Computer Methods and Programs in Biomedicine 282 (2026) 109383

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.sciencedirect.com/journal/computer-methods-and-programs-in-biomedicine

Tracheostomy weaning in patients with severe acquired brain injury: External validation of machine learning models

Piergiuseppe Liuzzi [a], Bahia Hakiki [a,b,*], Francesca Draghi [a], Agnese De Nisco [a], Anna Maria Romoli [a], Daniela Maccanti [a], Antonello Grippo [a], Rachele Burali [a], Alfonso Magliacano [a], Anna Estraneo [a], Angela Comanducci [c], Jorge Navarro [c], Caterina Tramonti [d], Valentina Carli [d], Pietro Balbi [d], Claudio Macchi [a], Francesca Cecchi [a,b], Andrea Mannini [a]

[a] IRCCS Fondazione Don Carlo Gnocchi ETS, Firenze, Italy
[b] Università di Firenze, Dipartimento di Medicina Sperimentale e Clinica, Firenze, Italy
[c] IRCCS Fondazione Don Carlo Gnocchi ETS, S. Maria Nascente, Milano, Italy
[d] Fondazione Don Carlo Gnocchi ETS, Polo Riabilitativo del Levante ligure, La Spezia, Italy

ARTICLE INFO

Keywords:
Severe acquired brain injury
Tracheostomy
Machine learning
Clinical decision support tools
Disorders of consciousness

ABSTRACT

Background and objective: Our primary aim was to externally validate previously developed machine-learning (ML) models for predicting the probability of tracheostomy decannulation after 3 months from admission to rehabilitation inpatient in patients with severe Acquired Brain Injury (sABI) using a new external and temporally-independent multicentric prospective dataset. A secondary aim was to evaluate the timing of decannulation and to assess model calibration and clinical net benefit.

Methods: External validation data was collected within the PRABI study, comprising 435 sABI patients admitted between January 2020 and April 2024 across four centers. A previously trained ensemble model and a AdaBoost SVR model were used to predict decannulation probability and timing on such external dataset, respectively. Performance was assessed using metrics such as accuracy, Area Under the Receiver Operating Characteristic curve (AUROC), sensitivity, and median absolute error.

Results: The external validation dataset included 402 patients. The ensemble model achieved an accuracy of 81.8 % and an AUROC of 0.85 for decannulation probability, with sensitivity and specificity of 76.4 % and 86.2 %, respectively. The AdaBoost SVR model predicted decannulation timing with a median absolute error of 26.2 days and an accuracy of 76.2 % when predictions were dichotomized at the 90-days threshold.

Conclusions: This study successfully validated the ML models on an independent dataset, demonstrating their robustness and generalizability. Accurate prediction of decannulation probability and timing is crucial for optimizing the management of sABI patients, reducing infection risks, enhancing recovery, and facilitating smoother transitions to home care. External validation is a critical step for ensuring the reliability of ML models in diverse clinical settings, paving the way for their integration into clinical practice.

1. Introduction

Severe acquired brain injury (sABI) encompasses traumatic, anoxic, vascular, or other brain damages causing coma lasting at least 24 h [1]. The high clinical complexity of sABI patients [2] often requires continuous monitoring. In particular, the placement of a tracheal cannula supports the respiratory drive and provides a method for the suction of secretions from the inferior respiratory areas. After the acute phase, patients with sABI are hospitalized in Intensive Rehabilitation Units (IRUs) where decannulation represents one of the key goals of the rehabilitation program [3]. Indeed, tracheostomy increases the respiratory infections risk, negatively impacts the behavioral condition of the patient [4], might impair physiological swallowing [5] and it might lead to bleeding and to tracheal stenosis [6]. In addition, cannula removal, by substantially reducing the care burden, may facilitate the patient's domiciliation.

* Corresponding author at: Centro IRCCS Don Gnocchi, Italy.
E-mail address: badiabahia.hakiki@unifi.it (B. Hakiki).

https://doi.org/10.1016/j.cmpb.2026.109383
Received 5 August 2025; Received in revised form 14 February 2026; Accepted 16 April 2026
Available online 17 April 2026

In a previous study, Mannini et al. [7] developed a predictive Machine Learning (ML) model targeting decannulation probability and timing in sABI survivors (within the rehabilitation inpatient) based on a combination of clinical and neurophysiological predictors. In both cases, higher consciousness levels assessed using the Coma recovery Scale revised (CRS-R), lower disability as measured by the Disability Rating Scale-DRS, and a non-pathological Body Mass Index were associated with tracheostomy removal, while anoxic etiology, the presence of acquired sub-tentorial lesions, and the need of mechanical ventilation were associated with a negative removal probability. The models achieved promising results with an accuracy of 84.8 % for decannulation probability and a median absolute error of 25.7 days for timing prediction.

External validation of a ML model is paramount for ensuring trustworthiness before clinical deployment and translation to daily routine [8–10]. It involves evaluating a model's performance on an independent dataset not used during its development, distant temporally and/or geographically, thereby ensuring its applicability across diverse patient populations and clinical settings. This process helps identify potential overfitting and biases, providing a more accurate estimate of a model's real-world performance. Despite its importance, recent studies have highlighted significant gaps in the conduct and reporting of external validation studies [10]. A systematic review assessing machine learning-based prediction models across various medical specialties revealed prevalent methodological shortcomings and a high risk of bias [9]. The review emphasized the necessity for rigorous validation practices and transparent reporting to enhance the credibility and utility of predictive models in clinical practice. In the context of decannulation prediction for sABI patients, the only available model, the DecaPreT prediction tool, was designed to estimate the probability of successful decannulation with a logistic regression model and was retrospectively validated [11]. Restricted to a cohort of patients with a traumatic brain injury, Khan et al. compared early versus late tracheostomy and reported associations with ICU length of stay, highlighting the relevance of timing decisions in the acute phase [12]. Likewise, Ashrafi et al. developed and compared multiple machine-learning models to predict ventilator-associated pneumonia in TBI using ICU electronic health record data, emphasizing prediction of acute respiratory complications and early intervention in critical care settings [13]. These studies address important but distinct ICU-focused questions (tracheostomy timing, VAP risk) in TBI populations. In contrast, the present study focuses on a different clinical decision and time window, i.e., tracheostomy decannulation during inpatient intensive rehabilitation, and evaluates both decannulation probability and timing in a broader etiology-related cohort of sABI patients.

Thus, this is the first-ever work targeting the temporally-distant external validation of a freely-available ML solution on multicenter prospective data in sABI patients, then embedded with interpretability assessments. This is paramount given that model transportability can be affected by differences in case-mix and tracheostomy-weaning practices across centers and time periods, and it needs to be systematically evaluated before considering real-world use [14,15]. Specifically, we sought (i) to perform a multicenter and temporal external validation of previously published models for decannulation probability and timing; (ii) to evaluate model performance in clinically relevant subgroups (e.g., patients with and without disorders of consciousness); and (iii) to assess calibration, clinical net benefit using decision-curve analysis, and the feasibility of integrating the models into bedside decision support.

2. Methods

2.1. Data collection, intended use, and outcome definition

Participants of the original dataset were admitted to the IRUs of the IRCCS Fondazione Don Carlo Gnocchi in Florence from August 1, 2012 to January 31, 2019 (protocol N. R17505). The adopted inclusion criteria were the admission diagnosis of sABI, age of 18 and above and presence of tracheostomy at admission [7].

