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

Version of Record: https://www.sciencedirect.com/science/article/pii/S0933365721001111
Manuscript_a8e0ac7b24afdcea8ac820fa02ee7950

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,*]

[a] *Université 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*
[b] *Data Analytic Program, University of California San Francisco, 101 Howard St, San Francisco, 94105, CA, United States of America*
[c] *Department of Anesthesiology, Critical Care and Burn Center, Lariboisière-Saint-Louis Hospital, AP-HP Nord, 1 Avenue Claude Vellefaux, Paris, 75010, France*
[d] *Université de Paris, Cardiovascular MArkers in Stress Conditions (MASCOT) - INSERM UMR-S 942, 41, boulevard de la Chapelle, Paris, 75010, France*
[e] *University of Ottawa Heart Institute, University of Ottawa, 40 Ruskin St, Ottawa, Ontario, Canada*
[f] *Department of biostastistics and medical information, Lariboisière-Saint-Louis Hospital, AP-HP Nord, 1 Avenue Claude Vellefaux, Paris, 75010, France*
[g] *Department 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
[1]both equally contributed

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

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

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

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

2. Methods

2.1. Datasets

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

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

2.2. Data partitioning

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

2.3. Periods'definition

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

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

2.4. Sampling periods and predicting on new patients

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

i. The model should not overfit to patients with more data, i.e. more periods

ii. The model should be aware of the correlation between periods coming from the same patient

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

2.5. Predictors and outcomes

Fixed predictors included baseline characteristics:

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

Two types of time-dependent characteristics were considered:

- Period-evolving treatment characteristics: status of mechanical ventilation; administration 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.

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

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

2.7. Model Optimization

2.7.1. Single-task learning

Let $\mathcal{D} = \{\{x^{(ip)}, y^{(ip)}_{MAP}, y^{(ij)}_{HR}\}^t_{p=1}\}^n_{i=1}$ 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 $f(.)$ 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, $f_{MAP}(.)$ and $f_{HR}(.)$. Then, ℓ_{MAP} and ℓ_{HR} are defined as follows:

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

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

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

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 $f(.)$ the joint clinical risk model. Hence, our global multi-task learning loss of $\mathcal{L}$ is defined as follows:

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

PDLs development was performed using PyTorch version 1.4 library.[43] We used Adam[44] for optimization. It is computationally efficient, has little memory requirements, is invariant to diagonal re-scaling of the gradients. The optimization of hyperparameters was based on the best models'performances on the tuning set. Thus, we fixed the learning rate to 0.003, the hidden size to 100 units, the number of epochs to 20, and the weight decay to 10^{-5}. Then best models were fitted on the two validation sets to report final performance results. Graphical representations of the final results were performed using R version 3.6.3 library.[45]

2.8. Assessment of performance

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

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

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

2.8.2. Acute hypotensive episodes prediction

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

i. "AHE", if observed average 5-min MAP ≤ 65

ii. "No AHE", either

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

i. "Very high", if predicted MAP ≤ 60

ii. "High", if $60 <$ predicted MAP ≤ 65

iii. "Moderate", if $65 <$ predicted MAP ≤ 70

iv. "Low", if predicted MAP > 70

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

3. Results

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

MAP prediction.. STL- and MTL-PDL performance for each time gap (5, 10, 15, 30, and 60-min) are presented in Fig. 4. When evaluated using the MIMIC-III validation set and the external validation cohort, the correlation coefficient R^2 (95% confidence interval) between the actual and the predicted MAP was consistently very close to 1 whatever the PDL architecture and the time gap. With a STL structure, the R^2 obtained in MIMIC-III validation 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 0.754 (0.7-0.8) in the external validation cohort. With a MTL structure, the R^2 ranged from 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 0.747 (0.692-0.794) in the external validation cohort. However, when compared to MTL-PDL, STL-PDL was consistently associated with a larger root mean square error (RMSE) (middle panel, Fig. 4). In MIMIC-III validation set, the RMSE for the STL-PDL was of 4.16, 4.73, 4.86, 6.15 and 7.44 for the 5, 10, 15, 30, and 60-min time gaps respectively. In contrast, the RMSE for the MTL-PDL was consistently lower: 3.93, 4.42, 4.77, 6.06 and 7.38 respectively. A similar pattern was observed in the external validation cohort: STL-PDL RMSE of 6.86, 7.24, 7.84, 10.03, and 13.42 at 5, 10, 15, 30, and 60-min respectively and for MTL-PDL 5.68, 6.39, 7.23, 9.44 and 12.14 respectively. Fig. 4, right panel illustrates the concordance between observed and predicted MAP by quantifying the limits of prediction agreement. In general, all differences between observed and predicted values lied within the 95% Limits Of Agreement (95% LOA) for each time gap and validation set. MTL-PDL was associated with more accurate predictions as illustrated by an average difference between observed and predicted MAP consistently closer to zero. In MIMIC-III validation set, the average difference between observed and predicted MAP for each time gap was of 0.59 [95% 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] 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 [-12.89-10.36], -0.26 [-14.73-14.2] for MTL-PDL. The average difference between observed and predicted MAP were larger in the external validation cohort than in the MIMIC-III validation dataset, but MTL-PDL was also superior to STL-PDL: average difference for the 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], 6.26 [-17.05-29.56] at 5, 10, 15, 30, and 60-min respectively and for the MTL-PDL 1.74 [-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 superiority of the MTL-PDL over STL-PDL was also confirmed using calibration plots (Fig. 5).

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

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

4. Discussion

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

Several AHE prediction models were developed over the past 20 years. In 2009, the 10th annual PhysioNet/Computers in cardiology challenge[49] was set to promote the development of methods for identifying ICU patients at imminent risk of AHE. During this challenge, multiple ML prediction models were proposed.[5,6,7,8,9,10,11,12] However, none of them achieved sufficient accuracy to be adopted in clinical practice. More recently, Hatib et al.[13] used a logistic regression model to predict hypotension based on 3,022 features extracted from the MAP waveform signal. Their model reached a sensitivity of 88% (95% CI, 85 to 90%) and a specificity of 87% (95% CI, 85 to 90%) but tended to underpredict the risk of hypotension in the higher-risk subgroups. Thus far, most predictive models used historical MAP values as their only input variable, ignoring other patient characteristics and/or time-dependent variables, e.g., heart rate, known to be highly correlated with the arterial BP. Our group[46] proposed to use multiple physiological signals in addition to patient and treatment characteristics to train an ensemble machine learning model to predict AHE. This model exhibited promising performance, with an area under the curve (AUC) of 0.890 (95% CI, 0.886 to 0.895). Kendale et al.[14] also used an ensemble learning model to predict hypotension following anesthesia induction using intraoperative vital signs, medications and comorbidities as features and obtained an AUC of 0.74 (95% CI, 0.72 to 0.77).

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

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

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

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

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

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

5. Conclusion

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

Acknowledgments

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

Author Contributions

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

Competing interests

The authors declare no competing interests.

Figures captions

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

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

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

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

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

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

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

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

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

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

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

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

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

References

[1] J.-L. Vincent, D. De Backer, Circulatory Shock, New England Journal of Medicine 369 (2013) 1726–1734.

[2] Y. Sakr, K. Reinhart, J. L. Vincent, C. L. Sprung, R. Moreno, V. M. Ranieri, D. De Backer, D. Payen, Does dopamine administration in shock influence outcome? Results of the Sepsis Occurrence in Acutely Ill Patients (SOAP) Study, Critical Care Medicine 34 (2006) 589–597.

[3] R. F. Wilson, J. A. Wilson, D. Gibson, W. J. Sibbald, Shock in the emergency department, Journal of the American College of Emergency Physicians 5 (9) (1976) 678–690.

[4] M. M. Levy, L. E. Evans, A. Rhodes, The Surviving Sepsis Campaign Bundle: 2018 update (jun 2018).

[5] P. Langley, S. King, D. Zheng, E. Bowers, K. Wang, J. Allen, A. Murray, Predicting acute hypotensive episodes from mean arterial pressure, in: Computers in Cardiology, 2009, Vol. 36, 2009, pp. 553–556.

[6] K. Jin, N. Stockbridge, Smoothing and discriminating MAP data, in: Computers in Cardiology, 2009, Vol. 36, 2009, pp. 633–636.

[7] F. Jousset, M. Lemay, J. Vesin, Computers in Cardiology / Physionet Challenge 2009: Predicting acute hypotensive episodes, in: Computers in Cardiology, 2009, Vol. 36, 2009, pp. 637–640.

[8] T. Ho, X. Chen, Utilizing histogram to identify patients using pressors for acute hypotension, in: Computers in Cardiology, 2009, Vol. 36, 2009, pp. 797–800.

[9] J. H. Henriques, T. R. Rocha, Prediction of Acute Hypotensive Episodes Using Neural Network Multi-models, in: Computers in Cardiology, 2009, Vol. 36, 2009, pp. 549–552.

[10] F. Chiarugi, I. Karatzanis, V. Sakkalis, I. Tsamardinos, T. Dermitzaki, M. Foukarakis, G. Vrouchos, Predicting the occurrence of acute hypotensive episodes: The PhysioNet Challenge, in: Computers in Cardiology, 2009, Vol. 36, 2009, pp. 621–624.

[11] X. Chen, D. Xu, Forecasting acute hypotensive episodes in intensive care patients based on a peripheral arterial blood pressure waveform, in: Computers in Cardiology, Vol. 36, 2009, pp. 545–548.

[12] P. Fournier, J. Roy, Acute hypotension episode prediction using information divergence for feature selection, and non-parametric methods for classification, in: Computers in Cardiology, 2009, Vol. 36, 2009, pp. 625–628.

