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

Ményssa Cherifa[1], Yannet Interian[2], Alice Blet[3,4], Matthieu Resche-Rigon[1], and Romain Pirracchio [*1,5]

[1] *ECSTRRA, Center of research in epidemiology and statistics, UMR INSERM 1153, University of Paris, Paris, France*

[2] *M.S. in Data Science Program, University of San Francisco, San Francisco, CA, United States of America*

[3] *Department of Anesthesiology, Critical Care and Burn Center, Lariboisière - Saint-Louis Hospitals, DMU Parabol, AP–HP Nord, University of Paris, France; Inserm UMR-S 942,Cardiovascular Markers in Stress Conditions (MASCOT), University of Paris, Paris, France*

[4] *University of Ottawa Heart Institute and University of Ottawa, Ottawa, Ontario, Canada*

[5] *Department of Anesthesia and Perioperative Medicine, Zuckerberg San Francisco General Hospital and Trauma Center, University of California San Francisco, San Francisco, CA, United States of America*

Abstract

In the intensive care unit (ICU), clinicians are trained to analyze simultaneously multiple physiological signals in order to understand and predict the evolution of a critical condition such as hemodynamic instability. We developed the Multi-task Learning Physiological Deep Learner (MTL-PDL), a deep learning algorithm that learns simultaneously from the mean arterial pressure (MAP) and of the heart rate (HR). Models were trained on 2,308 patients from the ICU medical information mart for intensive care version 3 (MIMIC-III) data set, and then externally validated on the surgical ICU of Lariboisière hospital (AP-HP, Paris, France).

In the external validation, our model exhibited excellent calibration as illustrated by a R-squared (R^2) of 0.747 (95% confidence interval, 0.692 to 0.794) and an 0.850 (95% confidence interval, 0.815 to 0.879) for the prediction of the MAP and HR respectively, up to 60-minutes ahead of time. When AHE is defined as a MAP below 65 mmHg during 5 minutes, our MTL-PDL reached a positive predictive value of 90% (for patients at very high risk) and a negative predictive value of 99.8% (for patients at low risk).

Based on its excellent prediction performance, the Physiological Deep Learner may allow to accurately predict far in advance the evolution of key hemodynamic parameters in critically ill patients. Using this algorithm in clinical practice could allow the clinician to proactively adjust the treatment and potentially avoid hypotension episodes and end-organ hypoperfusion.

Introduction

Shock is the clinical presentation of an acute circulatory failure that results in inadequate cellular oxygen supply and utilization.[1] It is a common condition that affects approximately one-third of the patients in the intensive care unit (ICU).[2] It is clinically characterized by a rapid decline in the mean arterial pressure (MAP).[1] Shock is considered to be a diagnostic and therapeutic emergency since any delay in treatment initiation may result in increased mortality.[3,4] Therefore, early identification of patients at risk of shock is of utmost importance.

Acute hypotension, defined as a rapid decline in mean arterial pressure (MAP), 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).[5–16] While most predictive models use only the MAP signal as input to predict the risk of hypotension, in practice, clinicians are trained to analyze simultaneously multiple physiological signals as well as additional information on patient overall condition and treatment[17] to 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, treatments received in the ICU, and several continuous physiological

*Corresponding author : Romain.Pirracchio@ucsf.edu

signals (including heart rate, blood pressure and pulse oximetry) 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 signals by using a multi-task learning (MTL) approach to improve the prediction of AHE in critically ill patients.

The goal of MTL is to join the simultaneous learning of 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 already been used to predict a variety of clinical outcomes, such as mortality,[20] hospital length of stay, time to the next medical visit[21] or the international classification of disease groupings.[22] Because of the well-describe interdependence between arterial blood pressure and heart rate,[23] we hypothesized that using MTL to lean jointly from the MAP and the HR would improve the performance of AHE prediction.

In the recent literature, AHE was defined as a MAP below 65 mmHg for at least 1 minute.[13,24,25] However, as previously highlighted by Chan et al.[26] this conventional definition of AHE based on a single cutoff value may not be suitable for individual patients. Normal blood pressure varies between individuals and patients may tolerate hypotension to various degrees before developing end-organ damages. To offer the possibility for the clinician to pick different blood pressure targets for different subgroups of patients, we wanted to develop an algorithm that wouldn't predict AHEs as defined by a specific threshold but would rather enable the prediction of the actual blood pressure value.

