



iDPP@CLEF 2022

Performance Evaluation Framework and Results

CLEF 2022, Bologna, Italy

Alessandro Guazzo

Department of Information Engineering
University of Padova





Amyotrophic lateral sclerosis (ALS)

- ▶ A neurological disease that causes the **degradation of motor neurons**
- ▶ A **fast progressing disease** that leads to increasing disability until death within 4-5 years since the diagnosis
- ▶ Lack of effective prognostic tools
 - ▷ Hard to understand how the disease will progress
- ▶ Clinicians require **tools to predict the disease progression** in order to suggest personalised therapies or indicate interventions

Designing and developing AI algorithms to predict the progression of the disease is key

iDPP@CLEF Tasks





Two tasks on **ALS progression** prediction

- ▶ **Task 1 aim:** rank subjects based on risk of occurrence of clinical events
 - ▷ **Sub-task 1a:** non-invasive ventilation (**NIV**) or **death**, whichever occurs first
 - *NIV and death were considered as **competing events (OR operation)** as if a patient dies NIV can no longer happen*
 - ▷ **Sub-task 1b:** percutaneous endoscopic gastrostomy (**PEG**) or **death**, whichever occurs first
 - *PEG and death were considered as **competing events (OR operation)** as if a patient dies PEG can no longer happen*
 - ▷ **Sub-task 1c: death**

- ▶ **Task 2 aim:** predict when clinical events will occur
 - ▷ **Sub-task 2a: NIV or death**
 - ▷ **Sub-task 2b: PEG or death**
 - ▷ **Sub-task 2c: death**



Runs

To evaluate whether **collecting additional data after the baseline** could improve performance, participants were asked to submit **two types of runs** for each task

- ▶ **M0 runs:** predictions obtained from a model trained using only data available at baseline
- ▶ **M6 runs:** predictions obtained from a model trained using all the data available until month 6



Run Example: Task 1

Patient Identifier

Risk score

0xd3ee37821226aa86382711f993370ef8	0.98906398792756	0	NIV upd_T1a_M0_S SVM
0xf529bd25b4480431bfbba5d6f2e28a74	0.98872565531425	1	NIV upd_T1a_M0_S SVM
0xe2d33655c094ef8a373c0039122d9a7e	0.986079286862359	2	NONE upd_T1a_M0_S SVM
0xd53604e8a14246f6ac13809f369bc9fe	0.98571327083338	3	NONE upd_T1a_M0_S SVM
0xf69395f0e9f8d7696c20d803a23345bd	0.984838635943417	4	NIV upd_T1a_M0_S SVM

Example of the format required for Task 1 submission runs



Run Example: Task 2

Patient Identifier

0xd32d8b8316d08d4c934e0db5b4cbbbab	6-12	NIV	upd_T2a_M6_Cox
0xf529bd25b4480431bfbba5d6f2e28a74	6-12	NONE	upd_T2a_M6_Cox
0xf1a0c725e956c72ee92590d8dcfefb58	6-12	NONE	upd_T2a_M6_Cox
0xf5eeae64002862bbbe0cb0d1667ffa56	12-18	NIV	upd_T2a_M6_Cox
0xdfd6ff34ba7f6abb38fa7ae9894d4d06	12-18	NIV	upd_T2a_M6_Cox

