MIMIC II Data Analysis

Ivana Malenica and Rachael Phillips

September, 2019

Contents

Overview

1 Overview of the Data

We only considered patients that had:

- at least 8 hours of data.
- at most 1 min time gap between two consecutive measurements.

We can see a list of all the covariates available, as well as the basic summary statistic for each below.

[1] “abpsys” “abpdias”
[3] “abpmean” “spo2”
[5] “imputed_abpmean” “imputed_abpsys_abpdias” [7] “hypo_event” “amine”
[9] “sedation” “ventilation”
[11] “rank_icu” “gender”
[13] “age” “sapsi_first”
[15] “sofa_first” “bmi”
[17] “care_unit” “admission_type_descr”
[19] “imputed_age” “imputed_bmi”
[21] “imputed_sofa” “imputed_sapsi”

Data Frame Summary

dat_summary

Dimensions: 327360 x 22 Duplicates: 3597

No

Variable

Stats / Values

Freqs (% of Valid)

Valid

Missing

1

abpsys [numeric]

Mean (sd) : 118.5 (25.1) min < med < max: 22 < 115.2 < 372.1 IQR (CV) : 30.7 (0.2)

2021 distinct values

327360 (100%)

0 (0%)

2

abpdias [numeric]

Mean (sd) : 60.4 (15) min < med < max: -14.4 < 58.6 < 298.6 IQR (CV) : 15.6 (0.2)

1335 distinct values

327360 (100%)

0 (0%)

3

abpmean [numeric]

Mean (sd) : 79.9 (17.1) min < med < max: 7.6 < 77.6 < 300.9 IQR (CV) : 19.4 (0.2)

1565 distinct values

327360 (100%)

0 (0%)

4

spo2 [numeric]

Mean (sd) : 89.8 (26.6) min < med < max: 0 < 98 < 100 IQR (CV) : 4.3 (0.3)

751 distinct values

327360 (100%)

0 (0%)

5

imputed_abpmean [factor]

1. 0
2. 1
   320383
   (
   97.9%
)
   6977
   (
   2.1%

)
327360 (100%)
0 (0%)
6
imputed_abpsys_abpdias [factor]
1. 0

3. 1
314251
(
96.0%
)
13109
(
4.0%
)
327360 (100%)
0 (0%)
7
hypo_event [factor]
1. 0

4. 1
284967
(
87.1%
)
42393
(
13.0%
)
327360 (100%)
0 (0%)
8
amine [factor]
1. 0

5. 1
172461
(
52.7%
)
154899
(
47.3%
)
327360 (100%)
0 (0%)
9
sedation [factor]
1. 0

6. 1
184670
(
56.4%
)
142690

(
43.6%
)
327360 (100%)
0 (0%)
10
ventilation [factor]
1. 0

7. 1
181813
(
55.5%
)
145547
(
44.5%
)
327360 (100%)
0 (0%)
11
rank_icu [factor]
1. 1

8. 10
9. 2
10. 3
11. 4
12. 5
13. 6
285600
(
87.2%
)
0
(
0.0%
)
31200
(
9.5%
)
7200
(
2.2%
)
1920
(
0.6%
)
960
(
0.3%
)
480
(

0.1%
)
327360 (100%)
0 (0%)
12
gender [factor]
1. F
14. M
129600
(
39.6%
)
197760
(
60.4%
)
327360 (100%)
0 (0%)
13
age [numeric]
Mean (sd) : 66.4 (15.4) min < med < max: 21 < 69 < 97 IQR (CV) : 21 (0.2)
75 distinct values
327360 (100%)
0 (0%)
14
sapsi_first [numeric]
Mean (sd) : 15.9 (4.9) min < med < max: 1 < 16 < 34 IQR (CV) : 6 (0.3)
32 distinct values
327360 (100%)
0 (0%)
15
sofa_first [numeric]
Mean (sd) : 7.7 (3.9) min < med < max: 0 < 8 < 20 IQR (CV) : 5 (0.5)
21 distinct values
327360 (100%)
0 (0%)
16
bmi [numeric]
Mean (sd) : 28.6 (5.6) min < med < max: 15.1 < 27.7 < 56.6 IQR (CV) : 3.9 (0.2)
438 distinct values
327360 (100%)
0 (0%)
17
care_unit [factor]
1. CCU
15. CSRU
16. MICU
92160
(
28.1%
)
139680
(
42.7%

