $1N$ | 0.71 (0.01) | 1.06 (0.23) | 0.17 | 85.51 (0010.92) |
| mtry + min.node.size + replace | 8 | 0.50 | $1N$ | 0.71 (0.01) | 1.05 (0.25) | 0.17 | 189.71 (0023.67) |
| mtry + min.node.size + splitrule | 8 | 0.50 | $1N$ | 0.72 (0.01) | 1.18 (0.26) | 0.24 | 273.10 (0042.92) |
| mtry + min.node.size + sample.fraction | 8 | 0.50 | $1N$ | 0.72 (0.01) | 1.07 (0.22) | 0.18 | 602.18 (0077.47) |
| mtry + min.node.size + replace + splitrule | 8 | 0.50 | $1N$ | 0.72 (0.01) | 1.15 (0.29) | 0.23 | 610.57 (0100.75) |
| mtry + min.node.size + sample.fraction + replace | 8 | 0.50 | $1N$ | 0.71 (0.01) | 1.06 (0.26) | 0.19 | 1296.39 (0160.06) |
| mtry + min.node.size + sample.fraction + splitrule | 8 | 0.50 | $1N$ | 0.72 (0.01) | 1.19 (0.29) | 0.25 | 1894.82 (0274.21) |
| mtry + min.node.size + sample.fraction + replace + splitrule | 8 | 0.50 | $1N$ | 0.72 (0.01) | 1.15 (0.29) | 0.23 | 4123.41 (0613.37) |
| None | 16 | 0.50 | $0.5N$ | 0.69 (0.01) | 1.04 (0.15) | 0.11 | 1.26 (0000.67) |
| mtry + min.node.size | 16 | 0.50 | $0.5N$ | 0.69 (0.01) | 1.16 (0.35) | 0.24 | 177.63 (0019.48) |
| mtry + min.node.size + replace | 16 | 0.50 | $0.5N$ | 0.69 (0.01) | 1.13 (0.31) | 0.23 | 403.39 (0042.74) |
| mtry + min.node.size + splitrule | 16 | 0.50 | $0.5N$ | 0.70 (0.01) | 1.23 (0.35) | 0.30 | 580.95 (0080.17) |
| mtry + min.node.size + sample.fraction | 16 | 0.50 | $0.5N$ | 0.69 (0.01) | 1.16 (0.31) | 0.25 | 1195.21 (0126.77) |
| mtry + min.node.size + replace + splitrule | 16 | 0.50 | $0.5N$ | 0.70 (0.01) | 1.18 (0.37) | 0.29 | 1352.95 (0213.08) |
| mtry + min.node.size + sample.fraction + replace | 16 | 0.50 | $0.5N$ | 0.69 (0.01) | 1.15 (0.32) | 0.26 | 2604.36 (0271.77) |
| mtry + min.node.size + sample.fraction + splitrule | 16 | 0.50 | $0.5N$ | 0.70 (0.01) | 1.23 (0.37) | 0.29 | 3839.93 (0493.26) |
| mtry + min.node.size + sample.fraction + replace + splitrule | 16 | 0.50 | $0.5N$ | 0.70 (0.01) | 1.16 (0.36) | 0.28 | 8493.39 (1151.49) |
| None | 16 | 0.50 | $1N$ | 0.71 (0.00) | 1.12 (0.12) | 0.14 | 2.39 (0000.53) |
| mtry + min.node.size | 16 | 0.50 | $1N$ | 0.71 (0.00) | 1.16 (0.23) | 0.21 | 411.09 (0046.37) |
| mtry + min.node.size + replace | 16 | 0.50 | $1N$ | 0.71 (0.01) | 1.16 (0.21) | 0.20 | 941.76 (0107.01) |
| mtry + min.node.size + splitrule | 16 | 0.50 | $1N$ | 0.71 (0.01) | 1.29 (0.24) | 0.28 | 1283.49 (0196.93) |
| mtry + min.node.size + sample.fraction | 16 | 0.50 | $1N$ | 0.71 (0.01) | 1.16 (0.21) | 0.21 | 2637.83 (0300.31) |
| mtry + min.node.size + replace + splitrule | 16 | 0.50 | $1N$ | 0.71 (0.01) | 1.23 (0.23) | 0.25 | 3024.97 (0501.31) |
| mtry + min.node.size + sample.fraction + replace | 16 | 0.50 | $1N$ | 0.71 (0.01) | 1.16 (0.22) | 0.20 | 5798.01 (0657.91) |
| mtry + min.node.size + sample.fraction + splitrule | 16 | 0.50 | $1N$ | 0.71 (0.01) | 1.27 (0.25) | 0.27 | 8201.16 (1178.16) |
| mtry + min.node.size + sample.fraction + replace + splitrule | 16 | 0.50 | $1N$ | 0.71 (0.01) | 1.22 (0.24) | 0.25 | 18350.58 (2833.03) |