Example of the format required for Task 2 submission runs

Predicted event time window

Variables





Variables

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

Static variables

Dynamic variables

Outcome variables

Performance Metrics

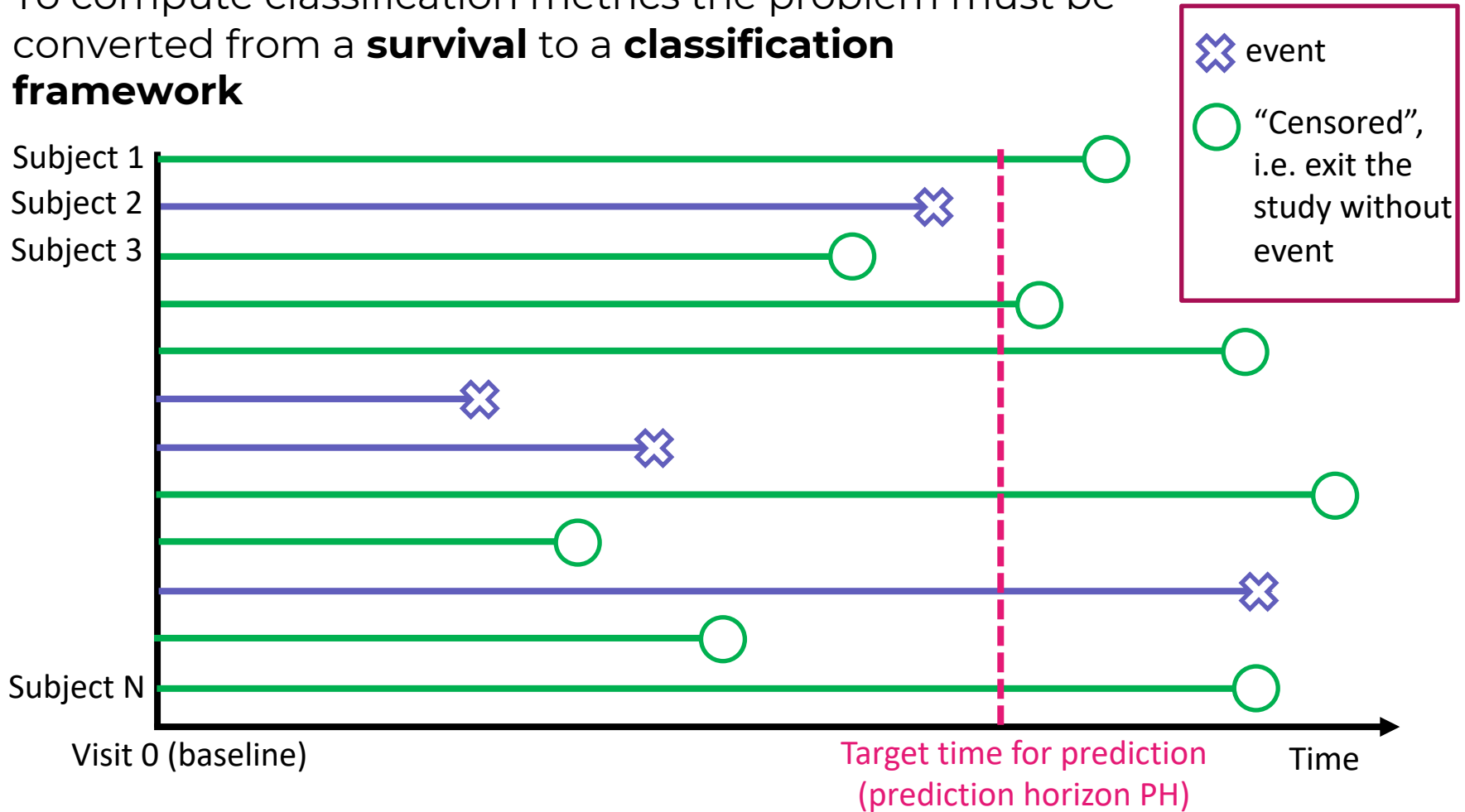
Task 1





From survival to classification

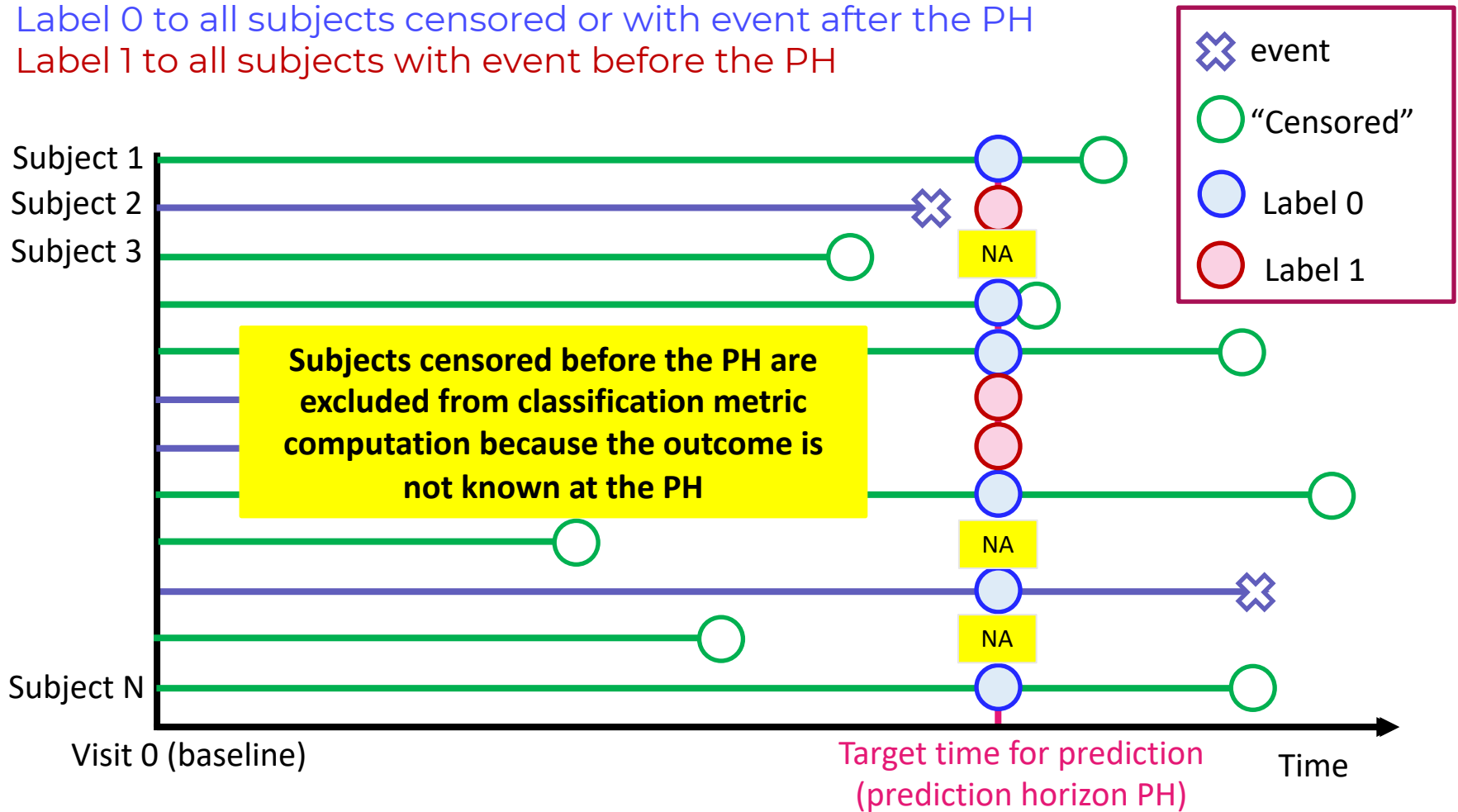
To compute classification metrics the problem must be converted from a **survival** to a **classification framework**





From survival to classification

Label 0 to all subjects censored or with event after the PH
Label 1 to all subjects with event before the PH

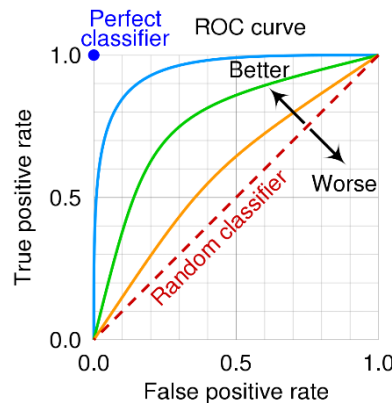




Area Under the Receiver Operating Characteristic Curve (AUROC)

The **AUROC** is a classification metric that evaluates how well a model **discriminates the positive class from the negative one**

It is computed from the ROC curve, obtained by plotting the **true positive rate (TP/P)** against the **false positive rate (1-TN/N)**



The AUROC ranges from 0 to 1, an higher value is associated to a better discrimination performance

A reference value is 0.5, that is the performance of a classifier that assigns labels randomly



Area Under the Receiver Operating Characteristic Curve (AUROC)

```
0xd3ee37821226aa86382711f993370ef8 0.98906398792756 0 NIV upd_T1a_M0_SSVM  
0xf529bd25b4480431bfbba5d6f2e28a74 0.98872565531425 1 NIV upd_T1a_M0_SSVM  
0xe2d33655c094ef8a373c0039122d9a7e 0.986079286862359 2 NONE upd_T1a_M0_SSVM  
0xd53604e8a14246f6ac13809f369bc9fe 0.98571327083338 3 NONE upd_T1a_M0_SSVM  
0xf69395f0e9f8d7696c20d803a23345bd 0.984838635943417 4 NIV upd_T1a_M0_SSVM
```

Example of the format required for Task 1 submission runs

To compute AUROC from the runs the **predicted risk scores** were compared against the binary ground truth (label 0 and label 1)

Considered prediction horizons: **12-18-24-30-36-48-60 months**



Brier score (BS)

The **Brier score** is a score function used to assess **model calibration** (i.e. whether or not predicted probabilities match expected probabilities)

It is computed with the following equation:

$$BS = \frac{1}{N} \sum_{i=1}^N (p_i - l_i)^2$$

Where:

- N is the number of subjects
- p_i is the predicted probability for subject i
- l_i is the ground truth for subject i

Small values of Brier score are associated to **better performance** as predicted probabilities are closer to the corresponding ground truths



Brier score (BS)

```
0xd3ee37821226aa86382711f993370ef8 0.98906398792756 0 NIV upd_T1a_M0_S SVM  
0xf529bd25b4480431bfbba5d6f2e28a74 0.98872565531425 1 NIV upd_T1a_M0_S SVM  
0xe2d33655c094ef8a373c0039122d9a7e 0.986079286862359 2 NONE upd_T1a_M0_S SVM  
0xd53604e8a14246f6ac13809f369bc9fe 0.98571327083338 3 NONE upd_T1a_M0_S SVM  
0xf69395f0e9f8d7696c20d803a23345bd 0.984838635943417 4 NIV upd_T1a_M0_S SVM
```

Example of the format required for Task 1 submission runs

To compute the Brier score from the runs the **risk scores** were compared against the binary ground truth (label 0 and label 1)

Considered prediction horizons: **12-18-24-30-36-48-60 months**



Concordance Index (C-index)

The **concordance index** (C-index) is a generalization of the AUROC that can **take into account censored subjects** and is thus more suitable to evaluate survival approaches

With the C-index model discrimination is assessed as the **ability of the model to assign higher risks to subject who will experience the event sooner**

$$\hat{C} = \frac{\sum_{i=1}^N \Delta_i \sum_{j=i+1}^N I(T_i^{obs} < T_j^{obs}) I(M_i > M_j)}{\sum_{i=1}^N \Delta_i \sum_{j=i+1}^N I(T_i^{obs} < T_j^{obs})}$$

With:

- Δ_i , binary variable, 1 if the subject i experienced the event at some point and 0 if censored
- M predicted risk score of a subject
- T censoring or event times