In this study, we are proposing to use a MTL approach to learn from patient individual physiology and improve the prediction of the MAP up to 60-min ahead of time. To validate our approach, we compared two different architectures for our Physiological Deep Learner (PDL): STL-PDL model, trained to predict MAP and HR separately, and MTL-PDL model 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 the surgical ICU of Lariboisière hospital (AP-HP, Paris, France) was used for external validation.

Results

Data Preparation Each patient ICU stay was decomposed in successive periods, as depicted in Fig. 1A. For a given period t, our objective was to predict the average MAP and the average heart rate (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) and sequential organ failure (SOFA) score, type of ICU,i.e, medical, surgical, cardiac, mixed), time-evolving treatment characteristics (including mechanical ventilation, vasopressors, and sedation medication) 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 from the MIMIC-III and Lariboisière cohort with no missing data in baseline characteristics, time-evolving characteristics, and physiological signals were selected. All patients with at least one complete set 3 successive windows (i.e observation, gap, prediction) with a time gap of at least 5 min were included in the analysis (Fig. 1B). From MIMIC-III, 2,308 patients (74,159 periods) qualified to be included in the analysis. Among them, 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. From the Lariboisière cohort (external validation cohort), 49 patients were included in the 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.

Model development and performance. MIMIC-III data was used to develop the PDL models. All models were evaluated on independent test sets from both the internal and external data sources. STL- and MTL-PDL had a similar architecture up to the last layer. We included patient identifier (id) as a variable to our models to associate each period to a specific patient and added gated recurrent units to the PDL architecture to effectively account for long-term dependency in the physiological time-series.

This baseline PDL framework was augmented with a specific learning architecture trained to predict the average 5-min MAP and HR either separately (STL) or jointly (MTL) (Fig. 2). MIMIC-III data was split as follows: 80% of the patients in MIMIC-III were used to train the algorithms, 10% to optimize the models, i.e., perform hyper-parameter search and the remaining 10% (MIMIC-III validation set) was used to evaluate the prediction performance. Model performance was also evaluated externally on the data from the Lariboisière cohort. The experimental workflow is detailed in Extended Data Fig. 1.

MAP prediction. Complete results on the performance of STL- and MTL-PDL for different time gaps (5, 10, 15, 30, and 60-min) are presented in Fig. 3. In MIMIC-III validation set as well as in the external cohort, when the task was to predict the MAP value averaged over 5 minutes, the correlation coefficient R^2 was very close to 1 whatever the PDL architecture and the time gap. However, when compared to MTL-PDL, STL-PDL was consistently associated with a larger root mean square error (RMSE) (middle panel,Fig. 3). 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. A similar pattern was observed with 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 5.68, 6.39, 7.23, 9.44 and 12.14 for MTL-PDL. On the right panel, Fig. 3 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, for the 5, 10, 15, 30, and 60-min time gaps, the average difference between observed and predicted MAP for STL-PDL 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] respectively, while for MTL-PDL 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]. 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. 4).

Prediction of acute hypotensive episodes. We also evaluated MTL-PDL performance to predict AHE defined as MAP (averaged over 5 consecutive minutes) below 65 mmHg. Based on the actual MAP prediction we defined 4 risk classes: "Very high risk" if the predicted predicted MAP (averaged over 5 min) is ≤ 60, "High risk" for $60 <$ predicted MAP ≤ 65, "Moderate risk" for $65 <$ predicted MAP ≤ 70 and "Low risk" for a predicted MAP > 70. Note that MTL-PDL was not trained to predicts these classes, but instead, we directly applied these definitions to the observed and predicted MAP values. Prediction performance (Fig. 5) was obtained by calculating for each predicted class, the positive predictive value = P(AHE | k) and the negative predictive value = P(No AHE | k), where k is the predicted risk class ("Very high", "High", "Moderate" or "Low"). In both validation sets, the higher the predicted risk, the higher the probability that of observing a MAP below 65 mmHg. The positive predictive value was 99% in MIMIC-III and 90% in the external validation cohort for the "Very high risk" class. The negative predictive values was 99.5% in MIMIC-III and 99.8% in in the external validation cohort for the "Low risk" class.