[13] F. Hatib, Z. Jian, S. Buddi, C. Lee, J. Settels, K. Sibert, J. Rinehart, M. Cannesson, Machine-learning Algorithm to Predict Hypotension Based on High-fidelity Arterial Pressure Waveform Analysis., Anesthesiology 129 (2018) 663–674.

[14] S. Kendale, P. Kulkarni, A. D. Rosenberg, J. Wang, Supervised Machine Learning Predictive Analytics for Prediction of Postinduction Hypotension, Anesthesiology 129 (2018) 675–688.

[15] R. Donald, T. Howells, I. Piper, I. Chambers, G. Citerio, P. Enblad, B. Gregson, K. Kiening, J. Mattern, P. Nilsson, Early warning of EUSIG-defined hypotensive events using a Bayesian artificial neural network, in: Intracranail Pressure and Brain Monitoring XIV, Vol. 114, 2012, pp. 39–44.

[16] S. Bhattacharya, V. Huddar, V. Rajan, C. K. Reddy, A dual boundary classifier for predicting acute hypotensive episodes in critical care., PloS one 13(2) (2018) e0193259.

[17] D. P. Barnaby, S. M. Fernando, K. J. Ferrick, C. L. Herry, A. J. Seely, P. E. Bijur, E. John Gallagher, Use of the low-frequency/high-frequency ratio of heart rate variability to predict short-term deterioration in emergency department patients with sepsis, Emergency Medicine Journal 35 (2) (2018) 96–102.

[18] M. Cherifa, R. Pirracchio, What every intensivist should know about Big Data and targeted machine learning in the intensive care unit, Rev Bras Ter Intensiva 31 (4) (2019) 444–446.

[19] R. A. Caruana, Multitask Learning: A Knowledge-Based Source of Inductive Bias, in: Machine Learning Proceedings 1993, Elsevier, 1993, pp. 41–48.

[20] I. S. Hofer, C. Lee, E. Gabel, P. Baldi, M. Cannesson, Article open development and validation of a deep neural network model to predict postoperative mortality, acute kidney injury, and reintubation using a single feature set, npj Digital Medicine (2020).

[21] Y. Si, K. Roberts, Deep Patient Representation of Clinical Notes via Multi-Task Learning for Mortality Prediction., AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science 2019 (2019) 779–788.

[22] E. Choi, M. T. Bahadori, J. Sun, Doctor AI: predicting clinical events via recurrent neural networks, CoRR abs/1511.05942 (2015).

[23] H. Harutyunyan, H. Khachatrian, D. C. Kale, G. Ver Steeg, A. Galstyan, Multitask learning and benchmarking with clinical time series data, Scientific data 6 (1) (2019) 96.

[24] M. C. Moghadam, E. M. K. Abad, N. Bagherzadeh, D. Ramsingh, G. P. Li, Z. N. Kain, A machine-learning approach to predicting hypotensive events in ICU settings, Computers in Biology and Medicine 118 (2020) 103626.

[25] K. Maheshwari, T. Shimada, J. Fang, I. Ince, E. J. Mascha, A. Turan, A. Kurz, D. I. Sessler, Hypotension Prediction Index software for management of hypotension during moderate- to high-risk noncardiac surgery: Protocol for a randomized trial, Trials 20 (1) (2019) 255.

[26] S. J. Davies, S. T. Vistisen, Z. Jian, F. Hatib, T. W. L. Scheeren, Ability of an Arterial Waveform Analysis–Derived Hypotension Prediction Index to Predict Future Hypotensive Events in Surgical Patients, Anesthesia & Analgesia 130 (2) (2020) 352–359.

[27] B. Chan, B. Chen, A. Sedghi, P. Laird, D. Maslove, P. Mousavi, Generalizable deep temporal models for predicting episodes of sudden hypotension in critically ill patients: a personalized approach, Scientific Reports (2020).

[28] A. E. Johnson, L. L. Pollard, Tom J.and Shen, L. wei H., M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, R. G. Mark, Mimic-iii, a freely accessible critical care database, Scientific Data 3 (2016) 160035.

[29] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C. K. Peng, H. E. Stanley, PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals., Circulation 101 (23) (jun 2000).

[30] E. Toulouse, B. Lafont, S. Granier, G. Mcgurk, J. E. Bazin, French legal approach to patient consent in clinical research, Anaesthesia Critical Care and Pain Medicine 39 (6) (2020) 883–885.

[31] P. H. C. Chen, Y. Liu, L. Peng, How to develop machine learning models for healthcare, Nature Materials 18 (5) (2019) 410–414.

[32] Y. Bengio, R. Ducharme, P. Vincent, C. Jauvin, A neural probabilistic language model, Journal of machine learning research 3 (2003) 1137–1155.

[33] Y. Li, L. Yao, C. Mao, A. Srivastava, X. Jiang, Y. Luo, Early prediction of acute kidney injury in critical care setting using clinical notes, Proceedings - 2018 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2018 2018 (2019) 683–686.

[34] J. Ye, L. Yao, J. Shen, R. Janarthanam, Y. Luo, Predicting mortality in critically ill patients with diabetes using machine learning and clinical notes, BMC Medical Informatics and Decision Making 20 (2020) 295.

[35] S. Barbieri, J. Kemp, O. Perez-Concha, S. Kotwal, M. Gallagher, A. Ritchie, L. Jorm, Benchmarking Deep Learning Architectures for Predicting Readmission to the ICU and Describing Patients-at-Risk, Scientific Reports 10 (1) (2020) 1–10.

[36] J. R. Le Gall, S. Lemeshow, F. Saulnier, A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study., JAMA 270 (1993) 2957–63.

[37] J. L. Vincent, R. Moreno, J. Takala, S. Willatts, A. De Mendonça, H. Bruining, C. K. Reinhart, P. M. Suter, L. G. Thijs, The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. On behalf of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine., Intensive care medicine 22 (1996) 707–10.

[38] K. Cho, B. van Merrienboer, Ç. Gülçehre, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using RNN encoder-decoder for statistical machine translation, CoRR abs/1406.1078 (2014).

[39] Y. Lecun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.

[40] S. Wu, G. Li, L. Deng, L. Liu, D. Wu, Y. Xie, L. Shi, L1 -Norm Batch Normalization for Efficient Training of Deep Neural Networks, IEEE Transactions on Neural Networks and Learning Systems 30 (7) (2019) 2043–2051.

[41] K. Hara, H. Shouno, D. Saito, Analysis of function of rectified linear unit used in deep learning Analysis of Bayesian approach of image restoration View project Deep Convolution Neural Network improvement View project Analysis of Function of Rectified Linear Unit Used in Deep learning, 2015 International Joint Conference on Neural Networks (IJCNN) (2015).

[42] R. Caruana, Multitask Learning, in: Learning to Learn, Springer US, 1998, pp. 95–133.

[43] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, A. Lerer, Automatic differentiation in pytorch, NIPS-W (2017).

[44] D. P. Kingma, J. Lei Ba, Adam: A method for stochastic optimization (2015). arXiv:1412.6980v9.

[45] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria (2020).

[46] M. Cherifa, A. Blet, A. Chambaz, E. Gayat, M. Resche-Rigon, R. Pirracchio, Prediction of an acute hypotensive episode during an ICU hospitalization with a super learner machine-learning algorithm, Anesthesia and Analgesia 130 (5) (2020) 1157–1166.

[47] S. L. Hyland, M. Faltys, M. Hüser, X. Lyu, T. Gumbsch, C. Esteban, C. Bock, M. Horn, M. Moor, B. Rieck, M. Zimmermann, D. Bodenham, K. Borgwardt, G. Rätsch, T. M. Merz, Early prediction of circulatory failure in the intensive care unit using machine learning, Nature Medicine 26 (3) (2020) 364–373.

[48] S. Bangdiwala, V. Shankar, The agreement chart, BMC medical research methodology 13 (2013) 97.