)
95520
(
29.2%
)
327360 (100%)
0 (0%)
18
admission_type_descr [factor]
1. ELECTIVE
17. EMERGENCY
18. URGENT
74400
(
22.7%
)
232320
(
71.0%
)
20640
(
6.3%
)
327360 (100%)
0 (0%)
19
imputed_age [factor]
1. 0
19. 1
326880
(
99.9%
)
480
(
0.1%
)
327360 (100%)
0 (0%)
20
imputed_bmi [factor]
1. 0
20. 1
228480
(
69.8%
)
98880
(
30.2%
)
327360 (100%)
0 (0%)

21
imputed_sofa [factor]
1. 0
21. 1
324000
(
99.0%
)
3360
(
1.0%
)
327360 (100%)
0 (0%)
22
imputed_sapsi [factor]
1. 0
22. 1
314400
(
96.0%
)
12960
(
4.0%
)
327360 (100%)
0 (0%)

We further explore the number of total hypotensive episodes experiences per each patient.

Data Frame Summary

df

Dimensions: 682 x 1 Duplicates: 481

No

Variable

Stats / Values

Freqs (% of Valid)

Valid

Missing

1

sum_all_events [numeric]

Mean (sd) : 62.2 (98.1) min < med < max: 0 < 17 < 480 IQR (CV) : 85 (1.6)

201 distinct values

682 (100%)

0 (0%)

Finally, we expore how many patients had at least one episode. 442 of the 698 subjects experienced at least one hypotensive event, and the outcome function `Y1` was used to specify hypotensive events. By definition, an hypotensive episode is defined as a 5 minute window with mean arterial pressure (MAP) below 62 mmHg.

Data Frame Summary

event_summary

Dimensions: 2 x 1 Duplicates: 0

No

Variable

Stats / Values

Freqs (% of Valid)

Valid

Missing

1

Number of Events [integer]

Min : 256 Mean : 341 Max : 426

256

:

1

(

50.0%

)

426

:

1

(

50.0%

)

2 (100%)

0 (0%)

2 Prepare Data for the Analysis

Below we list covariates we use for the further analysis. In particular, we can classify them as follows:

1. Baseline Covariates

```
##  [1] "gender"                "age"                   "care_unit"
##  [4] "admission_type_descr" "sapsi_first"           "sofa_first"
##  [7] "bmi"                   "rank_icu"              "imputed_age"
## [10] "imputed_bmi"           "imputed_sofa"          "imputed_sapsi"
```

1.Time-varying Covariates

```
## [1] "amine"           "sedation"        "ventilation"     "spo2"
## [5] "hr"              "abpmean"         "imputed_abpmean"
```

3 Build the Combined Super Learner

The combined online super learner also uses the individual super learner, which learns only from one sample at a time. For the individual super learner, we incorporate the above described covariates as well. In addition, we consider two different Cross-Validation schemes:

- Rolling Origin:
  - initial training set size 15 minutes
  - test set size 15 minutes
  - increase training set size by increments of 5 minutes
- Rolling Window:
  - each window size is 15 minutes
  - test set size 15 minutes
  - increase training set size by increments of 5 minutes

For the combined super learner, we incorporate a gap of 30 minutes between the last trained time point and the first prediction time point. If round hour, we include a gap of 0, due to the data-setup. Therefore, the prediction is for a 15 minute period 30 minutes in the future (since the last trained time-point).

As explored in previous simulations, we only consider the binary outcome, instead of the continuous (even though the combined Super Learner has support for both).

For the base learning library, we consider 8 variations of xgboost:

[1] “Lrnr_xgboost_20_1_4_0.001” “Lrnr_xgboost_20_1_8_0.001” [3] “Lrnr_xgboost_20_1_4_0.01” “Lrnr_xgboost_20_1_8_0.01” [5] “Lrnr_xgboost_20_1_4_0.1” “Lrnr_xgboost_20_1_8_0.1”
[7] “Lrnr_xgboost_20_1_4_0.2” “Lrnr_xgboost_20_1_8_0.2”

