Overview of iDPP@CLEF 2022: The Intelligent Disease Progression Prediction Challenge

Notebook for the iDPP Lab on Intelligent Disease Progression Prediction at CLEF 2022

Alessandro Guazzo[1*], Isotta Trescato[1*], Enrico Longato[1], Enidia Hazizaj[1], Dennis Dosso[1], Guglielmo Faggioli[1], Giorgio Maria Di Nunzio[1], Gianmaria Silvello[1], Martina Vettoretti[1], Erica Tavazzi[1], Chiara Roversi[1], Piero Fariselli[2], Sara C. Madeira[3], Mamede de Carvalho[3], Marta Gromicho[3], Adriano Chiò[2], Umberto Manera[2], Arianna Dagliati[4], Giovanni Birolo[2], Helena Aidos[3], Barbara Di Camillo[1] and Nicola Ferro[1]

[1]*University of Padua, Italy*
[2]*University of Turin, Italy*
[3]*University of Lisbon, Portugal*
[4]*University of Pavia, Italy*
[*] *These authors contributed equally*

Abstract

ALS is a severe chronic disease characterized by a progressive but variable impairment of neurological functions, characterized by high heterogeneity both in presentation features and rate of disease progression. As a consequence patients' needs are different, challenging both caregivers and clinicians. Indeed, the time of relevant events is variable, which is associated with uncertainty regarding the opportunity of critical interventions like non-invasive ventilation and gastrostomy, with implications on the quality of life of patients and their caregivers. For this reason, clinicians need tools able to support their decision in all phases of disease progression and underscore personalized therapeutic decisions.

The goal of iDPP@CLEF is to design and develop an evaluation infrastructure for AI algorithms able to: 1. better indicate intervention time; 2. stratify patients according to their phenotype and rate of disease progression; 3. predict progression rate in a probabilistic, time dependent fashion.

The participation in iDPP@CLEF was satisfactory, hinting at the interest of the community concerning the task. More so, the solutions identified by participants range over several different techniques and provided valid input to such a highly relevant domain as the prediction of the ALS progression.

CLEF 2022: Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy
alessandro.guazzo@phd.unipd.it (A. Guazzo); isotta.trescato@phd.unipd.it (I. Trescato); enrico.longato@unipd.it (E. Longato); enidia.hazizaj@studenti.unipd.it (E. Hazizaj); dennis.dosso@unipd.it (D. Dosso); guglielmo.faggioli@phd.unipd.it (G. Faggioli); giorgiomaria.dinunzio@unipd.it (G. M. Di Nunzio); gianmaria.silvello@unipd.it (G. Silvello); martina.vettoretti@unipd.it (M. Vettoretti); erica.tavazzi@phd.unipd.it (E. Tavazzi); chiara.roversi@studenti.unipd.it (C. Roversi); piero.fariselli@unito.it (P. Fariselli); sacmadeira@fc.ul.pt (S. C. Madeira); mamedemg@mail.telepac.pt (M. de Carvalho); mgromichosilva@medicina.ulisboa.pt (M. Gromicho); adriano.chio@unito.it (A. Chiò); umberto.manera@unito.it (U. Manera); arianna.dagliati@unipv.it (A. Dagliati); giovanni.birolo@unito.it (G. Birolo); haidos@fc.ul.pt (H. Aidos); barbara.dicamillo@unipd.it (B. Di Camillo); nicola.ferro@unipd.it (N. Ferro)
© 2022 Copyright for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3180/paper-88.pdf

1. Introduction

Amyotrophic Lateral Sclerosis (ALS) is a neurological disease that causes the progressive degeneration of the motor neurons that control voluntary muscles, resulting in an increasing impairment of motor and vital functions and leading to death usually within 4-5 years from the diagnosis. Likely resulting from a complex interplay of genetic and environmental factors, ALS is characterized by high heterogeneity in both symptoms and disease progression, especially in the early stages of the disease. This heterogeneity is partly responsible for the lack of effective prognostic tools in medical practice, as well as for the current absence of a therapy able to effectively slow down or reverse the disease course. On the one hand, patients need support for facing the psychological and economic burdens deriving from the uncertainty of how the disease will progress; on the other, clinicians require tools that may assist them throughout the patient's care, recommending tailored therapeutic decisions and providing alerts for urgently needed actions.
In order to improve the current diagnostic and prognostic situation, we should design and develop *Artificial Intelligence (AI)* algorithms be able to:

- stratify patients according to their phenotype, assessed all over the disease evolution;
- predict the progression of the disease in a probabilistic, time dependent fashion;
- better describe disease mechanisms.

The *Intelligent Disease Progression Prediction at CLEF (iDPP:CLEF)* lab[1] aims to design and develop an evaluation infrastructure for driving the development of such AI algorithms. By "evaluation infrastructure", we mean experimental collections, evaluation protocols, evaluation measures, ground-truth creation protocols, and so on. Indeed, in this context, it is fundamental, even if not so common yet, to develop shared approaches, promote the use of common benchmarks, foster the comparability and replicability of the experiments. Differently from previous challenges in the field, iDPP:CLEF addresses in a systematic way some issues related to the application of AI in clinical practice in ALS. Therefore, in addition to defining the risk scores based on the probability that an event will occur in the short or long term period, iDPP:CLEF also addresses the issue of providing information in a more structured and understandable way to clinicians.

The paper is organized as follows: Section 2 presents related challenges; Section 3 describes its tasks; Section 4 discusses the developed dataset; Section 5 explains the setup of the lab and introduces the participants; Section 6 introduces the evaluation measures adopted to score the runs; Section 7 analyzes the experimental results for the different tasks; finally, Section 8 draws some conclusions and outlooks some future work.

2. Related Challenges

To the best of our knowledge, within CLEF, there have been no other labs on this or similar topics before.

[1] https://brainteaser.health/open-evaluation-challenges/idpp-2022/

Outside CLEF, there have been a recent challenge on Kaggle[2] in 2021 and some older ones, the DREAM 7 ALS Prediction challenge[3] in 2012 and the DREAM ALS Stratification challenge[4] in 2015.

The Kaggle challenge used a mix of clinical and genomic data to seek for insights about the mechanisms of ALS and difference between people with ALS who progress faster versus those who develop it more slowly. The DREAM 7 ALS Prediction challenge [1] asked to use 3 months of ALS clinical trial information (months 0–3) to predict the future progression of the disease (months 3–12), expressed as the slope of change in *ALS Functional Rating Scale Revisited (ALSFRS-R)* [2], a functional scale that ranges between 0 and 40. The DREAM ALS Stratification challenge asked participants to stratify ALS patients into meaningful subgroups, to enable better understanding of patient profiles and application of personalized ALS treatments.

Differently from these previous challenges, iDPP:CLEF focuses on explainable AI and on temporal progression of the disease.

3. Tasks

iDPP:CLEF 2022 is the first edition of the lab and consists of pilot activities aimed both at an initial exploration of ALS progression prediction and at understanding of the refine and tune the labs itself for future iterations.

In particular, iDPP:CLEF targetes two kinds of activities:

1. preliminary and exploratory pilot tasks on disease progression prediction;
2. position papers on the explainability of the prediction algorithms.

Overall, this mix provides participants with the opportunity to make some hands-on experience with these data and provide feedback about the task design as well as to brainstorm on how to evaluate this kind of algorithms and, in particular, assess their explainability.

3.1. Pilot Task 1: Ranking Risk of Impairment

As shown in Figure 1, this task focuses on ranking of patients based on the risk of impairment in specific domains. More in detail, we use the ALSFRS-R scale to monitor speech, swallowing, handwriting, dressing/hygiene, walking and respiratory ability in time and ask participants to *rank patients based on time to event risk* of experiencing impairment in each specific domain.

More in detail, participants are asked to rank subjects based on the risk of early occurrence of

- **Task 1a**: *Non-Invasive Ventilation (NIV)* or (competing event) Death, whichever occurs first;
- **Task 1b**: *Percutaneous Endoscopic Gastrostomy (PEG)* or (competing event) Death, whichever occurs first;

[2] https://www.kaggle.com/alsgroup/end-als
[3] https://dreamchallenges.org/dream-7-phil-bowen-als-prediction-prize4life/
[4] https://dx.doi.org/10.7303/syn2873386.

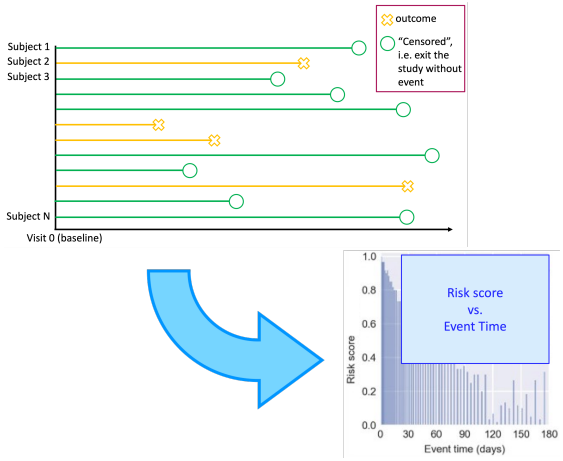


Figure 1: Task 1: from patients to ranking of patients based on time of event risk.

- **Task 1c**: Death.

For each of these tasks, participants are given a dataset containing 6 months of visits and are asked to rank patients on the risk of occurrence of one of the above events after month 6.

In particular, for each sub-task, we ask two type of submissions from participants:

- submissions using only data available until `Time 0`, i.e. the time of the first ALSFRS-R questionnaire;
- submissions using data available until `Month 6`.

Indeed, from the clinicians point of view, it is of interest to understand what they can say the first time they see the patient (`Time 0`) and what they can say if they collect additional data for the following 6 months.

3.2. Pilot Task 2: Predicting Time of Impairment

As shown in Figure 2, this task refines Task 1 asking participants to *predict when specific impairments will occur* (i.e. in the correct time-window). In this regard, we assess model calibration in terms of the ability of the proposed algorithms to estimate a probability of an event close to the true probability within a specified time-window.

In particular, participants are asked to predict the time of the event. Where the event is

- **Task 2a**: NIV or (competing event) Death, whichever occurs first;
- **Task 2b**: PEG or (competing event) Death, whichever occurs first;
- **Task 2c**: Death.

As in the previous case, for each sub-task, we ask two type of submissions from participants:

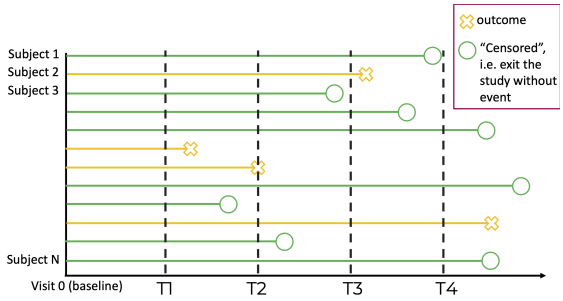


Figure 2: Task 2: from patients to time of impairment.

- submissions using only data available until `Time 0`, i.e. the time of the first ALSFRS-R questionnaire;
- submissions using data available until `Month 6`.

3.3. Position Papers Task 3: Explainability of AI algorithms

This task is not an evaluation challenge but rather a discussion on how to make these prediction algorithms explainable, also in a visual way.

Therefore, this task called for position papers to start a discussion on AI explainability including proposals on how the single patient data can be visualized in a multivariate fashion contextualizing its dynamic nature and the model predictions together with information on the predictive variables that most influence the prediction. We evaluated proposals of different visualization frameworks able to show the multivariate nature of the data and the model predictions in an explainable, possibly interactive, way.

Even if this task is not an evaluation challenge, authors of the papers are welcome to use the datasets provided by iDPP:CLEF, if they wish to give examples of their algorithms and solutions, or to explore the submissions made by other participants in iDPP:CLEF and apply their explainability techniques to them.

4. Dataset

iDPP:CLEF developed a dataset containing patient records from two clinical institutions in Turin, Italy, and in Lisbon, Portugal.

The dataset is fully anonymized, meaning that all the information which might reveal the identity of a patient, e.g. place of birth or city of residence, are removed; we also avoided absolute dates and made everything relative to `Time 0`, i.e. the date of the first ALSFRS-R questionnaire [2].

Table 1 summarizes the main features and variables available in the dataset. The following data are available for both the training and the test sets:

- the first available ALSFRS-R questionnaire a `Time 0` (both single question scores and total score).

Table 1
Main features of the iDPP:CLEF dataset.

Section	Sub-section	Variables
Baseline	Patient	Sex, Date of Birth
	ALS Onset	Date, Site
	Diagnosis	Date, Regions affected, Diagnostic Delay, FVC, BMI at diagnosis
Follow-up	Progression scores	ALSFRS-R, Rate of disease progression
	Tests	Hematologic tests, Muscle strength assessed by manual testing, Respiratory function tests
	Therapy	ALS treatments
	Other	Regions affected, Upper and lower motor neuron signs, Cognitive and neurophysiological changes
Clinical Events	History	BMI premorbid, Family history, Comorbidities, Previous surgery and trauma
	Interventions	Date of NIV, Date of PEG, Date of Tracheostomy
	Survival	Date of death
Lifestyle	Lifestyle	Working activity, Physical activity, History of smoking, Marital status, Education level

Thus, for example, time-of-onset and time-of-diagnosis are expressed as relative delta with respect to `Time 0` in months (also fractions);

- the slope of the ALSFRS-R score between time-of-onset and `Time 0` as:

$$slope = \frac{48 - \text{ALSFRS-R-score}\,(\texttt{Time 0})}{\texttt{Time 0} - \texttt{TimeOnset}}$$

- all the other static data, whose complete list is available at http://brainteaser.dei.unipd.it/challenges/idpp2022/assets/other/static-vars.txt
- visits , containing either other ALSFRS-R questionnaires or Spirometry, i.e. *Forced Vital Capacity (FVC)*. The complete list of variables for each visit is available at http://brainteaser.dei.unipd.it/challenges/idpp2022/assets/other/visits.txt.

We ensured that, for each patient, there are 6 months of data, so that predictions can be made using either only data available at `Time 0` or all the data available until month 6.

The following data are available only for the training set:

- Time of event (NIV, PEG, or DEATH); or
- Censoring time, i.e. time of the last available visit if none of the previous events occurs;

according to the following format:

```
0x4bed50627d141453da7499a7f6ae84ab 1 PEG 20.5
0x4d0e8370abe97d0fdedbded6787ebcfc 1 PEG 18.3
0x5bbf2927feefd8617b58b5005f75fc0d 1 DEATH 17.6
0x814ec836b32264453c04bb989f7825d4 0 NONE 37.4
0x71dabb094f55fab5fc719e348dffc85 1 PEG 8.2
...
```

where:

- Columns are separated by a white space;
- The first column is the `patient ID`, a 128 bit hex number (should be considered just as a string);
- The second column indicates whether the one of the above events occurred (1) or not (0);
- The third column is the occurred event. It comes from a controlled vocabulary and it can be either NIV, PEG, DEATH, or NONE;
- The fourth column is the time of the event, or the censoring time, from `Time 0` in months.

Training and test datasets follow a (roughly) 80%-20% proportion; more details about the split into training and test are provided below.

Both Task 1 and Task 2 use the same datasets but we prepared a separate dataset for each of the sub-tasks to make it simpler for participants to focus on a specific event to be predicted. Table 2 provides details about the created datasets.

Creation of the datasets

The full dataset contained approximately 4,800 records linked to patients, with around 20,000 ALSFRS-R questionnaires in total and 5,500 records concerning spirometries. The original data contain minor inconsistencies and typos. Therefore, we first process the data, removing records that are likely wrong or do not provide essential information to enable prediction. In terms of patient records we removed those presenting an unordered sequence of events (i.e., onset after diagnosis or diagnosis after death). Such event sequences are likely due to typos and other human errors, which result in wrong records that might introduce noise and spurious information in the final dataset.

Furthermore, a patient record was dropped if one or more of the following pieces of information were absent:

- onset or diagnosis dates;
- death date in records associated with dead patients;
- at least six months of historical ALSFRS-R questionnaires before an event (NIV, PEG, or (competing event) Death).

Table 2
Training and test datasets.

Training				
Sub-task	**Patients**	**ALSFRS-R**	**Spirometry**	**Outcome**
Sub-task a	1,454	3,668	1,189	• NIV: 675 patients (46.42%) • DEATH: 636 patients (43.74%) • NONE: 143 patients (9.83%)
Sub-task b	1,715	4,264	1,506	• PEG: 501 patients (29.21%) • DEATH: 969 patients (56.50%) • NONE: 245 patients (14.29%)
Sub-task c	1,756	4,366	1,536	• DEATH: 1,486 patients (84.62%) • NONE: 270 patients (15.38%)
Test				
Sub-task	**Patients**	**ALSFRS-R**	**Spirometry**	**Outcome**
Sub-task a	350	872	273	• NIV: 162 patients (46.29%) • DEATH: 152 patients (43.43%) • NONE: 36 patients (10.29%)
Sub-task b	430	1,049	361	• PEG: 120 patients (27.91%) • DEATH: 251 patients (58.37%) • NONE: 59 patients (13.72%)
Sub-task c	494	1,220	414	• DEATH: 417 patients (84.41%) • NONE: 77 patients (15.59%)

We adopt the filtering strategy mentioned above to grant that every record in the final dataset contains enough information to allow proper predictions.

Concerning the ALSFRS-R questionnaires, we removed those records that had one or more of the following problems:

- duplicate records;
- missing date;
- one or more of the ALSFRS-R items missing;
- ALSFRS-R reporting only the old 10^{th} item.

Table 3
Individuals and outcome types frequencies for each dataset released according to the task.

Variable	Dataset A	Dataset B	Dataset C
Number of subjects	1804	2145	2250
Outcome type	NIV: 837 Death: 788 Censoring: 169	PEG: 621 Death: 1220 Censoring: 304	Death: 1903 Censoring: 347

Furthermore, if one or more of the ALSFRS-R sub-scores or the total ALSFRS-R score do not agree with the sum of the associated ALSFRS-R items, we replace the value reported in the original dataset with the sum of the linked items. Finally, regarding the spirometries, we removed duplicated records, records with a missing date, and FVC percentage value.

Figure 3 illustrates a set of - synthetic - patients and their clinical history, describing whether they satisfy the conditions to be inserted into the dataset. By construction, the first ALSFRS visit (blue bullets) is considered as `Time 0`, while the moment of the previous spirometries (yellow bullets) and subsequent visits is indicated as the difference in months with respect to the reference ALSFRS.

- Patient 1 is inserted into the dataset, having a proper sequence of visits, questionnaires and events (at least six months of information before the first event).
- Patient 2, on the other hand, cannot be included in the dataset since they do not have enough information.
- For Patient 3, we observe that only four months passed between the first ALSFRS and the first event. Thus, even though we have 6 months of overall information (first spirometry to event), we cannot retain the record.
- Patient 4, regardless of the fact that they have a single ALSFRS, can be included in the dataset since the distance between the first ALSFRS and the event is above six months.
- Both patients 5 and 6 need to be excluded from further analyses: the former does not have six months of information before the first event, while the latter does not have enough history, regardless of the spirometry taken before the first ALSFRS.
- Patients 7 and 8, on the other hand, can be considered: the former has a proper clinical history, while the latter, even though they have a "censored" event, has more than six months of history.

Out of the 2559 original valid patient records, 2250 contained at least 6 months of information. Nevertheless, it is not possible to put all the patients on all the datasets. Assume for example a patient that underwent a NIV intervention at month 5. The record associated with such patient cannot be considered feasible for the Dataset A: it does not contain at least 6 months of information before the outcome. Nevertheless, if the very same patient undergoes a PEG event at month 7, the patient can be considered feasible to Dataset B (and therefore to Dataset C). Table

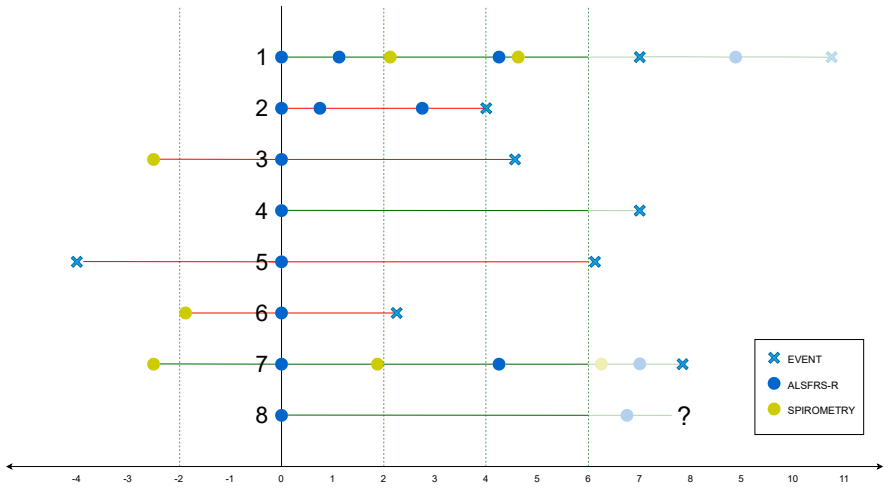


Figure 3: Sequences of events that allow (or forbid) a patient to be considered as suitable to belong to the dataset.

3 reports the number of valid patients present in each dataset, with the count of the labels that characterize them. Dataset A is the least populated: an NIV procedure is often required quite early in the course of the disease, and thus, for a large group of patients (755), it happened before six months of follow-up. Thus we were forced to exclude them since not enough information was available in those cases. We observe that dataset B contains 2145 individuals (414 were discarded). We observe an increase in the number of individuals considered suitable for this dataset: it is more likely that a PEG will be necessary much later in the progression of the disease, and thus more patients accrue more than six months of data before the outcome. Notice that patients that were labelled as NIV in dataset A, if they are suitable to enter dataset B (meaning that has at least six months of information before the PEG), can be labelled as either PEG, DEATH or NONE in the case they did not undergo any events other than the NIV for which they entered dataset A. Conversely, patients labelled as DEATH can either be labelled with PEG, in case they received a PEG after six months from the first ALSFRS-R and before their death, or DEATH in case they did not, but they cannot switch to class NONE. The converse is also true if we consider the relationship between datasets B and A. Finally, dataset C contains 2250 patients, with only 305 records discarded. Given the criteria used to construct Dataset C, it contains patients present in datasets A and B. Patients that received an NIV or a PEG after six months from the first ALSFRS-R have at least six months of information before their death. We also include new patients: those that had an NIV before six months from the first ALSFRS-R and those that had a PEG before six months but survived (or died) more than six months after the first ALSFRS-R. Patients that were not included in dataset C are those that survived less than six months after the first ALSFRS-R.

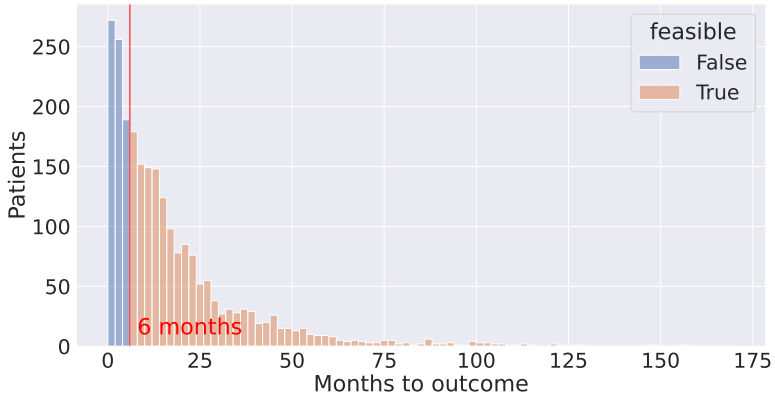
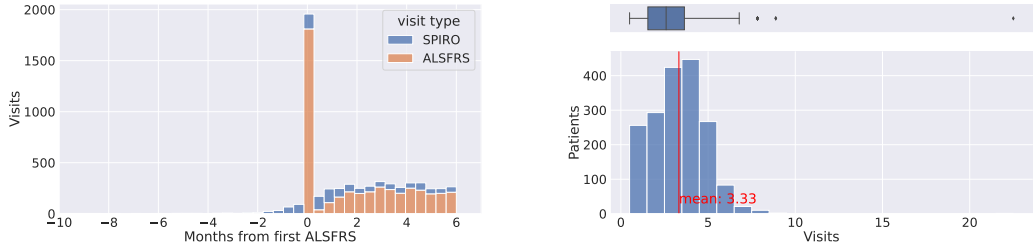


Figure 4: Distribution of the distance between the first visit and one event among NIV, Death or Censoring event (the one happening first) over the patient set.



(a) Number of visits by time span

(b) Number of visits by patient

Figure 5: Distribution of the visits included in dataset A.

4.0.1. NIV and Death

Concerning dataset A (NIV, Death or Censoring Events), we observe that 1804 patients satisfy the conditions to be included in the dataset (6 months of data between the first ALSFRS-R and the event).

Figure 4 reports the distribution of patients with respect to the distance from the event. We notice a very steep distribution, suggesting that the events often happen before month 6. Compared to Figure 7 and Figure 10, we argue that the NIV is the most likely event in the first six months. Both Figures 7 and 10 present a lower peak in the first part of the distribution.

Figure 5a reports the number of visits, according to their type, with respect to the time from the reference ALSFRS-R questionnaire. We notice a predictable increase in the number of visits with respect to the months passed since the diagnosis. Figure 5b illustrates the number of visits associated with each feasible patient that was included in the dataset. Interestingly, we notice that the mean number of visits in the considered six months is 3.3, with a slightly lower median number.

Finally, Figure 6. reports the number of patients incurring in a certain event for dataset A.

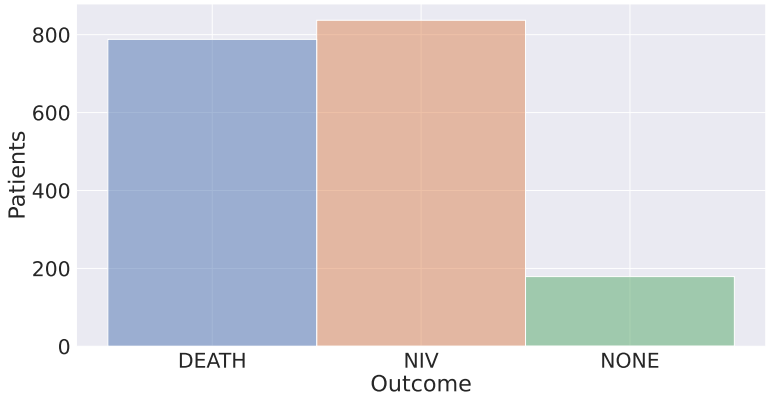


Figure 6: Distribution of events in Dataset A.

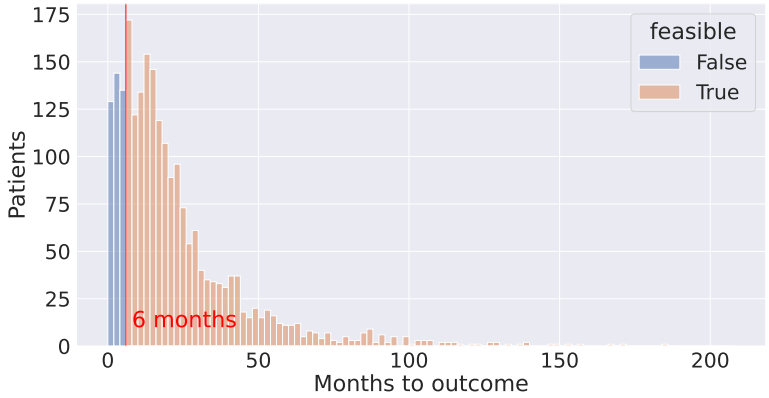
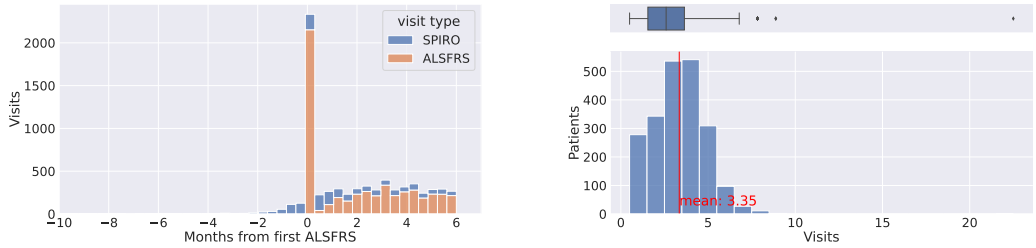


Figure 7: Distribution of the distance between the first visit and one event among PEG, Death or Censoring event (the one happening first) over the patient set.

Among patients included in dataset A, 788 are labelled with the event DEATH, 837 with the event NIV and 179 patients are labelled with NONE, which indicates the censoring event. If a patient incurred both NIV and Death, the event associated with that patient that needs to be predicted is only the NIV.

4.0.2. PEG and Death

Figure 7. reports the distribution of the distance to the event if we consider as feasible events the PEG, Death or the censoring event. As we noticed with the number of subjects that belong to each dataset, for what concerns dataset B, we have fewer subjects that have less than 6 months between their first ALSFRS-R and the event. In particular, we are able to accrue 2145 suitable subjects for this dataset.



(a) Number of visits by time span

(b) Number of visits by patient

Figure 8: Distribution of the visits included in dataset B.

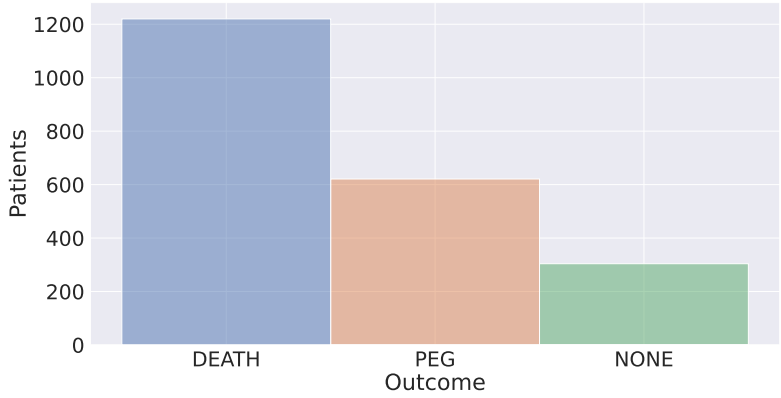


Figure 9: Distribution of events in Dataset B.

Following what we observed for dataset A, we report the number of visits according to their distance from the first ALSFRS-R (Figure 8a) as before, we notice an increase in the number of visits with the months passing. In this case, as Figure 8b. illustrates, we have on average 3.35 visits for each patient, with the vast majority of patients having between 2 and 3 visits.

Figure 9 exhibits the distribution of events in dataset B. Differently from dataset A, where the most common event was the NIV (the focus of the subtask) in this case, the most common event is DEATH. This indicates a generally increased probability of incurring death before receiving a PEG. From the perspective of an AI practitioner, it also indicates that models that were suited to task A are likely not usable trivially on this second dataset but would require a specific tuning in order to be applied on this second scenario.

