SUPPLEMENTARY MATERIAL

1 eMethods

1.1 Cross-cohort harmonisation

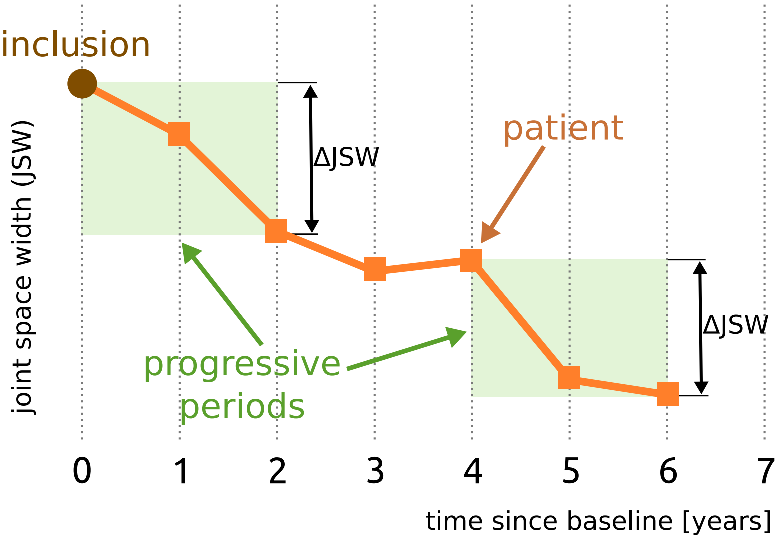
Based on the study protocols and data collection manuals, we identified semantic similarities (common data concepts) between CHECK and each of the target cohorts. The number of mapped attributes for each cohort is shown in eTable 1. Then we processed the syntax by either (1) transforming the attribute values in CHECK to a common lowest denominator set, or (2) transforming the other cohort attribute values to match the CHECK syntax (through mapping, re-categorization, aggregation, and scale alignment), always choosing the approach preserving more information. Finally, we performed a quality control exercise to detect mistakes (mainly cross tabulation and range/distribution comparison).

	patients	**attributes**	**mapped**
MUST	630	886	77–84
HOSTAS	538	130	45–50
DIGICOD	377	425	52–61
PROCOAC	983	288	40–57

eTable 1: Number of attributes mapped between each cohort and CHECK. A range is reported, as the exact number depends on the most recent visit timing and a number of attributes that were dropped due to the missing values (50% threshold was used).

1.2 Patient categories

To categorize patients we used one non-progressive category (**N**) and three progressive categories related to pain (**P**), structure (**S**), and combined pain and structure (**P+S**). The progression was analyzed multiple times per patient, in series of time windows, where a period between two visits was ≥ 2 years (see eFigure 1 for an example), which is representative of typical minimum timeline of a clinical trial. As a consequence, a patient may have progressive and non-progressive periods along the disease trajectory through time.



eFigure 1: Illustration of how the periods were used in the analysis. Example patient was included at year 0 of the cohort timeline and had measurements of joint space width (JSW) taken at every year until a knee replacement after year 6. Two periods marked with green boxes: $(0, 2)$ and $(4, 6)$, were assigned to the structural progression category, as the average change in JSW during these periods was above the pre-defined threshold.

When the progression could not be assessed due to missing values, the corresponding period was excluded from the dataset. When measurements were present for two knees, the most affected knee (with greater ΔJSW) was used to decide the progression status. Additionally, all periods following a knee replacement were removed.

1.3 Machine learning

1.3.1 Data preprocessing

Before the start of the harmonisation process we cleaned the original datasets as much as possible. We dropped duplicated columns, calculated missing visit dates or patient age using other data (e.g. visit date and age at baseline / birthday), split the values of multi-choice attributes into separate columns, removed comments, and converted text values to numerical categories.

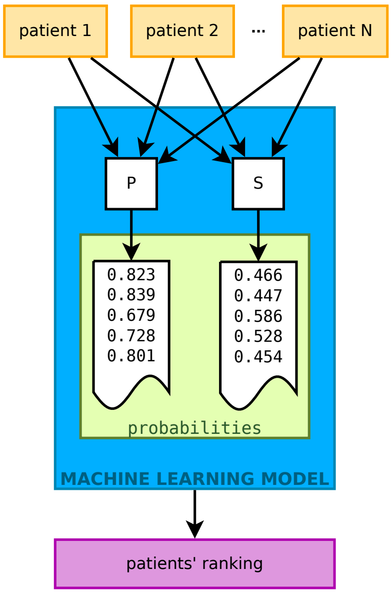
All attributes with more than 50% missing values and all periods with over 40% missing attributes were removed from each dataset. In addition, we removed attributes that did not vary at the start of each period (as they cannot be used to distinguishing between the categories), and attributes that could be exploited by the model, such as dates, visit numbers, and IDs. In attempt to fix some of the missing values, we filled forward attributes that do not change across visits (e.g. past diseases).

We assumed that all attributes with at most 10 different values are categorical, and we manually identified ordinal (e.g. *low, medium, high*) and continuous (e.g. *1.2, 2.4, 3.7*) attributes in the CHECK cohort variable guide.

Additional preprocessing was performed during the model training. The missing values were imputed, based on the training set only (to avoid information leaks from the test set). The imputation was carried out with the mode/mean value (for categorical/continuous attributes).

The final step after imputation was the one-hot encoding of nominal attributes (e.g. *left, right*). This encoding creates additional dummy attributes, one per category, and sets the one corresponding to the attribute value (the "hot one") to 1, and others to 0. All categorical attributes with more than 2 distinct values were encoded, unless they were known to be ordinal.

1.3.2 The *duo classifier*



eFigure 2: The *duo classifier* (Widera et al. 2020) uses a random forest algorithm to train two sub-models to independently predict probabilities of the pain (P) and structure (S) related progression. We implemented this classifier as a wrapper class on top of the sub-models that predicts one of the four class labels (N, P, S, P+S), but still provides independent P and S probabilities. These probabilities are later combined into a score that is used to rank the patients. Note that the probabilities do not represent the real-world likelihood of progression, but rather express the confidence of the model in its predictions. In other words, they need to be considered in context of the specific dataset being used to train the models.

1.3.3 Cross-validation

In all experiments, out-of-sample estimation of the algorithm performance was used. That is, some of the samples were kept hidden from the algorithm during training, and used later to test it. Specifically, we followed the standard 10-fold stratified cross-validation (CV) protocol, in which the instances are split into 10 approximately equal-sized parts (folds) and the split preserves the overall class distribution within each fold. Each fold is then used in turn as a test set, and the remaining 9 folds are used as a training set. To score the method performance, rather than averaging the scores across all 10 folds, we pooled the out-of-sample predictions together and calculated a single score.

The cross-validation was repeated 10 times with different partitions into folds. As random forest algorithm is not deterministic, we also repeated the model training (25 times) with different random seeds (the seeds remained constant across folds and cross-validation repeats).

To tune the configuration of the *duo classifier*, we performed an exhaustive search through 84 combinations of three random forest parameters: **number of trees** $\in [100, 200, 400, 600, 800, 1000]$, **maximum tree depth** $\in [4, 5, \ldots, 10]$, and **split quality criterion** $\in [gini, entropy]$ (standing for Gini impurity and information gain). On screening data, we limited the search to 60 combinations by using only Gini impurity as the split criterion, **number of trees** $\in [100, 200, 400, 600, 800, 1000]$, and **maximum tree depth** $\in [3, 4, \ldots, 12]$.

Because multiple algorithm parameters were tested, cross-validated performance of the best configuration is an optimistically biased estimate of the performance of the final model trained on all data. This “multiple induction” problem is conceptually equivalent to multiple hypothesis testing in statistics. To mitigate this, the focus was placed on the performance of the median model (rather than the best one) for each cross-validation repeat, and the median performance across all repeats.

1.3.4 F_1 score and class imbalance

The initial performance of the models was measured using the F_1 score (harmonic mean of positive predictive value (precision) and sensitivity (recall)). Since we needed to deal with four imbalanced categories, we used a macro-weighted multi-class variant of the F_1 score, in which per class F_1 scores are averaged and weighted by support (number of periods in each category).

1.3.5 Recruitment score

Despite its advantages, the F_1 score treats all mistakes as equal and does not represent the clinical preference for selection of patients in the **P+S** category. Therefore, we designed a specialised **recruitment score** directly based on the results of simulated selections. All simulations were based on the ranking derived from the model predictions through the use of a ranking function (see eSection 1.3.6). We tested 6 different selection sizes k representing an increasing number of patients, roughly corresponding to the number of patients expected to be enrolled from our cohorts. For each top k patients we calculated a sub-score (eEquation 1) as a weighted sum of the relative selection (r_i) in each category, normalised by the maximum weight. The weights c_i where used to model the clinical preference between the categories **N** < **P** < **S** < **P+S**. We used two kinds of c_i weights (see eTable 2): split weights (where non-progressive category has a negative weight) and progressive weights (where **S** and **P+S** categories are weighted much heavier). The total recruitment score was calculated as a weighted average of sub-scores for all selection sizes (eEquation 2). The weights s_k were centered around the $k \in [60, 90]$ range (see eTable 2).

	weights (c_i))	
category	**split**	**progressive**
N	-0.5	0
P	1.0	1
S	1.5	5
P+S	2.0	7

size (k)	weight (s_k)
10	0.50
30	0.75
60	1.00
90	1.00
120	0.75
150	0.50

$$sub\_score = \frac{1}{max_i(c_i)} \sum_i c_i * r_i \tag{1}$$

$$total\_score = \frac{\sum_k s_k * sub\_score(k)}{\sum_k s_k} \tag{2}$$

eTable 2: Weights used by the recruitment score.

1.3.6 Time factor in training of the selection models

The patient categories were defined for three periods between the CHECK visits at which the radiographic readings were taken: $(0, 2)$, $(2, 5)$, and $(5, 8)$. The difference in length (2 vs. 3 years) is negligable, as the categories have been defined using the average change per year, rather than a difference over the entire period.

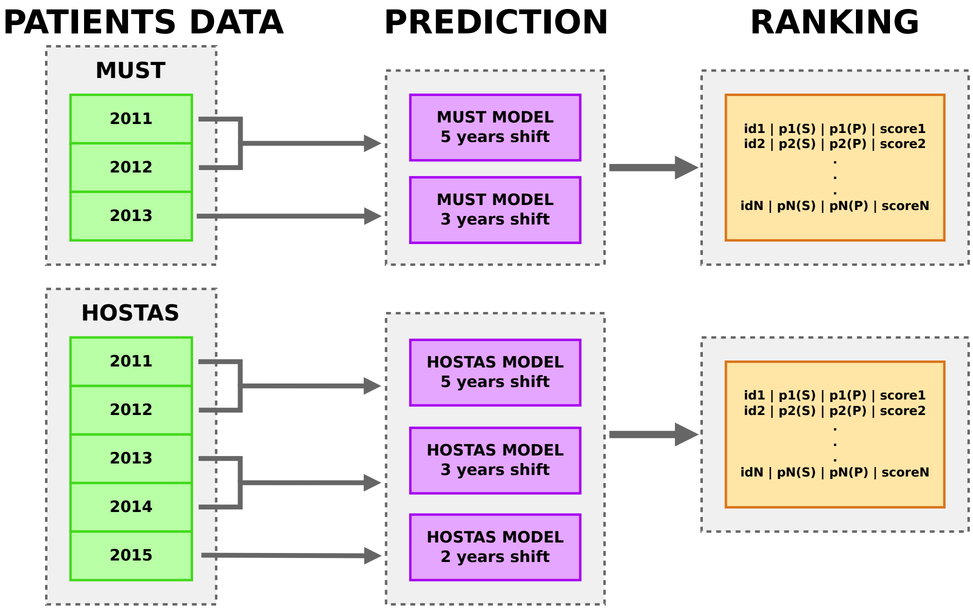
Additional complication arose from the fact, that for the majority of patients, even the most recent visits happened several years prior (see eTable 3). It required adjustment in training, to make sure that the patient category is not defined for a period immediately after the visit, but for a period shifted forward in time that matches the running time of the study. For example, when the most recent visit was two years prior, we used visit 0 data to predict the category of the $(2, 5)$ period (that is shifted 2 years into the future). Given the available periods, three different shifts were possible: 5 years ($0 \Rightarrow (5, 8)$), 3 years ($2 \Rightarrow (5, 8)$), and 2 years ($0 \Rightarrow (2, 5)$). For a few patients, no shift was needed as the data were recent enough, to directly use the original period categories in training.

year	’09	’10	’11	’12	’13	’14	’15	’16	’17
MUST	—	91	260	254	25	—	—	—	—
HOSTAS	7	5	37	21	85	79	132	135	37
DIGICOD	—	—	—	—	4	79	144	126	24
PROCOAC	36	44	47	41	53	132	298	220	—
shift	—	—	5y	5y	3y	3y	2y	2y	0y

eTable 3: The number of patients with the most recent visit in particular year and a corresponding shift of the target period used in training. Notice that the shift is not always exact, but only approximated due to the boundaries (begin/end years) of categorised CHECK periods.

1.3.7 Ranking function

To be able to rank the patients from each cohort, the **P** and **S** probabilities returned by the model were combined (see eFigure 3) into a single score using a **ranking function**.


PATIENTS DATA
PREDICTION
RANKING
MUST
2011
2012
2013
MUST MODEL 5 years shift
MUST MODEL 3 years shift
id1 | p1(S) | p1(P) | score1
id2 | p2(S) | p2(P) | score2
idN | pN(S) | pN(P) | scoreN
HOSTAS
2011
2012
2013
2014
2015
HOSTAS MODEL 5 years shift
HOSTAS MODEL 3 years shift
HOSTAS MODEL 2 years shift
id1 | p1(S) | p1(P) | score1
id2 | p2(S) | p2(P) | score2
idN | pN(S) | pN(P) | scoreN

eFigure 3: Illustration of the link between the cohort data and multi-model predictions used to construct the per cohort selection rankings. Only MUST and HOSTAS are shown, but similar procedure was followed for DIGICOD and PROCOAC.

The ranking function was designed to bias the selection towards the minority category (the **P+S** patients). In our experiments, we used three different ranking functions: direct sum (eEquation 3), scaled sum (eEquation 4), and sum of z-scores (eEquation 5). The idea behind the latter two, was to correct for differences in range between the two probability distributions (**P** range was found to be broader than **S**) and make them contribute equally to the ranking score.

$$score = P + S \quad (3) \qquad score = \frac{P}{max(P)} + \frac{S}{max(S)} \quad (4) \qquad score = \frac{P - \mu_p}{\sigma_p} + \frac{S - \mu_s}{\sigma_s} \quad (5)$$

Because different ranking functions could be used for each time shift, the scores were rescaled to the same range (using min-max normalisation) before being merged into a single cohort selection ranking.

As expected, the model prediction quality decreased with the increasing time shift. To represent the lower trust in the ranking score when a large shift is applied, we added a progressive shift-dependent non-linear penalty (eEquation 6) to the ranking score. In effect, patients with the oldest data were less likely to have a high rank.

$$penalty\_multiplier(x) = 1 - (9x + x^2)/200 \quad (6)$$

1.4 Enrolment process

The enrolment decisions were made weekly (or more often when needed) closely following the availability of the radiograph assessment data. We made additional effort to synchronize the start of the screening process at all recruitment centers, to maximize the number of patients in the ranking at the moment of first decisions. This was important, as the confidence in an enrolment decision increases with more patients in the ranking. That is, when many patients are already ranked, we can trust that a highly ranked new patient is likely to remain in the top of the ranking throughout the entire recruitment process, so a decision to enroll this person is straightforward. When the rank is close to the borderline between already enrolled and rejected patients, or the total number of ranked patients is low, it is best to delay the decision and wait for more data to increase the confidence. However, some sub-optimal decisions are unavoidable, as they are always made on partial ranking: never knowing how the next group of patients will be ranked. As a result, in retrospect some patients could have been rejected too soon if many future patients were ranked lower, or enrolled too early if many future patients were ranked higher.

2 eResults

2.1 Enrolment

recruitment center	all	invited	ranked	enrolled	rej. ratio
Oslo (MUST)	630	53	40	31	22.5%
Leiden (HOSTAS)	538	71	68	50	26.5%
Paris (DIGICOD)	377	29	29	20	31.0%
A Coruña (PROCOAC)	983	58	56	43	23.2%
Utrecht (CHECK)	1002	222	213	153	28.2%
TOTAL	3530	433	406	297	26.8%

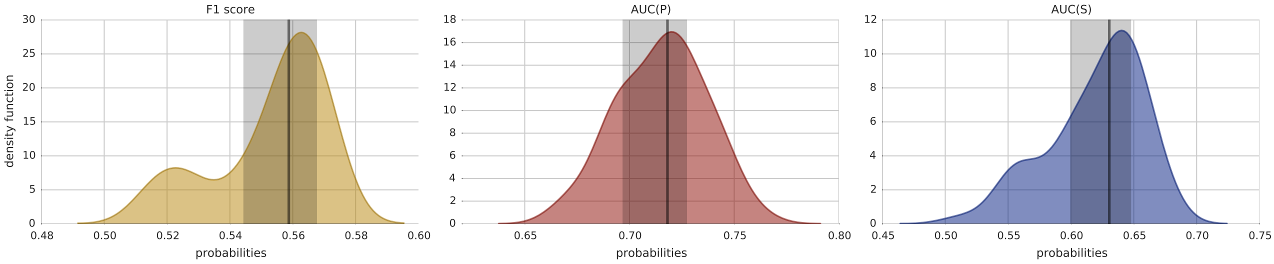
eTable 4: Recruitment process in numbers. The rejection ratio is the percentage of ranked patients not enrolled in the study.

2.2 Best model configurations with respect to recruitment score

F_1 score	configuration
0.5715	800-gini-2-d8
0.5715	400-gini-2-d8
0.571	600-gini-2-d8
...	
0.517	600-gini-2-d3
0.5165	200-gini-2-d3
0.5155	100-gini-2-d3

AUC(P)	configuration
0.76176	400-gini-2-d7
0.75076	600-gini-2-d7
0.74835	800-gini-2-d7
...	
0.67781	200-gini-2-d3
0.67145	200-gini-2-d4
0.66696	200-gini-2-d5

AUC(S)	configuration
0.67508	100-gini-2-d12
0.67263	600-gini-2-d10
0.67082	400-gini-2-d9
...	
0.54854	400-gini-2-d3
0.54734	800-gini-2-d3
0.51323	200-gini-2-d3



eFigure 4: Distribution of quality measures across all **screening model** configurations. The grey area shows interquartile range and thick vertical line indicates the median value. Tables show 3 top/bottom scores and corresponding model parameters.

ranking function: **sum** 0.55880 +0.00802

	N [1153]		**P** [298]		**S** [330]		**P+S** [111]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	0	0%	3	30%	0	0%	7	70%
30	3 +1	10%	9 -2	30%	5 +2	17%	13 -1	43%
60	13 -1	22%	21 +1	35%	7	12%	19	32%
90	19 -4	21%	34 +2	38%	12	13%	25 +2	28%
120	32	27%	44 -2	37%	15	12%	29 +2	24%
150	46 +2	31%	56	37%	19	13%	29 -2	19%

configuration:
`1000-gini-2-d6`

F_1 score: 0.55950 (29)
AUC(P): 0.72204 (26)
AUC(S): 0.62969 (31)

ranking function: **scaled** 0.56035 +0.01234

	N [1153]		**P** [298]		**S** [330]		**P+S** [111]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	0	0%	3	30%	0	0%	7	70%
30	3 +1	10%	9 -2	30%	5 +1	17%	13	43%
60	13 -1	22%	21	35%	8 +1	13%	18	30%
90	19 -4	21%	34 +1	38%	12	13%	25 +3	28%
120	30	25%	46 -2	38%	16 +1	13%	28 +1	23%
150	46 +1	31%	53 -1	35%	21 +1	14%	30 -1	20%

configuration:
`1000-gini-2-d6`

F_1 score: 0.55950 (29)
AUC(P): 0.72204 (26)
AUC(S): 0.62969 (31)

ranking function: **zscore** 0.53659 +0.00982

	N [1153]		**P** [298]		**S** [330]		**P+S** [111]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	0 -1	0%	3	30%	1 +1	10%	6	60%
30	6 +2	20%	8	27%	4 -1	13%	12 -1	40%
60	15 +1	25%	18 -1	30%	9 -1	15%	18 +1	30%
90	23 -2	26%	26 -4	29%	15 +1	17%	26 +5	29%
120	34 +2	28%	33 -6	28%	22 +1	18%	31 +3	26%
150	43	29%	45 -6	30%	27	18%	35 +6	23%

configuration:
`200-gini-2-d5`

F_1 score: 0.54350 (47)
AUC(P): 0.66696 (60)
AUC(S): 0.61694 (36)

(a) progressive weights

ranking function: **sum** 0.47798 +0.00787

	N [1153]		**P** [298]		**S** [330]		**P+S** [111]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	0	0%	3	30%	0	0%	7	70%
30	3 +1	10%	9 -2	30%	5 +2	17%	13 -1	43%
60	13 -1	22%	21 +1	35%	7	12%	19	32%
90	19 -4	21%	34 +2	38%	12	13%	25 +2	28%
120	32	27%	44 -2	37%	15	12%	29 +2	24%
150	46 +2	31%	56	37%	19	13%	29 -2	19%

configuration:
`1000-gini-2-d6`

F_1 score: 0.55950 (29)
AUC(P): 0.72204 (26)
AUC(S): 0.62969 (31)

ranking function: **scaled** 0.48095 +0.01700

	N [1153]		**P** [298]		**S** [330]		**P+S** [111]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	1 +1	10%	2 -1	20%	1 +1	10%	6 -1	60%
30	3 +1	10%	9 -2	30%	4	13%	14 +1	47%
60	11 -3	18%	19 -2	32%	11 +4	18%	19 +1	32%
90	22 -1	24%	33	37%	13 +1	14%	22	24%
120	32 +2	27%	42 -6	35%	20 +5	17%	26 -1	22%
150	45	30%	53 -1	35%	21 +1	14%	31	21%

configuration:
`100-gini-2-d6`

F_1 score: 0.55700 (36)
AUC(P): 0.69516 (47)
AUC(S): 0.64988 (13)

ranking function: **zscore** 0.48448 +0.01942

	N [1153]		**P** [298]		**S** [330]		**P+S** [111]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	0 -1	0%	3	30%	1 +1	10%	6	60%
30	6 +2	20%	8	27%	4 -1	13%	12 -1	40%
60	15 +1	25%	18 -1	30%	9 -1	15%	18 +1	30%
90	23 -2	26%	26 -4	29%	15 +1	17%	26 +5	29%
120	34 +2	28%	33 -6	28%	22 +1	18%	31 +3	26%
150	43	29%	45 -6	30%	27	18%	35 +6	23%

configuration:
`200-gini-2-d5`

F_1 score: 0.54350 (47)
AUC(P): 0.66696 (60)
AUC(S): 0.61694 (36)

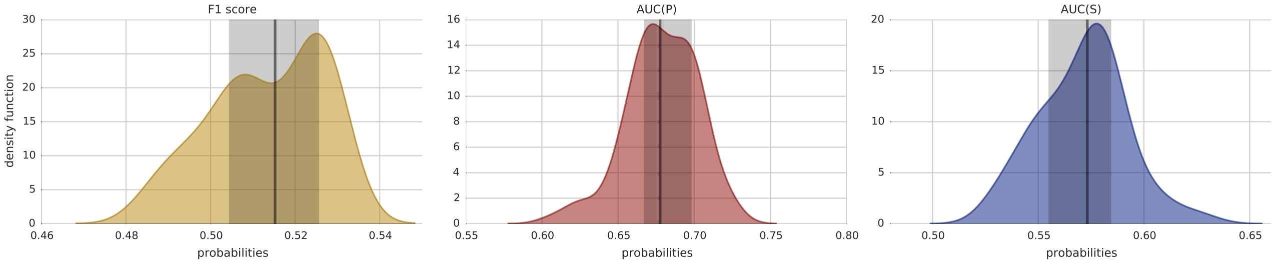
(b) split weights

eTable 5: Results of simulated recruitment for the **screening model** configurations with best recruitment score (given at the top of each table). The differences in recruitment compared to the model configuration with highest F_1 score are highlighted with green (positive) and red (negative) colour. For each of the ranking functions the used model configuration and its prediction quality measured with F_1 score and AUC is given for reference (rank of the measure among all configurations is given in brackets).