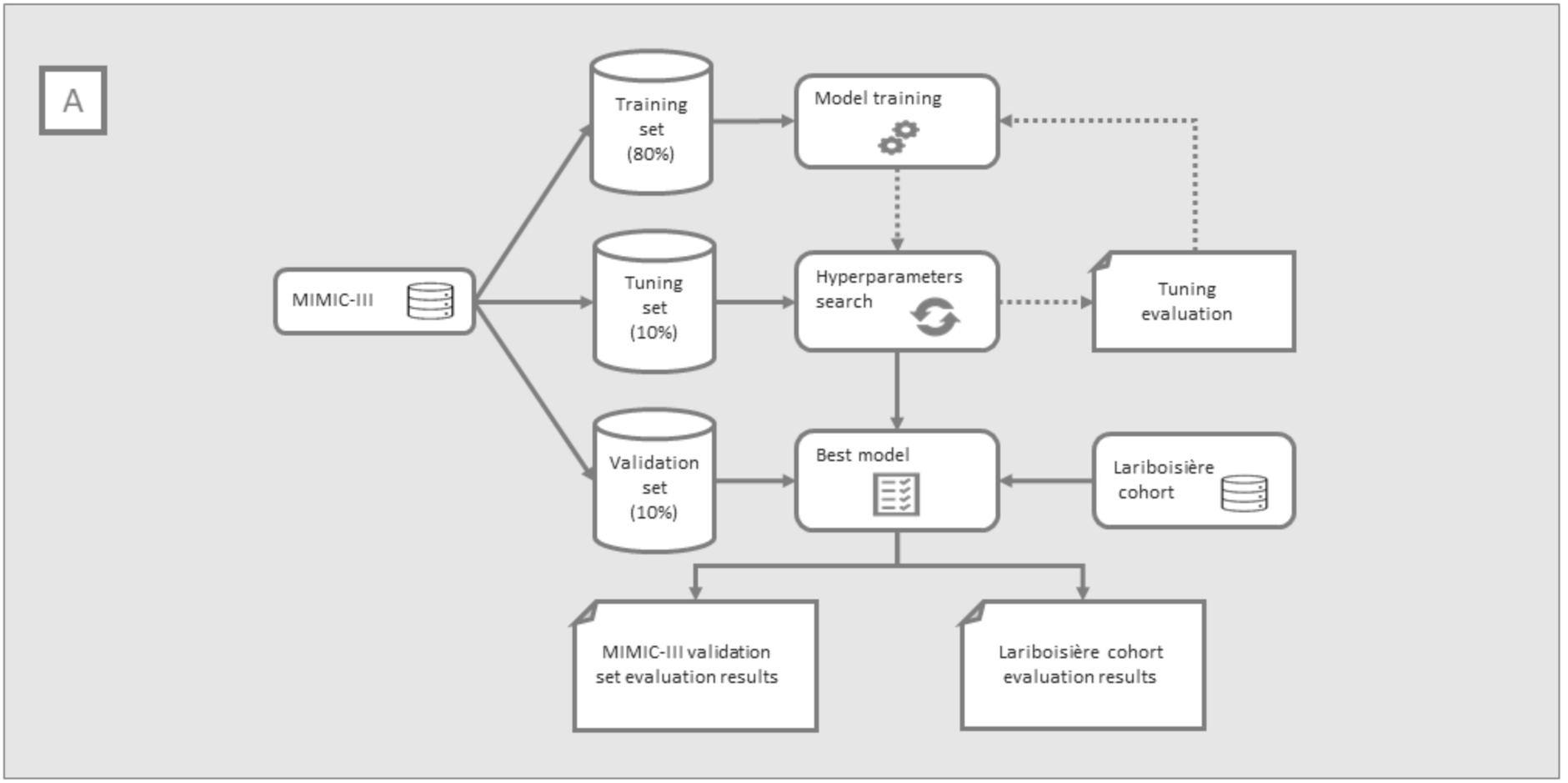
[49] G. Moody, L. Lehman, Predicting acute hypotensive episodes: The 10th annual physionet/computers in cardiology challenge, in: Computers in Cardiology, 2009, Vol. 36(5445351), 2009, pp. 541–544.

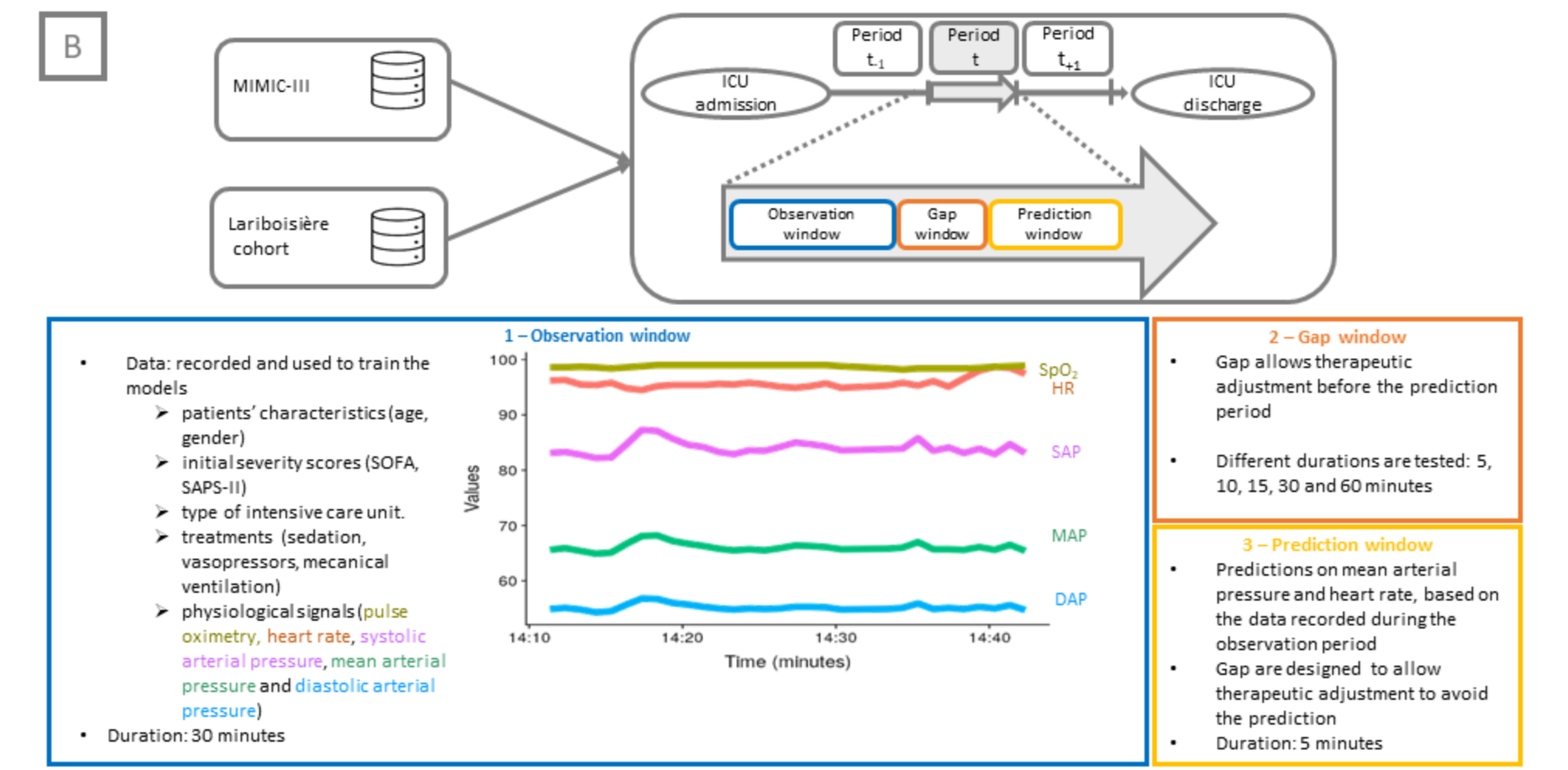
[50] S. Ruder, An overview of multi-task learning in deep neural networks, CoRR abs/1706.05098 (2017).

[51] E. Futier, J. Y. Lefrant, P. G. Guinot, T. Godet, E. Lorne, P. Cuvillon, S. Bertran, M. Leone, B. Pastene, V. Piriou, S. Molliex, J. Albanese, J. M. Julia, B. Tavernier, E. Imhoff, J. E. Bazin, J. M. Constantin, B. Pereira, S. Jaber, Effect of individualized

vs standard blood pressure management strategies on postoperative organ dysfunction 6among high-risk patients undergoing major surgery: A randomized clinical trial, JAMA - Journal of the American Medical Association (2017). doi:10.1001/jama.2017.14172.

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





1 – Observation window

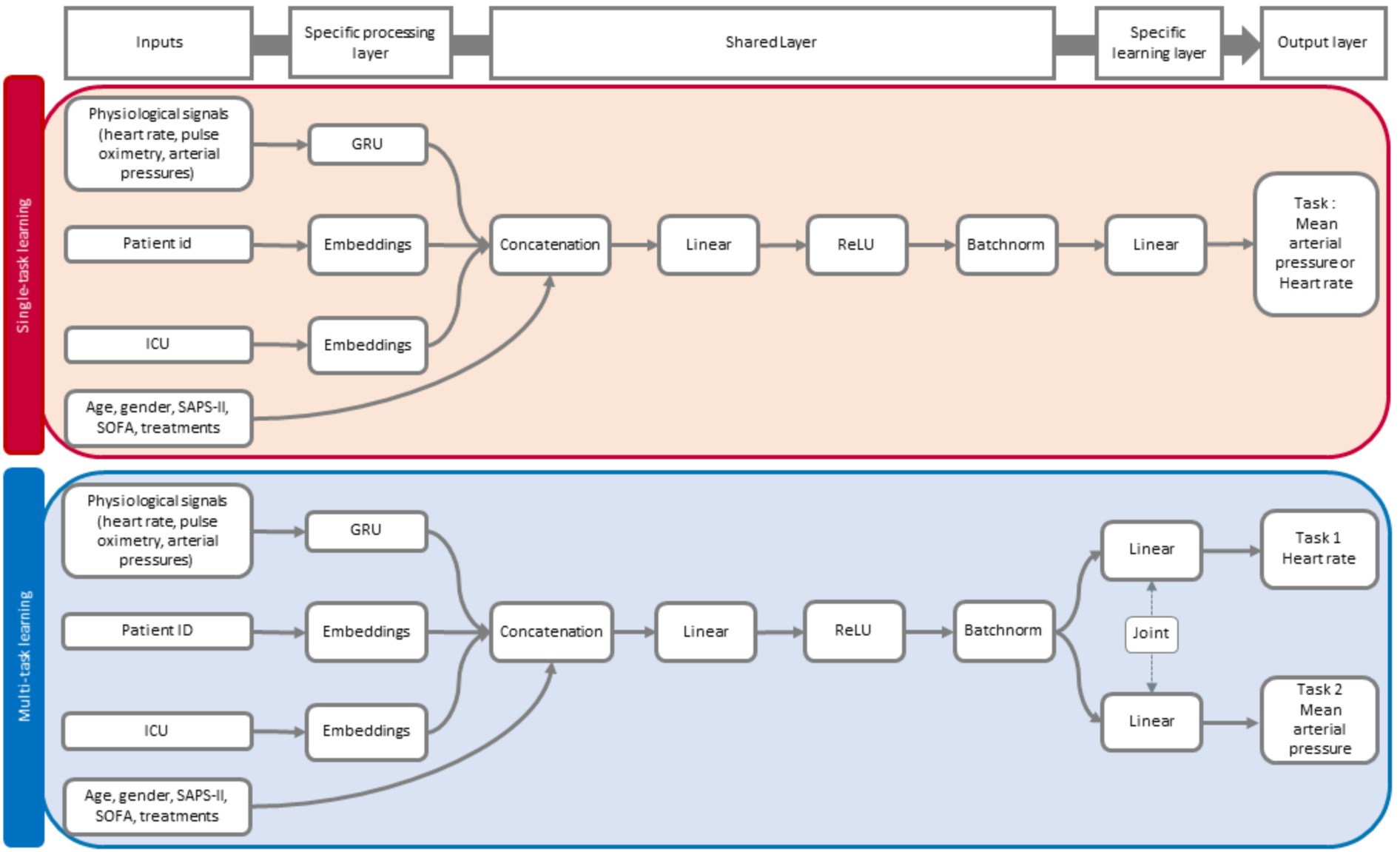
- Data: recorded and used to train the models
  - patients' characteristics (age, gender)
  - initial severity scores (SOFA, SAPS-II)
  - type of intensive care unit.
  - treatments (sedation, vasopressors, mecanical ventilation)
  - physiological signals (pulse oximetry, heart rate, systolic arterial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