4.0.3. Death and Censoring events

The final dataset, dataset C, concerns the death and censoring events and it is used in the tasks that require predicting such events. As shown by Figure 10. for what concerns the number of patients included in this dataset, we overcome both dataset A and dataset B: several patients that did not have the right features to be considered feasible before (at least 6 months

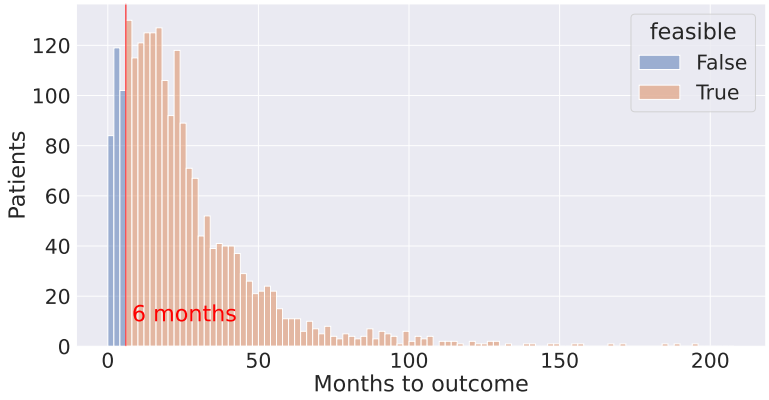
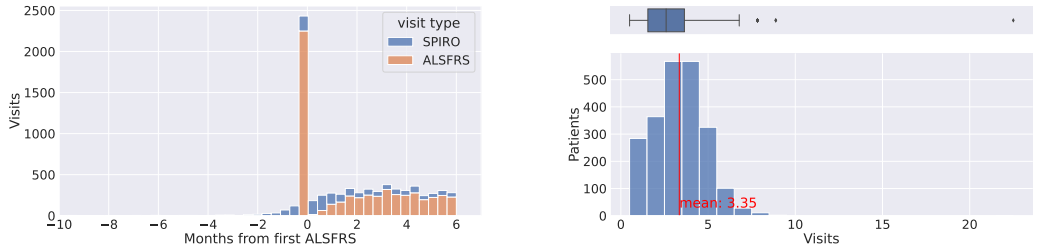


Figure 10: Distribution of the distance between the first visit and one event among Death or Censoring event (the one happening first) over the patient set.



(a) Number of visits by time span

(b) Number of visits by patient

Figure 11: Distribution of the visits included in dataset C.

of information from the first ALSFRS-R) can be included in this dataset. Dataset C counts 2250 records associated with as many patients.

Similarly to the previous datasets, Figure 11 illustrates the distribution of the visits, either with respect to the timespan or to the patients. As for the previous case, we have 3.35 visits per patient on average, with the mode on 3 visits and approximately 50% of the patients having between 2 and 3 visits.

Figure 12 illustrates how events distribute in dataset C. We have more deaths than censoring events. This is expected: historical data regards patients with ages up to more than 90 years followed sometimes for several years. As mentioned before, 1903 patients overwent the death event, while 347 incurred into the censoring event. Notice that, in this case, we have more censoring events than for what concerns previous datasets, since patients that incurred NIV and/or PEG but survived are considered suitable for the censoring event in this case.

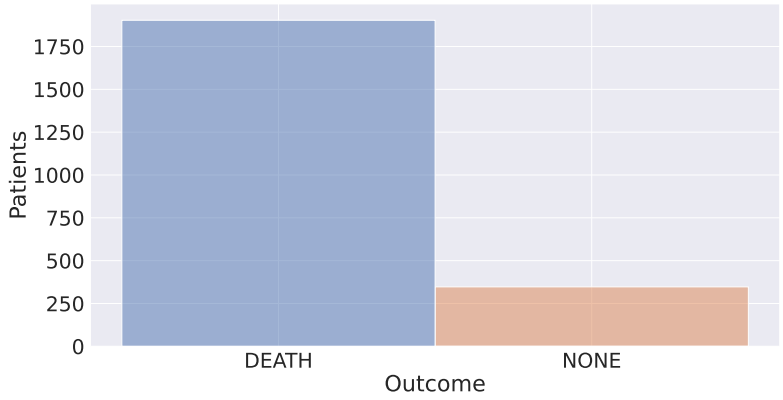


Figure 12: Distribution of events in Dataset c.

Split into training and test

Each of the three available datasets (sub-task a, b, and c) was split into a training set and a test set, with proportions 80% and 20%, respectively. The data were split stratifying the subjects according to outcome time and to the specific outcome type (i.e., *death, NIV, none* for sub-task a, *death, PEG, none* for sub-task b, and *death, none* for sub-task c). Stratifying by these two variables is instrumental to the fairness of the challenge as it forces an equal distribution of their levels across the two subsets. The simplest method to verify whether stratification has yielded a satisfactory split is to compare the distributions of the stratification variables (*outcome time* and *outcome type*) in each training/test pair. From the literature, certain variables are known to be particularly relevant for predicting events related to ALS progression [3], therefore, even though they were not included in the stratification criteria, *sex*, *age at onset*, *onset site*, *ALSFRS-F slope*, and the number of available visits in the first 6 months were also verified to be equally represented in the training and test sets. Tables 4, 5, and 6 report the variables' distributions for sub-task a, b, and c, respectively: the second column reports the distribution in the training sets and the third column in the test sets. Note that *outcome time* is measured in months from time 0, i.e., the first registered ALSFRS-R visit, while *age at onset* is measured in years. Since the distributions were similar, the training/test split provided to the participants met best-practice quality standards.

The distributions of the two variables used to stratify are also represented in Figures 13, 14 and 15 via bar plots for categorical variables (i.e., outcome types), and density plots for continuous variables (i.e., outcome time).

Table 4
Sub-task a, comparison between training and test populations. Continuous variables are presented as *median [1st - 3rd quartiles]*; discrete variables as *count (percentage on sample total)*, for each level.

	Training	**Test**
Number of subjects	1454	350
Outcome type	Death: 636 (44%) NIV: 675 (46%) Censoring: 143 (10%)	Death: 152 (43%) NIV: 162 (46%) Censoring: 36 (10%)
Outcome time	17.75 [11.14-30.99]	20.72 [11.25-36.76]
Sex	M: 743 (51%) F: 711 (49%)	M: 188 (54%) F: 16 (46%)
Age at onset	64.89 [55.66-70.76]	64.76 [56.66-71.58]
Onset site	Bulbar: 449 (31%) Axial: 3 (0.002%) Generalized: 4 (0.003%) Limbs: 998 (68%)	Bulbar: 105 (30%) Axial: 0 (0%) Generalized: 0 (0%) Limbs: 242 (70%)
ALSFRS-R slope	0.43 [0.24-0.79]	0.41 [0.23-0.80]
Number of available visits	2.00 [2.00-3.00]	3.00 [2.00-3.00]

Table 5
Sub-task b, comparison between training and test populations. Continuous variables are presented as *median [1st - 3rd quartiles]*; discrete variables as *count (percentage on sample total)*, for each level.

	Training	**Test**
Number of subjects	1715	430
Outcome type	Death: 969 (57%) 501 (29%) Censoring: 245 (14%)	Death: 251 (58%) NIV: 120 (28%) Censoring: 59 (14%)
Outcome time	19.97 [12.57-36.53]	21.82 [12.70-38.30]
Sex	M: 923 (54%) F: 792 (46%)	M: 241 (56%) F: 189 (44%)
Age at onset	65.14 [56.86-71.88]	64.83 [55.99-70.42]
Onset site	Bulbar: 499 (29%) Axial: 31 (2%) Generalized: 8 (0.5%) Limbs: 1177 (68.5%)	Bulbar: 125 (29%) Axial: 12 (3%) Generalized: 1 (0.2%) Limbs: 292 (68%)
ALSFRS-R slope	0.47 [0.25-0.84]	0.44 [0.24-0.85]
Number of available visits	2.00 [2.00-3.00]	2.00 [2.00-3.00]

Table 6
Sub-task c, comparison between training and test populations. Continuous variables are presented as *median [1st - 3rd quartiles]*; discrete variables as *count (percentage on sample total)*, for each level.

	Training	**Test**
Number of subjects	1756	494
Outcome type	Death: 1486 (85%) Censoring: 270 (15%)	Death: 417 (84%) Censoring: 77 (16%)
Outcome time	24.68 [14.42-41.84]	22.48 [13.72-38.91]
Sex	M: 930 (53%) F: 826 (47%)	M: 273 (55%) F: 221 (45%)
Age at onset	65.38 [58.27-72.18]	65.03 [57.02-70.86]
Onset site	Bulbar: 554 (31.5%) Axial: 32 (2%) Generalized: 4 (0.5%) Limbs: 1162 (66%)	Bulbar: 149 (30%) Axial: 13 (3%) Generalized: 1 (0.2%) Limbs: 331 (67%)
ALSFRS-R slope	0.49 [0.26-0.88]	0.45 [0.24-0.85]
Number of available visits	2.00 [2.00-3.00]	2.00 [2.00-3.00]

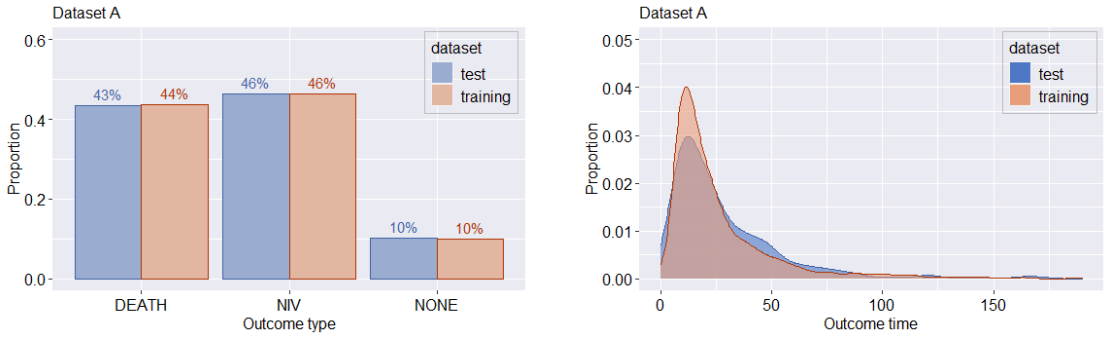


Figure 13: Comparison of the distributions of stratification variables for sub-task A: outcome type on the left and outcome time on the right. The distribution on the training set is in blue while that of the test set in orange.

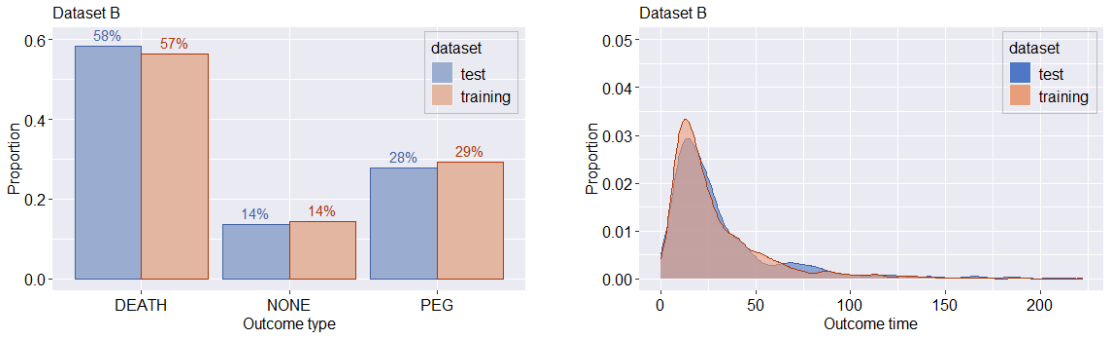


Figure 14: Comparison of the distributions of stratification variables for sub-task B: outcome type on the left and outcome time on the right. The distribution on the training set is in blue while that of the test set in orange.

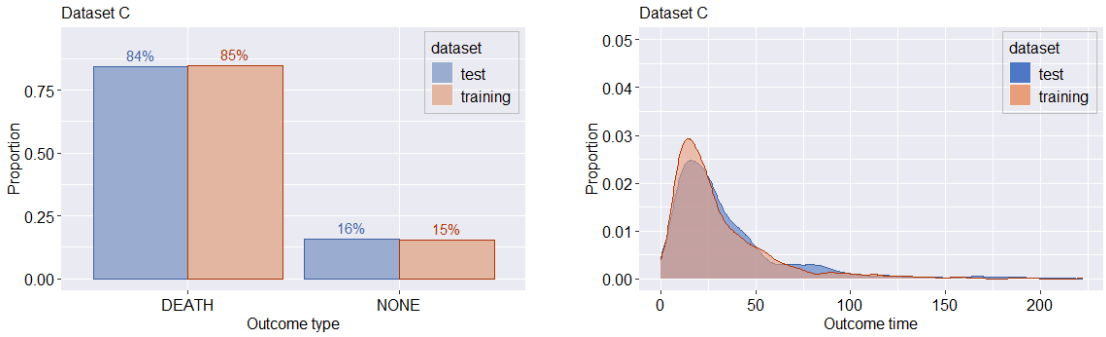


Figure 15: Comparison of the distributions of stratification variables for sub-task C: outcome type on the left and outcome time on the right. The distribution on the training set is in blue while that of the test set in orange.

5. Lab Setup and Participation

5.1. Guidelines

Participating teams were provided with the following guidelines:

- The runs should be submitted in a textual format in the participant repository, both described below;
- Each group can submit a maximum of 5 runs for each sub-task, thus amounting to maximum 15 runs for each of Task 1 and Task 2;
- For each task, participants are asked to submit two types of runs: either using only the information available at `Time 0` or using all the information available in the first 6 months.

Runs should be uploaded using the following name convention for their identifiers:

```
<teamname>_T<1|2><a|b|c>_<train>_<freefield>
```

where:

- `teamname` is the name of the participating team;
- `T<1|2><a|b|c>` is the identifier of the task the run is submitted to, e.g. T1b for Task 1, sub-task b;
- `train` is data window used to train the algorithm. It can be either `M0`, if only the data available at Time 0 have been used, or `M6` if all the data available in the first 6 months have been used;
- `freefield`is a free field that participants can use as they prefer.

For example, a complete run identifier may look like

```
upd_T2b_M6_survRF
```

where:

- `upd` is the University of Padua team;
- `T2b` means that the run is submitted for Task 2, sub-task b;
- `M6` means that the algorithm has been trained using all the data available in the first 6 months;
- `survRF` suggests that participants have used survival random forests as a prediction method.

Participant Repository

Participants are provided with an individual git repository for all the tasks they take part in. The repository contains the runs, resources, and possibly the code produced by each participant in order to promote reproducibility and open science. The repository is organised as follows:

- submission: this folder contains the runs submitted for the different tasks.
- score: this folder contains the performance scores of the submitted runs.
- code: this folder contains the source code of the developed system.
- resource: this folder contains any additional resources created during the participation.
- report: this folder contains the template for participant report.

The submission and score folders are organized into sub-folders for each task as follows:

- submission/task1: for the runs submitted to the first task. Similar structure for the other tasks.
- score/task1: for the performance scores of the runs submitted to the first task. Similar structure for the other tasks.

The goal of iDPP:CLEF is to speed up the creation of systems and resources for ALS progression prediction as well as openly share these systems and resources as much as possible. Therefore, participants are more than encouraged to share their code and any additional resources they have used or created.

All the contents of these repositories are released under the *Creative Commons Attribution-ShareAlike 4.0 International License*[5].

Task 1 Run Format

Runs had to be submitted as a text file with the following format:

```
0x4bed50627d141453da7499a7f6ae84ab 0.897 0 PEG upd_T1b_M6_survRF
0x4d0e8370abe97d0fdedbded6787ebcfc 0.773 1 PEG upd_T1b_M6_survRF
0x5bbf2927feefd8617b58b5005f75fc0d 0.773 2 DEATH upd_T1b_M6_survRF
0x814ec836b32264453c04bb989f7825d4 0.615 3 NONE upd_T1b_M6_survRF
0x71dabb094f55fab5fc719e348dffc85 0.317 4 PEG upd_T1b_M6_survRF
...
```

where:

- Columns are separated by a white space;
- The first column is the `patient ID`, a 128 bit hex number (should be considered just as a string);

[5] http://creativecommons.org/licenses/by-sa/4.0/

- The second column shows the prediction score that generated the ranking. It is expected to be a floating point number in the range $[0, 1]$. This score must be in descending (non-increasing) order;
- The third column is the rank of the patient by her/his risk of impairment, starting from 0. This is expected to be a strictly increasing integer number. It is important to include the rank so that we can handle tied scores (for a given run) in a uniform fashion;
- The fourth column is the predicted event. It comes from a controlled vocabulary and it can be either NIV, PEG, DEATH, or NONE. Note that, since each sub-task is focused on the prediction of a specific event (NIV, PEG, or DEATH), this column will contain that event or the competing event DEATH or NONE;
- The fifth column is the run identifier, according to the format described above. It must uniquely identify the participating team and the submitted run.

Task 2 Run Format

Runs had to be submitted as a text file with the following format:

```
0x4bed50627d141453da7499a7f6ae84ab 6-12 PEG upd_T2b_M6_survRF
0x4d0e8370abe97d0fdedbded6787ebcfc 18-24 PEG upd_T2b_M6_survRF
0x5bbf2927feefd8617b58b5005f75fc0d 24-30 DEATH upd_T2b_M6_survRF
0x814ec836b32264453c04bb989f7825d4 >36 NONE upd_T2b_M6_survRF
0x71dabb094f55fab5fc719e348dffc85 >36 PEG upd_T2b_M6_survRF
...
```

where:

- Columns are separated by a white space;
- The first column is the `patient ID`, a 128 bit hex number (should be considered just as a string);
- The second column shows the prediction window in months. Possible values are taken from a controlled vocabulary as follows:
  - `6-12`: the event will happen in the range of months $(6, 12]$;
  - `12-18`: the event will happen in the range of months $(12, 18]$;
  - `18-24`: the event will happen in the range of months $(18, 24]$;
  - `24-30`: the event will happen in the range of months $(24, 30]$;
  - `30-36`: the event will happen in the range of months $(30, 36]$;
  - `>36`: the event will happen in the range of months $(36, +\infty)$.
- The third column is the rank of the patient by her/his risk of impairment, starting from 0. It is important to include the rank so that we can handle tied scores (for a given run) in a uniform fashion;

Table 7
Teams participating in iDPP:CLEF 2022.

Team Name	Description	Country	Repository	Paper
BioHIT	National Centre for Scientific Research Demokritos (NCSR Demokritos)	Greece	https://bitbucket.org/brainteaser-health/idpp2022-biohit	–
CompBioMed	Department of Medical Sciences, University of Turin	Italy	https://bitbucket.org/brainteaser-health/idpp2022-compbiomed-unito	Pancotti et al. [4]
FCOOL	Faculty of Sciences of the University of Lisbon	Portugal	https://bitbucket.org/brainteaser-health/idpp2022-fcool	Branco et al. [5] and Nunes et al. [6]
LIG GETALP	Laboratoire d'Informatique de Grenoble, Université Grenoble Alpes	France	https://bitbucket.org/brainteaser-health/idpp2022-lig-getalp	Mannion et al. [7]
SBB	University of Padua	Italy	https://bitbucket.org/brainteaser-health/idpp2022-sbb	Trescato et al. [8]

Table 8
Break-down of the runs submitted by participants for each task and sub-task. Participation in Task 3 does not involve submission of runs and it is marked just with a tick.

Team Name	Total	Task 1			Task 2			Task 3
		a	b	c	a	b	c	
BioHIT	18	3	3	3	3	3	3	–
CompBioMed	40	8	8	6	6	6	6	–
FCOOL	15	–	–	–	5	5	5	✓
LIG GETALP	23	4	4	4	4	4	3	–
SBB	24	4	4	4	4	4	4	–
Total	**120**	**19**	**19**	**17**	**22**	**22**	**21**	

- The fourth column is the predicted event. It comes from a controlled vocabulary and it can be either NIV, PEG, DEATH, or NONE. Note that, since each sub-task is focused on the prediction of a specific event (NIV, PEG, or DEATH), this column will contain that event or the competing event DEATH or NONE;
- The fifth column is the run identifier, according to the format described above. It must uniquely identify the participating team and the submitted run.

5.2. Participants

Overall, 43 teams registered for participating in iDPP:CLEF but only 5 of them actually managed to submit runs for at least one of the offered tasks. Table 7 reports the details about the participating teams.

Table 8 provides breakdown of the number of runs submitted by each participant for each task and sub-task. Overall, we have received 120 runs which are roughly broken down evenly among the different tasks.

6. Evaluation Measures

iDPP:CLEF adopted several state-of-the-art evaluation measures to assess the performance of the prediction algorithms, among which:

- *ROC curve and/or the precision-recall curve (and area under the curve)* to show the trade-off between clinical sensitivity and specificity for every possible cut-off of the risk scores;
- *Concordance Index (C-index)* to summarize how well a predicted risk score describes an observed sequence of events.
- *E/O ratio and Brier Score* to assess whether or not the observed event rates match expected event rates in subgroups of the model population.
- *Specificity and recall* to assess, for each interval, the ability of the models of correctly identify true positives and true negatives.
- *Distance* to assess how far the predicted time interval was from the true time interval.

To ease the computation and reproducibility of the results, scripts for computing the measures is publicly available[6].

The next two sections provide details about the adopted measures for each Task.

6.1. Pilot Task 1: Ranking Risk of Impairment

The runs submitted for Task 1 were evaluated by means of Harrel's concordance index (C-index) [9], area under the receiver operating characteristic curve (AUROC) [10], and the Brier score (BS) [11]. The 95% confidence intervals of the C-index and the AUROC were also considered [12].

The C-index has an advantage over the other considered metrics (i.e., AUROC and BS) in that it can be used to evaluate model discrimination on the test sets regardless of censored data. According to the best practices in the field [13], before computing the C-index, a final censoring time equal to the last time-to-event in the training was set on each test set. This ensured consistency between Task 1's final results and those that might have been obtained by the participants during model development.

The AUROC and BS were computed at various prediction horizons (PHs). Specifically, seven clinically relevant PHs were considered, namely: 12, 18, 24, 30, 36, 48, and 60 months after the baseline. For each PH, the corresponding version of the test set comprised: all patients who experienced an event before the PH, and all patients who experienced an event or were censored after the PH as censored patients (and were, thus, censored at that PH). As the status of patients censored before the PH was, by definition, unknown, they were excluded from performance evaluation at that PH.

To contextualize the results obtained by the participants, each run was compared to the empirical lower bound established by the average performance of 100 random classifiers (i.e., such that their output was a random continuous number, uniformly sampled in the range $[0, 1]$).

[6] https://bitbucket.org/brainteaser-health/idpp2022-performance-computation

6.2. Pilot Task 2: Predicting Time of Impairment

To evaluate the predictions of Task 2, the selected evaluation metrics were: the specificity, the recall, and a measure of distance between the predicted and correct time intervals.

Confusion matrices were computed to derive specificity, i.e., the number of correct negative predictions divided by the total number of negatives, and recall, i.e., the ratio of correct positive predictions over the total predicted positives. Two types of confusion matrix were computed to evaluate two possible prediction goals: time interval prediction approach and label prediction approach.

For the time interval prediction approach the outcome times reported in the column *Time* of the published test sets were mapped to the corresponding interval ("6-12", "12-18", "18-24", "24-30", "30-36", or ">36" months). A conformance check was performed on the participants' predicted times: predictions in the time interval "0-6" were reassigned to the interval "6-12", i.e., the closest allowed interval. The confusion matrices reported the predicted time interval vs the true time interval, independently of the predicted event.

In the label prediction approach, maintaining the outcome time mapped as above, goodness of prediction was assessed by fixing the observation time to each time interval and evaluating the predicted label vs the true label. Let *true_label* be the label reported for the subject in the column *Type* of the published test sets, *this_interval_label* the corresponding label used in the confusion matrix of a specific time interval, *true_time_interval* the time interval in which the outcome happens, *observation_time* the time interval under evaluation, *predicted_label* the label predicted by the participants, *this_interval_predicted_label* the predicted label for the specific time interval, and *predicted_time_interval* the time interval in which the participants predicted the outcome. Then, the ground truth was constructed as follows.

- if *true_label* = "NONE", then *this_interval_label* = "NONE";
- if *true_time_interval* > *observation_time*, then *this_interval_label* = "NONE";
- if *true_time_interval* < *observation_time*, then *this_interval_label* = "CENSORED";
- if *true_time_interval* = *observation_time* and *true_label* = "NIV", then *this_interval_label* = "NIV";
- if *true_time_interval* = *observation_time* and *true_label* = "PEG", then *this_interval_label* = "PEG";
- if *true_time_interval* = *observation_time* and *true_label* = "DEATH", then *this_interval_label* = "DEATH";
- if *true_time_interval* = *observation_time* and *true_label* = "CENSORED", then *this_interval_label* = "CENSORED".

Similarly, the predicted labels were remapped as follows.

- if *predicted_label* = "NONE", then *this_interval_predicted_label* = "NONE";
- if *predicted_time_interval* > *observation_time*, then *this_interval_predicted_label* = "NONE";

- if *predicted_time_interval* < *observation_time*, then *this_interval_predicted_label* = "CENSORED";
- if *predicted_time_interval* = *observation_time* and *predicted_label* = "NIV", then *this_interval_predicted_label* = "NIV";
- if *predicted_time_interval* = *observation_time* and *predicted_label* = "PEG", then *this_interval_predicted_label* = "PEG";
- if *predicted_time_interval* = *observation_time* and *predicted_label* = "DEATH", then *this_interval_predicted_label* = "DEATH";
- if *predicted_time_interval* = *observation_time* and *predicted_label* = "CENSORED", then *this_interval_predicted_label* = "CENSORED".

This manipulation allowed a fair comparison of the predicted label vs the true label for each interval, actualized in comparing *this_interval_label* and *this_interval_predicted_label*.

Furthermore, a measure of distance between the predicted and correct time intervals, in months, was also considered (AbsDist). To compute the AbsDist, all the time intervals were replaced with the mean value of each interval (i.e., "6-12" was replaced with 9, "12-18" with 15, "18-24" with 21, "24-30" with 27, "30-36" with 33, and "$>$36" with 39). The difference between the predicted values and the true values was then computed as $meanValue_{predicted\,time\,interval} - meanValue_{true\,time\,interval}$. The obtained differences were, by construction, in the range $[-36; +36]$ where a smaller modulus corresponds to more accurate predictions. Negative values correspond to a events that occur before the predicted time and positive values to events that occur after. Finally, the AbsDist was obtained by averaging the differences absolute values.

To contextualize the results obtained by the participants, each run was compared to the performance of several synthetic runs, with the following characteristics:

- *min_interval*: a run in which the predicted time intervals are identical for all subjects, and fixed at the first possible time interval, i.e. "6-12";
- *max_interval*: a run in which the predicted time intervals are identical for all subjects, and fixed at the last possible time interval, i.e. "$>$36";
- *interval_18_24*: a run in which the predicted time intervals are identical for all subjects, and fixed at the time interval "18-24";
- *random_interval*: 100 randomly generated runs, but with the same distribution as the test set distribution (i.e., such that their output was sampled among the labels "6-12", "12-18", "18-24", "24-30", "30-36", "$>$36" following the same distribution of the true intervals);
- *inverse_distr_interval*: 100 randomly generated runs, but with an inverse distribution compared to the test set distribution (i.e., such that their output was sampled among the labels "6-12", "12-18", "18-24", "24-30", "30-36", "$>$36" following the inverse distribution of the true outcome);
- *corr_interval*: 100 correlated runs, with correlation coefficient to the true intervals ~ 0.7.

7. Results

For each task, we report here the analysis of the performance attained by the runs submitted by the Lab's participants according to the metrics described in Section 6.

7.1. Pilot Task 1: Ranking Risk of Impairment for ALS

We first determine the performance of the submitted runs in terms of C-index. As expected, the random classifiers yielded an average C-index of around 0.5 in each sub-task. Runs submitted by the BioHit team were comparable to those obtained by the random classifiers when considering sub-tasks a and b, meanwhile they were slightly better than random when considering sub-task c. In all three sub-tasks, runs submitted by other participants significantly outperformed the random classifiers with the CampBioMed team leading the pack (max C-index > 0.725). The complete list of figures concerning the c-index is available in Appendix A.

Secondly, we evaluate the submissions in terms of AUROC. The AUROC confirmed the results obtained when considering the C-index. Again, as expected, the random classifiers yielded an AUROC of around 0.5. Runs submitted by the BioHit team showed a discrimination that was comparable to the one of the random classifiers, and all runs submitted by other participants significantly outperformed the random classifiers. The CampBioMed and SBB teams obtained the best results when all the information available in the first 6 months was considered (M6 runs).

Across all sub-tasks, higher AUROC values were obtained when the PH was short (12-18 months) or long (48-60 months). This behaviour is mostly due to the trade-off between PH length and number of events at various time points. Short PHs tend to lead to increased predictive power as they are tightly correlated with the input data, however, they might be too close to the start of follow-up for many events to have been observed; on the contrary, longer PHs have a weaker link to the input data (harder to predict), but more events are available, leading to a more robust model training. The complete evaluation of the submitted runs in terms of AUROC is available in Appendix B.

Overall, model discrimination was acceptable, with C-index and AUROC values around 0.7 for all submitted models across all sub-tasks. Participants' runs performed better in sub-task b (prediction of PEG or death) than in the other sub-tasks according to all discrimination metrics.

As a final anylsis concerning pilot task 1, we compute the BS for the runs submitted for all the sub-tasks. The random classifier yielded a BS of around 0.325 regardless of the considered PH and sub-task as the random probability values were, on average, always well distributed in the range $[0, 1]$. Runs submitted by the CampBioMed team showed the best calibration at short PHs (12-24 months), while those submitted by the SBB team showed the worst one, mainly due to a consistent overestimation of the event probability. Other participants' runs, when considering a short PH, led to BS values that were comparable with those obtained by the random classifiers, as their models neither accurately predict the event probability, nor showed consistent under- or overestimation trends. Some runs submitted by the CampBioMed team, which had good calibration with a short PH, led to a poorer calibration at longer PHs (48-60 months). All other runs submitted by the participants significantly outperformed the random classifiers at 48-60 months by showing good calibration, with the SBB team leading the pack.

Overall, calibration was comparable across different sub-tasks. Instead, for most submitted runs, the BS decreased as the PH widened. This result suggest that the submitted models were trained without setting an artificial censoring time on the training set, leading to a better calibration in the long term, and an overestimation of short-term event probability. For most submitted runs, discrimination and calibration improved across all sub-tasks as dynamic variables were considered (M0 vs. M6), suggesting that, for ALS, collecting dynamic features for the initial six months, instead of using baseline features only, is likely to lead to better performance in describing disease progression. Results achieved by iDPP:CLEF participants in terms of BS is available in Appendix C.

Overall, for Task 1, runs submitted by the CampBioMed team were the best performing across the board; meanwhile, runs submitted by the BioHit team led to the lowest discrimination, but still yielded acceptable calibration at long PHs. Finally, the SBB and LIG GETALP teams obtained comparable results when considering runs obtained using all the information available in the first 6 months (M6 runs); meanwhile, when using only the information available at time 0 (M0 runs), runs submitted by the SBB team showed worse discrimination than those submitted by the LIG GETALP team.

7.2. Pilot Task 2: Predicting Time of Impairment for ALS

Appendix D and Appendix E contain respectively the specificity-recall plots for the time interval and the label prediction approaches.

The graph shows the specificity on the x-axis (from 1 to 0, left to right), and the recall on the y-axis (from 0 to 1, bottom to top). The ideal classifier would have specificity = 1 and recall = 1, and would therefore be located in the upper left corner: as a general guidance, the closer a run to the upper left corner, the better the classification obtained.

In all graphs, the synthetic runs with constant predictions (set to the minimum or maximum time interval) are located in the corners of the plot, depending on the time interval. As expected, the 100 runs with 70% correlation always form a cloud in the upper left corner, while the 200 randomly generated runs, 100 with the same distribution and 100 with the inverse distribution always remain in the lower left sector, with $1 > specificity > 0.5$ and $0 > recall > 0.5$.