4 Examine Results for Combined Super Learner

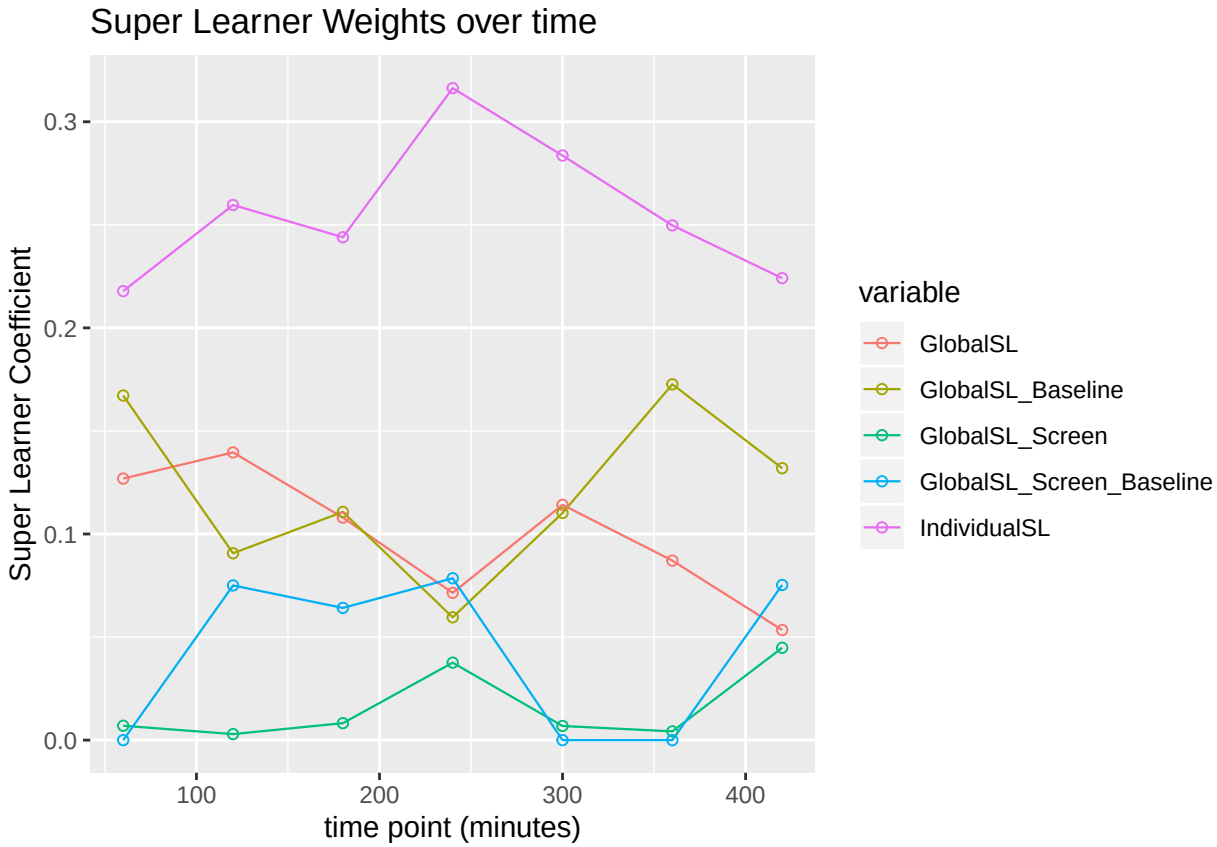
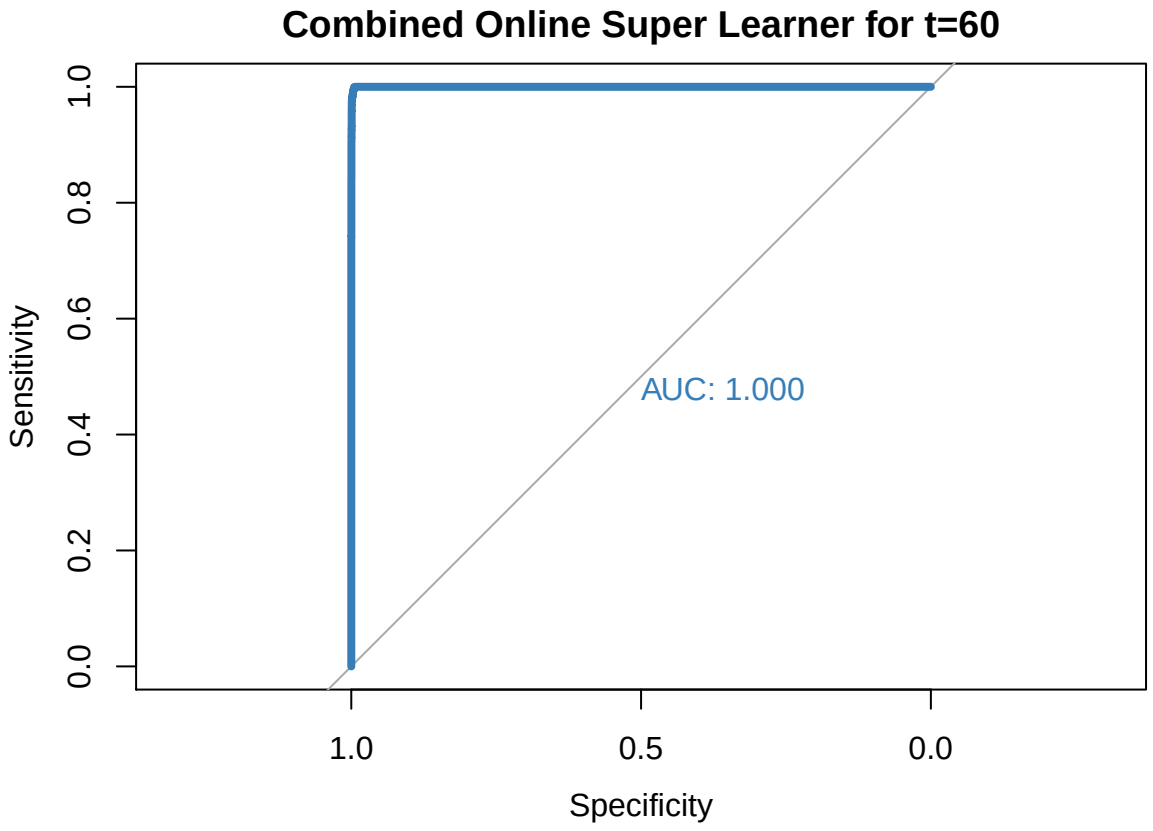


