



HAL
open science

The Physiological Deep Learner: First application of multitask deep learning to predict hypotension in critically ill patients

Ményssa Cherifa, Yannet Interian, Alice Blet, Matthieu Resche-Rigon,
Romain Pirracchio

► To cite this version:

Ményssa Cherifa, Yannet Interian, Alice Blet, Matthieu Resche-Rigon, Romain Pirracchio. The Physiological Deep Learner: First application of multitask deep learning to predict hypotension in critically ill patients. *Artificial Intelligence in Medicine*, 2021, 118, pp.102118. 10.1016/j.artmed.2021.102118 . hal-04182354

HAL Id: hal-04182354

<https://hal.science/hal-04182354>

Submitted on 22 Jul 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

The Physiological Deep Learner: first application of multitask deep learning to predict hypotension in critically ill patients

Ményssa Cherifa^a, Yannet Interian^b, Alice Blet^{c,d,1}, Matthieu Resche-Rigon^{a,f,1}, Romain Pirrachio^{a,g,1,*}

^aUniversité de Paris, ECSTRRA team, Center of research in epidemiology and statistics (CRESS) - INSERM UMR 1153, 1 Parvis Notre-Dame - Pl. Jean-Paul II, Paris, 75004, France

^bData Analytic Program, University of California San Francisco, 101 Howard St, San Francisco, 94105, CA, United States of America

^cDepartment of Anesthesiology, Critical Care and Burn Center, Lariboisière-Saint-Louis Hospital, AP-HP Nord, 1 Avenue Claude Vellefaux, Paris, 75010, France

^dUniversité de Paris, Cardiovascular MARKers in Stress Conditions (MASCOT) - INSERM UMR-S 942, 41, boulevard de la Chapelle, Paris, 75010, France

^eUniversity of Ottawa Heart Institute, University of Ottawa, 40 Ruskin St, Ottawa, Ontario, Canada

^fDepartment of biostatistics and medical information, Lariboisière-Saint-Louis Hospital, AP-HP Nord, 1 Avenue Claude Vellefaux, Paris, 75010, France

^gDepartment of Anesthesia and Perioperative Medicine, Zuckerberg San Francisco General Hospital and Trauma Center, University of California San Francisco, 1001 Potrero Ave, San Francisco, 94110, CA, United States of America

Abstract

Critical care clinicians are trained to analyze simultaneously multiple physiological parameters to predict critical conditions such as hemodynamic instability. We developed the Multitask Learning Physiological Deep Learner (MTL-PDL), a deep learning algorithm that predicts simultaneously the mean arterial pressure (MAP) and the heart rate (HR).

In an external validation dataset, our model exhibited very good calibration: R^2 of 0.747 (95% confidence interval, 0.692 to 0.794) and 0.850 (0.815 to 0.879) for respectively, MAP and HR prediction 60-minutes ahead of time. For acute hypotensive episodes defined as a MAP below 65 mmHg for 5 minutes, our MTL-PDL reached a predictive value of 90% for patients at very high risk (predicted MAP \leq 60 mmHg) and 2% for patients at low risk (predicted MAP $>$ 70 mmHg).

Based on its excellent prediction performance, the Physiological Deep Learner has the potential to help the clinician proactively adjust the treatment in order to avoid hypotensive episodes and end-organ hypoperfusion.

Keywords: Critical Care, Shock Hypotension, Multitask Learning, RNN

1. Introduction

Shock is the clinical presentation of an acute circulatory failure resulting in inadequate cellular oxygen supply and utilization.¹ It is a common condition that affects approximately

*Corresponding author

¹both equally contributed

4 one-third of the patients in the intensive care unit (ICU).² It is clinically characterized by
5 an acute hypotension, a rapid decline in the mean arterial pressure (MAP).¹ Shock is a
6 diagnostic and therapeutic emergency since any delay in treatment initiation may result in
7 increased morbi-mortality.^{3,4} Therefore, early identification of patients at risk of shock is of
8 utmost importance.

9 As acute hypotension is a key clinical symptom of shock,¹ many studies have attempted
10 to learn from the arterial blood pressure signal to develop prediction models for acute hy-
11 potensive episodes (AHE).^{5,6,7,8,9,10,11,12,13,14,15,16} Most predictive models only use the MAP
12 signal as input to predict the risk of hypotension. However, in practice, clinicians are trained
13 to analyze simultaneously multiple sources of information (including several physiological pa-
14 rameters, disease characteristics, treatments)¹⁷ to better stratify patient severity and predict
15 any forthcoming deterioration. In a previous work, we showed that an ensemble machine
16 learning model trained on baseline patient characteristics, severity scores, ICU treatments,
17 and several continuous physiological signals (including heart rate (HR), blood pressure (BP)
18 and pulse oximetry (SpO_2) was very accurate at predicting AHE up to 30 minutes (min) in
19 advance.¹⁸ In the present study, we propose to augment the concept of learning from multi-
20 ple physiological parameters by using a multi-task learning (MTL) approach to improve the
21 prediction of AHE in critically ill patients.

22 The goal of MTL is to mimic the intensivist’s behavior, namely to learn jointly and
23 simultaneously multiple related outcomes to enhance model performance across tasks, as
24 opposed to learning each task independently (a.k.a. single-task learning, STL).¹⁹ Multi-
25 task neural networks have been proposed to predict a variety of clinical outcomes, such as
26 postoperative mortality, acute kidney injury, and reintubation,²⁰ mortality,²¹ hospital length
27 of stay, time to the next medical visit²² or classification of disease groupings.²³ Because of
28 the well-described interdependence between arterial BP and HR,²⁴ we hypothesized that
29 applying MTL to jointly predict the MAP and the HR could improve the performance of
30 AHE prediction.

31 In the recent literature, AHE was defined as a MAP below 65 mmHg.^{13,25,26} However,
32 as previously highlighted by Chan et al.²⁷ this conventional definition of AHE based on a
33 single cutoff value may not be suitable for individual patients. Normal BP varies between
34 individuals and patients may tolerate hypotension to various degrees before developing end-
35 organ damages. To offer the possibility for the clinician to use individualized BP targets
36 (and thus individualized AHE definition), we developed an algorithm which primary task is
37 not to predict AHE as defined by a specific threshold but rather to predict the actual BP
38 value. In this study, we are proposing to use a MTL approach to predict the MAP value up
39 to 60-min ahead of time. To validate our approach, we compared two different architectures:
40 STL-PDL (Single-Task Learning - Physiological Deep Learner), trained to predict MAP and
41 HR separately, and MTL-PDL (Multi-Task Learning - Physiological Deep Learner) trained
42 to predict MAP and HR jointly. The data from the Medical Information Mart for Intensive
43 Care version 3 (MIMIC-III) waveform database matched subset (version 1.0) were used to
44 train the models. A cohort from Lariboisière hospital surgical ICU (AP-HP, Paris, France)
45 was used for external validation.

46 2. Methods

47 2.1. Datasets

48 The Medical Information Mart for Intensive Care version 3 (MIMIC-III) is a publicly
49 and freely available database including clinical data, physiologic measurements, treatment
50 administration and administrative data of ICU patients. The dataset comprises de-identified
51 data from patients admitted to any of the five ICUs of Boston’s Beth Israel deaconess medical
52 center (BIDMC, Boston, USA) for a period of seven years (2008-2014).²⁸ A unique identifier
53 number was attributed to each patient to match information available in the different tables.
54 Data collection was approved by the institutional review boards (IRB) of BIDMC and the
55 Massachusetts institute of technology (MIT, Cambridge, Massachusetts, USA). We specifi-
56 cally used the MIMIC-III waveform matched subset database (version 1.0), which contains
57 22,247 numeric records (recording of physiological signals every minute) matched and time-
58 aligned with 10,282 MIMIC-III clinical database records (global clinical information about
59 the ICU stay).²⁹

60 The Lariboisière cohort was used for external validation. This database includes clinical
61 data, physiological signals, treatment and administrative data prospectively consecutively
62 collected at the bedside over a two-year period (2017-2018) in the surgical ICU of Lariboisière
63 hospital (AP-HP, Paris, France). This cohort was notably built to be similar in structure
64 to MIMIC-III. This study received IRB approval (CE-SRLF 14-356), and signed informed
65 consent was waived.³⁰ Every patient was orally informed about inclusion in this database.

66 2.2. Data partitioning

67 Following recommendations from Chen and al.,³¹ 90% of the analyzed patients from
68 MIMIC-III were randomly assigned to the "development set": 80% of the patients allocated
69 to the training set (used to estimate model parameters), 10% to the tuning set (used to
70 perform hyper-parameter search) and the remaining 10% (MIMIC-III validation set) was
71 used to evaluate prediction performance. Model performance was also evaluated externally
72 on the data from the Lariboisière cohort. The experimental workflow is detailed in Fig. 1A

73 2.3. Periods’definition

74 Each patient ICU stay was divided into successive periods, as depicted in Fig. 1B For
75 a given period t , our objective was to predict the average MAP and the average HR values
76 observed during the last 5-min of this period (referred to as the *prediction window*) using
77 only the data from the first 30-min of the same period (referred to as the *observation win-*
78 *indow*). In clinical practice, such a prediction is only useful if it is made available sufficiently
79 in advance to allow for therapeutic adjustments. Thus, a time gap (referred to as the *gap*
80 *window*) was inserted between the *observation window* and the *prediction window*. Five time
81 gaps were tested: 5, 10, 15, 30, and 60-min. The following features were used for the predic-
82 tion task: baseline characteristics at ICU admission (age, gender, simplified acute physiology
83 score-II (SAPS-II), sequential organ failure assessment (SOFA) score, type of ICU, i.e., med-
84 ical, surgical, cardiac, mixed), time-evolving treatment characteristics (including mechanical
85 ventilation, vasopressors, and sedation), as well as the five following physiological signals
86 collected every minute: HR, pulse oximetry (SpO_2), systolic arterial pressure (SAP), dias-
87 tolic arterial pressure (DAP) and MAP. Patients or periods with missing clinical information

88 (baseline characteristics, time-evolving characteristics, severity scores, time-evolving treat-
89 ments) were excluded. Finally, only patients with at least one period with available data for
90 the 5-min time gap window were included in the analyses.

91 2.4. Sampling periods and predicting on new patients

92 As the length of ICU stay was different for each patient, our data were uneven in terms
93 of the number of periods per patient, creating two types of challenges:

- 94 i. The model should not overfit to patients with more data, i.e. more periods
- 95 ii. The model should be aware of the correlation between periods coming from the same
96 patient

97 To address (1), at the beginning of each iteration of the learning process (i.e., epoch) a new
98 balanced dataset was obtained by sampling with replacement the same number of periods
99 per patient. The median number of periods per patient included in the study was used
100 to determine the number of periods to be drawn per patient. Models were trained with
101 balanced datasets for multiple epochs. Thus, at each epoch, a different sample of the data
102 was used; consequently, all the data were used during the training process. Challenge (2)
103 was addressed by adding patient *id* as a variable in our models. In deep learning literature,
104 there is a long history of representing categorical variables with n-dimensional vectors.³²
105 This approach has been particularly useful for the representation of clinical data in modern
106 natural language processing to predict several ICU conditions.^{33,34} These vectors are learned
107 by the models together with other parameters during training. The matrix of vectors is
108 often referred to as an embedding layer.³⁵ In this paper, we train a embedding vector
109 representation for each patient *id* in the training set. As a result of training, patients with
110 similar predictors characteristics get a similar vector representation. Hence, by adding the
111 patient *id* to predictors and training a embedding vector for each patient *id*, we were able
112 to fully handle the correlation between periods coming from the same patient. However, in
113 validation and testing sets, patients were different than those of the training set. Thus, we
114 were unable to build vector representations for the new patients *ids*. To be able to predict
115 new patients we used the following approach. At each training process iteration (i.e epoch),
116 we overwrote 10% of the periods at random to have *id* 0. By doing that, the vector associated
117 with the *id* 0 was trained to be the "average user". Thus we could use this average user for
118 prediction of a new patient for model validation and testing.

119 2.5. Predictors and outcomes

120 Fixed predictors included baseline characteristics:

- 121 • Quantitative characteristics: age; initial severity scores: Simplified Acute Physiology
122 Score-II (SAPS-II)³⁶ and Sequential Organ Failure Score (SOFA)³⁷
- 123 • Categorical characteristics: gender; patient *id*; type of ICU

124 Two types of time-dependent characteristics were considered:

- 125 • Period-evolving treatment characteristics: status of mechanical ventilation; adminis-
126 tration of vasopressors and sedation medication

- Physiological signals: heart rate (HR); pulse oximetry (SpO_2); systolic (SAP); diastolic (DAP); mean arterial pressure (MAP)

We denoted quantitative characteristics by x_{cont} , categorical characteristics (except binary) by x_{cat} , gender and period-evolving treatment binary characteristics (collected every period) by x_{binary} and physiological signals (collected every min during the *30-min observation window*) by x_{series} . Thus, we defined a vector representation of predictors as $x = (x_{cont}, x_{cat}, x_{binary}, x_{series})$ associated with the two outcomes y_{MAP} and y_{HR} where y_{MAP} and y_{HR} represented respectively the MAP and HR averaged over the *5-min prediction window* for each patient period.

2.6. Model architecture

Deep learning models handle complex relationships between a large number of explanatory predictors and desired outputs, such as patient outcome.¹⁸ Based on a succession of layers (each layer receives its inputs from the previous one's outputs), deep learning models use backpropagation algorithms to update their internal parameters and optimize their predictions. The Physiological Deep Learner (PDL) was designed to predict the MAP and/or HR by mapping x to y_{MAP} and/or y_{HR} . Different PDL were implemented and compared, two single-task learning PDL (STL-PDL) predicting separately the MAP and the HR, and one Multi-task learning PDL (MTL-PDL) trained to jointly predict the MAP and HR.

To render the comparison between STL-PDL and MTL-PDL as fair as possible, all estimation processes were identical except the last step (Fig. 2). The PDL maps the vector of predictors $x = (x_{cont}, x_{cat}, x_{binary}, x_{series})$ to y_{MAP} and/or y_{HR} , and follows the following steps:

- i. x_{series} is input to a Gated Recurrent Unit (GRU)³⁸ a type of recurrent neural network (RNN) that outputs a vector r . For tasks that involve time series predictors, such as physiological signals overtime, it is often better to use RNNs. RNNs process an input time series one element at a time, maintaining in their hidden units a "state vector" that implicitly contains information about the history of all the past elements of the time series.³⁹
- ii. Each categorical variable in x_{cat} is input to an separate embedding layers given the outputs e_1 and e_2 . Embedding layers are initialized randomly and learned by the model in the optimization process. In particular models learn a vector representation for each patient id. This is particularly helpful to account for within-patient correlation between the periods of a same patient.
- iii. r , e_1 , e_2 , x_{binary} and x_{cont} get concatenated to form c
- iv. c gets fed into multiple linear regression layers and non-linear functions (ReLU) returning h .
- v. Two architectures depending on the algorithm:
 - (a) STL: h gets fed into a linear regression layer to predict MAP or HR independently.