Similarly, the dataset used for external validation derived from the PRABI study [16], a multisite (N = 4) prospective study, comprising sABI patients admitted in different sites of Fondazione Don Carlo Gnocchi between June 2020 and April 2024 (IRCCS Fondazione Don Carlo Gnocchi ONLUS, Firenze; Polo Specialistico Riabilitativo Fondazione Don Carlo Gnocchi ONLUS, Sant'Angelo Dei Lombardi; IRCCS Fondazione Don Carlo Gnocchi, Centro S. Maria Nascente, Milano; Fondazione Don Carlo Gnocchi, Polo Riabilitativo del Levante ligure, La Spezia) located in different regions of Italy. Similar to the original study, the external validation dataset included demographic data, clinical evaluations, and functional scales recorded within one week of admission. The analysis was decannulation outcome was collected at discharge from the Intensive Rehabilitation Unit. The intended use of the models is to support clinicians in expressing personalized prognosis estimates with families/caregivers, to complement bedside clinical decision-making regarding readiness to attempt tracheostomy decannulation and to help planning its expected timing during inpatient rehabilitation. For the classification task, the outcome was defined as successful tracheostomy decannulation occurring during the rehabilitation stay, with a maximum observed follow-up of 342 days from admission (median length of stay 112 days, interquartile range 65 days, maximum 342 days). Thus, the probability model estimates the likelihood of successful decannulation within this admission period. For the regression task, the outcome was the number of days from rehabilitation admission to tracheostomy decannulation among patients who were successfully decannulated. These time anchors were used consistently for both model development and external validation.

2.2. Ethics and declarations

The patients have been included after signing a written consent and have been recruited following the Helsinki Declaration and its later amendments. The ethical committee' approval protocol details for each center are: Florence: N. 16606_{OSS}; Santa Maria Nascente: N.21/2020/CeFdG/FC/SA; Sant'Angelo dei Lombardi: N. 1560 amendment N. 1888; La Spezia: N. 625/2020. The manuscript was prepared following TRIPOD+AI guidelines [17]. All patients from the PRABI study were included in the external validation dataset.

2.3. Data preparation and model definition

Included features were obtained from the feature selection results in Mannini et al. and are reported in detail in Supplementary Table 2 and in the *patient_input_sample.xlsx* file uploaded to the GitHub. Missing predictors were handled using K-nearest neighbor imputation fitted on the development cohort and then applied unchanged to the external cohort (released in the GitHub). External numerical data was normalized using the mean and standard deviation of the training data (released in the GitHub). On the project page, a dedicated python script termed *model_use.py* is provided to use the trained models, including the fitted preprocessing pipelines, by simply modifying the template *patient_input_sample.xlsx* file.

The previously derived optimization-derived hyperparameters [7] were used to deploy the models. For further details on hyperparameters optimization and validation, refer to Mannini et al. [7]. Exact hyperparameters values are provided as supplementary data to the manuscript for reproducibility. The first model, an ensemble composed by the weighted average between the predicted posterior probabilities of a Support Vector Machine (SVM) and Random Forest (RF), was used to tackle decannulation probability. The second model, an Adaptive Boosting (AdaBoost) with a Support Vector Regression (SVR) as a weak learner, predicted the decannulation timing. The classification and regression models were applied in a "frozen" state (trained as in Mannini et al.) without any refitting or hyperparameter tuning on the external

study data.

2.4. Model application and statistical evaluation

Models were tested with the prospective external dataset and embedded with Shapley Additive ExPlanations (SHAP [18]) to confirm the predictive power of the associated factors. For the classification model, uncertainty around performance estimates was quantified by computing 95 % confidence intervals for the area under the receiver operating characteristic curve (AUROC), sensitivity, specificity and F1-score. AUROC confidence intervals were obtained using DeLong's method, whereas confidence intervals for sensitivity, specificity and F1-score were derived by non-parametric bootstrap with 1000 resamples. For the regression model predicting decannulation timing, 95 % confidence intervals for the mean absolute error (MAE) and for the Pearson correlation coefficient between observed and predicted times were estimated using the same bootstrap procedure. Overall, such metric set allows direct evaluation of study objectives by i) evaluating external discrimination for the probability model under clinically meaningful trade-offs (sensitivity/specificity and F1), and (ii) quantifying the magnitude of prediction error (MAE, R2) for time-to-decannulation predictions. Calibration of the classification model was assessed by reporting the Brier score, calibration slope and intercept, and by inspecting a calibration plot (reliability curve) based on grouped predicted risks. Calibration addresses a different question from discrimination. Discrimination evaluates whether the model correctly ranks patients (i.e., assigns higher values to patients more likely to be decannulated), whereas calibration evaluates whether the model output can be interpreted as an accurate probability. In calibration-in-the-large, the intercept quantifies systematic over- or underestimation of the overall event probability (a positive intercept indicates predictions are, on average, too low and need an upward shift), whereas the slope quantifies whether predicted probabilities are too extreme or not extreme enough.

To assess clinical utility, we performed decision-curve analysis (DCA) on the externally validated classification model. For each patient, we calculated the predicted probability of successful decannulation and evaluated net benefit over a set of clinically relevant threshold probabilities. We focused on thresholds of 0.30, 0.40, 0.50, 0.60, and 0.70, chosen symmetrically around 0.50 so that, within a distance of 0.20 from this "equipoise" point, the model is evaluated in the probability range where clinical uncertainty is greatest and where decision support is most likely to influence practice. Net benefit at each threshold p_t was defined as $NB(p_t) = \frac{TP}{N} - \frac{FP}{N}\frac{p_t}{1-p_t}$, where TP and FP are the numbers of true and false positives and N is the sample size. Model net benefit was compared with "treat-all" and "treat-none" strategies. Clinically, each threshold probability p_t represents the minimum predicted likelihood of decannulation at which a clinician would consider it worthwhile to trigger an active decannulation pathway (e.g., prioritizing multidisciplinary evaluation, intensified weaning trials, and closer monitoring), rather than continuing standard cannula management and reassessment. Decision-curve analysis quantifies the expected clinical value of using the model at a given p_t by balancing the benefit of correctly identifying patients who will be decannulated (TP) against the harm and resource use associated with incorrectly flagging patients who will not (FP). Net benefit can be interpreted as the number of additional true-positive decisions per patient (or per 100 patients) achieved by using the model, after accounting for the implicit trade-off between missed opportunities and unnecessary escalation encoded by p_t.

Uncertainty around the model targeting decannulation time was quantified by residual bootstrap prediction intervals. Residuals from the training data were computed as $e_i = y_i^{true} - y_i^{pred}$. For each external-cohort subject, we generated 1000 bootstrap predictions by adding a residual sampled with replacement from the training residuals to the point prediction. The 2.5th and 97.5th percentiles of these bootstrap samples defined a 95 % prediction interval for each patient. We reported coverage (proportion of observed times falling inside the interval) and the distribution of interval widths. To evaluate the reliability of the regression model across the range of predictions and across centers, we additionally constructed a Q–Q reliability plot of standardized residuals. On the external cohort, residuals were standardized to zero mean and unit variance. For quantile levels from 5 % to 95 % we computed the empirical quantiles of these standardized residuals and plotted them against the corresponding theoretical quantiles of a standard normal distribution. Separate curves were drawn for the overall test set and for each center, thereby allowing visual assessment of center-specific bias or deviations from normality in the timing errors. Regression predictions were dichotomized using a threshold of 3 months from the event and used to compute a classification accuracy. We used 90 days (3 months) as a pragmatic early-versus-late threshold, consistent with the original study and with the typical time scale of inpatient rehabilitation planning. On average, 90 days after admission corresponds to 180 days after the event marking the transition from the subacute to the chronic phase of post-sABI outcomes. In the chronic phase, the probability of achieving rehabilitation goals, including decannulation, is significantly reduced.