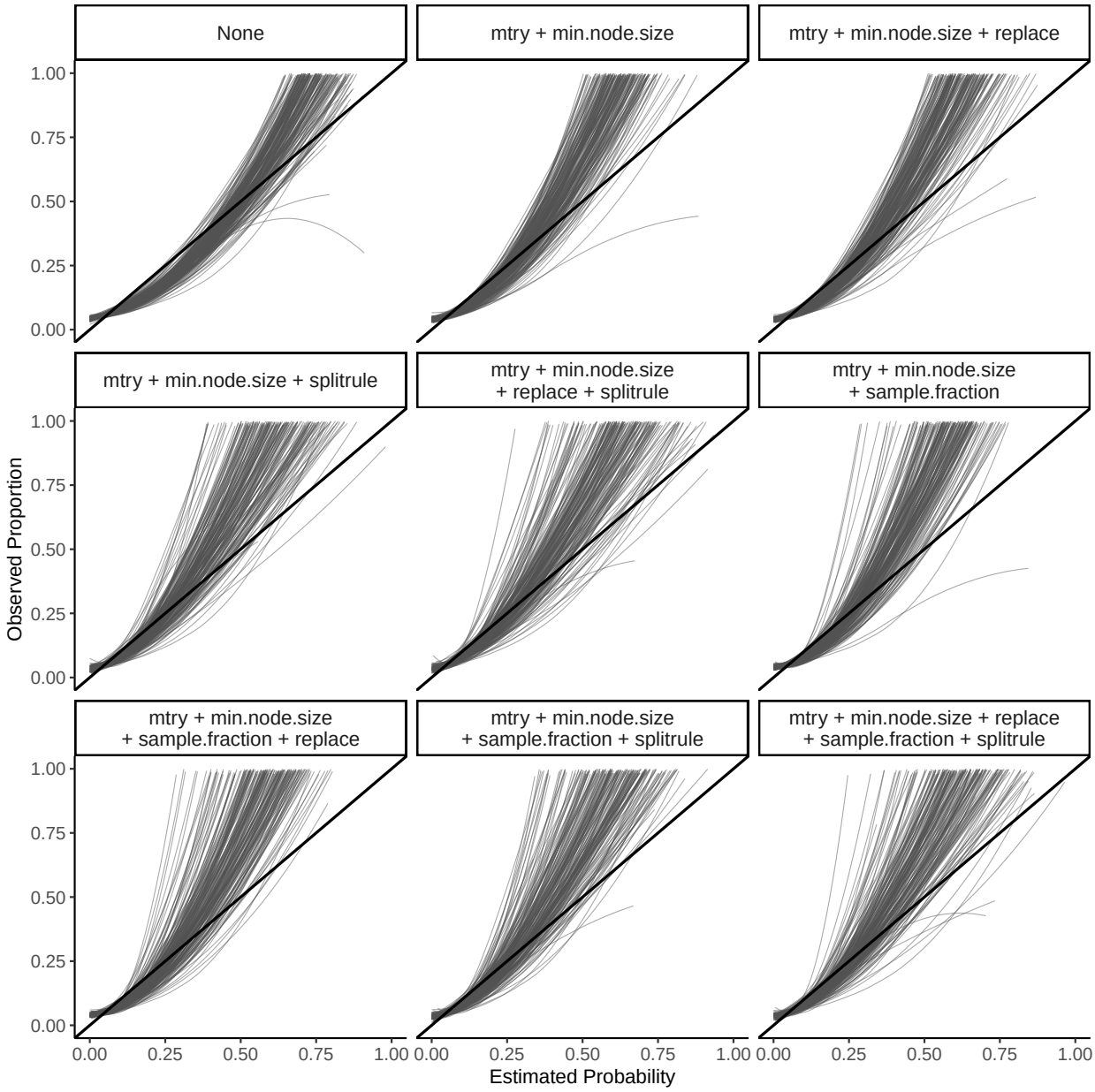


Figure S1: Calibration plots comparing hyperparameter combinations for every tenth dataset in scenarios where $EF = 0.1$.