As expected in the External Data Fig. 2 the agreement decreases as the time gap increases. However, it seems that the misclassification is always in favor of the risk class of AHE closest to the current class.

HR prediction. Similar results are provided in External Data Fig. 4 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. External Data Fig. 5 shows excellent and better calibration profile with MTL-PDL as compared to STL-PDL.

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 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 exhibited high level of positive and negative predictive values for the prediction of acute hypotensive episodes.

Several AHE prediction models were developed over the past 20 years. In 2009, the 10th annual PhysioNet/Computers in cardiology challenge[27] 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.[5–12] 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 subgroup. Thus far, most of the models used as only input variable previous MAP values, ignoring other patient characteristics and/or time-dependent variables, e.g., heart rate, known to be highly correlated with the arterial blood pressure. Our group[28] 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.[29] used gradient-boosted ensemble tree classifiers trained on 209 variables to predict with 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 the prediction of 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. Generally, MTL is used for two tasks formulations,i) prediction of separate outcomes and ii) identification of 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.[30] First, MTL works even if one of the outcomes is missing. We can still train the model to do both tasks at the same time. 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.[30] 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.[31] 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 since blood pressure varies between individuals as do their organ capacity to tolerate hypotension.[26] Accordingly, Futier et al.[32] showed among patients undergoing abdominal surgery, that targeting individualized systolic blood pressure goals reduced the risk of postoperative organ dysfunction. Chan et al.[26] 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 a example, Extended Data Fig. 6 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 already using in his/her clinical reasoning, 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 methods used to summarize the information from physiological time series.[28] A strength of the present study is that we used gated recurrent unit cells,[33] which are able to effectively retain long-term dependencies in time series. According to Le Cun et al.,[34] 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.[35] In this study, we only included patients from 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 manage 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.

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.

Methods

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 has comprises de-identified data from patient 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).[36] 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).[37]

The Lariboisière cohort is a database including clinical data, physiologic signals, treatment administration and administrative data prospectively and consecutively collected at the bedside over a two-year period (2017-2018) from the surgical ICU of Lariboisière hospital(AP-HP, Paris, France). This cohort was notably built to be consistent with MIMIC-III in order to be able to validate or replicate works based on it. This study was approved by the IRB of the « Société de Réanimation de Langue Française » (CE-SRLF 14-356), that exempted signed informed consent. Every patient was orally informed for its inclusion in this study.

Periods definition

From the admission to the discharge, each patient ICU stay was divided into periods. Based on the Leisman et al. recommendations for development and reporting of prediction models in critical care,[38] each period was divided into 3 successive windows: a *30-min observation* window, a varying *time gap* window and a *5-min prediction* window. To predict the averages 5-min MAP and 5-min HR in all *5-min prediction* windows, only data recorded during the *30-min observation* were exploited. Three different *time gap* window were considered: 5, 10, 15, 30, and 60-min between the observation and the prediction. The aim of such varying *time gap* was to emulate a framework in which clinicians will be able to introduce or perform some therapeutic adjustment. This setup is illustrated in Fig. 1A.

All patients from the MIMIC-III waveform database matched subset and Lariboisière cohort with no missing data in physiological signals were selected. Then, periods were defined for each patient. Secondarily, 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 (Fig. 1B).

Data partitioning

Following recommendations from Chen and al.,[39] 90% of the analyzed patients from MIMIC-III were randomly assigned to the "development set": 80% of the patients into the training set (used to estimate model parameters), and 10% into the tuning set (used to select model hyperparameters). The remaining 10% of the patients were assigned to the validation set for final evaluation. The latter was referred as the MIMIC-III validation set. Data from the Lariboisière cohort were exclusively used for external.