Table 1: **Risk for all different SLs considered**

	loss_online_SL	loss_regular_SL	loss_individual_SL
t=60	0.1858675402	0.2130141915	0.7316000287
t=120	0.1642030357	0.1921674220	0.5049301639
t=180	0.1835364538	0.2069144820	0.4828999435
t=240	0.1662170399	0.1998173694	0.3666914719
t=300	0.1917679758	0.2188869184	0.3407229895
t=360	0.1620297859	0.1886945592	0.3523024722
t=420	0.1782965104	0.2044575612	0.2996297284

4.1 AUC across various training times

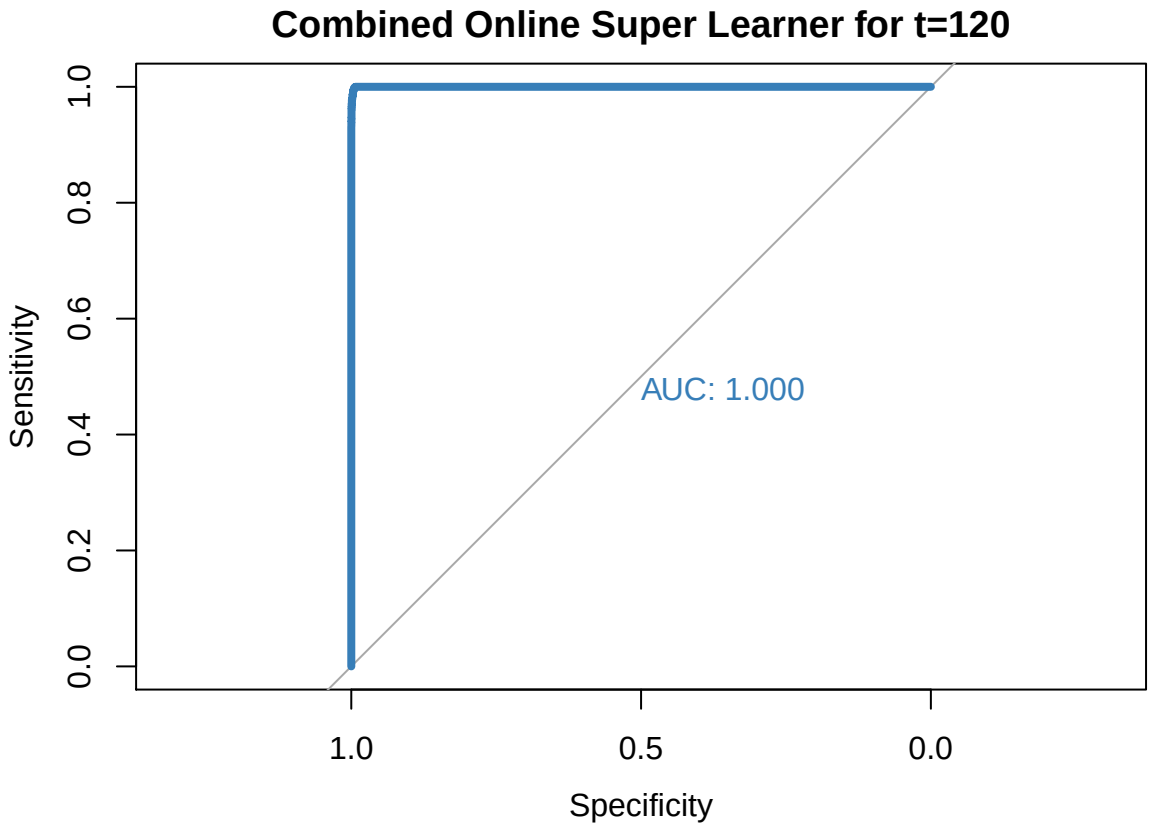