Across the different time intervals and the three sub-tasks, there is no homogeneity of performance among the participants. Overall, the two best performing teams were CompBioMed and FCOOL, outperformed in a few cases by SBB team.

There is a great variety of results across sub-tasks and observation times. It is always true that predicting the label *none* lead to a higher recall, compared to the prediction of *NIV*, *PEG*, and *death*. The best performing teams were, once again, CompBioMed and FCOOL.

Finally, we evaluate the AbsDist for all runs submitted for sub-tasks a, b and c. As expected, the *max_interval* run led to the worst result across all tasks, as most subjects' true time interval is smaller than the maximum one. Runs *random_interval*, *min_interval*, and *inverse_corr_interval* led to comparable distance values. Runs submitted by the BioHit team had AbsDist values comparable with the synthetic run *interval_18_24*, suggesting that their models might predict the average time interval for most subjects in all sub-tasks. All runs submitted by the other teams significantly outperformed the aforementioned synthetic runs with the CampBioMed and SBB teams leading the pack when considering sub-task a, and sub-tasks b and c, respectively.

Finally, the *corr_interval* synthetic run led to the smallest AbsDist value. Note, however, that this run was included only as an arbitrary reference to represent an excellent model, and its distance value was not strictly expected to be reached by any participant. AbsDist computed for all runs submitted can be found in Appendix F.

Predicting the correct event time interval proved to be a challenge for all teams, especially in terms of recall. However, almost all teams were able to obtain good AbsDist values as, on average, their models, despite not being able to precisely identify the correct time interval, tended to predict an interval that was immediately before or after the true one.

As observed for Task 1, runs performed better when considering all the information available in the first 6 months (M6 runs) rather than only the information available at time 0 (M0 runs).

7.3. Approaches

In this section, we provide a short summary of the approaches adopted by participants in iDPP:CLEF. There are two separate sub-sections, one for Task 1 and 2 focused on ALS progression prediction and the other for Task 3, on *eXplainable AI (XAI)* approaches for such kind of algorithms.

Task 1 and 2

BoiHIT explored the use of logistic regression, random forest classifiers, XGBoost, and LightGBM. Decision trees and boosting approaches were preferred due to their ability to deal with both categorical and numerical/continuous features and the interpretability they offer. Even if LightGBM was the model with the best performance, BoiHIT found out that this kind of approaches might not be appropriate for time dependent problems and that time to event analysis methods, such as survival analysis, might yield better results.

CompBioMed [4] considered three main approaches. The simplest one consisted on fitting a standard survival predictor separately for each event as outlined above for independent events, called Naive Multiple Event Survival (NMES). Another was the recently developed Deep Survival Machine (DSM), based on deep learning and capable of handling competing risks. Finally, they also proposed a time-aware classifier ensemble method, that also handles competing risks, called Time-Aware Classifier Ensemble (TACE). All the above approaches achieved comparable performance among them. Only the TACE models appeared to be slightly worse than the rest in when using 6 months of data. Moreover, no clear advantage of the DSM models, that specifically handles competing risks, was observed with respect to the NMES models, which treat all events, as if they were independent.

FCOOL [5] proposes a hierarchical approach, with a first-stage event prediction, followed by specialized models predicting the time window to a particular event. The procedure is three-fold: first, it creates patient snapshots based on clustering with constraints, thus organizing patient records in an efficient manner. Second, it uses a pattern-based approach that incorporates recent advances on temporal pattern mining to the context of classification. This approach performs end-stage event prediction while allowing the entire patient's medical history to be considered. Finally, exploiting the predictions from the previous step, specialized models are learned using the original features to predict the time window to an event. This two-stage prediction approach

aimed to promote homogeneity and lessen the impact of class imbalance, in comparison to performing one single multilabel task.

LIG GETALP [7] employed Cox's proportional hazards model to the task of ranking the risk of impairment, using the gradient boosting learning strategy The output of the time-independent part of the survival function calculated by the gradient boosting survival analysis method is then mapped to the interval (0, 1), via a sigmoid function. To estimate the time-to-event, LIG GETALP used a regression model based on Accelerated Gradient Boosting (AGB). This being a standard regression model, it does not take censoring into account and Mannion et al. uses class predictions based on the Task 1 survival model to "censor" the time-to-event predictions.

SBB [8] considered three survival analysis methods, namely: Cox, SSVM, and RSF. They were chosen to represent a broad spectrum of baseline models including parametric (SSVM), semiparametric (Cox), linear (Cox, SSVM), and nonlinear (RSF) models. The Cox model and the RSF can only output risk scores, which can be used to address Task 1 by ranking ALS patients according to their risk of impairment, but do not provide a straightforward solution to predicting Task 2's time of impairment. To extend these approaches to Task 2, the predicted time of impairment for a given patient was selected as the median predicted time to impairment, i.e., the time at which the estimated survival function crossed the 0.5 threshold. Instead, the SSVM can be used either as a ranker or a time regressor depending on how the risk ratio hyperparameter is set during model training. Here, the SSVM was initially trained as a time regressor to address Task 2 directly. Then, its predicted times were converted into risk scores in the range [0-1], as requested by the challenge rules, via Platt scaling.

Task 3

Nunes et al. [6] proposes a novel approach that generates semantic similarity-based explanations for patient-level predictions. The underlying idea is to explain the prediction for one patient by considering aspect-oriented semantic similarity with other relevant patients based on the most important features used by ML approaches or selected by users. To build rich and easy to understand semantic-similarity based explanations, Nunes et al. developed five steps: (1) the enrichment of the Brainteaser Ontology [14] through integration of other biomedical ontologies; (2) the semantic annotation of patients (if not already available); (3) the similarity calculation between patients; (4) selection of the set of patients to explain a specific prediction; and (5) the visualization of the generated similarity-based explanations.

Buonocore et al. [15] trained a set of 4 well-known classifiers to predict death occurrence: Gradient Boosting (using XGB implementation), Random Forest, Logistic Regression and Multilayer perceptron. For the XAI methods Buonocore et al. focused our attention on three different methods for post-hoc, model-agnostic, local explainability, selecting SHAP, LIME and AraucanaXAI. Then, Buonocore et al. evaluated and compared XAI approaches in terms of a set of metrics defined in previous research on XAI in healthcare: *identity*: if there are two identical instances, they must have the same explanations; *fidelity*: concordance of the predictions between the XAI surrogate model and the original ML model; *separability*: if there are 2 dissimilar instances, they must have dissimilar explanations; *time*: average time required by the XAI method to output an explanation across the entire test set. The quantitative evaluation of the three different XAI methods did not reveal definitive superior performance of one of the approaches, albeit SHAP

seems to be the better overall performing algorithm. However the explainability evaluation metrics are not all that is needed to thoroughly assess the multifaceted construct of what constitutes a "good" explanation in XAI in healthcare.

8. Conclusions and Future Work

iDPP:CLEF is a new pilot activity focusing on predicting the temporal progression of ALS and on the explainability of the AI algorithms for such prediction.

We developed 3 datasets containing anonymized patient data from two medical institutions, one in Turin and the other in Lisbon, for the prediction of NIV, PEG, or death.

Out of 43 registered participants, 5 managed to submit a total of 120 runs, evenly spread across the offered tasks. Participants adopted a range of approaches, including various types of survival analysis, also using deep learning techniques. For the XAI of the prediction algorithms they used both semantic-similarity based techniques and state-of-art post-hoc and model-agnostic XAI approaches.

For this initial iteration of the lab, iDPP:CLEF focus on ALS progression prediction. Possible, future cycles will be extended to *Multiple Sclerosis (MS)*, another chronic disease, impairing neurological functions. Moreover, we plan to extend the datasets to also include data from environmental sensor, e.g. concerning pollution.

Acknowledgments

The work reported in this paper has been partially supported by the BRAINTEASER[7] project (contract n. GA101017598), as a part of the European Union's Horizon 2020 research and innovation programme.

References

[1] R. Küffner, N. Zach, R. Norel, J. Hawe, D. Schoenfeld, L. Wang, G. Li, L. Fang, L. Mackey, O. Hardiman, M. Cudkowicz, A. Sherman, G. Ertaylan, M. Grosse-Wentrup, T. Hothorn, J. van Ligtenberg, J. H. Macke, T. Meyer, B. Schölkopf, L. Tran, R. Vaughan, G. Stolovitzky, M. L. Leitner, Crowdsourced analysis of clinical trial data to predict amyotrophic lateral sclerosis progression, Nature Biotechnology 33 (2015) 51–57.

[2] J. M. Cedarbaum, N. Stambler, E. Malta, C. Fuller, D. Hilt, B. Thurmond, A. Nakanishi, The ALSFRS-R: a revised ALS functional rating scale that incorporates assessments of respiratory function, Journal of the Neurological Sciences 169 (1999) 13–21.

[3] A. Chio, G. Logroscino, O. Hardiman, R. Swingler, D. Mitchell, E. Beghi, B. G. Traynor, E. Consortium, et al., Prognostic factors in als: a critical review, Amyotrophic lateral sclerosis 10 (2009) 310–323.

[4] C. Pancotti, G. Birolo, T. Sanavia, C. Rollo, P. Fariselli, Multi-Event Survival Prediction for Amyotrophic Lateral Sclerosis, in: [16], 2022.

[7] https://brainteaser.health/

[5] R. Branco, D. Soares, A. S. Martins, E. Auletta, E. N. Castanho, S. Nunes, F. Serrano, R. T. Sousa, C. Pesquita, S. C. Madeira, H. Aidos, Hierarchical Modelling for ALS Prognosis: Predicting the Progression Towards Critical Events, in: [16], 2022.
[6] S. Nunes, R. T. Sousa, F. Serrano, R. Branco, D. F. Soares, A. S. Martins, E. Auletta, E. N. Castanho, S. C. Madeira, H. Aidos, C. Pesquita, Explaining Artificial Intelligence Predictions of Disease Progression with Semantic Similarity, in: [16], 2022.
[7] A. Mannion, T. Chevalier, D. Schwab, L. Goeuriot, Predicting the Risk of & Time to Impairment for ALS patients, in: [16], 2022.
[8] I. Trescato, A. Guazzo, E. Longato, E. Hazizaj, C. Roversi, E. Tavazzi, M. Vettoretti, B. Di Camillo, Baseline Machine Learning Approaches To Predict Amyotrophic Lateral Sclerosis Disease Progression, in: [16], 2022.
[9] J. Harrell, Frank E., R. M. Califf, D. B. Pryor, K. L. Lee, R. A. Rosati, Evaluating the Yield of Medical Tests, JAMA 247 (1982) 2543–2546.
[10] J. A. Hanley, B. J. McNeil, The meaning and use of the area under a receiver operating characteristic (roc) curve., Radiology 143 (1982) 29–36. PMID: 7063747.
[11] G. W. Brier, Verification of Forecasts Expressed in Terms of Probability, Monthly Weather Review 78 (1950) 1–3.
[12] M. J. Pencina, R. B. D'Agostino, Overall c as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation, Statistics in Medicine 23 (2004) 2109–2123.
[13] E. Longato, M. Vettoretti, B. Di Camillo, A practical perspective on the concordance index for the evaluation and selection of prognostic time-to-event models, Journal of Biomedical Informatics 108 (2020) 103496:1–103496:9.
[14] M. Bettin, G. M. Di Nunzio, D. Dosso, G. Faggioli, N. Ferro, N. Marchetti, G. Silvello, Deliverable 9.1 – Project ontology and terminology, including data mapper and RDF graph builder, BRAINTEASER, EU Horizon 2020, Contract N. GA101017598. https://brainteaser.health/, 2021.
[15] T. M. Buonocore, G. Nicora, A. Dagliati, E. Parimbelli, Evaluation of XAI on ALS 6-months mortality prediction, in: [16], 2022.
[16] G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), CLEF 2022 Working Notes, CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, 2022.

A. Pilot task 1: C-index

Figures 16 to 18 show the C-index with its 95% confidence intervals computed for all runs submitted by participants and for the 100 random classifiers (last row) for sub-tasks a, b, and c, respectively.

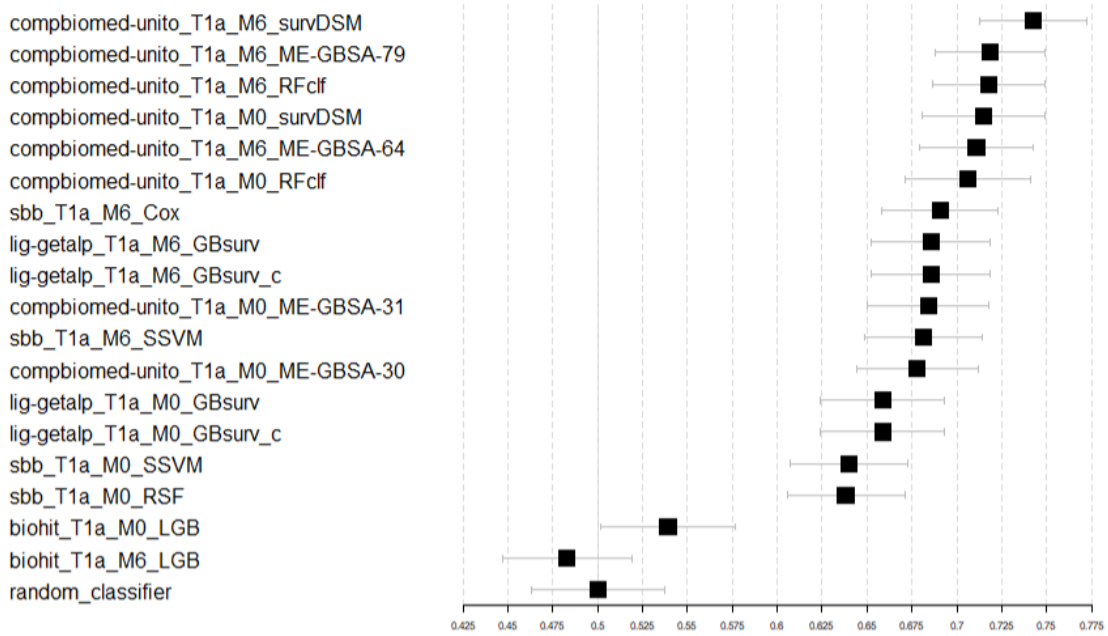


Figure 16: Sub-task a C-index computed for all submitted runs. The bars in the plot show the 95% confidence intervals. The average C-index of 100 random classifiers is reported in the last row.

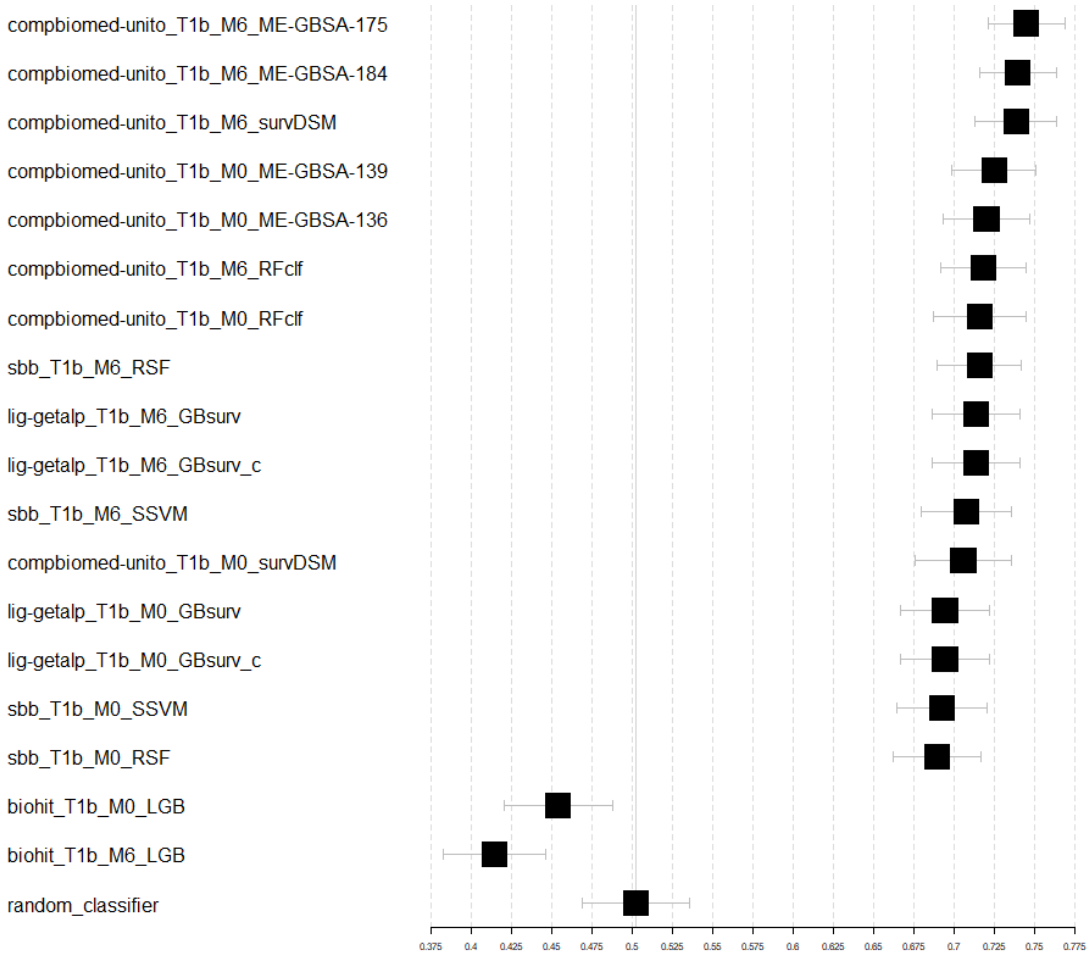


Figure 17: Sub-task b C-index computed for all submitted runs. The bars in the plot show the 95% confidence intervals. The average C-index of 100 random classifiers is reported in the last row.

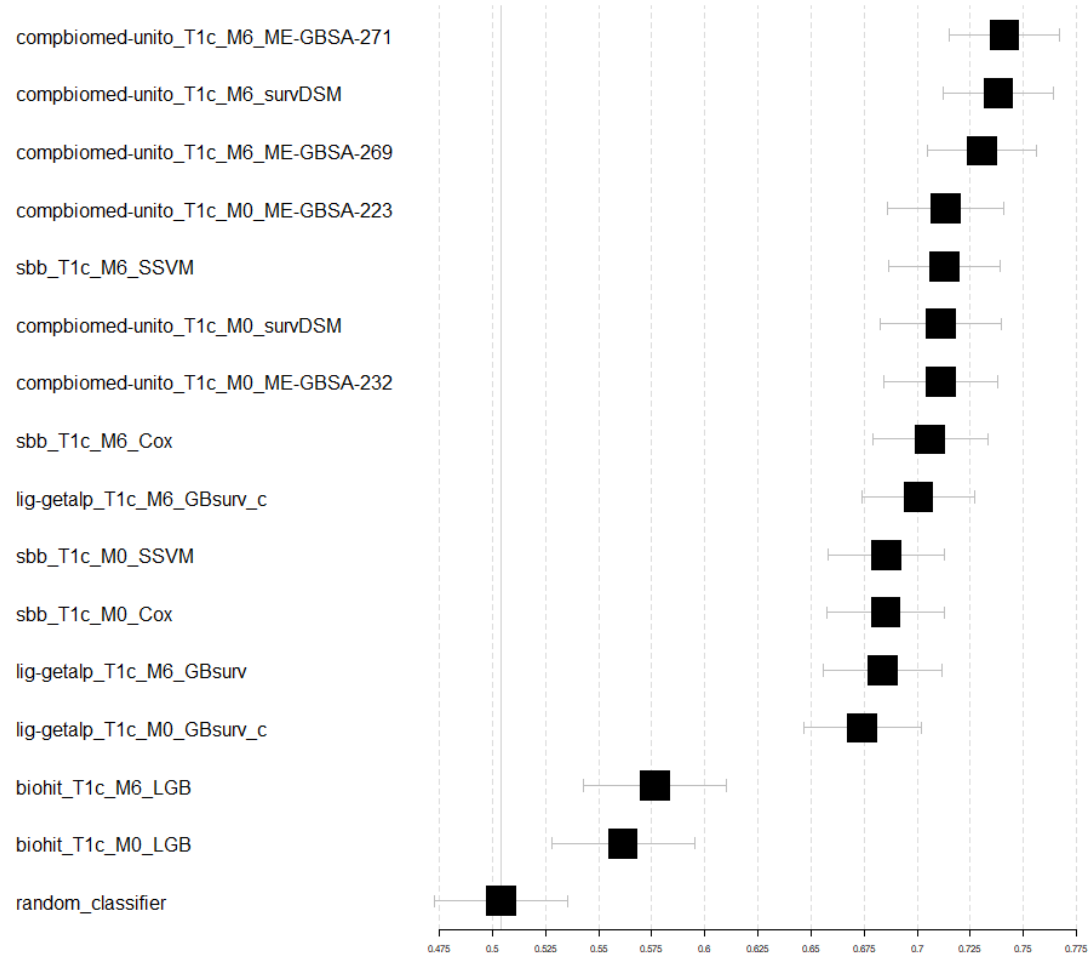


Figure 18: Sub-task c C-index computed for all submitted runs. The bars in the plot show the 95% confidence intervals. The average C-index of 100 random classifiers is reported in the last row.

B. Pilot task 1: AUROC

Figures 19 to 39 show the AUROC with its 95% confidence intervals computed for all runs submitted for sub-tasks a, b, and c at all the considered PHs (12, 18, 24, 30, 36, 48, and 60 months). The average AUROC of the 100 random classifiers is reported in the the last row of the figures.

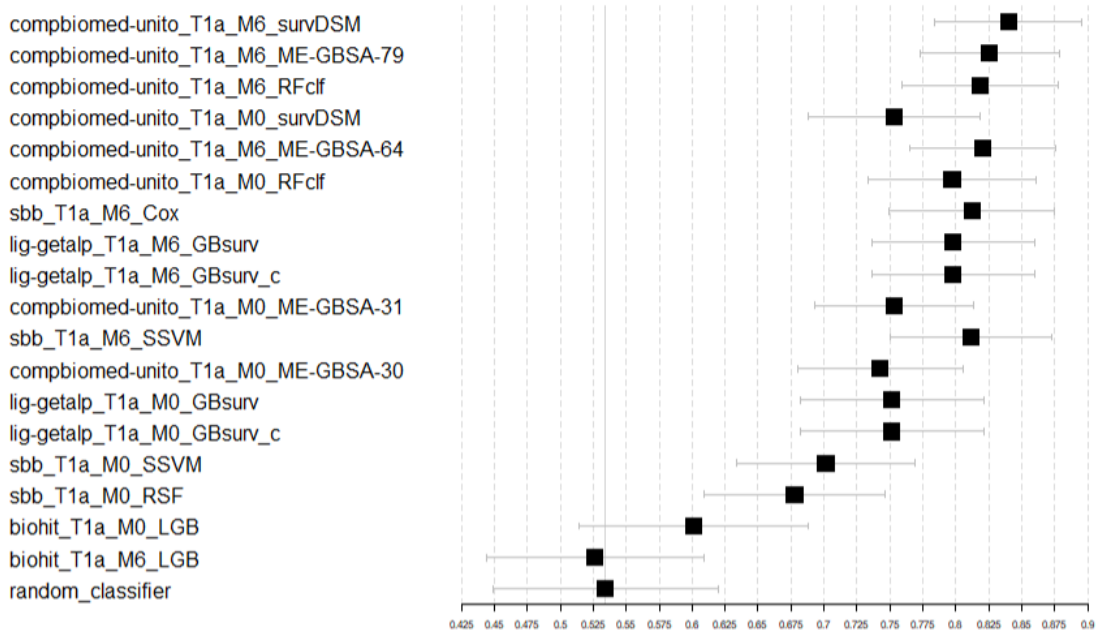


Figure 19: Sub-task a AUROC computed for all submitted runs with a 12-months PH. The bars in the plot show the 95% confidence intervals. The average 12-months AUROC of 100 random classifiers is reported in the last row.

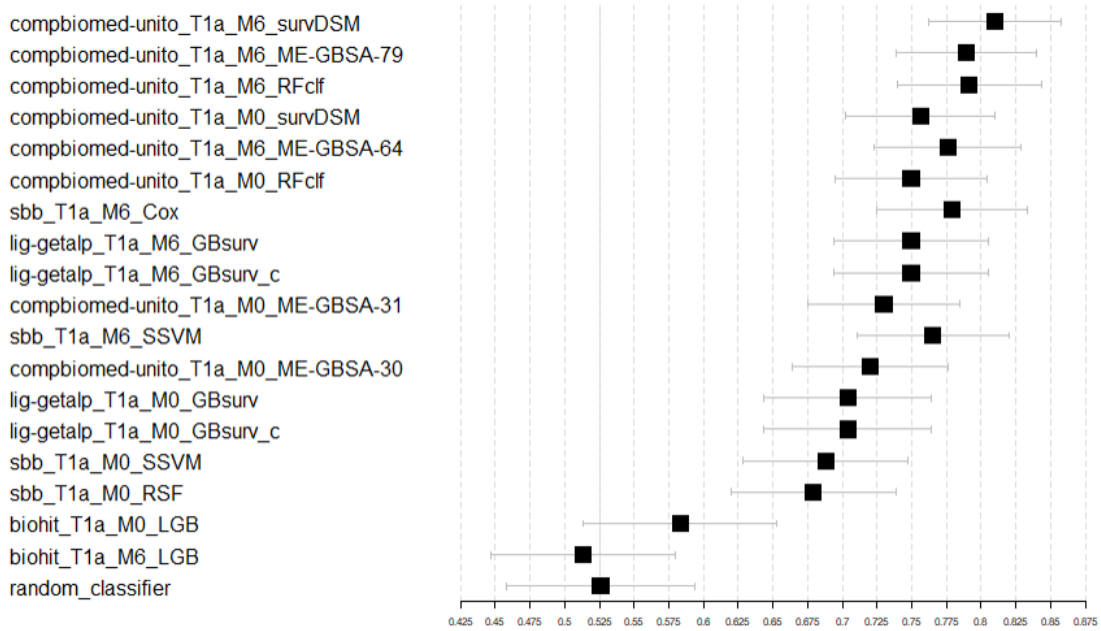


Figure 20: Sub-task a AUROC computed for all submitted runs with a 18-months PH. The bars in the plot show the 95% confidence intervals. The average 18-months AUROC of 100 random classifiers is reported in the last row.

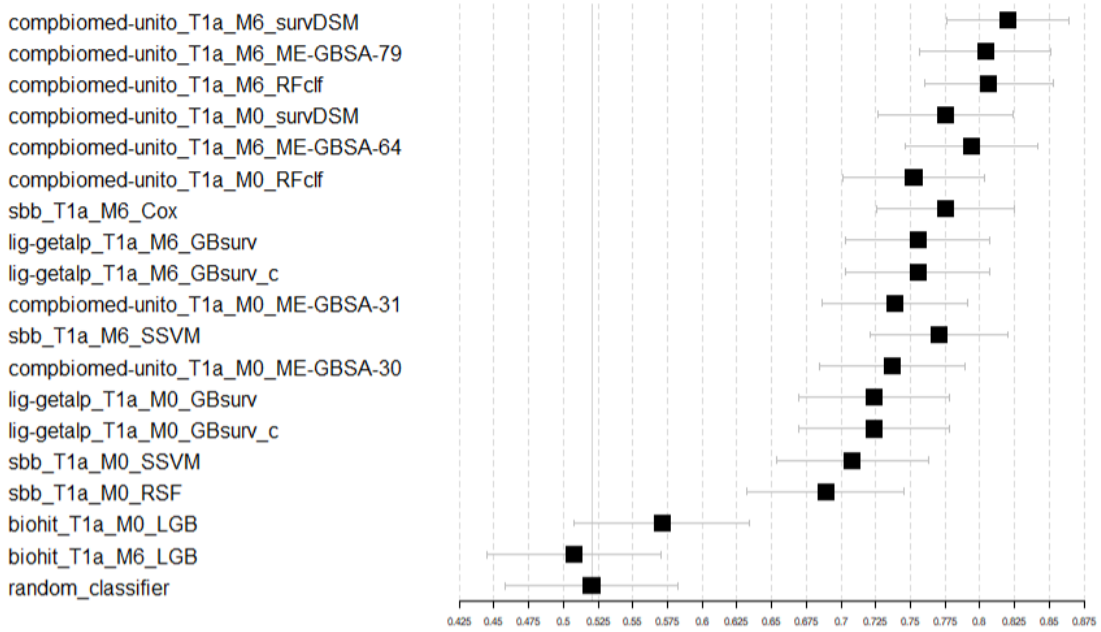


Figure 21: Sub-task a AUROC computed for all submitted runs with a 24-months PH. The bars in the plot show the 95% confidence intervals. The average 24-months AUROC of 100 random classifiers is reported in the last row.

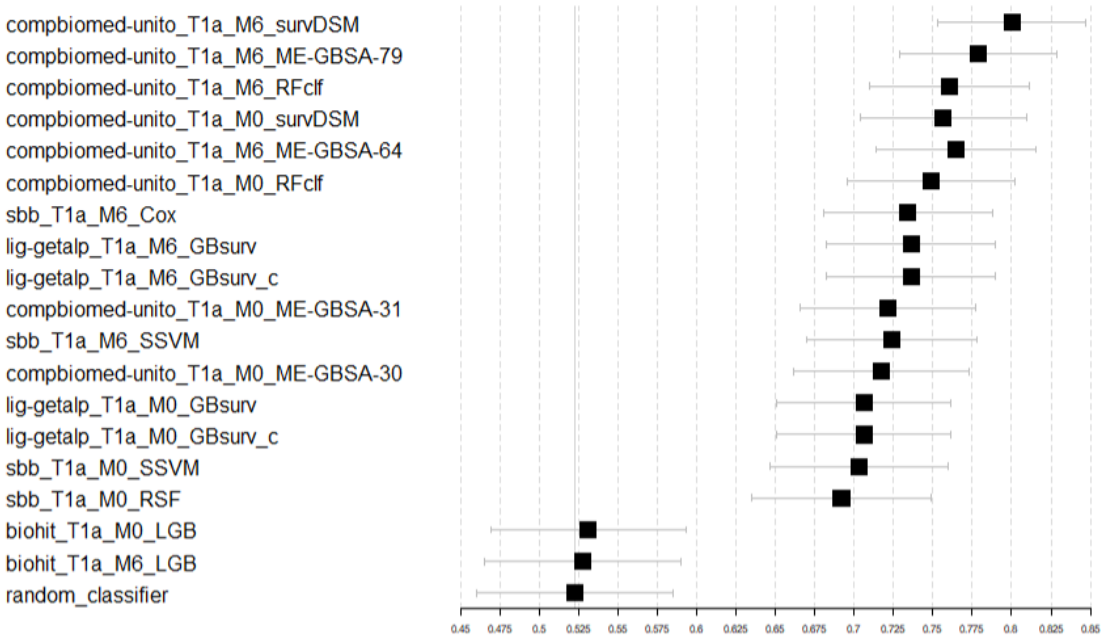


Figure 22: Sub-task a AUROC computed for all submitted runs with a 30-months PH. The bars in the plot show the 95% confidence intervals. The average 30-months AUROC of 100 random classifiers is reported in the last row.