C-index = 1 indicates perfect concordance

C-index = 0.5 represents a random prediction



Concordance Index (C-index)

```
0xd3ee37821226aa86382711f993370ef8 0.98906398792756 0 NIV upd_T1a_M0_S SVM  
0xf529bd25b4480431bfbba5d6f2e28a74 0.98872565531425 1 NIV upd_T1a_M0_S SVM  
0xe2d33655c094ef8a373c0039122d9a7e 0.986079286862359 2 NONE upd_T1a_M0_S SVM  
0xd53604e8a14246f6ac13809f369bc9fe 0.98571327083338 3 NONE upd_T1a_M0_S SVM  
0xf69395f0e9f8d7696c20d803a23345bd 0.984838635943417 4 NIV upd_T1a_M0_S SVM
```

Example of the format required for Task 1 submission runs

To compute the C-Index from the runs the **predicted risk scores** were used to assess whether **subjects with higher predicted scores experienced the event sooner**

Performance Metrics

Task 2





Confusion Matrix, Recall, Specificity

To check whether the **time interval** in which the event of interest occurs was **correctly predicted**, recall $\frac{TP}{TP+FN}$ and specificity $\frac{TN}{TN+FP}$ were derived from the confusion matrix

The confusion matrix compares the true event time windows with those predicted by the models

		True/Actual		
		6-12	12-18	18-24 ...
Predicted	6-12	4	6	3
	12-18	1	2	0
	18-24 ...	1	2	6

Rows of the matrix correspond to predicted windows

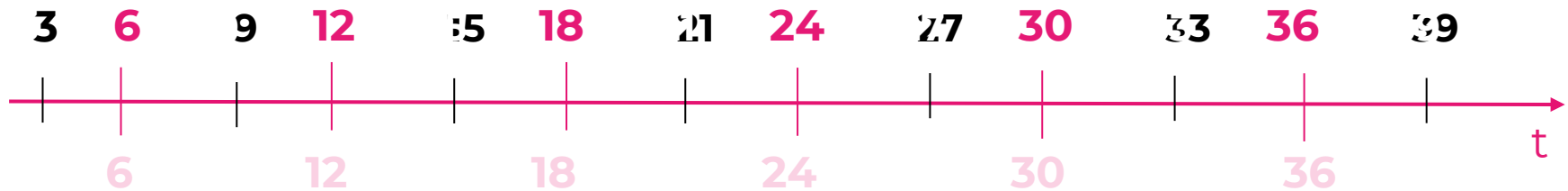
Columns of the matrix correspond to true windows



Absolute Distance (AbsDist)

A measure of **distance between the predicted and correct time intervals**

All the time intervals were replaced with the mean value of each interval



the distance was computed as the absolute value of

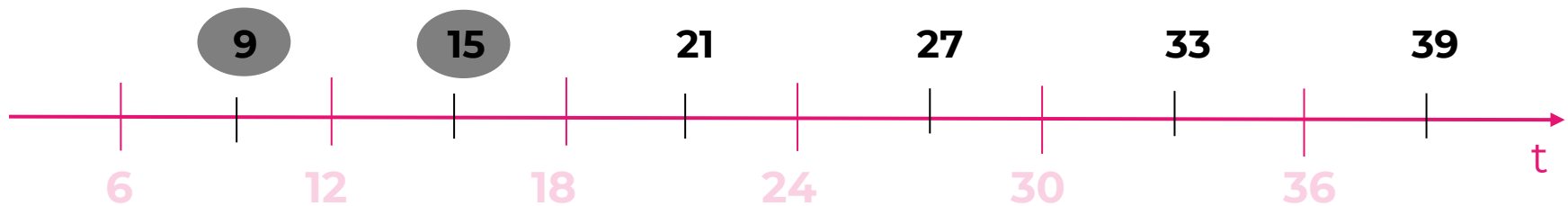
$$(mid\ point_{predicted}) - (mid\ point_{true})$$



Absolute Distance (AbsDist)

A measure of **distance between the predicted and correct time intervals**, was also considered

All the time intervals were replaced with the mean value of each interval



Example:

Predicted time window -> **6-12 months**

(mid point)_{predicted} -> **9 months**

True time window -> **12-18 months**

(mid point)_{true} -> **15 months**

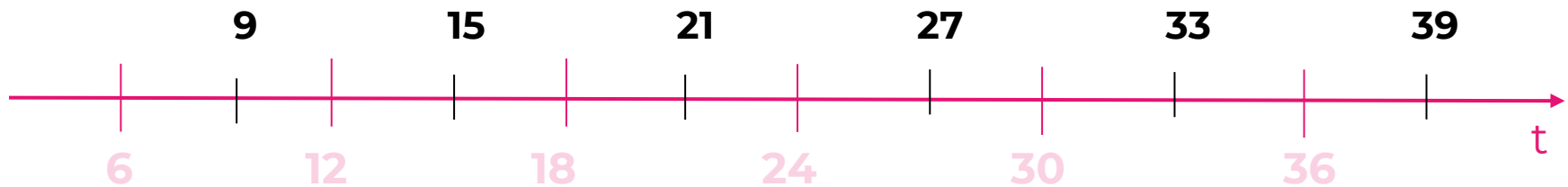
$$\text{AbsDist} = |(mid\ point_{predicted}) - (mid\ point_{true})| = |9-15| = 6\ \text{months}$$



Absolute Distance (AbsDist)

A measure of **distance between the predicted and correct time intervals**, was also considered

All the time intervals were replaced with the mean value of each interval



The final metric was obtained by averaging the absolute differences **across all subjects**

$$\frac{1}{N_{subj}} \sum_{i=1}^{N_{subj}} |(mid\ point_{predicted})_i - (mid\ point_{true})_i|$$