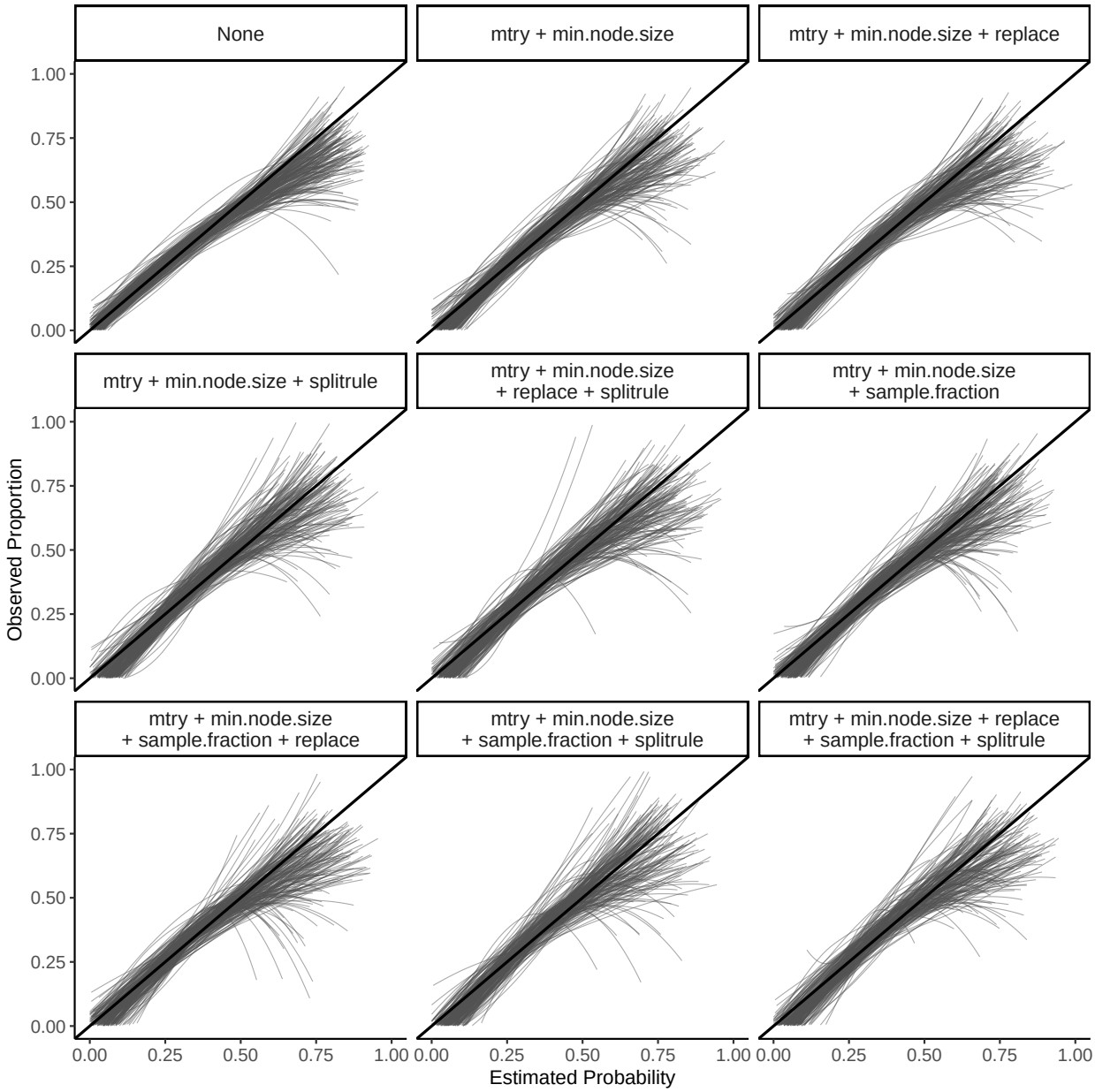


Figure S2: Calibration plots comparing hyperparameter combinations for every tenth dataset in scenarios where $EF = 0.3$.

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 | 61.66 (010.39) |
| Brier score | 8 | 0.10 | $0.5N$ | 0.71 (0.02) | 0.82 (0.19) | 0.30 | 61.59 (010.21) |
| Calibration slope | 8 | 0.10 | $0.5N$ | 0.71 (0.02) | 0.82 (0.19) | 0.32 | 62.25 (010.16) |
| AUC | 8 | 0.10 | $0.5N$ | 0.70 (0.02) | 0.74 (0.30) | 0.49 | 61.88 (010.35) |
| Classification accuracy | 8 | 0.10 | $0.5N$ | 0.70 (0.02) | 0.68 (0.26) | 0.49 | 61.80 (010.38) |
| Cohen's Kappa | 8 | 0.10 | $0.5N$ | 0.69 (0.02) | 0.49 (0.18) | 0.76 | 61.87 (010.25) |
| Calibration intercept | 8 | 0.10 | $1N$ | 0.72 (0.01) | 0.90 (0.11) | 0.16 | 133.02 (019.18) |
| Logarithmic loss | 8 | 0.10 | $1N$ | 0.72 (0.01) | 0.91 (0.13) | 0.16 | 132.20 (018.92) |
| Calibration slope | 8 | 0.10 | $1N$ | 0.72 (0.01) | 0.90 (0.14) | 0.17 | 133.04 (019.33) |
| Brier score | 8 | 0.10 | $1N$ | 0.72 (0.01) | 0.88 (0.18) | 0.21 | 132.19 (018.93) |
| AUC | 8 | 0.10 | $1N$ | 0.72 (0.01) | 0.86 (0.24) | 0.32 | 132.58 (019.01) |
| Classification accuracy | 8 | 0.10 | $1N$ | 0.71 (0.01) | 0.74 (0.21) | 0.38 | 132.44 (019.10) |
| Cohen's Kappa | 8 | 0.10 | $1N$ | 0.71 (0.01) | 0.58 (0.14) | 0.59 | 132.55 (019.09) |
| Calibration intercept | 16 | 0.10 | $0.5N$ | 0.72 (0.01) | 1.08 (0.13) | 0.13 | 438.81 (069.00) |
| Logarithmic loss | 16 | 0.10 | $0.5N$ | 0.72 (0.01) | 1.08 (0.17) | 0.15 | 436.92 (069.45) |
| Brier score | 16 | 0.10 | $0.5N$ | 0.71 (0.01) | 1.05 (0.24) | 0.17 | 437.44 (069.91) |
| Calibration slope | 16 | 0.10 | $0.5N$ | 0.71 (0.01) | 1.03 (0.26) | 0.18 | 438.02 (069.78) |
| AUC | 16 | 0.10 | $0.5N$ | 0.71 (0.01) | 1.05 (0.19) | 0.22 | 437.19 (069.85) |
| Classification accuracy | 16 | 0.10 | $0.5N$ | 0.71 (0.01) | 0.75 (0.23) | 0.38 | 437.87 (070.17) |
| Cohen's Kappa | 16 | 0.10 | $0.5N$ | 0.70 (0.01) | 0.55 (0.15) | 0.62 | 437.55 (069.15) |
| Brier score | 16 | 0.10 | $1N$ | 0.72 (0.01) | 1.08 (0.21) | 0.14 | 1120.78 (163.62) |
| Calibration intercept | 16 | 0.10 | $1N$ | 0.72 (0.01) | 1.13 (0.12) | 0.14 | 1123.30 (161.76) |
| Calibration slope | 16 | 0.10 | $1N$ | 0.72 (0.01) | 1.04 (0.22) | 0.14 | 1125.29 (163.32) |
| Logarithmic loss | 16 | 0.10 | $1N$ | 0.72 (0.01) | 1.13 (0.13) | 0.15 | 1122.20 (166.33) |
| AUC | 16 | 0.10 | $1N$ | 0.72 (0.01) | 1.11 (0.16) | 0.18 | 1124.40 (161.56) |
| Classification accuracy | 16 | 0.10 | $1N$ | 0.72 (0.01) | 0.80 (0.18) | 0.28 | 1122.93 (161.93) |
| Cohen's Kappa | 16 | 0.10 | $1N$ | 0.71 (0.01) | 0.65 (0.11) | 0.45 | 1123.55 (162.53) |