Sampling periods and predicting on new patients

From the admission to the discharge, each patient ICU stay was divided into successive periods. As each patient has distinct length of stay in the ICU, our data were uneven in terms of the number of periods per patient. While it is suitable to maximize the information from the training data, two challenges rose up:

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

ii. Our model should be aware of the correlations between periods from the same patient

To address (i), at the beginning of each iteration of the learning process (epoch) a new balance dataset is obtained by sampling the same number of patient's periods with replacement therefore each patient

has the same number of periods. The median number of periods per patient included in the study is used to determined the number of periods to be drawn per patient. Models are trained with balance datasets for multiple epochs and each epoch used different sample of the data; consequently all the data are used during the models' training. Challenge (ii) was addressed by adding patient *id* as a variable to our models. Our models train an embedding layer[40] to learn patient *id* representations. Each patient is represented by a n-vector of reals. These vectors are learned by the models together with other parameters during training. Patients with similar characteristics will get similar vectors. This approach is analogous to the use of vectors as to representation words in modern natural language processing.[41] By adding the patient *id* to predictors and training a vector for each *id*, we fully handle the correlation between periods of a same patient. Nevertheless, in validation and testing sets, patients are different than those of the training set. Thus, we are unable to build vector representations for theses new patient *ids*. To be able to predict on new patients we use the following approach. At each training epoch we overwrite 10% of the periods at random to have *id* 0. By doing that, the vector associated with the *id* 0 is trained to be the "average user". Thus at model's validation and testing time we can use this average user for prediction of a new patient.

Predictors and outcomes

Fixed predictors included baseline characteristics:

- Quantitative characteristics: age; initial severity scores: simplified acute physiology score-II (SAPS-II)[42] and sequential organ failure score (SOFA)[43]
- Categorical characteristics: gender; patient id; type of ICU

Two types of time-dependent characteristics were considered:

- Period-evolving treatment characteristics: status of mechanical ventilation; concomitant administration of vasopressors; 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 binaries by x_{cat}, gender and period-evolving treatment binary characteristic (collected every period) by x_{binary} and physiological signals (collected every min during the *30-min observation window*) by x_{series}. Thus, we get 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} represent the averages 5-min MAP and 5-min HR calculated across the *5-min prediction window* respectively for each patient period.

Model architecture and optimization

Deep learning models handle complex relationships between a large number of explanatory predictors and desired outputs, such as patient outcomes.[18] Based on a succession of layers (each layer receives its inputs from the previous one's outputs), these models use backpropagation algorithm to update their internal parameters and optimize their predictions. The Physiological Deep Learner (PDL) was designed to predicts the averages 5-min MAP and/or 5-min HR by mapping x to y_{MAP} and/or y_{HR}. Different PDL for the outcomes prediction were implemented and compared, two single-task learning PDL (STL-PDL) to separately predict the averages 5-min MAP and the 5-min HR respectively, and one Multi-task learning PDL (MTL-PDL) trained to jointly predict MAP and HR.

To make the comparison between STL-PDL and MTL-PDL as fair as possible all estimation processes are identical except the last step (Fig. 2). In the specific data processing layer, physiological signals were fed into the gated recurrent unit (GRU) cells. GRU[33] is a type of recurrent neural network, able to effectively retain long-term dependencies in sequential data. It encodes general information about the whole temporal patient ICU stay by considering longitudinal changes in the patient physiological signals. Also, multi-layer gates enable for regulating the information to be kept or discarded. The GRU's input corresponds to x_{series}. The output vector of the GRU can be interpreted as a unique representation of the patient's physiological signals trajectory. Each categorical variable in x_{cat}, were mapped into two separate embedding layers. Embedding layers were randomly initialized and learned by the model during the optimization process. Note that with the first embedding, we learned the vector representation of each patient. Processed data from the previous operations were concatenated together with x_{binary} and

x_{cont} variables. The concatenated vector was included into a linear layer followed by a rectified linear units (ReLU) layer[44] and a batch normalization (batchnorm) layer.[45] The outputs from these layers were either fed separately into two linear layers to independently predict MAP and HR or to one linear layer to jointly predict MAP or HR.