165 (b) MTL: h gets fed into linked linear regression layers to predict MAP and HR
 166 jointly.

To make the comparison between STL-PDL and MTL-PDL as fair as possible all models are identical until the last step. Here a summary of all steps for the two PDL’s architectures:

Single-task learning	Multi-task learning
$r = GRU(x_{series})$	$r = GRU(x_{series})$ (1)
$e_1 = EMB_1(x_{cat}[1])$	$e_1 = EMB_1(x_{cat}[1])$ (2)
$e_2 = EMB_2(x_{cat}[2])$	$e_2 = EMB_2(x_{cat}[2])$ (3)
$c = (r, e_1, e_2, x_{binary}, x_{cont})$	$c = (r, e_1, e_2, x_{binary}, x_{cont})$ (4)
$h = Batchnorm(ReLU(Linear1(c)))$	$h = Batchnorm(ReLU(Linear1(c)))$ (5)
$\hat{y}_{MAP} = Linear2(h)$ or $\hat{y}_{HR} = Linear2(h)$	$(\hat{y}_{MAP}; \hat{y}_{HR}) = jointLinear(h)$ (6)

167 where GRU is a gated recurrent unit; $x_{cat}[1]$ and $x_{cat}[2]$ are patient id and type of ICU
 168 respectively; EMB_1 and EMB_2 are two embedding layers; \hat{y} is either y_{MAP} or y_{HR} ; $Linear1$,
 169 $Linear2$ and $jointLinear$, are 3 linear regression layers; Batchnorm is batch-normalization
 170 layer;⁴⁰ $ReLU(x)$ is the rectified linear unit function.⁴¹

171 2.7. Model Optimization

172 2.7.1. Single-task learning

Let $\mathcal{D} = \{\{x^{(ip)}, y_{MAP}^{(ip)}, y_{HR}^{(ij)}\}_{p=1}^t\}_{i=1}^n$ be the set of training observations, where t , the number of periods and n the number of patients. The basic idea is to find out a relationship $f(\cdot)$ between x and y (either y_{MAP} or y_{HR}), which represents the clinical risk model that takes x as input and outputs \hat{y} as predictions. Thus, the model is specified by $\hat{y} = f(x)$. To predict MAP and HR values, we considered two separate STL-PDL, one for each task, and used mean squared error (MSE) as loss function. Concretely, let ℓ_{MAP} and ℓ_{HR} be the MSE for the outcomes y_{MAP} and y_{HR} , respectively and \hat{y}_{MAP} and \hat{y}_{HR} the predictions output by the two separated clinical risk models, $f_{MAP}(\cdot)$ and $f_{HR}(\cdot)$. Then, ℓ_{MAP} and ℓ_{HR} are defined as follows:

$$\ell_{MAP}(f_{MAP})(\mathcal{D}) = \frac{1}{N} \sum_{i=1}^N (y_{MAP}^{(i)} - \hat{y}_{MAP}^{(i)})^2, \quad (7)$$

$$\ell_{HR}(f_{HR})(\mathcal{D}) = \frac{1}{N} \sum_{i=1}^N (y_{HR}^{(i)} - \hat{y}_{HR}^{(i)})^2, \quad (8)$$

173 where $N = t \times n$ corresponds to the total number of observations. Finally, the parameter
 174 estimation of each clinical risk model (i.e., tacks) is determined by minimizing the MSE loss
 175 function.

176 2.7.2. Multi-task learning

In contrast, MTL involves jointly estimating several prediction models.⁴² The intuition is that a joint estimation can do better than an independent estimation of the tasks sharing similarities. **MTL refers to the optimization of a global loss function. Therefore, the**

minimization of the global MTL cost function allows the simultaneous estimation of the parameters associated with the MAP and those associated with the HR based on common predictors. Thus, the final estimated parameters reflect the understanding found to predict the two tasks together. They are consequently different from those estimated independently with single-task models. In that case, we can treat our problem as building one prediction model for the two tasks, and use mean squared error (MSE) as loss function. Concretely, Let \mathcal{L} be the global MSE for the outcomes y_{MAP} and y_{HR} and \hat{y}_{MAP} and \hat{y}_{HR} the predictions output by $f(\cdot)$ the joint clinical risk model. Hence, our global multi-task learning loss of \mathcal{L} is defined as follows:

$$\mathcal{L}(f)(\mathcal{D}) = \frac{1}{2N} \left\{ \sum_{i=1}^N (y_{MAP}^{(i)} - \hat{y}_{MAP}^{(i)})^2 + (y_{HR}^{(i)} - \hat{y}_{HR}^{(i)})^2 \right\} \quad (9)$$

177 PDLs development was performed using PyTorch version 1.4 library.⁴³ We used Adam⁴⁴
 178 for optimization. It is computationally efficient, has little memory requirements, is invariant
 179 to diagonal re-scaling of the gradients. The optimization of hyperparameters was based on
 180 the best models' performances on the tuning set. Thus, we fixed the learning rate to 0.003,
 181 the hidden size to 100 units, the number of epochs to 20, and the weight decay to 10^{-5} .
 182 Then best models were fitted on the two validation sets to report final performance results.
 183 Graphical representations of the final results were performed using R version 3.6.3 library.⁴⁵

184 2.8. Assessment of performance

185 2.8.1. Evaluation of the averages prediction of MAP and HR

Graphical representations of the final results were performed using R version 3.6.3 library.⁴⁵ To assess models' performance in both validation sets, R-squared (R^2) together with its 95% confidence interval (95%CI) and root mean square error (RMSE), defined as

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (\hat{y}^{(i)} - y^{(i)})^2}, \quad (10)$$

186 where $\hat{y}^{(i)}$ is a generic outcome predicted and $y^{(i)}$ a generic outcome observed, were computed.
 187 Differences between observed and predicted outcomes against observed outcomes were plotted
 188 to quantify 95% limits of agreement (95% LOA) predictions. Finally, calibration plots
 189 were plotted. To do so, patients were grouped into observed and predicted outcomes deciles.
 190 Within each decile, the true mean per decile defined as the average of observed 5-min MAP
 191 or 5-min HR values was computed. Similarly, the predicted mean per decile defined as the
 192 average 5-min MAP or 5-min HR values was also computed. Then, each couple of means
 193 was plotted according to time gaps for each validation set. Thus, the closer the line to the
 194 diagonal, the better the calibration.

195 2.8.2. Acute hypotensive episodes prediction

196 To indicate what would be the performance of an AHE alert device based on our findings,
 197 we defined a threshold for the MAP at 65 mmHg as this threshold is commonly used to
 198 defined AHE.^{13,46,47} Therefore, we classified the patients using the following rule:

- 199 i. "AHE", if observed average 5-min MAP ≤ 65
- 200 ii. "No AHE", either

201 Note that our MTL-PDL was not trained to predict this binary outcome, but we directly
202 defined classes of predicted risk of AHE on the continuous predicted 5-min MAP from the
203 MTL-PDL according to the following rule:

- 204 i. "Very high", if predicted MAP ≤ 60
- 205 ii. "High", if $60 < \text{predicted MAP} \leq 65$
- 206 iii. "Moderate", if $65 < \text{predicted MAP} \leq 70$
- 207 iv. "Low", if predicted MAP > 70

208 We examined the performance of the MTL-PDL by comparing the observed classes to the
209 predicted classes. Relying on the crossing matrices, we calculated for each predicted classes
210 ("Very high", "High", "Moderate" or "Low"), the probability of AHE knowing the class,
211 $P(AHE|k)$ and the probability of no AHE knowing the class, $P(NoAHE|k)$, where k is the
212 predicted risk class. Moreover, by applying the same risk classes definitions to the observed
213 and predicted MAP values we displayed Bangdiwala’s agreement chart for each validation set
214 and each time gap. This chart assesses the concordance between two methods of measure-
215 ment of ordinal categorical data.⁴⁸ Thus, it gives an overview of misclassification between
216 observed and predicted classes. For each class, exact agreement between the observed and
217 predicted is represented by a rectangle filled with the bleakest color. Partial agreement is
218 reached when the closest class is predicted instead of the actual class. It is represented by
219 an intermediate color between exact and no agreement. No agreement is obtained when the
220 farthest class is predicted instead of the actual class. It is represented by the lightest color.
221 The more the diagonal goes through the corners of the rectangles, the greater the global
222 agreement.

223 3. Results

224 *Data Preparation.* Within MIMIC-III and Lariboisière cohort, all patients with complete
225 data (baseline characteristics, time-evolving characteristics and physiological signals) were
226 selected. Among them, all patients with at least one complete set of 3 successive windows
227 (i.e. observation, gap, prediction) with a time gap of at least 5 min were included in the
228 analysis (Fig. 3). From MIMIC-III, 2,308 patients (74,159 periods) qualified to be included
229 in the analysis, among which, 2,290 patients (62,951 periods) still had data available when
230 increasing the time gap to 10 min, 2,261 patients (52,413 periods) with a time gap of 15 min,
231 2,153 patients (34,499 periods) with a time gap of 30 min and 1,996 patients (17,870 periods)
232 with a time gap of 60 min. Forty nine patients from Lariboisière cohort were included in the
233 external validation analysis. All 49 patients had data available with 5, 10, 15 and 30-min
234 time gaps, representing a total of 1,417, 1,226, 1,024, and 629 periods respectively. Only 43
235 patients (295 periods) had data available with a time gap of 60 min. Patients characteristics
236 are summarized in Table 1.

237 *MAP prediction..* STL- and MTL-PDL performance for each time gap (5, 10, 15, 30, and
 238 60-min) are presented in Fig. 4. When evaluated using the MIMIC-III validation set and the
 239 external validation cohort, the correlation coefficient R^2 (95% confidence interval) between
 240 the actual and the predicted MAP was consistently very close to 1 whatever the PDL archi-
 241 tecture and the time gap. With a STL structure, the R^2 obtained in MIMIC-III validation
 242 set ranged from 0.954 (0.952-0.956) to 0.839 (0.824-0.853), and from 0.937 (0.93-0.943) to
 243 0.754 (0.7-0.8) in the external validation cohort. With a MTL structure, the R^2 ranged from
 244 0.958 (0.956-0.96) to 0.833 in (0.817-0.847) in MIMIC-III, and from 0.952 (0.946-0.956) to
 245 0.747 (0.692-0.794) in the external validation cohort. However, when compared to MTL-
 246 PDL, STL-PDL was consistently associated with a larger root mean square error (RMSE)
 247 (middle panel, Fig. 4). In MIMIC-III validation set, the RMSE for the STL-PDL was of
 248 4.16, 4.73, 4.86, 6.15 and 7.44 for the 5, 10, 15, 30, and 60-min time gaps respectively. In
 249 contrast, the RMSE for the MTL-PDL was consistently lower: 3.93, 4.42, 4.77, 6.06 and 7.38
 250 respectively. A similar pattern was observed in the external validation cohort: STL-PDL
 251 RMSE of 6.86, 7.24, 7.84, 10.03, and 13.42 at 5, 10, 15, 30, and 60-min respectively and for
 252 MTL-PDL 5.68, 6.39, 7.23, 9.44 and 12.14 respectively. Fig. 4, right panel illustrates the
 253 concordance between observed and predicted MAP by quantifying the limits of prediction
 254 agreement. In general, all differences between observed and predicted values lied within the
 255 95% Limits Of Agreement (95% LOA) for each time gap and validation set. MTL-PDL was
 256 associated with more accurate predictions as illustrated by an average difference between
 257 observed and predicted MAP consistently closer to zero. In MIMIC-III validation set, the
 258 average difference between observed and predicted MAP for each time gap was of 0.59 [95%
 259 LOA = -7.48-8.67], -0.06 [-9.34-9.22], 0.85 [-8.54-10.24], 1.52 [-10.15-13.2], 1.69 [-12.52-15.9]
 260 for STL-PDL, while it was of 0.28 [-7.41-7.97], 0.12 [-8.54-8.77], -0.27 [-9.59-9.06], -1.26 [-
 261 12.89-10.36], -0.26 [-14.73-14.2] for MTL-PDL. The average difference between observed and
 262 predicted MAP were larger in the external validation cohort than in the MIMIC-III val-
 263 idation dataset, but MTL-PDL was also superior to STL-PDL: average difference for the
 264 STL-PDL was 2.81 [-9.47-15.09], 2.33 [-11.1-15.76], 2.55 [-11.99-17.09], 3.15 [-15.53-21.84],
 265 6.26 [-17.05-29.56] at 5, 10, 15, 30, and 60-min respectively and for the MTL-PDL 1.74 [-
 266 8.85-12.34], 0.04 [-12.49-12.57], 0.19 [-13.98-14.37], 0.80 [-17.64-19.24] 1.54 [-22.1-25.18]. The
 267 superiority of the MTL-PDL over STL-PDL was also confirmed using calibration plots (Fig.
 268 5).

269 *Prediction of acute hypotensive episodes..* In Fig. 6 are displayed predictive values for risk of
 270 acute hypotensive episodes. In both validation sets, the higher the predicted risk, the higher
 271 the probability of observing a MAP below 65 mmHg. $P(AHE|k)$ was 99% in MIMIC-III
 272 and 90% in the external validation cohort for the "Very high risk" class. $P(NoAHE|k)$ was
 273 99.5% in MIMIC-III and 99.8% in the external validation cohort for the "Low risk" class.
 274 Finally, by applying the same risk classes definitions to the observed and predicted MAP
 275 values, we assessed the MTL-PDL misclassification error in both validation set (Fig. 7 and
 276 8). As expected, the agreement decreased as the time gap increased. However, it seems that
 277 the misclassification always goes in the direction of partial agreement (i.e. the closest class
 278 predicted instead of the current class) rather than total disagreement (i.e. the farthest class
 279 predicted instead of the current class) between observed and predicted AHE class risk.

280 *HR prediction..* Comparable results are provided in Fig. 9 for HR prediction. Similar to
281 MAP prediction, MTL-PDL was found to outperform STL-PDL for HR prediction, especially
282 with 30 and 60-min time gaps. In both internal and external validation sets, the R^2 was
283 similar and close to 1. RMSE was consistently lower with MTL-PDL except for 5 and 10-
284 min gaps in the external validation cohort where there was no difference between the two
285 PDL architectures. Fig. 10 shows excellent and better calibration profile with MTL-PDL as
286 compared to STL-PDL.