2.5. Data availability

Anonymized data can be obtained by reasonable request from any qualified investigator, whilst trained models, pre-processors, and trained imputers are freely distributed together with a python script running the predictions process at the following GitHub link (https://github.com/PeppeDrone/DoCDecannulation).

3. Results

3.1. Cohort and outcome prevalence

The developmental dataset was retrieved from Mannini et al. [7] and was used for model training using hyperparameters found in Mannini et al. For what concerns the external dataset, four-hundred thirty-five patients were enrolled in the PRABI study, of which 408 (93.8 %) patients were admitted to the IRU with a tracheostomy.

After removing patients with a missing outcome (Fig. 1), the external validation dataset resulted in a cohort of 402 patients (238 males; 106 traumatic, 30 anoxic, 182 hemorrhagic, 70 ischemic, 15 other; median age 68 years [IQR = 19]); median time post-injury 45 days [IQR = 25]; median Disability Rating Scale, DRS of 22 points [IQR = 5]). Out of 402 (Florence: 67 %, Milan: 10 %, Sant'Angelo dei Lombardi: 18 %, La Spezia: 5 %), 193 patients had a diagnosis of a pDoC at admission (73 in an Unresponsive Wakefulness Syndrome-UWS and 120 in a Minimally Conscious State-MCS) with a median CRS-R value of 8 points [IQR = 7]. At T1, 235 patients were successfully decannulated (58.3 %) coherent with the prevalence in the original validation dataset (56.9 %). In comparison with the first cohort, the second showed a significant increase of patients without a pDoC, and therefore a decrease of disability levels, in the external validation dataset ($p < 0.05$, see Supplementary Table 1) with no significant differences in age, sex, level of cognitive functions, and feeding supports.

3.2. Decannulation probability: discrimination accuracy

For both models the adopted features were the ones which retained univariate significance in the developmental dataset [7] (N = 16 and N = 13 respectively for probability and timing predictions) and are reported in Supplementary Table 2. All summarized results are also reported in Supplementary Table 3.

The ensemble model predicted decannulation probability with an accuracy of 81.8 % (95 %CI 0.784–0.853) and an AUROC of 0.860 (95 % CI: 0.822–0.897) on the external dataset (Fig. 2A and 2B, Table 1). Sensitivity for predicting successful decannulation was 0.821 (95 % CI

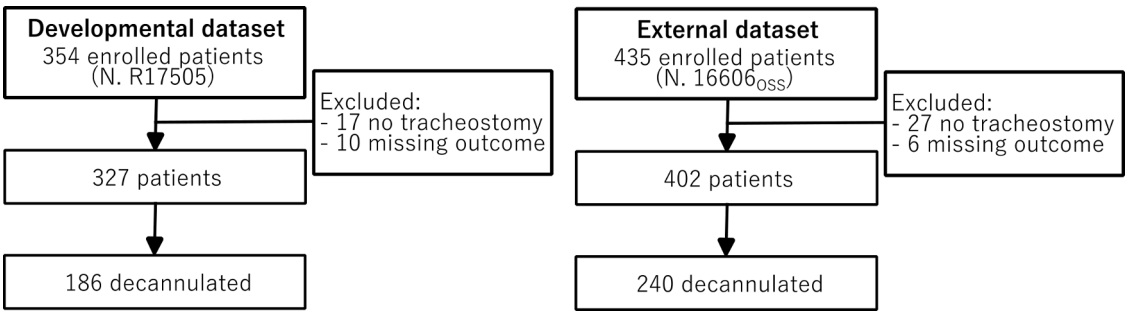

Fig. 1. Admission and enrollment flow-chart of the retrospective developmental study and the PRABI multicentric study. For both datasets Ethical Committee numbers were reported (R, Retrospective, OSS, Observational.