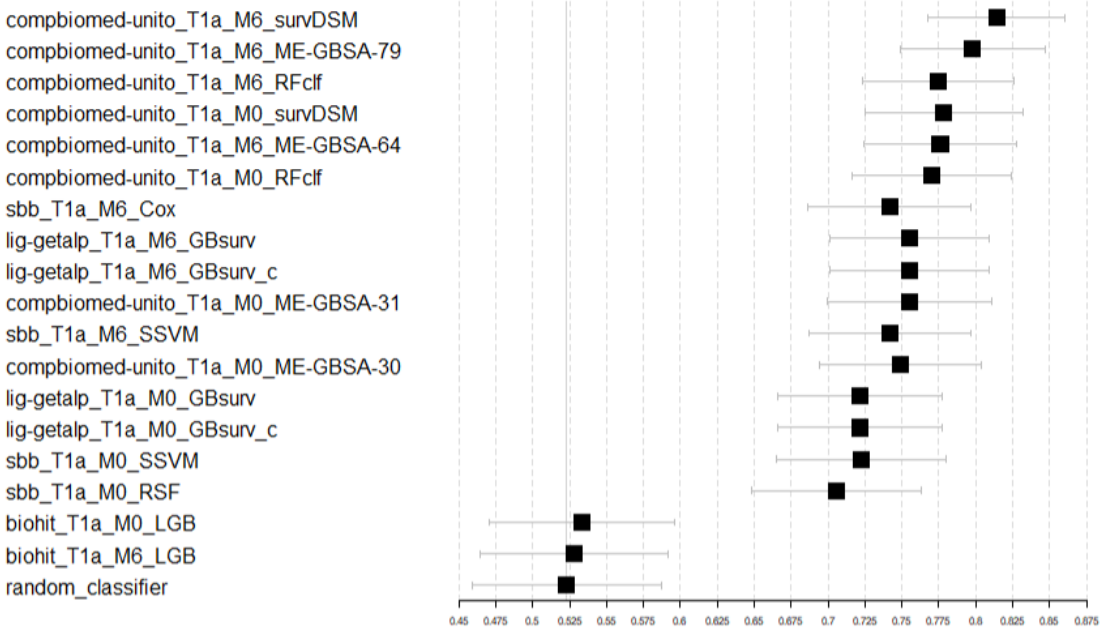


Figure 23: Sub-task a AUROC computed for all submitted runs with a 36-months PH. The bars in the plot show the 95% confidence intervals. The average 36-months AUROC of 100 random classifiers is reported in the last row.

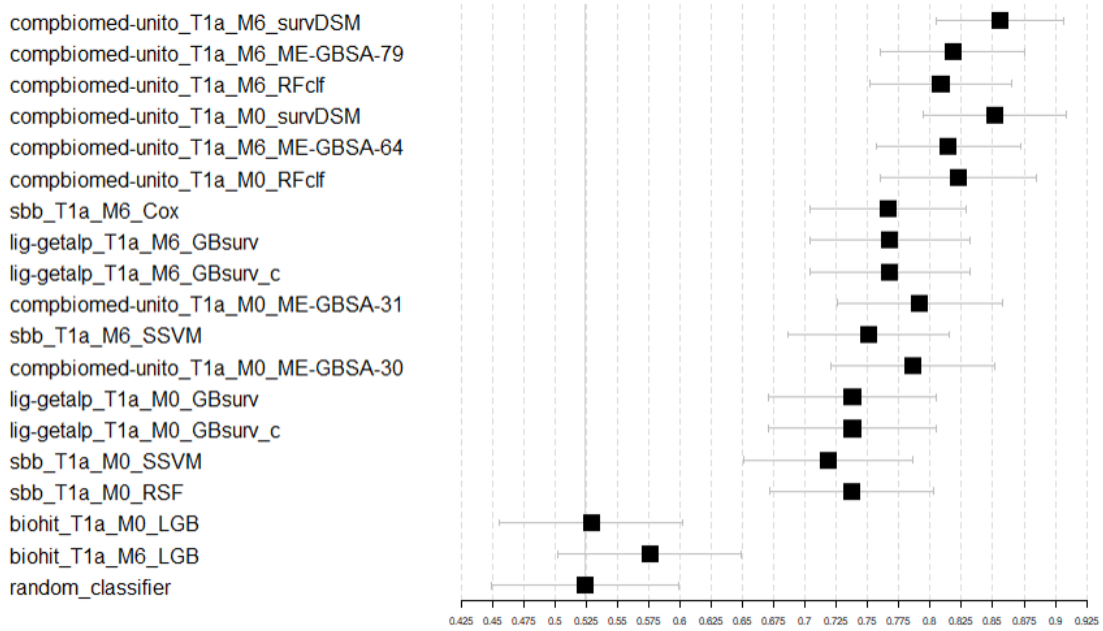


Figure 24: Sub-task a AUROC computed for all submitted runs with a 48-months PH. The bars in the plot show the 95% confidence intervals. The average 48-months AUROC of 100 random classifiers is reported in the last row.

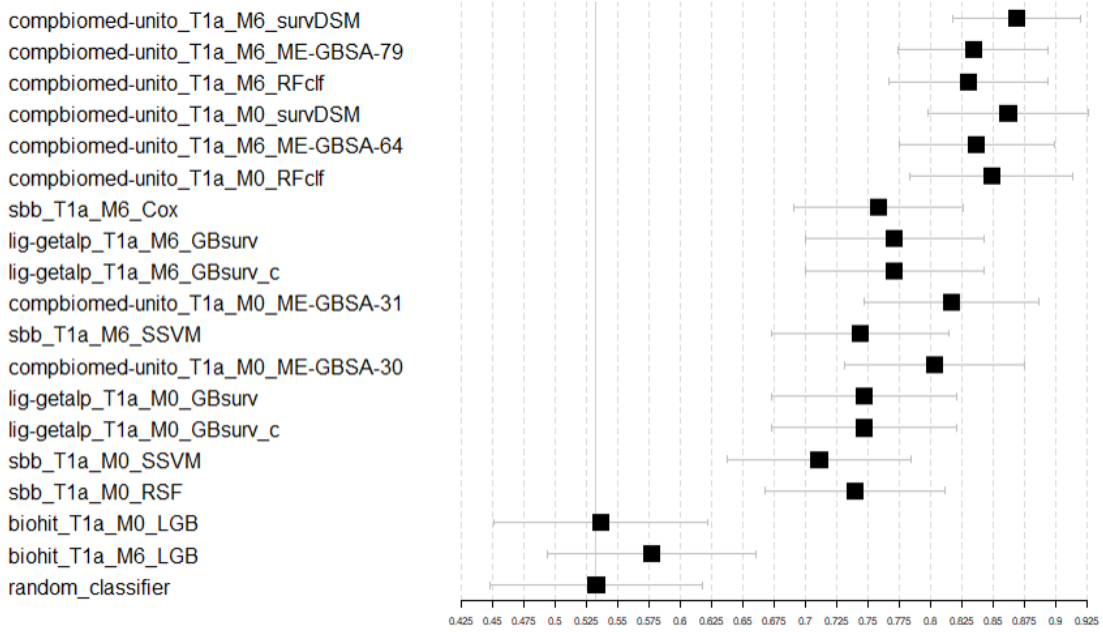


Figure 25: Sub-task a AUROC computed for all submitted runs with a 60-months PH. The bars in the plot show the 95% confidence intervals. The average 60-months AUROC of 100 random classifiers is reported in the last row.

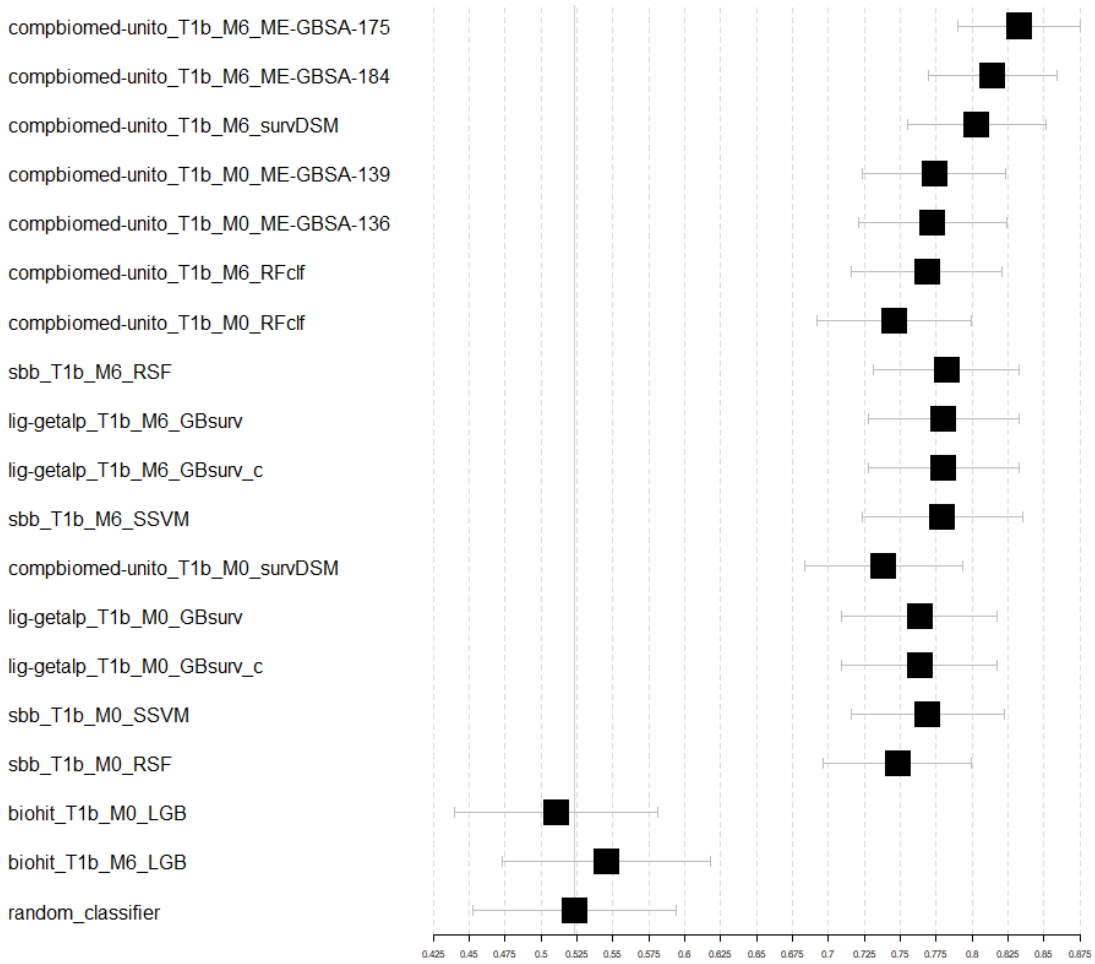


Figure 26: Sub-task b AUROC computed for all submitted runs with a 12-months PH. The bars in the plot show the 95% confidence intervals. The average 12-months AUROC of 100 random classifiers is reported in the last row.

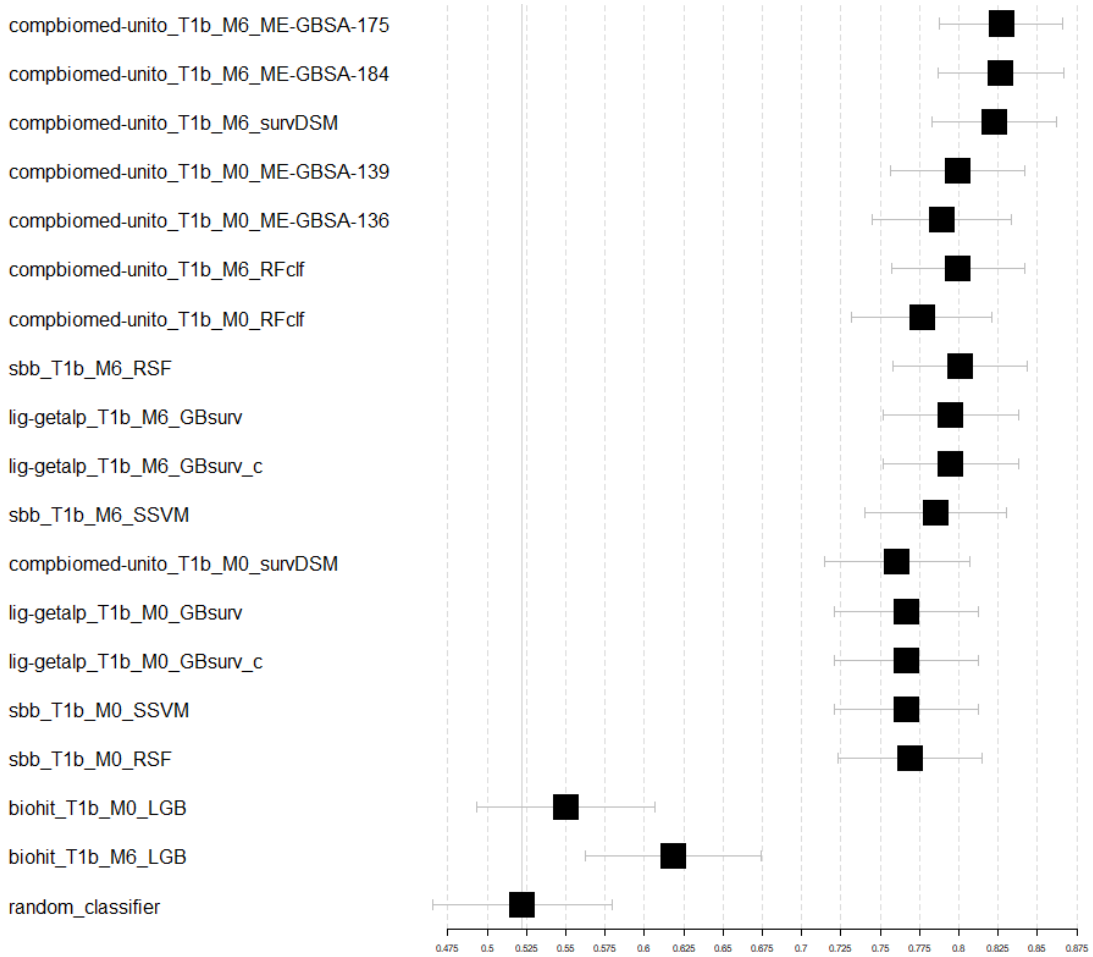


Figure 27: Sub-task b AUROC computed for all submitted runs with a 18-months PH. The bars in the plot show the 95% confidence intervals. The average 18-months AUROC of 100 random classifiers is reported in the last row.

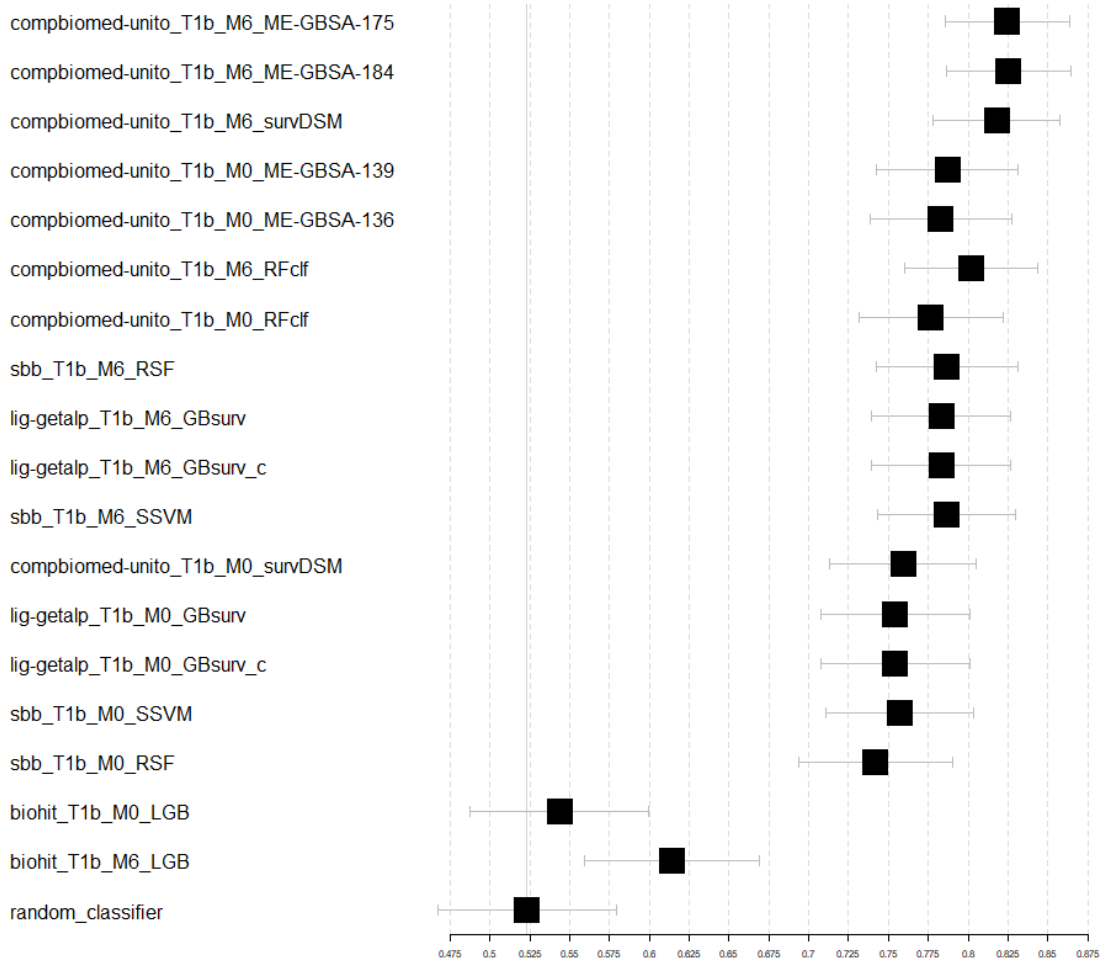


Figure 28: Sub-task b AUROC computed for all submitted runs with a 24-months PH. The bars in the plot show the 95% confidence intervals. The average 24-months AUROC of 100 random classifiers is reported in the last row.

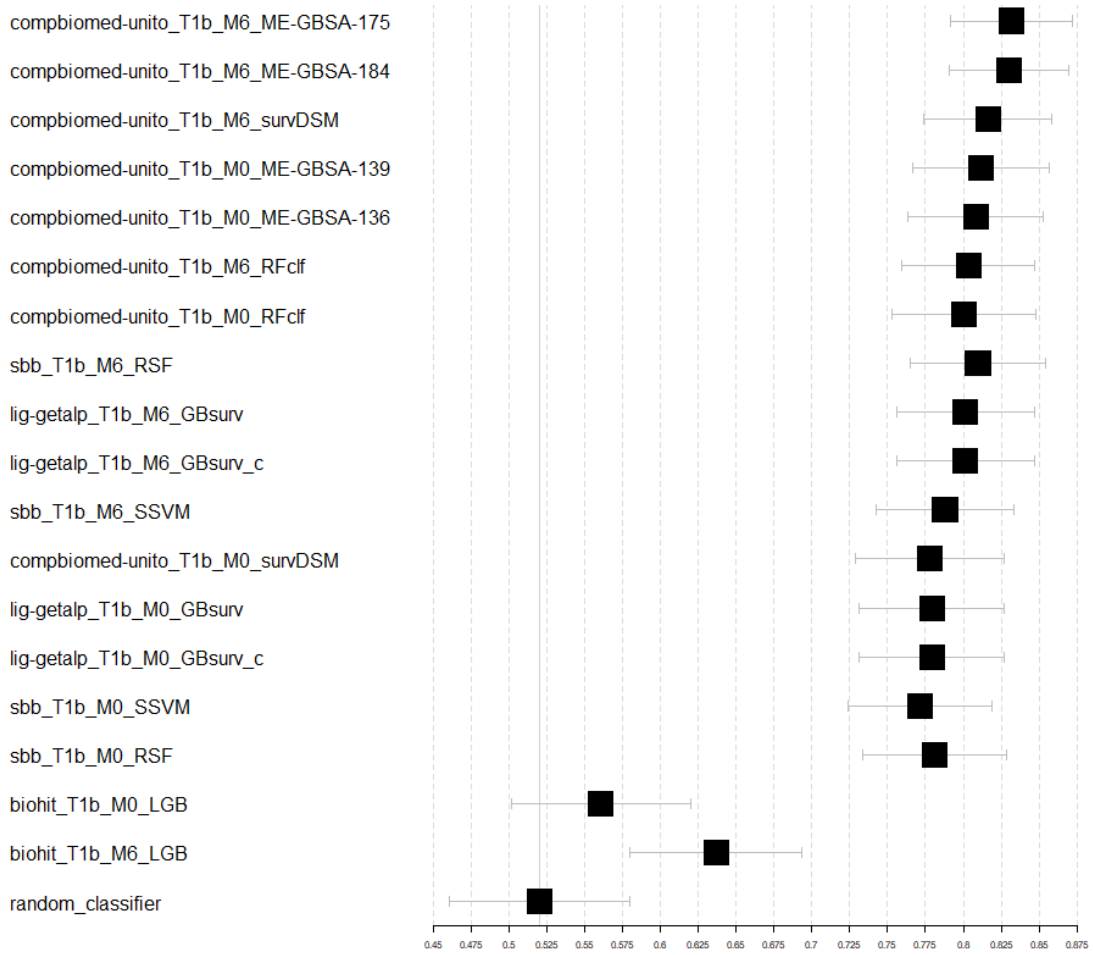


Figure 29: Sub-task b AUROC computed for all submitted runs with a 30-months PH. The bars in the plot show the 95% confidence intervals. The average 30-months AUROC of 100 random classifiers is reported in the last row.

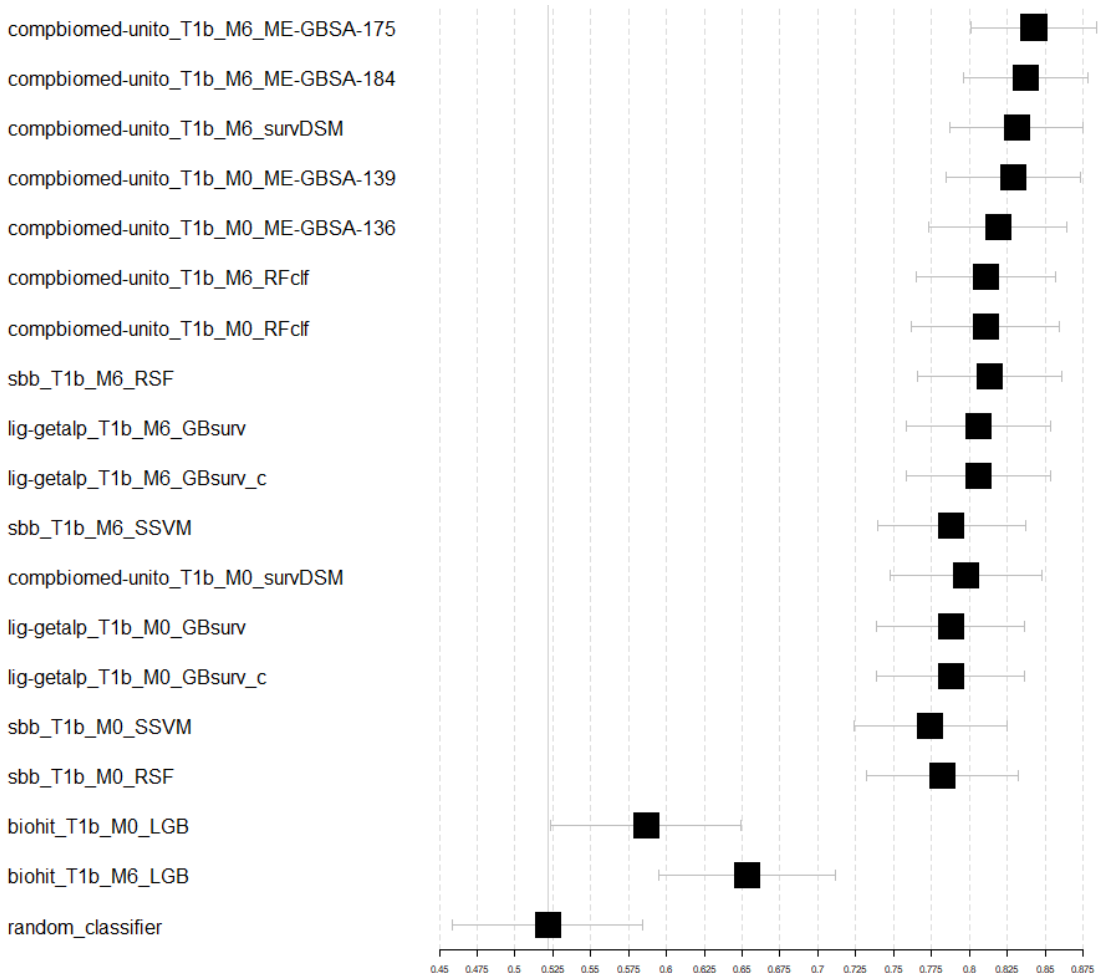


Figure 30: Sub-task b AUROC computed for all submitted runs with a 36-months PH. The bars in the plot show the 95% confidence intervals. The average 36-months AUROC of 100 random classifiers is reported in the last row.

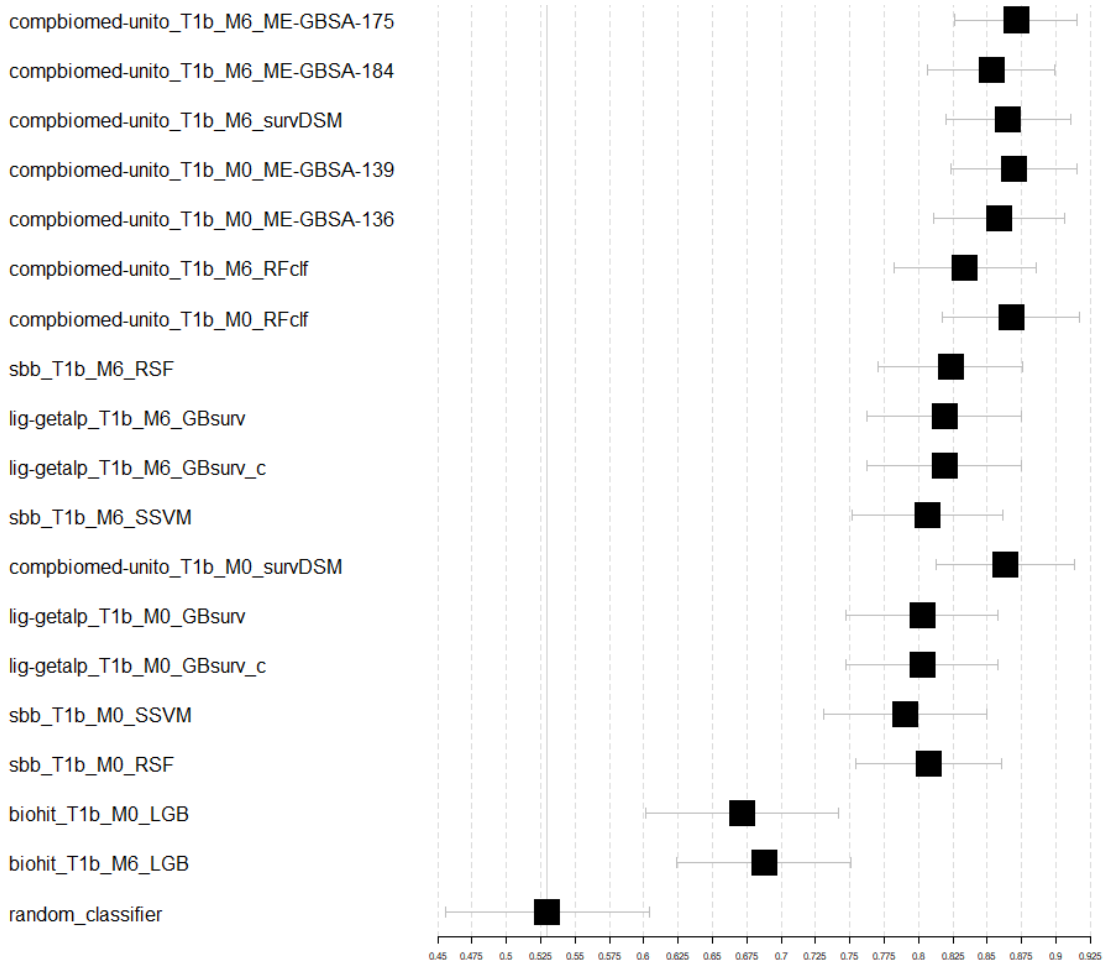


Figure 31: Sub-task b AUROC computed for all submitted runs with a 48-months PH. The bars in the plot show the 95% confidence intervals. The average 48-months AUROC of 100 random classifiers is reported in the last row.

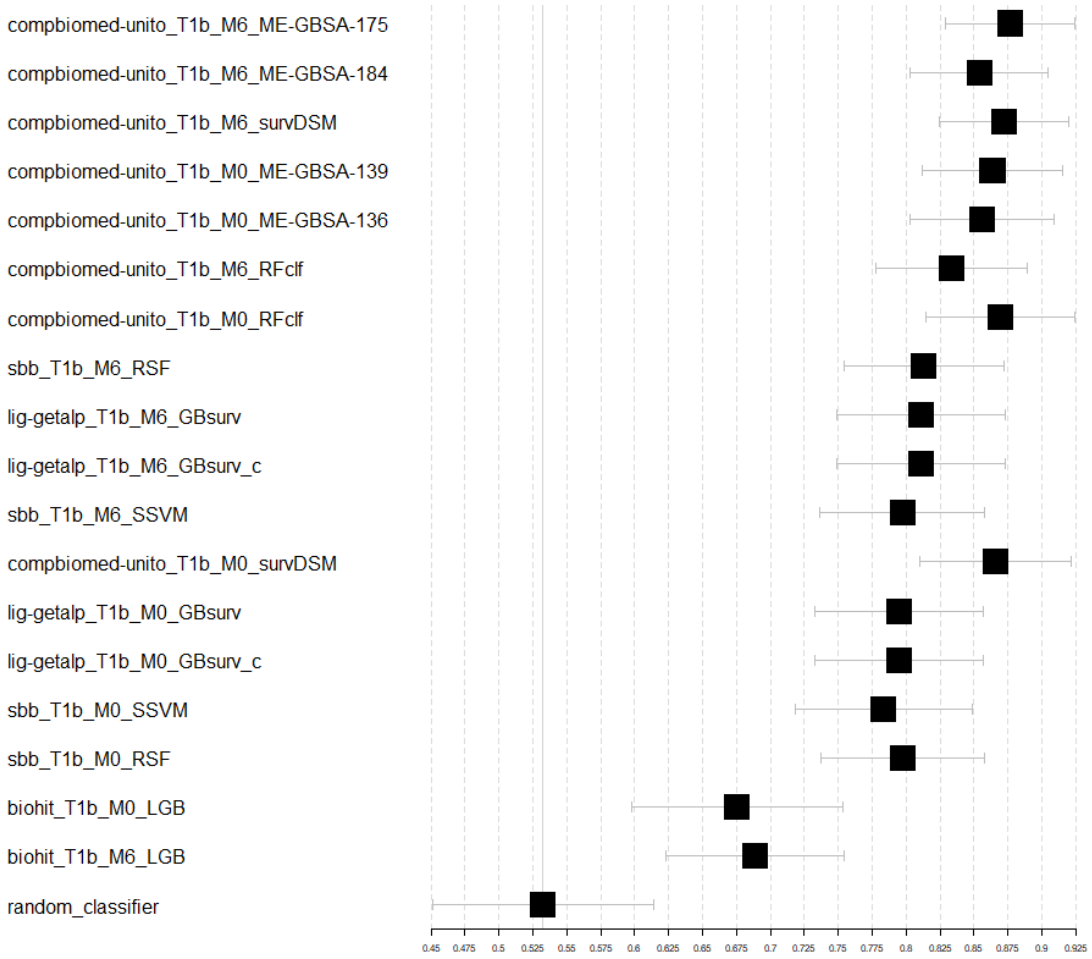


Figure 32: Sub-task b AUROC computed for all submitted runs with a 60-months PH. The bars in the plot show the 95% confidence intervals. The average 60-months AUROC of 100 random classifiers is reported in the last row.