287 4. Discussion

288 We developed the Physiological Deep Learner that processes baseline characteristics and
289 multiple continuous physiological signals to accurately predict the evolution of the MAP and
290 the HR in critically ill patients. More precisely, the major novelty of this study was the use
291 of a MTL architecture to improve the prediction performance by jointly modeling MAP and
292 HR. This learning framework is similar to the way clinicians are trained to jointly analyze
293 the evolution of the HR and the MAP given their close physiological interdependence. To
294 render this new prediction tool useful in clinical practice, we trained the Physiological Deep
295 Learner to predict the MAP and the HR with incremental time gaps, up to 60-min ahead of
296 time. Compared to a more traditional STL-PDL approach, our MTL-PDL achieved better
297 performance, with better calibration profile and fewer errors. In addition, the Physiological
298 Deep Learner was able to predict with high accuracy the occurrence or not of an acute
299 hypotensive episode.

300 Several AHE prediction models were developed over the past 20 years. In 2009, the 10th
301 annual PhysioNet/Computers in cardiology challenge⁴⁹ was set to promote the development
302 of methods for identifying ICU patients at imminent risk of AHE. During this challenge, mul-
303 tiple ML prediction models were proposed.^{5,6,7,8,9,10,11,12} However, none of them achieved
304 sufficient accuracy to be adopted in clinical practice. More recently, Hatib et al.¹³ used a
305 logistic regression model to predict hypotension based on 3,022 features extracted from the
306 MAP waveform signal. Their model reached a sensitivity of 88% (95% CI, 85 to 90%) and
307 a specificity of 87% (95% CI, 85 to 90%) but tended to underpredict the risk of hypotension
308 in the higher-risk subgroups. Thus far, most predictive models used historical MAP values
309 as their only input variable, ignoring other patient characteristics and/or time-dependent
310 variables, e.g., heart rate, known to be highly correlated with the arterial BP. Our group⁴⁶
311 proposed to use multiple physiological signals in addition to patient and treatment charac-
312 teristics to train an ensemble machine learning model to predict AHE. This model exhibited
313 promising performance, with an area under the curve (AUC) of 0.890 (95% CI, 0.886 to
314 0.895). Kendale et al.¹⁴ also used an ensemble learning model to predict hypotension fol-
315 lowing anesthesia induction using intraoperative vital signs, medications and comorbidities
316 as features and obtained an AUC of 0.74 (95% CI, 0.72 to 0.77).

317 Very recently, Hyland et al.⁴⁷ used gradient-boosted ensemble tree classifiers trained on
318 209 variables to predict circulatory failure in critically ill patients. As expected based on
319 physiological knowledge, this study reported that HR was among the top-5 most important
320 predictors for circulatory failure. Based on the idea that MAP and HR are intrinsically cor-
321 related, we developed the Physiological Deep Learner using a multi task learning approach.
322 MTL is generally used for i) the prediction of separate outcomes or ii) to identify separate

323 subpopulations. Our formulation falls into the first category, where HR and MAP prediction
324 were defined as the two different tasks. There are several expected benefits to MTL.⁴² First,
325 MTL works even if one of the outcomes is missing. Thus, one can still train the model to
326 do both tasks at the same time and allow easy implementation in real practice. The other
327 advantage is data amplification. When we consider two tasks with independent noise added
328 to their training signals, both profit from computing a hidden layer feature \mathcal{F} of the inputs.
329 Determining both can optimize the learning of \mathcal{F} by averaging \mathcal{F} across the different noise
330 processes.⁴² In addition, focusing on one task carries the risk of overfitting while learning
331 to predict MAP and HR values jointly is associated with increased generalizability.⁵⁰ This
332 was confirmed in the present study, where we were able to integrate MTL to jointly predict
333 MAP and HR up to 60 min in advance and found high calibration and accuracy even when
334 tested in an external dataset.

335 In most studies on hypotension prediction in the ICU, AHE is defined as a binary status
336 based on a single MAP threshold. This binary approach carries some limitations. First, def-
337 initions are often heterogeneous across studies. Second and most importantly, a definition
338 based on a single cutoff value may not be suitable for individual patients. Indeed, blood
339 pressure varies between individuals, as does individual organ capacity to tolerate hypoten-
340 sion.²⁷ Accordingly, Futier et al.⁵¹ showed among patients undergoing abdominal surgery,
341 that targeting individualized systolic blood pressure goals reduced the risk of postoperative
342 organ dysfunction. Chan et al.²⁷ introduced the concept of a patient-specific definition of
343 AHE based on the use of two moving averages of MAP recordings in which the outcome of
344 interest was defined as a 20% drop in the averages. The Physiological Deep Learner goes
345 even beyond that since it was trained to predict the actual blood pressure rather than any
346 binary transform of the MAP. As an example, Fig. 11 shows individual MAP and HR pre-
347 dictions for four different patients with a time gap of 15 minutes. Our goal was to develop a
348 more clinically meaningful algorithm by i) providing the clinician with an information, i.e.
349 the predicted actual MAP, similar to the one he/she is using in his/her clinical practice (i.e.
350 the actual MAP), and ii) leaving to the clinician the latitude to interpret this prediction and
351 classify it or not as a possible hypotensive episode.

352 Finally, most previous studies applied different methods of features extraction to physio-
353 logical signals time series to summarize them into finite values. However, in doing so, a large
354 part of the information is being lost. In a previous study, we demonstrated how sensitive
355 deep learning models are to the method used to summarize the information from physiologi-
356 cal time series.⁴⁶ A strength of the present study is that we used gated recurrent unit cells,³⁸
357 which are able to effectively retain long-term dependencies in time series. According to Le
358 Cun et al.,³⁹ this is the most optimal way to encode temporal information about the entire
359 patient ICU stay since it preserves the longitudinal changes and the original time-dependent
360 order in patient physiological signals.

361 Our study carries some limitations. Although appealing, our results will need to be
362 confirmed in a larger validation set. Indeed, the external validation dataset was relatively
363 limited in size. Real-life data from bedside monitors and electronic medical systems are prone
364 to missing values, errors and artifacts, adding significant noise to the data.⁵² In this study,
365 we only included patients with complete data and particularly complete physiological time-
366 series. However, missingness is likely to be informative in some ICU patients. Therefore, our
367 algorithm may lack generalizability to patients presenting a lot of missing data. In future

368 iterations of our algorithms, we will need to include a more robust approach to managing
369 missing values. We were not able to provide prediction intervals around MAP and HR
370 predicted values. However we are confident that this will be possible in the near future.
371 Producing valid prediction intervals for machine learning models is an active area of research
372 within our group. Finally, in this iteration of the Physiological Deep Learner, we gave the
373 same weight to each prediction task. In the future, weighting differently the two tasks to
374 reflect their relative clinical importance may result in better prediction performance for the
375 primary task.

376 **5. Conclusion**

377 The Physiological Deep Learner trained to predict simultaneously the mean arterial blood
378 pressure and the heart rate up to 60 min in advance, demonstrated very good performance
379 both internally and externally. Although further prospective validation is needed, these re-
380 sults support the use of a deep learning model with multitask learning structure to learn from
381 multiple physiological signals in the ICU. Based on this result, we believe that algorithms
382 such as the Physiological Deep Learner will help the clinician to predict the evolution of key
383 physiological features at the bedside and thereby allow them to adapt their treatment and
384 avoid critical events. This hypothesis remains to be tested in a prospective manner.

385 **Acknowledgments**

386 We thank our colleagues Pr Alexandre Mebazaa, Pr Etienne Gayat, and Dr Fabrice
387 Vallée from the Department of Anesthesia Burn and Critical Care, University Hospitals
388 Saint-Louis - Lariboisière, AP-HP, Paris, France, who provided insight, expertise and access
389 to the Lariboisière cohort that greatly assisted the research.

390 **Author Contributions**

391 M.C conceived the study, wrote the study protocol, led the data management, developed
392 the Physiological Deep Learner, conducted the analyses, and wrote the manuscript. Y.I
393 conceived the study, developed the Physiological Deep Learner, conducted the analyses, and
394 wrote the manuscript. M.R.R, and R.P conceived the study, wrote the study protocol, and
395 wrote the manuscript. A.B assisted with the interpretation and figures' draft and reviewed
396 the manuscript. All authors contributed to data interpretation, critically reviewed, and
397 approved the manuscript before submission.

398 **Competing interests**

399 The authors declare no competing interests.

400 **Figures captions**

401 **Fig. 1| Periods definition and learning framework process.** **A.** From MIMIC-III,
402 80% of the patients were randomly assigned to the training set, 10% to the tuning set, and
403 the remaining 10% to the validation set. The latter corresponds to the MIMIC-III validation
404 set. Data from the Lariboisière cohort were exclusively used for external validation of the
405 models. Note that the allocation of the data in the different sets was performed in such a
406 way that all periods of the same patient were assigned to the same set. MIMIC-III, Medi-
407 cal Information Mart for Intensive Care III. **B.** Patients from MIMIC-III and Lariboisière
408 cohort, have their ICU stay divided, from the admission to the discharge into periods of
409 the same duration. Each period was divided into 3 successive windows (Observation, Gap,
410 Prediction). To predict the average 5-min MAP and HR of the prediction window, only data
411 recorded during the observation window were used.

412
413 **Fig. 2| Comparison between single-task learning and multi-task learning.** Each
414 input variable is treated differently by our model during the specific processing layer when it
415 is necessary. Then, they are concatenated and fed into successive layers until the output. In
416 single-task learning, the output corresponds to the prediction of one outcome while in multi-
417 task learning, the outputs correspond to two distinct outcome predictions. ID, identifier;
418 ICU, Intensive Care Unit; Linear, linear regression; SOFA, Sequential Organ Failure As-
419 sessment; SAPS-II, Simplified Acute Physiology Score; GRU, Gated Recurrent Unit; ReLU,
420 Rectified Linear Unit; Batchnorm; Batch normalization.

421
422 **Fig. 3| Flow-chart of patients selection.** All patients with no missing data on phys-
423 iological signals and clinical information from the MIMIC-III and Lariboisière cohort were
424 selected. Then, only patients with at least one period with a time gap of 5 min were in-
425 cluded. ICU, Intensive Care Unit; MIMIC-III, Medical Information Mart for Intensive Care
426 III databases; SAPS-II, Simplified Acute Physiology Score; SOFA, Sequential Organ Failure
427 Assessment; HR, Heart Rate; SpO_2 , pulse oximetry; MAP, Mean Arterial Pressure, DAP,
428 Diastolic Arterial Pressure, SAP, Systolic Arterial Pressure.

429
430 **Fig. 4| Models performances to predict the value of MAP averaged over 5 min.**
431 **Left**, R^2 together with its 95% confidence interval were computed to measure the linear
432 regression agreement between observed and predicted. As its value can vary from 0 to 1,
433 a focus has been done to see the results properly. **Middle**, For each validation set and
434 architecture, we calculated root mean square error(RMSE). Note, the closer RMSE is to 0,
435 the better it is. **Right**, Differences between observed and predicted values against observed
436 values were represented. The plain line represents the average difference and the dotted
437 lines the 95% limits of agreement (95% LOA). The closer the average difference is to 0, the
438 better the performance is. MAP, Mean Arterial Pressure, MIMIC-III, Medical Information
439 Mart for Intensive Care III; STL, Single-Task Learning; MTL, Multi-Task Learning; R^2 ,
440 R-squared; RMSE, Root Mean Square Error.

441
442 **Fig. 5| Calibration plots for the value of MAP averaged over 5 min.** Patients
443 were grouped into deciles. Within each decile, the average observed and predicted MAP

444 is calculated. The first corresponds to the observed mean per decile and the latter to the
445 predicted mean per decile. Each couple of mean is plotting according to the time gap for
446 both validation sets. The closer the line is to the diagonal, the better the calibration is.
447 MAP, Mean Arterial Pressure, MIMIC-III, Medical Information Mart for Intensive Care III.

448

449 **Fig. 6| Predictive values for acute hypotensive episodes prediction.** For each pre-
450 dicted classes ("Very high", "High", "Moderate" or "Low"), the probability of AHE knowing
451 the class, $P(AHE|k)$ and the probability of no AHE knowing the class, $P(NoAHE|k)$, where
452 k is the predicted risk class were calculated. Note that in the Lariboisière data, some values
453 are missing due to the absence of patients in the category. AHE, Acute Hypotensive Episode;
454 MIMIC-III, Medical Information Mart for Intensive Care III.

455

456 **Fig. 7| Agreement plots for acute hypotensive episodes on the MIMIC-III vali-**
457 **dation set.** This chart gives an overview of classification/misclassification between observed
458 and predicted risk class of AHE ("Very high", "High", "Moderate" or "Low"). The exact
459 agreement between the observed and predicted is obtained when the rectangle is filled with
460 the bleakest color. The partial agreement is obtained when the closest class is predicted
461 instead of the current. It is represented by an intermediate color between exact and no
462 agreement. No agreement is obtained when a class farther than the very next class are
463 predicted instead of the current class. It is represented by the lightest color. The more
464 the diagonal goes through the corners of the rectangles, the greeter the global agreement is.
465 MIMIC-III, Medical Information Mart for Intensive Care III.

466

467 **Fig. 8| Agreement plots for acute hypotensive episodes on Lariboisière cohort.**
468 This chart gives an overview of classification/misclassification between observed and pre-
469 dicted risk class of AHE ("Very high", "High", "Moderate" or "Low"). The exact agreement
470 between the observed and predicted is obtained when the rectangle is filled with the bleakest
471 color. The partial agreement is obtained when the closest class is predicted instead of the
472 current. It is represented by an intermediate color between exact and no agreement. No
473 agreement is obtained when a class farther than the very next class are predicted instead of
474 the current class. It is represented by the lightest color. The more the diagonal goes through
475 the corners of the rectangles, the greeter the global agreement is.

476

477 **Fig. 9| Models performances to predict the value of HR averaged over 5 min.**
478 **Left**, R^2 together with its 95% confidence interval were computed to measure the linear
479 regression agreement between observed and predicted. As its value can vary from 0 to 1,
480 a focus has been done to see the results properly. **Middle**, For each validation set and
481 architecture, we calculated root mean square error(RMSE). Note, the closer RMSE is to 0,
482 the better it is. **Right**, Differences between observed and predicted values against observed
483 values were represented. The plain line represents the average difference and the dotted lines
484 the 95% limits of agreement (95% LOA). The closer the average difference is to 0, the better
485 the performance is. HR, Heart Rate; MIMIC-III, Medical Information Mart for Intensive
486 Care III; STL, Single-Task Learning; MTL, Multi-Task Learning; R^2 , R-squared; RMSE,
487 Root Mean Square Error.