F_1 score	configuration
0.531	1000-gini-2-d6
0.5305	400-gini-2-d6
0.53	1000-entropy-2-d6
...	
0.487	600-gini-2-d10
0.486	1000-gini-2-d10
0.4855	800-gini-2-d10

AUC(P)	configuration
0.72447	100-gini-2-d10
0.72223	600-gini-2-d7
0.71798	400-entropy-2-d10
...	
0.62554	200-entropy-2-d7
0.62238	100-gini-2-d8
0.60731	100-entropy-2-d10

AUC(S)	configuration
0.62874	100-gini-2-d7
0.62108	100-entropy-2-d9
0.61126	200-entropy-2-d8
...	
0.53129	600-gini-2-d9
0.5312	100-entropy-2-d5
0.52607	400-gini-2-d8



eFigure 5: Distribution of quality measures across all **CHECK selection model (2 years shift)** configurations. The grey area shows interquartile range and thick vertical line indicates the median value. Tables show 3 top/bottom scores and corresponding model parameters.

ranking function: **sum** 0.45005 +0.01600

	N [589]		**P** [128]		**S** [205]		**P+S** [59]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	2	20%	5 -1	50%	1	10%	2 +1	20%
30	5	17%	14 +1	47%	2 -1	7%	9	30%
60	13 -1	22%	25	42%	7 -1	12%	15 +2	25%
90	24 -3	27%	35 +1	39%	15 +2	17%	16	18%
120	39 +1	32%	41 -1	34%	20 -1	17%	20 +1	17%
150	54	36%	46 -2	31%	30 +3	20%	20 -1	13%

configuration:
200-entropy-2-d8

F_1 score: 0.51750 (38)
AUC(P): 0.70572 (7)
AUC(S): 0.61126 (3)

ranking function: **scaled** 0.44934 +0.01890

	N [589]		**P** [128]		**S** [205]		**P+S** [59]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	0 -2	0%	7 +1	70%	1	10%	2 +1	20%
30	7 +2	23%	11 -2	37%	5 +2	17%	7 -2	23%
60	14 -1	23%	25 +1	42%	9 +1	15%	12 -1	20%
90	26 -1	29%	31 -3	34%	16 +3	18%	17 +1	19%
120	38 -1	32%	41 -1	34%	22 +2	18%	19	16%
150	54	36%	47 +1	31%	30 +2	20%	19 -3	13%

configuration:
100-gini-2-d5

F_1 score: 0.52350 (29)
AUC(P): 0.70094 (14)
AUC(S): 0.58517 (18)

ranking function: **zscore** 0.44890 +0.02752

	N [589]		**P** [128]		**S** [205]		**P+S** [59]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	2 -1	20%	4 -1	40%	2 +1	20%	2 +1	20%
30	4 -2	13%	17 +4	57%	5 +2	17%	4 -4	13%
60	13 -3	22%	23 +2	38%	10 +1	17%	14	23%
90	26	29%	30 -3	33%	16 +2	18%	18 +1	20%
120	40 +1	33%	37 -2	31%	24 +1	20%	19	16%
150	52	35%	46	31%	31 +1	21%	21 -1	14%

configuration:
100-entropy-2-d9

F_1 score: 0.50650 (59)
AUC(P): 0.69862 (20)
AUC(S): 0.62108 (2)

(a) progressive weights

ranking function: **sum** 0.36628 +0.01805

	N [589]		**P** [128]		**S** [205]		**P+S** [59]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	1 -1	10%	4 -2	40%	1	10%	4 +3	40%
30	9 +4	30%	11 -2	37%	3	10%	7 -2	23%
60	18 +4	30%	21 -4	35%	10 +2	17%	11 -2	18%
90	27	30%	32 -2	36%	15 +2	17%	16	18%
120	37 -1	31%	41 -1	34%	23 +2	19%	19	16%
150	54	36%	48	32%	26 -1	17%	22 +1	15%

configuration:
100-gini-2-d8

F_1 score: 0.51450 (43)
AUC(P): 0.62238 (83)
AUC(S): 0.53503 (80)

ranking function: **scaled** 0.37420 +0.02643

	N [589]		**P** [128]		**S** [205]		**P+S** [59]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	2	20%	4 -2	40%	1	10%	3 +2	30%
30	6 +1	20%	13	43%	2 -1	7%	9	30%
60	14 -1	23%	24	40%	7 -1	12%	15 +2	25%
90	26 -1	29%	32 -2	36%	15 +2	17%	17 +1	19%
120	39	32%	41 -1	34%	20	17%	20 +1	17%
150	53 -1	35%	46	31%	31 +3	21%	20 -2	13%

configuration:
200-entropy-2-d8

F_1 score: 0.51750 (38)
AUC(P): 0.70572 (7)
AUC(S): 0.61126 (3)

ranking function: **zscore** 0.38403 +0.03131

	N [589]		**P** [128]		**S** [205]		**P+S** [59]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	2 -1	20%	3 -2	30%	2 +1	20%	3 +2	30%
30	5 -1	17%	12 -1	40%	5 +2	17%	8	27%
60	17 +1	28%	21	35%	8 -1	13%	14	23%
90	29 +3	32%	26 -7	29%	19 +5	21%	16 -1	18%
120	45 +6	38%	35 -4	29%	21 -2	18%	19	16%
150	56 +4	37%	44 -2	29%	28 -2	19%	22	15%

configuration:
400-entropy-2-d8

F_1 score: 0.51750 (39)
AUC(P): 0.66829 (60)
AUC(S): 0.58640 (15)

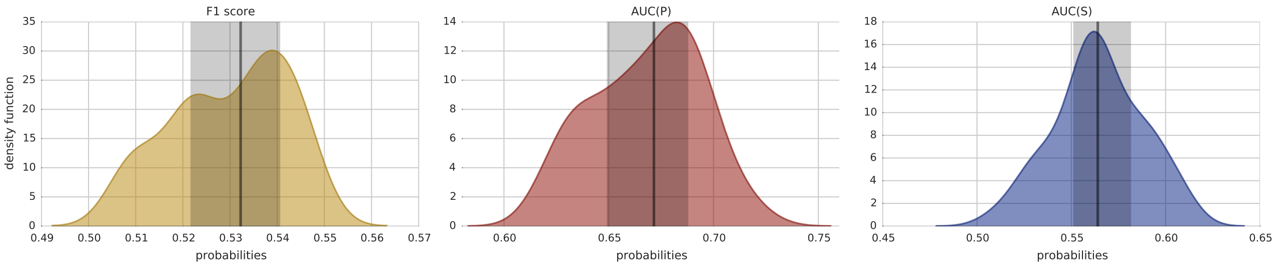
(b) split weights

eTable 6: Results of simulated recruitment for the **CHECK selection model (2 years shift)** configurations with best recruitment score (given at the top of each table). The differences in recruitment compared to the model configuration with highest F_1 score are highlighted with green (positive) and red (negative) colour. For each of the ranking functions the used model configuration and its prediction quality measured with F_1 score and AUC is given for reference (rank of the measure among all configurations is given in brackets).

F_1 score	configuration
0.5475	1000-gini-2-d7
0.547	800-gini-2-d7
0.5465	600-gini-2-d7
...	
0.509	1000-gini-2-d4
0.5085	800-entropy-2-d4
0.508	100-entropy-2-d4

AUC(P)	configuration
0.72235	200-gini-2-d5
0.7162	200-entropy-2-d9
0.71169	400-gini-2-d5
...	
0.62271	800-entropy-2-d7
0.61949	1000-entropy-2-d9
0.61639	200-gini-2-d8

AUC(S)	configuration
0.61265	400-gini-2-d8
0.61033	100-entropy-2-d8
0.6094	1000-entropy-2-d4
...	
0.51835	600-entropy-2-d10
0.51455	800-gini-2-d9
0.50736	200-gini-2-d6



eFigure 6: Distribution of quality measures across all **PROCOAC selection model (2 years shift)** configurations. The grey area shows interquartile range and thick vertical line indicates the median value. Tables show 3 top/bottom scores and corresponding model parameters.

ranking function: **sum** 0.45012 +0.05129

	N [600]		**P** [131]		**S** [209]		**P+S** [59]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	1 -3	10%	6 +2	60%	2	20%	1 +1	10%
30	5 -2	17%	12 +2	40%	6 -1	20%	7 +1	23%
60	13 -2	22%	28 +4	47%	9 -1	15%	10 -1	17%
90	24 -1	27%	37 +2	41%	14 -3	16%	15 +2	17%
120	36 -3	30%	45 +4	38%	22	18%	17 -1	14%
150	50 -5	33%	52 +4	35%	27 -1	18%	21 +2	14%

configuration:
`1000-entropy-2-d4`

F_1 score: 0.50950 (81)
AUC(P): 0.70460 (5)
AUC(S): 0.60940 (3)

ranking function: **scaled** 0.45904 +0.06262

	N [600]		**P** [131]		**S** [209]		**P+S** [59]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	1 -3	10%	5 +1	50%	3 +1	30%	1 +1	10%
30	6 -2	20%	11 +1	37%	5 -1	17%	8 +2	27%
60	13	22%	25 +1	42%	11 -2	18%	11 +1	18%
90	21 -5	23%	36 +3	40%	19 +2	21%	14	16%
120	37 -5	31%	42 +3	35%	24 +3	20%	17 -1	14%
150	53 -1	35%	48	32%	30 +1	20%	19	13%

configuration:
`100-entropy-2-d4`

F_1 score: 0.50800 (84)
AUC(P): 0.65652 (57)
AUC(S): 0.59390 (10)

ranking function: **zscore** 0.43789 +0.05213

	N [600]		**P** [131]		**S** [209]		**P+S** [59]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	2 -1	20%	4	40%	3 +1	30%	1	10%
30	7 -1	23%	9	30%	8 +2	27%	6 -1	20%
60	17 -1	28%	18 -2	30%	13	22%	12 +3	20%
90	25 -7	28%	29 +4	32%	20 +3	22%	16	18%
120	37 -8	31%	38 +2	32%	25 +4	21%	20 +2	17%
150	51 -4	34%	46 +2	31%	33 +2	22%	20	13%

configuration:
`400-entropy-2-d4`

F_1 score: 0.50950 (79)
AUC(P): 0.69462 (12)
AUC(S): 0.59230 (11)

(a) progressive weights

ranking function: **sum** 0.36328 +0.02876

	N [600]		**P** [131]		**S** [209]		**P+S** [59]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	2 -2	20%	5 +1	50%	3 +1	30%	0	0%
30	6 -1	20%	10	33%	6 -1	20%	8 +2	27%
60	11 -4	18%	25 +1	42%	13 +3	22%	11	18%
90	24 -1	27%	35	39%	15 -2	17%	16 +3	18%
120	38 -1	32%	43 +2	36%	22	18%	17 -1	14%
150	52 -3	35%	51 +3	34%	26 -2	17%	21 +2	14%

configuration:
`200-entropy-2-d5`

F_1 score: 0.52200 (62)
AUC(P): 0.67133 (43)
AUC(S): 0.56952 (32)

ranking function: **scaled** 0.37050 +0.03481

	N [600]		**P** [131]		**S** [209]		**P+S** [59]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	1 -3	10%	5 +1	50%	3 +1	30%	1 +1	10%
30	6 -2	20%	11 +1	37%	5 -1	17%	8 +2	27%
60	13	22%	25 +1	42%	11 -2	18%	11 +1	18%
90	21 -5	23%	36 +3	40%	19 +2	21%	14	16%
120	37 -5	31%	42 +3	35%	24 +3	20%	17 -1	14%
150	53 -1	35%	48	32%	30 +1	20%	19	13%

configuration:
`100-entropy-2-d4`

F_1 score: 0.50800 (84)
AUC(P): 0.65652 (57)
AUC(S): 0.59390 (10)

ranking function: **zscore** 0.38412 +0.03549

	N [600]		**P** [131]		**S** [209]		**P+S** [59]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	2 -1	20%	4	40%	3 +1	30%	1	10%
30	7 -1	23%	9	30%	8 +2	27%	6 -1	20%
60	17 -1	28%	18 -2	30%	13	22%	12 +3	20%
90	25 -7	28%	29 +4	32%	20 +3	22%	16	18%
120	37 -8	31%	38 +2	32%	25 +4	21%	20 +2	17%
150	51 -4	34%	46 +2	31%	33 +2	22%	20	13%

configuration:
`400-entropy-2-d4`

F_1 score: 0.50950 (79)
AUC(P): 0.69462 (12)
AUC(S): 0.59230 (11)

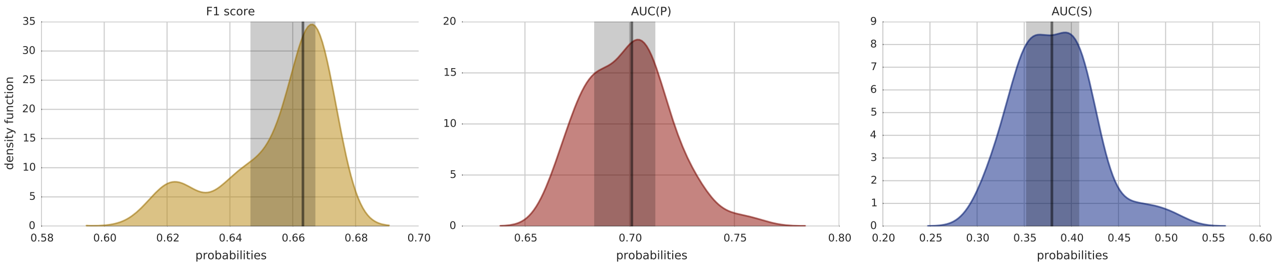
(b) split weights

eTable 7: Results of simulated recruitment for the **PROCOAC selection model (2 years shift)** configurations with best recruitment score (given at the top of each table). The differences in recruitment compared to the model configuration with highest F_1 score are highlighted with green (positive) and red (negative) colour. For each of the ranking functions the used model configuration and its prediction quality measured with F_1 score and AUC is given for reference (rank of the measure among all configurations is given in brackets).

$\mathbf{F_1}$ **score**	**configuration**
0.671	600-gini-2-d8
0.671	800-gini-2-d8
0.671	1000-gini-2-d8
…	
0.6195	400-entropy-2-d4
0.6185	200-entropy-2-d4
0.614	100-entropy-2-d4

AUC(P)	**configuration**
0.7579	200-entropy-2-d10
0.74864	600-entropy-2-d10
0.73559	800-entropy-2-d10
…	
0.66532	400-entropy-2-d4
0.66488	100-gini-2-d6
0.66409	600-entropy-2-d7

AUC(S)	**configuration**
0.50995	100-gini-2-d5
0.49726	100-entropy-2-d10
0.47931	200-entropy-2-d8
…	
0.31457	600-gini-2-d9
0.31423	600-entropy-2-d7
0.30115	100-entropy-2-d5



eFigure 7: Distribution of quality measures across all **PROCOAC selection model (3 years shift)** configurations. The grey area shows interquartile range and thick vertical line indicates the median value. Tables show 3 top/bottom scores and corresponding model parameters.

ranking function: **sum** 0.19963 +0.04667

	N [695]		**P** [115]		**S** [109]		**P+S** [24]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	4 -1	40%	6 +2	60%	0	0%	0 -1	0%
30	12 -4	40%	14 +3	47%	2	7%	2 +1	7%
60	26 -5	43%	27 +6	45%	3 -1	5%	4	7%
90	41 -5	46%	38 +2	42%	7 +3	8%	4	4%
120	62	52%	44	37%	10	8%	4	3%
150	78 -1	52%	52 +1	35%	13 +1	9%	7 -1	5%

configuration:
600-gini-2-d4

$\mathbf{F_1}$ **score:** 0.62550 (75)
AUC(P): 0.67687 (72)
AUC(S): 0.39203 (35)

ranking function: **scaled** 0.17176 +0.07243

	N [695]		**P** [115]		**S** [109]		**P+S** [24]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	4 -3	40%	6 +4	60%	0	0%	0 -1	0%
30	13 -4	43%	14 +4	47%	2	7%	1	3%
60	31 -3	52%	22 +3	37%	3 -1	5%	4 +1	7%
90	45 -7	50%	34 +4	38%	7 +3	8%	4	4%
120	62 -7	52%	45 +4	38%	8 +4	7%	5 -1	4%
150	77 -8	51%	54 +8	36%	12 +1	8%	7 -1	5%

configuration:
200-entropy-2-d4

$\mathbf{F_1}$ **score:** 0.61850 (83)
AUC(P): 0.67400 (76)
AUC(S): 0.38778 (39)

ranking function: **zscore** 0.10859 +0.07952

	N [695]		**P** [115]		**S** [109]		**P+S** [24]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	5 -3	50%	5 +4	50%	0 -1	0%	0	0%
30	16 -4	53%	12 +4	40%	1	3%	1	3%
60	34 -6	57%	21 +5	35%	2 -1	3%	3 +2	5%
90	52 -6	58%	27 +2	30%	4	4%	7 +4	8%
120	73 -3	61%	34 +2	28%	6 +1	5%	7	6%
150	91 -3	61%	42 +3	28%	9	6%	8	5%

configuration:
600-entropy-2-d4

$\mathbf{F_1}$ **score:** 0.62050 (79)
AUC(P): 0.68470 (61)
AUC(S): 0.45113 (6)

(a) progressive weights

ranking function: **sum** 0.16284 +0.01514

	N [695]		**P** [115]		**S** [109]		**P+S** [24]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	5	50%	4	40%	0	0%	1	10%
30	12 -4	40%	13 +2	43%	3 +1	10%	2 +1	7%
60	26 -5	43%	26 +5	43%	4	7%	4	7%
90	43 -3	48%	38 +2	42%	5 +1	6%	4	4%
120	60 -2	50%	46 +2	38%	10	8%	4	3%
150	81 +2	54%	51	34%	12	8%	6 -2	4%

configuration:
1000-gini-2-d5

$\mathbf{F_1}$ **score:** 0.64700 (63)
AUC(P): 0.69741 (46)
AUC(S): 0.34269 (72)

ranking function: **scaled** 0.14999 +0.01797

	N [695]		**P** [115]		**S** [109]		**P+S** [24]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	5 -2	50%	3 +1	30%	1 +1	10%	1	10%
30	17	57%	10	33%	2	7%	1	3%
60	32 -2	53%	23 +4	38%	2 -2	3%	3	5%
90	47 -5	52%	31 +1	34%	7 +3	8%	5 +1	6%
120	62 -7	52%	44 +3	37%	8 +4	7%	6	5%
150	83 -2	55%	50 +4	33%	11	7%	6 -2	4%

configuration:
200-entropy-2-d5

$\mathbf{F_1}$ **score:** 0.63850 (71)
AUC(P): 0.70149 (42)
AUC(S): 0.41199 (20)

ranking function: **zscore** 0.13388 +0.02925

	N [695]		**P** [115]		**S** [109]		**P+S** [24]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	4 -4	40%	4 +3	40%	1	10%	1 +1	10%
30	18 -2	60%	10 +2	33%	1	3%	1	3%
60	39 -1	65%	16	27%	3	5%	2 +1	3%
90	54 -4	60%	27 +2	30%	6 +2	7%	3	3%
120	73 -3	61%	32	27%	9 +4	8%	6 -1	5%
150	94	63%	36 -3	24%	12 +3	8%	8	5%

configuration:
100-entropy-2-d6

$\mathbf{F_1}$ **score:** 0.65200 (60)
AUC(P): 0.66777 (81)
AUC(S): 0.33462 (75)

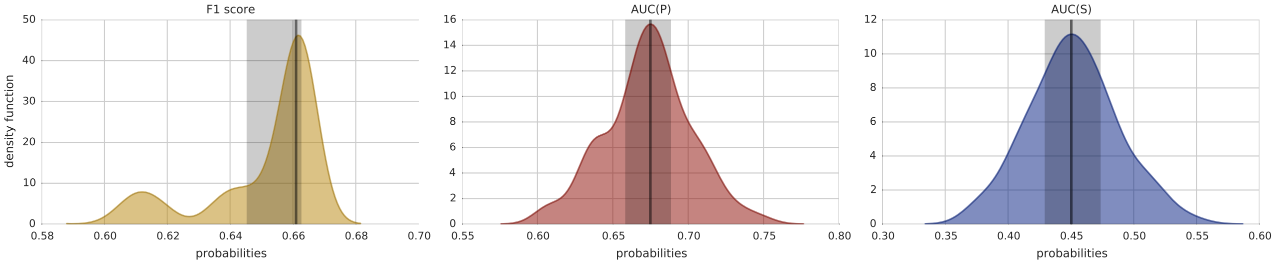
(b) split weights

eTable 8: Results of simulated recruitment for the **PROCOAC selection model (3 years shift)** configurations with best recruitment score (given at the top of each table). The differences in recruitment compared to the model configuration with highest F_1 score are highlighted with green (positive) and red (negative) colour. For each of the ranking functions the used model configuration and its prediction quality measured with F_1 score and AUC is given for reference (rank of the measure among all configurations is given in brackets).