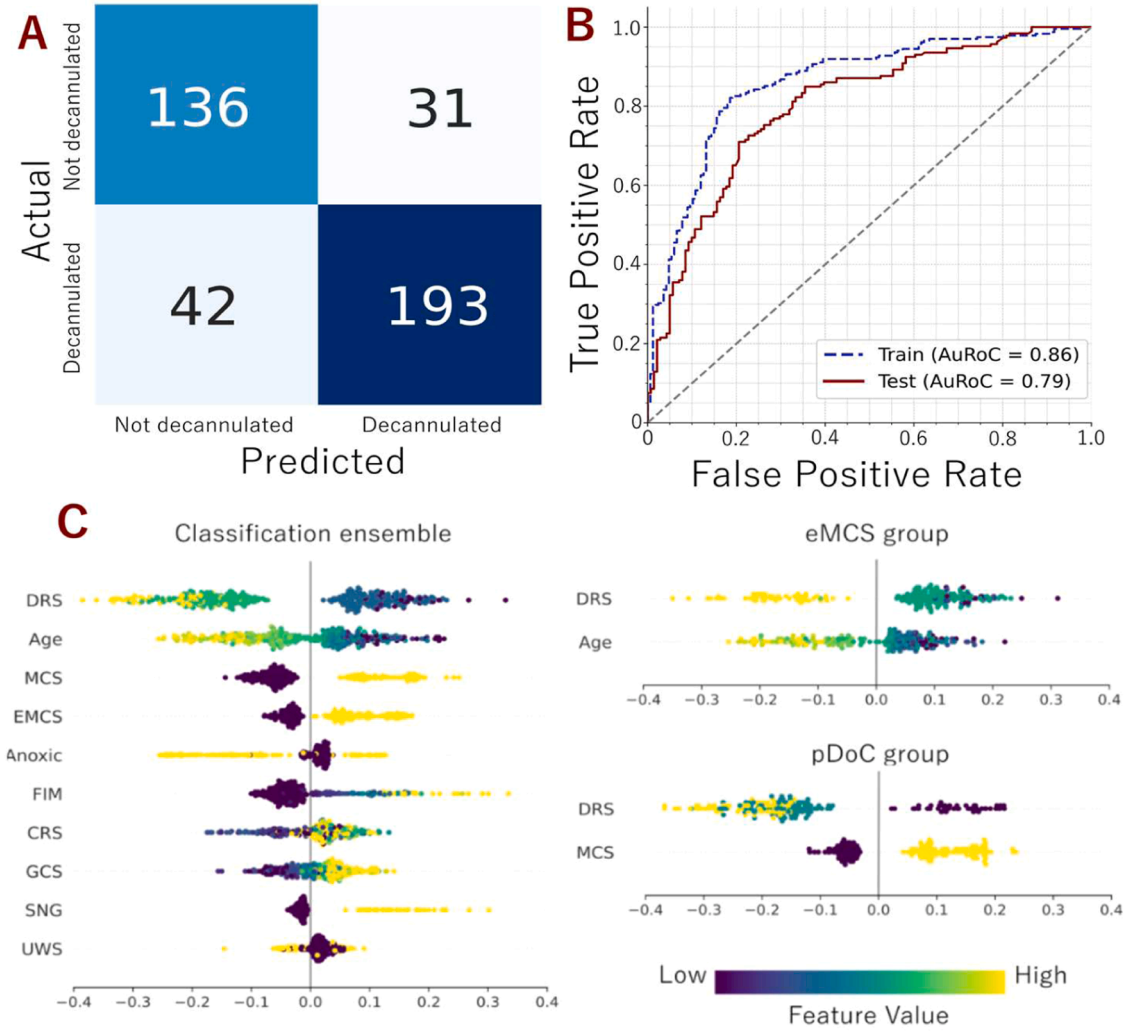

Fig. 2. Confusion matrix (panel A) and ROC curve (panel B) for the decannulation probability classification. In panel C, SHAP analysis of the most important predictors were reported for the entire cohort (left) and of the two most important predictors for the eMCS and pDoC sub-cohorts (right). On the x-axis the SHAP value (importance) of the variable is reported and the color of the points codes the value of the specific features (e.g., high ~ old age). Each point represents a patient and features are ordered according to the average SHAP value of the cohort. **Legend**. AUROC: Area under the Receiver Operator Characteristic; pDoC: prolonged Disorders of Consciousness; eMCS: emergence from Minimally Conscious State; SHAP: SHapley Additive exPlanations.

0.773–0.868; 82.1 %, 95 % CI 77.3–86.8 %), and specificity for predicting non-decannulation was 0.814 (95 % CI 0.754–0.870; 81.4 %, 95 % CI 75.4–87.0 %). The weighted F1-score was 0.819 (95 % CI 0.784–0.855). SHAP analysis (averaged between the two models embedded in the classification solution) showed that the two most important predictors were related to disability levels (in terms of DRS) and age (Fig. 2C). In particular, plots illustrate the contribution of these variables to the model's decision-making process, where higher DRS values, indicative of more severe disability, were associated with a lower likelihood of decannulation. Similarly, age influenced the model's predictions, with an older age being associated to a worse outcome. Similarly, being either an MCS or and EMCS patient (compared to UWS), as

well as a higher CRS-R total score at admission to the IRU and a higher GCS value during the acute phase were associated to better decannulation outcomes. Conversely, an anoxic etiology and a lower functional independence (Functional Independent Measure [19]) were both associated with a worse recovery. Grouping patients in the external validation group according to the presence of a pDoC, showed how feature importance of patients without a pDoC was coherent with the one derived from the general sABI cohort (i.e., DRS first, age second). On the other hand, in pDoC patients (193, 47.8 %) consciousness level retained a higher weight than age in comparison with the whole and with the No-DoC cohort coherently with recent literature [14,20].

3.3. Decannulation probability: clinical utility and calibration

To quantify the potential clinical utility of the ensemble as a bedside decision support tool, we performed Decision Curve Analysis (DCA, Fig. 3A) on the external test set. Across a clinically relevant range of decision thresholds, the model showed positive net benefit compared with both "treat-all" and "treat-none" strategies. At threshold probabilities of 0.30, 0.40, 0.50, 0.60, and 0.70, the net benefit of the ensemble was 0.4097 (95 % CI 0.3436–0.4727), 0.3507 (95 % CI 0.2786–0.4237), 0.4030 (95 % CI 0.3408–0.4627), 0.1468 (95 % CI 0.1107–0.1816), and 0.0075 (95 % CI 0.0000–0.0174), respectively. At the same thresholds, the net benefit of a "treat-all" strategy was 0.4065, 0.3076, 0.1692, –0.0386, and –0.3847, while "treat-none" consistently yielded zero net benefit. Thus, for thresholds between 0.3 and 0.6, the model provides a meaningful incremental net benefit over trivial strategies. Calibration analysis (Fig. 3B) yielded a calibration slope of 3.09 and an intercept of 0.245, indicating more extreme risk estimates, despite great generalizability in patients' ranking (Brier score on the overall test-set of 0.205, AUROC > 0.85).

3.4. Decannulation probability: center-wise stability

Grouping the external test predictions by center (inside versus outside from Florence), no substantial differences were detected in performance metrics. In particular, in patients outside Florence, the ensemble achieved an accuracy of 0.841 (sensitivity 0.786, specificity 0.937, F1-score 0.844, Brier score of 0.194) and an AUROC (with De Long) of 0.801 (95 %CI 0.797–0.945). In the Florence center an accuracy of 0.807 (sensitivity 0.841, specificity 0.765, and F1-score 0.807, Brier score of 0.210) was achieved with an AUROC of 0.781 (95 %CI 0.763–0.891). Net benefit was also computed center-wise at clinically relevant thresholds. For Florence net benefit was 0.375 and 0.367 at thresholds 0.3 and 0.5, respectively, while out-of-Florence net benefit was approximately 0.481 and 0.477 at the same thresholds. Centers-specific curves are reported together with the global (test) DCA curve in Fig. 3. Center-wise Brier scores were comparable between the out-of-Florence center (Brier 0.194) and the Florence center (Brier 0.210), as well as calibration slopes (2.55 and 2.77, respectively) and positive intercepts (0.097 and 0.074), confirming overconfident probability estimates across sites. Taken together, these findings suggest that while the ensemble is reliable for ranking patients and supporting binary decisions, its absolute decannulation risk taken as event-related probabilities may benefit from local recalibration.

3.5. Decannulation timing: accuracy

Median decannulation timing was found to be 99 days [IQR = 46] from the brain injury (97 days [IQR = 78] in Mannini et al. [7]). The AdaBoost SVR model predicted the decannulation timing with a median absolute error of 25.53 days (95 % CI 21.56–28.00) with an IQR of absolute errors was 26.78 days (95 % CI 22.28–33.27, Table 1), slightly higher than the original study but still demonstrating substantial predictive accuracy (moderate-to-good correlation, with correlation coefficient R = 0.57). Converting the regression predictions in a binarized classification resulted in an accuracy of 76.2 % (specificity of 82.63 %, sensitivity 71.43 %).

3.6. Decannulation timing: reliability and center-wise stability

To complement average error measures and estimate the target

Table 1
Comparison of classification and regression metrics for the prediction of decannulation probability and timing. Models' performances on the internal validation dataset [7] and on the external validation dataset are reported in two separate columns. Median and IQR are reported for the Median Absolute Error of the regression problem.

	Original dataset	External dataset
A: Prediction of decannulation probability		
Accuracy	84.8 %	81.8 %
AUROC	0.85	0.79
Sensitivity	84.6 %	82.1 %
Specificity	85.0 %	81.4 %
F1-score	81.5 %	81.9 %
B: Prediction of decannulation timing		
Median Absolute Error	25.7 [IQR = 25.6]	25.5 [IQR = 26.8]
Dichotomized accuracy	77.5 %	76.2 %

Legend. AUROC: Area Under the Receiver Operator Characteristic; IQR: Inter-Quartile Range.