488

489 **Fig. 10| Calibration plots for the value of HR averaged over 5 min.** Patients
 490 were grouped into deciles. Within each decile, the average observed and predicted MAP
 491 is calculated. The first corresponds to the observed mean per decile and the latter to the
 492 predicted mean per decile. Each couple of mean is plotting according to the time gap for
 493 both validation sets. The closer the line is to the diagonal, the better the calibration is. HR,
 494 Heart Rate; MIMIC-III, Medical Information Mart for Intensive Care III.

495
 496 **Fig. 11| Physiological Deep Learner’s predictions examples** Individual predictions
 497 of MAP and HR for four different patient’s periods with a time gap of 15 minutes are pre-
 498 sented. **A** and **B** correspond to patients with average observed 5-min MAP below 65 mmHg
 499 and **C** and **D** to patients with average observed 5-min MAP greater than 65 mmHg. HR,
 500 heart rate; MAP, mean arterial pressure.

502 **Tables**

Table 1: Patients characteristics

Variables	MIMIC-III	Lariboisière cohort
Number of patients	2,308	49
Age	66 [56-76]	56 [49-68]
Gender (Female)	884 (39.1%)	22 (44.9%)
Status at admission		
SAPS-II score	28[21-36]	41 [21-59]
SOFA score	3[1-5]	7 [4-10]
Site		
CCU	410 (17.8%)	
CSRU	812 (35.2%)	
MICU	366 (15.9%)	
SICU	529 (22.9%)	49 (100%)
TSICU	191 (8.4%)	
Organ-support therapies		
Vasopressors	310 (13.4)	18 (36.7%)
Sedation	861 (37.3%)	27 (55.1%)
Mechanical ventilation	1,218(52.8%)	32 (65.3%)

All patients from both dataset with no missing data: baseline characteristics, time-evolving characteristics (i.e., organ-support therapies), and physiological signals considered in the analyses. Continuous variables are presented as median [InterQuartile Range]; binary or categorical variables as count (%). MIMIC-III, Medical Information Mart for Intensive Care III; SAPS-II, Simplified Acute Physiology Score II; SOFA, Sequential Organ Failure Assessment; CCU: Cardiac Care Unit; CSRU: Cardiac Surgery Recovery Unit, MICU: Medical Intensive Care Unit; SICU: Surgical Intensive Care Unit; TSICU: Trauma Surgical Intensive Care Unit

503 **References**

- 504 ¹ J.-L. Vincent, D. De Backer, Circulatory Shock, *New England Journal of Medicine* 369
505 (2013) 1726–1734.
- 506 ² Y. Sakr, K. Reinhart, J. L. Vincent, C. L. Sprung, R. Moreno, V. M. Ranieri, D. De
507 Backer, D. Payen, Does dopamine administration in shock influence outcome? Results of
508 the Sepsis Occurrence in Acutely Ill Patients (SOAP) Study, *Critical Care Medicine* 34
509 (2006) 589–597.
- 510 ³ R. F. Wilson, J. A. Wilson, D. Gibson, W. J. Sibbald, Shock in the emergency department,
511 *Journal of the American College of Emergency Physicians* 5 (9) (1976) 678–690.
- 512 ⁴ M. M. Levy, L. E. Evans, A. Rhodes, The Surviving Sepsis Campaign Bundle: 2018 update
513 (jun 2018).
- 514 ⁵ P. Langley, S. King, D. Zheng, E. Bowers, K. Wang, J. Allen, A. Murray, Predicting acute
515 hypotensive episodes from mean arterial pressure, in: *Computers in Cardiology, 2009*,
516 Vol. 36, 2009, pp. 553–556.
- 517 ⁶ K. Jin, N. Stockbridge, Smoothing and discriminating MAP data, in: *Computers in Car-*
518 *diology, 2009*, Vol. 36, 2009, pp. 633–636.
- 519 ⁷ F. Jousset, M. Lemay, J. Vesin, *Computers in Cardiology / Physionet Challenge 2009:*
520 *Predicting acute hypotensive episodes*, in: *Computers in Cardiology, 2009*, Vol. 36, 2009,
521 pp. 637–640.
- 522 ⁸ T. Ho, X. Chen, Utilizing histogram to identify patients using pressors for acute hypoten-
523 sion, in: *Computers in Cardiology, 2009*, Vol. 36, 2009, pp. 797–800.
- 524 ⁹ J. H. Henriques, T. R. Rocha, Prediction of Acute Hypotensive Episodes Using Neural
525 Network Multi-models, in: *Computers in Cardiology, 2009*, Vol. 36, 2009, pp. 549–552.
- 526 ¹⁰ F. Chiarugi, I. Karatzanis, V. Sakkalis, I. Tsamardinos, T. Dermitzaki, M. Foukarakis,
527 G. Vrouchos, Predicting the occurrence of acute hypotensive episodes: The PhysioNet
528 Challenge, in: *Computers in Cardiology, 2009*, Vol. 36, 2009, pp. 621–624.
- 529 ¹¹ X. Chen, D. Xu, Forecasting acute hypotensive episodes in intensive care patients based
530 on a peripheral arterial blood pressure waveform, in: *Computers in Cardiology*, Vol. 36,
531 2009, pp. 545–548.
- 532 ¹² P. Fournier, J. Roy, Acute hypotension episode prediction using information divergence
533 for feature selection, and non-parametric methods for classification, in: *Computers in*
534 *Cardiology, 2009*, Vol. 36, 2009, pp. 625–628.
- 535 ¹³ F. Hatib, Z. Jian, S. Buddi, C. Lee, J. Settels, K. Sibert, J. Rinehart, M. Cannesson,
536 Machine-learning Algorithm to Predict Hypotension Based on High-fidelity Arterial Pres-
537 sure Waveform Analysis., *Anesthesiology* 129 (2018) 663–674.

- 538 ¹⁴ S. Kendale, P. Kulkarni, A. D. Rosenberg, J. Wang, Supervised Machine Learning Pre-
539 dictive Analytics for Prediction of Postinduction Hypotension, *Anesthesiology* 129 (2018)
540 675–688.
- 541 ¹⁵ R. Donald, T. Howells, I. Piper, I. Chambers, G. Citerio, P. Enblad, B. Gregson, K. Kien-
542 ing, J. Mattern, P. Nilsson, Early warning of EUSIG-defined hypotensive events using a
543 Bayesian artificial neural network, in: *Intracranial Pressure and Brain Monitoring XIV*,
544 Vol. 114, 2012, pp. 39–44.
- 545 ¹⁶ S. Bhattacharya, V. Huddar, V. Rajan, C. K. Reddy, A dual boundary classifier for pre-
546 dicting acute hypotensive episodes in critical care., *PloS one* 13(2) (2018) e0193259.
- 547 ¹⁷ D. P. Barnaby, S. M. Fernando, K. J. Ferrick, C. L. Herry, A. J. Seely, P. E. Bijur, E. John
548 Gallagher, Use of the low-frequency/high-frequency ratio of heart rate variability to pre-
549 dict short-term deterioration in emergency department patients with sepsis, *Emergency*
550 *Medicine Journal* 35 (2) (2018) 96–102.
- 551 ¹⁸ M. Cherifa, R. Pirracchio, What every intensivist should know about Big Data and tar-
552 geted machine learning in the intensive care unit, *Rev Bras Ter Intensiva* 31 (4) (2019)
553 444–446.
- 554 ¹⁹ R. A. Caruana, Multitask Learning: A Knowledge-Based Source of Inductive Bias, in:
555 *Machine Learning Proceedings 1993*, Elsevier, 1993, pp. 41–48.
- 556 ²⁰ I. S. Hofer, C. Lee, E. Gabel, P. Baldi, M. Cannesson, Article open development and
557 validation of a deep neural network model to predict postoperative mortality, acute kidney
558 injury, and reintubation using a single feature set, *npj Digital Medicine* (2020).
- 559 ²¹ Y. Si, K. Roberts, Deep Patient Representation of Clinical Notes via Multi-Task Learning
560 for Mortality Prediction., *AMIA Joint Summits on Translational Science proceedings*.
561 *AMIA Joint Summits on Translational Science 2019* (2019) 779–788.
- 562 ²² E. Choi, M. T. Bahadori, J. Sun, Doctor AI: predicting clinical events via recurrent neural
563 networks, *CoRR* abs/1511.05942 (2015).
- 564 ²³ H. Harutyunyan, H. Khachatrian, D. C. Kale, G. Ver Steeg, A. Galstyan, Multitask learn-
565 ing and benchmarking with clinical time series data, *Scientific data* 6 (1) (2019) 96.
- 566 ²⁴ M. C. Moghadam, E. M. K. Abad, N. Bagherzadeh, D. Ramsingh, G. P. Li, Z. N. Kain, A
567 machine-learning approach to predicting hypotensive events in ICU settings, *Computers*
568 *in Biology and Medicine* 118 (2020) 103626.
- 569 ²⁵ K. Maheshwari, T. Shimada, J. Fang, I. Ince, E. J. Mascha, A. Turan, A. Kurz, D. I.
570 Sessler, Hypotension Prediction Index software for management of hypotension during
571 moderate- to high-risk noncardiac surgery: Protocol for a randomized trial, *Trials* 20 (1)
572 (2019) 255.

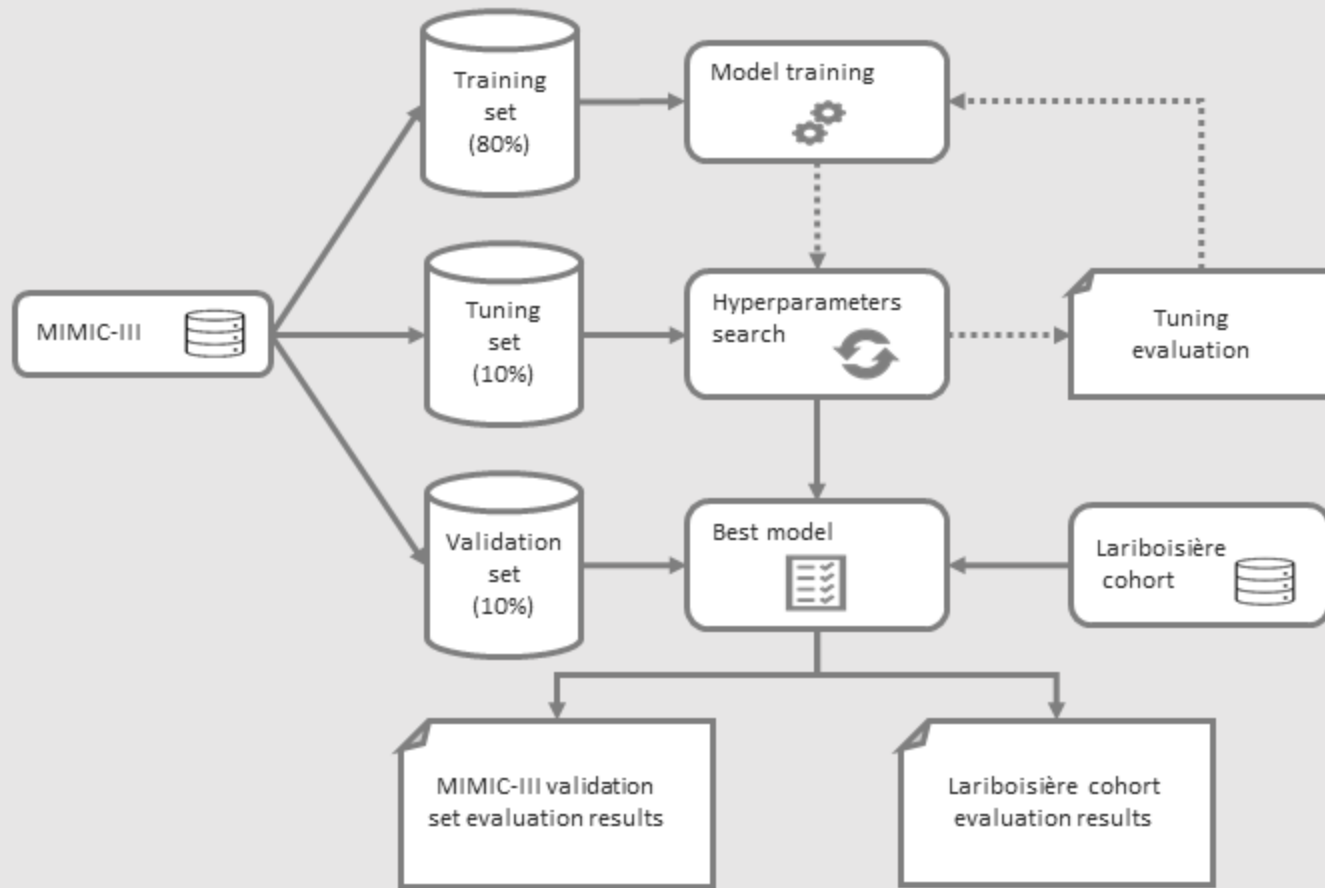
- 573 ²⁶ S. J. Davies, S. T. Vistisen, Z. Jian, F. Hatib, T. W. L. Scheeren, Ability of an Arterial
574 Waveform Analysis-Derived Hypotension Prediction Index to Predict Future Hypotensive
575 Events in Surgical Patients, *Anesthesia & Analgesia* 130 (2) (2020) 352–359.
- 576 ²⁷ B. Chan, B. Chen, A. Sedghi, P. Laird, D. Maslove, P. Mousavi, Generalizable deep
577 temporal models for predicting episodes of sudden hypotension in critically ill patients: a
578 personalized approach, *Scientific Reports* (2020).
- 579 ²⁸ A. E. Johnson, L. L. Pollard, Tom J. and Shen, L. wei H., M. Feng, M. Ghassemi, B. Moody,
580 P. Szolovits, L. Anthony Celi, R. G. Mark, Mimic-iii, a freely accessible critical care
581 database, *Scientific Data* 3 (2016) 160035.
- 582 ²⁹ A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E.
583 Mietus, G. B. Moody, C. K. Peng, H. E. Stanley, PhysioBank, PhysioToolkit, and Phys-
584 ioNet: components of a new research resource for complex physiologic signals., *Circulation*
585 101 (23) (jun 2000).
- 586 ³⁰ E. Toulouse, B. Lafont, S. Granier, G. Mcgurk, J. E. Bazin, French legal approach to
587 patient consent in clinical research, *Anaesthesia Critical Care and Pain Medicine* 39 (6)
588 (2020) 883–885.
- 589 ³¹ P. H. C. Chen, Y. Liu, L. Peng, How to develop machine learning models for healthcare,
590 *Nature Materials* 18 (5) (2019) 410–414.
- 591 ³² Y. Bengio, R. Ducharme, P. Vincent, C. Jauvin, A neural probabilistic language model,
592 *Journal of machine learning research* 3 (2003) 1137–1155.
- 593 ³³ Y. Li, L. Yao, C. Mao, A. Srivastava, X. Jiang, Y. Luo, Early prediction of acute kidney
594 injury in critical care setting using clinical notes, *Proceedings - 2018 IEEE International*
595 *Conference on Bioinformatics and Biomedicine, BIBM 2018* 2018 (2019) 683–686.
- 596 ³⁴ J. Ye, L. Yao, J. Shen, R. Janarthnam, Y. Luo, Predicting mortality in critically ill
597 patients with diabetes using machine learning and clinical notes, *BMC Medical Informatics*
598 *and Decision Making* 20 (2020) 295.
- 599 ³⁵ S. Barbieri, J. Kemp, O. Perez-Concha, S. Kotwal, M. Gallagher, A. Ritchie, L. Jorm,
600 Benchmarking Deep Learning Architectures for Predicting Readmission to the ICU and
601 Describing Patients-at-Risk, *Scientific Reports* 10 (1) (2020) 1–10.
- 602 ³⁶ J. R. Le Gall, S. Lemeshow, F. Saulnier, A new Simplified Acute Physiology Score (SAPS
603 II) based on a European/North American multicenter study., *JAMA* 270 (1993) 2957–63.
- 604 ³⁷ J. L. Vincent, R. Moreno, J. Takala, S. Willatts, A. De Mendonça, H. Bruining, C. K.
605 Reinhart, P. M. Suter, L. G. Thijs, The SOFA (Sepsis-related Organ Failure Assessment)
606 score to describe organ dysfunction/failure. On behalf of the Working Group on Sepsis-
607 Related Problems of the European Society of Intensive Care Medicine., *Intensive care*
608 *medicine* 22 (1996) 707–10.