Results Overview

Task 1





Results Overview Task 1

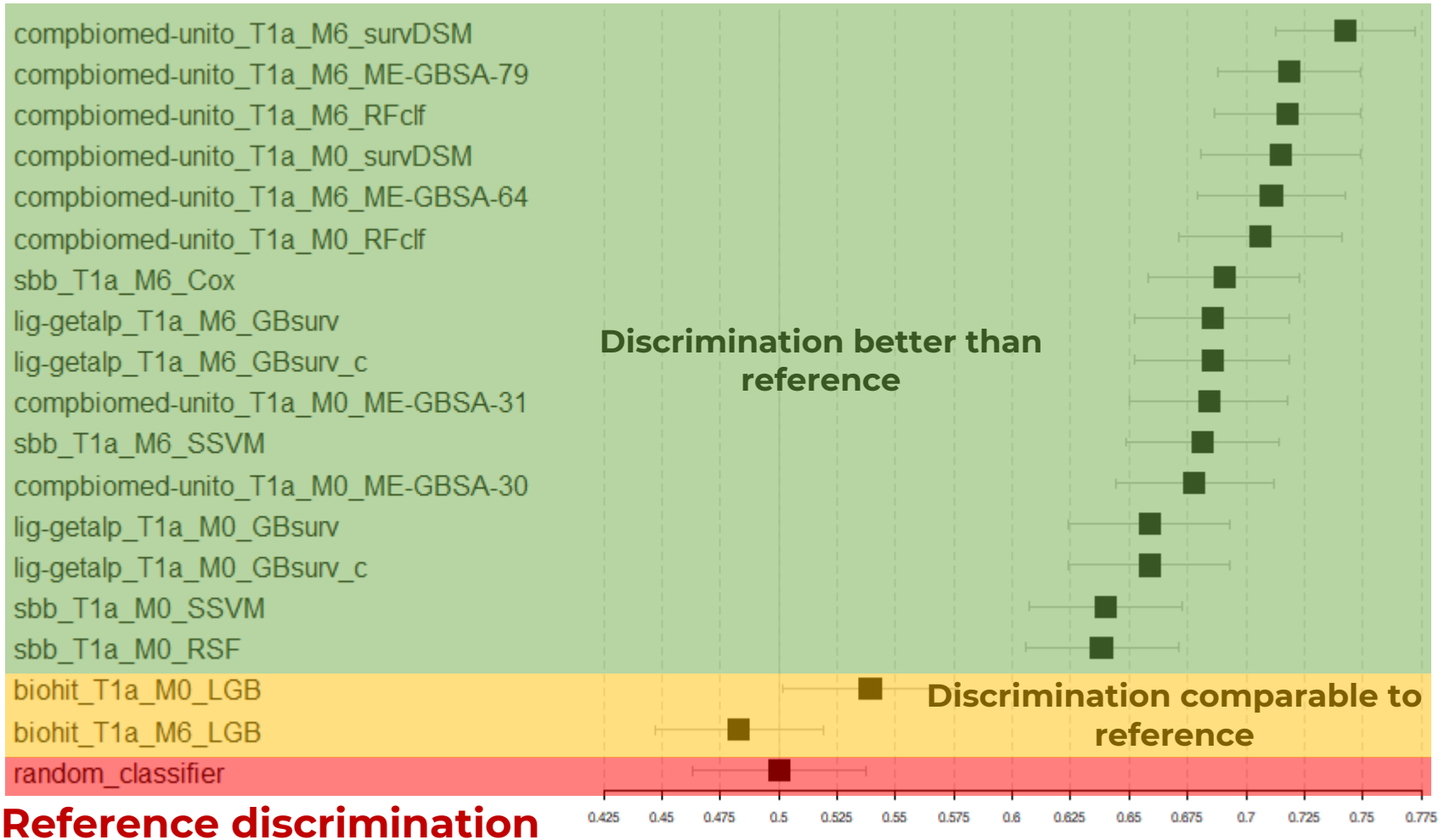
Risk prediction

To evaluate **models developed by participants against a reference**, results obtained from **submitted runs** were compared against the average performance of **100 random classifiers**

Each random classifier was obtained by **assigning risk scores randomly** sampling from a **uniform distribution** ranging from 0 to 1



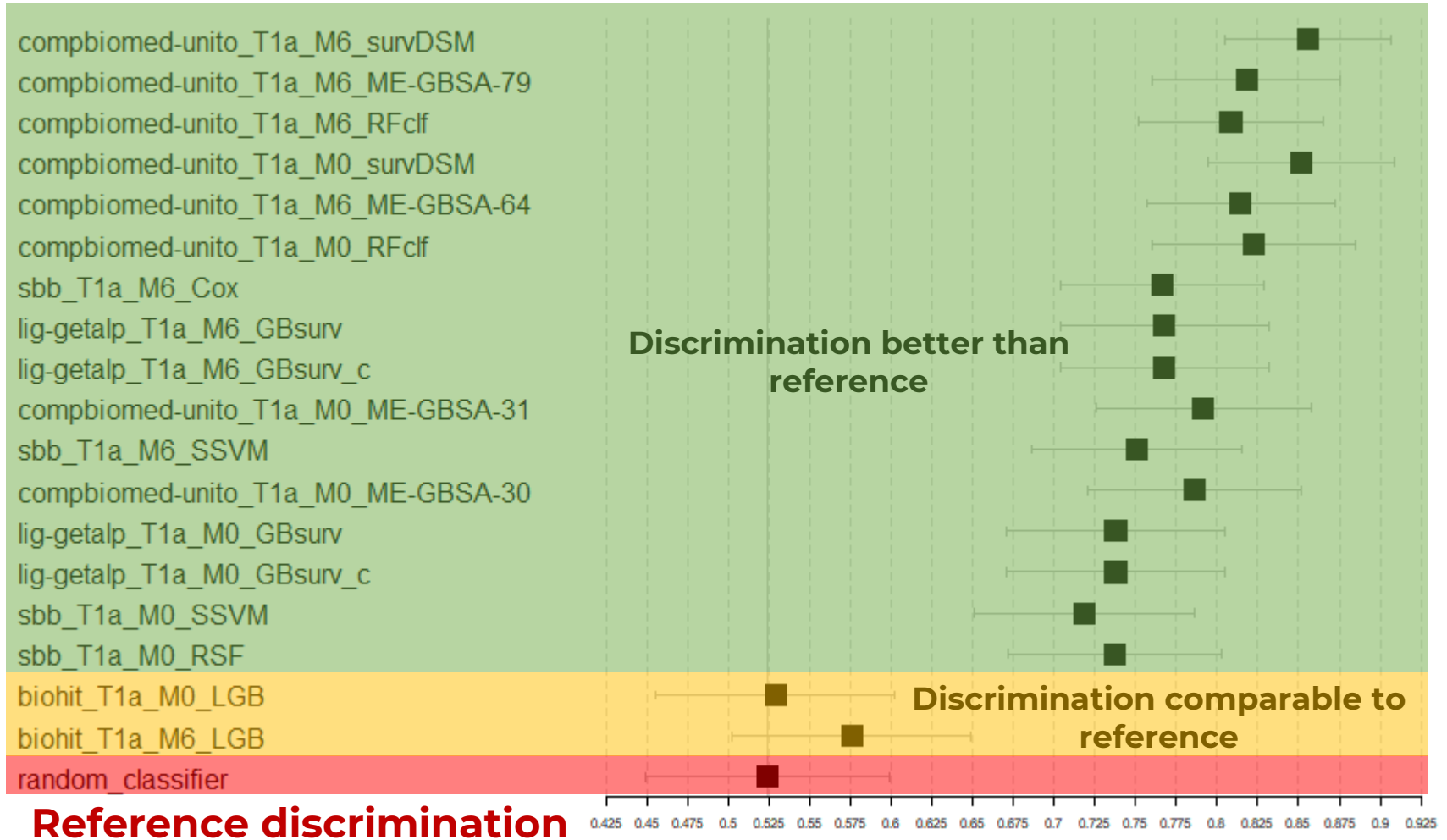
Discrimination C-index Sub-task 1a (NIV or Death)





Discrimination

48-months AUROC Sub-task 1a (NIV or Death)





Results Overview Task 1

Risk prediction (discrimination)

M6 runs (obtained by considering all data available until month 6) led to **better discrimination** than **M0 runs** (obtained by considering only data available at baseline)

Performance differences seem to be more dependent on **task interpretation and data manipulation** rather than methodologic approach