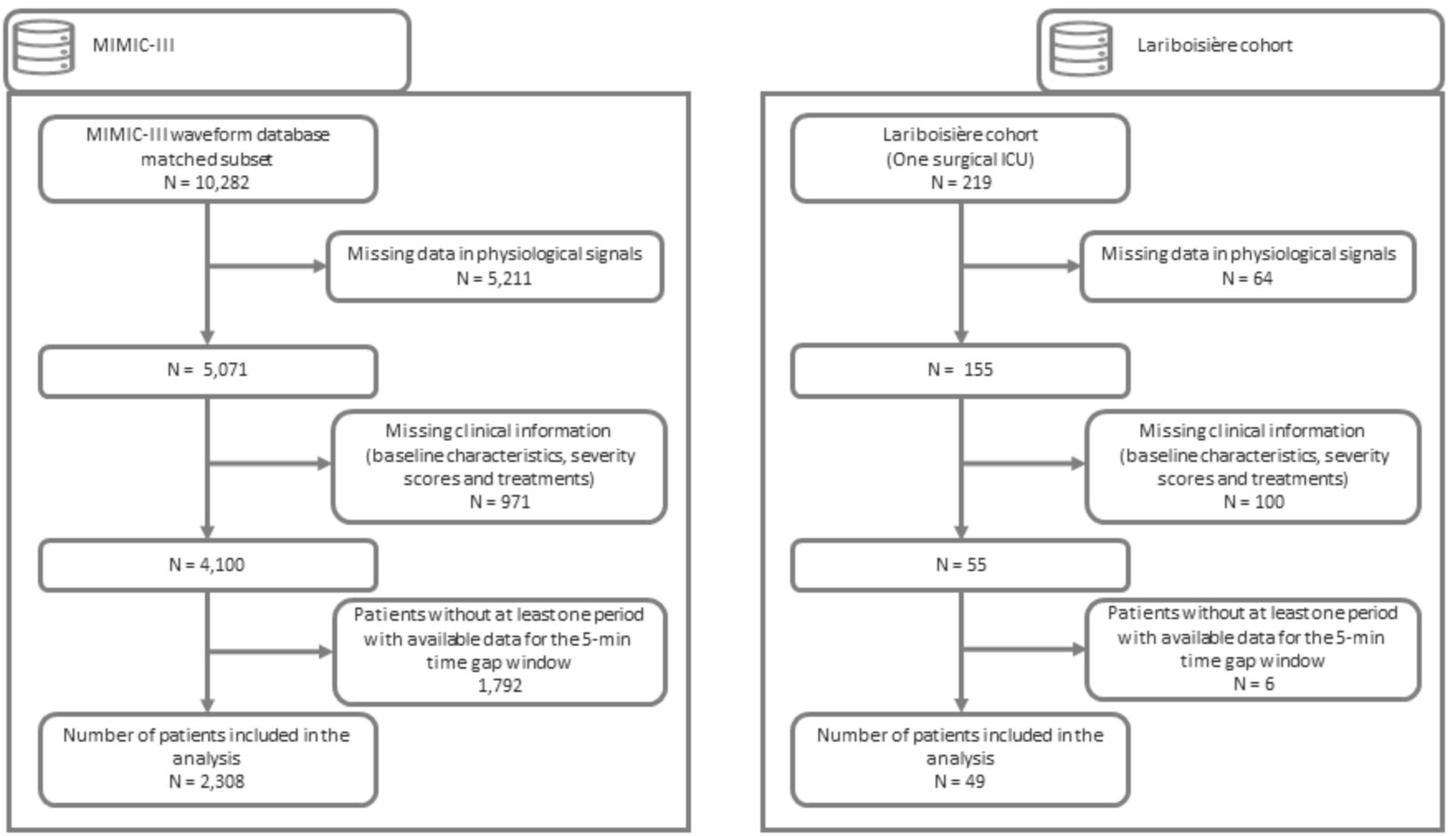
2 – Gap window

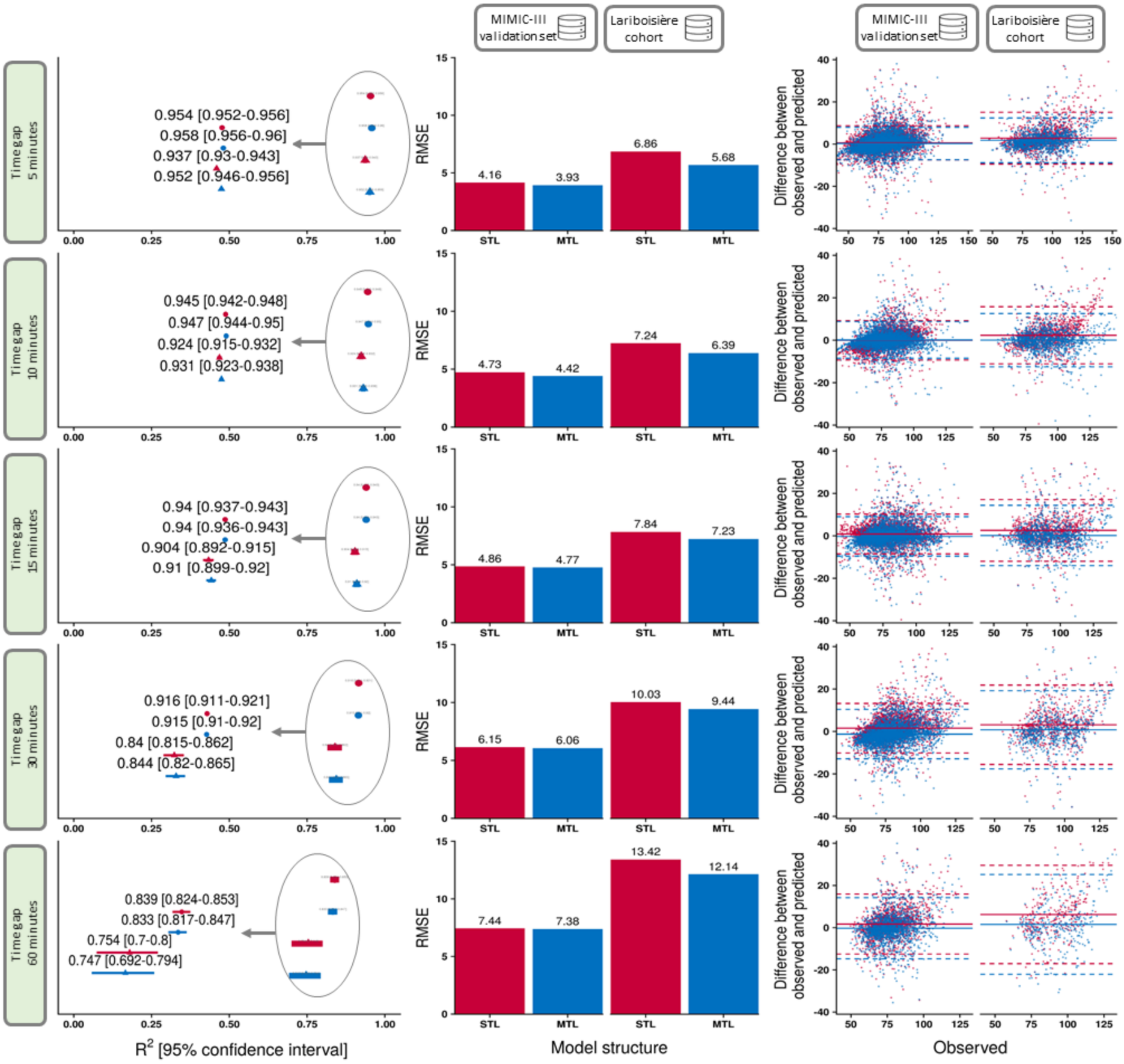
- Gap allows therapeutic adjustment before the prediction period
- Different durations are tested: 5, 10, 15, 30 and 60 minutes