- 609 ³⁸ K. Cho, B. van Merriënboer, Ç. Gülçehre, F. Bougares, H. Schwenk, Y. Bengio, Learn-
610 ing phrase representations using RNN encoder-decoder for statistical machine translation,
611 CoRR abs/1406.1078 (2014).
- 612 ³⁹ Y. Lecun, Y. Bengio, G. Hinton, Deep learning, *Nature* 521 (7553) (2015) 436–444.
- 613 ⁴⁰ S. Wu, G. Li, L. Deng, L. Liu, D. Wu, Y. Xie, L. Shi, L1 -Norm Batch Normalization for
614 Efficient Training of Deep Neural Networks, *IEEE Transactions on Neural Networks and*
615 *Learning Systems* 30 (7) (2019) 2043–2051.
- 616 ⁴¹ K. Hara, H. Shouno, D. Saito, Analysis of function of rectified linear unit used in deep
617 learning Analysis of Bayesian approach of image restoration View project Deep Convolu-
618 tion Neural Network improvement View project Analysis of Function of Rectified Linear
619 Unit Used in Deep learning, 2015 International Joint Conference on Neural Networks
620 (IJCNN) (2015).
- 621 ⁴² R. Caruana, Multitask Learning, in: *Learning to Learn*, Springer US, 1998, pp. 95–133.
- 622 ⁴³ A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison,
623 L. Antiga, A. Lerer, Automatic differentiation in pytorch, NIPS-W (2017).
- 624 ⁴⁴ D. P. Kingma, J. Lei Ba, Adam: A method for stochastic optimization (2015).
625 arXiv:1412.6980v9.
- 626 ⁴⁵ R Core Team, R: A Language and Environment for Statistical Computing, R Foundation
627 for Statistical Computing, Vienna, Austria (2020).
- 628 ⁴⁶ M. Cherifa, A. Blet, A. Chambaz, E. Gayat, M. Resche-Rigon, R. Pirracchio, Prediction of
629 an acute hypotensive episode during an ICU hospitalization with a super learner machine-
630 learning algorithm, *Anesthesia and Analgesia* 130 (5) (2020) 1157–1166.
- 631 ⁴⁷ S. L. Hyland, M. Faltys, M. Hüser, X. Lyu, T. Gumbsch, C. Esteban, C. Bock, M. Horn,
632 M. Moor, B. Rieck, M. Zimmermann, D. Bodenham, K. Borgwardt, G. Rätsch, T. M. Merz,
633 Early prediction of circulatory failure in the intensive care unit using machine learning,
634 *Nature Medicine* 26 (3) (2020) 364–373.
- 635 ⁴⁸ S. Bangdiwala, V. Shankar, The agreement chart, *BMC medical research methodology* 13
636 (2013) 97.
- 637 ⁴⁹ G. Moody, L. Lehman, Predicting acute hypotensive episodes: The 10th annual phy-
638 sionet/computers in cardiology challenge, in: *Computers in Cardiology*, 2009, Vol.
639 36(5445351), 2009, pp. 541–544.
- 640 ⁵⁰ S. Ruder, An overview of multi-task learning in deep neural networks, CoRR
641 abs/1706.05098 (2017).
- 642 ⁵¹ E. Futier, J. Y. Lefrant, P. G. Guinot, T. Godet, E. Lorne, P. Cuvillon, S. Bertran,
643 M. Leone, B. Pastene, V. Piriou, S. Molliex, J. Albanese, J. M. Julia, B. Tavernier,
644 E. Imhoff, J. E. Bazin, J. M. Constantin, B. Pereira, S. Jaber, Effect of individualized

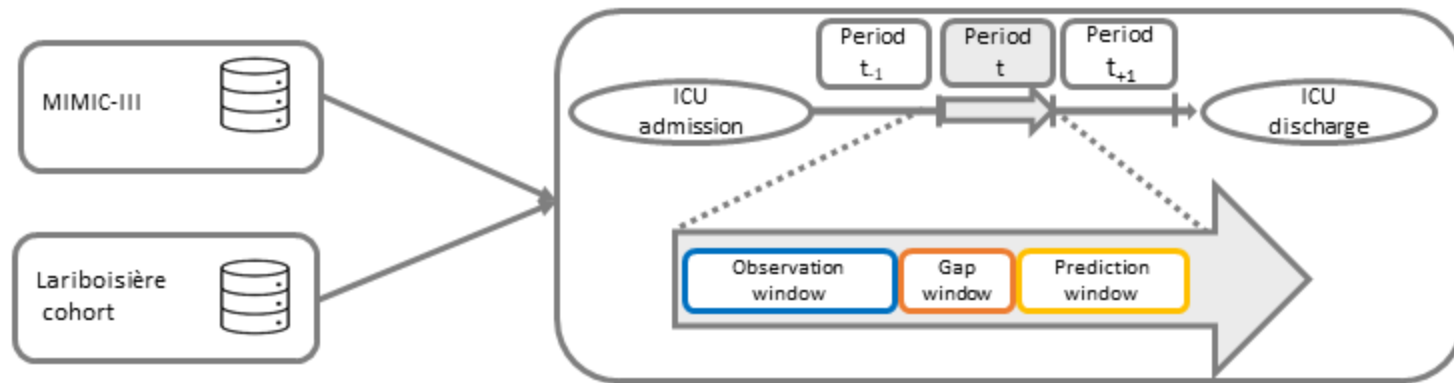
645 vs standard blood pressure management strategies on postoperative organ dysfunction
646 among high-risk patients undergoing major surgery: A randomized clinical trial, JAMA
647 - Journal of the American Medical Association (2017). doi:10.1001/jama.2017.14172.

648 ⁵² A. Meyer, D. Zverinski, B. Pfahringer, J. Kempfert, T. Kuehne, S. H. Sündermann,
649 C. Stamm, T. Hofmann, V. Falk, C. Eickhoff, Machine learning for real-time prediction
650 of complications in critical care: a retrospective study, *The Lancet Respiratory Medicine*
651 6 (12) (2018) 905–914.

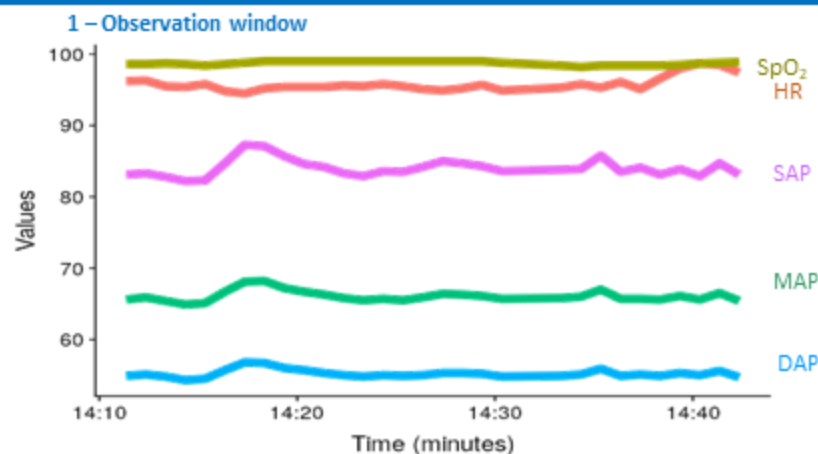
A



B



- Data: recorded and used to train the models
 - patients' characteristics (age, gender)
 - initial severity scores (SOFA, SAPS-II)
 - type of intensive care unit.
 - treatments (sedation, vasopressors, mechanical ventilation)
 - physiological signals (pulse oximetry, heart rate, systolic arterial pressure, mean arterial pressure and diastolic arterial pressure)
- Duration: 30 minutes

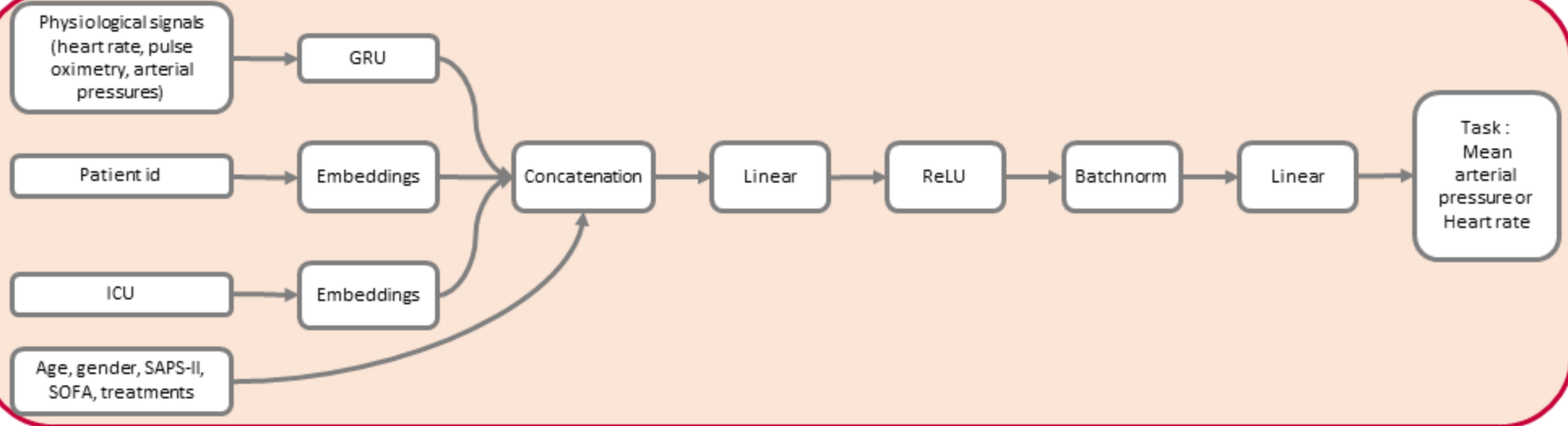


- 2 - Gap window
- Gap allows therapeutic adjustment before the prediction period
 - Different durations are tested: 5, 10, 15, 30 and 60 minutes

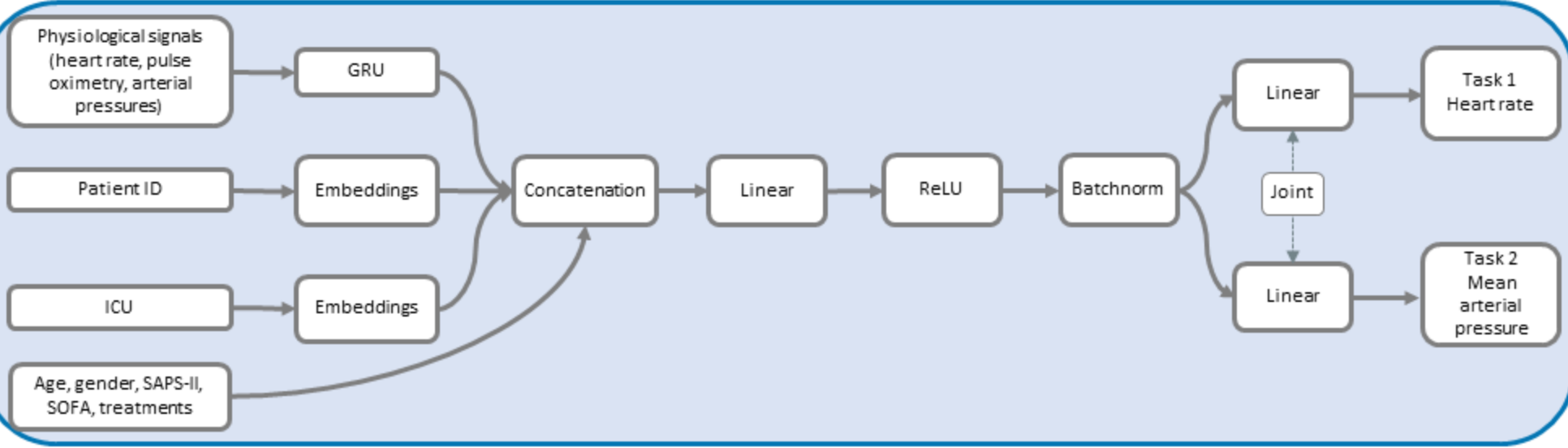
- 3 - Prediction window
- Predictions on mean arterial pressure and heart rate, based on the data recorded during the observation period
 - Gap are designed to allow therapeutic adjustment to avoid the prediction
 - Duration: 5 minutes