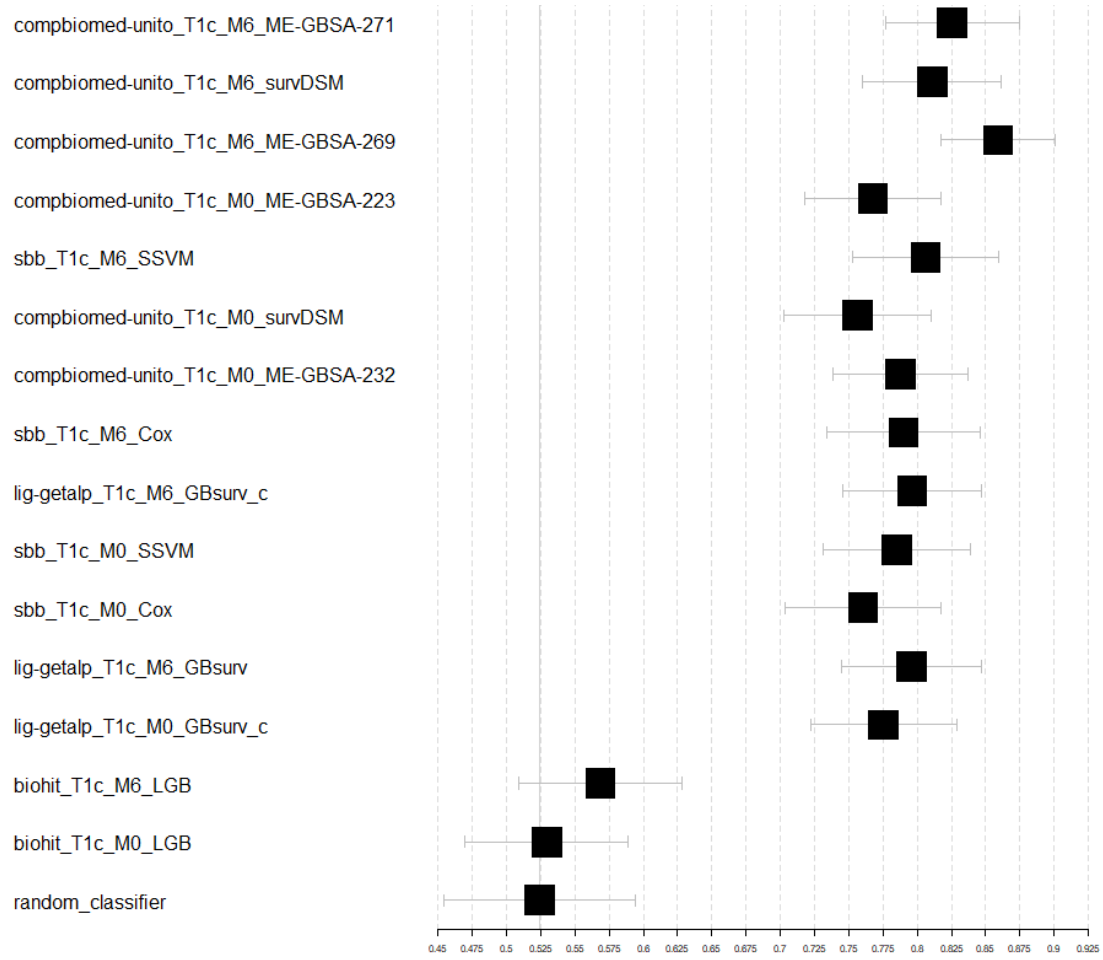


Figure 33: Sub-task c AUROC computed for all submitted runs with a 12-months PH. The bars in the plot show the 95% confidence intervals. The average 12-months AUROC of 100 random classifiers is reported in the last row.

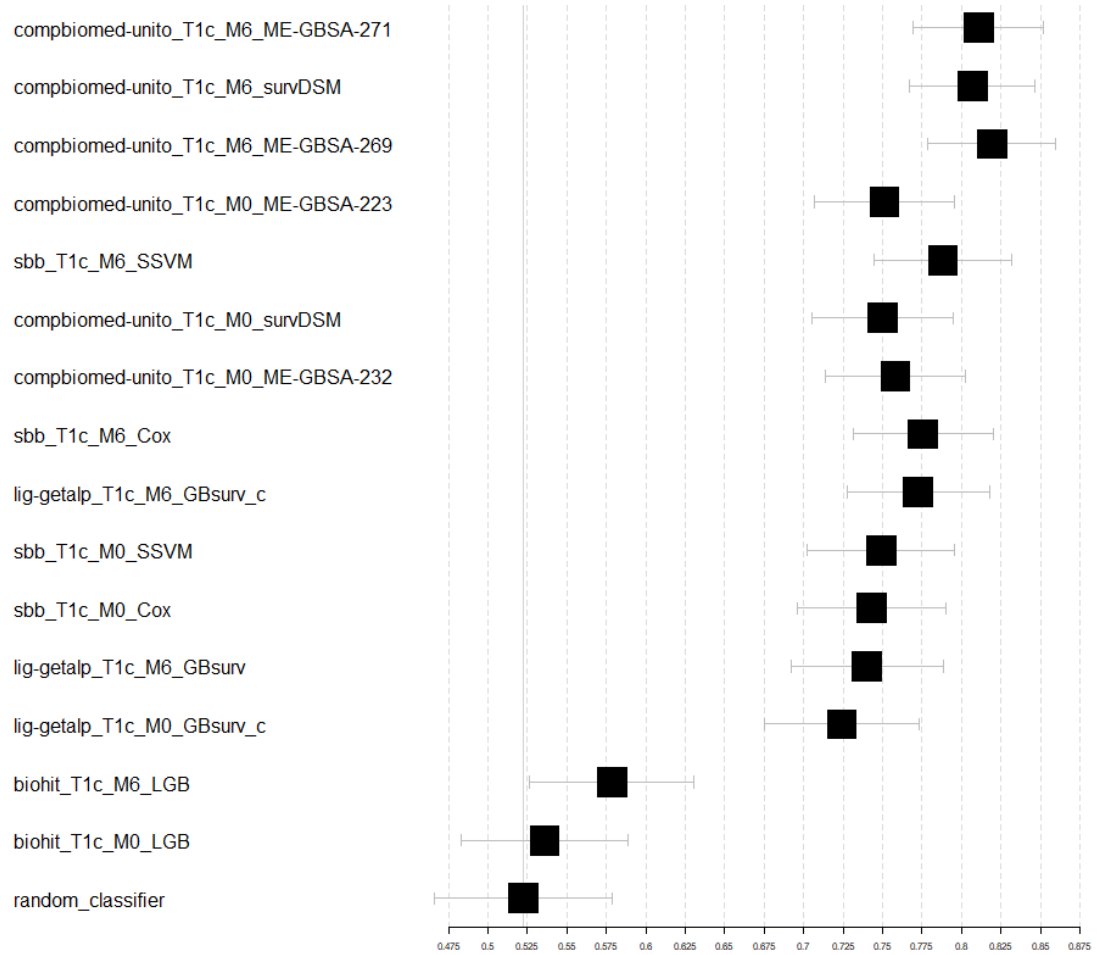


Figure 34: Sub-task c AUROC computed for all submitted runs with a 18-months PH. The bars in the plot show the 95% confidence intervals. The average 18-months AUROC of 100 random classifiers is reported in the last row.

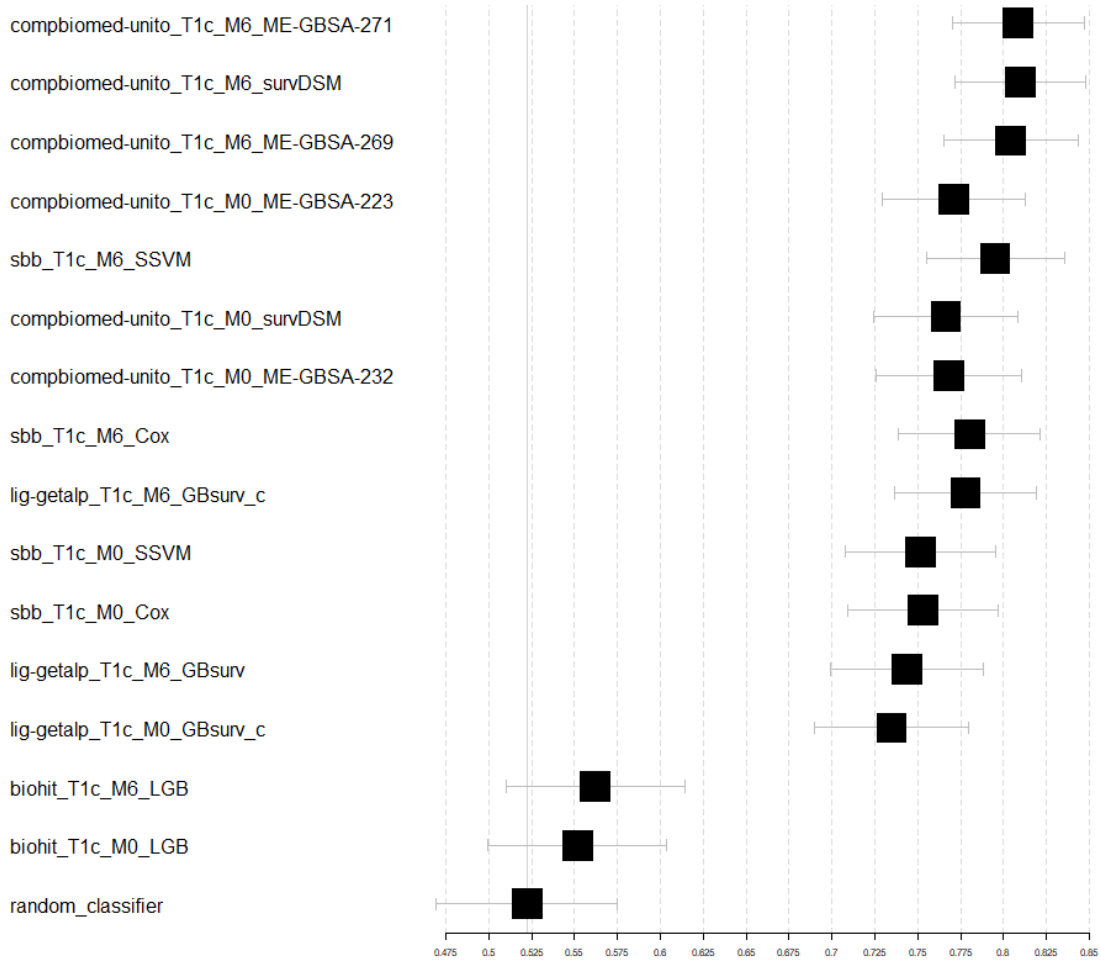


Figure 35: Sub-task c AUROC computed for all submitted runs with a 24-months PH. The bars in the plot show the 95% confidence intervals. The average 24-months AUROC of 100 random classifiers is reported in the last row.

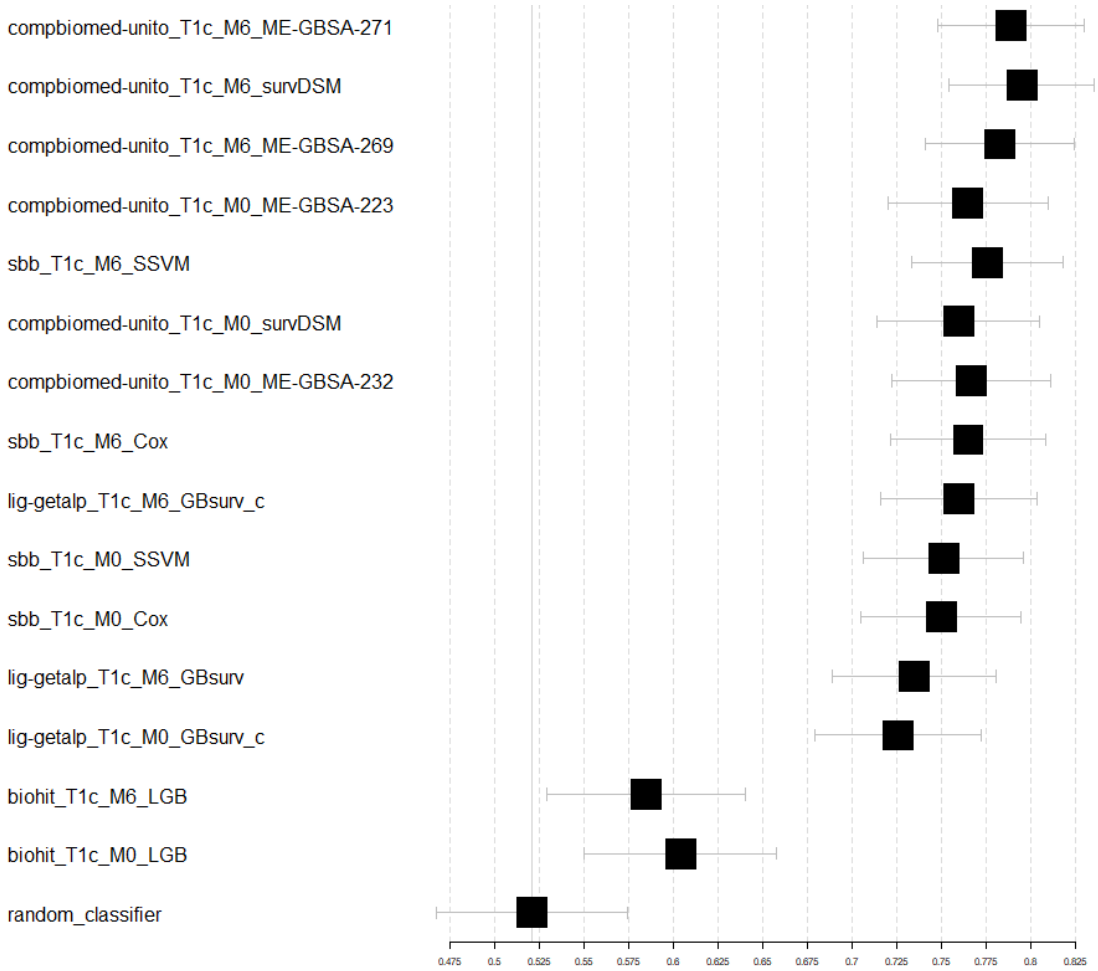


Figure 36: Sub-task c AUROC computed for all submitted runs with a 30-months PH. The bars in the plot show the 95% confidence intervals. The average 30-months AUROC of 100 random classifiers is reported in the last row.

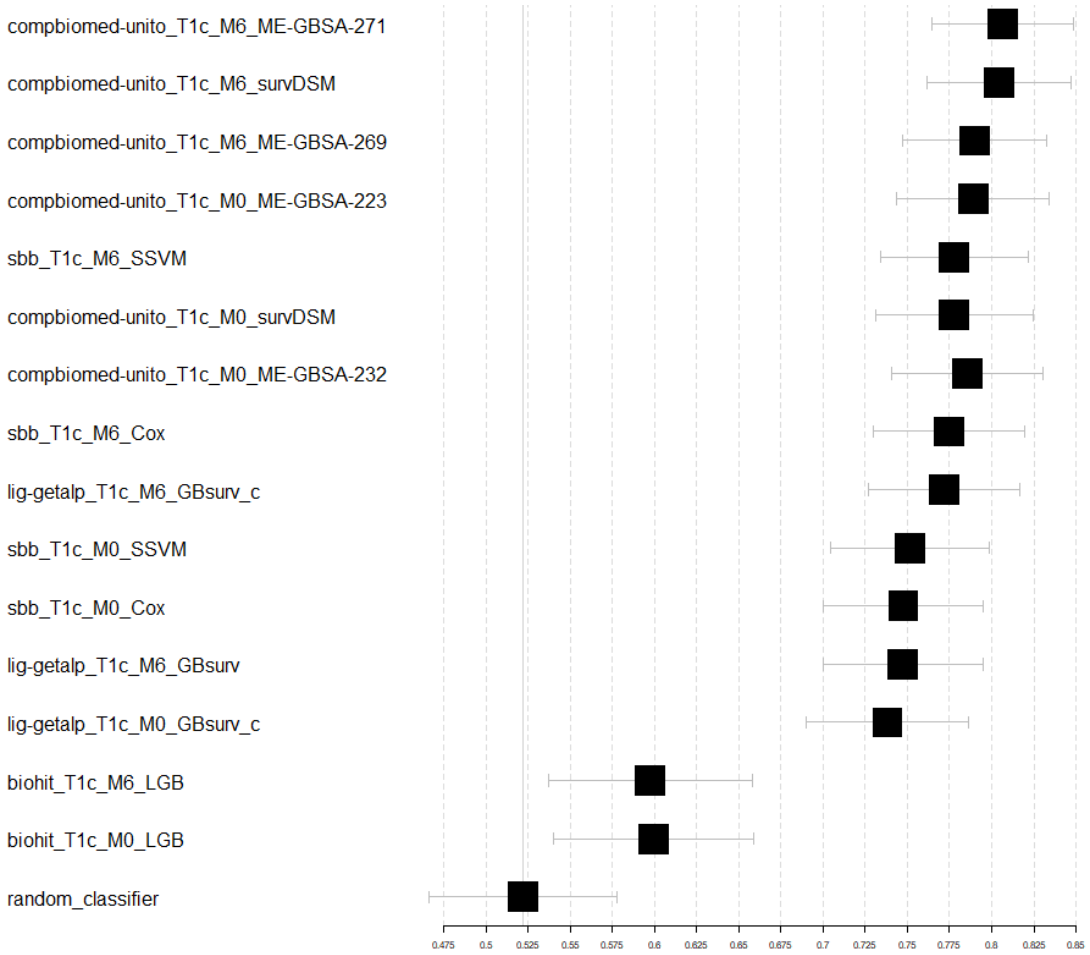


Figure 37: Sub-task c AUROC computed for all submitted runs with a 36-months PH. The bars in the plot show the 95% confidence intervals. The average 36-months AUROC of 100 random classifiers is reported in the last row.

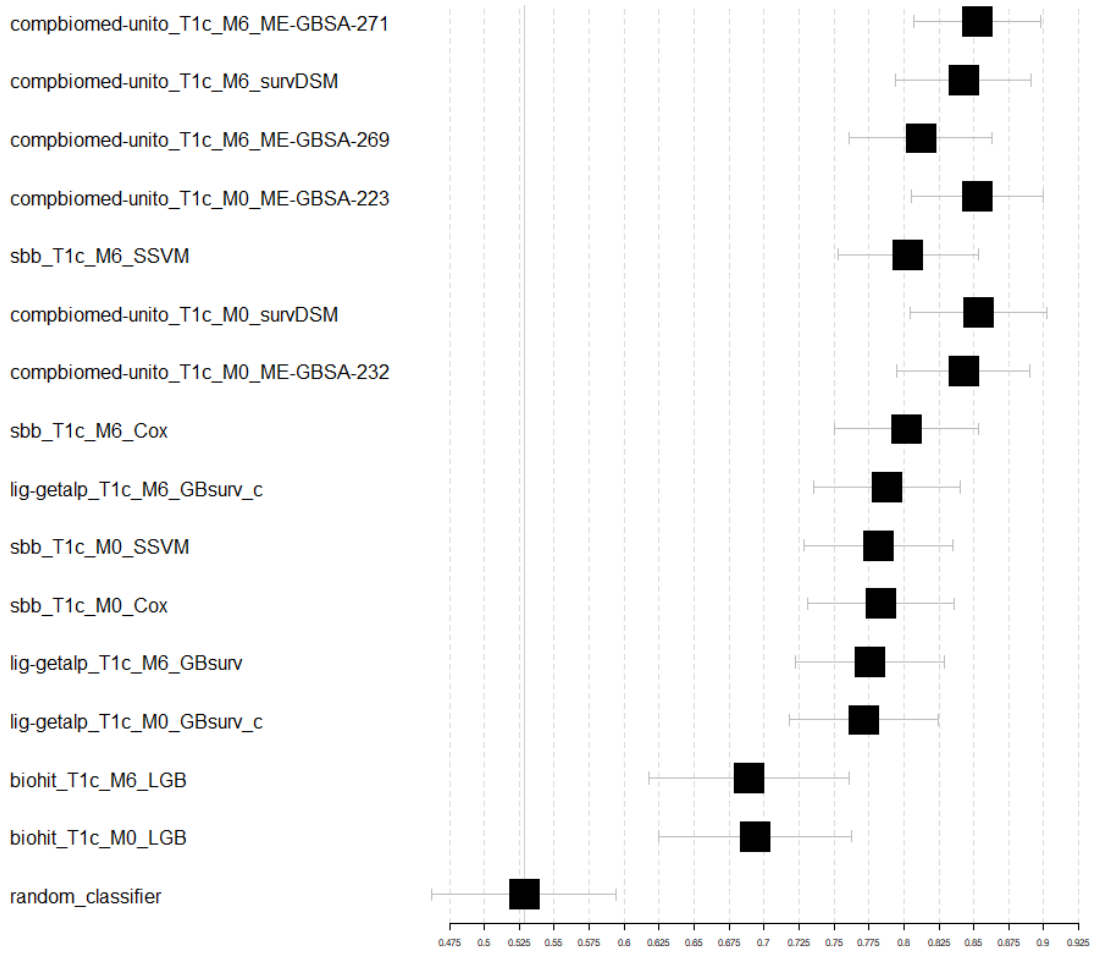


Figure 38: Sub-task c AUROC computed for all submitted runs with a 48-months PH. The bars in the plot show the 95% confidence intervals. The average 48-months AUROC of 100 random classifiers is reported in the last row.

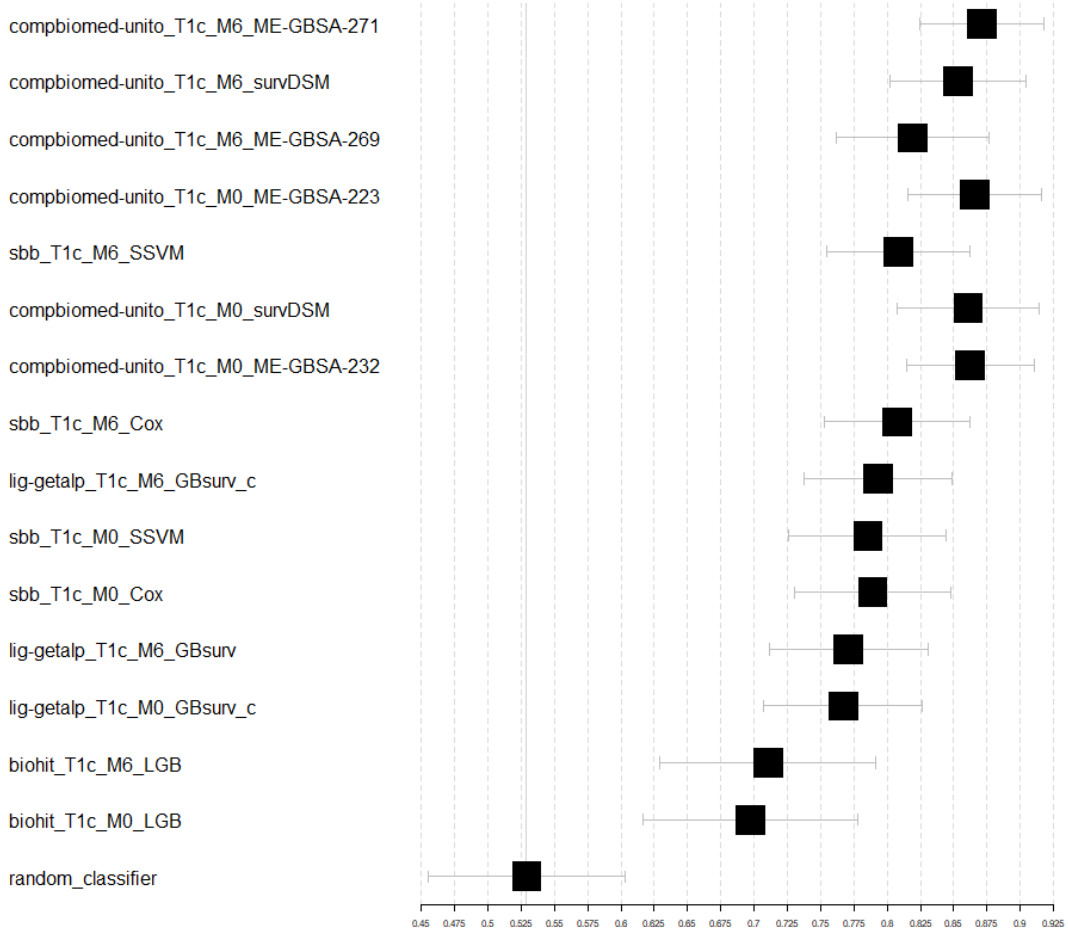


Figure 39: Sub-task c AUROC computed for all submitted runs with a 60-months PH. The bars in the plot show the 95% confidence intervals. The average 60-months AUROC of 100 random classifiers is reported in the last row.

C. Pilot task 1: BS

Figures 40 to 60 show the BS computed for all runs submitted for sub-tasks a, b, and c at all the considered PHs (12, 18, 24, 30, 36, 48, and 60 months). The average BS of the 100 random classifiers is reported in the last row of the figures.

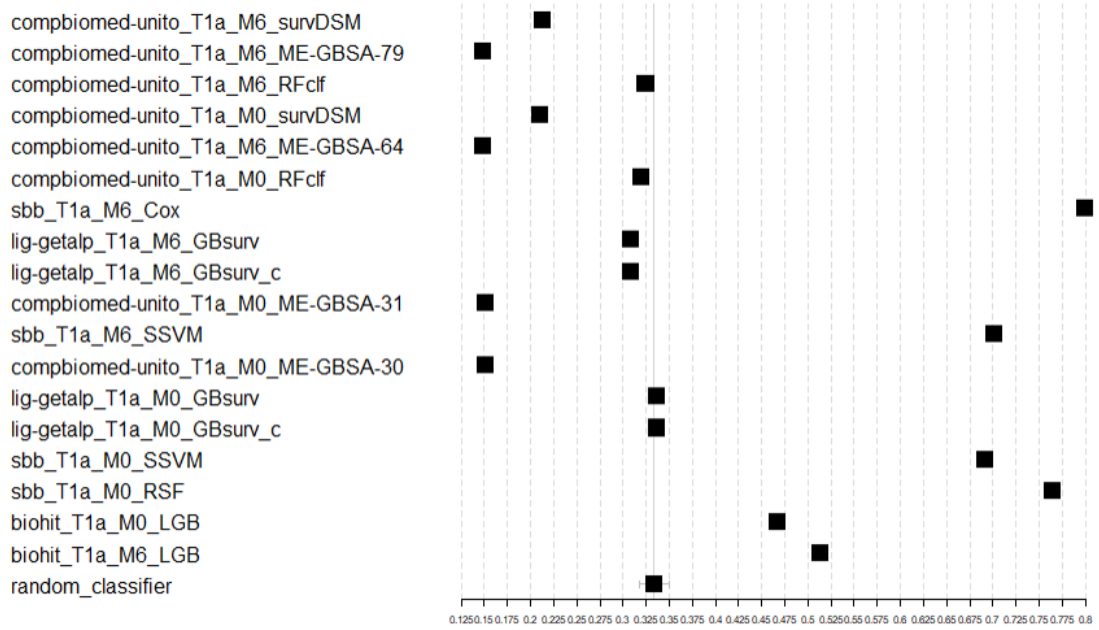


Figure 40: Sub-task a BS computed for all submitted runs with a 12-months PH. The random classifier average 12-months BS is reported in the last row with its 95% confidence intervals.

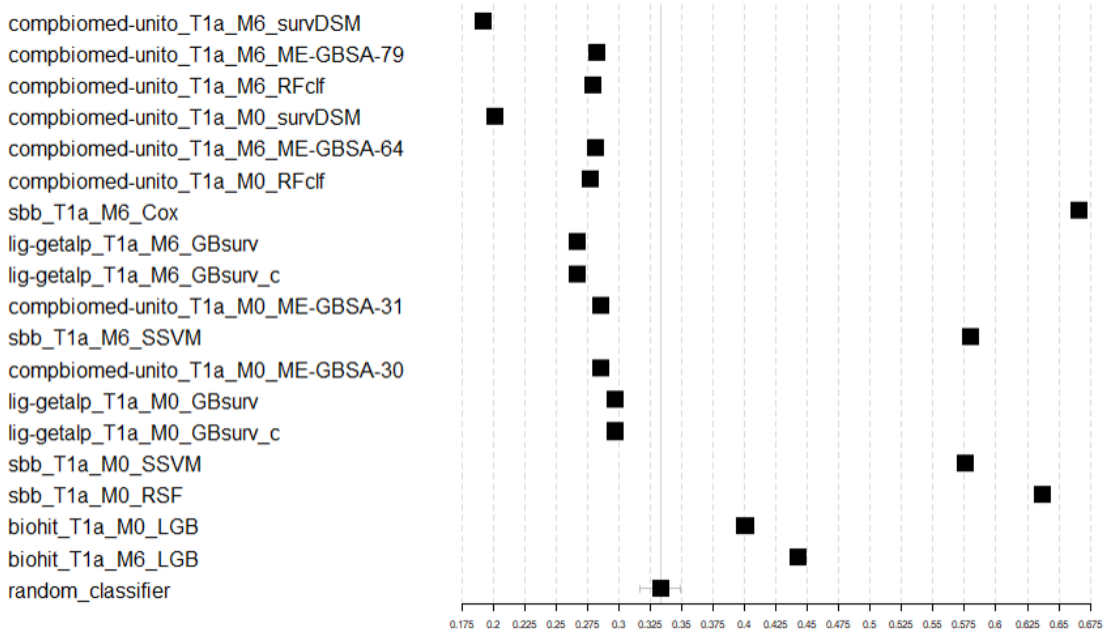


Figure 41: Sub-task a BS computed for all submitted runs with a 18-months PH. The random classifier average 18-months BS is reported in the last row with its 95% confidence intervals.