Given some training data $D = \{\{x^{(ip)}, y^{(ip)}_{MAP}, y^{(ij)}_{HR}\}^{t}_{p=1}\}^{n}_{i=1}$, t, the number of periods and n the number of patients, STL-PDL would independently optimizes the mean square error functions (MSE) $\frac{1}{N}\sum_{i=1}^{N}(\hat{y}^{(i)}_{MAP} - y^{(i)}_{MAP})^2$ and $\frac{1}{N}\sum_{i=1}^{N}(\hat{y}^{(i)}_{HR} - y^{(i)}_{HR})^2$ where $N = t \times n$, $\hat{y}^{(i)}_{.}$ and $y^{(i)}_{.}$ correspond to the total number of observations, the predicted and the observed outcomes respectively. The MTL-PDL jointly optimizes the MSE function $\frac{1}{N}\left\{\sum_{i=1}^{N}(\hat{y}^{(i)}_{MAP} - y^{(i)}_{MAP})^2 + (\hat{y}^{(i)}_{HR} - y^{(i)}_{HR})^2\right\}$. PDLs development was performed using *PyTorch* version 1.4 library.[46] We used *Adam*[47] for optimization, because of its computational efficiency, its low memory requirements, invariance to diagonal re-scaling of the gradients. (See also Supplementary Material).

Assessment of performance

Evaluation of the averages prediction of MAP and HR

Graphical representations of the final results were performed using R version 3.6.3 library[48] and were reported according the type of PDL's architecture and time gaps. To assess how well the models predict on both validation sets, R-squared (R^2) together with its 95% confidence interval (95%CI) and root mean square error (RMSE), defined as $\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}^{(i)} - y^{(i)})^2}$ 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 between observed and predicted outcomes 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 were computed. Similarly, the predicted mean per decile defined as the average 5-min MAP or 5-min HR values were also computed. Then, each couple of mean was plotted according to time gaps for both validation sets. Thus, the closer line was to the diagonal, better the calibration was.

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.[13,28,29] MTL-PDL was not trained to predict this binary outcome, nevertheless we applied the following classification as alert tool. We classified classes patients according to their AHE status using following rules:

i. "AHE", if observed average 5-min MAP $\leq$ 65

ii. "No AHE", either

We defined classes of predicted risk of AHE according to the following rules:

i. "Very high", if predicted average 5-min MAP $\leq$ 60

ii. "High", if 60 $<$ predicted average 5-min MAP $\leq$ 65

iii. "Moderate", if 65 $<$ predicted average 5-min MAP $\leq$ 70

iv. "Low", if predicted average 5-min MAP $>$ 70

We examined the performance of the MTL-PDL by crossing the observed classes to the predicted classes. Relying on the crossing matrices, we calculated the following metrics for each predicted classes:

- P(AHE | k)
- P(No AHE | k)

where k corresponds to either "Very high", "High", "Moderate" or "Low" predicted classes. Moreover, we displayed Bangdiwala's agreement chart for both validation sets and each time gap. This chart assesses the concordance between two methods of measurement of ordinal categorical data.[49] Thus, it gives an overview of misclassification between observed classes and predicted classes. Here we applied

the rules of the predicted classes to the observed ones. For each class, 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 the farthest class is 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.

References

[1] Jean-Louis Vincent and Daniel De Backer. Circulatory Shock. *New England Journal of Medicine*, 369:1726–1734, 2013.

[2] Yasser Sakr, Konrad Reinhart, Jean Louis Vincent, Charles L. Sprung, Rui Moreno, V. Marco Ranieri, Daniel De Backer, and Didier Payen. Does dopamine administration in shock influence outcome? Results of the Sepsis Occurrence in Acutely Ill Patients (SOAP) Study. *Critical Care Medicine*, 34:589–597, 2006.

[3] Robert F. Wilson, Jacqueline A. Wilson, Dennis Gibson, and William J. Sibbald. Shock in the emergency department. *Journal of the American College of Emergency Physicians*, 5(9):678–690, 1976.

[4] Mitchell M. Levy, Laura E. Evans, and Andrew Rhodes. The Surviving Sepsis Campaign Bundle: 2018 update, jun 2018.