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 | 39.14 (06.86) |
| Logarithmic loss | 8 | 0.30 | $0.5N$ | 0.68 (0.02) | 0.95 (0.24) | 0.22 | 38.38 (06.67) |
| Brier score | 8 | 0.30 | $0.5N$ | 0.68 (0.02) | 0.94 (0.25) | 0.24 | 38.29 (06.69) |
| Calibration slope | 8 | 0.30 | $0.5N$ | 0.68 (0.02) | 0.92 (0.27) | 0.25 | 39.13 (06.91) |
| AUC | 8 | 0.30 | $0.5N$ | 0.68 (0.02) | 0.87 (0.29) | 0.32 | 38.63 (06.79) |
| Classification accuracy | 8 | 0.30 | $0.5N$ | 0.68 (0.02) | 0.84 (0.31) | 0.33 | 38.52 (06.65) |
| Cohen's Kappa | 8 | 0.30 | $0.5N$ | 0.67 (0.02) | 0.68 (0.22) | 0.46 | 38.52 (06.62) |
| Calibration intercept | 8 | 0.30 | $1N$ | 0.70 (0.01) | 1.12 (0.19) | 0.17 | 70.13 (10.88) |
| Logarithmic loss | 8 | 0.30 | $1N$ | 0.70 (0.01) | 1.06 (0.24) | 0.17 | 69.22 (10.62) |
| AUC | 8 | 0.30 | $1N$ | 0.70 (0.01) | 1.01 (0.25) | 0.18 | 69.56 (10.60) |
| Brier score | 8 | 0.30 | $1N$ | 0.70 (0.01) | 1.04 (0.26) | 0.19 | 69.25 (10.72) |
| Calibration slope | 8 | 0.30 | $1N$ | 0.70 (0.01) | 1.03 (0.28) | 0.19 | 70.02 (10.84) |
| Classification accuracy | 8 | 0.30 | $1N$ | 0.70 (0.01) | 0.92 (0.28) | 0.23 | 69.50 (10.91) |
| Cohen's Kappa | 8 | 0.30 | $1N$ | 0.70 (0.01) | 0.79 (0.18) | 0.30 | 69.66 (11.28) |
| Calibration slope | 16 | 0.30 | $0.5N$ | 0.68 (0.01) | 1.06 (0.34) | 0.22 | 190.02 (26.57) |
| Classification accuracy | 16 | 0.30 | $0.5N$ | 0.68 (0.01) | 0.99 (0.33) | 0.23 | 188.90 (25.97) |
| Brier score | 16 | 0.30 | $0.5N$ | 0.69 (0.01) | 1.16 (0.31) | 0.23 | 188.27 (25.60) |
| Logarithmic loss | 16 | 0.30 | $0.5N$ | 0.69 (0.01) | 1.16 (0.34) | 0.23 | 188.45 (25.85) |
| AUC | 16 | 0.30 | $0.5N$ | 0.69 (0.01) | 1.18 (0.29) | 0.24 | 188.85 (25.93) |
| Calibration intercept | 16 | 0.30 | $0.5N$ | 0.69 (0.01) | 1.27 (0.21) | 0.27 | 190.27 (26.14) |
| Cohen's Kappa | 16 | 0.30 | $0.5N$ | 0.67 (0.01) | 0.81 (0.17) | 0.28 | 188.71 (25.83) |
| Calibration slope | 16 | 0.30 | $1N$ | 0.70 (0.01) | 1.04 (0.23) | 0.17 | 433.23 (56.57) |
| Cohen's Kappa | 16 | 0.30 | $1N$ | 0.69 (0.01) | 0.90 (0.17) | 0.18 | 431.02 (55.22) |
| Classification accuracy | 16 | 0.30 | $1N$ | 0.70 (0.01) | 1.02 (0.29) | 0.20 | 432.34 (60.67) |
| Brier score | 16 | 0.30 | $1N$ | 0.70 (0.01) | 1.17 (0.27) | 0.22 | 429.33 (54.24) |
| Logarithmic loss | 16 | 0.30 | $1N$ | 0.70 (0.01) | 1.15 (0.25) | 0.22 | 431.65 (57.99) |
| AUC | 16 | 0.30 | $1N$ | 0.70 (0.01) | 1.26 (0.23) | 0.26 | 431.01 (55.35) |
| Calibration intercept | 16 | 0.30 | $1N$ | 0.70 (0.01) | 1.35 (0.18) | 0.31 | 432.13 (55.46) |