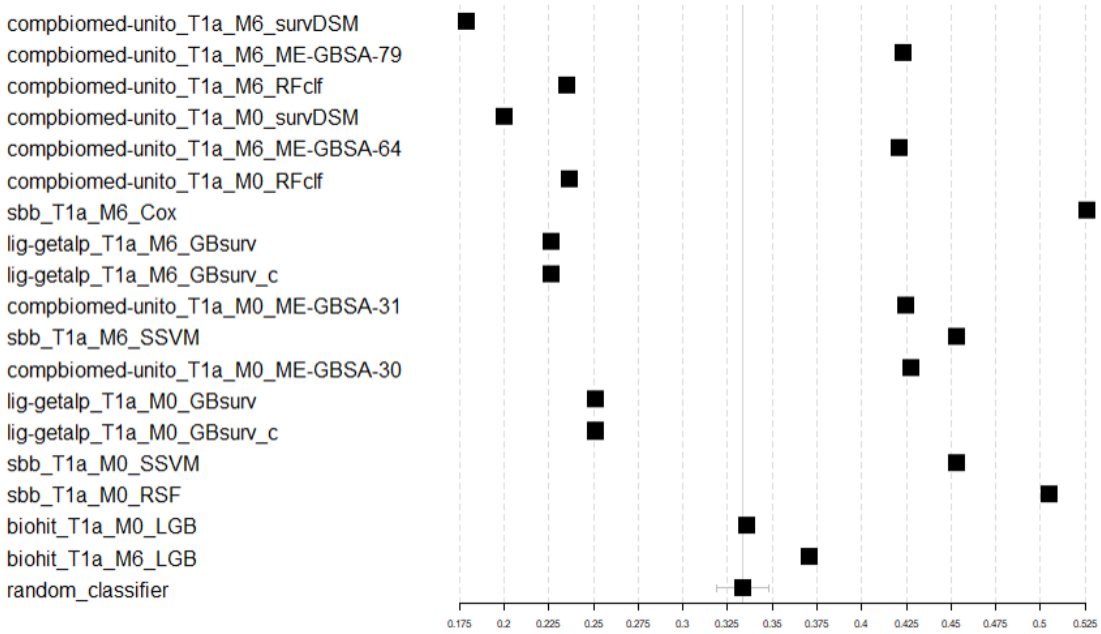


Figure 42: Sub-task a BS computed for all submitted runs with a 24-months PH. The random classifier average 24-months BS is reported in the last row with its 95% confidence intervals.

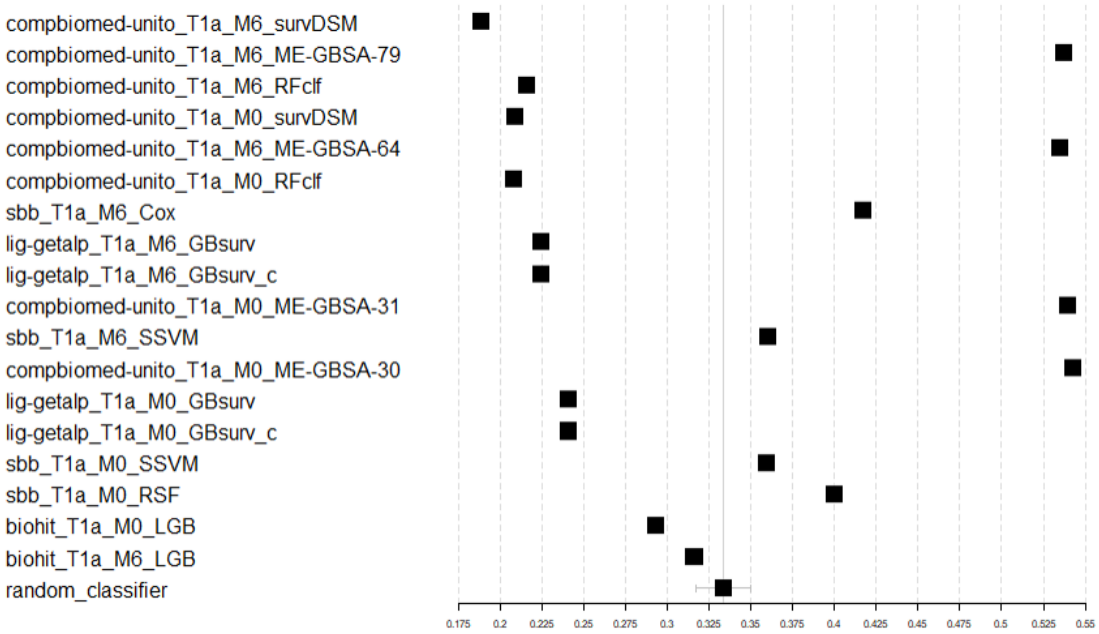


Figure 43: Sub-task a BS computed for all submitted runs with a 30-months PH. The random classifier average 30-months BS is reported in the last row with its 95% confidence intervals.

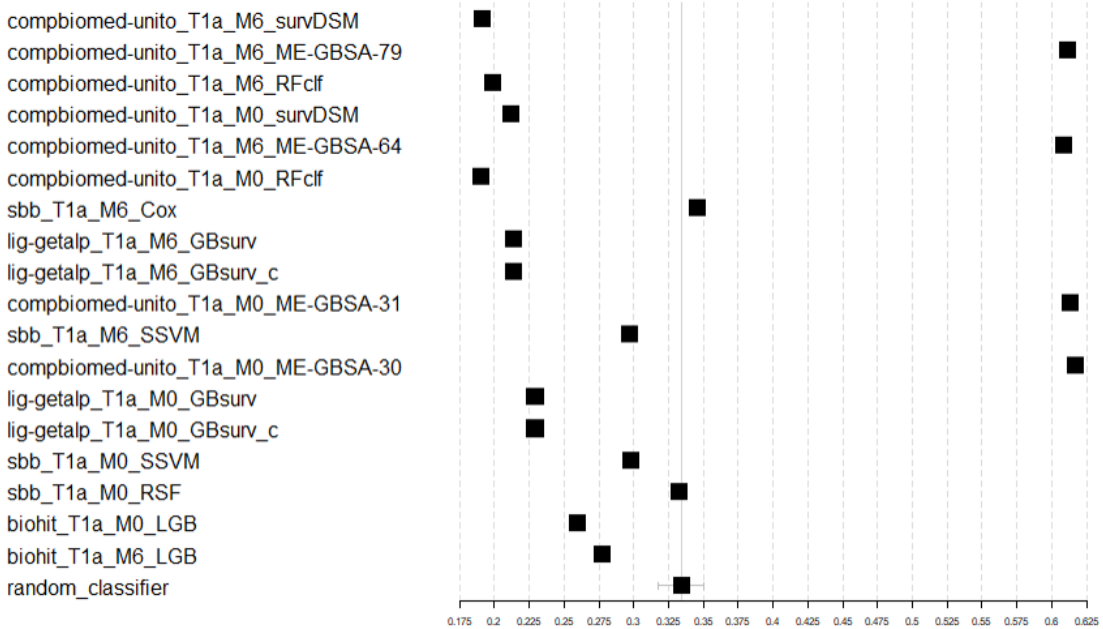


Figure 44: Sub-task a BS computed for all submitted runs with a 36-months PH. The random classifier average 36-months BS is reported in the last row with its 95% confidence intervals.

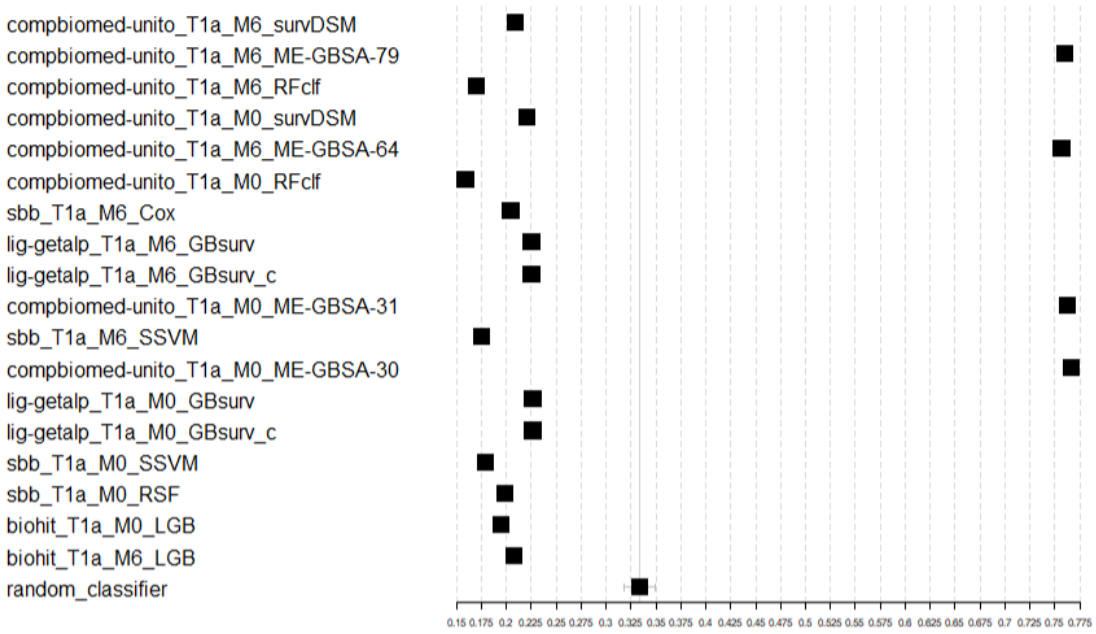


Figure 45: Sub-task a BS computed for all submitted runs with a 48-months PH. The random classifier average 48-months BS is reported in the last row with its 95% confidence intervals.

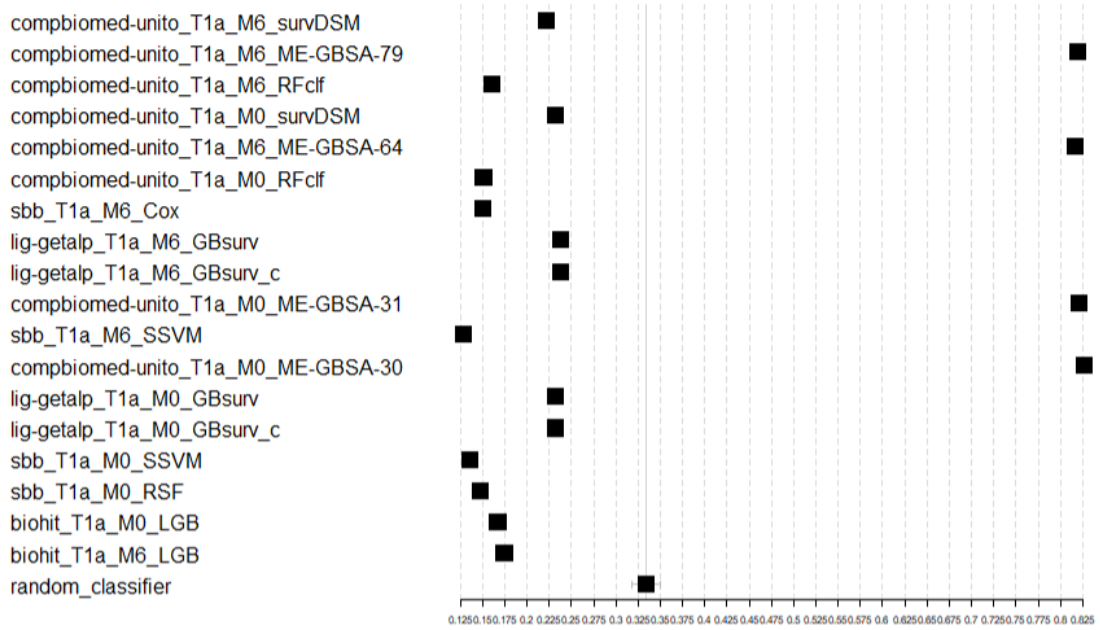


Figure 46: Sub-task a BS computed for all submitted runs with a 60-months PH. The random classifier average 60-months BS is reported in the last row with its 95% confidence intervals.

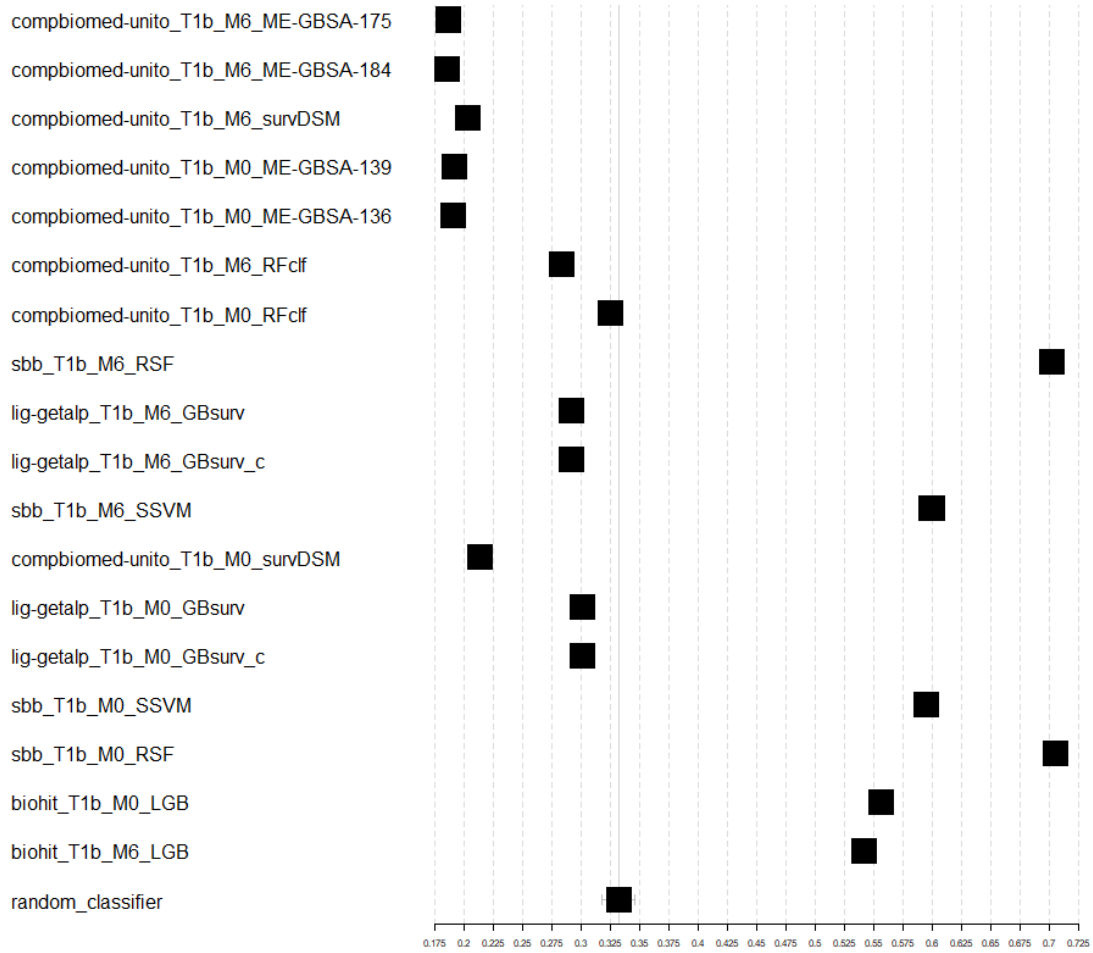


Figure 47: Sub-task b BS computed for all submitted runs with a 12-months PH. The random classifier average 12-months BS is reported in the last row with its 95% confidence intervals.

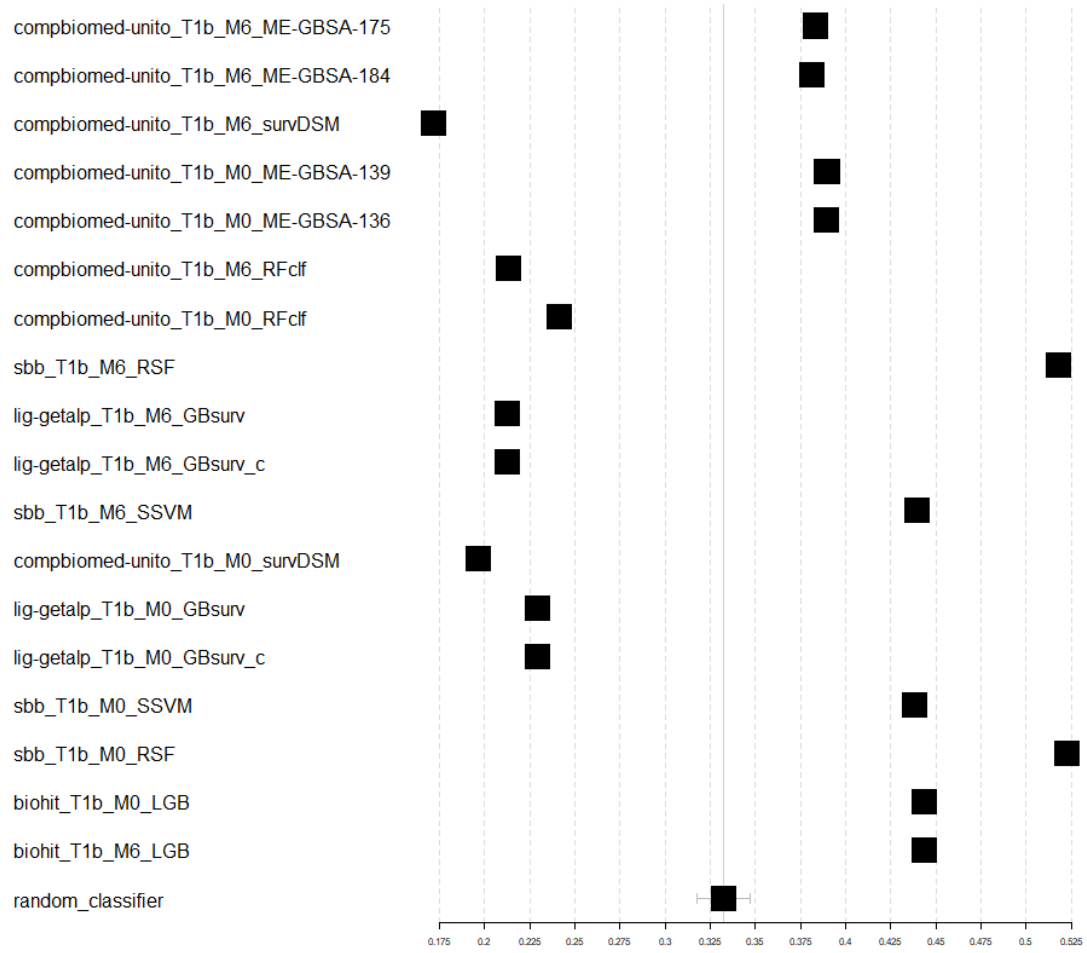


Figure 48: Sub-task b BS computed for all submitted runs with a 18-months PH. The random classifier average 18-months BS is reported in the last row with its 95% confidence intervals.

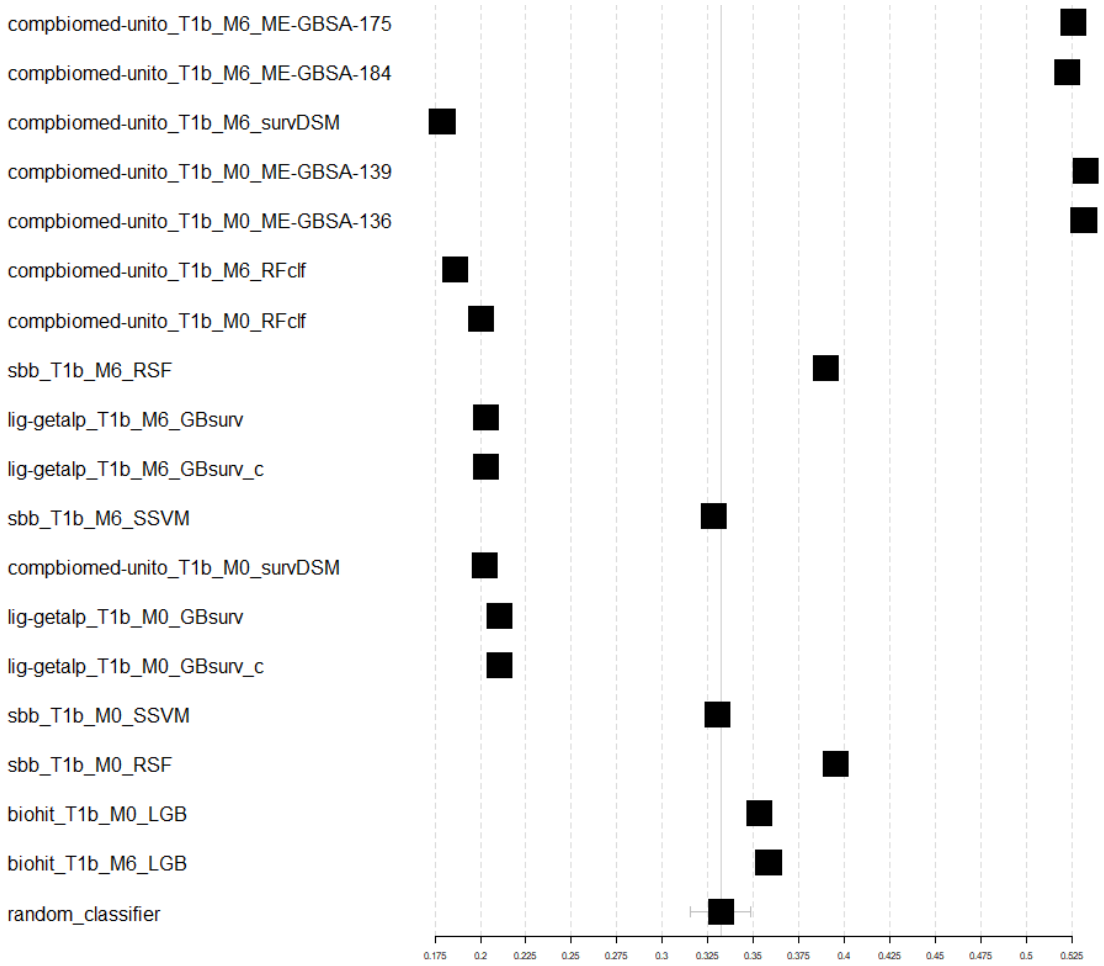


Figure 49: Sub-task b BS computed for all submitted runs with a 24-months PH. The random classifier average 24-months BS is reported in the last row with its 95% confidence intervals.

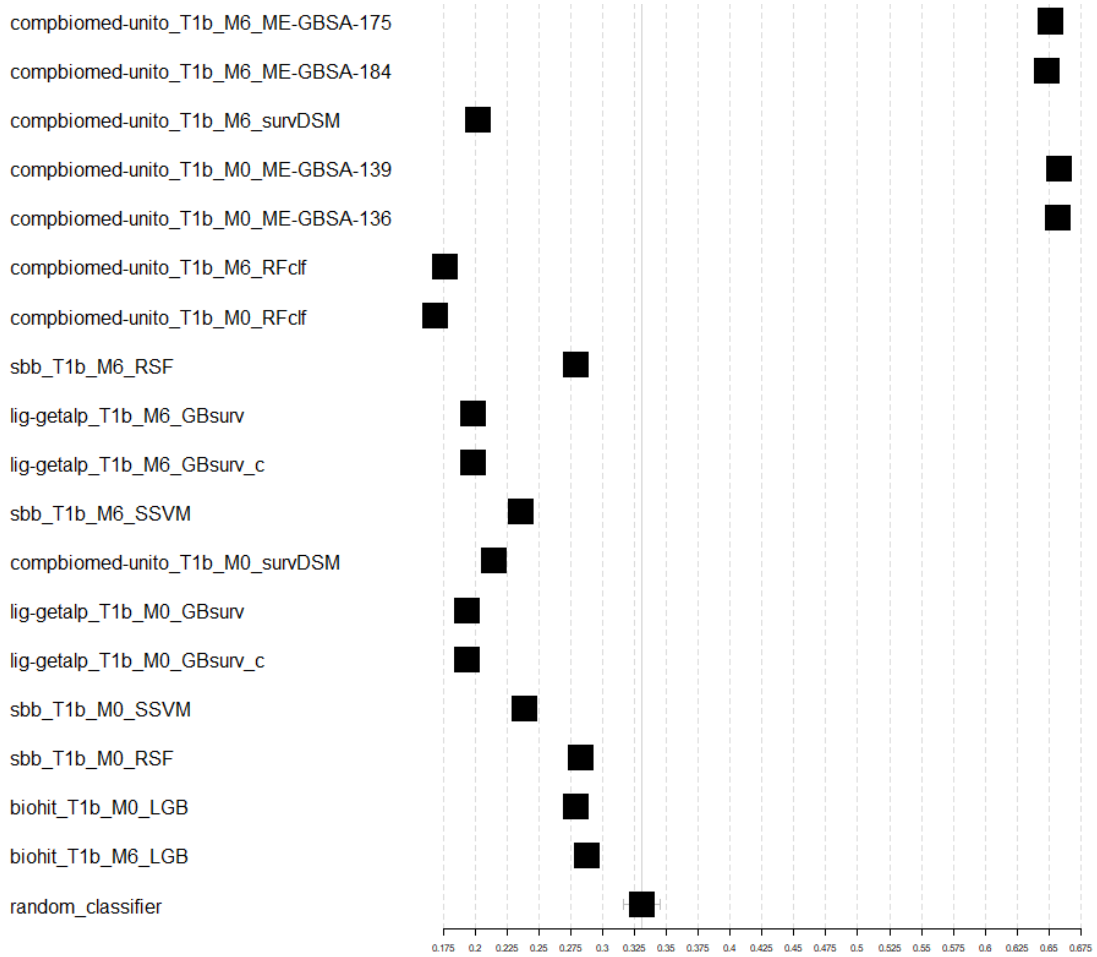


Figure 50: Sub-task b BS computed for all submitted runs with a 30-months PH. The random classifier average 30-months BS is reported in the last row with its 95% confidence intervals.

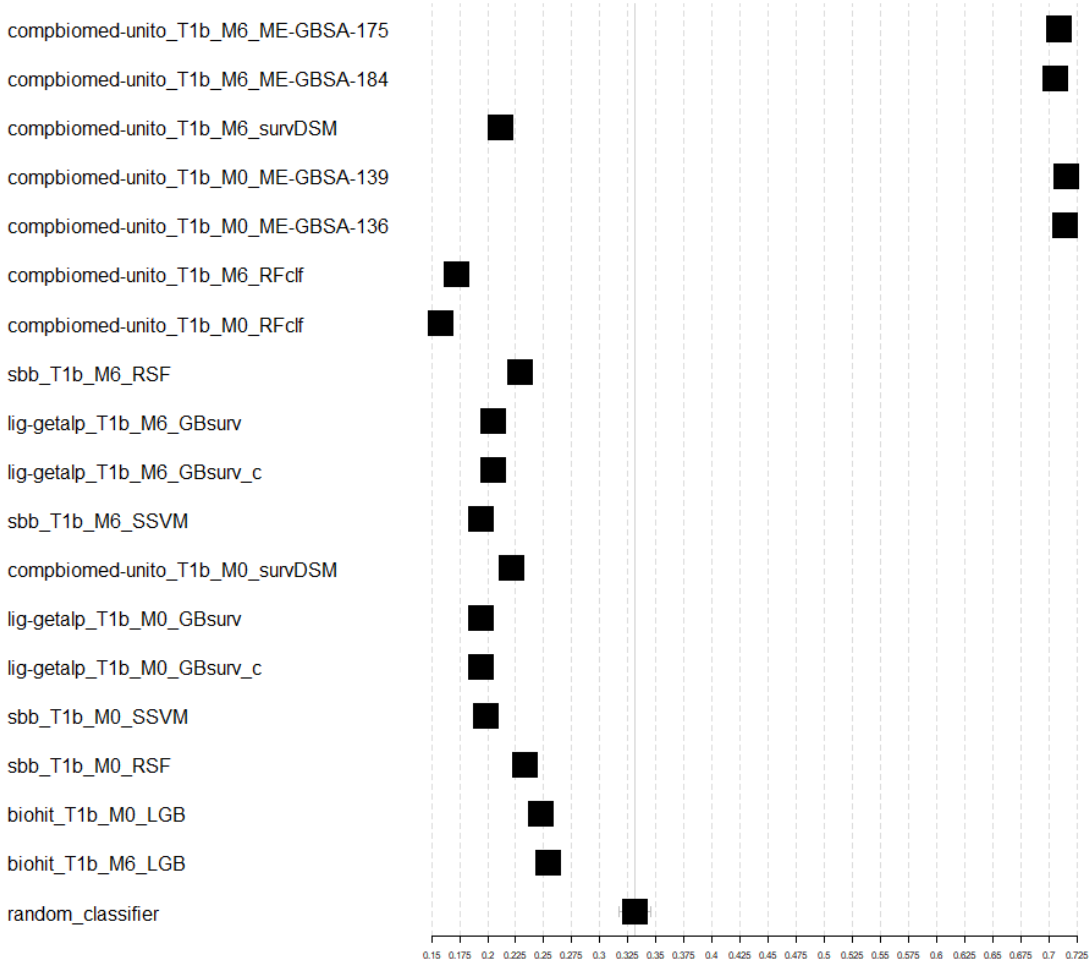


Figure 51: Sub-task b BS computed for all submitted runs with a 36-months PH. The random classifier average 36-months BS is reported in the last row with its 95% confidence intervals.

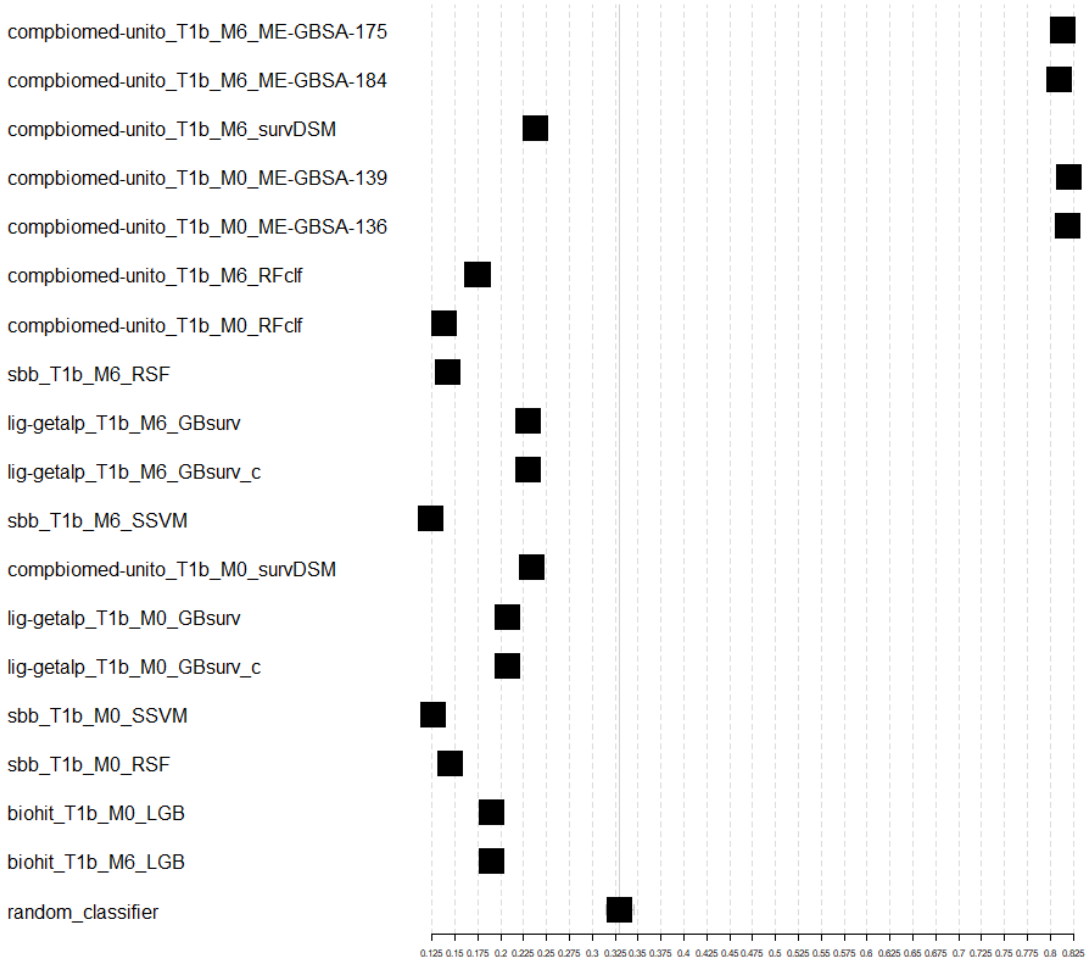


Figure 52: Sub-task b BS computed for all submitted runs with a 48-months PH. The random classifier average 48-months BS is reported in the last row with its 95% confidence intervals.