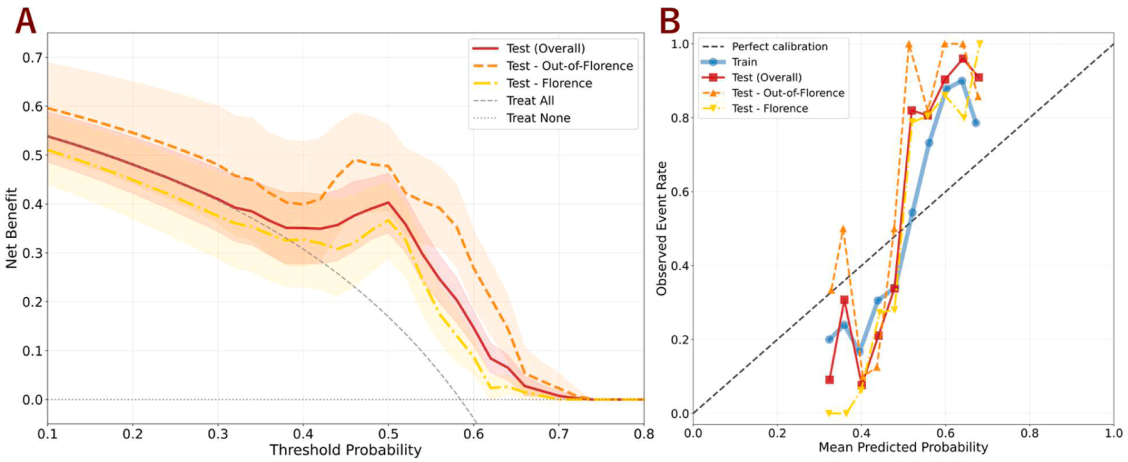

Fig. 3. DCA (panel A) and calibration curves (panel B) for the decannulation probability ensemble model. In both analysis, external test-set center-wise results are also reported. **Legend**. DCA: Decision Curve Analysis.

uncertainty, we computed 95 % prediction intervals (PIs) for the timing model using a residual bootstrap approach. Prediction intervals were derived from the training residual distribution and then applied to each test prediction. In the external timing test set (N = 223), 92.4 % of true decannulation times fell within the nominal 95 % prediction intervals, indicating slightly conservative but reasonable coverage (expected coverage 95 %), with a median of the PI distribution of 195.3 days (similar in patients from and outside Florence, respectively 0.959 and 0.853). Reliability analysis (Fig. 4B) showed that standardized residuals in the external cohort were approximately aligned with the theoretical normal quantiles, with only mild departures in the upper tail, consistent with a small proportion of patients experiencing substantially longer decannulation times than predicted. When stratifying by center, the Q–Q curves for Florence and out-of-Florence patients largely overlapped, indicating similar calibration of timing errors across sites and no clear center-specific bias towards early or late predictions. In line with this, analyzing separately patients from Florence and out-of-Florence yielded a median absolute error of 25.60 days [IQR 22.09 days] and 24.79 days [IQR 30.40 days], respectively, suggesting comparable accuracy of the timing model across centers despite residual variability in individual predictions.

4. Discussion

4.1. Main findings and clinical implications

In this multicenter prospective study, we externally validated previously developed ML models for predicting both the probability and timing of tracheostomy decannulation in patients with sABI. This is, to our knowledge, the first external validation of freely available ML models specifically designed for decannulation prediction in this population, addressing a relevant gap in the literature. The classification model maintained good discrimination and calibration on a temporally and geographically distinct cohort, while the timing model achieved a clinically meaningful prediction of the number of days from rehabilitation admission to decannulation. Taken together, these results support the potential role of the proposed models as a clinically interpretable prognostic decision-support tool to estimate decannulation readiness and to assist planning during inpatient rehabilitation. However, this external validation study does not provide evidence for autonomous clinical decision-making or direct intervention driven by the model outputs. The model is intended to complement, not replace, expert assessment: final decisions regarding decannulation timing and airway management remain under the responsibility of the multidisciplinary clinical team, integrating bedside evaluation, local protocols, and evolving clinical context. Before considering routine clinical deployment, prospective multicenter impact studies are needed to demonstrate benefit on patient-centered outcomes and workflow, alongside appropriate governance procedures and, where required, local recalibration to ensure that numerical outputs are interpretable for probability communication. Decannulation in sABI patients is a pivotal milestone, as it reduces tracheostomy-related infection risks (e.g. tracheitis, stoma infections, pneumonia), mitigates long-term complications such as tracheal stenosis, tracheomalacia, vocal cord injury and accidental decannulation, and facilitates behavioural engagement, communication and swallowing, ultimately improving overall recovery and quality of life [15,21]. Timely decannulation also alleviates caregiver burden and supports smoother transitions to home care. In this context, accurate prediction of both decannulation probability and timing can help clinicians prioritize candidates for decannulation attempts, better allocate resources, and manage expectations of patients and families. Importantly, the externally validated models run in milliseconds on a standard computer, meaning that risk estimates can be generated in real time at the bedside and integrated into routine multidisciplinary decision-making.

Our findings show that the externally validated classification model retains good accuracy for decannulation prediction, while the regression model exhibits a median absolute timing error of approximately 26 days. When interpreted against the typical time scale of decannulation in this population (median ≈99 days from rehabilitation admission) and the observed rehabilitation length of stay (median 112 days; IQR 65; maximum 342 days), this error corresponds to a relative deviation on the order of weeks. In the context of sABI rehabilitation, where decannulation timing is frequently influenced by evolving neurological recovery, respiratory complications, secretion burden, swallowing trajectory, intercurrent infections, and operational constraints, week-level uncertainty is expected and does not necessarily preclude clinical usefulness for medium-term planning (e.g., prioritizing follow-up

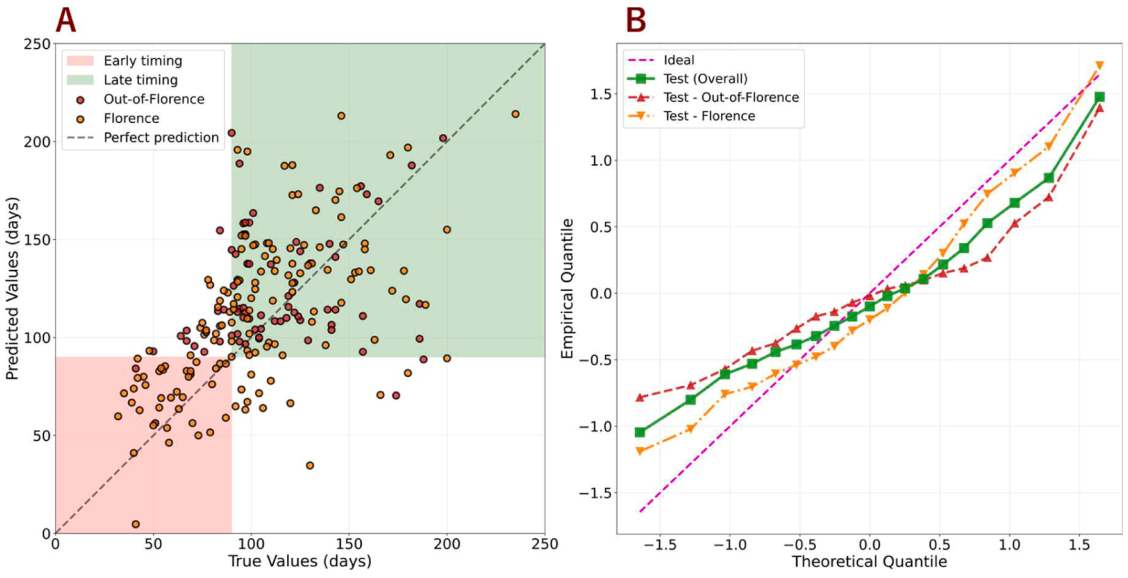

Fig. 4. Scatter plot indicating the predictions of the regression model targeting decannulation timing (panel A). The background quadrants highlight early (<90 days) and late (≥90 days) predictions, and points are colored according to recruiting center (Florence vs out-of-Florence). Reliability (Q–Q) plot (panel B) of standardized residuals, comparing the empirical quantiles of prediction errors with the theoretical quantiles of a standard normal distribution. The dashed line indicates perfect agreement between observed and theoretical quantiles. **Legend**. DCA: Decision Curve Analysis.