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<br>Mean (SD) | Calibration slope<br>Median (IQR) | RMSD(slope) | Runtime (seconds)<br>Mean (SD) |
|---|---|---|---|---|---|---|---|
| Calibration intercept | 8 | 0.50 | $0.5N$ | 0.69 (0.01) | 1.10 (0.22) | 0.18 | 44.47 (07.12) |
| Brier score | 8 | 0.50 | $0.5N$ | 0.69 (0.01) | 1.01 (0.32) | 0.21 | 43.57 (06.97) |
| Logarithmic loss | 8 | 0.50 | $0.5N$ | 0.69 (0.01) | 1.01 (0.30) | 0.21 | 43.63 (07.07) |
| AUC | 8 | 0.50 | $0.5N$ | 0.69 (0.01) | 0.97 (0.27) | 0.22 | 43.93 (07.03) |
| Calibration slope | 8 | 0.50 | $0.5N$ | 0.69 (0.01) | 0.98 (0.29) | 0.23 | 44.49 (07.15) |
| Classification accuracy | 8 | 0.50 | $0.5N$ | 0.69 (0.01) | 0.90 (0.27) | 0.24 | 43.74 (06.93) |
| Cohen's Kappa | 8 | 0.50 | $0.5N$ | 0.69 (0.01) | 0.91 (0.30) | 0.25 | 43.73 (07.05) |
| AUC | 8 | 0.50 | $1N$ | 0.71 (0.01) | 1.05 (0.24) | 0.16 | 82.53 (15.78) |
| Logarithmic loss | 8 | 0.50 | $1N$ | 0.71 (0.01) | 1.03 (0.24) | 0.16 | 81.71 (10.69) |
| Classification accuracy | 8 | 0.50 | $1N$ | 0.71 (0.01) | 1.02 (0.25) | 0.17 | 81.68 (10.54) |
| Brier score | 8 | 0.50 | $1N$ | 0.71 (0.01) | 1.03 (0.23) | 0.17 | 81.84 (10.68) |
| Calibration slope | 8 | 0.50 | $1N$ | 0.71 (0.01) | 1.03 (0.27) | 0.18 | 82.55 (10.86) |
| Cohen's Kappa | 8 | 0.50 | $1N$ | 0.71 (0.01) | 0.97 (0.25) | 0.18 | 82.02 (10.59) |
| Calibration intercept | 8 | 0.50 | $1N$ | 0.71 (0.01) | 1.19 (0.17) | 0.21 | 82.52 (10.74) |
| Calibration slope | 16 | 0.50 | $0.5N$ | 0.69 (0.01) | 1.03 (0.32) | 0.22 | 165.70 (22.18) |
| Cohen's Kappa | 16 | 0.50 | $0.5N$ | 0.69 (0.01) | 1.08 (0.34) | 0.22 | 164.47 (21.47) |
| Classification accuracy | 16 | 0.50 | $0.5N$ | 0.69 (0.01) | 1.10 (0.37) | 0.23 | 164.19 (21.39) |
| AUC | 16 | 0.50 | $0.5N$ | 0.69 (0.01) | 1.19 (0.32) | 0.24 | 164.75 (22.15) |
| Logarithmic loss | 16 | 0.50 | $0.5N$ | 0.69 (0.01) | 1.15 (0.31) | 0.24 | 164.25 (21.77) |
| Brier score | 16 | 0.50 | $0.5N$ | 0.69 (0.01) | 1.17 (0.33) | 0.25 | 164.17 (21.73) |
| Calibration intercept | 16 | 0.50 | $0.5N$ | 0.70 (0.01) | 1.32 (0.23) | 0.31 | 165.59 (22.46) |
| Calibration slope | 16 | 0.50 | $1N$ | 0.71 (0.01) | 1.05 (0.22) | 0.17 | 372.49 (46.57) |
| Classification accuracy | 16 | 0.50 | $1N$ | 0.71 (0.01) | 1.16 (0.29) | 0.23 | 370.64 (46.50) |
| Brier score | 16 | 0.50 | $1N$ | 0.71 (0.01) | 1.19 (0.26) | 0.23 | 369.97 (46.03) |
| Cohen's Kappa | 16 | 0.50 | $1N$ | 0.71 (0.01) | 1.17 (0.30) | 0.23 | 371.31 (48.20) |
| Logarithmic loss | 16 | 0.50 | $1N$ | 0.71 (0.01) | 1.19 (0.25) | 0.23 | 370.51 (46.13) |
| AUC | 16 | 0.50 | $1N$ | 0.71 (0.01) | 1.25 (0.28) | 0.26 | 371.55 (47.34) |
| Calibration intercept | 16 | 0.50 | $1N$ | 0.71 (0.01) | 1.40 (0.23) | 0.34 | 372.42 (46.31) |