F_1 score	**configuration**
0.665	400-entropy-2-d9
0.6645	200-entropy-2-d9
0.6645	800-entropy-2-d9
...	
0.6095	600-entropy-2-d4
0.608	200-entropy-2-d4
0.6045	100-entropy-2-d4

AUC(P)	**configuration**
0.74779	100-entropy-2-d10
0.73464	800-gini-2-d10
0.72941	1000-gini-2-d10
...	
0.61348	400-entropy-2-d8
0.60744	200-gini-2-d6
0.60453	100-entropy-2-d7

AUC(S)	**configuration**
0.54454	100-entropy-2-d8
0.52455	100-entropy-2-d7
0.52258	200-gini-2-d5
...	
0.37827	1000-entropy-2-d8
0.37705	400-entropy-2-d7
0.37631	200-entropy-2-d10



eFigure 8: Distribution of quality measures across all **PROCOAC selection model (5 years shift)** configurations. The grey area shows interquartile range and thick vertical line indicates the median value. Tables show 3 top/bottom scores and corresponding model parameters.

ranking function: **sum** 0.12759 +0.08711

	N [748]		**P** [115]		**S** [109]		**P+S** [25]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	3 -4	30%	6 +3	60%	1 +1	10%	0	0%
30	15 -2	50%	12 +2	40%	2	7%	1	3%
60	36 -3	60%	17	28%	4 +1	7%	3 +2	5%
90	53 -5	59%	26 +3	29%	7 +1	8%	4 +1	4%
120	72 -11	60%	33 +8	28%	10 +2	8%	5 +1	4%
150	90 -10	60%	42 +10	28%	12 -1	8%	6 +1	4%

configuration:
`200-gini-2-d4`

F_1 score: 0.61450 (77)
AUC(P): 0.67888 (35)
AUC(S): 0.47582 (19)

ranking function: **scaled** 0.11106 +0.10766

	N [748]		**P** [115]		**S** [109]		**P+S** [25]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	4 -3	40%	5 +2	50%	1 +1	10%	0	0%
30	16 -4	53%	10 +3	33%	3 +1	10%	1	3%
60	36 -7	60%	18 +6	30%	3 -1	5%	3 +2	5%
90	55 -8	61%	24 +5	27%	7	8%	4 +3	4%
120	72 -11	60%	31 +6	26%	12 +2	10%	5 +3	4%
150	93 -10	62%	38 +6	25%	13 +1	9%	6 +3	4%

configuration:
`100-gini-2-d4`

F_1 score: 0.61100 (78)
AUC(P): 0.67596 (40)
AUC(S): 0.45560 (37)

ranking function: **zscore** 0.06046 +0.09153

	N [748]		**P** [115]		**S** [109]		**P+S** [25]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	6 -2	60%	3 +1	30%	1 +1	10%	0	0%
30	18 -4	60%	10 +4	33%	1	3%	1	3%
60	39 -6	65%	15 +5	25%	4	7%	2 +1	3%
90	56 -10	62%	19 +4	21%	11 +3	12%	4 +3	4%
120	80 -6	67%	25 +4	21%	11	9%	4 +2	3%
150	101 -7	67%	30 +2	20%	15 +3	10%	4 +2	3%

configuration:
`100-gini-2-d4`

F_1 score: 0.61100 (78)
AUC(P): 0.67596 (40)
AUC(S): 0.45560 (37)

(a) progressive weights

ranking function: **sum** 0.14726 +0.04191

	N [748]		**P** [115]		**S** [109]		**P+S** [25]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	6 -1	60%	1 -2	10%	2 +2	20%	1 +1	10%
30	17	57%	9 -1	30%	3 +1	10%	1	3%
60	35 -4	58%	17	28%	5 +2	8%	3 +2	5%
90	56 -2	62%	24 +1	27%	6	7%	4 +1	4%
120	78 -5	65%	30 +5	25%	8	7%	4	3%
150	99 -1	66%	37 +5	25%	9 -4	6%	5	3%

configuration:
`100-gini-2-d9`

F_1 score: 0.66200 (35)
AUC(P): 0.66996 (49)
AUC(S): 0.41570 (70)

ranking function: **scaled** 0.14087 +0.04530

	N [748]		**P** [115]		**S** [109]		**P+S** [25]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	4 -3	40%	5 +2	50%	1 +1	10%	0	0%
30	16 -4	53%	10 +3	33%	3 +1	10%	1	3%
60	36 -7	60%	18 +6	30%	3 -1	5%	3 +2	5%
90	55 -8	61%	24 +5	27%	7	8%	4 +3	4%
120	72 -11	60%	31 +6	26%	12 +2	10%	5 +3	4%
150	93 -10	62%	38 +6	25%	13 +1	9%	6 +3	4%

configuration:
`100-gini-2-d4`

F_1 score: 0.61100 (78)
AUC(P): 0.67596 (40)
AUC(S): 0.45560 (37)

ranking function: **zscore** 0.12757 +0.04001

	N [748]		**P** [115]		**S** [109]		**P+S** [25]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	6 -2	60%	3 +1	30%	1 +1	10%	0	0%
30	18 -4	60%	10 +4	33%	1	3%	1	3%
60	39 -6	65%	15 +5	25%	4	7%	2 +1	3%
90	56 -10	62%	19 +4	21%	11 +3	12%	4 +3	4%
120	80 -6	67%	25 +4	21%	11	9%	4 +2	3%
150	101 -7	67%	30 +2	20%	15 +3	10%	4 +2	3%

configuration:
`100-gini-2-d4`

F_1 score: 0.61100 (78)
AUC(P): 0.67596 (40)
AUC(S): 0.45560 (37)

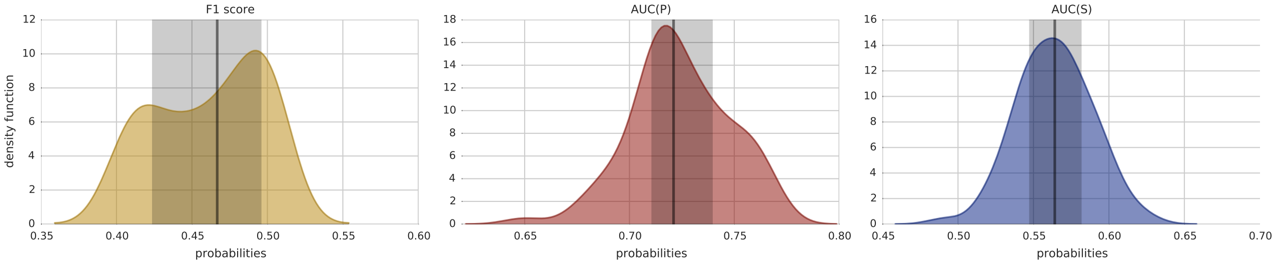
(b) split weights

eTable 9: Results of simulated recruitment for the **PROCOAC selection model (5 years shift)** configurations with best recruitment score (given at the top of each table). The differences in recruitment compared to the model configuration with highest F_1 score are highlighted with green (positive) and red (negative) colour. For each of the ranking functions the used model configuration and its prediction quality measured with F_1 score and AUC is given for reference (rank of the measure among all configurations is given in brackets).

F_1 score	configuration
0.5075	600-gini-2-d10
0.5075	200-gini-2-d10
0.5075	800-gini-2-d10
…	
0.4055	200-entropy-2-d4
0.4055	400-entropy-2-d4
0.4055	600-gini-2-d4

AUC(P)	configuration
0.77101	400-gini-2-d9
0.7691	200-gini-2-d9
0.76522	1000-gini-2-d9
…	
0.67686	800-gini-2-d5
0.67596	100-gini-2-d5
0.6497	200-entropy-2-d4

AUC(S)	configuration
0.62573	100-entropy-2-d9
0.61445	600-gini-2-d9
0.61146	400-gini-2-d9
…	
0.51838	100-gini-2-d10
0.51642	1000-entropy-2-d4
0.49083	800-gini-2-d4



eFigure 9: Distribution of quality measures across all **DIGICOD selection model (no shift)** configurations. The grey area shows interquartile range and thick vertical line indicates the median value. Tables show 3 top/bottom scores and corresponding model parameters.

ranking function: **sum** 0.40107 +0.05096

	N [1749]		**P** [358]		**S** [591]		**P+S** [160]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	0 -2	0%	3	30%	2	20%	5 +2	50%
30	7 -1	23%	12 +2	40%	3 -1	10%	8	27%
60	19 -4	32%	25 +7	42%	4 -4	7%	12 +1	20%
90	33 -5	37%	35 +12	39%	6 -9	7%	16 +2	18%
120	52 -3	43%	43 +13	36%	7 -12	6%	18 +2	15%
150	65 -8	43%	49 +11	33%	16 -7	11%	20 +4	13%

configuration: `800-entropy-2-d6`

F_1 score: 0.44600 (54)
AUC(P): 0.73757 (23)
AUC(S): 0.57591 (25)

ranking function: **scaled** 0.40534 +0.10619

	N [1749]		**P** [358]		**S** [591]		**P+S** [160]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	0 -3	0%	3 +1	30%	2	20%	5 +2	50%
30	6 -3	20%	15 +5	50%	3 -1	10%	6 -1	20%
60	18 -9	30%	25 +11	42%	4 -6	7%	13 +4	22%
90	35 -9	39%	33 +13	37%	7 -7	8%	15 +3	17%
120	50 -12	42%	41 +17	34%	12 -8	10%	17 +3	14%
150	63 -12	42%	46 +13	31%	20 -4	13%	21 +3	14%

configuration: `400-entropy-2-d5`

F_1 score: 0.42200 (71)
AUC(P): 0.73463 (29)
AUC(S): 0.59744 (8)

ranking function: **zscore** 0.37788 +0.07958

	N [1749]		**P** [358]		**S** [591]		**P+S** [160]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	1 -2	10%	2	20%	2	20%	5 +2	50%
30	8 -1	27%	12 +2	40%	2 -2	7%	8 +1	27%
60	21 -6	35%	23 +9	38%	5 -5	8%	11 +2	18%
90	37 -7	41%	29 +9	32%	10 -4	11%	14 +2	16%
120	51 -12	42%	36 +13	30%	16 -4	13%	17 +3	14%
150	65 -10	43%	43 +11	29%	22 -3	15%	20 +2	13%

configuration: `400-entropy-2-d5`

F_1 score: 0.42200 (71)
AUC(P): 0.73463 (29)
AUC(S): 0.59744 (8)

(a) progressive weights

ranking function: **sum** 0.34905 +0.01120

	N [1749]		**P** [358]		**S** [591]		**P+S** [160]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	1 -1	10%	4 +1	40%	2	20%	3	30%
30	7 -1	23%	12 +2	40%	4	13%	7 -1	23%
60	25 +2	42%	15 -3	25%	9 +1	15%	11	18%
90	38	42%	21 -2	23%	16 +1	18%	15 +1	17%
120	53 -2	44%	27 -3	22%	23 +4	19%	17 +1	14%
150	65 -8	43%	39 +1	26%	28 +5	19%	18 +2	12%

configuration: `100-gini-2-d10`

F_1 score: 0.50600 (6)
AUC(P): 0.72094 (43)
AUC(S): 0.51838 (82)

ranking function: **scaled** 0.35374 +0.03652

	N [1749]		**P** [358]		**S** [591]		**P+S** [160]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	1 -2	10%	2	20%	2	20%	5 +2	50%
30	8 -1	27%	10	33%	2 -2	7%	10 +3	33%
60	22 -5	37%	22 +8	37%	4 -6	7%	12 +3	20%
90	34 -10	38%	34 +14	38%	7 -7	8%	15 +3	17%
120	47 -15	39%	40 +16	33%	15 -5	12%	18 +4	15%
150	61 -14	41%	46 +13	31%	23 -1	15%	20 +2	13%

configuration: `200-gini-2-d6`

F_1 score: 0.44600 (53)
AUC(P): 0.70588 (69)
AUC(S): 0.56553 (42)

ranking function: **zscore** 0.35132 +0.03387

	N [1749]		**P** [358]		**S** [591]		**P+S** [160]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	2 -1	20%	1 -1	10%	2	20%	5 +2	50%
30	11 +2	37%	8 -2	27%	3 -1	10%	8 +1	27%
60	26 -1	43%	17 +3	28%	6 -4	10%	11 +2	18%
90	36 -8	40%	24 +4	27%	16 +2	18%	14 +2	16%
120	55 -8	46%	30 +7	25%	18 -2	15%	17 +3	14%
150	70 -5	47%	35 +3	23%	27 +2	18%	18	12%

configuration: `100-gini-2-d8`

F_1 score: 0.48500 (30)
AUC(P): 0.74470 (19)
AUC(S): 0.56772 (37)

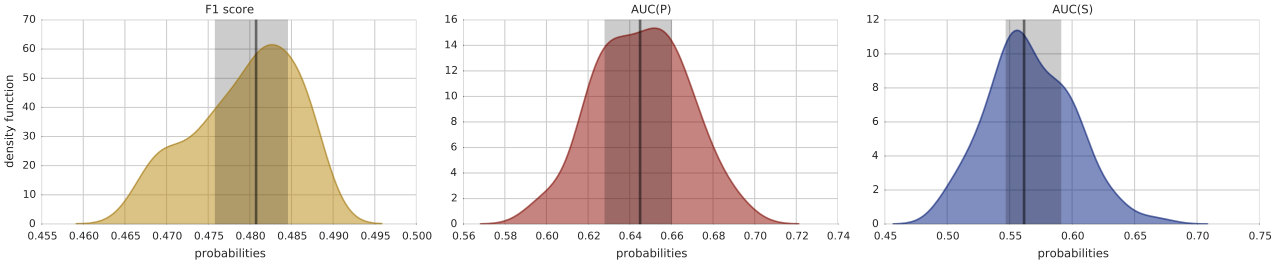
(b) split weights

eTable 10: Results of simulated recruitment for the **DIGICOD selection model (no shift)** configurations with best recruitment score (given at the top of each table). The differences in recruitment compared to the model configuration with highest F_1 score are highlighted with green (positive) and red (negative) colour. For each of the ranking functions the used model configuration and its prediction quality measured with F_1 score and AUC is given for reference (rank of the measure among all configurations is given in brackets).

F_1 score	**configuration**
0.488	800-entropy-2-d7
0.4875	1000-gini-2-d6
0.4875	600-gini-2-d6
...	
0.4675	100-gini-2-d4
0.4675	200-entropy-2-d4
0.467	100-entropy-2-d4

AUC(P)	**configuration**
0.69284	600-entropy-2-d7
0.69242	1000-gini-2-d9
0.68539	600-gini-2-d9
...	
0.60083	100-gini-2-d10
0.59814	600-gini-2-d5
0.59669	100-entropy-2-d5

AUC(S)	**configuration**
0.66614	100-entropy-2-d5
0.63768	200-gini-2-d9
0.63185	100-gini-2-d9
...	
0.50369	200-gini-2-d6
0.50276	800-gini-2-d8
0.49902	100-entropy-2-d7



eFigure 10: Distribution of quality measures across all **DIGICOD selection model (2 years shift)** configurations. The grey area shows interquartile range and thick vertical line indicates the median value. Tables show 3 top/bottom scores and corresponding model parameters.

ranking function: **sum** 0.27076 +0.04159

	N [601]		**P** [131]		**S** [210]		**P+S** [59]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	3	30%	5 -1	50%	1 +1	10%	1	10%
30	13 -1	43%	11 -1	37%	4 +3	13%	2 -1	7%
60	26 -3	43%	18 +2	30%	8	13%	8 +1	13%
90	39 -3	43%	24 +1	27%	15	17%	12 +2	13%
120	57 -5	48%	30	25%	20 +3	17%	13 +2	11%
150	70 -10	47%	36 +3	24%	29 +4	19%	15 +3	10%

configuration:
`400-entropy-2-d4`

F_1 score: 0.46900 (79)
AUC(P): 0.64130 (48)
AUC(S): 0.58921 (23)

ranking function: **scaled** 0.27646 +0.07134

	N [601]		**P** [131]		**S** [210]		**P+S** [59]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	3 -1	30%	6	60%	0	0%	1 +1	10%
30	13 -2	43%	10 +1	33%	4 +2	13%	3 -1	10%
60	25 -4	42%	17 +1	28%	9 +1	15%	9 +2	15%
90	39 -7	43%	24 +3	27%	15 +2	17%	12 +2	13%
120	57 -7	48%	29 +1	24%	20 +3	17%	14 +3	12%
150	71 -8	47%	35	23%	29 +6	19%	15 +2	10%

configuration:
`400-entropy-2-d4`

F_1 score: 0.46900 (79)
AUC(P): 0.64130 (48)
AUC(S): 0.58921 (23)

ranking function: **zscore** 0.29147 +0.09761

	N [601]		**P** [131]		**S** [210]		**P+S** [59]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	3 -2	30%	4 -1	40%	1 +1	10%	2 +2	20%
30	13 -2	43%	9	30%	3 +1	10%	5 +1	17%
60	26 -2	43%	17	28%	10 +1	17%	7 +1	12%
90	38 -10	42%	23 +3	26%	17 +4	19%	12 +3	13%
120	53 -12	44%	28 +4	23%	25 +6	21%	14 +2	12%
150	73 -7	49%	33 +1	22%	28 +3	19%	16 +3	11%

configuration:
`100-gini-2-d4`

F_1 score: 0.46750 (82)
AUC(P): 0.64346 (45)
AUC(S): 0.59745 (16)

(a) progressive weights

ranking function: **sum** 0.27167 +0.04236

	N [601]		**P** [131]		**S** [210]		**P+S** [59]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	4 +1	40%	3 -3	30%	1 +1	10%	2 +1	20%
30	14	47%	10 -2	33%	3 +2	10%	3	10%
60	23 -6	38%	20 +4	33%	9 +1	15%	8 +1	13%
90	40 -2	44%	23	26%	15	17%	12 +2	13%
120	59 -3	49%	28 -2	23%	20 +3	17%	13 +2	11%
150	76 -4	51%	33	22%	25	17%	16 +4	11%

configuration:
`100-gini-2-d4`

F_1 score: 0.46750 (82)
AUC(P): 0.64346 (45)
AUC(S): 0.59745 (16)

ranking function: **scaled** 0.27675 +0.05614

	N [601]		**P** [131]		**S** [210]		**P+S** [59]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	3 -1	30%	4 -2	40%	1 +1	10%	2 +2	20%
30	16 +1	53%	7 -2	23%	2	7%	5 +1	17%
60	25 -4	42%	18 +2	30%	9 +1	15%	8 +1	13%
90	43 -3	48%	23 +2	26%	12 -1	13%	12 +2	13%
120	53 -11	44%	30 +2	25%	23 +6	19%	14 +3	12%
150	75 -4	50%	33 -2	22%	27 +4	18%	15 +2	10%

configuration:
`100-entropy-2-d4`

F_1 score: 0.46700 (84)
AUC(P): 0.64518 (42)
AUC(S): 0.59520 (18)

ranking function: **zscore** 0.29239 +0.07340

	N [601]		**P** [131]		**S** [210]		**P+S** [59]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	3 -2	30%	4 -1	40%	1 +1	10%	2 +2	20%
30	13 -2	43%	9	30%	3 +1	10%	5 +1	17%
60	26 -2	43%	17	28%	10 +1	17%	7 +1	12%
90	38 -10	42%	23 +3	26%	17 +4	19%	12 +3	13%
120	53 -12	44%	28 +4	23%	25 +6	21%	14 +2	12%
150	73 -7	49%	33 +1	22%	28 +3	19%	16 +3	11%

configuration:
`100-gini-2-d4`

F_1 score: 0.46750 (82)
AUC(P): 0.64346 (45)
AUC(S): 0.59745 (16)

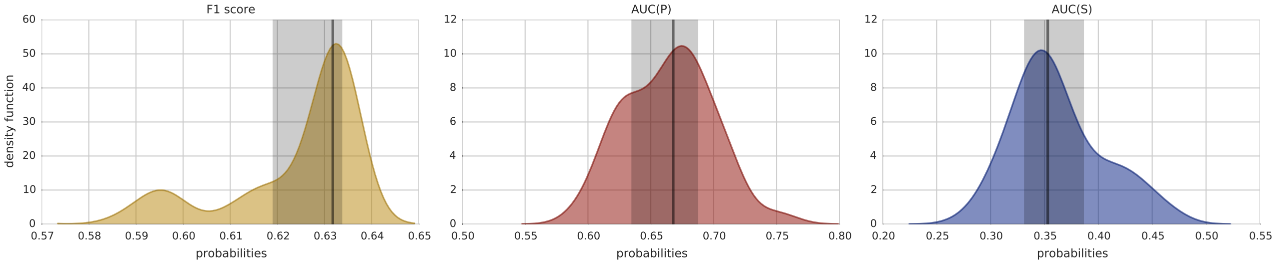
(b) split weights

eTable 11: Results of simulated recruitment for the **DIGICOD selection model (2 years shift)** configurations with best recruitment score (given at the top of each table). The differences in recruitment compared to the model configuration with highest F_1 score are highlighted with green (positive) and red (negative) colour. For each of the ranking functions the used model configuration and its prediction quality measured with F_1 score and AUC is given for reference (rank of the measure among all configurations is given in brackets).