[5] P. Langley, S.T. King, D. Zheng, E.J. Bowers, K. Wang, J. Allen, and A. Murray. Predicting acute hypotensive episodes from mean arterial pressure. In *Computers in Cardiology, 2009*, volume 36, pages 553–556, 2009.

[6] K. Jin and N. Stockbridge. Smoothing and discriminating MAP data. In *Computers in Cardiology, 2009*, volume 36, pages 633–636, 2009.

[7] F. Jousset, M. Lemay, and J.M. Vesin. Computers in Cardiology / Physionet Challenge 2009: Predicting acute hypotensive episodes. In *Computers in Cardiology, 2009*, volume 36, pages 637–640, 2009.

[8] Thomas Ho and X. Chen. Utilizing histogram to identify patients using pressors for acute hypotension. In *Computers in Cardiology, 2009*, volume 36, pages 797–800, 2009.

[9] J H Henriques and T R Rocha. Prediction of Acute Hypotensive Episodes Using Neural Network Multi-models. In *Computers in Cardiology, 2009*, volume 36, pages 549–552, 2009.

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

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

[12] P.a. Fournier and J.F. Roy. Acute hypotension episode prediction using information divergence for feature selection, and non-parametric methods for classification. In *Computers in Cardiology, 2009*, volume 36, pages 625–628, 2009.

[13] Feras Hatib, Zhongping Jian, Sai Buddi, Christine Lee, Jos Settels, Karen Sibert, Joseph Rinehart, and Maxime Cannesson. Machine-learning Algorithm to Predict Hypotension Based on High-fidelity Arterial Pressure Waveform Analysis. *Anesthesiology*, 129:663–674, 2018.

[14] Samir Kendale, Prathamesh Kulkarni, Andrew D. Rosenberg, and Jing Wang. Supervised Machine Learning Predictive Analytics for Prediction of Postinduction Hypotension. *Anesthesiology*, 129:675–688, 2018.

[15] Rob Donald, Tim Howells, Ian Piper, I Chambers, G Citerio, P Enblad, B Gregson, K Kiening, J Mattern, and P Nilsson. Early warning of EUSIG-defined hypotensive events using a Bayesian artificial neural network. In *Intracranail Pressure and Brain Monitoring XIV*, volume 114, pages 39–44, 2012.

[16] Sakyajit Bhattacharya, Vijay Huddar, Vaibhav Rajan, and Chandan K Reddy. A dual boundary classifier for predicting acute hypotensive episodes in critical care. *PloS one*, 13(2):e0193259, 2018.

[17] Douglas P. Barnaby, Shannon M. Fernando, Kevin J. Ferrick, Christophe L. Herry, Andrew J.E. Seely, Polly E. Bijur, and 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):96–102, feb 2018.

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

[19] Richard A. Caruana. Multitask Learning: A Knowledge-Based Source of Inductive Bias. In *Machine Learning Proceedings 1993*, pages 41–48. Elsevier, 1993.

[20] Yuqi Si and Kirk 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*, 2019:779–788, 2019.

[21] Edward Choi, Mohammad Taha Bahadori, and Jimeng Sun. Doctor AI: predicting clinical events via recurrent neural networks. *CoRR*, abs/1511.05942, 2015.

[22] Hrayr Harutyunyan, Hrant Khachatrian, David C. Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific data*, 6(1):96, jun 2019.

[23] Mina Chookhachizadeh Moghadam, Ehsan Masoumi Khalil Abad, Nader Bagherzadeh, Davinder Ramsingh, Guann Pyng Li, and Zeev N. Kain. A machine-learning approach to predicting hypotensive events in ICU settings. *Computers in Biology and Medicine*, 118:103626, mar 2020.

[24] Kamal Maheshwari, Tetsuya Shimada, Jonathan Fang, Ilker Ince, Edward J. Mascha, Alparslan Turan, Andrea Kurz, and Daniel 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):255, may 2019.

[25] Simon James Davies, Simon Tilma Vistisen, Zhongping Jian, Feras Hatib, and Thomas 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):352–359, feb 2020.

[26] Brandon Chan, Brian Chen, Alireza Sedghi, Philip Laird, David Maslove, and Parvin Mousavi. Generalizable deep temporal models for predicting episodes of sudden hypotension in critically ill patients: a personalized approach. *Scientific Reports*, 2020.