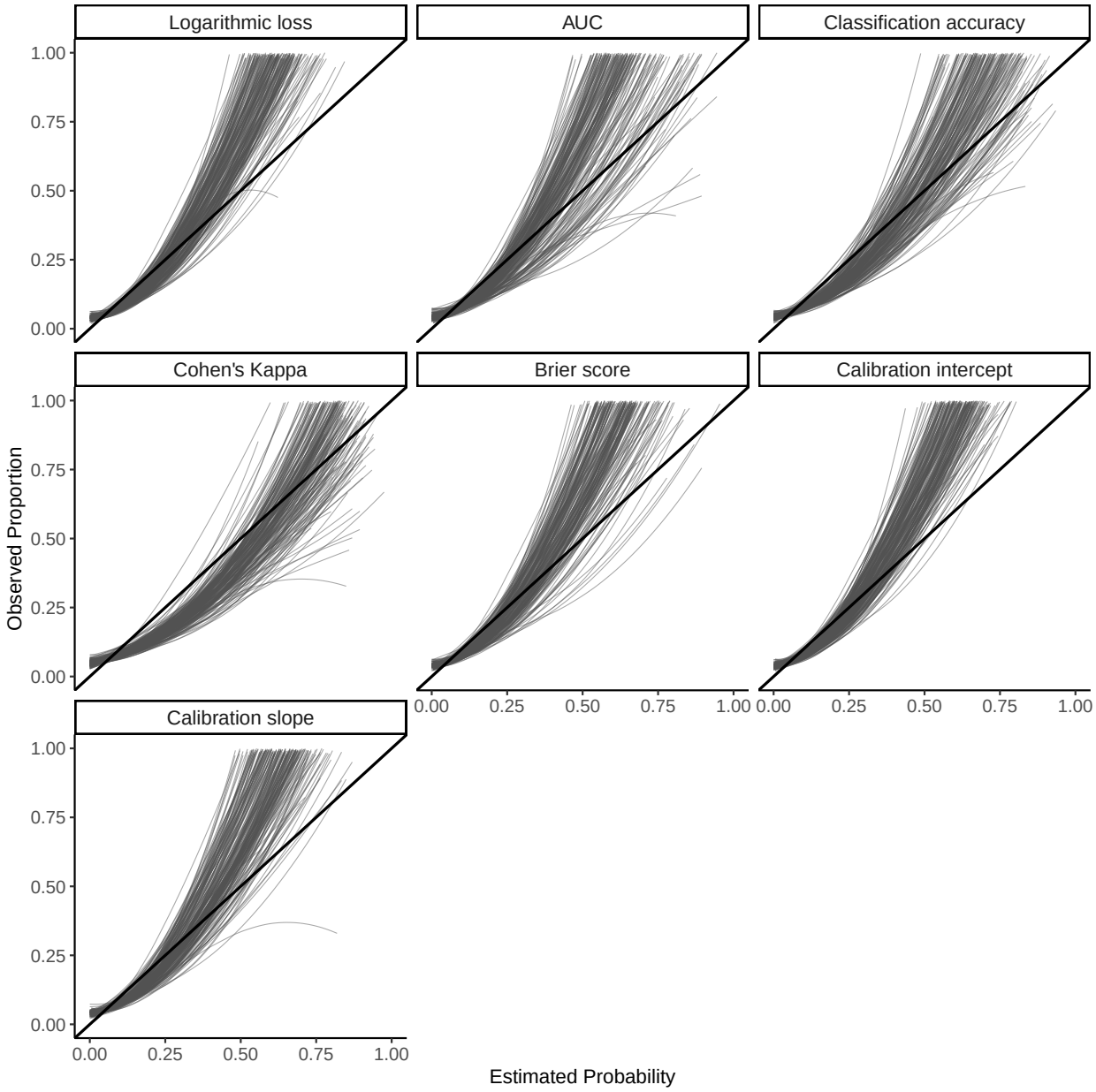


Figure S3: Calibration plots comparing optimisation criteria for every tenth dataset in scenarios where $EF = 0.1$.