F_1 score	configuration
0.635	400-entropy-2-d9
0.635	800-gini-2-d9
0.635	1000-gini-2-d8
...	
0.592	200-entropy-2-d4
0.59	100-gini-2-d4
0.5875	100-entropy-2-d4

AUC(P)	configuration
0.75403	600-entropy-2-d6
0.74839	200-entropy-2-d6
0.72026	1000-entropy-2-d6
...	
0.60554	100-gini-2-d10
0.59794	600-gini-2-d4
0.59259	600-gini-2-d10

AUC(S)	configuration
0.4697	800-entropy-2-d10
0.46506	600-gini-2-d10
0.44963	100-entropy-2-d5
...	
0.29548	1000-entropy-2-d4
0.29349	400-gini-2-d5
0.27736	400-entropy-2-d6



eFigure 11: Distribution of quality measures across all **DIGICOD selection model (3 years shift)** configurations. The grey area shows interquartile range and thick vertical line indicates the median value. Tables show 3 top/bottom scores and corresponding model parameters.

ranking function: **sum** 0.11305 +0.09372

	N [695]		**P** [115]		**S** [109]		**P+S** [24]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	5	50%	5	50%	0	0%	0	0%
30	15 -4	50%	11 +2	37%	3 +2	10%	1	3%
60	33 -10	55%	20 +9	33%	5 +1	8%	2	3%
90	54 -10	60%	26 +11	29%	7	8%	3 -1	3%
120	71 -16	59%	32 +13	27%	14 +4	12%	3 -1	2%
150	93 -19	62%	37 +16	25%	17 +4	11%	3 -1	2%

configuration:
200-gini-2-d4

F_1 score: 0.59600 (77)
AUC(P): 0.67962 (28)
AUC(S): 0.35275 (43)

ranking function: **scaled** 0.09109 +0.09426

	N [695]		**P** [115]		**S** [109]		**P+S** [24]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	6	60%	4	40%	0	0%	0	0%
30	16 -5	53%	10 +2	33%	3 +2	10%	1 +1	3%
60	34 -9	57%	20 +9	33%	5 +1	8%	1 -1	2%
90	54 -11	60%	24 +10	27%	9 +2	10%	3 -1	3%
120	74 -14	62%	30 +11	25%	13 +4	11%	3 -1	2%
150	97 -15	65%	35 +15	23%	15 +1	10%	3 -1	2%

configuration:
200-gini-2-d4

F_1 score: 0.59600 (77)
AUC(P): 0.67962 (28)
AUC(S): 0.35275 (43)

ranking function: **zscore** 0.05304 +0.05882

	N [695]		**P** [115]		**S** [109]		**P+S** [24]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	7 +1	70%	3 -1	30%	0	0%	0	0%
30	19 -1	63%	9 +1	30%	1 -1	3%	1 +1	3%
60	34 -10	57%	20 +8	33%	5 +2	8%	1	2%
90	57 -8	63%	23 +8	26%	9 +3	10%	1 -3	1%
120	79 -10	66%	27 +10	22%	13 +3	11%	1 -3	1%
150	97 -17	65%	34 +15	23%	16 +3	11%	3 -1	2%

configuration:
200-gini-2-d4

F_1 score: 0.59600 (77)
AUC(P): 0.67962 (28)
AUC(S): 0.35275 (43)

(a) progressive weights

ranking function: **sum** 0.13653 +0.02929

	N [695]		**P** [115]		**S** [109]		**P+S** [24]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	5	50%	5	50%	0	0%	0	0%
30	16 -3	53%	9	30%	4 +3	13%	1	3%
60	38 -5	63%	11	18%	9 +5	15%	2	3%
90	60 -4	67%	18 +3	20%	10 +3	11%	2 -2	2%
120	83 -4	69%	23 +4	19%	12 +2	10%	2 -2	2%
150	102 -10	68%	27 +6	18%	17 +4	11%	4	3%

configuration:
100-gini-2-d6

F_1 score: 0.62450 (59)
AUC(P): 0.69898 (13)
AUC(S): 0.39538 (18)

ranking function: **scaled** 0.12788 +0.02951

	N [695]		**P** [115]		**S** [109]		**P+S** [24]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	6	60%	4	40%	0	0%	0	0%
30	16 -5	53%	10 +2	33%	3 +2	10%	1 +1	3%
60	34 -9	57%	20 +9	33%	5 +1	8%	1 -1	2%
90	54 -11	60%	24 +10	27%	9 +2	10%	3 -1	3%
120	74 -14	62%	30 +11	25%	13 +4	11%	3 -1	2%
150	97 -15	65%	35 +15	23%	15 +1	10%	3 -1	2%

configuration:
200-gini-2-d4

F_1 score: 0.59600 (77)
AUC(P): 0.67962 (28)
AUC(S): 0.35275 (43)

ranking function: **zscore** 0.11869 +0.02362

	N [695]		**P** [115]		**S** [109]		**P+S** [24]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	5 -1	50%	4	40%	1 +1	10%	0	0%
30	19 -1	63%	7 -1	23%	4 +2	13%	0	0%
60	43 -1	72%	11 -1	18%	6 +3	10%	0 -1	0%
90	64 -1	71%	15	17%	9 +3	10%	2 -2	2%
120	84 -5	70%	20 +3	17%	12 +2	10%	4	3%
150	106 -8	71%	25 +6	17%	14 +1	9%	5 +1	3%

configuration:
100-entropy-2-d8

F_1 score: 0.63200 (40)
AUC(P): 0.66417 (44)
AUC(S): 0.44677 (4)

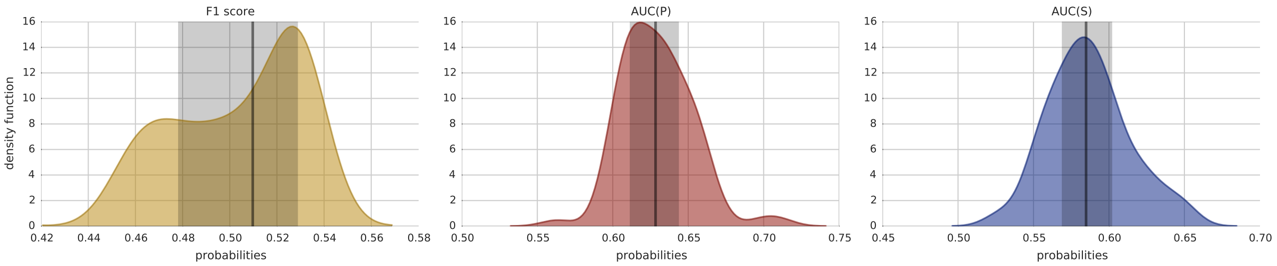
(b) split weights

eTable 12: Results of simulated recruitment for the **DIGICOD selection model (3 years shift)** configurations with best recruitment score (given at the top of each table). The differences in recruitment compared to the model configuration with highest F_1 score are highlighted with green (positive) and red (negative) colour. For each of the ranking functions the used model configuration and its prediction quality measured with F_1 score and AUC is given for reference (rank of the measure among all configurations is given in brackets).

F_1 score	configuration
0.534	1000-gini-2-d9
0.534	800-gini-2-d9
0.5335	600-gini-2-d9
…	
0.458	200-entropy-2-d4
0.458	400-entropy-2-d4
0.4555	100-entropy-2-d4

AUC(P)	configuration
0.71063	100-entropy-2-d8
0.69878	200-entropy-2-d10
0.66968	1000-gini-2-d10
…	
0.59698	800-gini-2-d4
0.59521	1000-gini-2-d8
0.56287	100-gini-2-d7

AUC(S)	configuration
0.65308	200-entropy-2-d8
0.64805	200-gini-2-d5
0.64666	200-entropy-2-d5
…	
0.54374	100-gini-2-d6
0.52782	1000-entropy-2-d8
0.52771	100-entropy-2-d6



eFigure 12: Distribution of quality measures across all **HOSTAS selection model (no shift)** configurations. The grey area shows interquartile range and thick vertical line indicates the median value. Tables show 3 top/bottom scores and corresponding model parameters.

ranking function: **sum** 0.42769 +0.13967

	N [1733]		**P** [358]		**S** [591]		**P+S** [160]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	1 -3	10%	6	60%	1 +1	10%	2 +2	20%
30	6 -6	20%	18 +5	60%	1 -1	3%	5 +2	17%
60	14 -6	23%	31 +3	52%	4 -3	7%	11 +6	18%
90	22 -12	24%	43 +4	48%	9	10%	16 +8	18%
120	37 -10	31%	52 +8	43%	10 -3	8%	21 +5	18%
150	53 -5	35%	59 +11	39%	11 -6	7%	27	18%

configuration:
`400-entropy-2-d4`

F_1 score: 0.45800 (83)
AUC(P): 0.62820 (43)
AUC(S): 0.56992 (62)

ranking function: **scaled** 0.40155 +0.16160

	N [1733]		**P** [358]		**S** [591]		**P+S** [160]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	2 -4	20%	6 +2	60%	1 +1	10%	1 +1	10%
30	8 -6	27%	15 +5	50%	2	7%	5 +1	17%
60	16 -10	27%	26 +5	43%	5 -3	8%	13 +8	22%
90	25 -14	28%	41 +9	46%	6 -4	7%	18 +9	20%
120	38 -10	32%	49 +8	41%	11 -6	9%	22 +8	18%
150	52 -7	35%	58 +11	39%	15 -8	10%	25 +4	17%

configuration:
`600-entropy-2-d4`

F_1 score: 0.45850 (81)
AUC(P): 0.62516 (46)
AUC(S): 0.62772 (8)

ranking function: **zscore** 0.30406 +0.06586

	N [1733]		**P** [358]		**S** [591]		**P+S** [160]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	4	40%	3	30%	2	20%	1	10%
30	11 -5	37%	10 +3	33%	5 +1	17%	4 +1	13%
60	21 -8	35%	21 +6	35%	11 +1	18%	7 +1	12%
90	38 -6	42%	26 +2	29%	14	16%	12 +4	13%
120	55	46%	31 +2	26%	17 -4	14%	17 +2	14%
150	65	43%	42 +3	28%	22 -5	15%	21 +2	14%

configuration:
`100-entropy-2-d5`

F_1 score: 0.47050 (72)
AUC(P): 0.61668 (57)
AUC(S): 0.55021 (79)

(a) progressive weights

ranking function: **sum** 0.31392 +0.09032

	N [1733]		**P** [358]		**S** [591]		**P+S** [160]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	2 -2	20%	6	60%	1 +1	10%	1 +1	10%
30	6 -6	20%	16 +3	53%	1 -1	3%	7 +4	23%
60	17 -3	28%	25 -3	42%	4 -3	7%	14 +9	23%
90	22 -12	24%	44 +5	49%	9	10%	15 +7	17%
120	35 -12	29%	52 +8	43%	13	11%	20 +4	17%
150	48 -10	32%	59 +11	39%	16 -1	11%	27	18%

configuration:
`100-gini-2-d4`

F_1 score: 0.45950 (78)
AUC(P): 0.60619 (73)
AUC(S): 0.60128 (23)

ranking function: **scaled** 0.31180 +0.08828

	N [1733]		**P** [358]		**S** [591]		**P+S** [160]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	3 -3	30%	5 +1	50%	1 +1	10%	1 +1	10%
30	8 -6	27%	13 +3	43%	3 +1	10%	6 +2	20%
60	16 -10	27%	27 +6	45%	5 -3	8%	12 +7	20%
90	29 -10	32%	36 +4	40%	7 -3	8%	18 +9	20%
120	37 -11	31%	47 +6	39%	13 -4	11%	23 +9	19%
150	50 -9	33%	58 +11	39%	17 -6	11%	25 +4	17%

configuration:
`100-gini-2-d4`

F_1 score: 0.45950 (78)
AUC(P): 0.60619 (73)
AUC(S): 0.60128 (23)

ranking function: **zscore** 0.29884 +0.03957

	N [1733]		**P** [358]		**S** [591]		**P+S** [160]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	4	40%	3	30%	1 -1	10%	2 +1	20%
30	13 -3	43%	6 -1	20%	4	13%	7 +4	23%
60	29	48%	14 -1	23%	8 -2	13%	9 +3	15%
90	42 -2	47%	21 -3	23%	16 +2	18%	11 +3	12%
120	56 +1	47%	26 -3	22%	23 +2	19%	15	12%
150	70 +5	47%	36 -3	24%	26 -1	17%	18 -1	12%

configuration:
`200-entropy-2-d5`

F_1 score: 0.47200 (71)
AUC(P): 0.60272 (78)
AUC(S): 0.64666 (3)

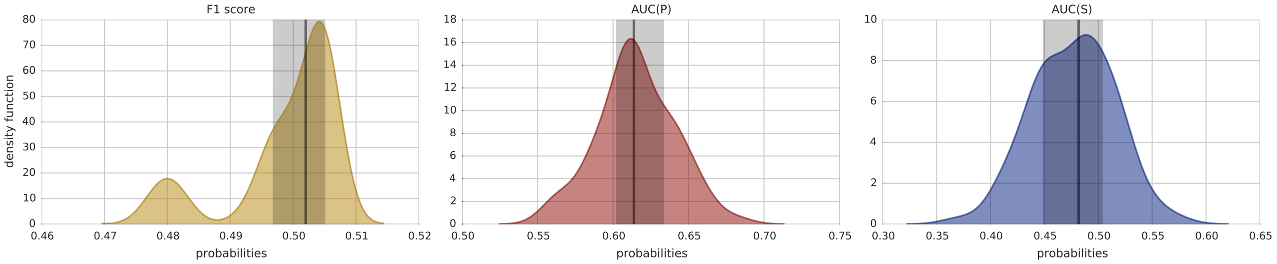
(b) split weights

eTable 13: Results of simulated recruitment for the **HOSTAS selection model (no shift)** configurations with best recruitment score (given at the top of each table). The differences in recruitment compared to the model configuration with highest F_1 score are highlighted with green (positive) and red (negative) colour. For each of the ranking functions the used model configuration and its prediction quality measured with F_1 score and AUC is given for reference (rank of the measure among all configurations is given in brackets).

F_1 score	configuration
0.5065	800-gini-2-d7
0.5065	1000-gini-2-d7
0.5065	1000-gini-2-d6
...	
0.478	200-entropy-2-d4
0.4775	100-entropy-2-d4
0.4775	100-gini-2-d4

AUC(P)	configuration
0.683	400-entropy-2-d10
0.6682	600-gini-2-d10
0.6617	200-entropy-2-d10
...	
0.56448	200-gini-2-d9
0.5615	200-entropy-2-d8
0.55471	100-gini-2-d4

AUC(S)	configuration
0.57127	200-gini-2-d9
0.55159	800-entropy-2-d9
0.54481	800-gini-2-d10
...	
0.40815	200-entropy-2-d4
0.40785	200-gini-2-d4
0.37153	400-gini-2-d4



eFigure 13: Distribution of quality measures across all **HOSTAS selection model (2 years shift)** configurations. The grey area shows interquartile range and thick vertical line indicates the median value. Tables show 3 top/bottom scores and corresponding model parameters.

ranking function: **sum** 0.33606 +0.05119

	N [598]		**P** [131]		**S** [208]		**P+S** [59]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	1 -2	10%	7 +1	70%	1	10%	1 +1	10%
30	9 -2	30%	13 +1	43%	5 -1	17%	3 +2	10%
60	25 +2	42%	19 -1	32%	9 -3	15%	7 +2	12%
90	35 -5	39%	30 +4	33%	14	16%	11 +1	12%
120	46 -6	38%	38 +4	32%	20	17%	16 +2	13%
150	59 -6	39%	47 +8	31%	25 -2	17%	19	13%

configuration:
`100-entropy-2-d4`

F_1 score: 0.47750 (83)
AUC(P): 0.64205 (12)
AUC(S): 0.48170 (44)

ranking function: **scaled** 0.30863 +0.05184

	N [598]		**P** [131]		**S** [208]		**P+S** [59]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	2 -3	20%	7 +3	70%	0 -1	0%	1 +1	10%
30	11	37%	13 +1	43%	4 -2	13%	2 +1	7%
60	25	42%	19	32%	8 -5	13%	8 +5	13%
90	36 -4	40%	28 +2	31%	16 +2	18%	10	11%
120	46 -7	38%	38 +5	32%	23 +3	19%	13 -1	11%
150	62 -6	41%	44 +7	29%	26 -1	17%	18	12%

configuration:
`100-entropy-2-d4`

F_1 score: 0.47750 (83)
AUC(P): 0.64205 (12)
AUC(S): 0.48170 (44)

ranking function: **zscore** 0.30153 +0.04397

	N [598]		**P** [131]		**S** [208]		**P+S** [59]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	2 -3	20%	7 +3	70%	1	10%	0	0%
30	11	37%	13 +2	43%	4 -3	13%	2 +1	7%
60	24	40%	19 -2	32%	12	20%	5 +2	8%
90	37 -2	41%	28 +2	31%	14 -2	16%	11 +2	12%
120	48 -6	40%	36 +3	30%	22 +2	18%	14 +1	12%
150	63 -9	42%	46 +11	31%	26 +1	17%	15 -3	10%

configuration:
`100-gini-2-d4`

F_1 score: 0.47750 (84)
AUC(P): 0.55471 (84)
AUC(S): 0.49832 (26)

(a) progressive weights

ranking function: **sum** 0.28131 +0.02704

	N [598]		**P** [131]		**S** [208]		**P+S** [59]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	3	30%	4 -2	40%	1	10%	2 +2	20%
30	9 -2	30%	12	40%	6	20%	3 +2	10%
60	25 +2	42%	18 -2	30%	11 -1	18%	6 +1	10%
90	41 +1	46%	26	29%	14	16%	9 -1	10%
120	53 +1	44%	33 -1	28%	19 -1	16%	15 +1	12%
150	69 +4	46%	36 -3	24%	26 -1	17%	19	13%

configuration:
`200-entropy-2-d10`

F_1 score: 0.50050 (51)
AUC(P): 0.66170 (3)
AUC(S): 0.50763 (19)

ranking function: **scaled** 0.27571 +0.03105

	N [598]		**P** [131]		**S** [208]		**P+S** [59]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	4 -1	40%	3 -1	30%	2 +1	20%	1 +1	10%
30	10 -1	33%	13 +1	43%	6	20%	1	3%
60	20 -5	33%	21 +2	35%	15 +2	25%	4 +1	7%
90	37 -3	41%	25 -1	28%	18 +4	20%	10	11%
120	56 +3	47%	30 -3	25%	21 +1	18%	13 -1	11%
150	74 +6	49%	35 -2	23%	26 -1	17%	15 -3	10%

configuration:
`400-entropy-2-d10`

F_1 score: 0.50100 (50)
AUC(P): 0.68300 (1)
AUC(S): 0.51902 (11)

ranking function: **zscore** 0.27582 +0.03119

	N [598]		**P** [131]		**S** [208]		**P+S** [59]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	3 -2	30%	4	40%	1	10%	2 +2	20%
30	9 -2	30%	12 +1	40%	6 -1	20%	3 +2	10%
60	26 +2	43%	17 -4	28%	11 -1	18%	6 +3	10%
90	44 +5	49%	24 -2	27%	13 -3	14%	9	10%
120	57 +3	48%	30 -3	25%	20	17%	13	11%
150	72	48%	32 -3	21%	28 +3	19%	18	12%

configuration:
`200-entropy-2-d10`

F_1 score: 0.50050 (51)
AUC(P): 0.66170 (3)
AUC(S): 0.50763 (19)

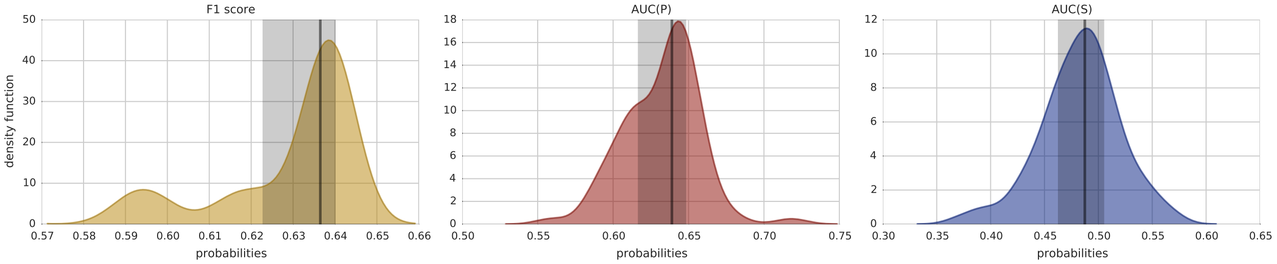
(b) split weights

eTable 14: Results of simulated recruitment for the **HOSTAS selection model (2 years shift)** configurations with best recruitment score (given at the top of each table). The differences in recruitment compared to the model configuration with highest F_1 score are highlighted with green (positive) and red (negative) colour. For each of the ranking functions the used model configuration and its prediction quality measured with F_1 score and AUC is given for reference (rank of the measure among all configurations is given in brackets).