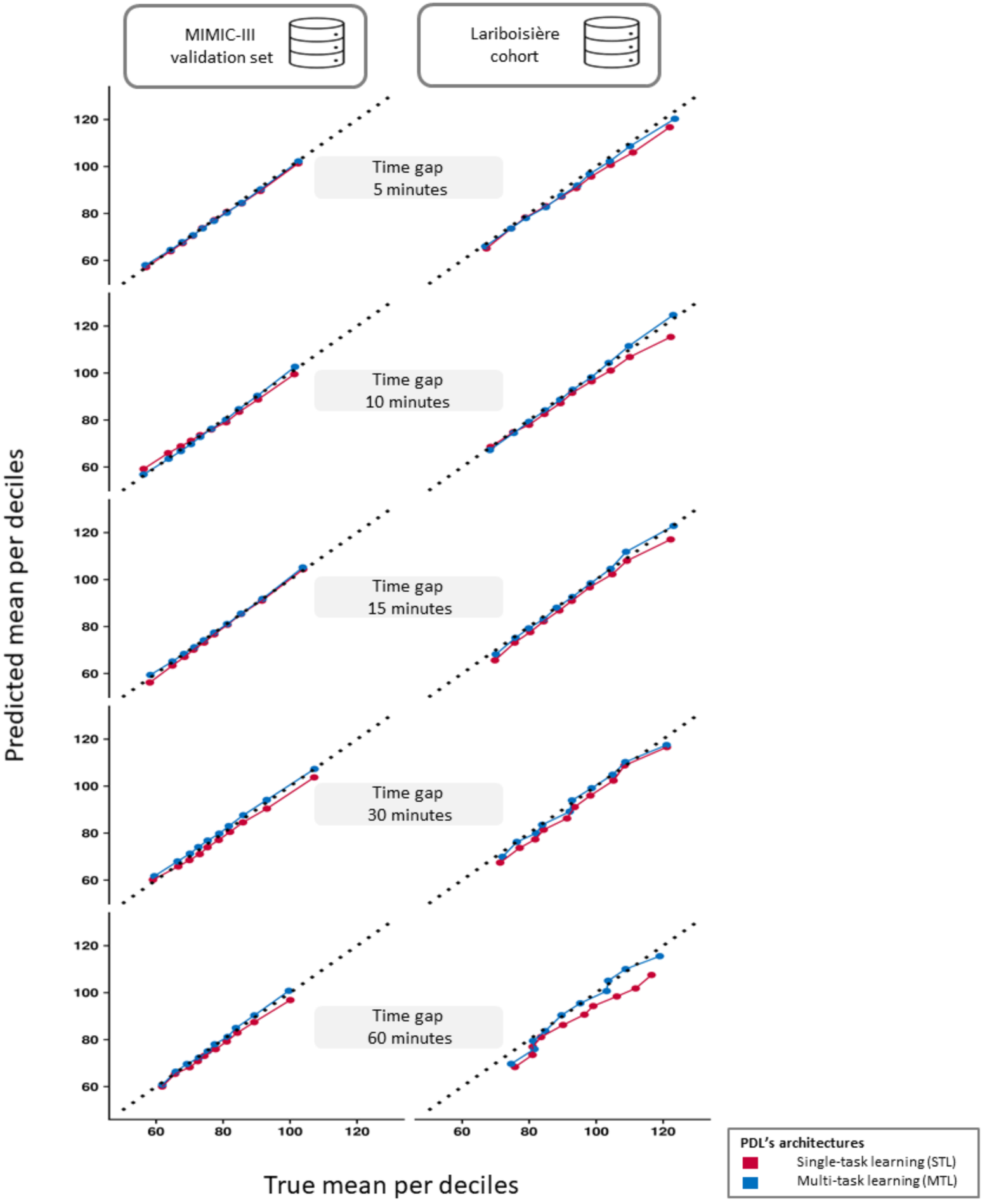
3 – Prediction window

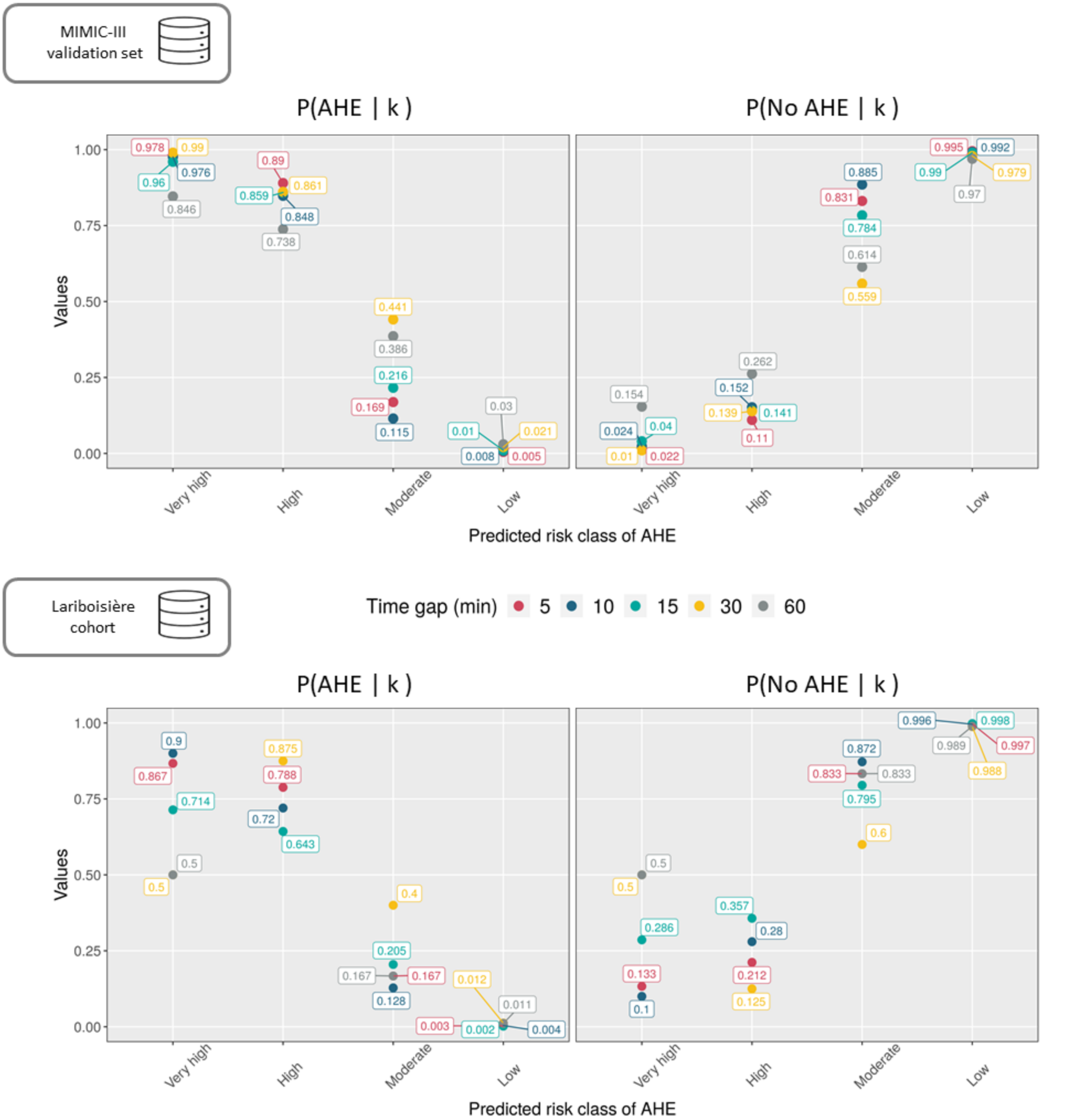
- Predictions on mean arterial pressure and heart rate, based on the data recorded during the observation period
- Gap are designed to allow therapeutic adjustment to avoid the prediction
- Duration: 5 minutes