[27] G.B. Moody and L.H. Lehman. Predicting acute hypotensive episodes: The 10th annual physionet/computers in cardiology challenge. In *Computers in Cardiology, 2009*, volume 36(5445351), pages 541–544, 2009.

[28] Ményssa Cherifa, Alice Blet, Antoine Chambaz, Etienne Gayat, Matthieu Resche-Rigon, and Romain Pirracchio. Prediction of an acute hypotensive episode during an ICU hospitalization with a super learner machine-learning algorithm. *Anesthesia and Analgesia*, 130(5):1157–1166, 2020.

[29] Stephanie L. Hyland, Martin Faltys, Matthias Hüser, Xinrui Lyu, Thomas Gumbsch, Cristóbal Esteban, Christian Bock, Max Horn, Michael Moor, Bastian Rieck, Marc Zimmermann, Dean Bodenham, Karsten Borgwardt, Gunnar Rätsch, and Tobias M. Merz. Early prediction of circulatory failure in the intensive care unit using machine learning. *Nature Medicine*, 26(3):364–373, mar 2020.

[30] Rich Caruana. Multitask Learning. In *Learning to Learn*, pages 95–133. Springer US, 1998.

[31] Sebastian Ruder. An overview of multi-task learning in deep neural networks. *CoRR*, abs/1706.05098, 2017.

[32] Emmanuel Futier, Jean Yves Lefrant, Pierre Gregoire Guinot, Thomas Godet, Emmanuel Lorne, Philippe Cuvillon, Sebastien Bertran, Marc Leone, Bruno Pastene, Vincent Piriou, Serge Molliex, Jacques Albanese, Jean Michel Julia, Benoit Tavernier, Etienne Imhoff, Jean Etienne Bazin, Jean Michel Constantin, Bruno Pereira, and Samir Jaber. Effect of individualized vs standard blood pressure management strategies on postoperative organ dysfunction among high-risk patients undergoing major surgery: A randomized clinical trial. *JAMA - Journal of the American Medical Association*, 2017.

[33] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *CoRR*, abs/1406.1078, 2014.

[34] Yann Lecun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[35] Alexander Meyer, Dina Zverinski, Boris Pfahringer, Jörg Kempfert, Titus Kuehne, Simon H. Sündermann, Christof Stamm, Thomas Hofmann, Volkmar Falk, and Carsten Eickhoff. Machine learning for real-time prediction of complications in critical care: a retrospective study. *The Lancet Respiratory Medicine*, 6(12):905–914, dec 2018.

[36] Alistair E.W. Johnson, Lu Lehman Pollard, Tom J.and Shen, Li wei H., Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3:160035, 2016.

[37] 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, and H. E. Stanley. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation*, 101(23), jun 2000.

[38] Daniel E. Leisman, Michael O. Harhay, David J. Lederer, Michael Abramson, Alex A. Adjei, Jan Bakker, Zuhair K. Ballas, Esther Barreiro, Scott C. Bell, Rinaldo Bellomo, Jonathan A. Bernstein, Richard D. Branson, Vito Brusasco, James D. Chalmers, Sudhansu Chokroverty, Giuseppe Citerio, Nancy A. Collop, Colin R. Cooke, James D. Crapo, Gavin Donaldson, Dominic A. Fitzgerald, Emma Grainger, Lauren Hale, Felix J. Herth, Patrick M. Kochanek, Guy Marks, J. Randall Moorman, David E. Ost, Michael Schatz, Aziz Sheikh, Alan R. Smyth, Iain Stewart, Paul W. Stewart, Erik R. Swenson, Ronald Szymusiak, Jean Louis Teboul, Jean Louis Vincent, Jadwiga A. Wedzicha, and David M. Maslove. Development and Reporting of Prediction Models: Guidance for Authors from Editors of Respiratory, Sleep, and Critical Care Journals. *Critical Care Medicine*, 48(5):623–633, 2020.

[39] Po Hsuan Cameron Chen, Yun Liu, and Lily Peng. How to develop machine learning models for healthcare. *Nature Materials*, 18(5):410–414, may 2019.

