Neurocrit Care
https://doi.org/10.1007/s12028-026-02637-6

NEUROCRITICAL CARE SOCIETY

ORIGINAL WORK

Time-Stamped Annotations in High-Frequency Physiological Data: Evaluating an Approach for Assessing Annotation Plausibility in the CENTER-TBI Dataset

Sara Turella[1,2*], Erta Beqiri[3], Stefan Yu Bögli[3,4], Bogdan Ianosi[1,2], Ihsane Olakorede[3], Tommaso Zoerle[5,6], Jeanette Tas[1,2†], Raimund Helbok[1,2†] and Peter Smielewski[3†] on behalf of the Smart Neuromonitoring to Support Precision Medicine in Acute Central Nervous Injury (SOPRANI) collaborators and the CENTER-TBI High-Resolution Sub-Study Participants and Investigators

© 2026 The Author(s)

Abstract

Background/Objective: High-resolution neuromonitoring data with time-stamped clinical annotations offer valuable insights into treatment responses, though reliability is limited by manual documentation. This study aims to develop and demonstrate a three-step methodological framework to evaluate the plausibility of time-stamped clinical annotations using a high-frequency dataset. The framework integrates visual inspection of reference annotations and automated classification based on predefined physiological criteria.

Methods: Annotated interventions in the High-Resolution Collaborative European Neuro Trauma Effectiveness Research in Traumatic Brain Injury (CENTER-TBI) dataset were retrospectively analysed. Step 1 included visually inspecting physiotherapy and suctioning annotations to assess the overall plausibility of annotations at the patient level. In Step 2, an intracranial pressure (ICP)-based classification was applied to the period before osmotherapy annotations. Rejected annotations were those without sustained intracranial hypertension (ICP > 20 mm Hg for ≥ 5 min) beforehand. Step 3 classified the accepted annotations as *effective* (ICP reduction ≥ 10 mm Hg or normalisation) or *ineffective* on the basis of the post-annotation ICP trend.

Results: Across 205 patients, 15,455 annotated interventions were identified. Visual inspection (Step 1) classified 90.2% of files as having moderate or high evidence of an annotation–signal relationship and 9.8% as having low evidence. The automated analysis (Step 2) identified 388 osmotherapy annotations in 76 patients; 140 (36.1%) were

*Correspondence: sara.turella91@gmail.com
†Jeanette Tas, Raimund Helbok and Peter Smielewski have contributed equally to this work.
[1] Department of Neurology, Kepler University Hospital, Johannes Kepler University Linz, Linz, Austria
Full list of author information is available at the end of the article

The Smart Neuromonitoring to Support Precision Medicine in Acute Central Nervous Injury (SOPRANI) collaborators and the CENTER-TBI High-Resolution Sub-Study Participants and Investigators

Springer

rejected, and 248 (63.9%) were retained. Step 3 found that, among valid events, 67.7% were effective and 32.3% ineffective. Rejected events were more frequent in low-evidence files (57.3%%) than in high-evidence ones (18.0%, $p < 0.001$).

Conclusions: This study provides the first systematic evaluation of time-stamped annotations in the CENTER-TBI high-resolution dataset. Concordant findings obtained using visual inspection and automated classification supports the credibility of both approaches and illustrates a framework for evaluating physiological plausibility, demonstrated here for osmotherapy but applicable to other annotated interventions. External validation is needed to prove the generalisability of the proposed algorithm.

Keywords: Traumatic brain injury, Osmotherapy, Clinical data annotation, CENTER-TBI, Neuromonitoring data, Neurocritical care

Introduction

Continuous high-resolution neuromonitoring enables the detection of rapid physiological changes and treatment responses in neurocritical care. Its integration into research frameworks has supported the development of individualised care strategies and predictive models [1, 2]. Despite these advances, a key limitation in using physiological datasets is the scarcity of associated annotations of treatments and nursing procedures, making it difficult to interpret the sources of captured physiological variability. Such annotated clinical interventions are rarely recorded in electronic record systems, and those entered manually by bedside staff—usually in clinical research settings—are prone to inaccuracies, inconsistencies and documentation delays [3, 4], which may ultimately reduce rather than improve data interpretability. This is particularly critical for analyses that require precise temporal alignment between interventions and physiological responses, such as causal inference or data-driven models training. For instance, supervised machine learning models assume accurate training labels, a condition rarely met in clinical settings. As noted by Sylolypavan et al. [5], such inconsistencies can compromise model performance, leading to incorrect predictions and undermining the reliability of decision-support systems in clinical practice.

Although probabilistic approaches have been proposed to handle imprecise labels [6], their mathematical complexity and lack of accessible implementations have limited their use in practice. Consequently, most studies treat annotation imprecision as a study limitation [7], rather than addressing it through systematic validation or correction. This highlights a broader methodological gap: the absence of practical frameworks to assess clinical annotations in high-resolution datasets. While it is unfeasible to establish a definitive "ground truth" for retrospective clinical annotations, assessing the internal consistency and physiological plausibility is a valuable and attainable goal.