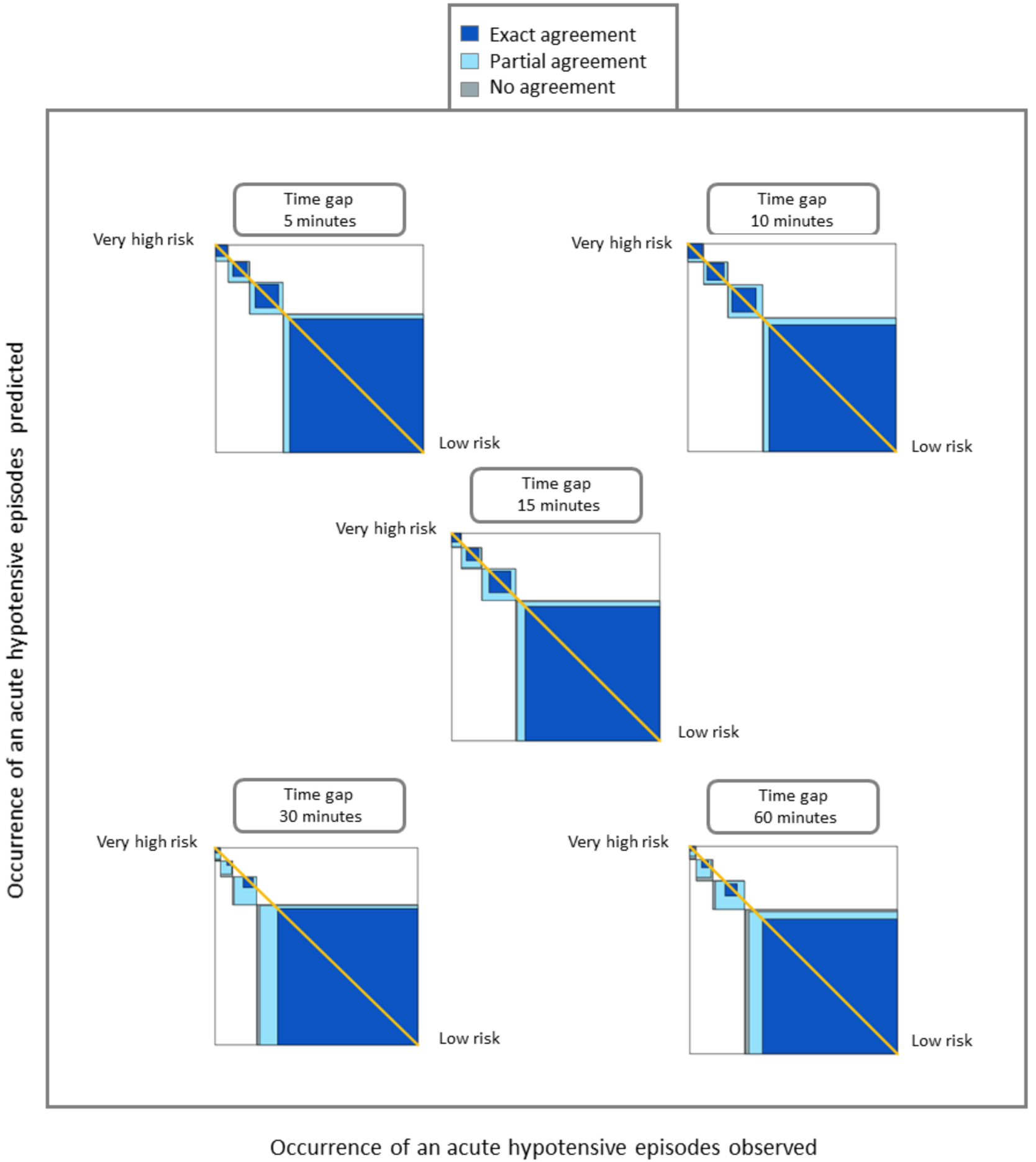

Inputs
Specific processing layer
Shared Layer
Specific learning layer
Output layer
Single-task learning
Physiological signals (heart rate, pulse oximetry, arterial pressures)
GRU
Patient id
Embeddings
ICU
Embeddings
Age, gender, SAPS-II, SOFA, treatments
Concatenation
Linear
ReLU
Batchnorm
Linear
Task : Mean arterial pressure or Heart rate
Multi-task learning
Physiological signals (heart rate, pulse oximetry, arterial pressures)
GRU
Patient ID
Embeddings
ICU
Embeddings
Age, gender, SAPS-II, SOFA, treatments
Concatenation
Linear
ReLU
Batchnorm
Linear
Joint
Linear
Task 1 Heart rate
Task 2 Mean arterial pressure

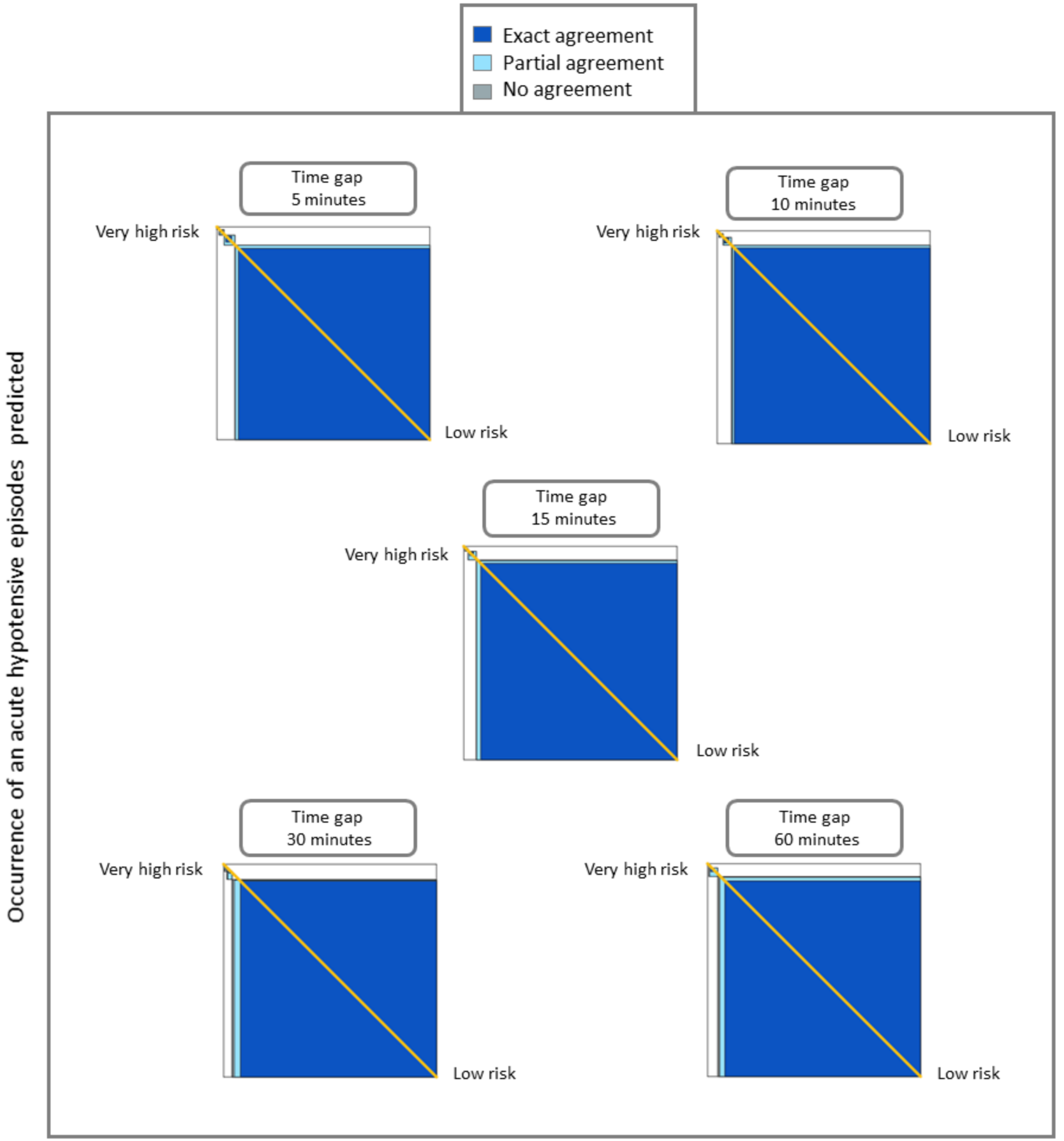

MIMIC-III
MIMIC-III waveform database matched subset N = 10,282
Missing data in physiological signals N = 5,211
N = 5,071
Missing clinical information (baseline characteristics, severity scores and treatments) N = 971
N = 4,100
Patients without at least one period with available data for the 5-min time gap window 1,792
Number of patients included in the analysis N = 2,308
Lariboisière cohort
Lariboisière cohort (One surgical ICU) N = 219
Missing data in physiological signals N = 64
N = 155
Missing clinical information (baseline characteristics, severity scores and treatments) N = 100
N = 55
Patients without at least one period with available data for the 5-min time gap window N = 6
Number of patients included in the analysis N = 49