[40] Sebastiano Barbieri, James Kemp, Oscar Perez-Concha, Sradha Kotwal, Martin Gallagher, Angus Ritchie, and Louisa Jorm. Benchmarking Deep Learning Architectures for Predicting Readmission to the ICU and Describing Patients-at-Risk. *Scientific Reports*, 10(1):1–10, dec 2020.

[41] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of machine learning research*, 3:1137–1155, 2003.

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

[43] J L Vincent, R Moreno, J Takala, S Willatts, A De Mendonça, H Bruining, C K Reinhart, P M Suter, and 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*, 22:707–10, 1996.

[44] Kazuyuki Hara, Hayaru Shouno, and Daisuke 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. In *2015 International Joint Conference on Neural Networks (IJCNN)*, 2015.

[45] Shuang Wu, Guoqi Li, Lei Deng, Liu Liu, Dong Wu, Yuan Xie, and Luping Shi. L1 -Norm Batch Normalization for Efficient Training of Deep Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems*, 30(7):2043–2051, jul 2019.

[46] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017.

[47] Diederik P Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization, 2015.

[48] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020.

[49] Shrikant Bangdiwala and Viswanathan Shankar. The agreement chart. *BMC medical research methodology*, 13:97, 07 2013.

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 in either baseline characteristics, time-evolving characteristics (i.e, organ-support therapies), and physiological signals, were included 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

Figures legends

Fig. 1| Periods definition and flow-chart of patients selection. A, Patients from MIMIC-III and Lariboisière cohort, have their ICU stay divided, from the admission to the discharge, into periods and each period was divided into 3 successive windows. To predict the average 5-min MAP and HR, only data recorded during the observation were used. **B**, All patients from the MIMIC-III and Lariboisière cohort were selected if there were no missing data in either physiological signals and clinical information. 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.

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.

Extended Data Fig. 1| Learning framework process. 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 split of the data was done 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 databases

Fig. 3| 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. 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. 4| Calibration plots for the value of MAP averaged over 5 min. Patients were grouped into deciles. Into each decile, the average observed and predicted MAP is calculated. The first corresponds to the true 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. MIMIC-III, medical information mart for intensive care III.

Fig. 5| Positive and negative predictive values for acute hypotensive episodes prediction. Note that in the Lariboisière data, some values are missing due to the absence of patients in the category. Acute hypotensive episode; MIMIC-III, medical information mart for intensive care III.

Extended Data Fig. 2| Agreement plots for acute hypotensive episodes on the MIMIC-III validation set. This chart gives an overview of misclassification between observed and predicted risk class of AHE defined as "Very high" if predicted average 5-min MAP $\leq$ 60, "High" if 60 $<$ predicted average 5-min MAP $\leq$ 65, "Moderate" if 65 $<$ predicted average 5-min MAP $\leq$ 70 and "Low" if predicted average 5-min MAP $>$ 70. 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 the farthest class is 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.

Extended Data Fig. 3| Agreement plots for acute hypotensive episodes on Lariboisière cohort. This chart gives an overview of misclassification between observed and predicted risk class of AHE defined as "Very high" if predicted average 5-min MAP ≤ 60, "High" if $60 <$ predicted average 5-min MAP ≤ 65, "Moderate" if $65 <$ predicted average 5-min MAP ≤ 70 and "Low" if predicted average 5-min MAP > 70. 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 the farthest class is 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.

Extended Data Fig. 4| 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. 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.

Extended Data Fig. 5| Calibration plots for the value of HR averaged over 5 min. Patients were grouped into deciles. Into each decile, the average observed and predicted MAP is calculated. The first corresponds to the true 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. MIMIC-III, medical information mart for intensive care III.

Extended Data Fig. 6| Physiological Deep Learner's predictions examples Individual predictions of MAP and HR for four different patient's period with a time gap of 15 minutes are presented. **A** and **B** correspond to patients with average 5-min MAP below 65 mmHg and **C** and **D** to patients with average 5-min MAP greater than 65 mmHg. HR, heart rate; MAP, mean arterial pressure.