assessments, organizing stepwise weaning trials, and supporting transition planning). However, the individual-level prediction intervals are often wide, indicating substantial patient-specific uncertainty around the point estimate. This limits the model's utility for fine-grained scheduling or resource allocation at the individual level (e.g., booking a specific procedure slot or predicting an exact discharge milestone). Accordingly, the timing model should be interpreted as providing a probabilistic time range rather than a date-level forecast: it can support expectation-setting and approximate planning over weeks-to-months, while decisions about exact scheduling should remain anchored to serial bedside reassessment and local operational considerations.

4.2. Interpretability and subgroup analyses

To enhance interpretability and clinical trust, we complemented global performance metrics with explainability analyses and subgroup evaluations. SHAP analysis consistently identified baseline Disability Rating Scale (DRS) and age as the most influential predictors of decannulation probability, in line with previous evidence linking greater disability and older age to prolonged tracheostomy dependence [22–24]. The DRS captures both cognitive impairment and functional limitations, combining items overlapping with the Glasgow Coma Scale (e.g. eye opening, communication, motor response) and additional dimensions such as feeding, toileting, grooming, ability to live independently and employability [24–26]. Lower Functional Independence Measure (FIM) scores at admission were similarly associated with a reduced likelihood of decannulation, which is coherent with prior work demonstrating that higher functional independence correlates with better decannulation outcomes [22,27].

Acute-phase Glasgow Coma Scale scores further supported the importance of early neurological severity, with higher GCS being associated with increased decannulation probability. This confirms that initial consciousness level and acute severity can influence mid-term tracheostomy outcomes [28]. Anoxic aetiology was, as expected, linked to poorer decannulation outcomes, reflecting the well-known association between anoxic injury and less favorable recovery trajectories. A particularly informative finding emerged from subgroup analysis: when stratifying patients into those with prolonged disorders of consciousness (pDoC) and those in an emerged minimally conscious state (eMCS), feature importance partially re-ordered. In the overall cohort and in eMCS patients, disability and age remained the dominant predictors of decannulation. In contrast, among pDoC patients, consciousness level outweighed age, underscoring the pivotal role of improving awareness for achieving tracheostomy removal in this group. This observation aligns with prior studies showing that patients who exhibit signs of consciousness have a higher probability of successful decannulation [11,15]. Collectively, these interpretability and subgroup analyses help explain why the model behaves as it does and directly support our conclusion that it captures clinically meaningful determinants of decannulation.

4.3. External validation and performance stability

As expected in real-world external validation, we observed a slight performance drop compared with the original internal results [9,10, 29–31].This attenuation is typical when transporting predictive models to new populations and reflects both temporal drift and differences in case-mix and practice patterns. In our cohort, three factors likely contribute to the observed differences. First, there were shifts in patient demographics and injury characteristics between the development and external-validation cohorts, including an increased proportion of hemorrhagic etiologies and an overall reduction in patients with severe pDoC. These changes are coherent with the impact of the COVID-19 pandemic on intensive rehabilitation case-mix [32] and arguably make the prediction problem more challenging, increasing the value of demonstrating robust performance under altered conditions. Second, center imbalance and practice heterogeneity may influence performance: approximately 67 % of patients in the external cohort were recruited from the Florence center, yet the remaining centers contributed patients managed under partially different rehabilitation protocols and tracheostomy weaning criteria. This heterogeneity is not fully represented in the original monocentric development data and may introduce calibration differences across sites.

4.4. Calibration and implications for probability interpretation

In order to mimic realistic deployment of a "frozen" decision-support system, we applied the same preprocessing pipeline used during model development to the external cohort, using imputation and scaling parameters learned on the original dataset. While this approach correctly reflects how a pre-trained model would be implemented in clinical practice, it may sub-optimally adapt to center-specific or temporal shifts, with potential impact on calibration. Centre-wise analyses confirmed that performance remained good across all sites, despite showing slight probability miscalibration. In particular, the model preserved ranking capability (good discrimination between recovery and non-recovery) while exhibiting a shifted probability mapping when transported to the external cohort. In particular, the positive intercept and the slope larger than one shows that raw outputs behave like risk scores systematically shifted and not directly interpretable as bedside probabilities. Nonetheless, the preserved discrimination supports the use of the model for ranking and for binary prediction of decannulation probability. If the intended use requires communicating absolute probabilities or linking thresholds to expected event rates, a pragmatic solution is local recalibration (e.g., logistic recalibration updating the intercept alone, or both intercept and slope), which can be performed using a modest local sample without retraining the underlying model. These center effects and the modest drop in performance therefore support, rather than weaken, our main conclusion: the models retain useful accuracy and clinical utility under realistic distribution shifts, while also highlighting the importance of explicit external validation before clinical deployment.

4.5. Clinical interpretation of decision thresholds

In this study, the decision thresholds used in DCA can be interpreted as explicit *action triggers* within the inpatient rehabilitation workflow. A threshold probability p_t represents the minimum predicted likelihood of successful decannulation during the rehabilitation stay at which a clinician would consider it worthwhile to initiate an active decannulation pathway such as prioritizing multidisciplinary airway and swallowing evaluation, scheduling intensified capping/weaning trials and respiratory physiotherapy, and planning closer monitoring, rather than continuing standard cannula management with routine reassessment. In this context, lower thresholds (e.g., 0.30–0.40) reflect a strategy that tolerates more false alarms in order to avoid missing candidates who could benefit from earlier evaluation, whereas higher thresholds (e.g., 0.60–0.70) reflect a more conservative strategy that reserves escalation for patients with stronger predicted readiness, thereby limiting unnecessary resource use. Importantly, decision-curve analysis does not claim that the model 'decides' decannulation; rather, it evaluates whether using the model to prioritize escalation versus routine management yields more correct escalations than blanket strategies. The finding of positive net benefit over thresholds 0.30–0.60 (Fig. 3A) indicates that, in this clinically uncertain probability range, model-guided prioritization would be expected to identify additional true candidates for active evaluation per 100 patients, after accounting for the implicit trade-off between missed opportunities and unnecessary escalation encoded by p_t.