F_1 score	configuration
0.6425	800-entropy-2-d7
0.6425	1000-entropy-2-d7
0.6425	600-entropy-2-d8
...	
0.592	800-entropy-2-d4
0.591	600-entropy-2-d4
0.588	100-entropy-2-d4

AUC(P)	configuration
0.71814	200-gini-2-d4
0.68033	400-gini-2-d5
0.6714	400-entropy-2-d6
...	
0.58772	200-gini-2-d8
0.58379	100-entropy-2-d6
0.55908	200-entropy-2-d8

AUC(S)	configuration
0.56884	100-entropy-2-d9
0.56165	1000-entropy-2-d4
0.54714	200-gini-2-d8
...	
0.39344	100-gini-2-d10
0.39319	400-gini-2-d10
0.37248	100-gini-2-d5



eFigure 14: Distribution of quality measures across all **HOSTAS selection model (3 years shift)** configurations. The grey area shows interquartile range and thick vertical line indicates the median value. Tables show 3 top/bottom scores and corresponding model parameters.

ranking function: **sum** 0.19560 +0.07704

	N [694]		P [115]		S [109]		P+S [24]	
size	abs	rel	abs	rel	abs	rel	abs	rel
10	3 -1	30%	6 +3	60%	1 -1	10%	0 -1	0%
30	15 -5	50%	12 +6	40%	1 -2	3%	2 +1	7%
60	28 -7	47%	22 +6	37%	6 -1	10%	4 +2	7%
90	45 -6	50%	27 +2	30%	12	13%	6 +4	7%
120	61 -11	51%	36 +5	30%	17 +4	14%	6 +2	5%
150	82 -6	55%	43 +4	29%	19 +2	13%	6	4%

configuration: `200-gini-2-d4`

F_1 score: 0.59650 (77)
AUC(P): 0.71814 (1)
AUC(S): 0.44851 (72)

ranking function: **scaled** 0.18626 +0.06672

	N [694]		P [115]		S [109]		P+S [24]	
size	abs	rel	abs	rel	abs	rel	abs	rel
10	2 -2	20%	7 +4	70%	1 -1	10%	0 -1	0%
30	13 -6	43%	13 +6	43%	3	10%	1	3%
60	33 -1	55%	18 +1	30%	6 -2	10%	3 +2	5%
90	46 -6	51%	28 +3	31%	12 +1	13%	4 +2	4%
120	65 -7	54%	34 +3	28%	16 +2	13%	5 +2	4%
150	83 -8	55%	43 +6	29%	18	12%	6 +2	4%

configuration: `400-gini-2-d4`

F_1 score: 0.59800 (75)
AUC(P): 0.64773 (22)
AUC(S): 0.50533 (21)

ranking function: **zscore** 0.16194 +0.07330

	N [694]		P [115]		S [109]		P+S [24]	
size	abs	rel	abs	rel	abs	rel	abs	rel
10	4 -1	40%	4 +2	40%	2	20%	0 -1	0%
30	14 -6	47%	11 +4	37%	4 +2	13%	1	3%
60	29 -8	48%	20 +6	33%	9 +1	15%	2 +1	3%
90	50 -2	56%	24	27%	13	14%	3 +2	3%
120	69 -6	58%	32 +5	27%	16	13%	3 +1	2%
150	92 -8	61%	37 +7	25%	18 +1	12%	3	2%

configuration: `600-gini-2-d4`

F_1 score: 0.59750 (76)
AUC(P): 0.64886 (21)
AUC(S): 0.51809 (14)

(a) progressive weights

ranking function: **sum** 0.18507 +0.01962

	N [694]		P [115]		S [109]		P+S [24]	
size	abs	rel	abs	rel	abs	rel	abs	rel
10	4	40%	3	30%	2	20%	1	10%
30	15 -5	50%	11 +5	37%	3	10%	1	3%
60	31 -4	52%	20 +4	33%	6 -1	10%	3 +1	5%
90	47 -4	52%	25	28%	13 +1	14%	5 +3	6%
120	68 -4	57%	33 +2	28%	14 +1	12%	5 +1	4%
150	85 -3	57%	41 +2	27%	19 +2	13%	5 -1	3%

configuration: `400-gini-2-d5`

F_1 score: 0.62300 (62)
AUC(P): 0.68033 (2)
AUC(S): 0.48519 (44)

ranking function: **scaled** 0.18534 +0.02295

	N [694]		P [115]		S [109]		P+S [24]	
size	abs	rel	abs	rel	abs	rel	abs	rel
10	5 +1	50%	3	30%	1 -1	10%	1	10%
30	18 -1	60%	6 -1	20%	3	10%	3 +2	10%
60	34	57%	14 -3	23%	9 +1	15%	3 +2	5%
90	52	58%	21 -4	23%	14 +3	16%	3 +1	3%
120	67 -5	56%	33 +2	28%	15 +1	12%	5 +2	4%
150	90 -1	60%	35 -2	23%	20 +2	13%	5 +1	3%

configuration: `100-gini-2-d7`

F_1 score: 0.64200 (4)
AUC(P): 0.62220 (55)
AUC(S): 0.51008 (18)

ranking function: **zscore** 0.17900 +0.02526

	N [694]		P [115]		S [109]		P+S [24]	
size	abs	rel	abs	rel	abs	rel	abs	rel
10	5	50%	3 +1	30%	1 -1	10%	1	10%
30	14 -6	47%	11 +4	37%	4 +2	13%	1	3%
60	30 -7	50%	19 +5	32%	9 +1	15%	2 +1	3%
90	47 -5	52%	27 +3	30%	13	14%	3 +2	3%
120	69 -6	58%	31 +4	26%	17 +1	14%	3 +1	2%
150	92 -8	61%	35 +5	23%	19 +2	13%	4 +1	3%

configuration: `600-entropy-2-d4`

F_1 score: 0.59100 (83)
AUC(P): 0.63906 (42)
AUC(S): 0.54160 (5)

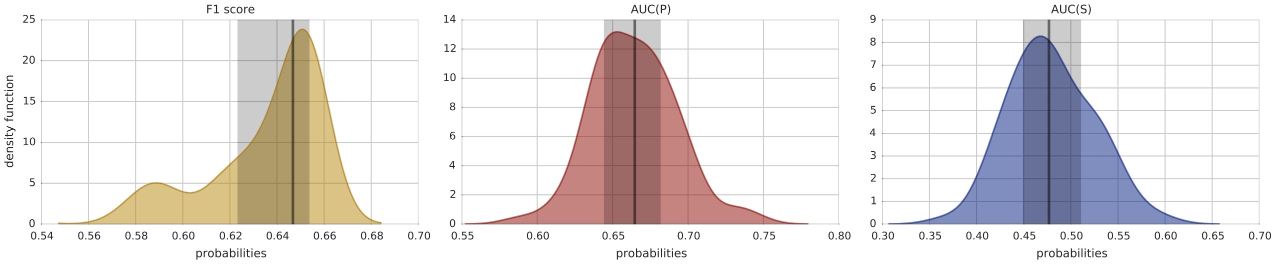
(b) split weights

eTable 15: Results of simulated recruitment for the **HOSTAS selection model (3 years shift)** configurations with best recruitment score (given at the top of each table). The differences in recruitment compared to the model configuration with highest F_1 score are highlighted with green (positive) and red (negative) colour. For each of the ranking functions the used model configuration and its prediction quality measured with F_1 score and AUC is given for reference (rank of the measure among all configurations is given in brackets).

$\mathbf{F_1}$ **score**	**configuration**
0.655	800-entropy-2-d9
0.655	600-entropy-2-d10
0.655	800-entropy-2-d10
...	
0.584	1000-entropy-2-d4
0.582	200-entropy-2-d4
0.5765	100-entropy-2-d4

AUC(P)	**configuration**
0.74372	200-entropy-2-d9
0.73141	1000-entropy-2-d5
0.73035	800-entropy-2-d7
...	
0.6127	800-gini-2-d6
0.60945	200-gini-2-d6
0.58808	100-entropy-2-d9

AUC(S)	**configuration**
0.59901	400-gini-2-d7
0.57552	600-gini-2-d8
0.56576	200-gini-2-d4
...	
0.40368	200-gini-2-d7
0.39942	800-entropy-2-d10
0.36552	100-entropy-2-d9



eFigure 15: Distribution of quality measures across all **HOSTAS selection model (5 years shift)** configurations. The grey area shows interquartile range and thick vertical line indicates the median value. Tables show 3 top/bottom scores and corresponding model parameters.

ranking function: **sum** 0.06818 +0.10810

	N [746]		**P** [114]		**S** [109]		**P+S** [24]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	5 -1	50%	4	40%	1 +1	10%	0	0%
30	17 -6	57%	10 +3	33%	3 +3	10%	0	0%
60	40 -5	67%	16 +3	27%	3 +1	5%	1 +1	2%
90	56 -14	62%	25 +9	28%	4 +1	4%	5 +4	6%
120	77 -13	64%	31 +10	26%	5 -1	4%	7 +4	6%
150	99 -7	66%	37 +10	25%	7 -5	5%	7 +2	5%

configuration:
`600-gini-2-d4`

$\mathbf{F_1}$ **score:** 0.59400 (76)
AUC(P): 0.69614 (10)
AUC(S): 0.54806 (5)

ranking function: **scaled** 0.04772 +0.09849

	N [746]		**P** [114]		**S** [109]		**P+S** [24]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	5 -1	50%	4	40%	1 +1	10%	0	0%
30	18 -4	60%	9 +2	30%	3 +2	10%	0	0%
60	39 -8	65%	16 +4	27%	3 +2	5%	2 +2	3%
90	60 -10	67%	23 +8	26%	4	4%	3 +2	3%
120	84 -9	70%	27 +10	22%	5 -3	4%	4 +2	3%
150	102 -15	68%	34 +13	23%	7 -2	5%	7 +4	5%

configuration:
`200-entropy-2-d4`

$\mathbf{F_1}$ **score:** 0.58200 (83)
AUC(P): 0.63957 (71)
AUC(S): 0.48456 (37)

ranking function: **zscore** -0.00040 +0.05108

	N [746]		**P** [114]		**S** [109]		**P+S** [24]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	6	60%	4	40%	0	0%	0	0%
30	21 -2	70%	8 +2	27%	1	3%	0	0%
60	42 -3	70%	12 -1	20%	5 +3	8%	1 +1	2%
90	64 -7	71%	18 +4	20%	6 +2	7%	2 +1	2%
120	86 -8	72%	23 +6	19%	8	7%	3 +2	2%
150	108 -8	72%	29 +7	19%	10 +1	7%	3	2%

configuration:
`200-entropy-2-d4`

$\mathbf{F_1}$ **score:** 0.58200 (83)
AUC(P): 0.63957 (71)
AUC(S): 0.48456 (37)

(a) progressive weights

ranking function: **sum** 0.12160 +0.05692

	N [746]		**P** [114]		**S** [109]		**P+S** [24]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	5 -1	50%	4	40%	1 +1	10%	0	0%
30	19 -4	63%	8 +1	27%	2 +2	7%	1 +1	3%
60	39 -6	65%	17 +4	28%	2	3%	2 +2	3%
90	56 -14	62%	23 +7	26%	6 +3	7%	5 +4	6%
120	78 -12	65%	28 +7	23%	8 +2	7%	6 +3	5%
150	100 -6	67%	35 +8	23%	9 -3	6%	6 +1	4%

configuration:
`100-entropy-2-d4`

$\mathbf{F_1}$ **score:** 0.57650 (84)
AUC(P): 0.62597 (79)
AUC(S): 0.45562 (59)

ranking function: **scaled** 0.10946 +0.04649

	N [746]		**P** [114]		**S** [109]		**P+S** [24]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	8 +2	80%	2 -2	20%	0	0%	0	0%
30	20 -2	67%	8 +1	27%	1	3%	1 +1	3%
60	42 -5	70%	12	20%	4 +3	7%	2 +2	3%
90	63 -7	70%	17 +2	19%	5 +1	6%	5 +4	6%
120	80 -13	67%	23 +6	19%	10 +2	8%	7 +5	6%
150	105 -12	70%	25 +4	17%	13 +4	9%	7 +4	5%

configuration:
`100-entropy-2-d6`

$\mathbf{F_1}$ **score:** 0.63100 (60)
AUC(P): 0.62474 (81)
AUC(S): 0.46465 (50)

ranking function: **zscore** 0.10307 +0.03935

	N [746]		**P** [114]		**S** [109]		**P+S** [24]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	8 +2	80%	2 -2	20%	0	0%	0	0%
30	22 -1	73%	6	20%	1	3%	1 +1	3%
60	42 -3	70%	11 -2	18%	5 +3	8%	2 +2	3%
90	63 -8	70%	16 +2	18%	8 +4	9%	3 +2	3%
120	85 -9	71%	22 +5	18%	10 +2	8%	3 +2	2%
150	106 -10	71%	26 +4	17%	12 +3	8%	6 +3	4%

configuration:
`100-entropy-2-d6`

$\mathbf{F_1}$ **score:** 0.63100 (60)
AUC(P): 0.62474 (81)
AUC(S): 0.46465 (50)

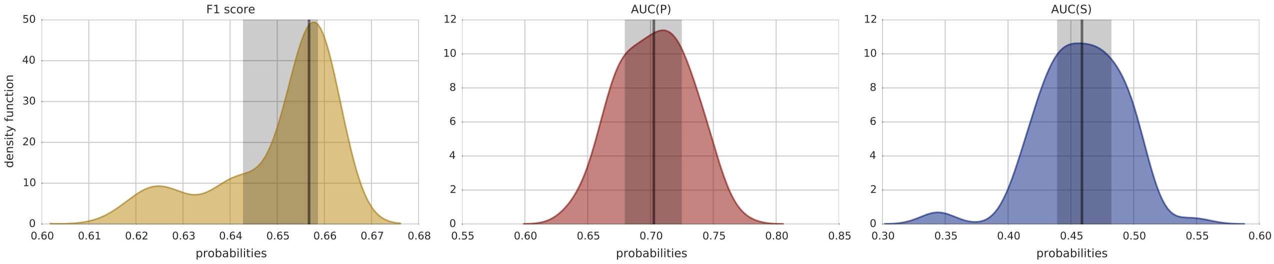
(b) split weights

eTable 16: Results of simulated recruitment for the **HOSTAS selection model (5 years shift)** configurations with best recruitment score (given at the top of each table). The differences in recruitment compared to the model configuration with highest F_1 score are highlighted with green (positive) and red (negative) colour. For each of the ranking functions the used model configuration and its prediction quality measured with F_1 score and AUC is given for reference (rank of the measure among all configurations is given in brackets).

F_1 score	configuration
0.661	600-entropy-2-d9
0.661	800-entropy-2-d9
0.661	800-entropy-2-d10
…	
0.621	600-entropy-2-d4
0.6205	200-entropy-2-d4
0.617	100-entropy-2-d4

AUC(P)	configuration
0.76807	1000-entropy-2-d8
0.75185	600-gini-2-d9
0.75038	600-entropy-2-d8
…	
0.65255	800-gini-2-d4
0.64235	600-gini-2-d4
0.63654	100-gini-2-d4

AUC(S)	configuration
0.54707	200-entropy-2-d4
0.5064	400-gini-2-d5
0.50515	600-gini-2-d5
…	
0.40769	100-gini-2-d10
0.34551	100-entropy-2-d9
0.3428	400-gini-2-d10



eFigure 16: Distribution of quality measures across all **MUST selection model (3 years shift)** configurations. The grey area shows interquartile range and thick vertical line indicates the median value. Tables show 3 top/bottom scores and corresponding model parameters.

ranking function: **sum** 0.21637 +0.04482

	N [685]		**P** [112]		**S** [109]		**P+S** [24]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	2 -2	20%	6 +1	60%	1	10%	1 +1	10%
30	14	47%	13 +1	43%	2	7%	1 -1	3%
60	27 -4	45%	22 +1	37%	7 +2	12%	4 +1	7%
90	42 -3	47%	35 +1	39%	9 +2	10%	4	4%
120	63 +1	52%	40 -1	33%	11	9%	6	5%
150	76 -7	51%	51 +4	34%	15 +1	10%	8 +2	5%

configuration:
200-gini-2-d9

F_1 score: 0.65800 (32)
AUC(P): 0.69630 (50)
AUC(S): 0.42949 (70)

ranking function: **scaled** 0.19582 +0.06123

	N [685]		**P** [112]		**S** [109]		**P+S** [24]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	4	40%	5	50%	1	10%	0	0%
30	11 -7	37%	15 +7	50%	4 +1	13%	0 -1	0%
60	27 -5	45%	24 +3	40%	7 +2	12%	2	3%
90	46 -3	51%	30 -2	33%	11 +4	12%	3 +1	3%
120	62 -2	52%	40	33%	12 -1	10%	6 +3	5%
150	78 -7	52%	49 +6	33%	15 -2	10%	8 +3	5%

configuration:
400-gini-2-d4

F_1 score: 0.62750 (76)
AUC(P): 0.68317 (60)
AUC(S): 0.46928 (34)

ranking function: **zscore** 0.19267 +0.08769

	N [685]		**P** [112]		**S** [109]		**P+S** [24]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	4 -2	40%	3	30%	3 +2	30%	0	0%
30	11 -7	37%	12 +2	40%	6 +4	20%	1 +1	3%
60	29 -6	48%	21 +3	35%	8 +3	13%	2	3%
90	48 +1	53%	29 -1	32%	9 -2	10%	4 +2	4%
120	63 -4	52%	37	31%	13 -1	11%	7 +5	6%
150	88 -2	59%	40 +1	27%	15 -2	10%	7 +3	5%

configuration:
100-gini-2-d4

F_1 score: 0.62300 (79)
AUC(P): 0.63654 (84)
AUC(S): 0.48970 (16)

(a) progressive weights

ranking function: **sum** 0.19522 +0.03691

	N [685]		**P** [112]		**S** [109]		**P+S** [24]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	3 -1	30%	4 -1	40%	1	10%	2 +2	20%
30	13 -1	43%	11 -1	37%	3 +1	10%	3 +1	10%
60	29 -2	48%	21	35%	6 +1	10%	4 +1	7%
90	46 +1	51%	32 -2	36%	7	8%	5 +1	6%
120	63 +1	52%	43 +2	36%	8 -3	7%	6	5%
150	78 -5	52%	49 +2	33%	16 +2	11%	7 +1	5%

configuration:
200-entropy-2-d10

F_1 score: 0.65950 (14)
AUC(P): 0.73027 (14)
AUC(S): 0.43250 (66)

ranking function: **scaled** 0.18013 +0.03788

	N [685]		**P** [112]		**S** [109]		**P+S** [24]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	5 +1	50%	3 -2	30%	2 +1	20%	0	0%
30	12 -6	40%	13 +5	43%	4 +1	13%	1	3%
60	27 -5	45%	24 +3	40%	7 +2	12%	2	3%
90	44 -5	49%	30 -2	33%	11 +4	12%	5 +3	6%
120	62 -2	52%	40	33%	12 -1	10%	6 +3	5%
150	78 -7	52%	50 +7	33%	14 -3	9%	8 +3	5%

configuration:
100-gini-2-d4

F_1 score: 0.62300 (79)
AUC(P): 0.63654 (84)
AUC(S): 0.48970 (16)

ranking function: **zscore** 0.19044 +0.05669

	N [685]		**P** [112]		**S** [109]		**P+S** [24]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	4 -2	40%	3	30%	3 +2	30%	0	0%
30	11 -7	37%	12 +2	40%	6 +4	20%	1 +1	3%
60	29 -6	48%	21 +3	35%	8 +3	13%	2	3%
90	48 +1	53%	29 -1	32%	9 -2	10%	4 +2	4%
120	63 -4	52%	37	31%	13 -1	11%	7 +5	6%
150	88 -2	59%	40 +1	27%	15 -2	10%	7 +3	5%

configuration:
100-gini-2-d4

F_1 score: 0.62300 (79)
AUC(P): 0.63654 (84)
AUC(S): 0.48970 (16)

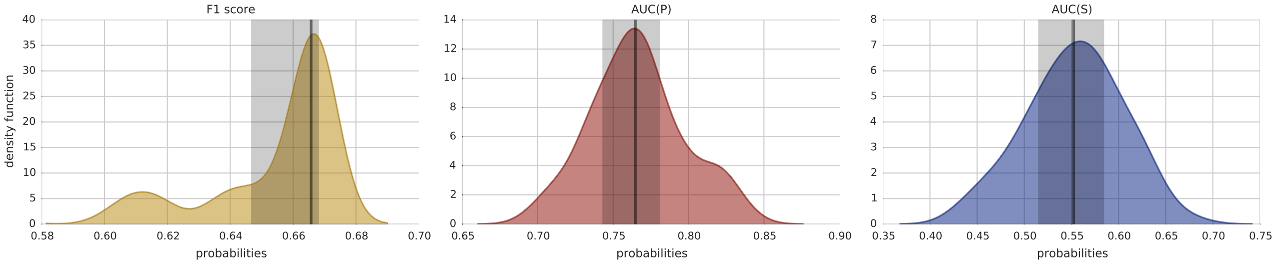
(b) split weights

eTable 17: Results of simulated recruitment for the **MUST selection model (3 years shift)** configurations with best recruitment score (given at the top of each table). The differences in recruitment compared to the model configuration with highest F_1 score are highlighted with green (positive) and red (negative) colour. For each of the ranking functions the used model configuration and its prediction quality measured with F_1 score and AUC is given for reference (rank of the measure among all configurations is given in brackets).