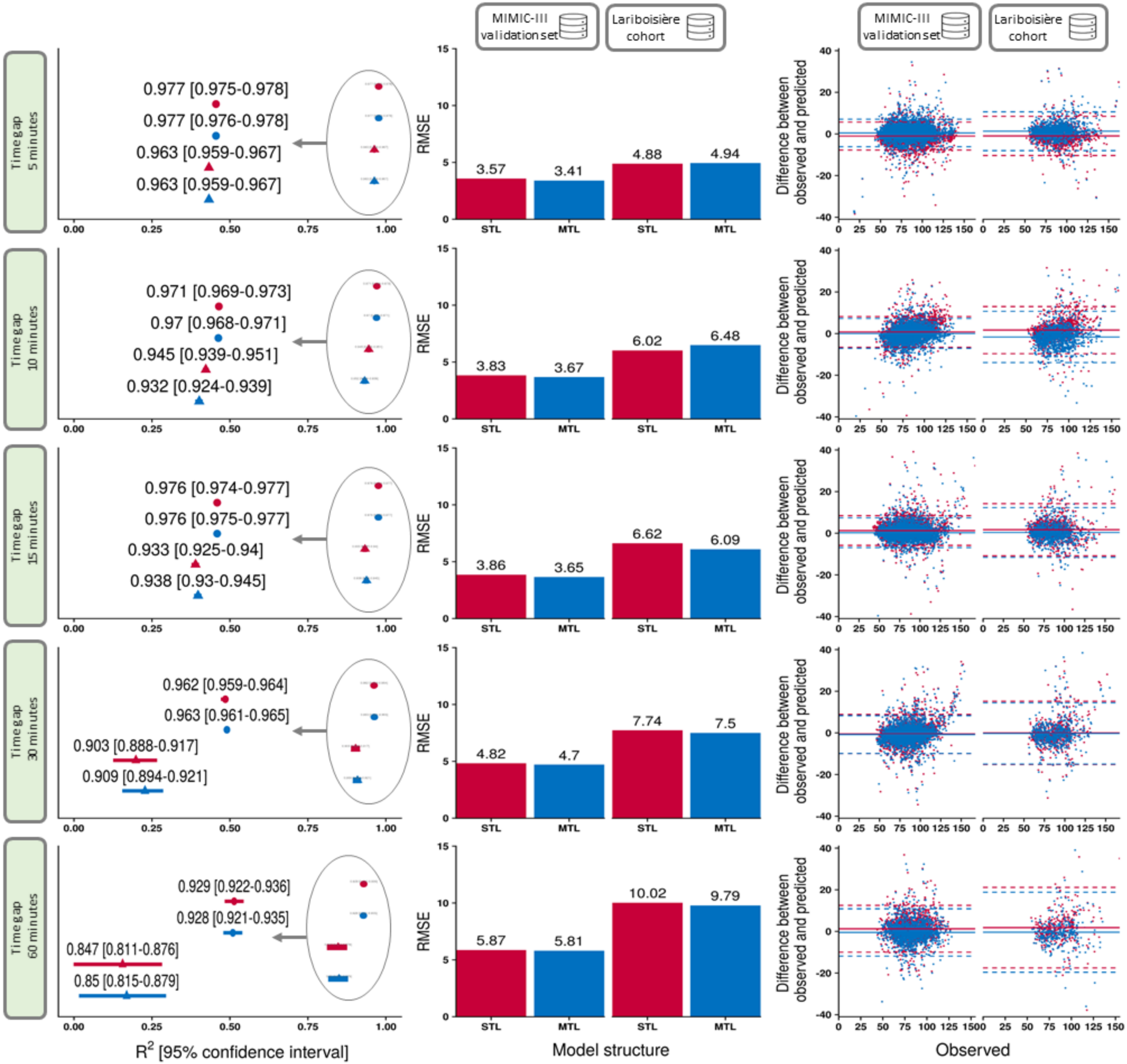

MIMIC-III validation set
Lariboisière cohort
MIMIC-III validation set
Lariboisière cohort
Time gap 5 minutes
Time gap 10 minutes
Time gap 15 minutes
Time gap 30 minutes
Time gap 60 minutes
0.954 [0.952-0.956]
0.958 [0.956-0.96]
0.937 [0.93-0.943]
0.952 [0.946-0.956]
0.945 [0.942-0.948]
0.947 [0.944-0.95]
0.924 [0.915-0.932]
0.931 [0.923-0.938]
0.94 [0.937-0.943]
0.94 [0.936-0.943]
0.904 [0.892-0.915]
0.91 [0.899-0.92]
0.916 [0.911-0.921]
0.915 [0.91-0.92]
0.84 [0.815-0.862]
0.844 [0.82-0.865]
0.839 [0.824-0.853]
0.833 [0.817-0.847]
0.754 [0.7-0.8]
0.747 [0.692-0.794]
0.00 0.25 0.50 0.75 1.00
RMSE
4.16 3.93 6.86 5.68
4.73 4.42 7.24 6.39
4.86 4.77 7.84 7.23
6.15 6.06 10.03 9.44
7.44 7.38 13.42 12.14
STL MTL STL MTL
Difference between observed and predicted
50 75 100 125 150
50 75 100 125
R² [95% confidence interval]
Model structure
Observed
PDL's architectures
Single-task learning (STL)
Multi-task learning (MTL)
Datasets
MIMIC-III validation set
Lariboisière cohort

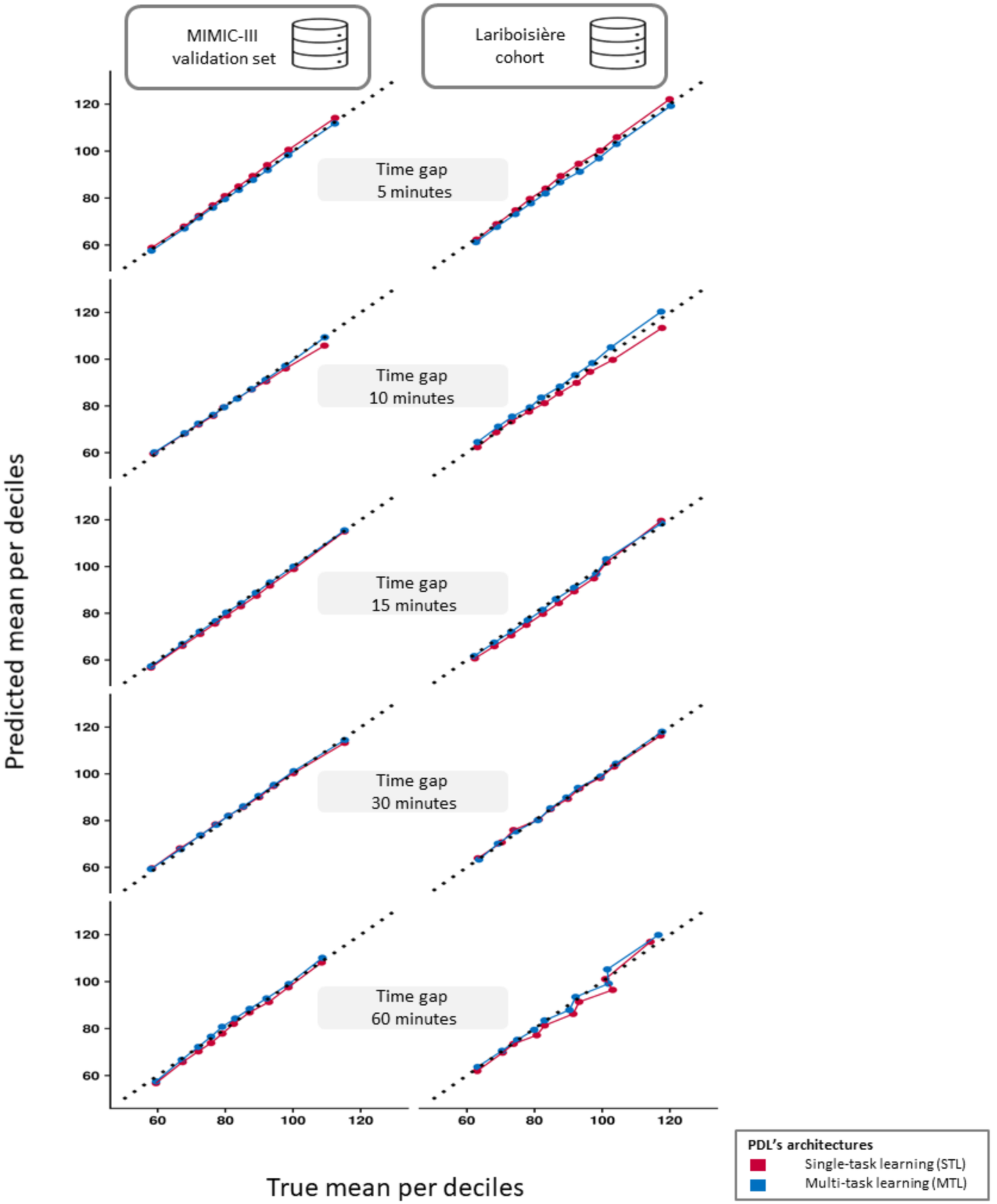

MIMIC-III validation set
Lariboisière cohort
Time gap 5 minutes
Time gap 10 minutes
Time gap 15 minutes
Time gap 30 minutes
Time gap 60 minutes
Predicted mean per deciles
True mean per deciles
60
80
100
120
PDL's architectures
Single-task learning (STL)
Multi-task learning (MTL)


MIMIC-III validation set
P(AHE | k)
P(No AHE | k)
Values
1.00
0.75
0.50
0.25
0.00
0.978
0.99
0.976
0.96
0.846
0.89
0.861
0.859
0.848
0.738
0.441
0.386
0.216
0.169
0.115
0.03
0.01
0.021
0.008
0.005
0.995
0.992
0.99
0.979
0.97
0.885
0.831
0.784
0.614
0.559
0.262
0.154
0.152
0.139
0.141
0.11
0.024
0.04
0.01
0.022
Very high
High
Moderate
Low
Predicted risk class of AHE
Lariboisière cohort
Time gap (min) 5 10 15 30 60
P(AHE | k)
P(No AHE | k)
Values
1.00
0.75
0.50
0.25
0.00
0.9
0.867
0.714
0.5
0.5
0.875
0.788
0.72
0.643
0.4
0.205
0.167
0.167
0.128
0.012
0.011
0.003
0.002
0.004
0.996
0.998
0.989
0.997
0.988
0.872
0.833
0.833
0.795
0.6
0.5
0.5
0.357
0.286
0.28
0.133
0.212
0.1
0.125
Very high
High
Moderate
Low
Predicted risk class of AHE


Exact agreement
Partial agreement
No agreement
Time gap
5 minutes
Very high risk
Low risk
Time gap
10 minutes
Very high risk
Low risk
Time gap
15 minutes
Very high risk
Low risk
Time gap
30 minutes
Very high risk
Low risk
Time gap
60 minutes
Very high risk
Low risk
Occurrence of an acute hypotensive episodes predicted
Occurrence of an acute hypotensive episodes observed


Exact agreement
Partial agreement
No agreement
Time gap
5 minutes
Very high risk
Low risk
Time gap
10 minutes
Very high risk
Low risk
Time gap
15 minutes
Very high risk
Low risk
Time gap
30 minutes
Very high risk
Low risk
Time gap
60 minutes
Very high risk
Low risk
Occurrence of an acute hypotensive episodes predicted
Occurrence of an acute hypotensive episodes observed