```
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```



Call: roc.default(response = calc_t60$truth, predictor = calc_t60$pred_f_in, plot = TRUE, col = "#377eb8", lwd = 4, print.auc = TRUE, main = "Combined Online Super Learner for t=60")

Data: calc_t60$pred_f_in in 8845 controls (calc_t60$truth 0) < 1385 cases (calc_t60$truth 1). Area under the curve: 0.9999

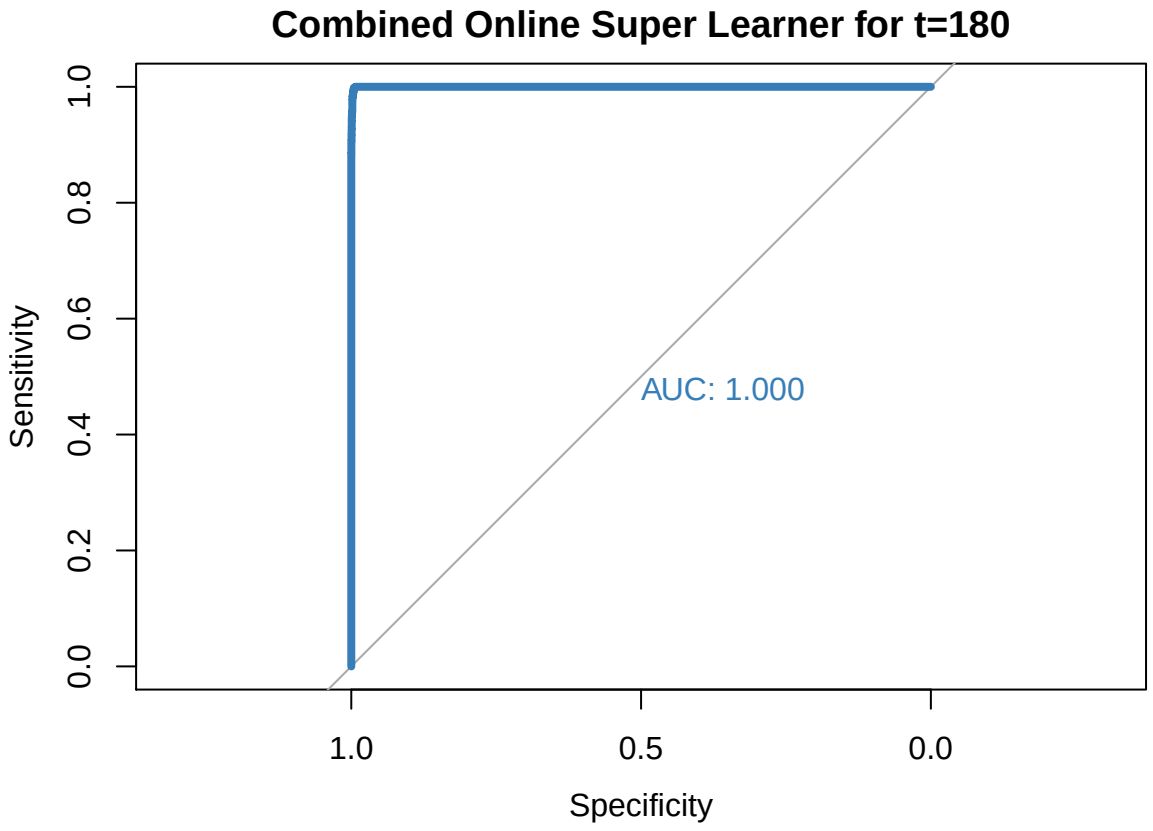
```
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```