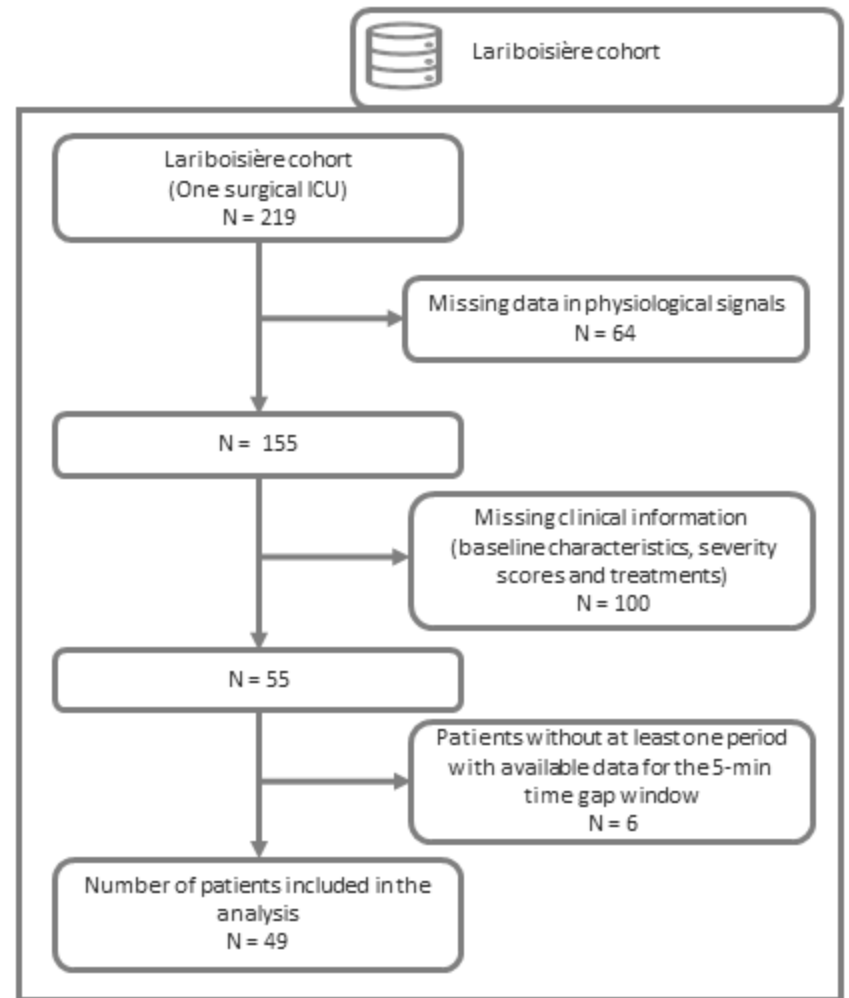
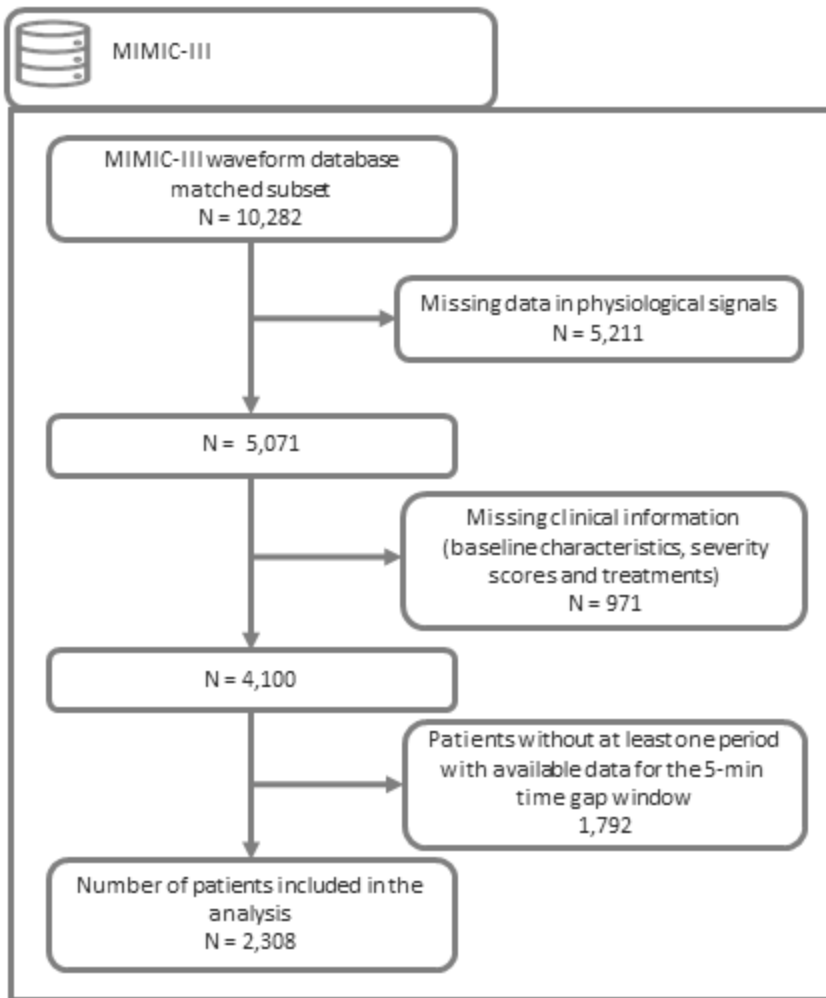


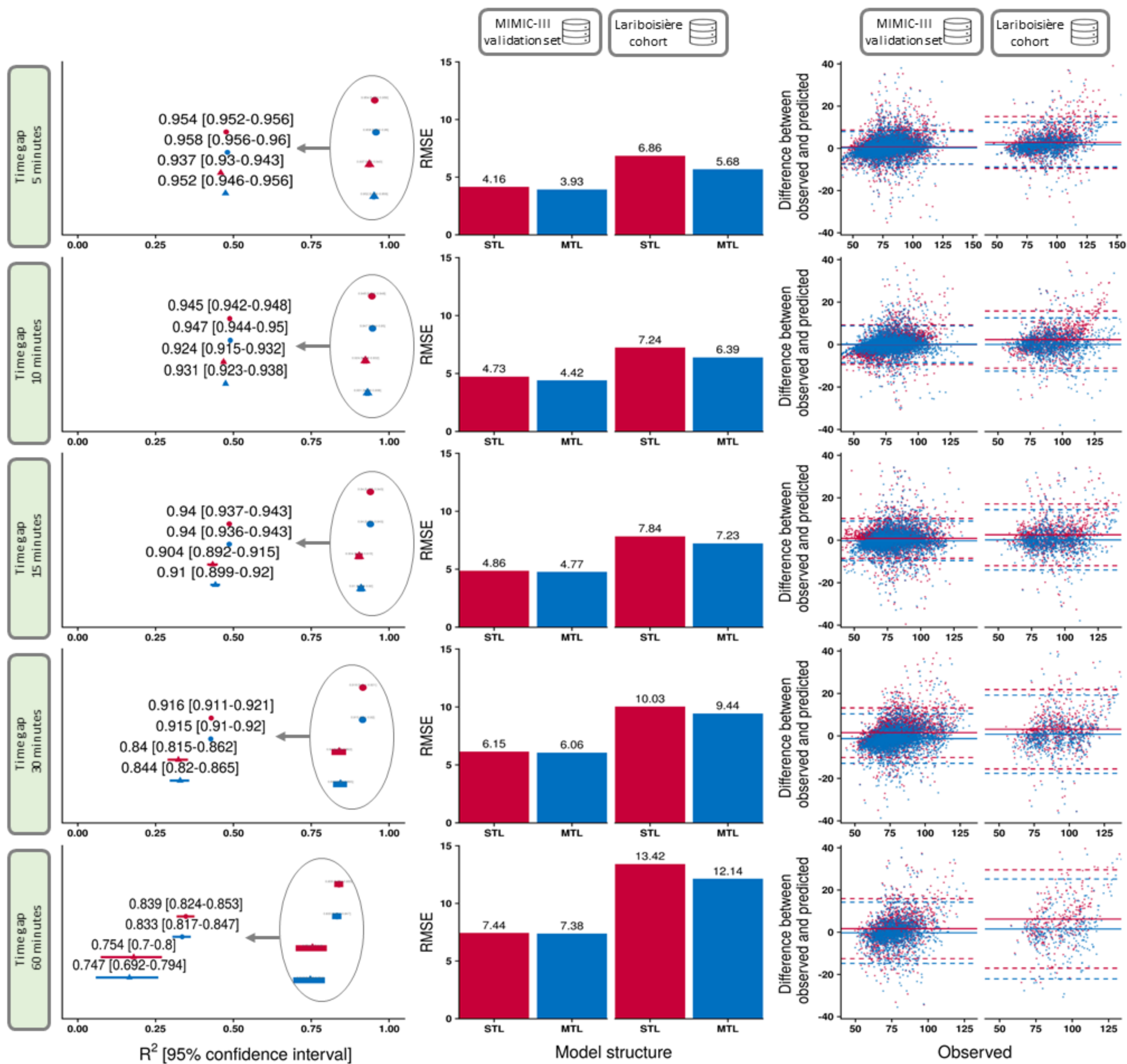
Single-task learning



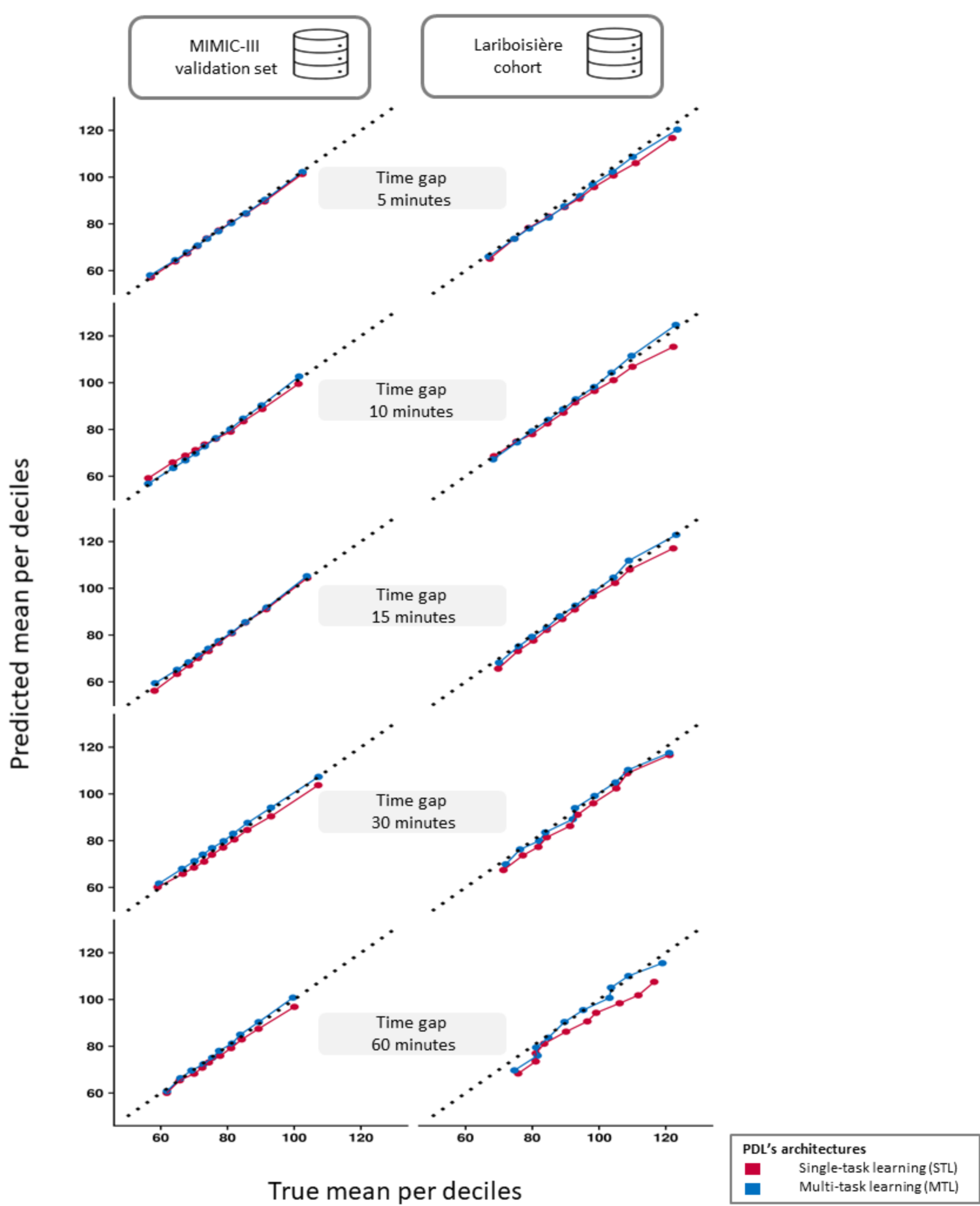
Multi-task learning







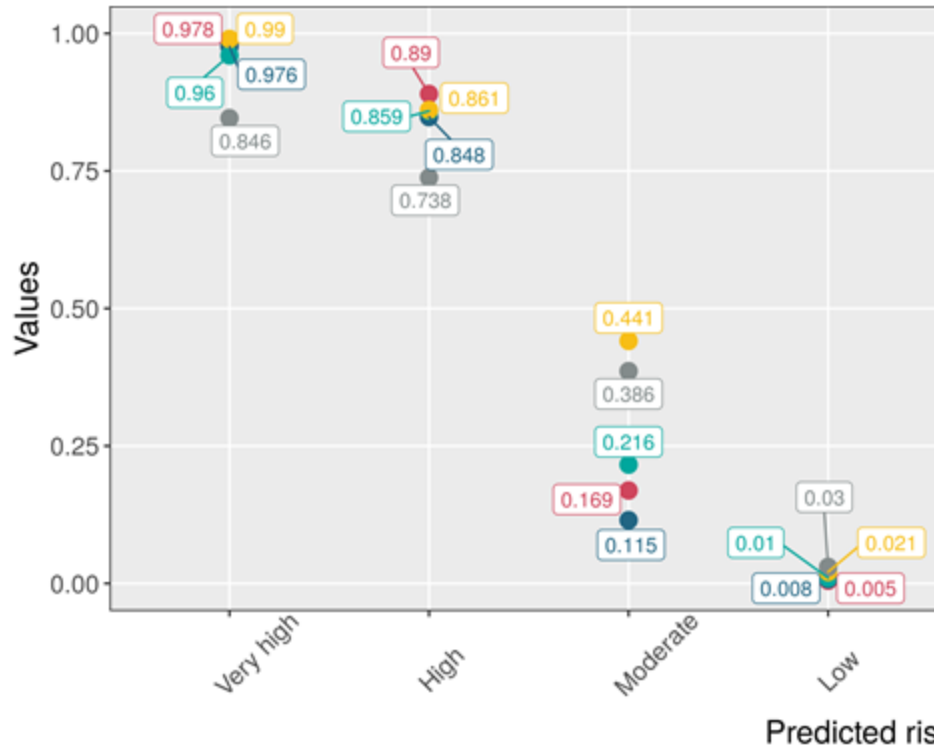
PDL's architectures ■ Single-task learning (STL) ■ Multi-task learning (MTL)
Datasets ○ MIMIC-III validation set △ Lariboisière cohort