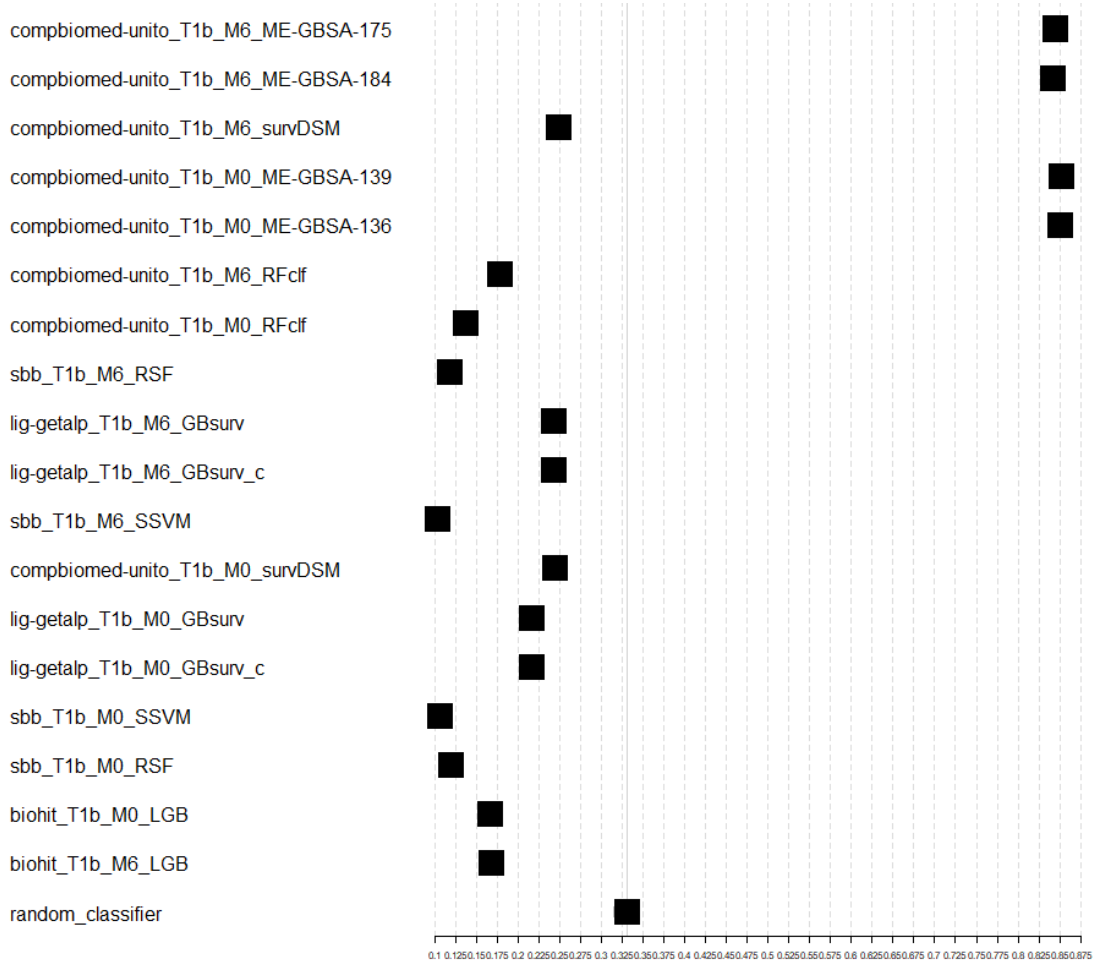


Figure 53: Sub-task b BS computed for all submitted runs with a 60-months PH. The random classifier average 60-months BS is reported in the last row with its 95% confidence intervals.

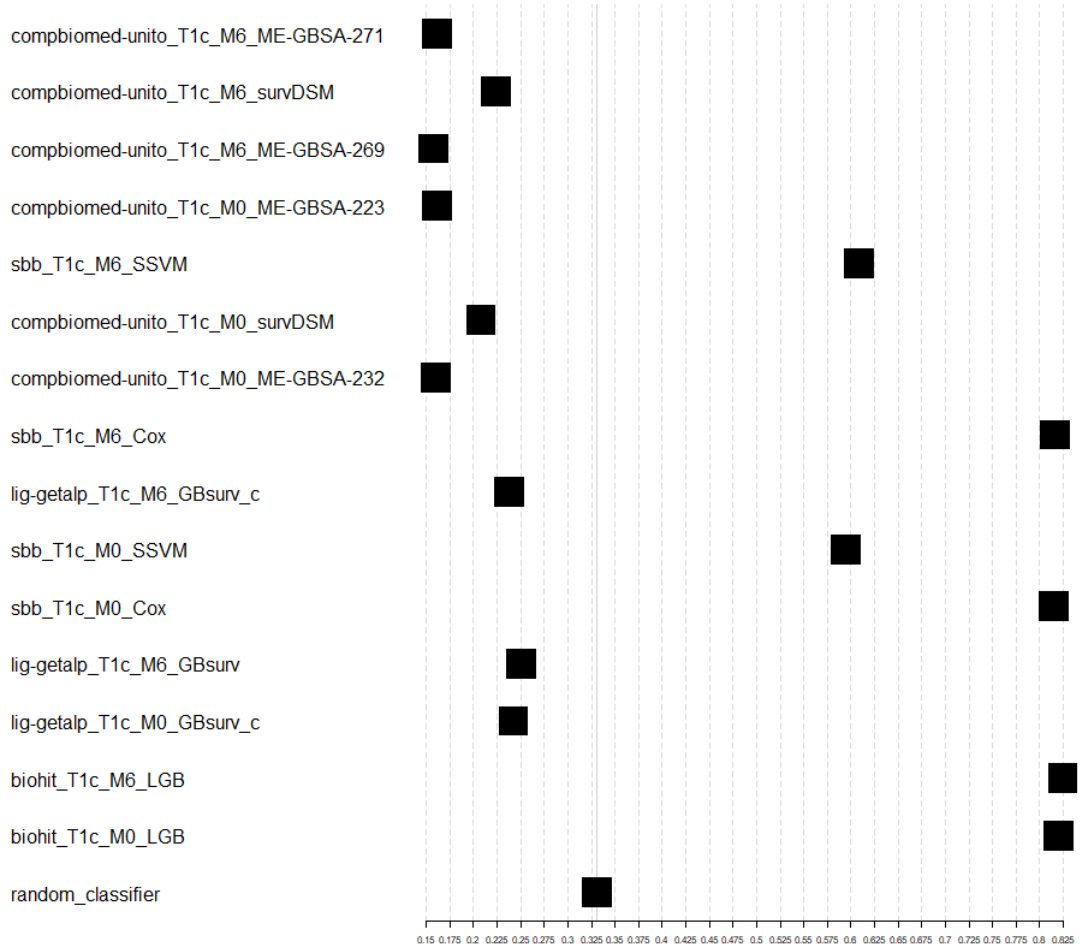


Figure 54: Sub-task c BS computed for all submitted runs with a 12-months PH. The random classifier average 12-months BS is reported in the last row with its 95% confidence intervals.

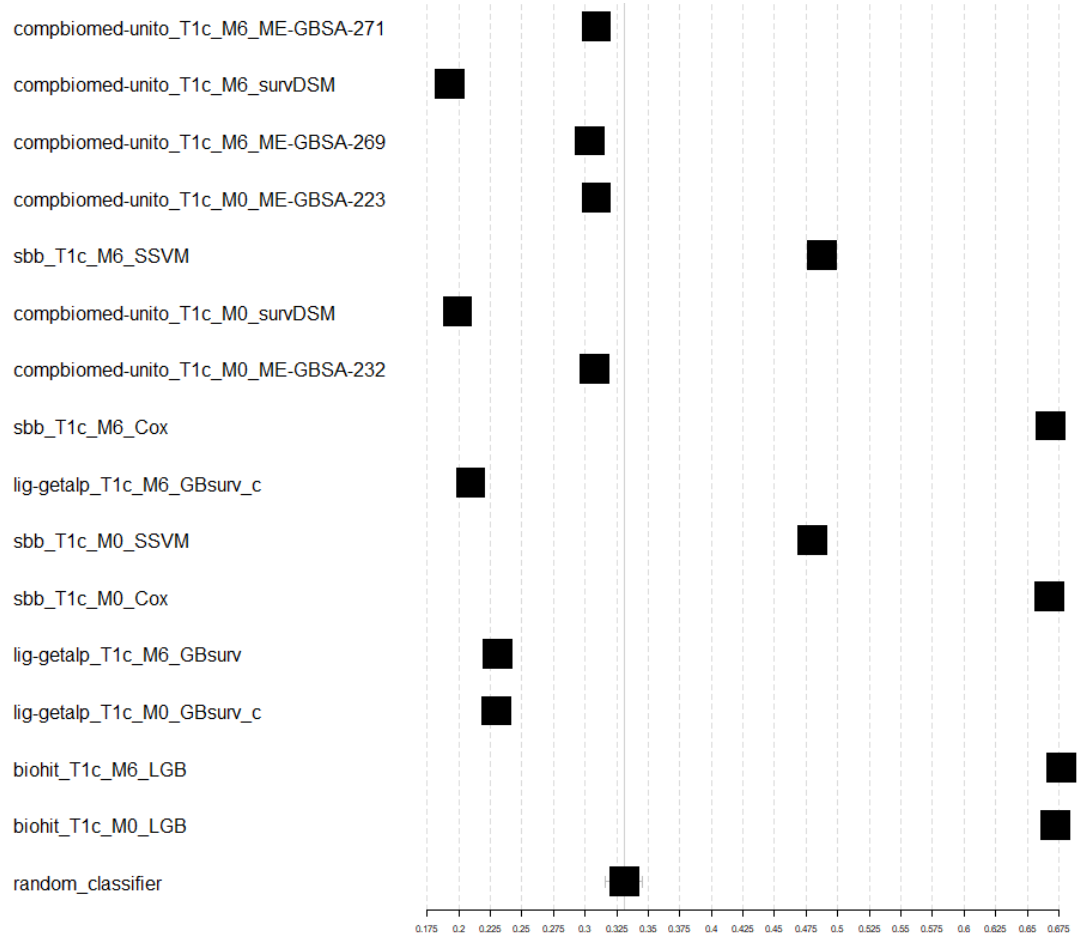


Figure 55: Sub-task c BS computed for all submitted runs with a 18-months PH. The random classifier average 18-months BS is reported in the last row with its 95% confidence intervals.

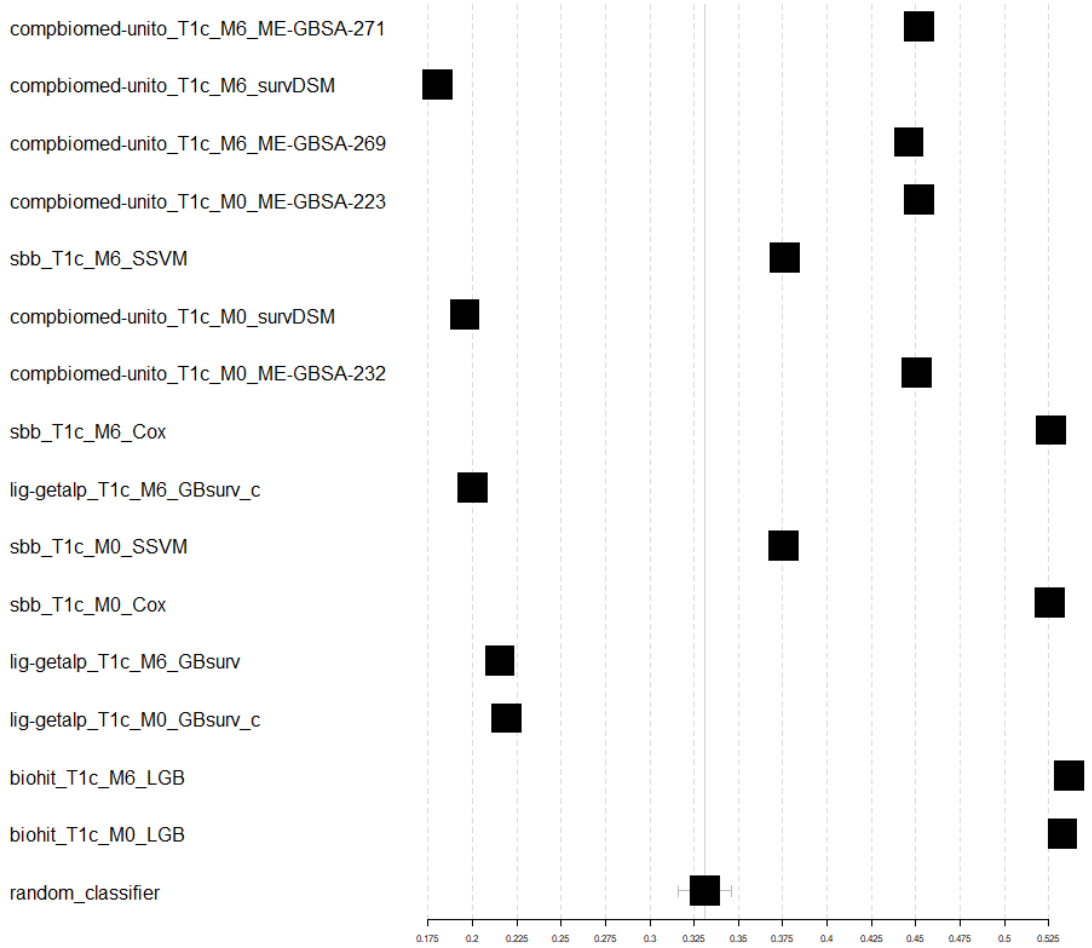


Figure 56: Sub-task c BS computed for all submitted runs with a 24-months PH. The random classifier average 24-months BS is reported in the last row with its 95% confidence intervals.

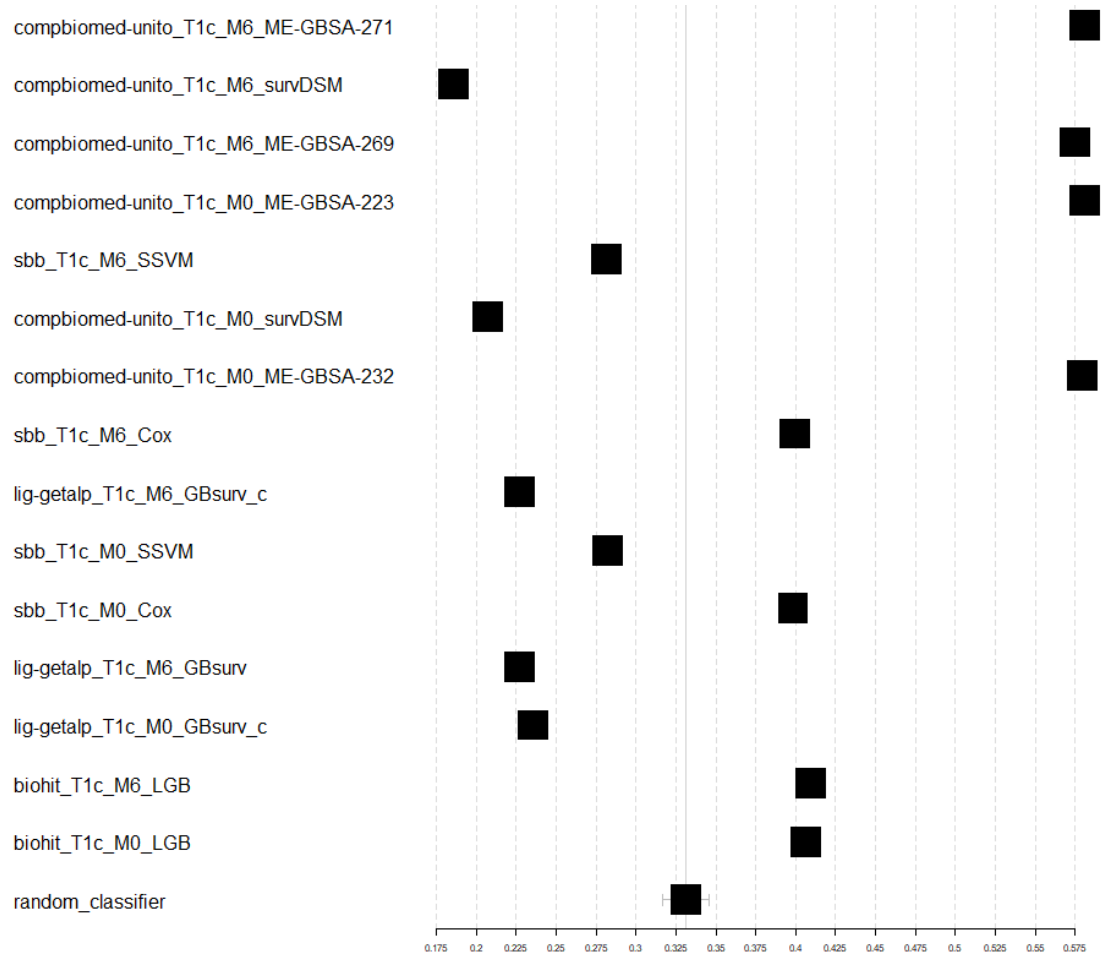


Figure 57: Sub-task c BS computed for all submitted runs with a 30-months PH. The random classifier average 30-months BS is reported in the last row with its 95% confidence intervals.

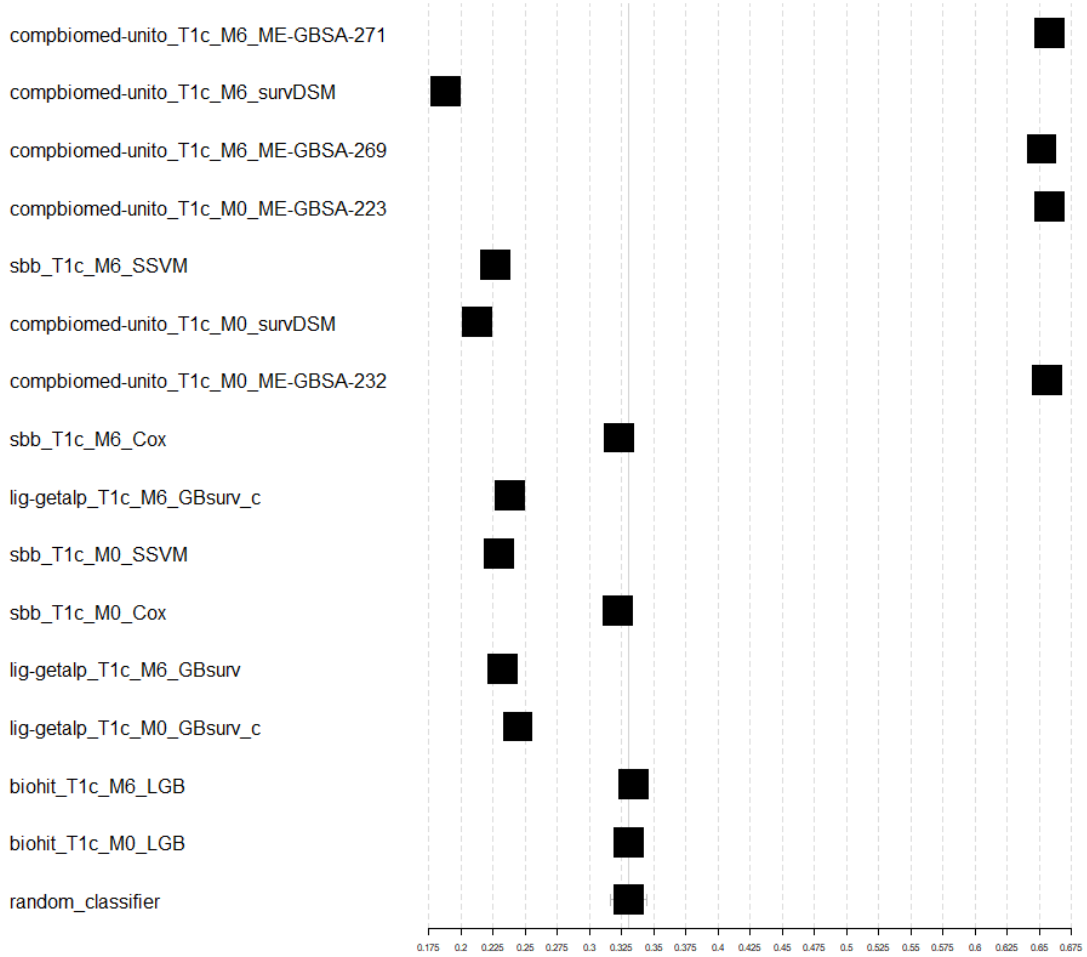


Figure 58: Sub-task c BS computed for all submitted runs with a 36-months PH. The random classifier average 36-months BS is reported in the last row with its 95% confidence intervals.

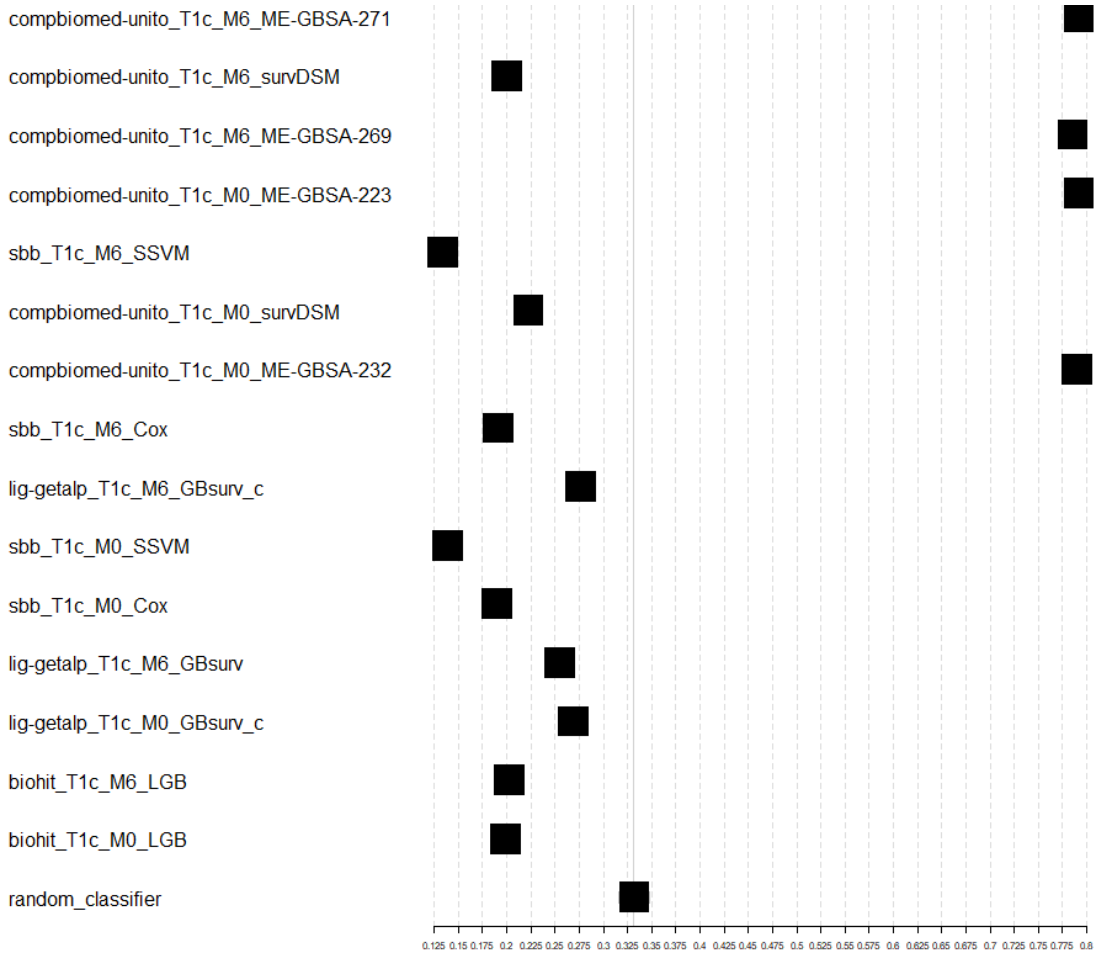


Figure 59: Sub-task c BS computed for all submitted runs with a 48-months PH. The random classifier average 48-months BS is reported in the last row with its 95% confidence intervals.

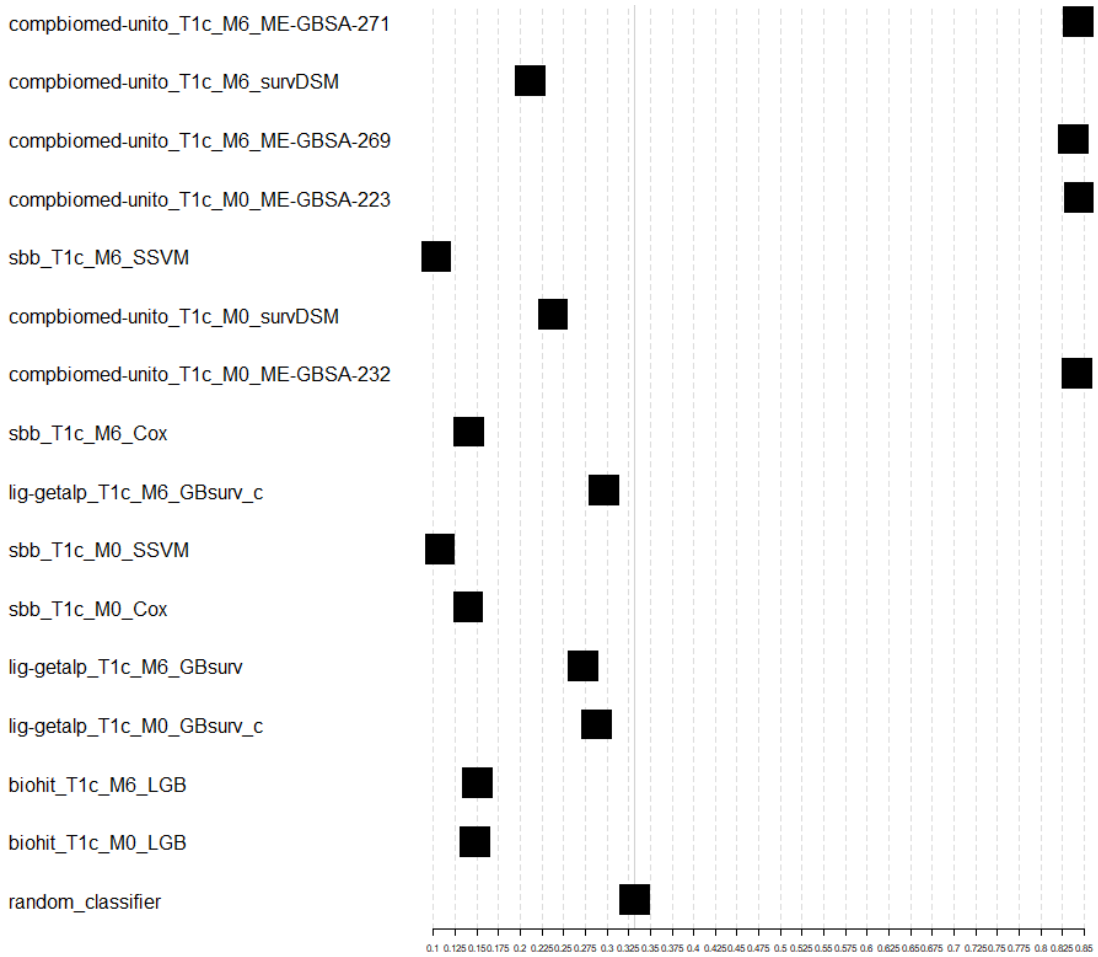
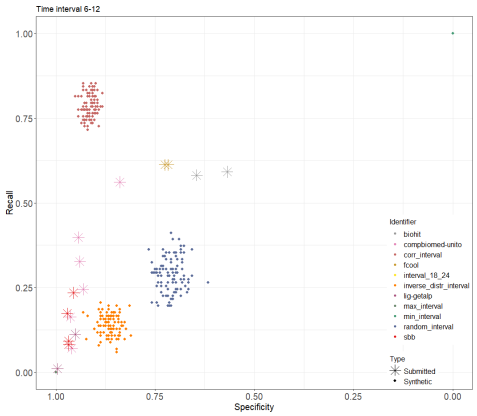


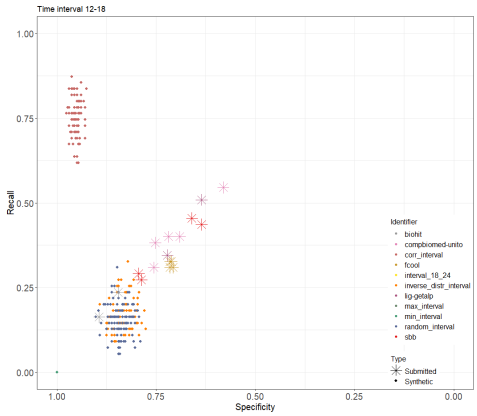
Figure 60: Sub-task c BS computed for all submitted runs with a 60-months PH. The random classifier average 60-months BS is reported in the last row with its 95% confidence intervals.

D. Pilot task 2: Time Interval Prediction Approach

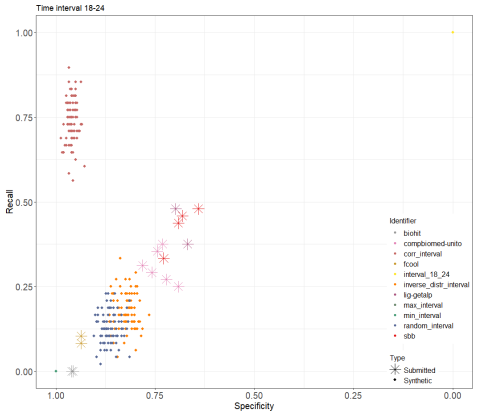
Figures 61, 62, and 63 show the specificity-recall plots for the time interval prediction approach. The 18 graphs include all time intervals and sub-tasks and display all participants' runs plus all the synthetic runs.



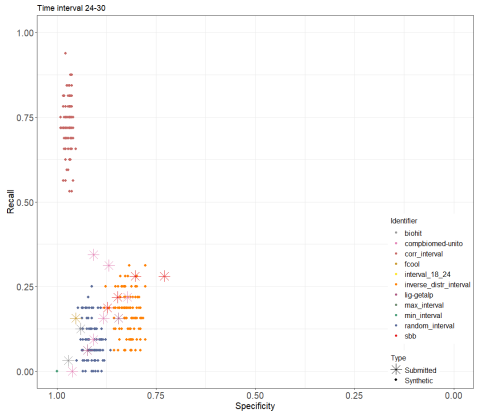
(a) Time interval 6-12.