Call: roc.default(response = calc_t120$truth$truth, predictor = calc_t120$pred_f in$pred, plot = TRUE, col = "#377eb8", lwd = 4, print.auc = TRUE, main = "Combined Online Super Learner for t=120")

Data: calc_t120$pred_f in$pred in 8885 controls (calc_t120$truth$truth 0) < 1345 cases (calc_t120$truth$truth 1). Area under the curve: 0.9999

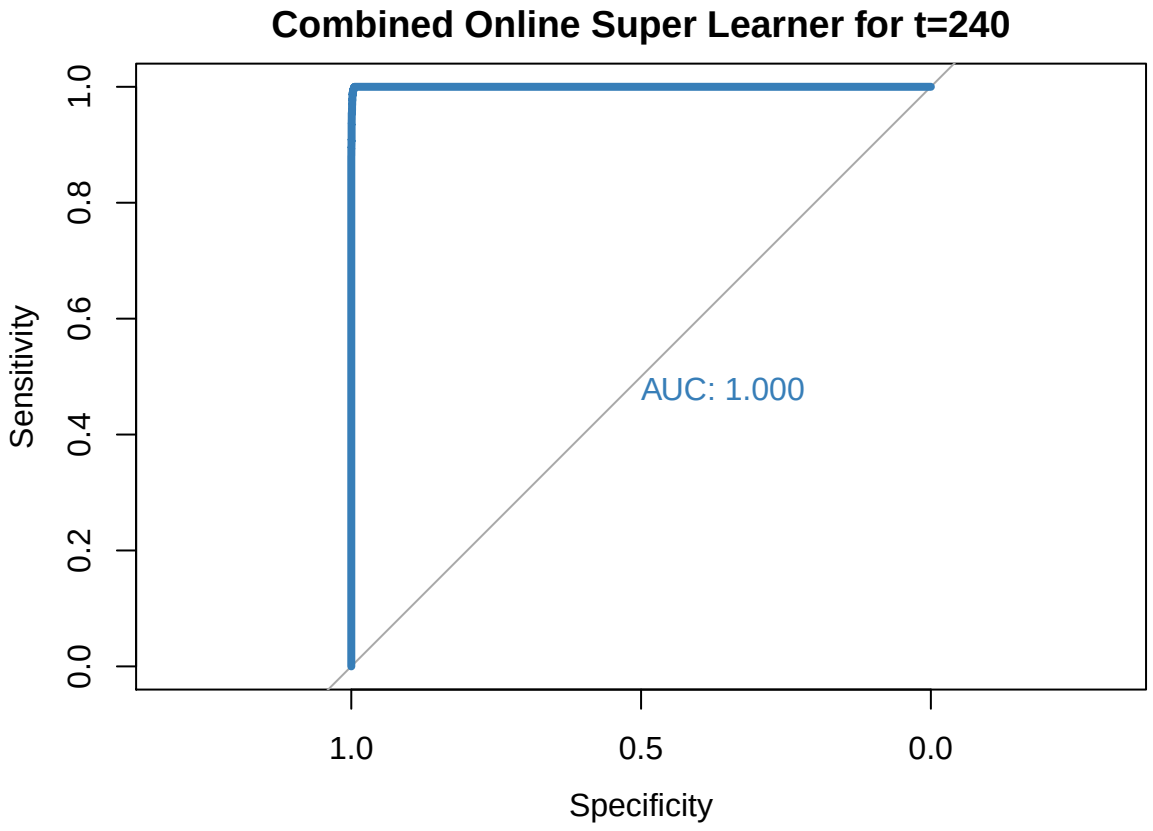
```
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```



Call: roc.default(response = calc_t180$truth$truth, predictor = calc_t180$pred_f in$pred, plot = TRUE, col = "#377eb8", lwd = 4, print.auc = TRUE, main = "Combined Online Super Learner for t=180")

Data: calc_t180$pred_f in$pred in 8916 controls (calc_t180$truth$truth 0) < 1314 cases (calc_t180$truth$truth 1). Area under the curve: 0.9998

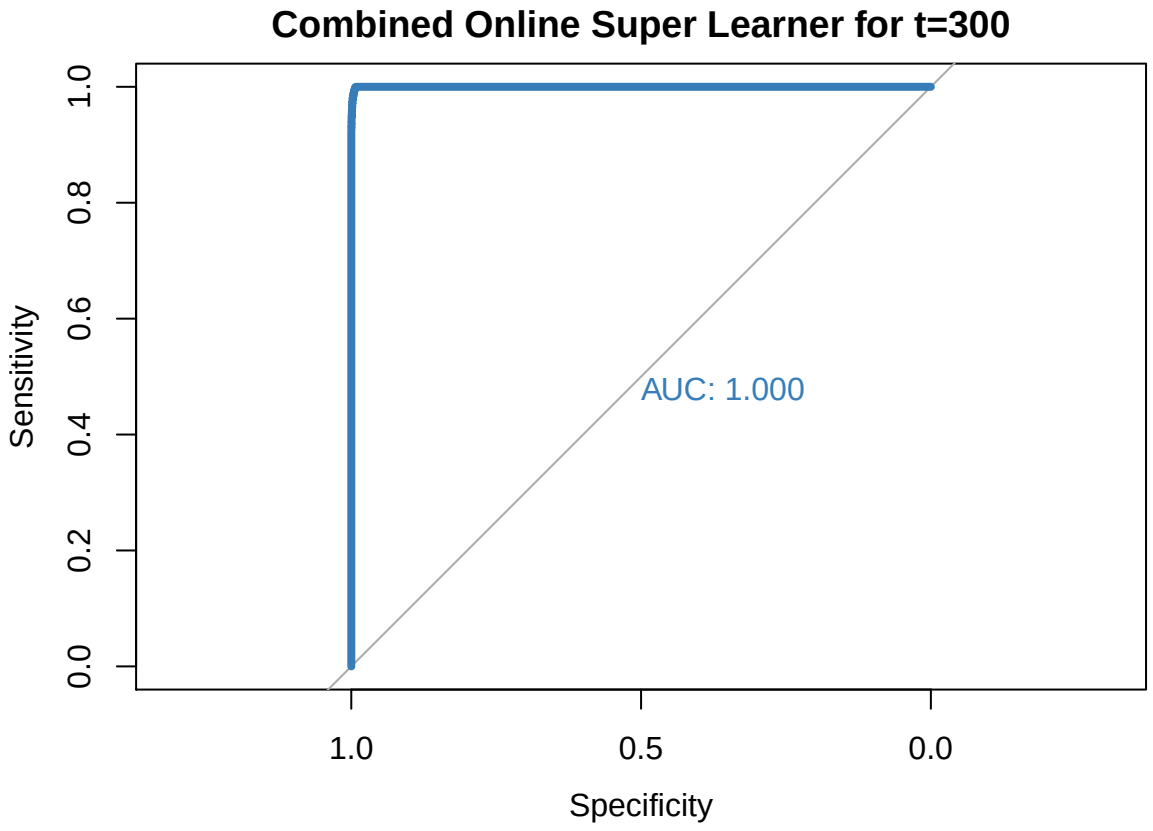
```
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```



Call: roc.default(response = calc_t240$truth$truth, predictor = calc_t240$pred_f in$pred, plot = TRUE, col = "#377eb8", lwd = 4, print.auc = TRUE, main = "Combined Online Super Learner for t=240")