F_1 score	**configuration**
0.6695	1000-entropy-2-d9
0.6695	100-gini-2-d9
0.6695	600-gini-2-d8
...	
0.6075	400-entropy-2-d4
0.606	200-entropy-2-d4
0.602	100-entropy-2-d4

AUC(P)	**configuration**
0.83968	200-entropy-2-d7
0.82861	200-gini-2-d10
0.82806	200-gini-2-d8
...	
0.70792	100-gini-2-d6
0.70761	400-entropy-2-d4
0.69667	100-gini-2-d7

AUC(S)	**configuration**
0.67518	100-entropy-2-d7
0.64823	1000-entropy-2-d5
0.63406	400-entropy-2-d9
...	
0.45221	800-gini-2-d7
0.43645	200-gini-2-d10
0.43481	100-gini-2-d4



eFigure 17: Distribution of quality measures across all **MUST selection model (5 years shift)** configurations. The grey area shows interquartile range and thick vertical line indicates the median value. Tables show 3 top/bottom scores and corresponding model parameters.

ranking function: **sum** 0.16384 +0.05031

	N [733]		**P** [113]		**S** [108]		**P+S** [24]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	1	10%	9 +1	90%	0 -1	0%	0	0%
30	13 -3	43%	16 +3	53%	0 -1	0%	1 +1	3%
60	30 -6	50%	23 +3	38%	4 +2	7%	3 +1	5%
90	54 -5	60%	26 +2	29%	5 +1	6%	5 +2	6%
120	72 -3	60%	33 +4	28%	8 -1	7%	7	6%
150	90 -5	60%	38 +1	25%	13 +3	9%	9 +1	6%

configuration: `100-gini-2-d8`

F_1 score: 0.66850 (15)
AUC(P): 0.80261 (11)
AUC(S): 0.51358 (65)

ranking function: **scaled** 0.11148 +0.06709

	N [733]		**P** [113]		**S** [108]		**P+S** [24]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	2 -1	20%	8 +1	80%	0	0%	0	0%
30	16 -1	53%	13 +2	43%	0 -2	0%	1 +1	3%
60	35 -6	58%	18 +2	30%	5 +2	8%	2 +2	3%
90	56 -6	62%	22 +2	24%	9 +3	10%	3 +1	3%
120	75 -12	62%	31 +8	26%	11 +5	9%	3 -1	2%
150	101 -9	67%	33 +9	22%	12 +2	8%	4 -2	3%

configuration: `100-gini-2-d8`

F_1 score: 0.66850 (15)
AUC(P): 0.80261 (11)
AUC(S): 0.51358 (65)

ranking function: **zscore** 0.06883 +0.05417

	N [733]		**P** [113]		**S** [108]		**P+S** [24]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	3 -1	30%	7 +1	70%	0	0%	0	0%
30	19	63%	11 +2	37%	0 -2	0%	0	0%
60	38 -4	63%	15	25%	6 +3	10%	1 +1	2%
90	58 -8	64%	21 +3	23%	9 +4	10%	2 +1	2%
120	78 -11	65%	27 +8	22%	11 +3	9%	4	3%
150	105 -4	70%	30 +5	20%	11	7%	4 -1	3%

configuration: `100-gini-2-d8`

F_1 score: 0.66850 (15)
AUC(P): 0.80261 (11)
AUC(S): 0.51358 (65)

(a) progressive weights

ranking function: **sum** 0.14543 +0.02477

	N [733]		**P** [113]		**S** [108]		**P+S** [24]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	3 +2	30%	7 -1	70%	0 -1	0%	0	0%
30	13 -3	43%	14 +1	47%	1	3%	2 +2	7%
60	33 -3	55%	19 -1	32%	4 +2	7%	4 +2	7%
90	55 -4	61%	22 -2	24%	8 +4	9%	5 +2	6%
120	76 +1	63%	29	24%	9	8%	6 -1	5%
150	96 +1	64%	35 -2	23%	11 +1	7%	8	5%

configuration: `100-gini-2-d9`

F_1 score: 0.66950 (2)
AUC(P): 0.77615 (27)
AUC(S): 0.54445 (49)

ranking function: **scaled** 0.12381 +0.02870

	N [733]		**P** [113]		**S** [108]		**P+S** [24]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	2 -1	20%	8 +1	80%	0	0%	0	0%
30	16 -1	53%	13 +2	43%	0 -2	0%	1 +1	3%
60	35 -6	58%	18 +2	30%	5 +2	8%	2 +2	3%
90	56 -6	62%	22 +2	24%	9 +3	10%	3 +1	3%
120	75 -12	62%	31 +8	26%	11 +5	9%	3 -1	2%
150	101 -9	67%	33 +9	22%	12 +2	8%	4 -2	3%

configuration: `100-gini-2-d8`

F_1 score: 0.66850 (15)
AUC(P): 0.80261 (11)
AUC(S): 0.51358 (65)

ranking function: **zscore** 0.12087 +0.03332

	N [733]		**P** [113]		**S** [108]		**P+S** [24]	
size	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**	**abs**	**rel**
10	5 +1	50%	4 -2	40%	1 +1	10%	0	0%
30	15 -4	50%	12 +3	40%	3 +1	10%	0	0%
60	39 -3	65%	15	25%	5 +2	8%	1 +1	2%
90	62 -4	69%	17 -1	19%	9 +4	10%	2 +1	2%
120	83 -6	69%	22 +3	18%	13 +5	11%	2 -2	2%
150	105 -4	70%	26 +1	17%	16 +5	11%	3 -2	2%

configuration: `100-entropy-2-d5`

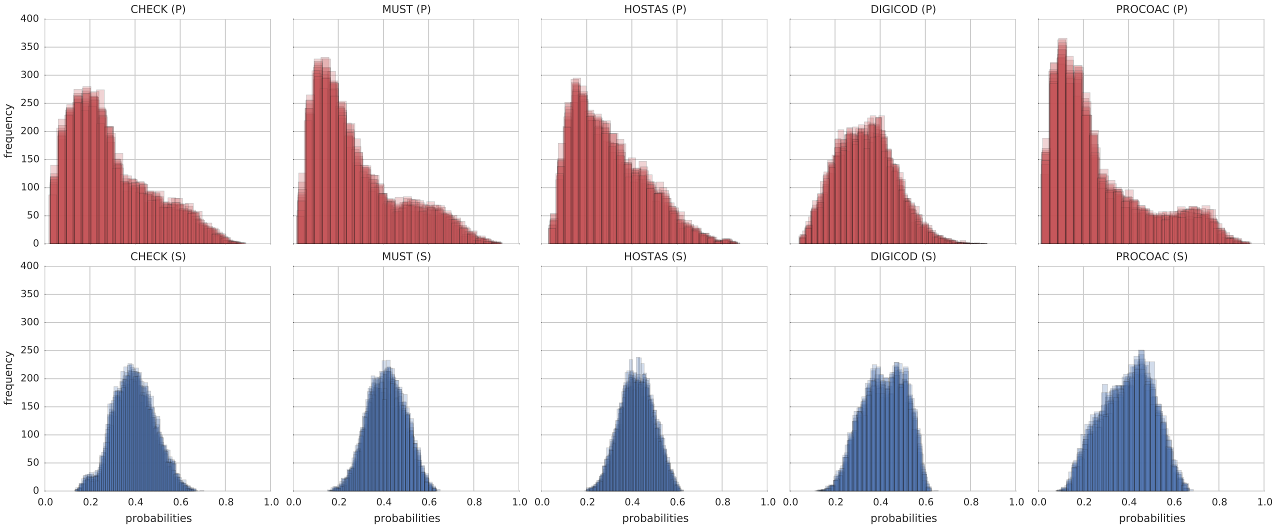
F_1 score: 0.63300 (72)
AUC(P): 0.74166 (65)
AUC(S): 0.58055 (25)

(b) split weights

eTable 18: Results of simulated recruitment for the **MUST selection model (5 years shift)** configurations with best recruitment score (given at the top of each table). The differences in recruitment compared to the model configuration with highest F_1 score are highlighted with green (positive) and red (negative) colour. For each of the ranking functions the used model configuration and its prediction quality measured with F_1 score and AUC is given for reference (rank of the measure among all configurations is given in brackets).

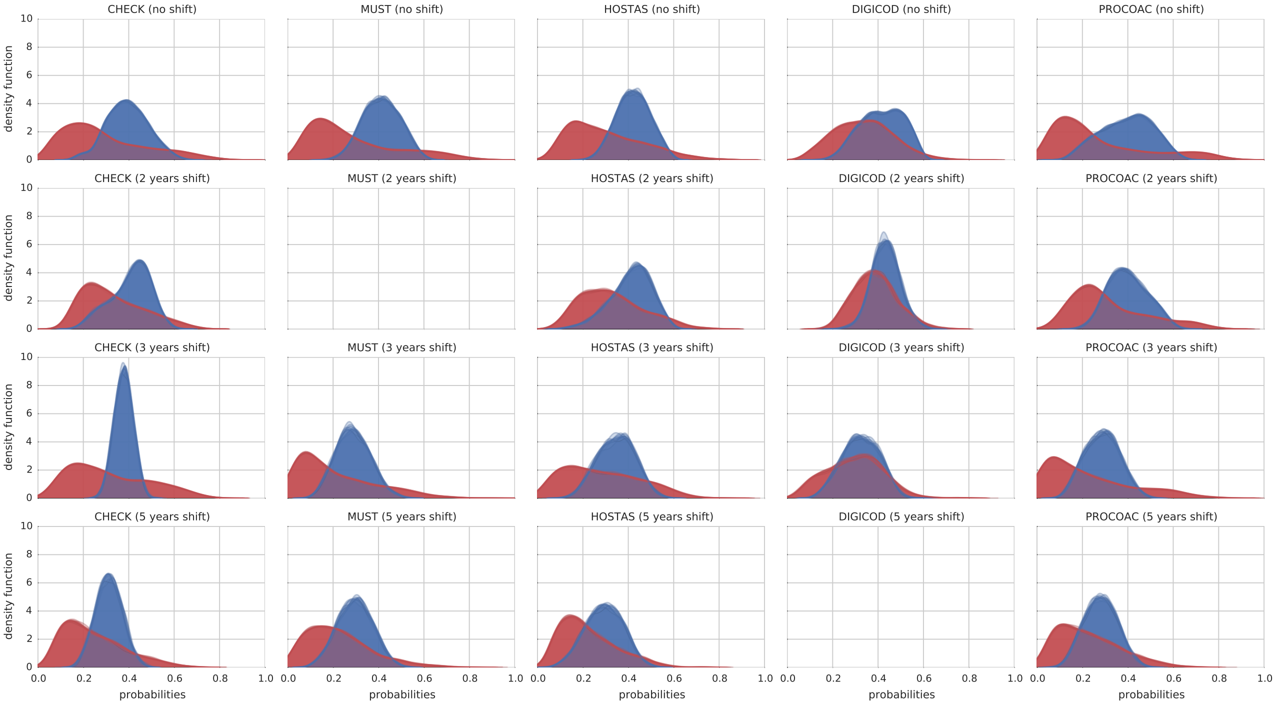
2.3 Probabilities and selection models confidence

eFigure 18 compares the distribution of probabilities returned by the selection models when no time shift is used. For CHECK, MUST and HOSTAS these distributions were relatively similar for both **P** and **S** labels, but differed more for DIGICOD (mostly for P) and PROCOAC (mostly for S). These differences are a result of reduced set of attributes used by the models, limited to what was possible to map and harmonise to CHECK. However, the differences are not large enough to invalidate our approach, and mainly affect the probabilities related to pain.



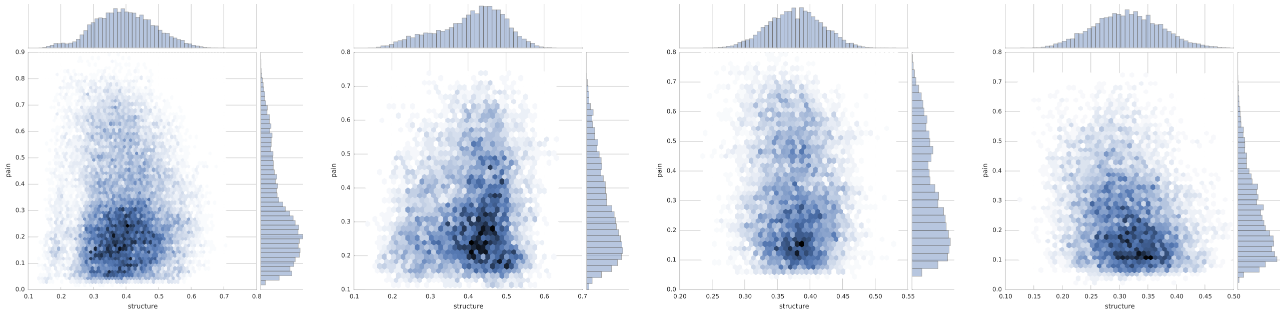
eFigure 18: Distribution of model selection probabilities for all cohorts. In each panel, the histograms for all CV-repeats are superimposed on top of each other.

eFigure 19 illustrates the effect of the time shift. The selection model probabilities tends to get more narrow (especially for **S** label) with increasing shift. The maximum probability decreases and the entire distributions lean more towards 0, which represent reduced confidence the models have in the prediction made for increasing shift.



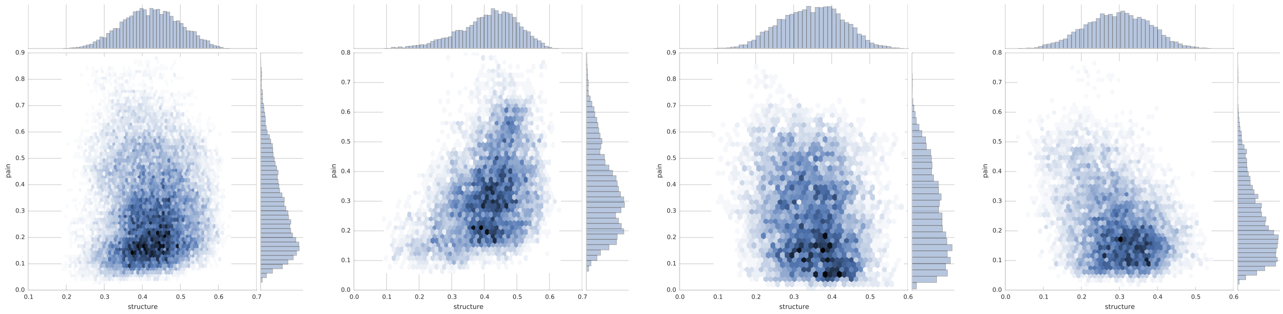
eFigure 19: Change in probability distributions with increasing time shift. The **P** probablity is plot in red, and **S** in blue. In each panel, the kernel density functions for all CV-repeats are superimposed on top of each other. Empty panel is shown when a particular shift was not applicable.

eFigure 20–24 show the joint distribution of probabilities returned by the selection models for the training data (harmonised CHECK cohort) for increasing shift.



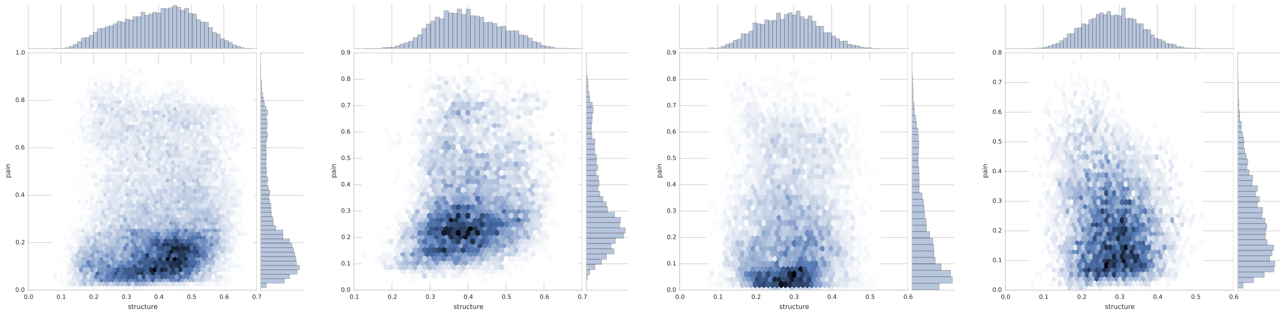
(a) no shift **(b)** 2 years shift **(c)** 3 years shift **(d)** 5 years shift

eFigure 20: Joint probability distribution of **CHECK** models confidence in predicting **P** and **S** labels on training set for increasing shift.



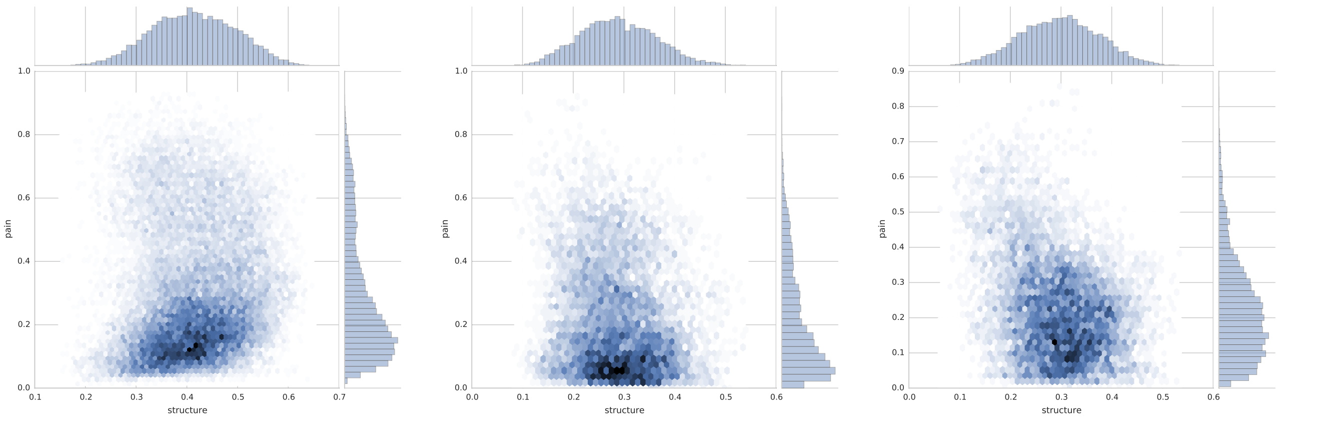
(a) no shift **(b)** 2 years shift **(c)** 3 years shift **(d)** 5 years shift

eFigure 21: Joint probability distribution of **HOSTAS** models confidence in predicting **P** and **S** labels on training set for increasing shift.



(a) no shift **(b)** 2 years shift **(c)** 3 years shift **(d)** 5 years shift

eFigure 22: Joint probability distribution of **PROCOAC** models confidence in predicting **P** and **S** labels on training set for increasing shift.

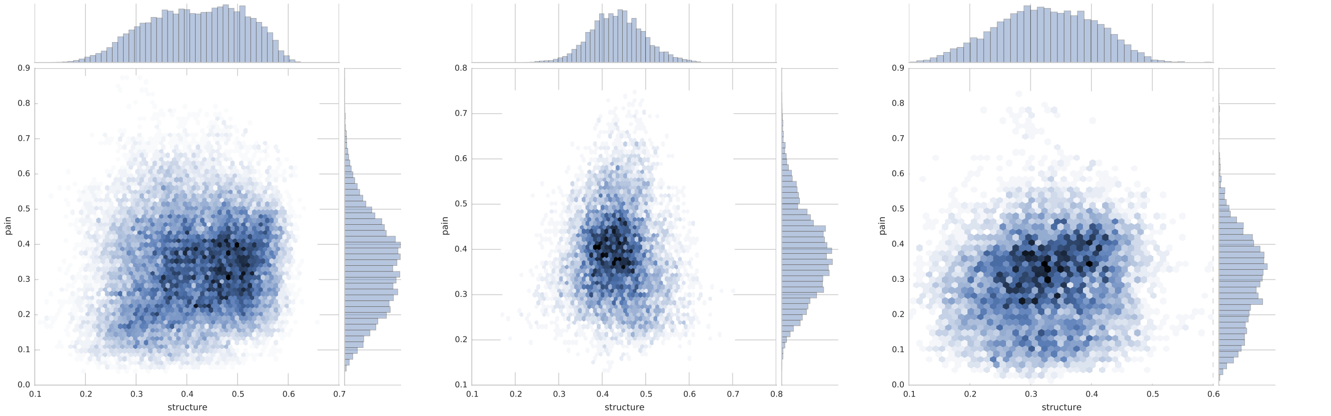


(a) no shift

(b) 3 years shift

(c) 5 years shift

eFigure 23: Joint probability distribution of **MUST** models confidence in predicting **P** and **S** labels on training set for increasing shift.



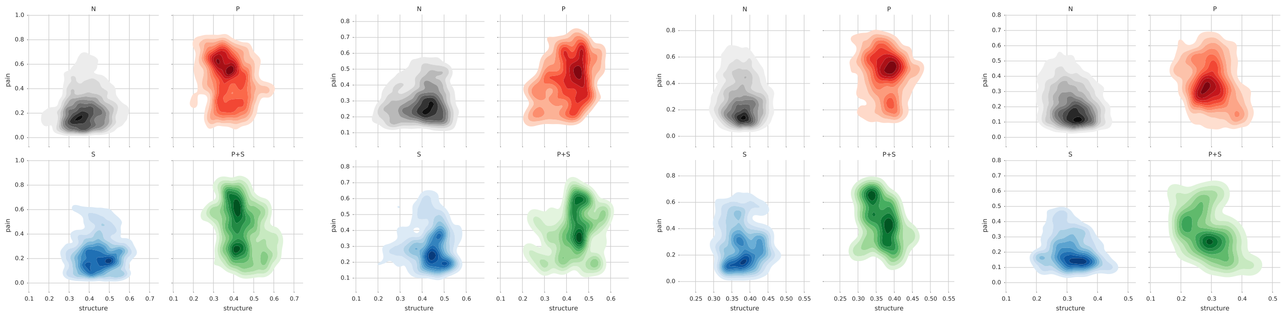
(a) no shift

(b) 2 years shift

(c) 3 years shift

eFigure 24: Joint probability distribution of **DIGICOD** models confidence in predicting **P** and **S** labels on training set for increasing shift.

The following figures show the joint distribution divided into categories. The models confidence in predictions for non-progressive periods (**N**) remains unchanged with increasing shift, but for others, the distribution peaks slowly move towards lower valued regions.