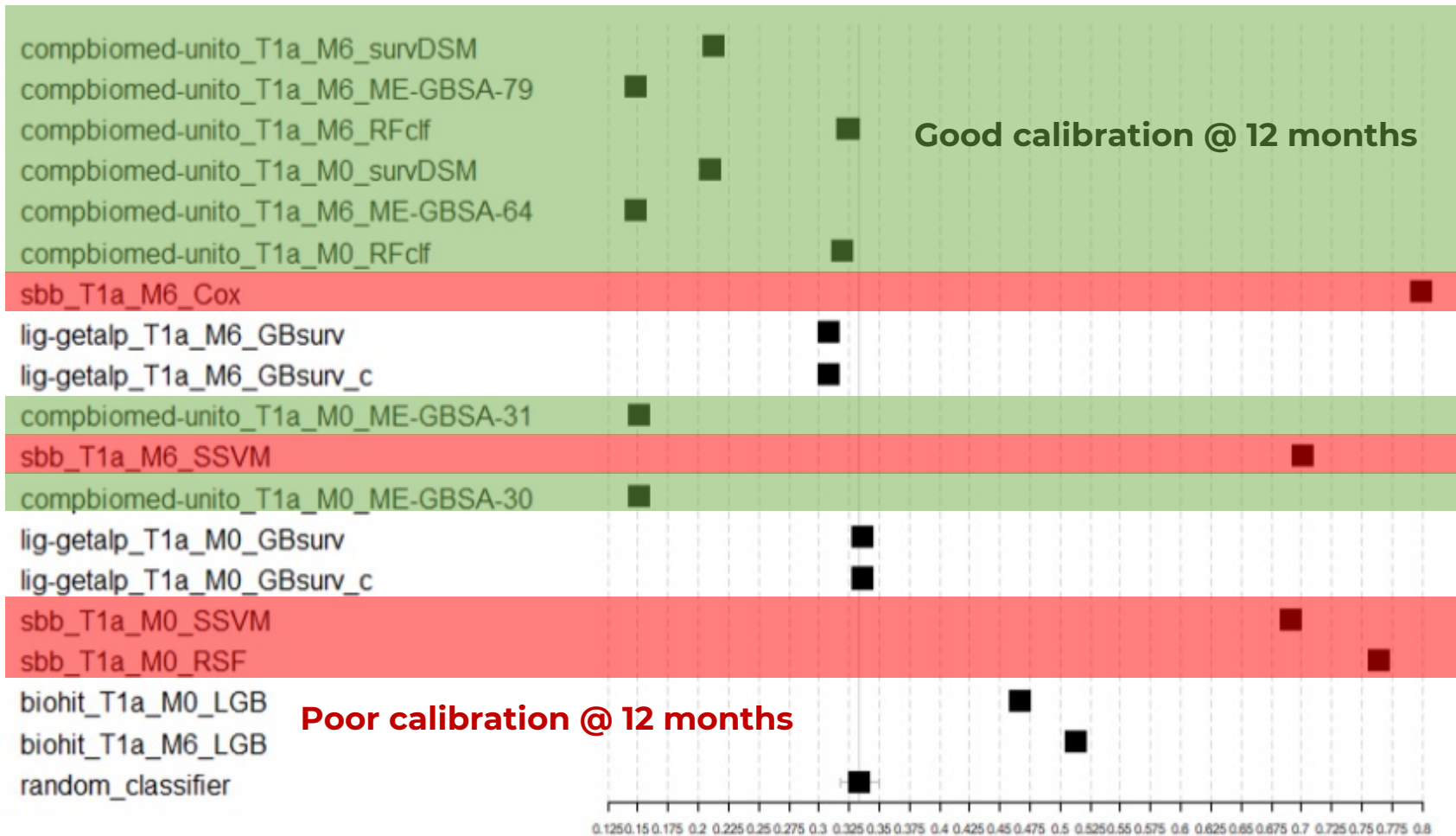
Results improve when:

- ▶ Using a **survival framework**
- ▶ Obtaining relevant features from **dynamic data (e.g., slope, min, max, mean)**



Calibration

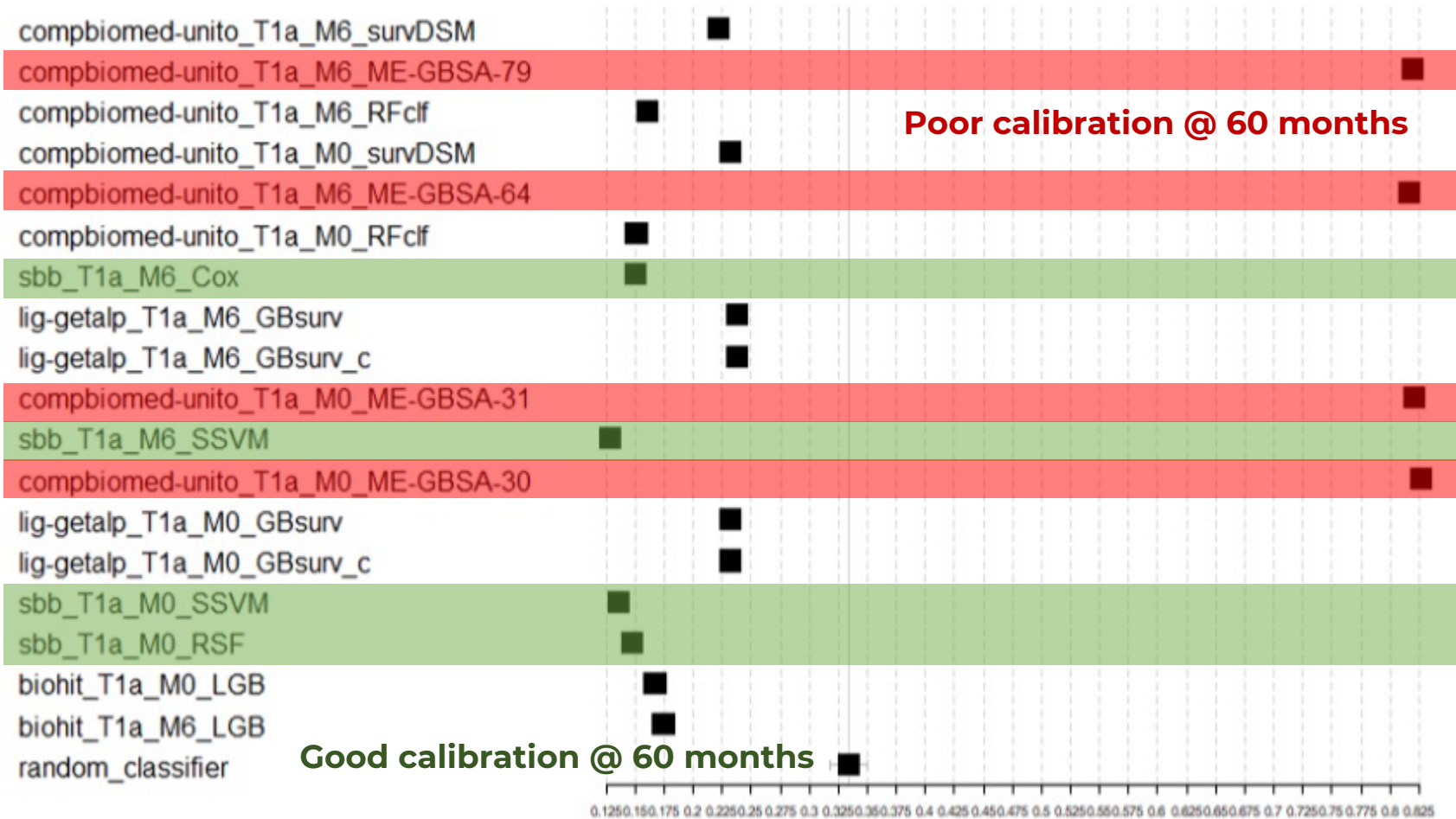
12-months Brier Score Sub-task 1a (NIV or Death)





Calibration

60-months Brier Score Sub-task 1a (NIV or Death)





Results Overview Task 1

Risk prediction (calibration)

M6 runs (obtained by considering all data available until month 6) led to **better calibration** than **M0 runs** (obtained by considering only data available at baseline)

Calibration performance strongly **depends on the considered prediction horizon** as results depend on how the models were trained

Participants may perform a censoring step before training the model to **calibrate the model to the cumulative incidence at a given prediction horizon** or use **recalibration techniques**