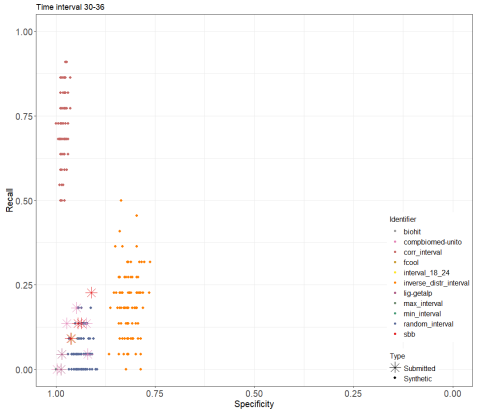
(b) Time interval 12-18.



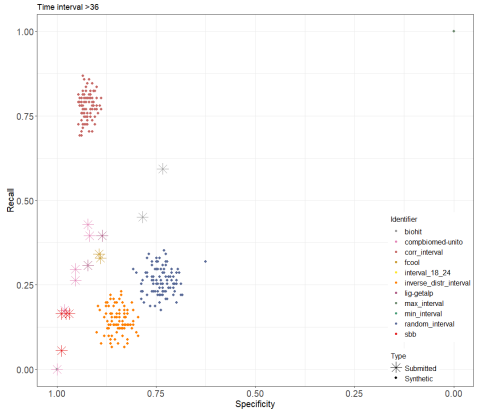
(c) Time interval 18-24.



(d) Time interval 24-30.

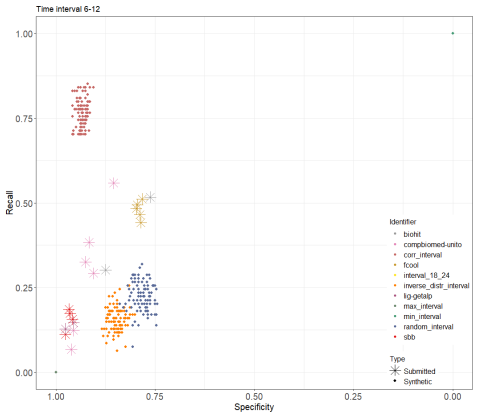


(e) Time interval 30-36.

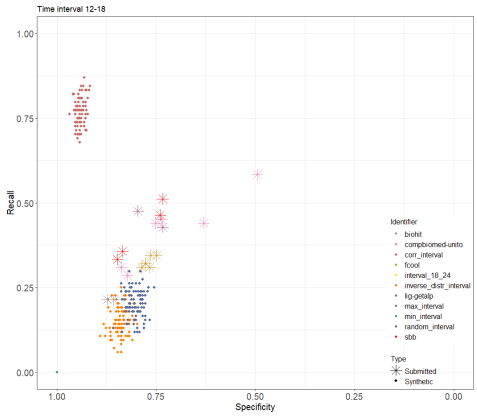


(f) Time interval >36.

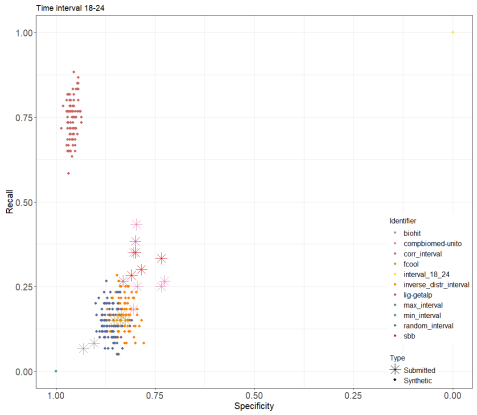
Figure 61: Time interval prediction approach. Specificity-recall plot, sub-task a.



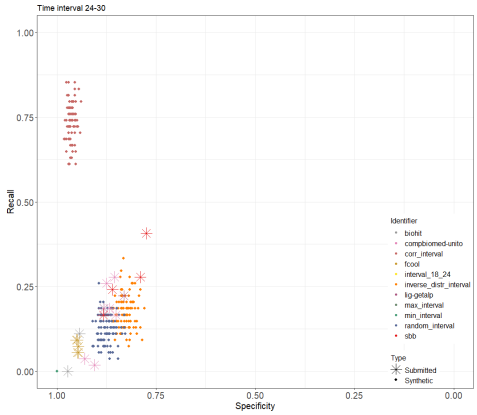
(a) Time interval 6-12.



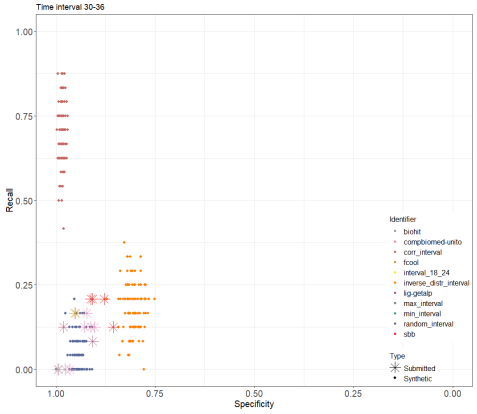
(b) Time interval 12-18.



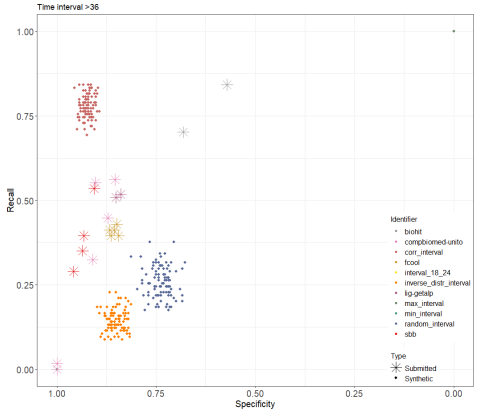
(c) Time interval 18-24.



(d) Time interval 24-30.

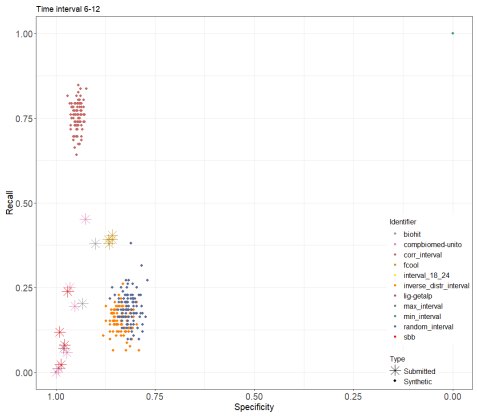


(e) Time interval 30-36.

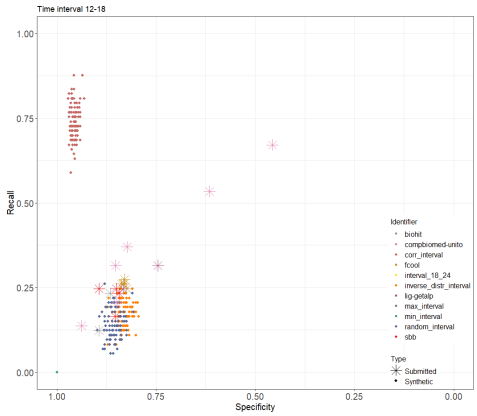


(f) Time interval >36.

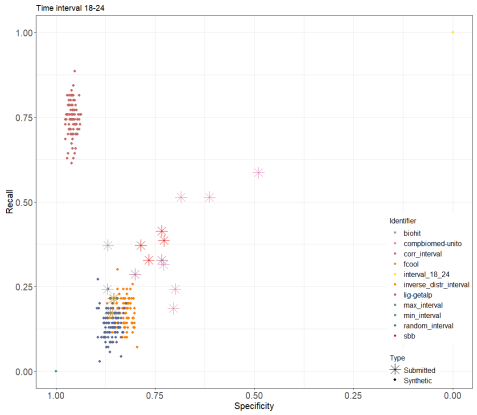
Figure 62: Time interval prediction approach. Specificity-recall plot, sub-task b.



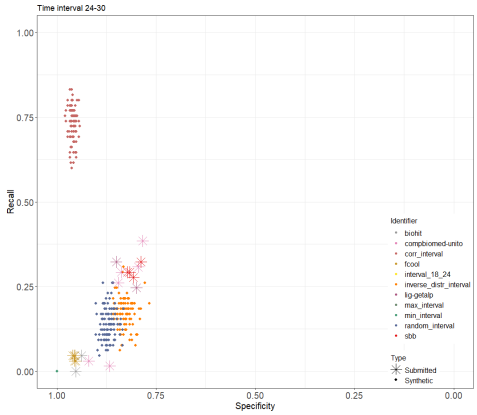
(a) Time interval 6-12.



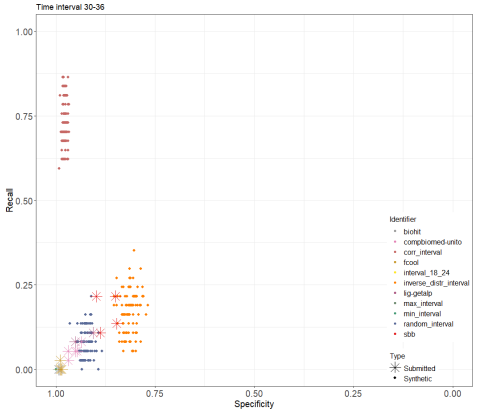
(b) Time interval 12-18.



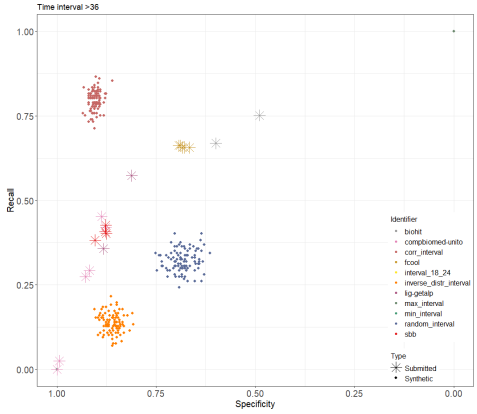
(c) Time interval 18-24.



(d) Time interval 24-30.



(e) Time interval 30-36.

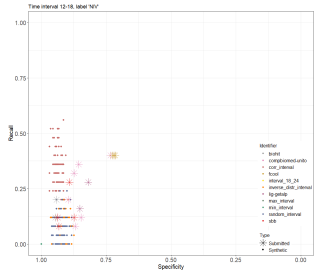


(f) Time interval >36.

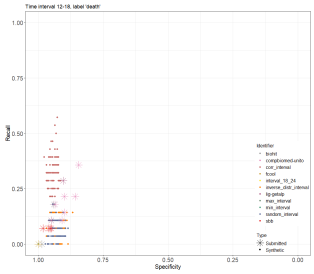
Figure 63: Time interval prediction approach. Specificity-recall plot, sub-task c.

E. Pilot task 2: Label Prediction Approach

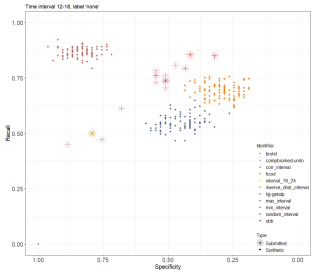
Figures 64, 65, and 66 show the specificity-recall plots for the label prediction approach. In every Figure, each row corresponds to an observation time ("12-18", "18-24", "24-30", "30-36") and each column to a label (*NIV*, *none*, and *death* for sub-task a; *PEG*, *none*, and *death* for sub-task b; *none*, and *death* for sub-task c).



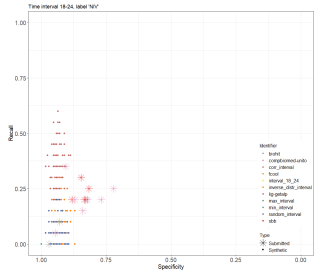
(a) Observation time 12-18, label "NIV".



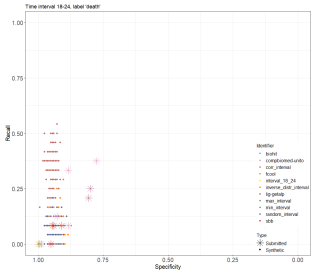
(b) Observation time 12-18, label "death".



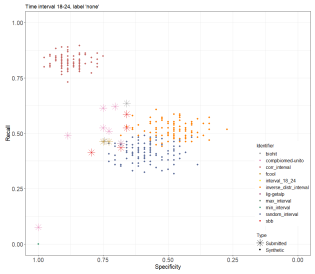
(c) Observation time 12-18, label "none".



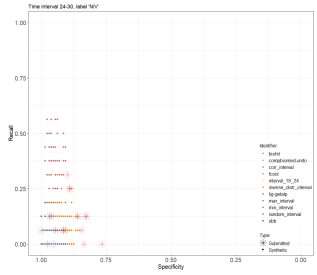
(d) Observation time 18-24, label "NIV".



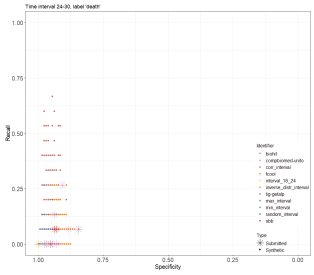
(e) Observation time 18-24, label "death".



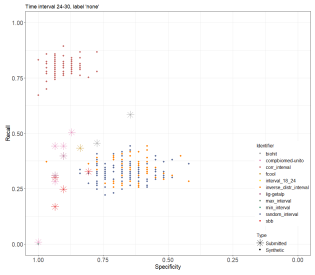
(f) Observation time 18-24, label "none".



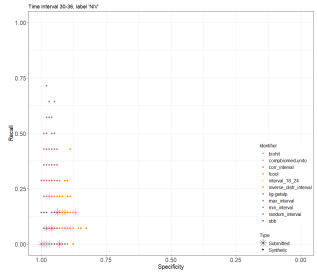
(g) Observation time 24-30, label "NIV".



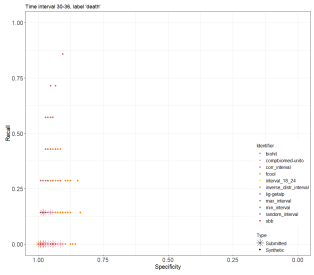
(h) Observation time 24-30, label "death".



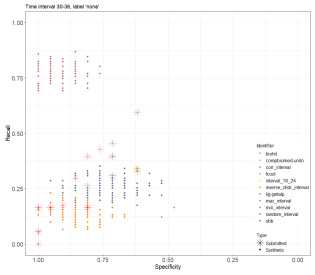
(i) Observation time 24-30, label "none".



(j) Observation time 30-36, label "NIV".

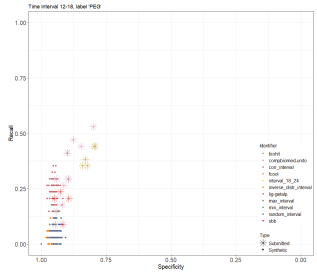


(k) Observation time 30-36, label "death".

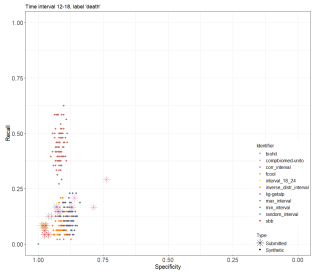


(l) Observation time 30-36, label "none".

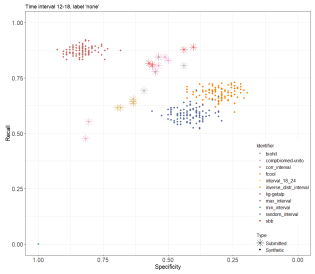
Figure 64: Label prediction approach. Specificity-recall plot, sub-task a. Each row corresponds to an observation time ("12-18", "18-24", "24-30", "30-36") and each column to a label ("NIV", "death", "none")



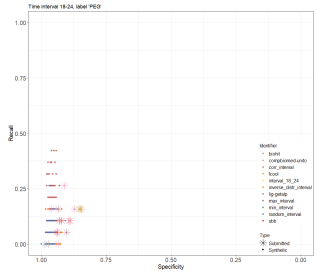
(a) Observation time 12-18, label "PEG".



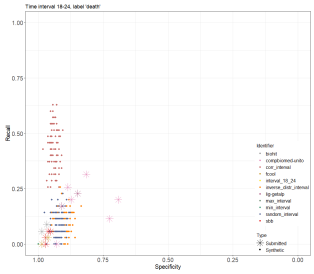
(b) Observation time 12-18, label "death".



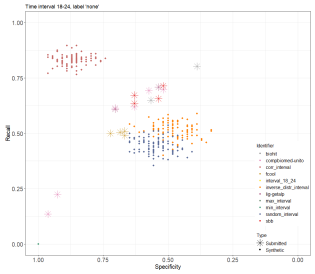
(c) Observation time 12-18, label "none".



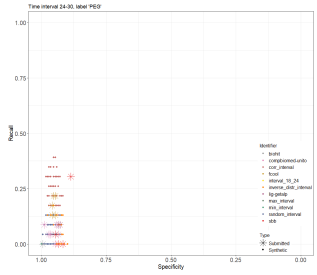
(d) Observation time 18-24, label "PEG".



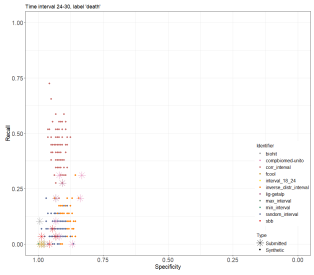
(e) Observation time 18-24, label "death".



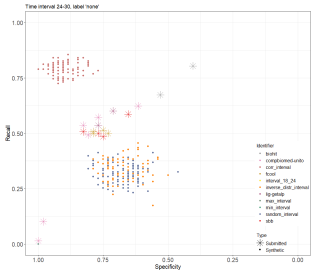
(f) Observation time 18-24, label "none".



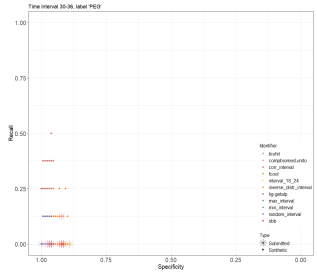
(g) Observation time 24-30, label "PEG".



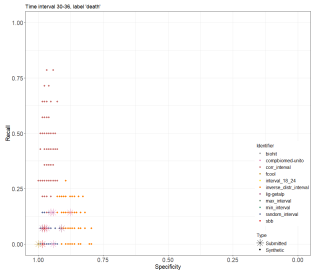
(h) Observation time 24-30, label "death".



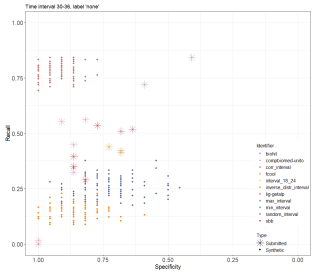
(i) Observation time 24-30, label "none".



(j) Observation time 30-36, label "PEG".

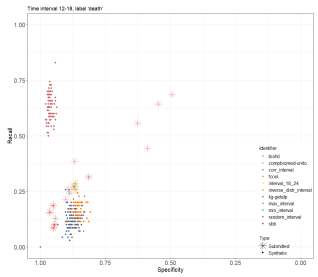


(k) Observation time 30-36, label "death".

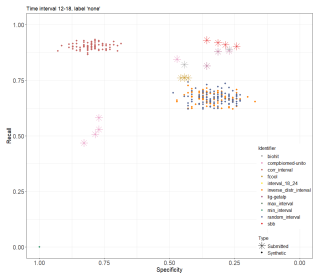


(l) Observation time 30-36, label "none".

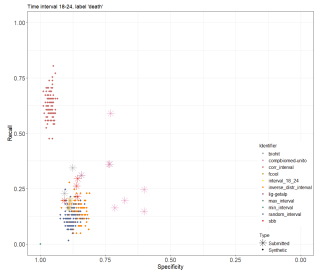
Figure 65: Label prediction approach. Specificity-recall plot, sub-task b. Each row corresponds to an observation time ("12-18", "18-24", "24-30", "30-36") and each column to a label ("PEG", "death", "none")



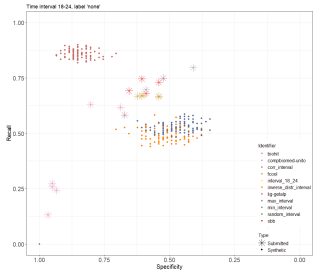
(a) Observation time 12-18, label “death”.



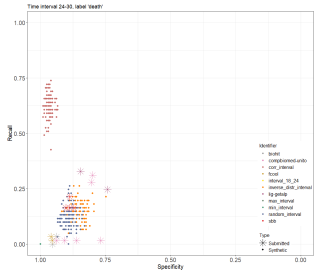
(b) Observation time 12-18, label “none”.



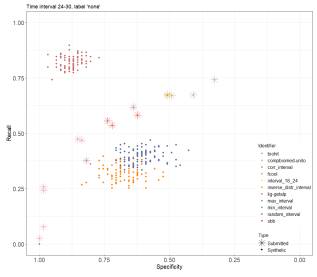
(c) Observation time 18-24, label “death”.



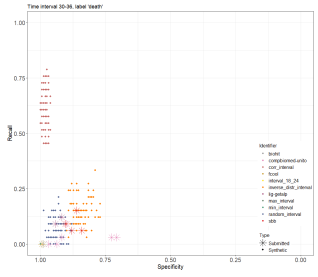
(d) Observation time 18-24, label “none”.



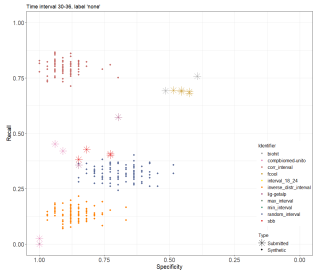
(e) Observation time 24-30, label “death”.



(f) Observation time 24-30, label “none”.



(g) Observation time 30-36, label “death”.



(h) Observation time 30-36, label “none”.

Figure 66: Label prediction approach. Specificity-recall plot, sub-task c. Each row corresponds to an observation time (“12-18”, “18-24”, “24-30”, “30-36”) and each column to a label (“death”, “none”)

F. Pilot task 2: AbsDist

Figures 67 to 69 show the AbsDist computed for all runs submitted for sub-tasks a, b, and c. The average AbsDist of the synthetic runs is reported as well.

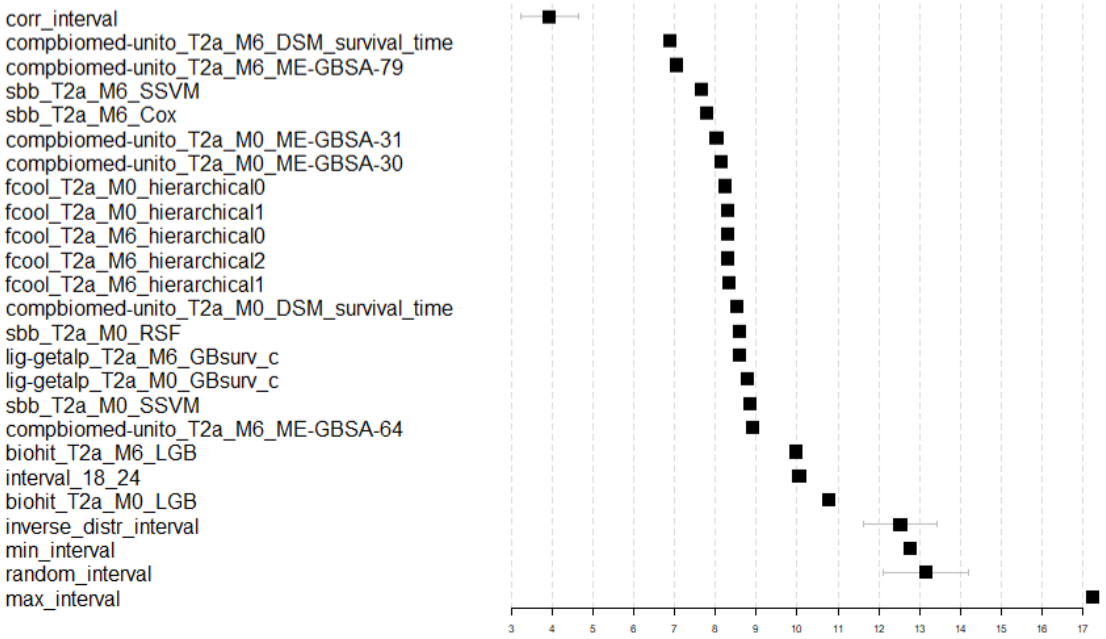


Figure 67: Sub-task a AbsDist computed for all submitted and synthetic runs. The AbsDist of *corr_interval*, *inverse_distr_interval*, and *random_interval* is the average with 95% confidence intervals computed on the corresponding 100 randomly generated runs.

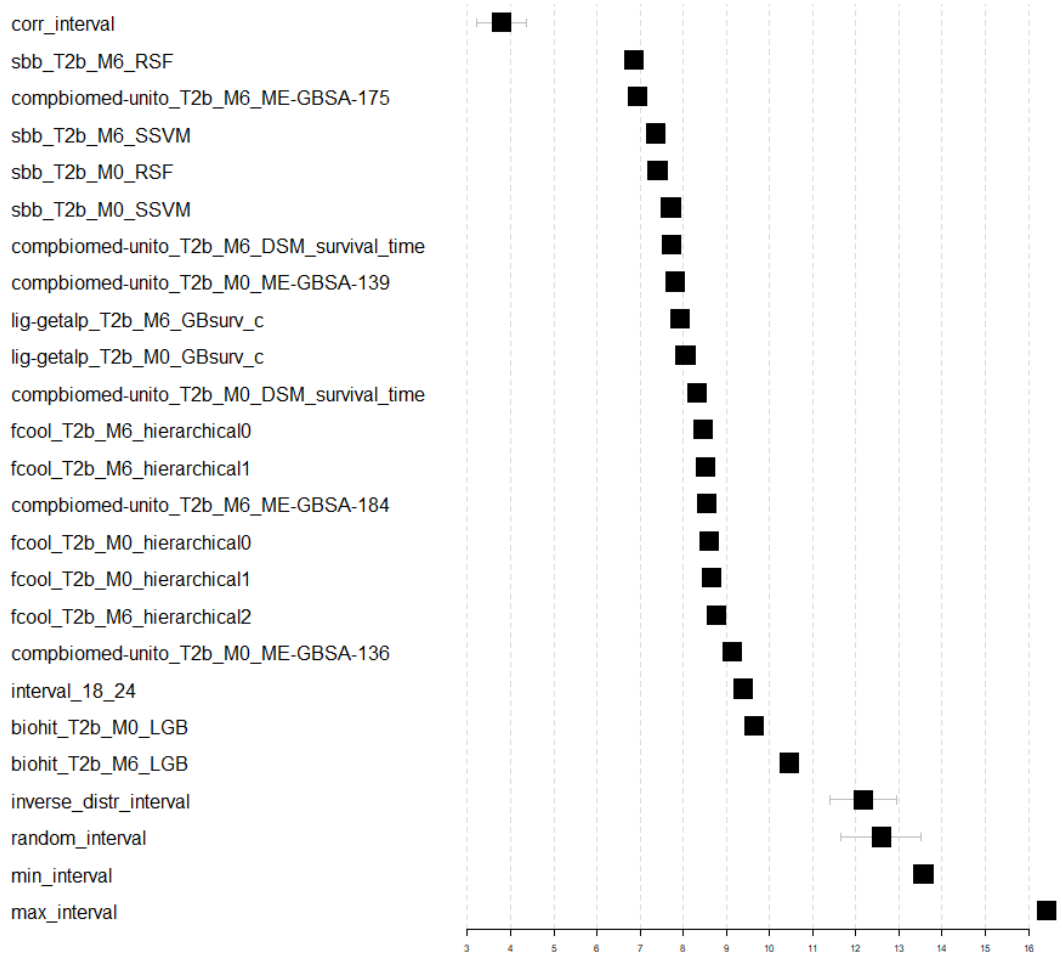


Figure 68: Sub-task b AbsDist computed for all submitted and synthetic runs. The AbsDist of *corr_interval*, *inverse_distr_interval*, and *random_interval* is the average with 95% confidence intervals computed on the corresponding 100 randomly generated runs.

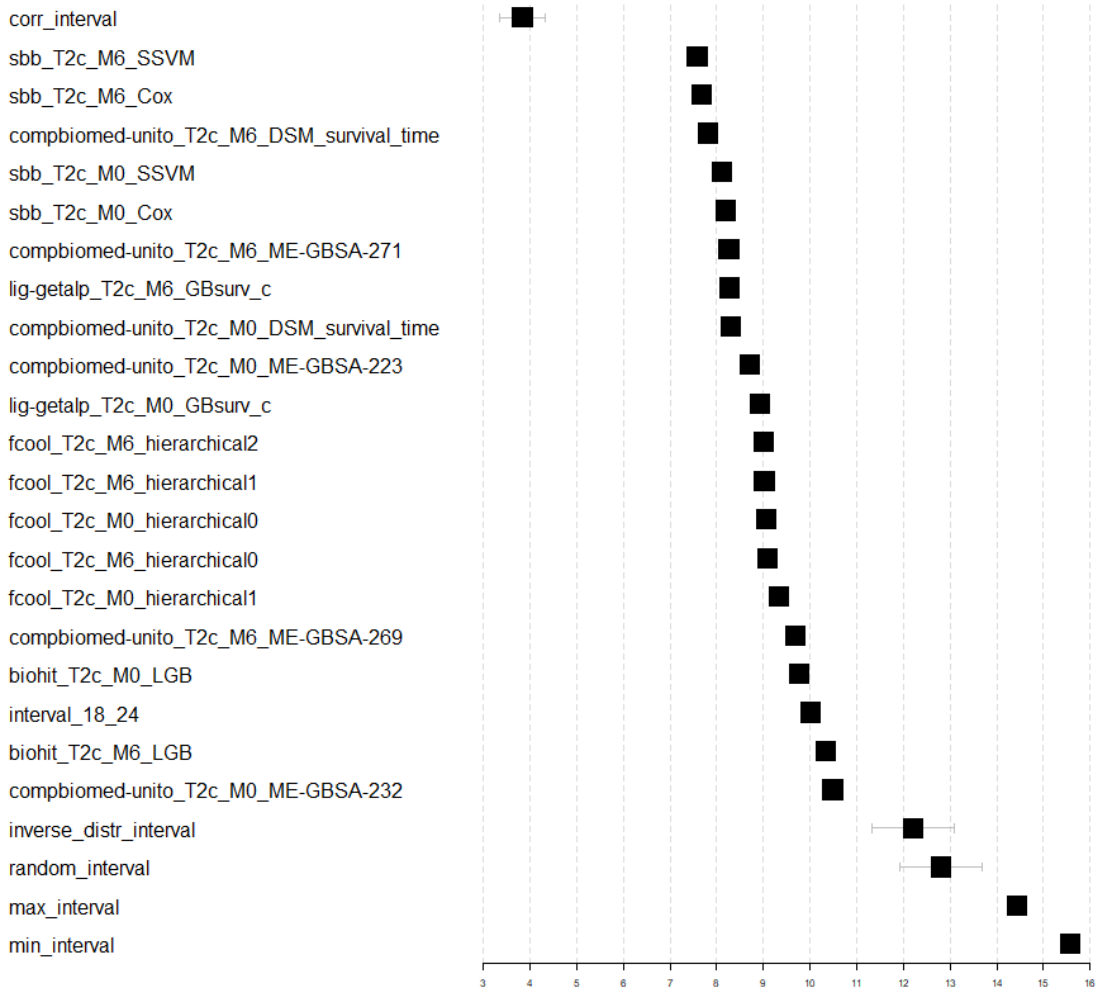


Figure 69: Sub-task c AbsDist computed for all submitted and synthetic runs. The AbsDist of *corr_interval*, *inverse_distr_interval*, and *random_interval* is the average with 95% confidence intervals computed on the corresponding 100 randomly generated runs.