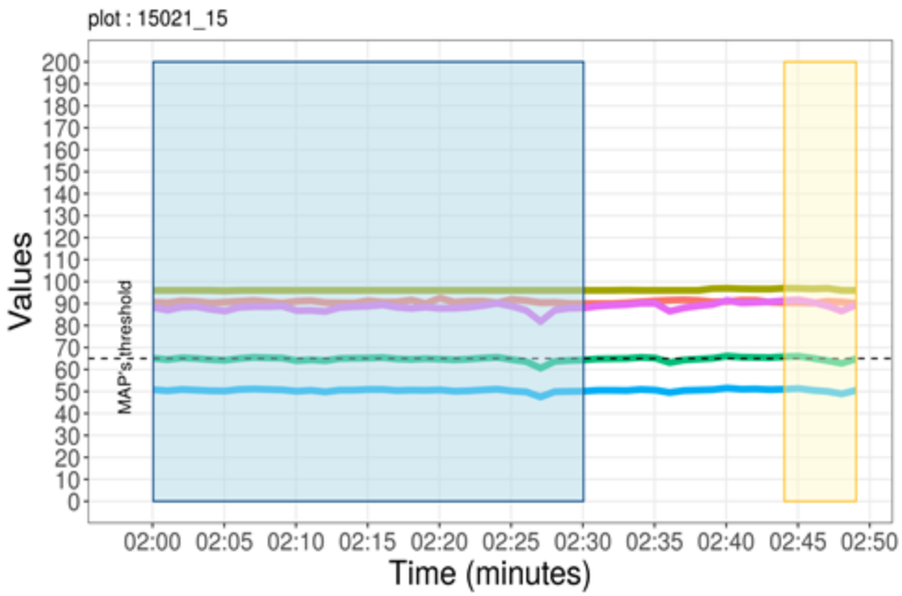

MIMIC-III validation set
Lariboisière cohort
Time gap 5 minutes
0.977 [0.975-0.978]
0.977 [0.976-0.978]
0.963 [0.959-0.967]
0.963 [0.959-0.967]
Time gap 10 minutes
0.971 [0.969-0.973]
0.97 [0.968-0.971]
0.945 [0.939-0.951]
0.932 [0.924-0.939]
Time gap 15 minutes
0.976 [0.974-0.977]
0.976 [0.975-0.977]
0.933 [0.925-0.94]
0.938 [0.93-0.945]
Time gap 30 minutes
0.962 [0.959-0.964]
0.963 [0.961-0.965]
0.903 [0.888-0.917]
0.909 [0.894-0.921]
Time gap 60 minutes
0.929 [0.922-0.936]
0.928 [0.921-0.935]
0.847 [0.811-0.876]
0.85 [0.815-0.879]
0.00 0.25 0.50 0.75 1.00
R2 [95% confidence interval]
RMSE
3.57 3.41 4.88 4.94
3.83 3.67 6.02 6.48
3.86 3.65 6.62 6.09
4.82 4.7 7.74 7.5
5.87 5.81 10.02 9.79
STL MTL STL MTL
Model structure
Difference between observed and predicted
0 25 50 75 100 125 150
Observed
PDL's architectures
Single-task learning (STL)
Multi-task learning (MTL)
Datasets
MIMIC-III validation set
Lariboisière cohort


MIMIC-III
validation set
Lariboisière
cohort
Time gap
5 minutes
Time gap
10 minutes
Time gap
15 minutes
Time gap
30 minutes
Time gap
60 minutes
Predicted mean per deciles
True mean per deciles
60
80
100
120
PDL's architectures
Single-task learning (STL)
Multi-task learning (MTL)

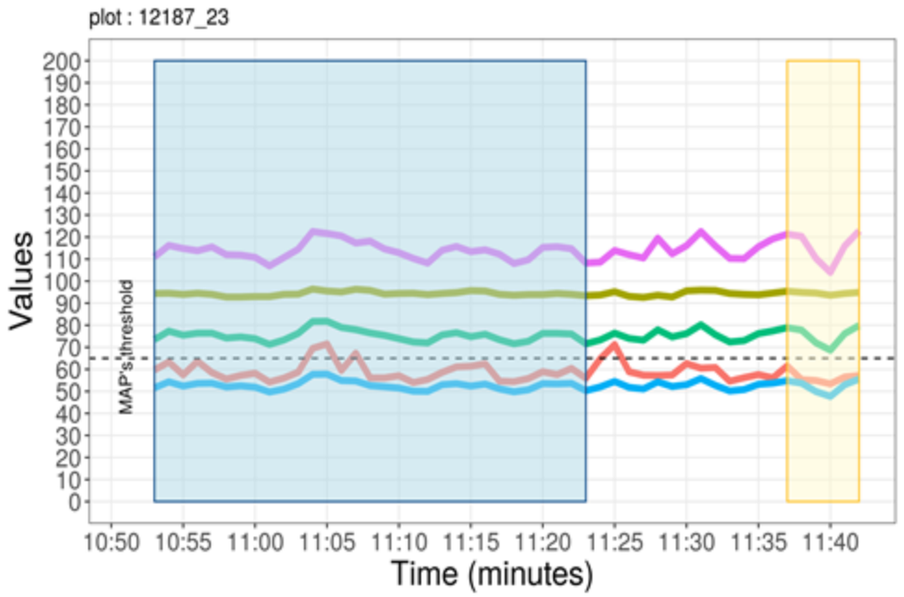
A

30-min observation window
5-min prediction window

C



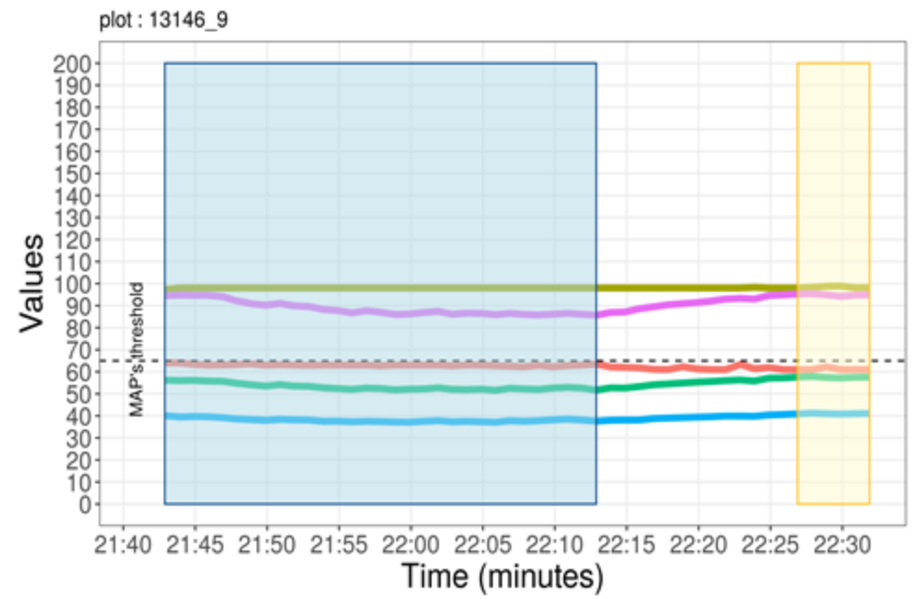
Average 5-min values in prediction window:
- Observed MAP: 64.54 - Predicted MAP: 64.56
- Observed HR: 90.48 - Predicted HR: 91.19



Average 5-min values in prediction window:
- Observed MAP: 74.92 - Predicted MAP: 74.58
- Observed HR: 75.72 - Predicted HR: 72.82

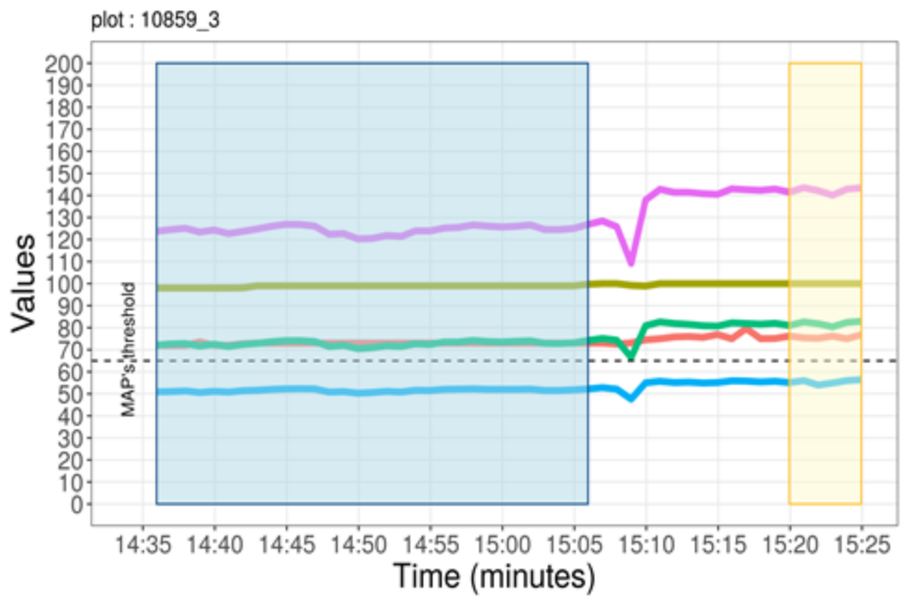
Physiological signals
- Heart rate
- Pulse oximetry
- Mean arterial pressure
- Diastolic arterial pressure
- Systolic arterial pressure

B



Average 5-min values in prediction window:
- Observed MAP: 57.52 - Predicted MAP: 57.85
- Observed HR: 61.24 - Predicted HR: 65.34

D



Average 5-min values in prediction window:
- Observed MAP: 82.02 - Predicted MAP: 73.71
- Observed HR: 75.72 - Predicted HR: 72.82