4.6. Strengths and limitations

This study has several strengths that directly address current gaps in the field. First, it provides a prospective multicenter external validation of ML models for decannulation prediction in sABI, using a temporally distinct cohort from four independent rehabilitation centers. Second, the models rely on clinically interpretable variables that are routinely collected in rehabilitation settings, and their predictions can be computed within milliseconds, enabling bedside-speed inference and seamless integration into multidisciplinary clinical workflows. Third, the trained models and processing code are made freely available, promoting transparency, reproducibility, and the possibility for independent investigators to test, adapt and recalibrate the tools in other healthcare systems. At the same time, important limitations must be acknowledged. Although we used decision-curve analysis and calibration assessment to characterize the potential clinical utility of the model, we did not conduct a prospective interventional study to evaluate its actual impact on clinical workflows, decision-making behavior, or patient outcomes. DCA, calibration and performance metrics alone cannot substitute for prospective trials; future interventional studies are needed to establish whether integrating this CDSS into routine care improves decannulation timing and personalization of rehabilitation pathways in the chronic phase. Lastly, the ethical and managerial complexity of defining interventional studies where clinical decisions is supported by AI in such a complex population as patients with sABI, inherently calls for further multicentric prospective validations.

Moreover, the models were deliberately developed on a restricted set of variables, defined in Mannini et al.7, and included in the database of the PRABI study (NCT04495192) selected to enhance feasibility and generalizability. While this choice facilitates implementation, it also implies that potentially relevant but unmeasured confounders were not captured and may influence decannulation outcomes. Examples include differences in the personalized rehabilitation plan (e.g., respiratory physiotherapy strategies and intensity including the postural reeducation, the use of respiratory secretion drainage equipment, the sialorrhea management and the swallowing rehabilitation), previous and concomitant comorbidities (e.g., chronic pulmonary or cardiac disease, malnutrition, diabetes, neuropathies), and intercurrent complications arising during the rehabilitation stay (e.g., respiratory and urinary infections, sepsis, bedsores, postural changes due to severe spasticity, epileptic seizures and paroxysmal sympathetic hyperactivity, cognitive, behavioural and communication alterations). Because these factors may vary across centers and over time, they can contribute to residual error and heterogeneity in external validation, and may partly explain shifts in probability calibration. This limitation may be particularly relevant for the timing model, as clinical events and operational constraints can delay or accelerate the decannulation date even when eventual decannulation remains feasible. Future studies should collect these variables prospectively and assess whether incorporating them improves transportability and individual-level timing accuracy.

Further, the median timing prediction error of ~26 days, although acceptable in the context of long rehabilitation stays (median PI width distribution of 195.3), may be less actionable in settings where length of stay is short or decisions about decannulation must be made within narrow time windows. In those environments, the classification model for decannulation probability, complemented by clinical judgement, may be more readily applicable than the exact timing predictions. Finally, despite the multicenter design, our cohort remains drawn from a specific national healthcare system (Italian) and rehabilitation network (Fondazione Don Gnocchi). However, the included centers belong to four different regional healthcare systems, allowing for some variability given by the inherent regional nature of the Italian healthcare system. Additional validation in other countries and organizational settings will be important to confirm the generalizability of our findings.

5. Conclusion

This study provides the first prospective, temporally-independent, multicenter external validation of freely available machine-learning models for predicting both the probability and timing of tracheostomy decannulation in patients with severe acquired brain injury. The models demonstrated robust discrimination, acceptable calibration, clinically meaningful timing performance and consistent behavior across centers, while preserving interpretability through the use of established clinical scales and SHAP-based explanations. These properties support their use as a bedside clinical decision support system to inform decannulation readiness and planning, with the potential to optimize management, reduce complications and facilitate more efficient transitions to home care. By combining rigorous external validation with open distribution of the trained models and code, this work represents a key step towards reliable and transparent use of ML-based tools for tracheostomy weaning in real-world rehabilitation practice, while also highlighting the need for future prospective interventional studies to quantify their impact on patient outcomes and workflows.

Ethics statement

The patients have been included after signing a written consent and have been recruited following the Helsinki Declaration and its later amendments. The ethical committee' approval protocol details for each center are: Florence: N. 16606_{OSS}; Santa Maria Nascente: N.21/2020/CeFdG/FC/SA; Sant'Angelo dei Lombardi: N. 1560 amendment N. 1888; La Spezia: N. 625/2020. The manuscript was prepared following TRIPOD+AI guidelines [15].

CRediT authorship contribution statement

Piergiuseppe Liuzzi: Writing – original draft, Software, Formal analysis, Conceptualization. **Bahia Hakiki:** Writing – review & editing, Supervision, Investigation, Funding acquisition, Conceptualization. **Francesca Draghi:** Writing – original draft, Data curation, Conceptualization. **Agnese De Nisco:** Writing – original draft, Data curation. **Anna Maria Romoli:** Writing – review & editing, Investigation, Data curation. **Daniela Maccanti:** Data curation. **Antonello Grippo:** Writing – review & editing, Methodology, Data curation. **Rachele Burali:** Data curation. **Alfonso Magliacano:** Writing – original draft, Data curation. **Anna Estraneo:** Writing – review & editing, Supervision, Data curation. **Angela Comanducci:** Writing – review & editing, Data curation. **Jorge Navarro:** Writing – review & editing, Data curation. **Caterina Tramonti:** Writing – original draft, Data curation. **Valentina Carli:** Writing – original draft, Data curation. **Pietro Balbi:** Writing – review & editing, Data curation. **Claudio Macchi:** Writing – review & editing, Project administration. **Francesca Cecchi:** Writing – review & editing, Supervision. **Andrea Mannini:** Writing – review & editing, Supervision, Resources, Methodology, Conceptualization.

Declaration of competing interest

None Declared.

Acknowledgments

This study was funded by the Italian Ministry of Health under the "Ricerca Corrente" program; the "Ricerca Finalizzata" program thorugh the project "Clinical validatiOn of an AI-based DEcision support system for Robotic upper limb rehabilitation in patients with Stroke. A CO-AIDER Study" (RF-2021-12375472) and under the frame of Transforming Health and Care Systems, THCS, (NeuroRehab4EU, GA N° 101095654 of the EU Horizon Europe Research and Innovation Programme)"

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.cmpb.2026.109383.

References

[1] R. Formisano, Severe acquired brain injury and high specialty neurorehabilitation needs, Neurol. Sci. 42 (1) (2021) 347–348, https://doi.org/10.1007/s10072-020-04587-z.

[2] Multi-center study on overall clinical complexity of patients with prolonged disorders of consciousness of different etiologies: brain injury, No 1 - Get Access. Accessed July 12, https://www.tandfonline.com/doi/full/10.1080/02699052.2020.1861652, 2024.

[3] B. Hakiki, F. Draghi, S. Pancani, et al., Decannulation after a severe acquired brain injury, Arch. Phys. Med. Rehabil. 101 (11) (2020) 1906–1913, https://doi.org/10.1016/j.apmr.2020.05.004.

[4] K.L. Christopher, Tracheostomy decannulation, Respir. Care 50 (4) (2005) 538–541.

[5] A.Y. Bicego, K. Lejoly, A. Maudoux, et al., Swallowing in disorders of consciousness, Rev. Neurol. (Paris) 170 (10) (2014). Accessed June 3, 2024, https://orbi.uliege.be/bitstream/2268/178549/1/NEUROL_1324_ProofAvecCOrrectionsEditeur.pdf.

[6] J.H. Law, K. Barnhart, W. Rowlett, O. de la Rocha, S. Lowenberg, Increased frequency of obstructive airway abnormalities with long-term tracheostomy, Chest 104 (1) (1993) 136–138, https://doi.org/10.1378/chest.104.1.136.