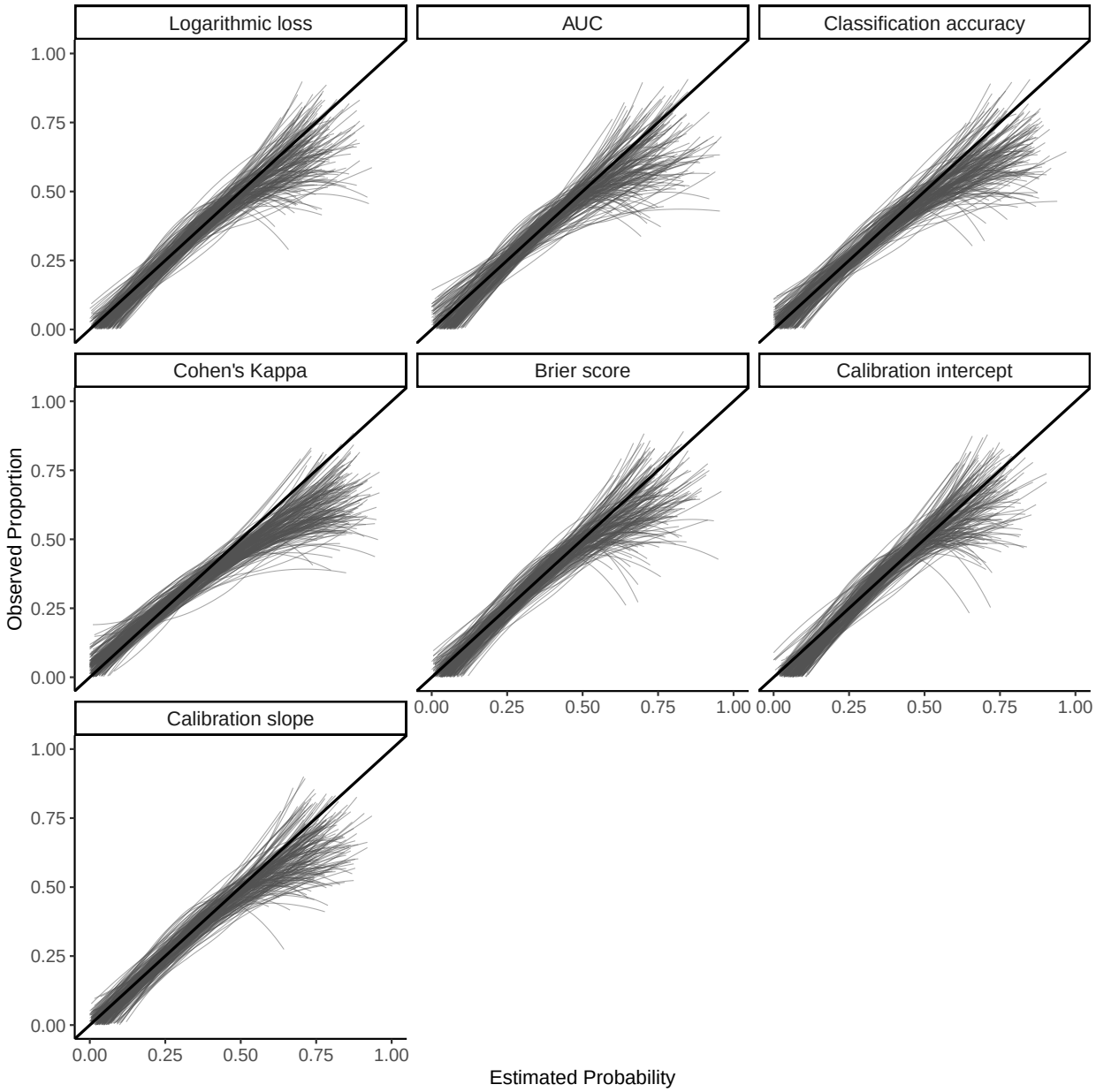


Figure S4: Calibration plots comparing optimisation criteria for every tenth dataset in scenarios where $EF = 0.3$.