Data: calc_t240$pred_f in$pred in 8956 controls (calc_t240$truth$truth 0) < 1274 cases (calc_t240$truth$truth 1). Area under the curve: 0.9999

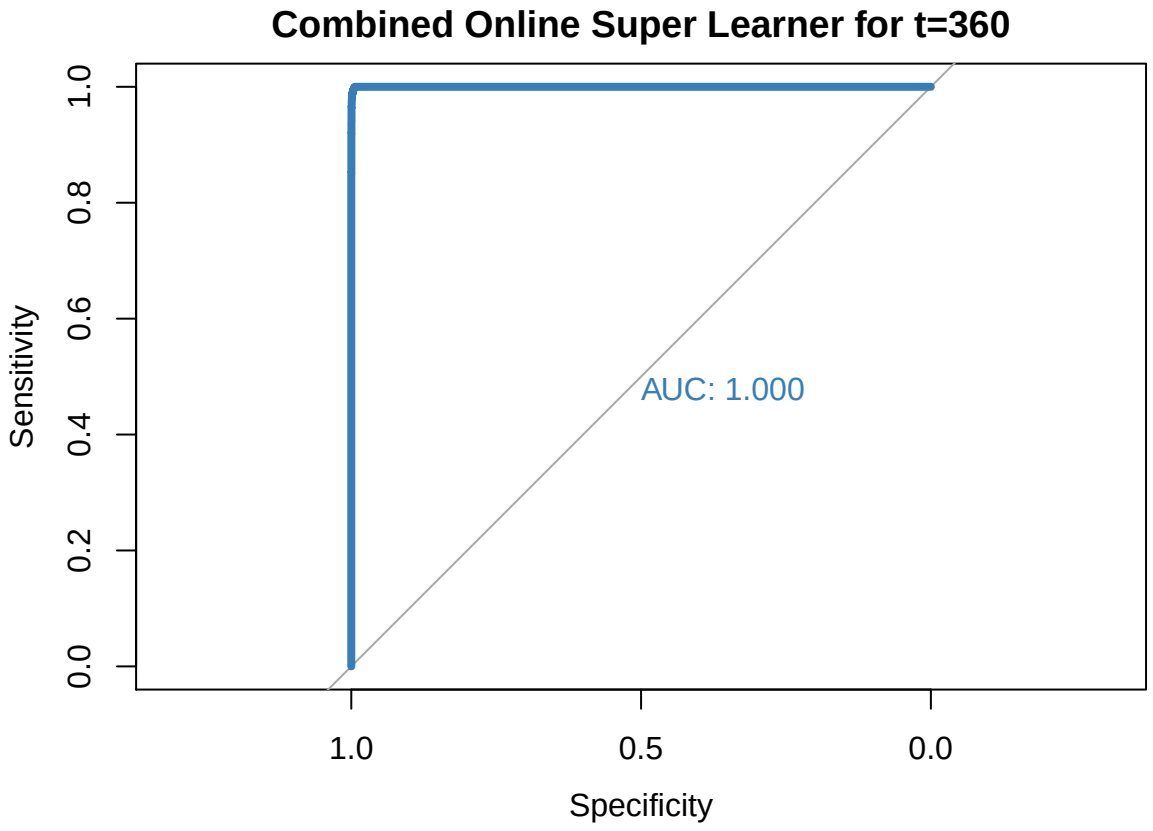
```
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```



Call: roc.default(response = calc_t300$truth$truth, predictor = calc_t300$pred_f in$pred, plot = TRUE, col = "#377eb8", lwd = 4, print.auc = TRUE, main = "Combined Online Super Learner for t=300")

Data: calc_t300$pred_f in$pred in 8796 controls (calc_t300$truth$truth 0) < 1434 cases (calc_t300$truth$truth 1). Area under the curve: 0.9999

```
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```



Call:
roc.default(response = calc_t360$truth$truth, predictor = calc_t360$pred_f in$pred, plot = TRUE, col = "#377eb8", lwd = 4, print.auc = TRUE, main = "Combined Online Super Learner for t=360")

Data: calc_t360$pred_f in$pred in 8805 controls (calc_t360$truth$truth 0) < 1425 cases (calc_t360$truth$truth 1). Area under the curve: 0.9999