Results Overview

Task 2

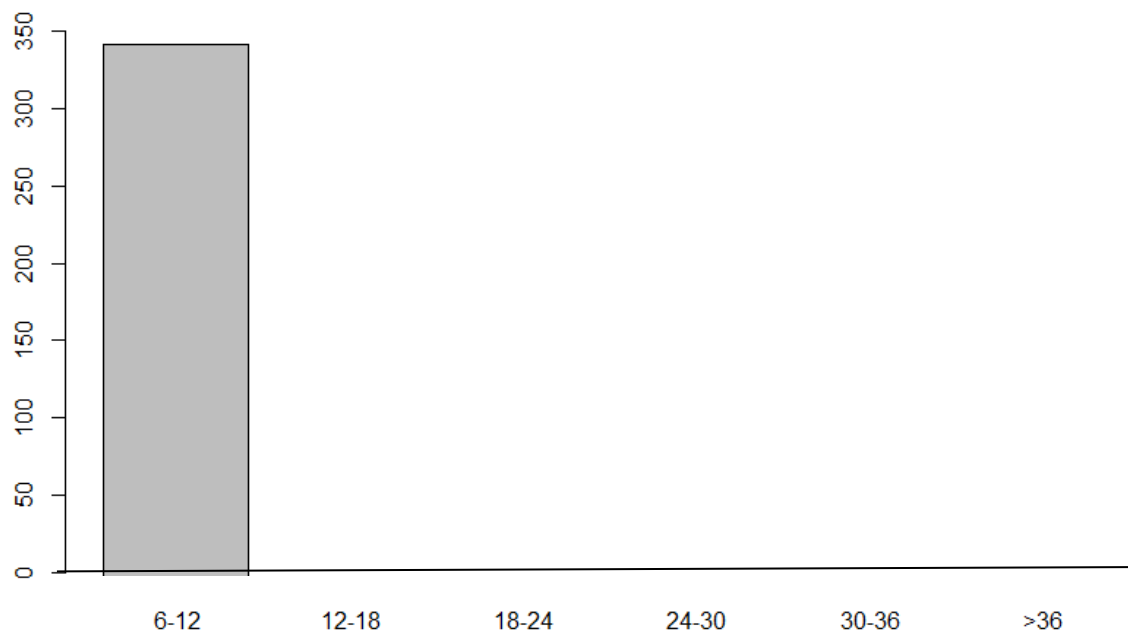




Results Overview T2

To evaluate **models developed by participants against a reference**, results obtained from **submitted runs** were compared against several synthetic runs

- ▶ **min_interval:** predicted time intervals are identical to the minimum one

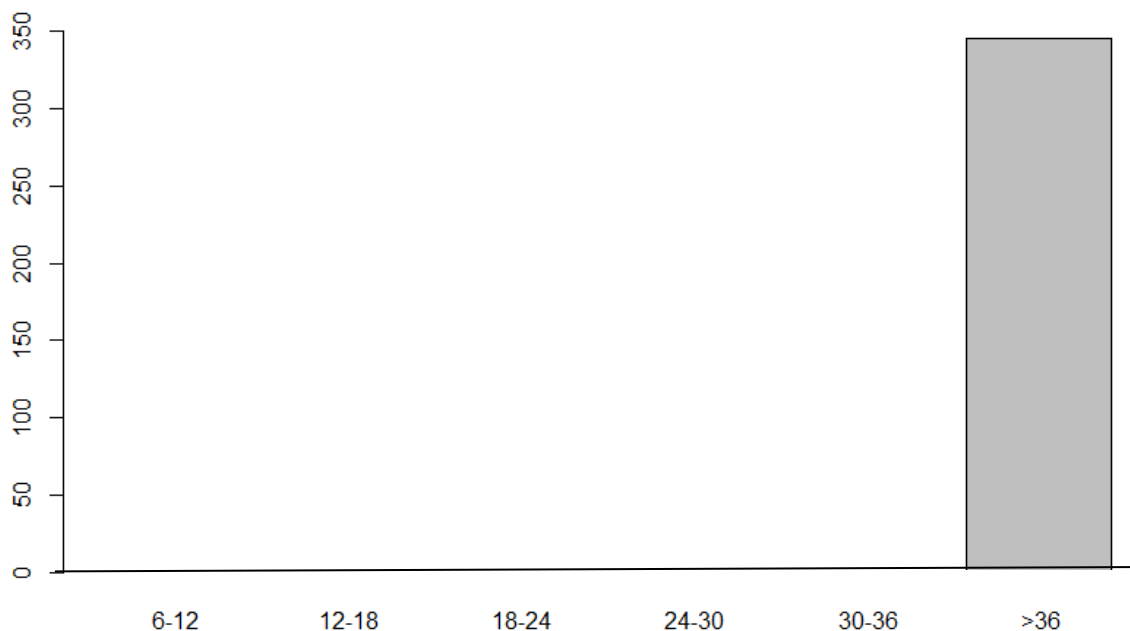




Results Overview T2

To evaluate **models developed by participants against a reference**, results obtained from **submitted runs** were compared against several synthetic runs

- ▶ **max_interval:** predicted time intervals are identical to the maximum one

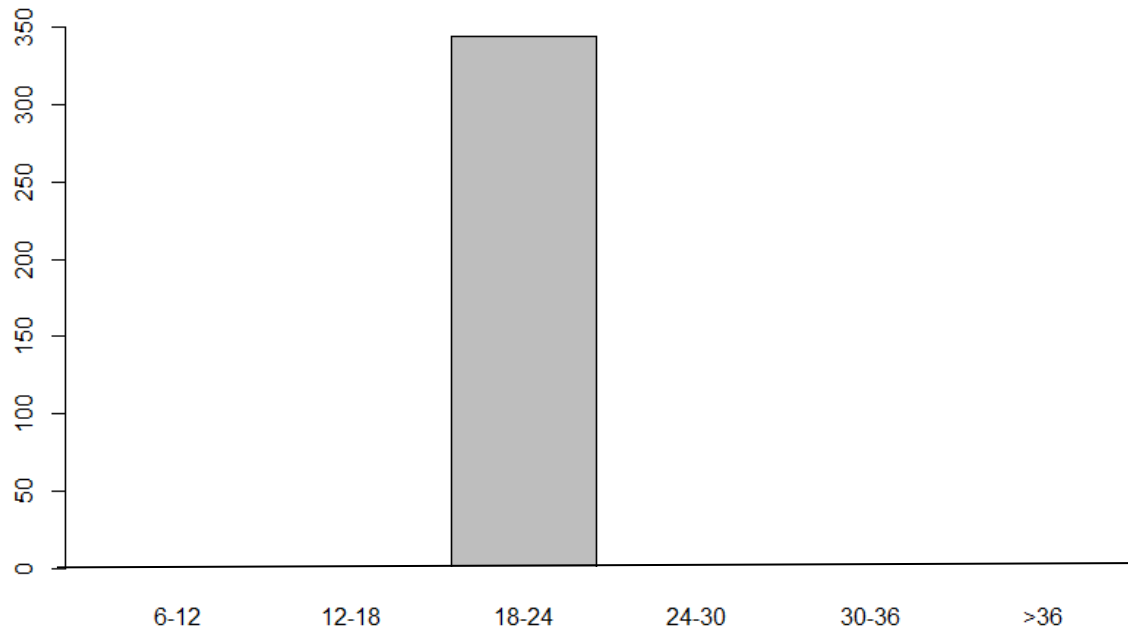




Results Overview T2

To evaluate **models developed by participants against a reference**, results obtained from **submitted runs** were compared against several synthetic runs