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) |

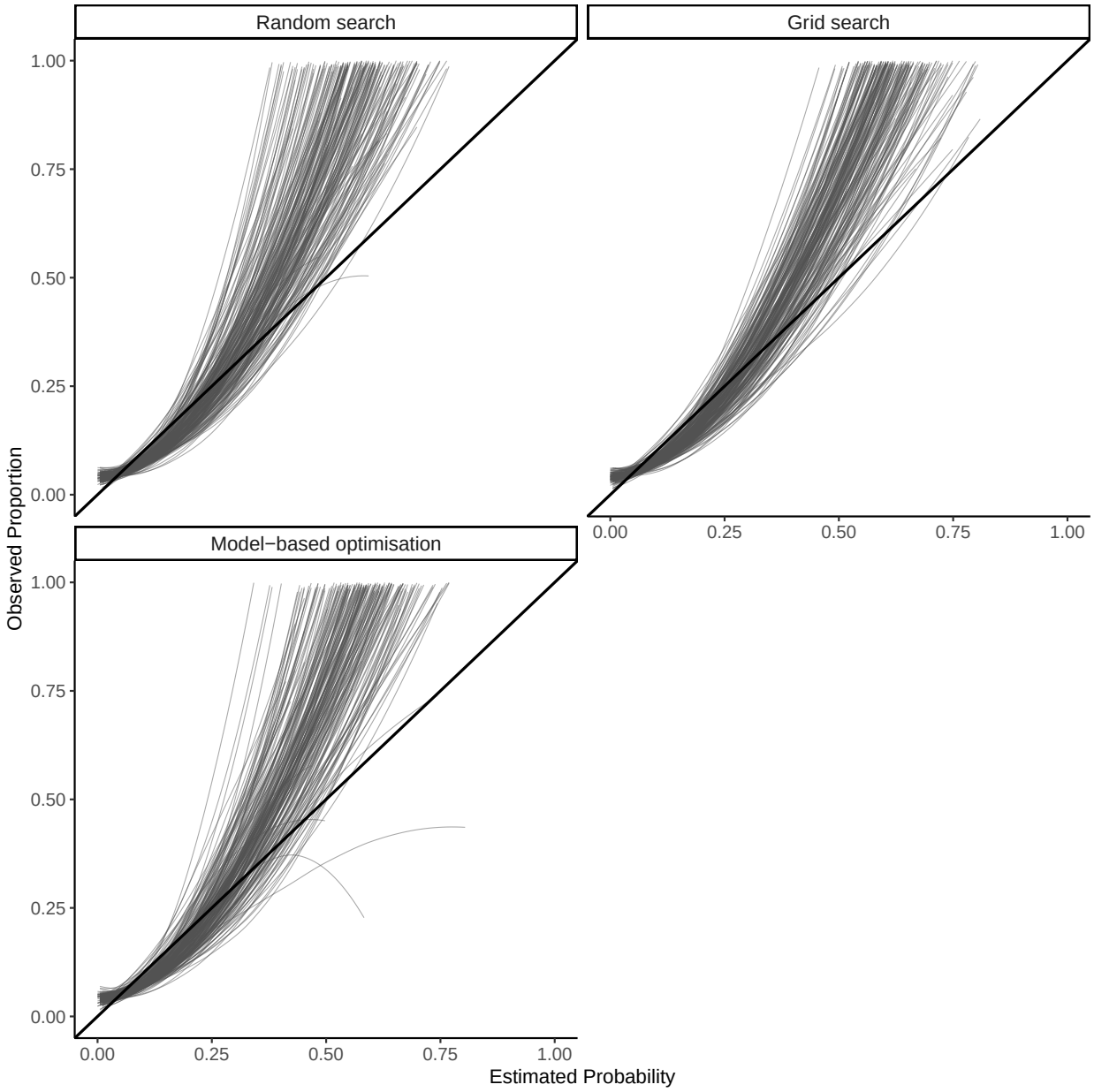


Figure S5: Calibration plots comparing hyperparameter search algorithms for every tenth dataset in scenarios where $EF = 0.1$.

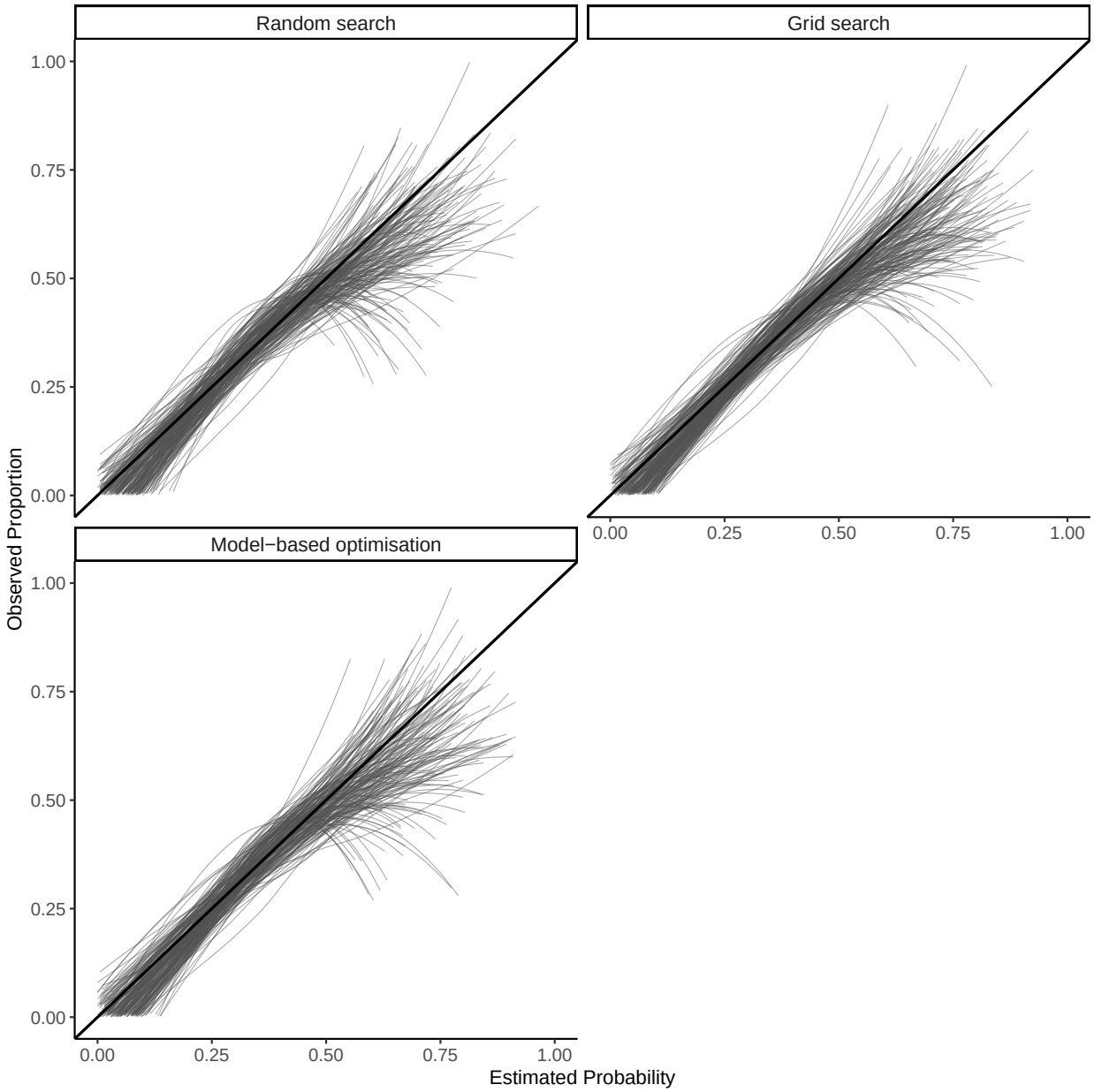


Figure S6: Calibration plots comparing hyperparameter search algorithms for every tenth dataset in scenarios where $EF = 0.3$.