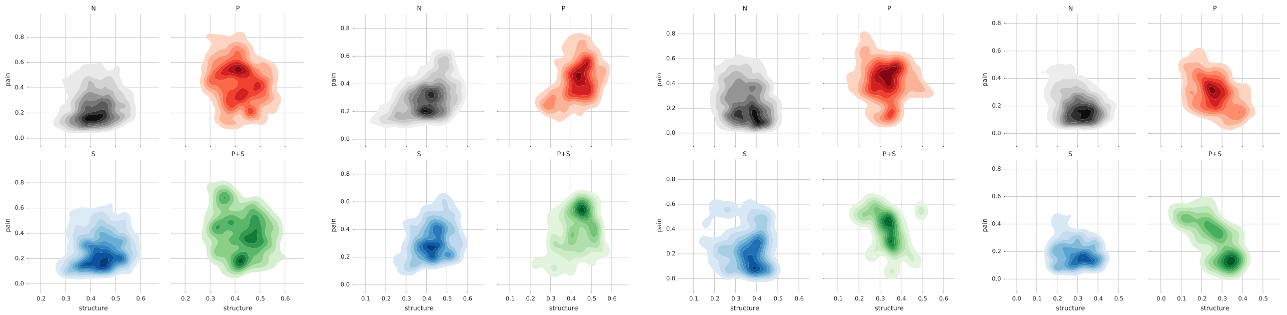
(a) no shift

(b) 2 years shift

(c) 3 years shift

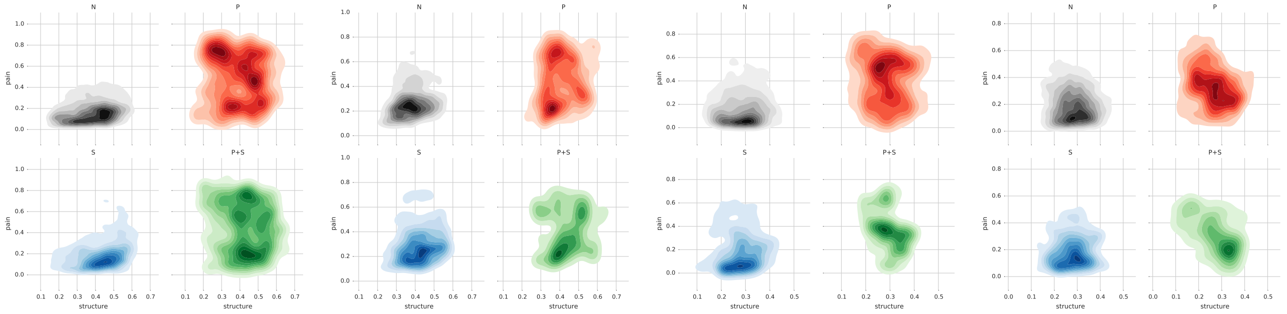
(d) 5 years shift

eFigure 25: Joint probability distribution of **CHECK** models confidence in predicting **P** and **S** labels on training set for increasing shift divided by the period category.



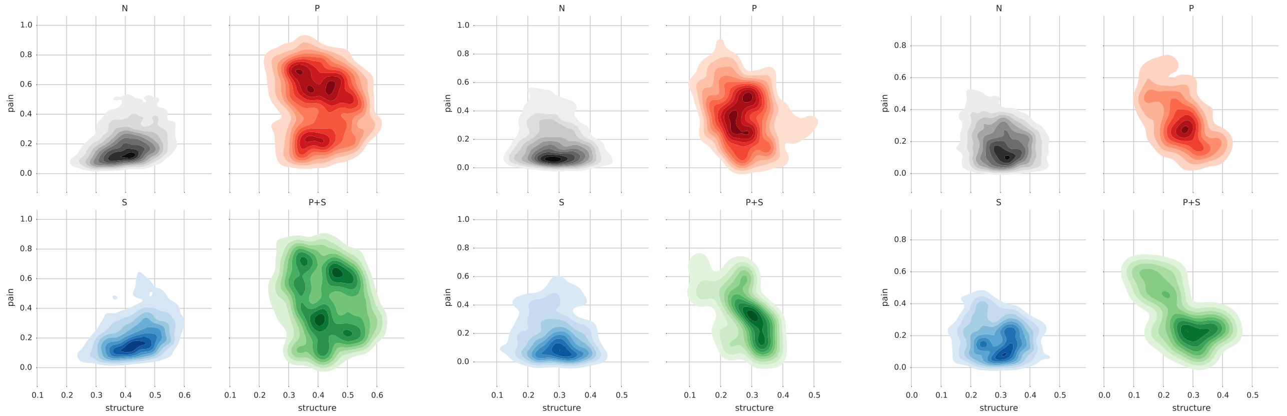
(a) no shift **(b)** 2 years shift **(c)** 3 years shift **(d)** 5 years shift

eFigure 26: Joint probability distribution of **HOSTAS** models confidence in predicting **P** and **S** labels on training set for increasing shift divided by the period category.



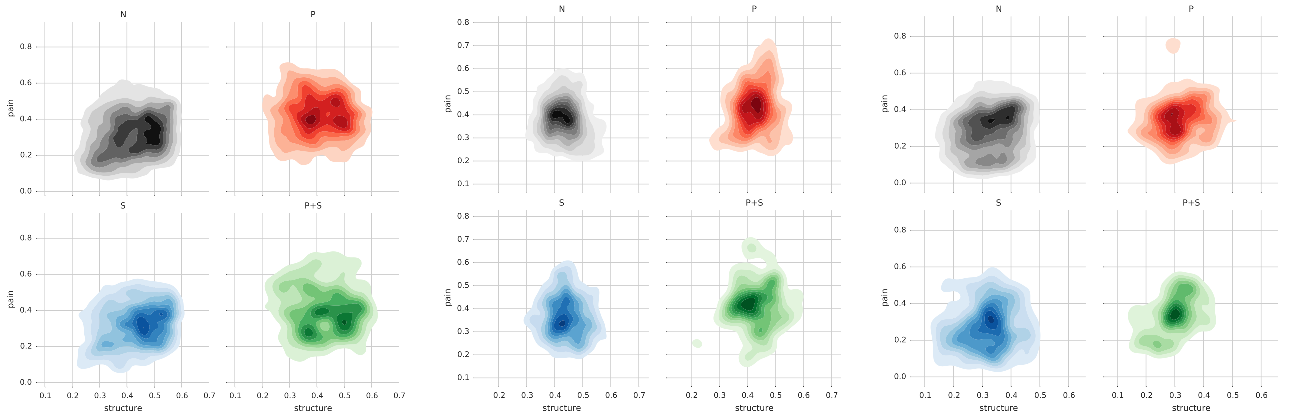
(a) no shift **(b)** 2 years shift **(c)** 3 years shift **(d)** 5 years shift

eFigure 27: Joint probability distribution of **PROCOAC** models confidence in predicting **P** and **S** labels on training set for increasing shift divided by the period category.



(a) no shift **(b)** 3 years shift **(c)** 5 years shift

eFigure 28: Joint probability distribution of **MUST** models confidence in predicting **P** and **S** labels on training set for increasing shift divided by the period category.



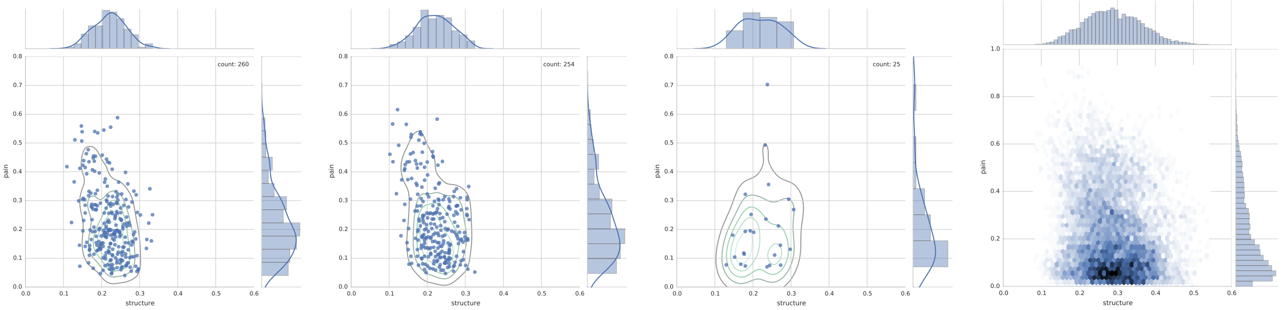
(a) no shift

(b) 2 years shift

(c) 3 years shift

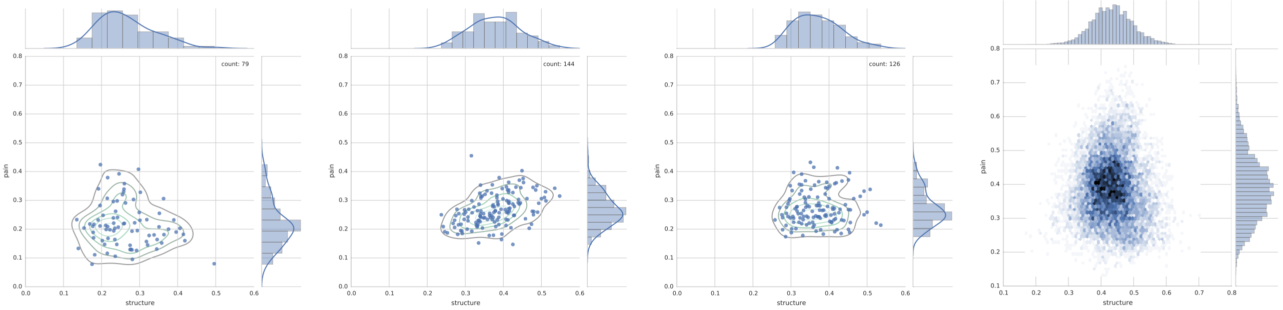
eFigure 29: Joint probability distribution of **DIGICOD** models confidence in predicting **P** and **S** labels on training set for increasing shift divided by the period category.

The following figures show the influence of time shift on probabilities returned by the selection models for the harmonised cohort data (real patients data, rather than harmonised CHECK periods). With 2–3 year shift, the distributions remain roughly equivalent to what was seen in the training set, but for visits more than 4 years prior, the distributions start to differ substantially. This indicates that in such cases, the prediction of the selection models might not be reliable.



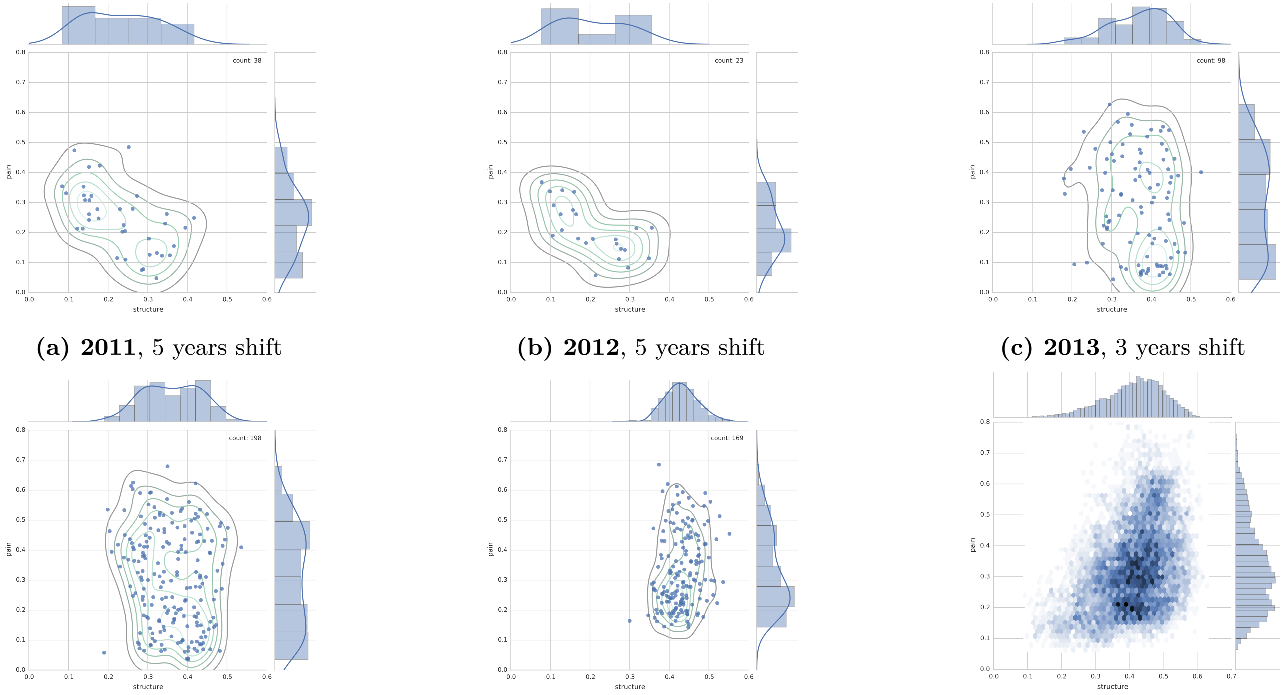
(a) **2011**, 5 years shift **(b)** **2012**, 5 years shift **(c)** **2013**, 3 years shift **(d)** training set, 3 years shift

eFigure 30: Joint probability distribution of **MUST** selection models applied to harmonised patient data. Each panel shows results of a model with a specific shift, applied to data from a specific visit year, except the last panel, where for comparison the distribution on the training data (from eFigure 23) is shown.



(a) **2014**, 3 years shift **(b)** **2015**, 2 years shift **(c)** **2016**, 2 years shift **(d)** training set, 2 years shift

eFigure 31: Joint probability distribution of **DIGICOD** selection models applied to harmonised patient data. Each panel shows results of a model with a specific shift, applied to data from a specific visit year, except the last panel, where for comparison the distribution on the training data (from eFigure 24) is shown.



(a) **2011**, 5 years shift

(b) **2012**, 5 years shift

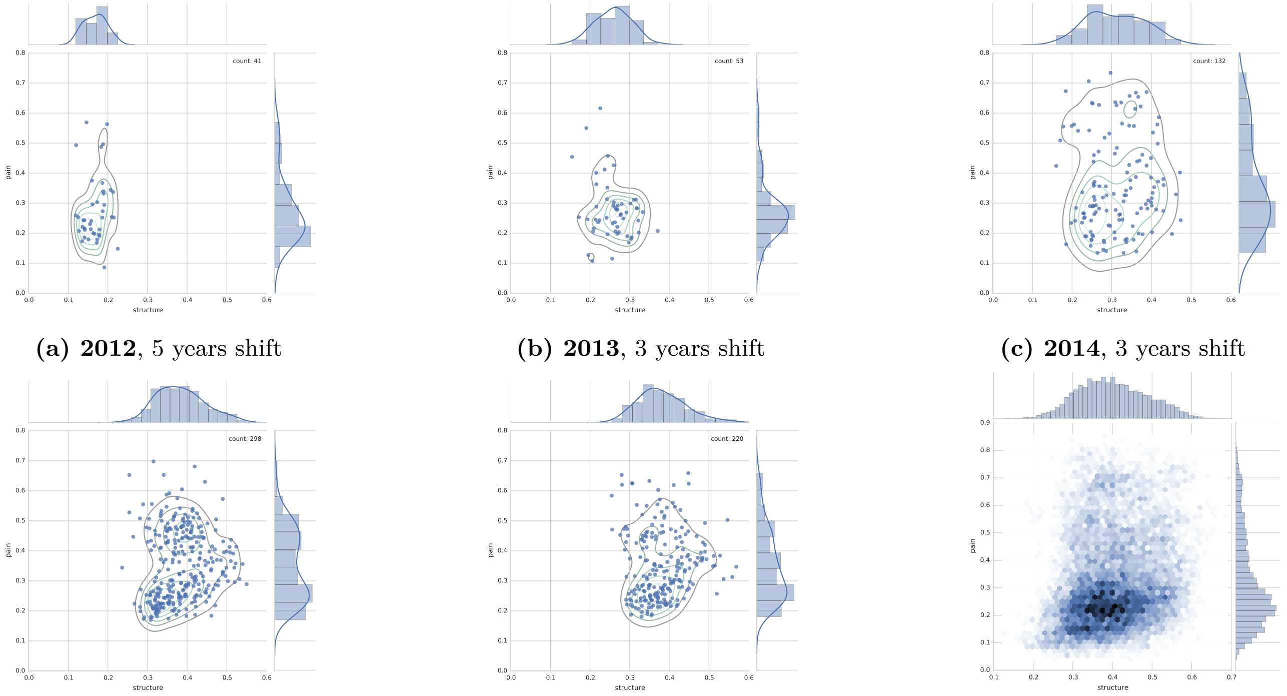
(c) **2013**, 3 years shift

(d) **2014**, 3 years shift

(e) **2015**, 2 years shift

(f) training set, 2 years shift

eFigure 32: Joint probability distribution of **HOSTAS** selection models applied to harmonised patient data. Each panel shows results of a model with a specific shift, applied to data from a specific visit year, except the last panel, where for comparison the distribution on the training data (from eFigure 21) is shown.



(a) **2012**, 5 years shift

(b) **2013**, 3 years shift

(c) **2014**, 3 years shift

(d) **2015**, 2 years shift

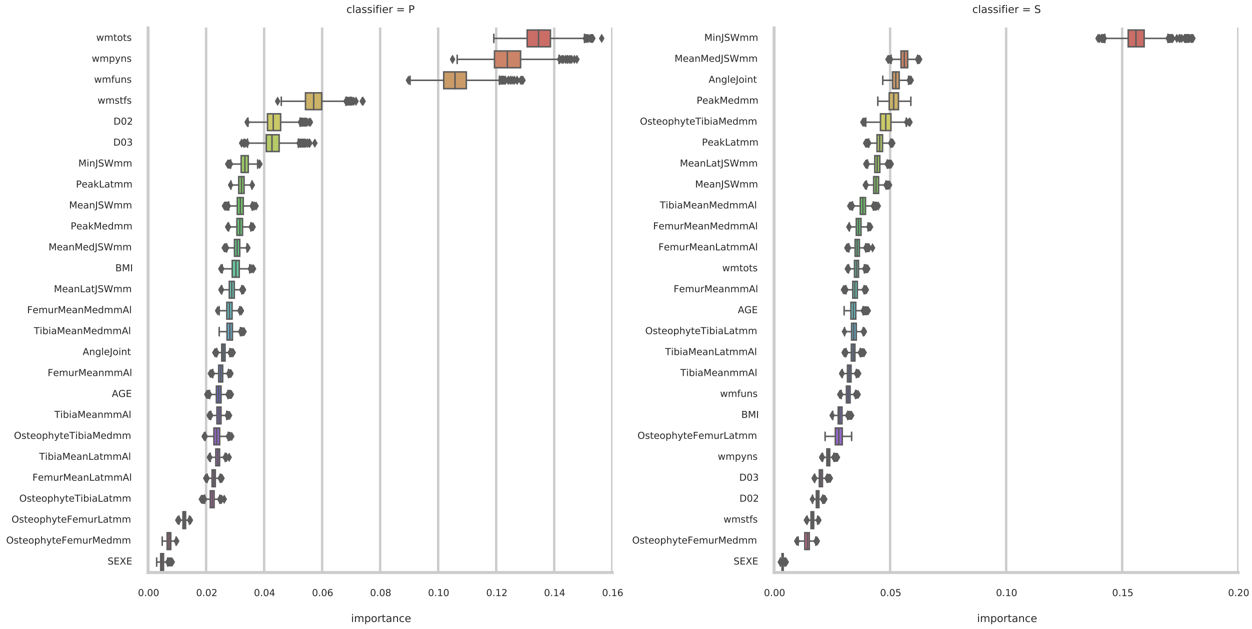
(e) **2016**, 2 years shift

(f) training set, 2 years shift

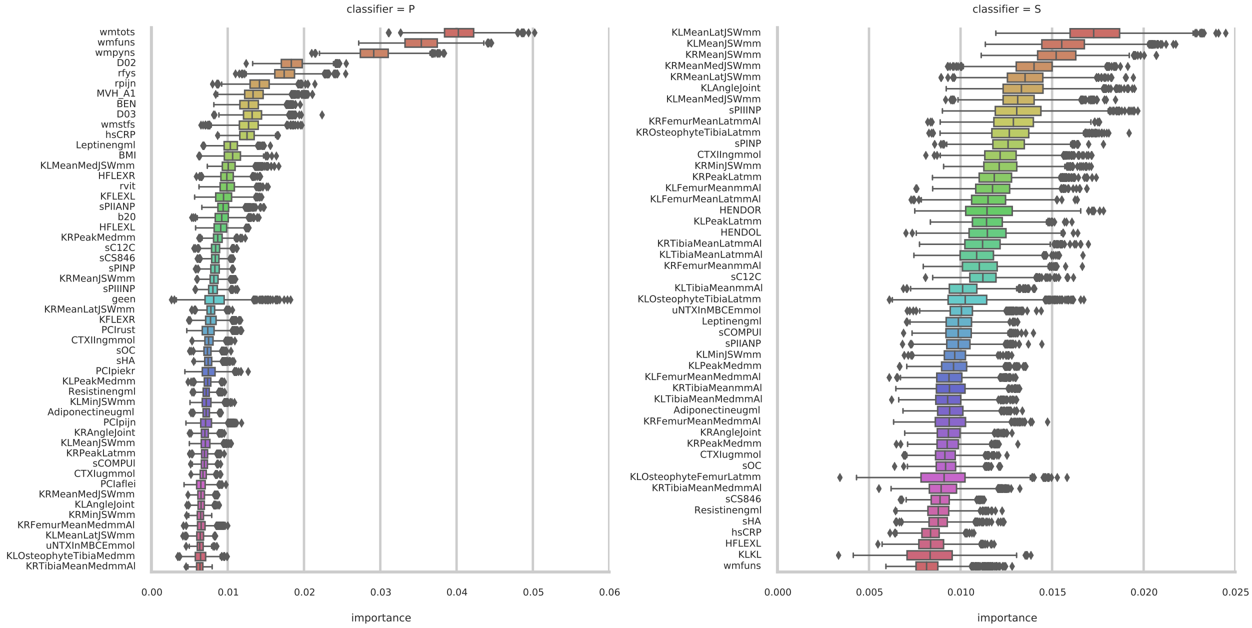
eFigure 33: Joint probability distribution of **PROCOAC** selection models applied to harmonised patient data. Each panel shows results of a model with a specific shift, applied to data from a specific visit year, except the last panel, where for comparison the distribution on the training data (from eFigure 22) is shown.