- ▶ **interval_18_24**: predicted time intervals are identical to the middle one

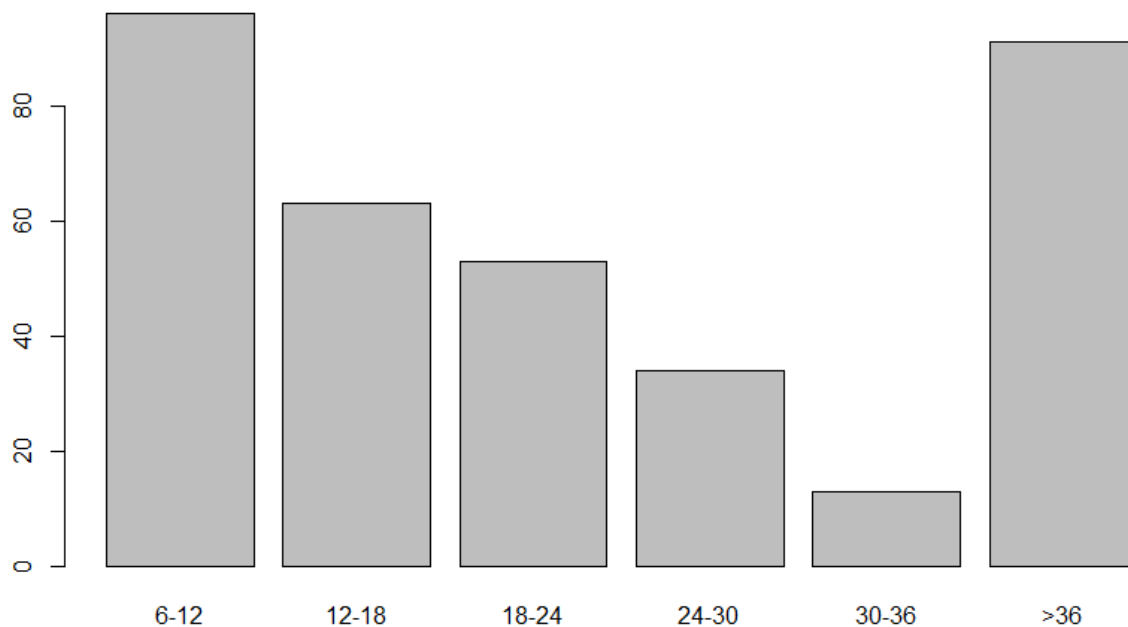




Results Overview T2

To evaluate **models developed by participants against a reference**, results obtained from **submitted runs** were compared against several synthetic runs

- ▶ **random_interval**: 100 randomly generated runs with the same distribution as the test set distribution

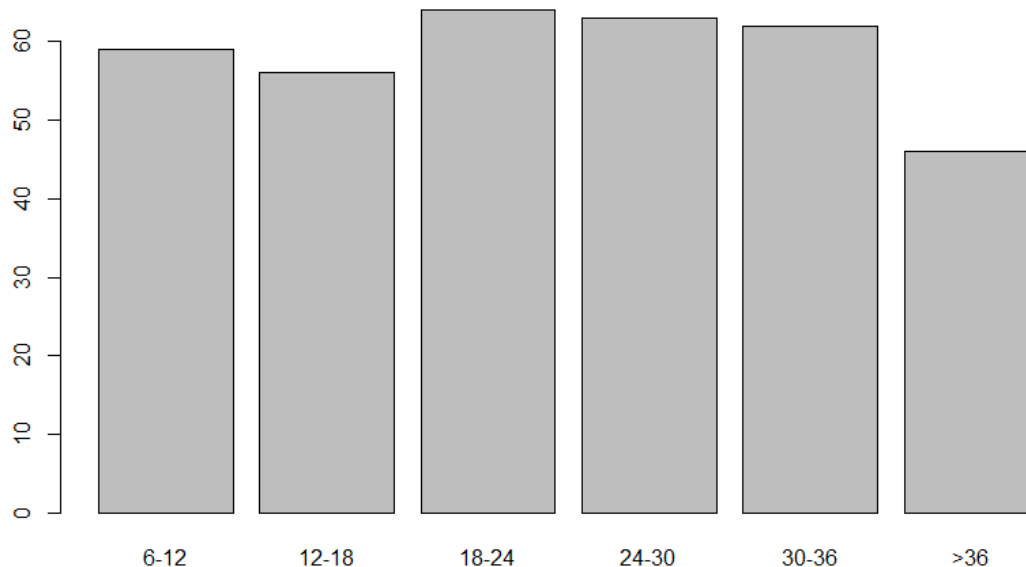




Results Overview T2

To evaluate **models developed by participants against a reference**, results obtained from **submitted runs** were compared against several synthetic runs

- ▶ **Inverse_distr_interval:** 100 randomly generated runs, with an inverse distribution compared to the test set distribution