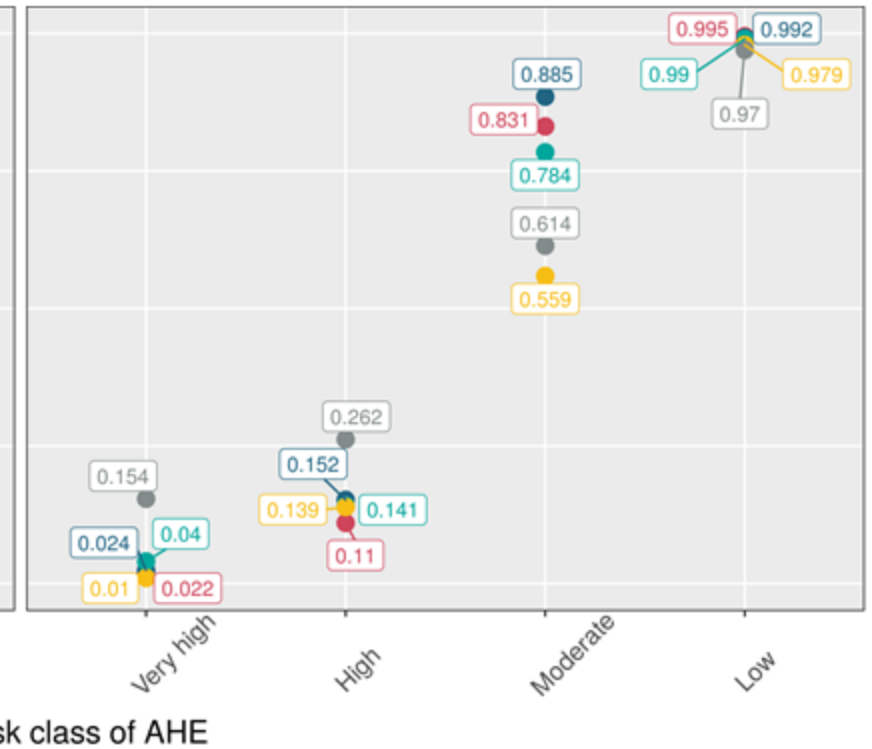
MIMIC-III
validation set



$P(\text{AHE} | k)$



$P(\text{No AHE} | k)$

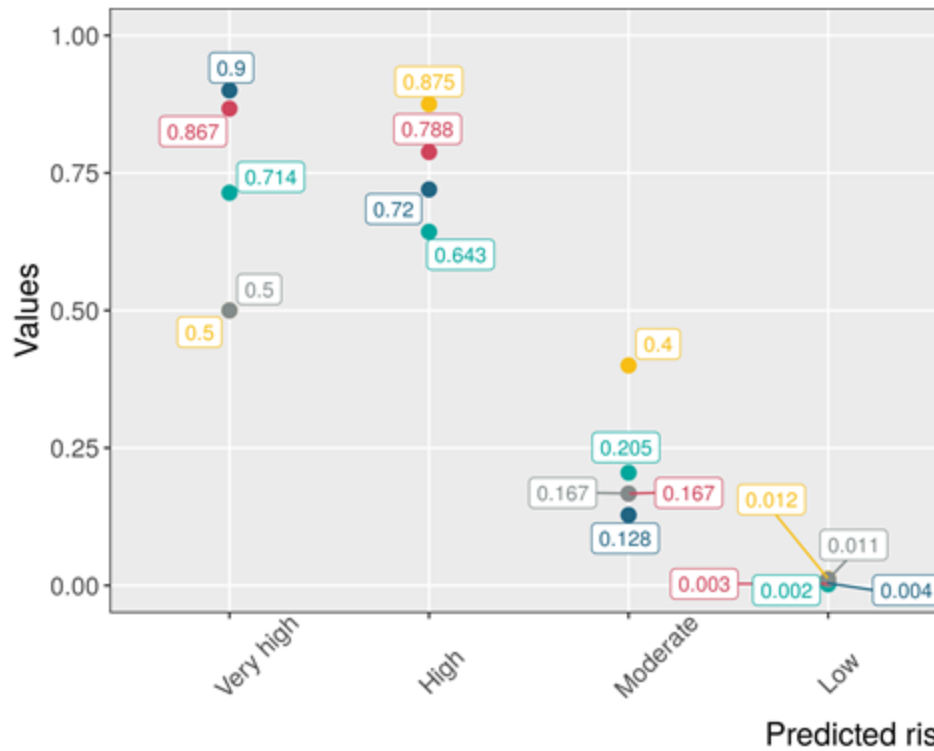


Lariboisière
cohort

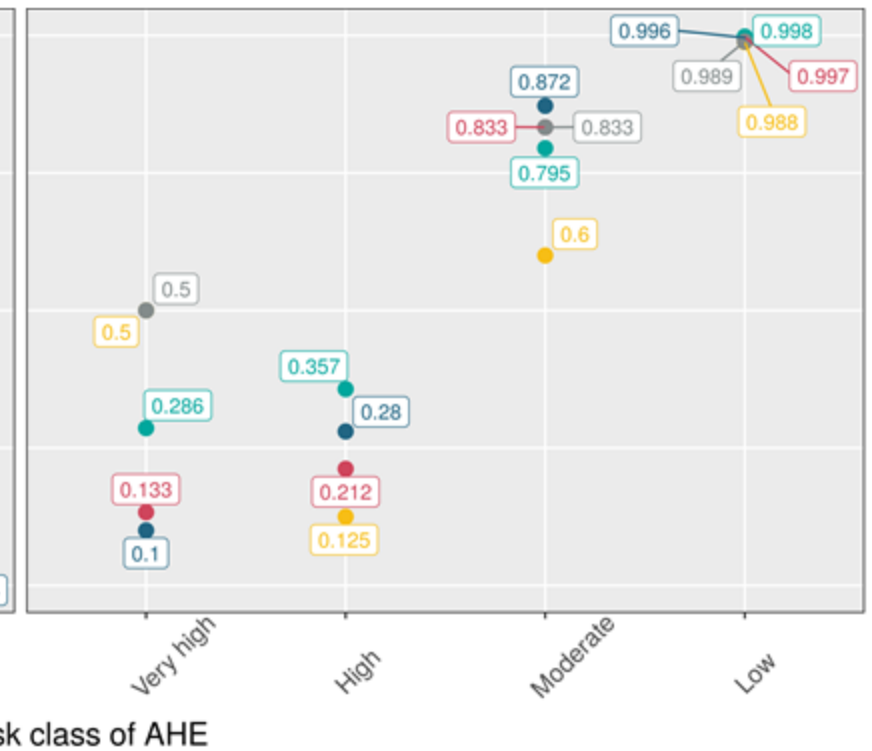


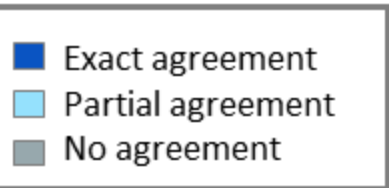
Time gap (min) ● 5 ● 10 ● 15 ● 30 ● 60

$P(\text{AHE} | k)$

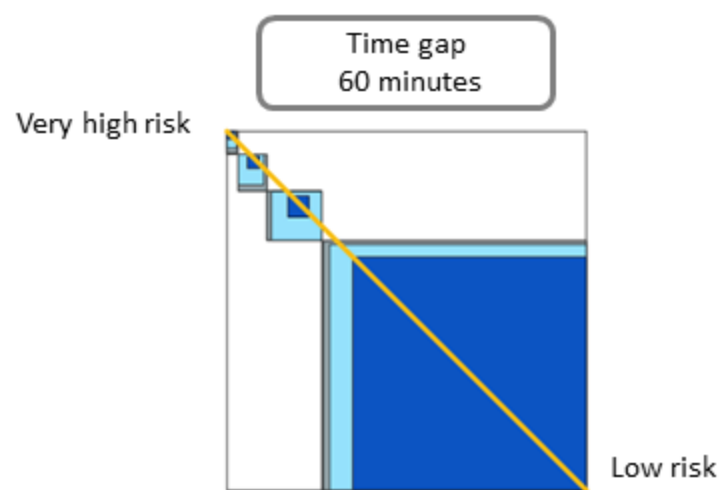
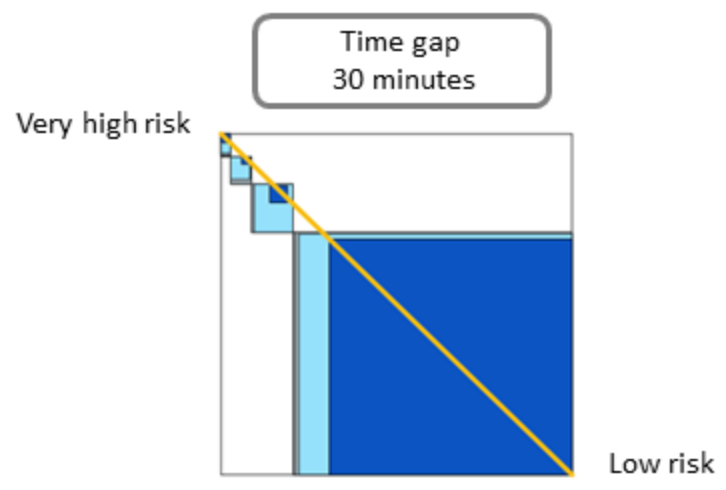
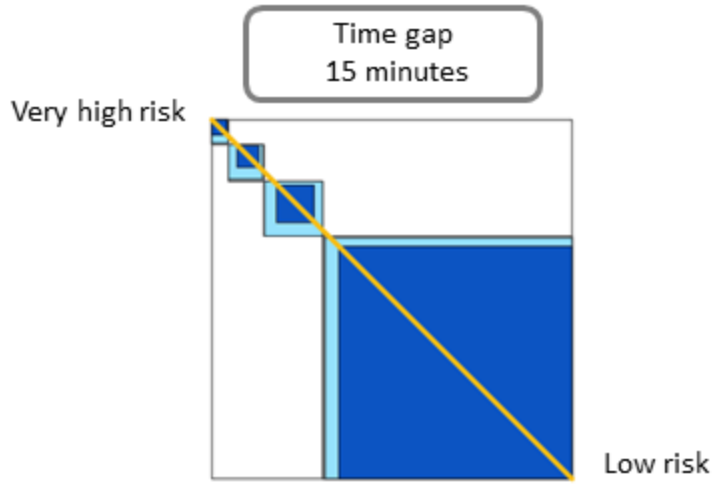
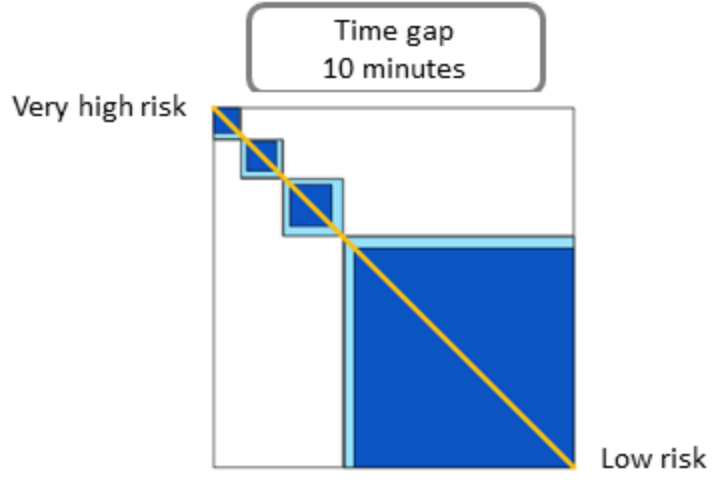
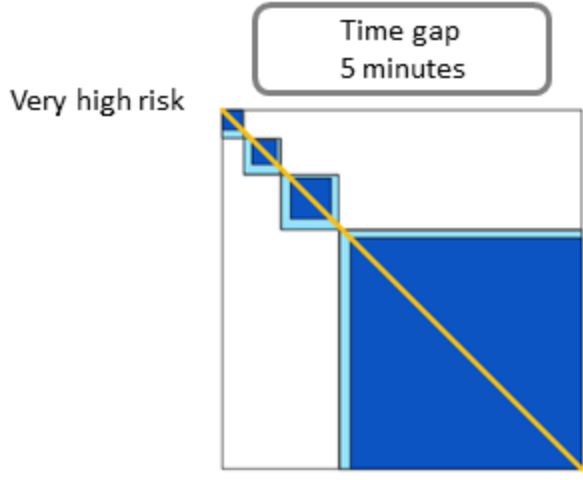