[7] A. Mannini, B. Hakiki, P. Liuzzi, et al., Data-driven prediction of decannulation probability and timing in patients with severe acquired brain injury, Comput. Methods Programs Biomed. 209 (2021) 106345, https://doi.org/10.1016/j.cmpb.2021.106345.

[8] N. Badjatia, J. Podell, R.B. Felix, et al., Machine learning approaches to prognostication in traumatic brain injury, Curr. Neurol. Neurosci. Rep. 25 (1) (2025) 19, https://doi.org/10.1007/s11910-025-01405-x.

[9] Navarro C.L.A., Damen J.A.A., Takada T., et al. Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. Published online October 20, 2021. doi:10.1136/bmj.n2281.

[10] F. Cabitza, A. Campagner, F. Soares, et al., The importance of being external. Methodological insights for the external validation of machine learning models in medicine, Comput. Methods Programs Biomed. 208 (2021) 106288, https://doi.org/10.1016/j.cmpb.2021.106288.

[11] C. Reverberi, F. Lombardi, M. Lusuardi, A. Pratesi, M. Di Bari, Development of the decannulation prediction tool in patients with dysphagia after acquired brain injury, J. Am. Med. Dir. Assoc. 20 (4) (2019) 470–475, https://doi.org/10.1016/j.jamda.2018.09.022, e1.

[12] Q.Y. Khan, H.A. Farooqi, D.M. Khalid, et al., Role of early vs late tracheostomy in severe traumatic brain injury outcome to check in hospital icu stay, Health Sci. J. 4 (1) (2025) 46–51, https://doi.org/10.59365/hsj.4(1).2025.184.

[13] N. Ashrafi, A. Abdollahi, K. Alaei, M. Pishgar, Enhanced prediction of ventilator-associated pneumonia in patients with traumatic brain injury using advanced machine learning techniques, Sci. Rep. 15 (1) (2025) 11363, https://doi.org/10.1038/s41598-025-95779-0.

[14] F. Draghi, S. Pancani, A.D. Nisco, et al., Implications of the consciousness State on decannulation in patients with a prolonged disorder of consciousness, Arch. Phys. Med. Rehabil. 0 (0) (2024), https://doi.org/10.1016/j.apmr.2024.05.006.

[15] B. Hakiki, S. Pancani, F. Draghi, et al., Decannulation and improvement of responsiveness in patients with disorders of consciousness, Neuropsychol. Rehabil. 32 (4) (2022) 520–536, https://doi.org/10.1080/09602011.2020.1833944.

[16] B. Hakiki, I. Donnini, A.M. Romoli, et al., Clinical, neurophysiological, and genetic predictors of recovery in patients with severe acquired brain injuries (PRABI): a study protocol for a longitudinal observational study, Front. Neurol. 13 (2022) 711312, https://doi.org/10.3389/fneur.2022.711312.

[17] G.S. Collins, K.G.M. Moons, P. Dhiman, et al., TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods, BMJ 385 (2024) e078378, https://doi.org/10.1136/bmj-2023-078378.

[18] S.M. Lundberg, S.I. Lee, A unified approach to interpreting model predictions, in: Advances in Neural Information Processing Systems, 30, Curran Associates, Inc., 2017. Accessed December 8, 2022, https://papers.nips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

[19] R.A. Keith, C.V. Granger, B.B. Hamilton, F.S. Sherwin, The functional independence measure: a new tool for rehabilitation, Adv. Clin. Rehabil. 1 (1987) 6–18.

[20] E. Mélotte, A. Maudoux, R. Panda, et al., Links between swallowing and consciousness: a narrative review, Dysphagia 38 (1) (2023) 42–64, https://doi.org/10.1007/s00455-022-10452-2.

[21] J. Bösel, Tracheostomy in stroke patients, Curr. Treat. Options. Neurol. 16 (1) (2013) 274, https://doi.org/10.1007/s11940-013-0274-1.

[22] T. Gallice, E. Cugy, O. Branchard, P. Dehail, G. Moucheboeuf, Predictive factors for successful decannulation in patients with tracheostomies and brain injuries: a systematic review, Dysphagia 39 (4) (2024) 552–572, https://doi.org/10.1007/s00455-023-10646-2.

[23] C. Perin, R. Meroni, V. Rega, G. Braghetto, C.G. Cerri, Parameters influencing tracheostomy decannulation in patients undergoing rehabilitation after severe acquired brain injury (sABI), Int. Arch. Otorhinolaryngol. 21 (4) (2017) 382–389, https://doi.org/10.1055/s-0037-1598654.

[24] P. Liuzzi, A. Grippo, S. Campagnini, et al., Merging clinical and EEG biomarkers in an elastic-net regression for disorder of consciousness prognosis prediction, IEEE Trans. Neural. Syst. Rehabil. Eng. Publ. IEEE Eng. Med. Biol. Soc. 30 (2022) 1504–1513, https://doi.org/10.1109/TNSRE.2022.3178801.

[25] K.M. Hall, N. Mann, W.M.J. High, J. Wright, J.S. Kreutzer, D. Wood, Functional measures after traumatic brain injury: ceiling effects of FIM, FIM+FAM, DRS, and CIQ, J. Head Trauma Rehabil. 11 (5) (1996) 27.

[26] A. Deepika, B.I. Devi, D. Shukla, Predictive validity of disability rating scale in determining functional outcome in patients with severe traumatic brain injury, Neurol. India 65 (1) (2017) 83–86, https://doi.org/10.4103/0028-3886.198228.

[27] Y. Chen, G. Aishan, S. Fan, et al., Predictors of long-term decannulation in patients with disorders of consciousness, Front. Neurol. 14 (2023), https://doi.org/10.3389/fneur.2023.1099307.

[28] R.D. Zafonte, F.M. Hammond, N.R. Mann, D.L. Wood, K.L. Black, S.R. Millis, Relationship between Glasgow coma scale and functional outcome, Am. J. Phys. Med. Rehabil. 75 (5) (1996) 364–369, https://doi.org/10.1097/00002060-199609000-00012.

[29] T. Ohnuma, S. Uchino, Prediction models and their external validation studies for mortality of patients with acute kidney injury: a systematic review, PLoS One 12 (2017), https://doi.org/10.1371/journal.pone.0169341.

[30] S. Moazemi, S. Vahdati, J. Li, et al., Artificial intelligence for clinical decision support for monitoring patients in cardiovascular ICUs: a systematic review, Front. Med. 10 (2023), https://doi.org/10.3389/fmed.2023.1109411.

[31] W. Liu, L. Laranjo, H. Klimis, et al., Machine-learning versus traditional approaches for atherosclerotic cardiovascular risk prognostication in primary prevention cohorts: a systematic review and meta-analysis, Eur. Heart. J. Qual. Care Clin. Outcomes. 9 (4) (2023) 310–322, https://doi.org/10.1093/ehjqcco/qcad017.

[32] B. Hakiki, A. Grippo, M. Scarpino, et al., Effects of COVID-19 pandemic on intensive rehabilitation after severe acquired brain injuries, Neurol. Sci. Off J. Ital. Neurol. Soc. Ital Soc. Clin. Neurophysiol. 43 (2) (2022) 791–798, https://doi.org/10.1007/s10072-021-05709-x.