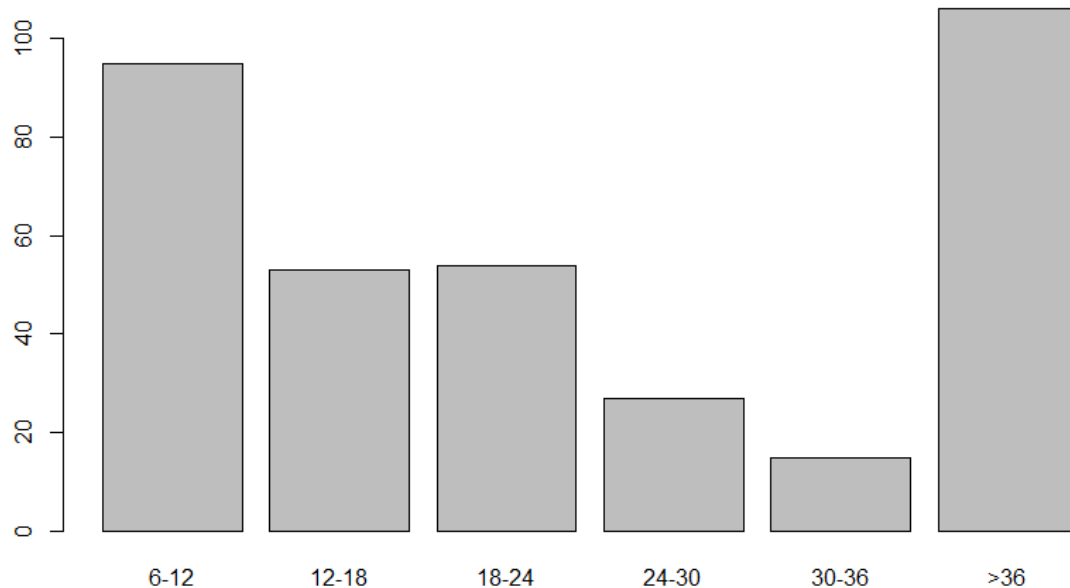




Results Overview T2

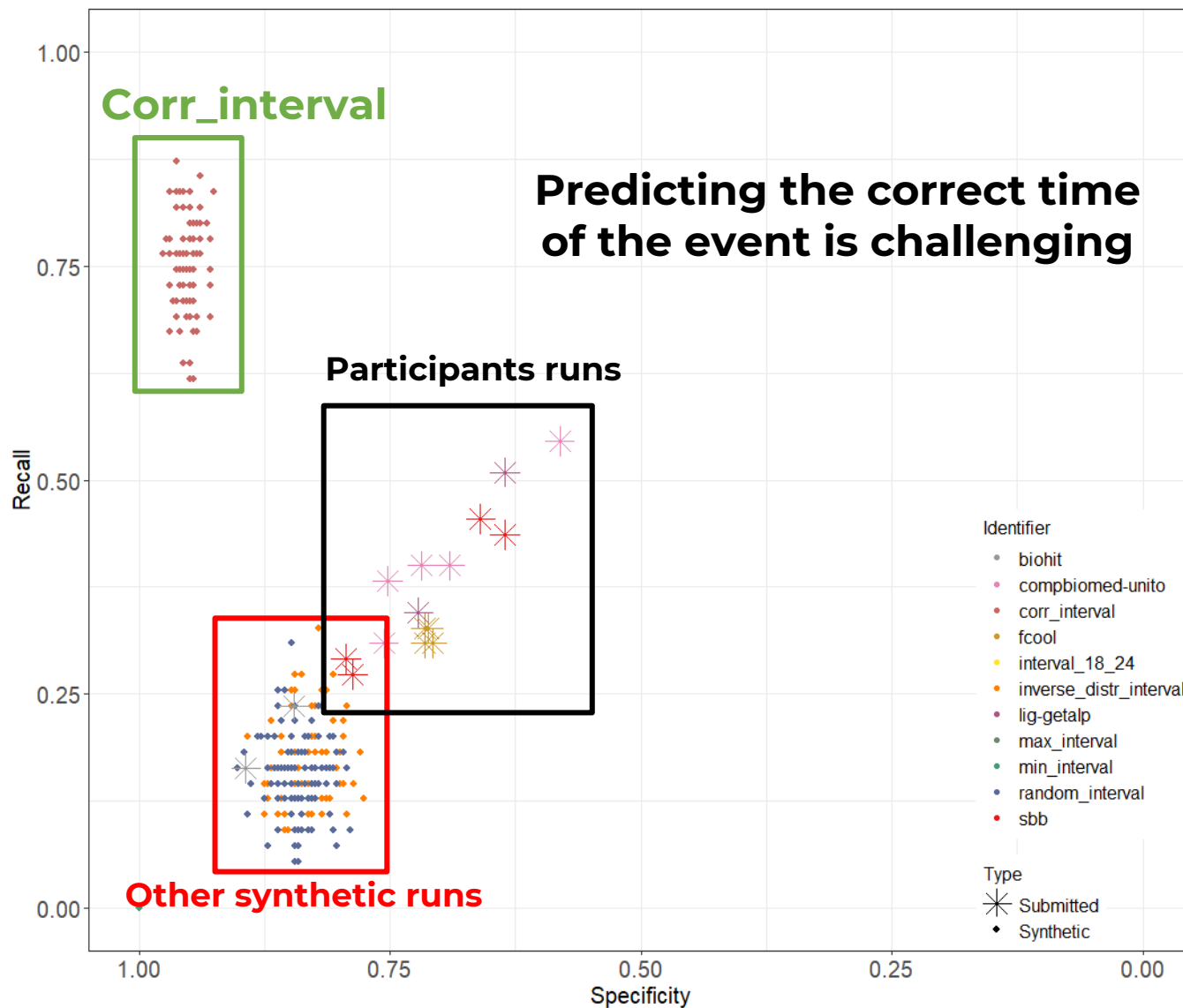
To evaluate **models developed by participants against a reference**, results obtained from **submitted runs** were compared against several synthetic runs

- ▶ **corr_interval:** 100 correlated runs, with correlation coefficient to the true intervals ~ 0.7





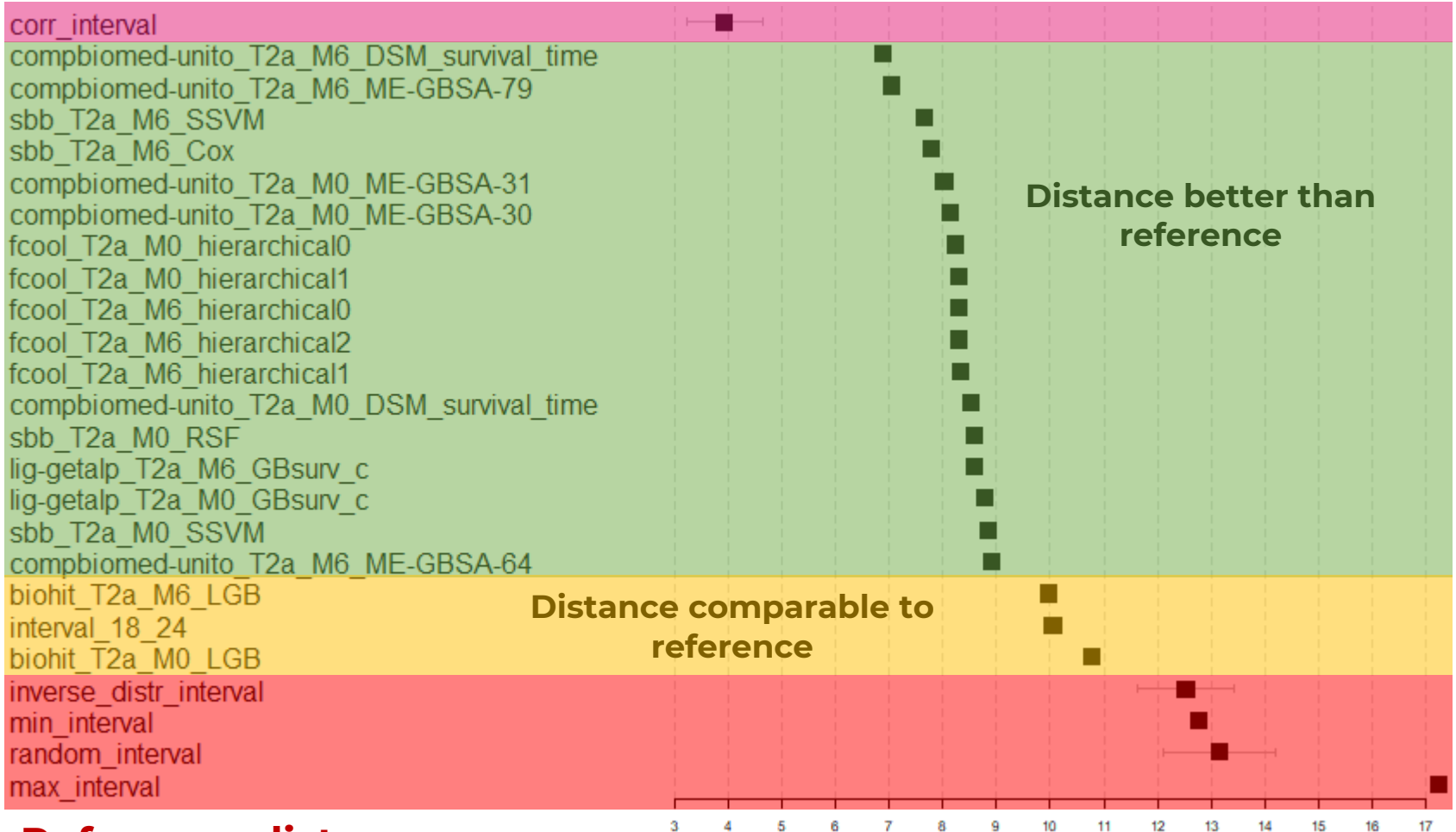
Recall & Specificity Sub-task 2a (NIV or Death) Time Interval 12-18





Absolute Distance Sub-task 2a (NIV or Death)

Optimal benchmark not reached



Reference distance



Results Overview T2

M6 runs (obtained by considering all data available until month 6) led to **better performance** than **M0 runs** (obtained by considering only data available at baseline)

For participants it was **challenging to predict the correct time interval of the event**, however, as absolute distance values were on average around 7 months, they were **predicting adjacent windows**



Thank You



Alessandro Guazzo

Department of Information Engineering
University of Padova



alessandro.guazzo@phd.unipd.it



@sysbiobigunipd





- ▶ 43 teams registered but only 5 submitted... why?
 - ▷ Difficult/complex task ?
 - ▷ Not clear survival / classification framework ?
 - ▷ Not clear how to cope with competing risks ?
 - ▷ Other aspects ?

- ▶ Next year same challenge with air pollution data
 - ▷ Same tasks...
 - ▷ Is air pollution affecting prognosis ?
 - ▷ Making more explicit the importance of **predicting probability of events within subsequent time windows** ?