$P(\text{No AHE} | k)$





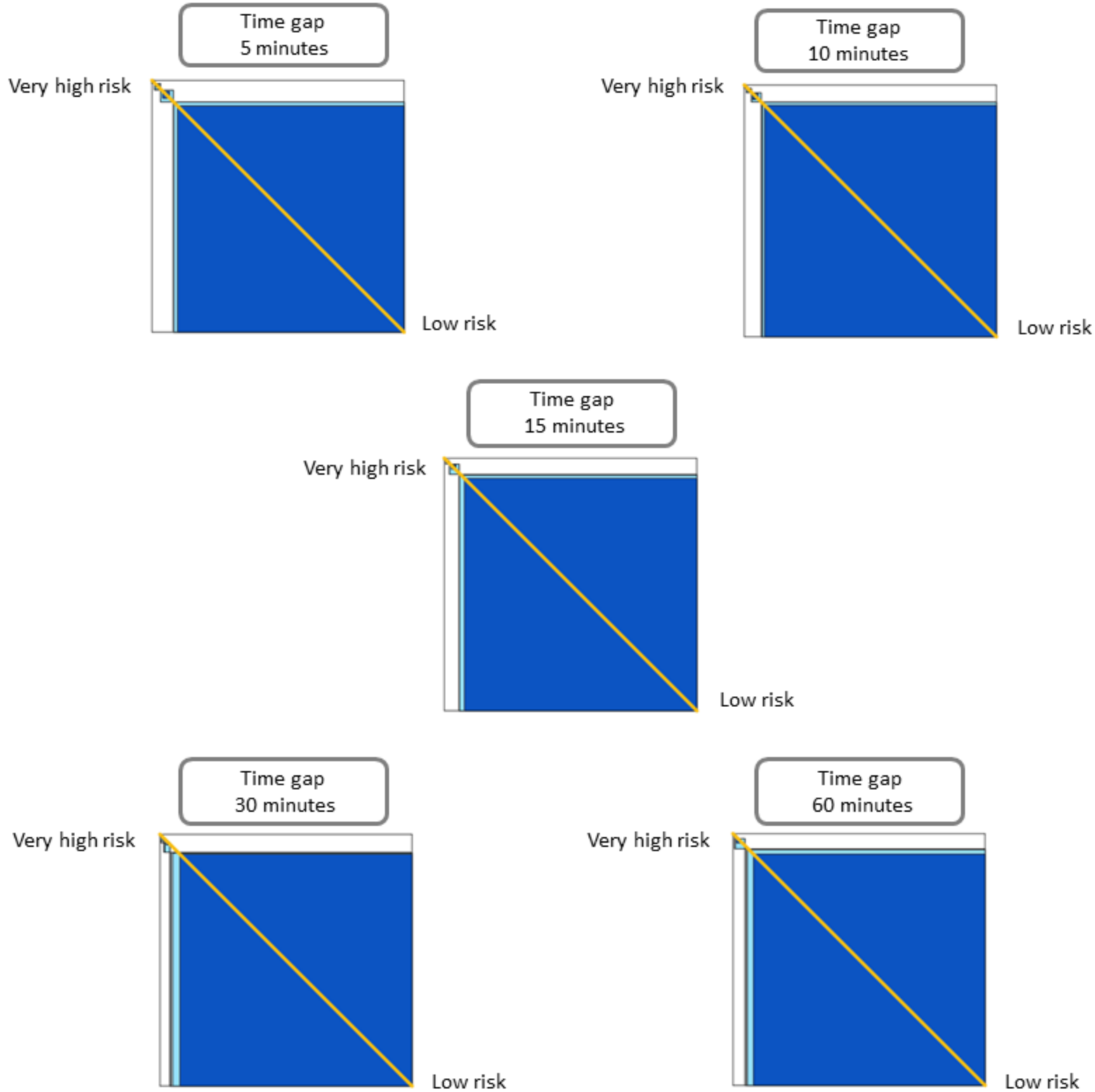
Occurrence of an acute hypotensive episodes predicted



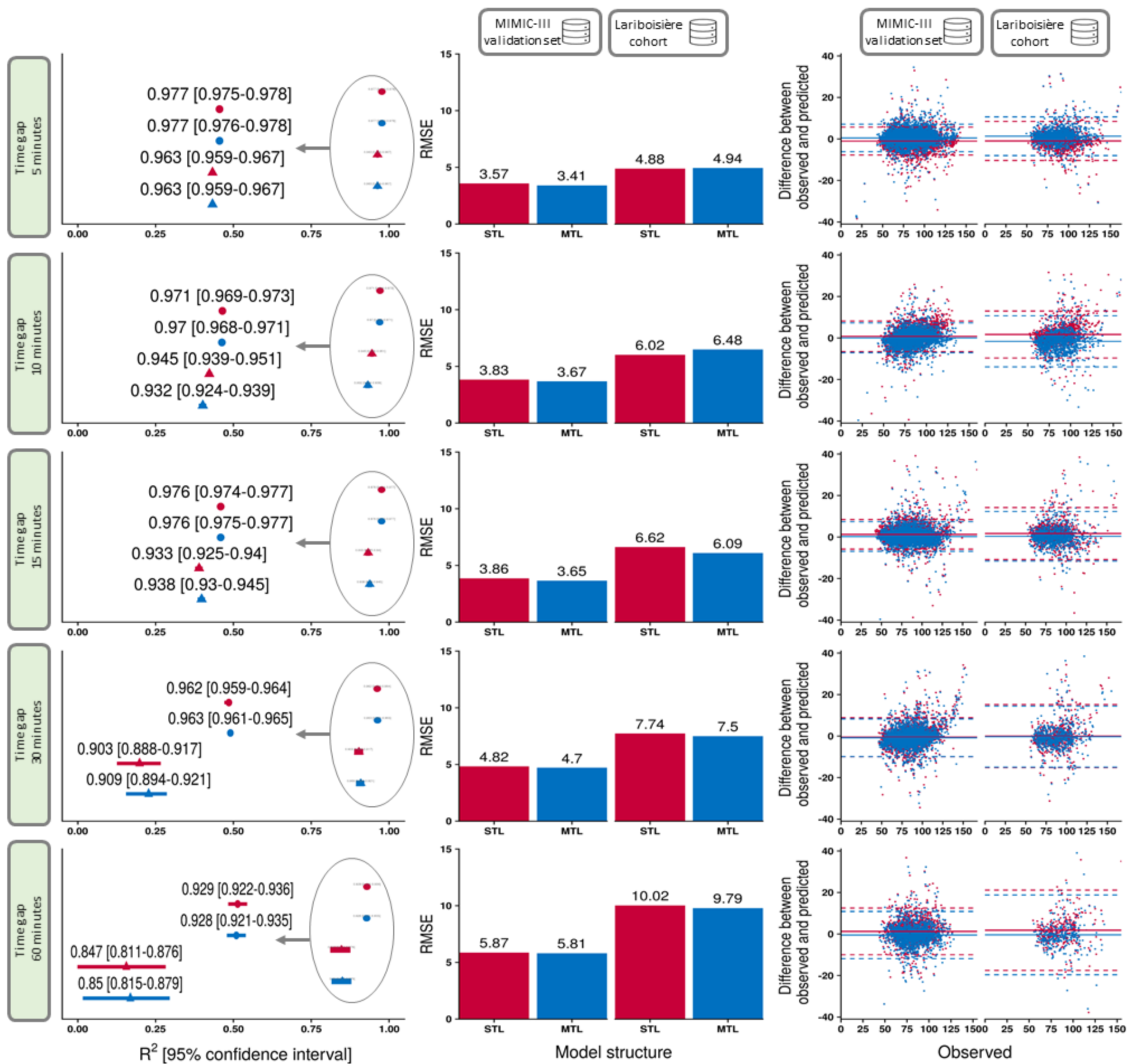
Occurrence of an acute hypotensive episodes observed

- Exact agreement
- Partial agreement
- No agreement

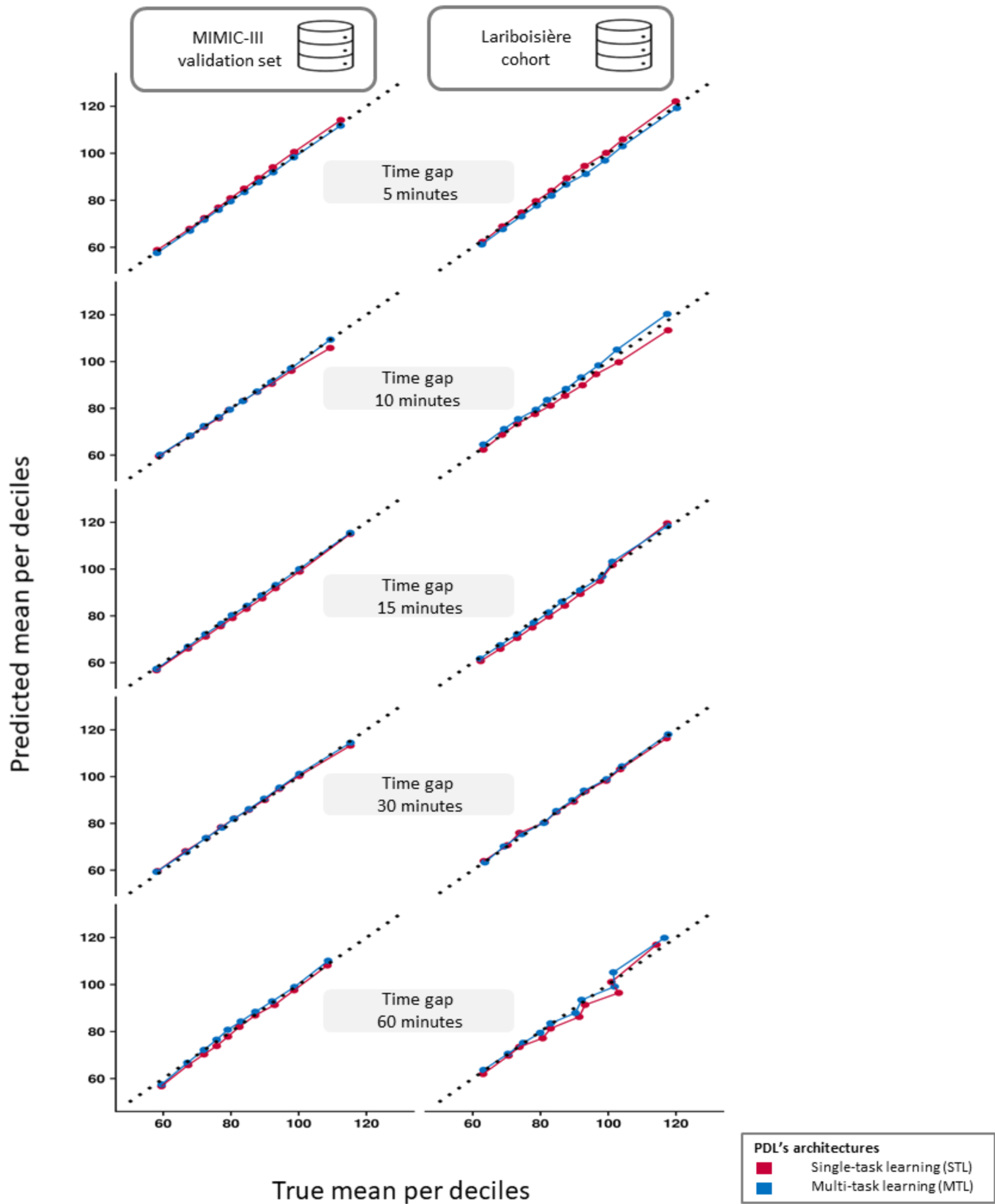
Occurrence of an acute hypotensive episodes predicted



Occurrence of an acute hypotensive episodes observed



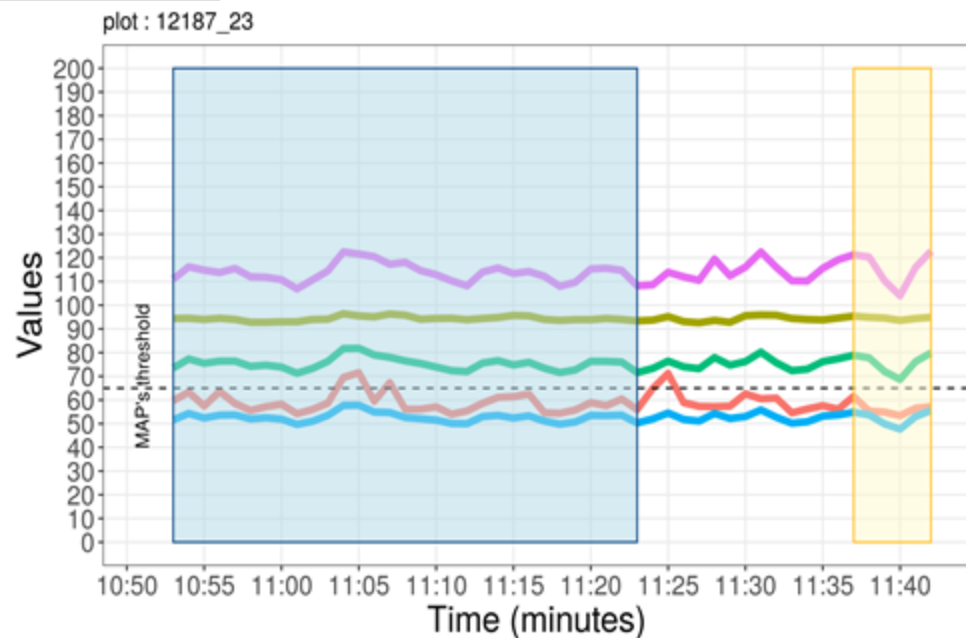
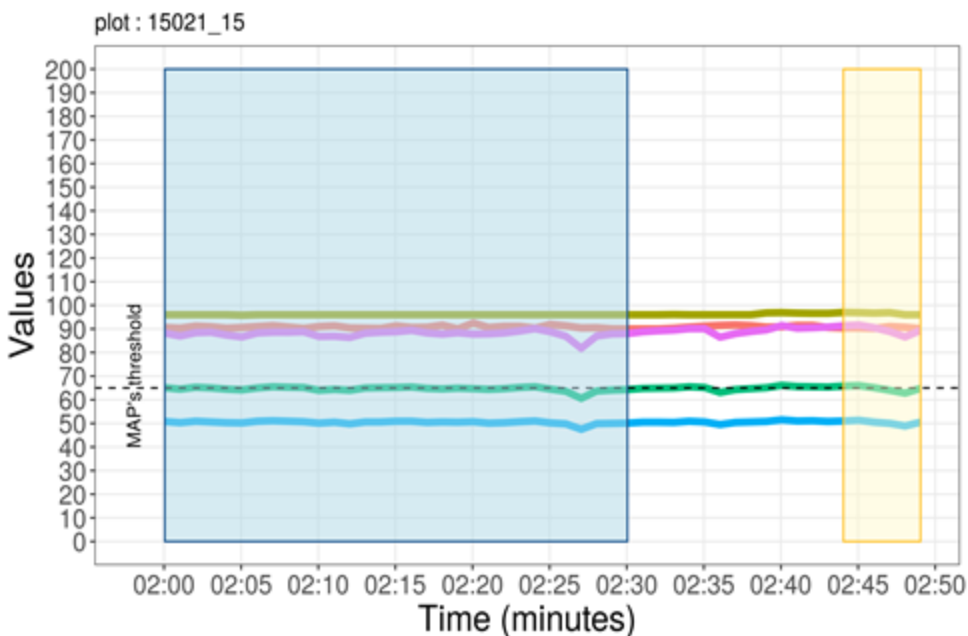
PDL's architectures ■ Single-task learning (STL) ■ Multi-task learning (MTL)
Datasets ○ MIMIC-III validation set △ Lariboisière cohort



A

30-min observation window
5-min prediction window

C



Physiological signals

Heart rate
Pulse oximetry
Mean arterial pressure
Diastolic arterial pressure
Systolic arterial pressure

B

D