2.4 Model interpretation — attribute importance

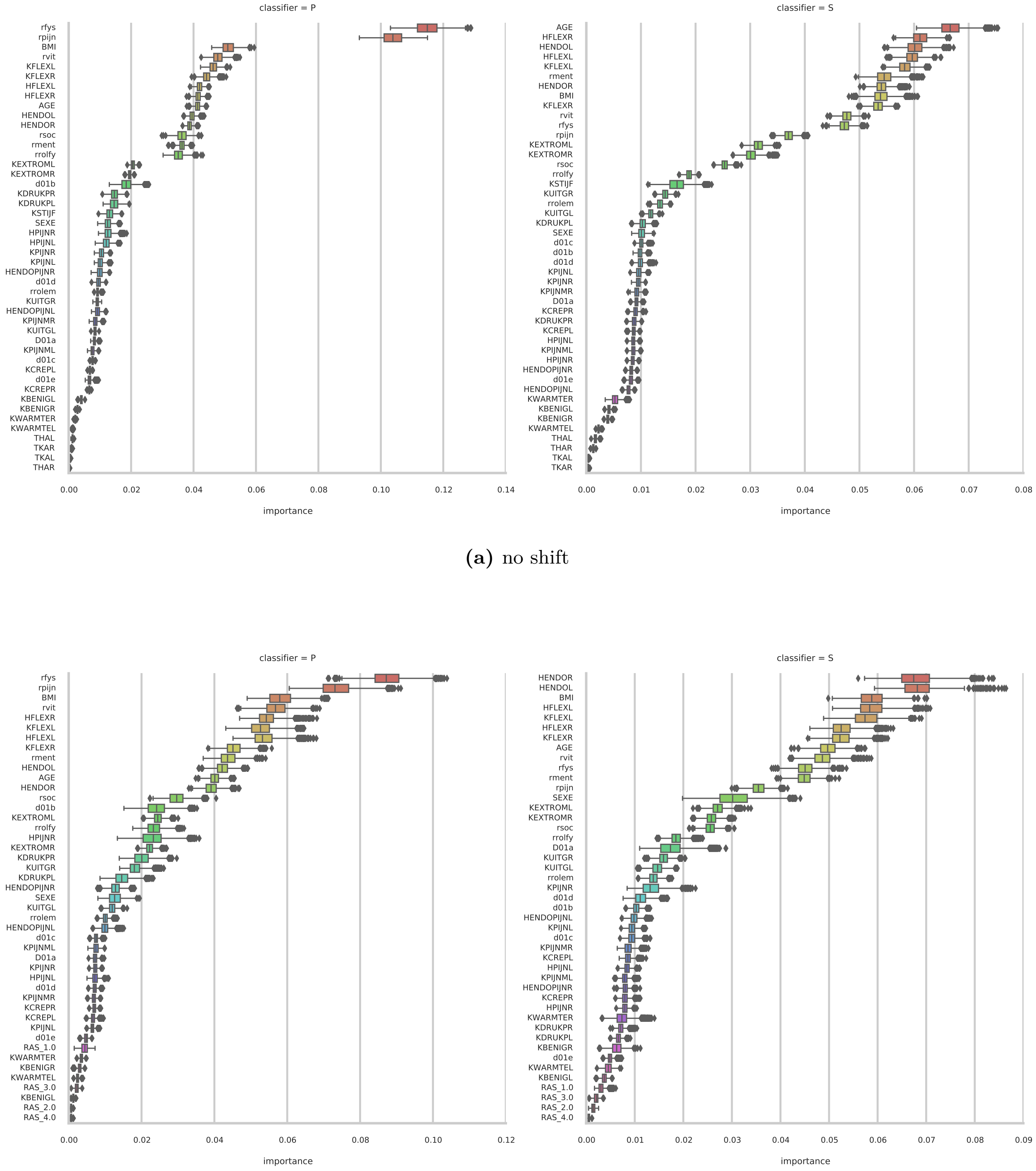
The following figures show the relative importance of the top 50 attributes for all models used in the selection process (the screening model and all the cohort-specific models). The more important an attribute is, the more impact it has on the model output. Two outputs are analysed independently: a probability of **pain-related progression** and a probability of **structure-related progression**. These correspond to the outputs of the two sub-models used by the *duo classifier*.



eFigure 34: Relative importance of top attributes used by the **screening model**. The two panels show the importance as impact on the probability of progression returned by different sub-models: P (pain-related) on the left, and S (structure-related) on the right. Attributes are listed in order of importance (descending).



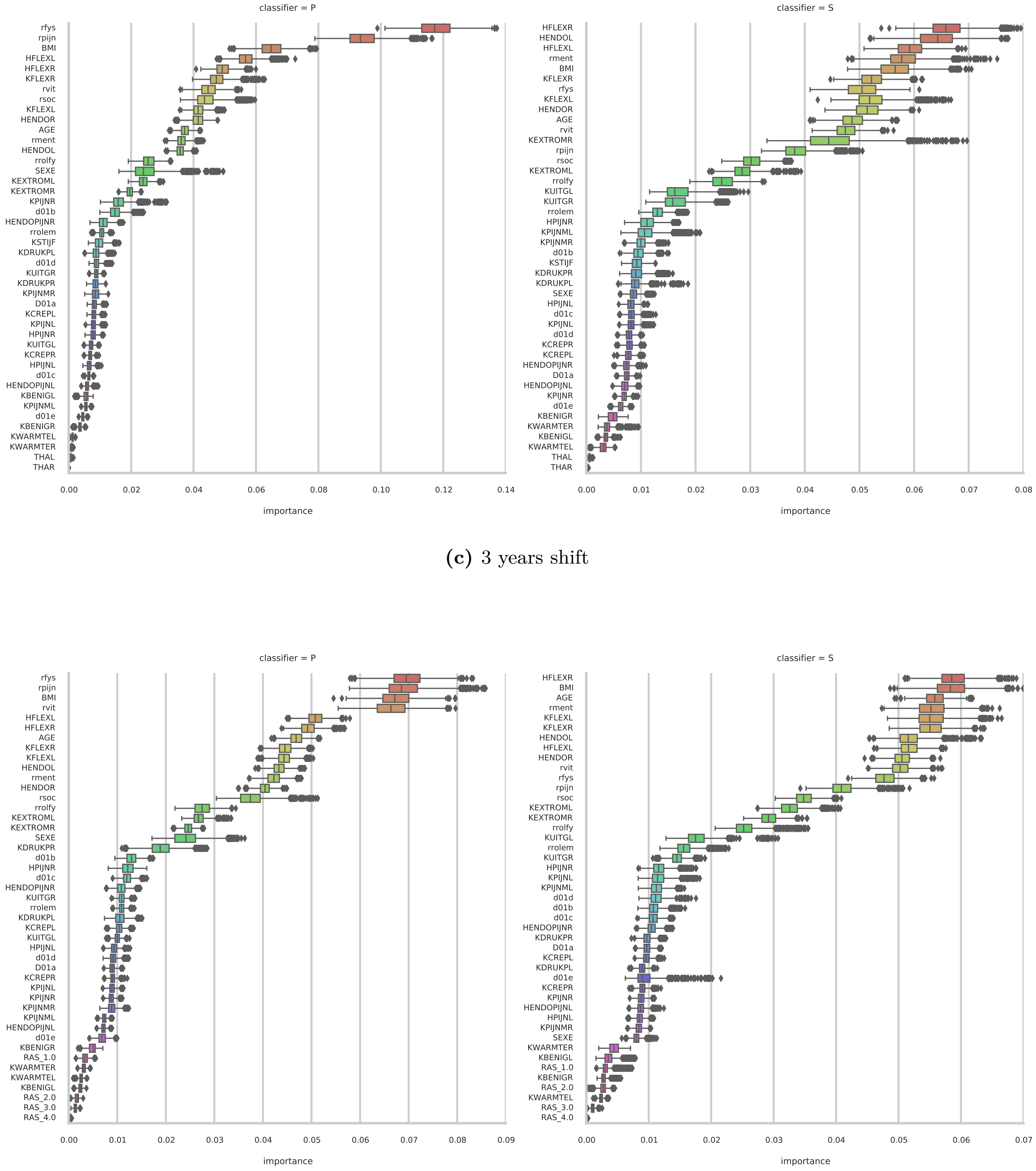
eFigure 35: Relative importance of top attributes used by the **CHECK selection model** (2 years shift). The two panels show the importance as impact on the probability of progression returned by different sub-models: P (pain-related) on the left, and S (structure-related) on the right. Attributes are listed in order of importance (descending).



(a) no shift

(b) 2 years shift

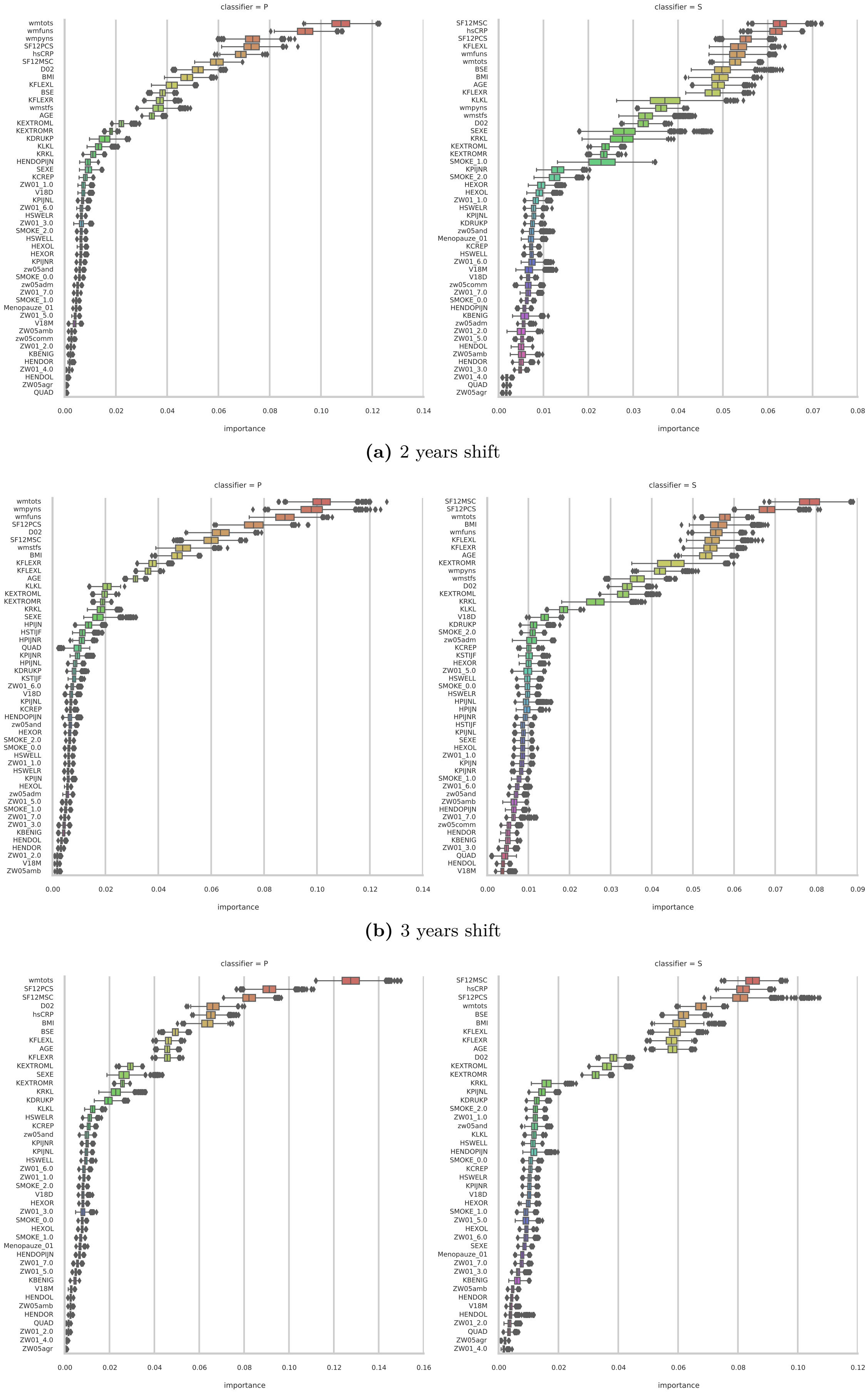
eFigure 36: Relative importance of top attributes used by the **HOSTAS selection models**. The two panels show the importance as impact on the probability of progression returned by different sub-models: pain-related (P) on the left, and structure-related (S) on the right. Attributes are listed in order of importance (descending).



(c) 3 years shift

(d) 5 years shift

eFigure 36: Relative importance of top attributes used by the **HOSTAS selection models**. The two panels show the importance as impact on the probability of progression returned by different sub-models: pain-related (P) on the left, and structure-related (S) on the right. Attributes are listed in order of importance (descending).

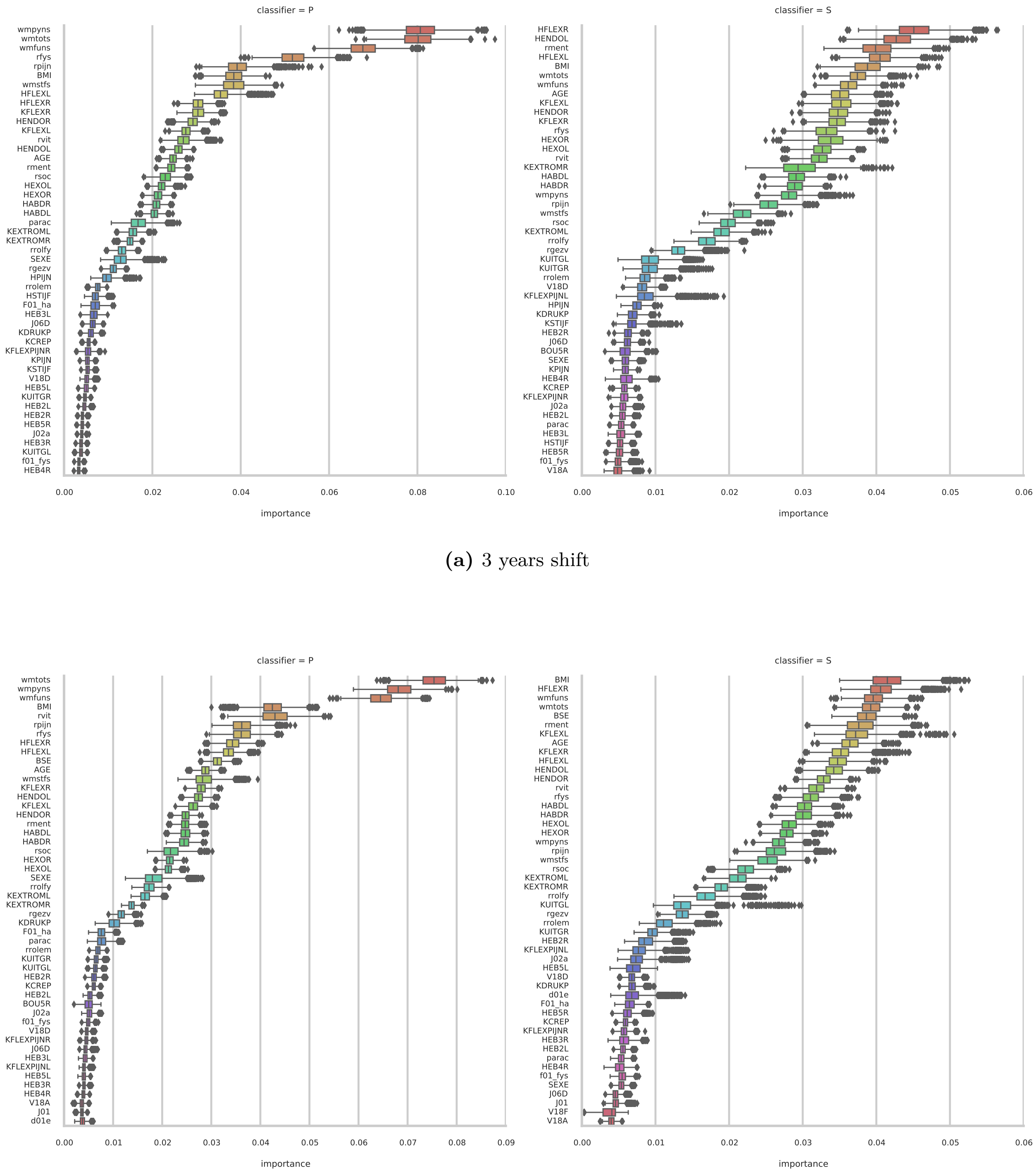


(a) 2 years shift

(b) 3 years shift

(c) 5 years shift

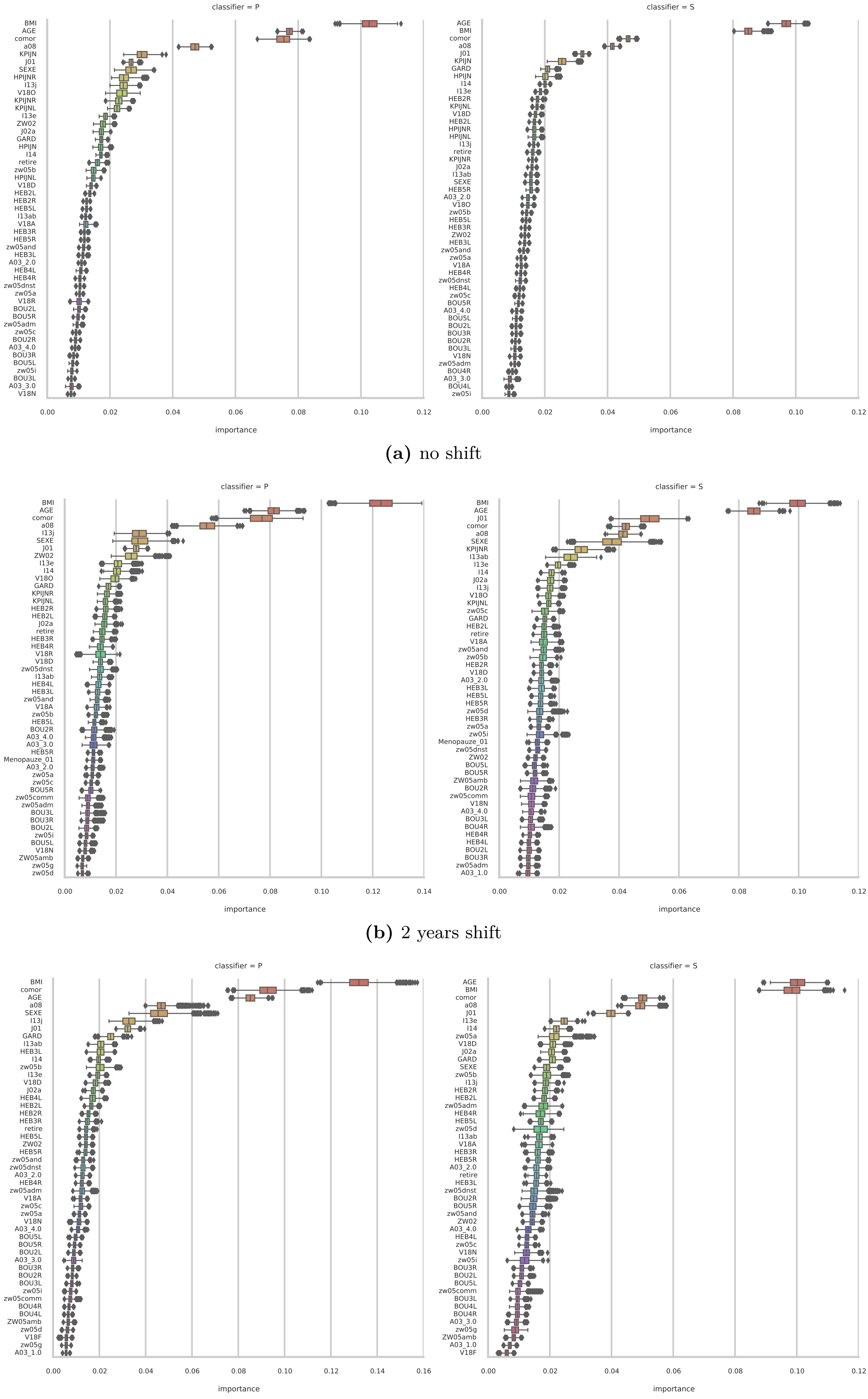
eFigure 37: Relative importance of top attributes used by the **PROCOAC selection models**. The two panels show the importance as impact on the probability of progression returned by different sub-models: pain-related (P) on the left, and structure-related (S) on the right. Attributes are listed in order of importance (descending).



(a) 3 years shift

(b) 5 years shift

eFigure 38: Relative importance of top attributes used by the **MUST selection models**. The two panels show the importance as impact on the probability of progression returned by different sub-models: pain-related (P) on the left, and structure-related (S) on the right. Attributes are listed in order of importance (descending).



(a) no shift

(b) 2 years shift

(c) 3 years shift

eFigure 39: Relative importance of top attributes used by the **DIGICOD selection models**. The two panels show the importance as impact on the probability of progression returned by different sub-models: pain-related (P) on the left, and structure-related (S) on the right. Attributes are listed in order of importance (descending).