In the Collaborative European Neuro Trauma Effectiveness Research in Traumatic Brain Injury (CENTER-TBI) Study, predefined intervention markers were available to standardise bedside annotation of clinical interventions. Instead of free-text or centre-specific annotations, staff could select from nine predefined event categories, such as osmotherapy, suctioning and physiotherapy, presented on a touch screen interface (Fig. 1). This fixed structure aimed to encourage consistent and comparable annotations across different centres [8]. Nevertheless, the process remained manual and dependent on clinical workflows, leaving it vulnerable to human error. Moreover, annotations could also be entered retrospectively, further contributing to inaccuracy and temporal imprecision, despite the best efforts of the bedside team. Concerns regarding the reliability of these annotations were later raised within the CENTER-TBI high-resolution substudy committee, and to date, no event-based analyses of physiological data from this cohort have been published.

In this project, the first of its kind, we developed and applied a three-step methodological framework to evaluate the plausibility of clinical annotations in the CENTER-TBI high-resolution dataset, using osmotherapy as a representative intervention. Specifically, the framework aims to identify annotated interventions that are inconsistent with the expected physiological context and response patterns. We hypothesised that combining visual inspection of reference annotations at the file level with automated classification of osmotherapy annotations at the event level would provide an internal consistency check of annotation plausibility. As illustrated in Fig. 2, suctioning and physiotherapy/handling are often accompanied by short-timescale, recognisable perturbations in monitoring signals. Therefore, these bedside procedures can serve as informative reference events when assessing annotation plausibility.

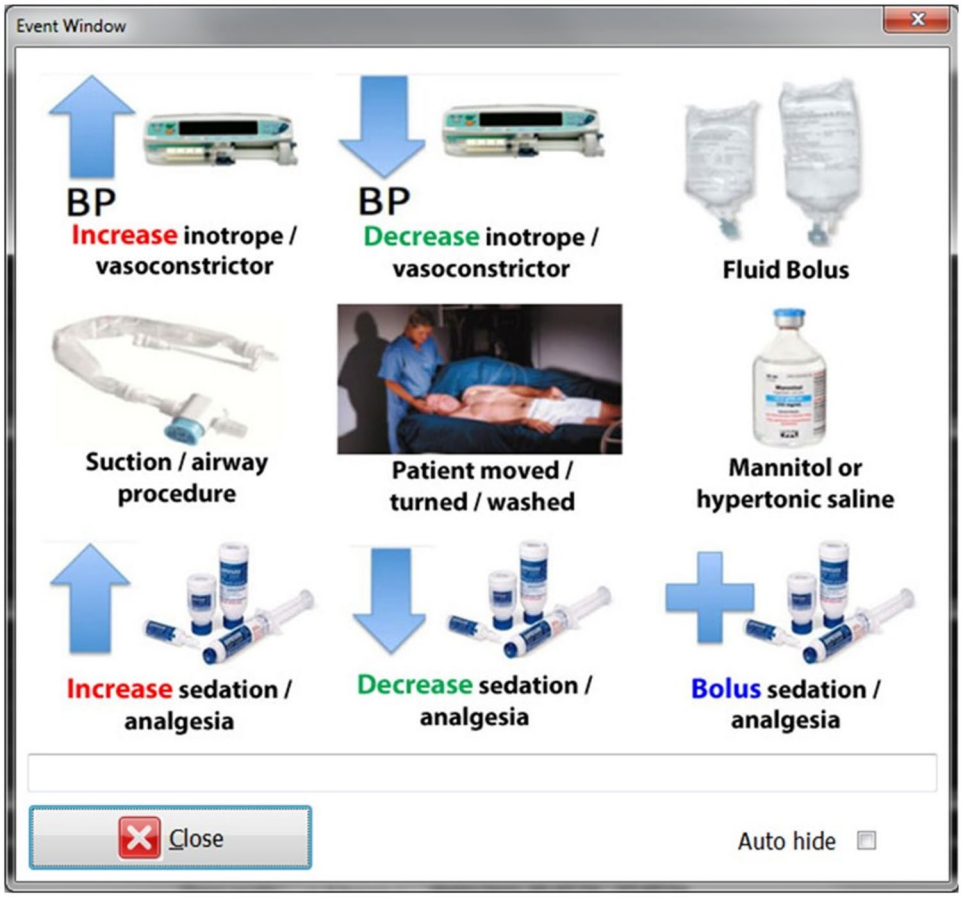

Fig. 1 The ICM+ annotation form, permanently available on the bedside laptop screen, allowing quick event entry through the touch-screen interface. ICM+, Intensive Care Monitoring software (Cambridge Enterprise Ltd)

Demonstrating this framework for osmotherapy, a well-defined intervention with quantifiable physiological effects, also illustrates a broader methodological concept that can inform the evaluation of other annotations in high-resolution neuromonitoring datasets.

Methods

Study Design

This study is a retrospective analysis of CENTER-TBI high-resolution ICU data, including high-frequency physiological waveforms and time-stamped annotations from adult patients (≥ 18 years old) with traumatic brain injury (TBI) who underwent neuro-monitoring in the intensive care unit (ICU), prospectively enrolled between January 2015 and December 2017 across 21 centres in the European Union (EU) as part of the CENTER-TBI study [9, 10]. The dataset included continuous arterial blood pressure (ABP), intraparenchymal ICP and electrocardiogram (ECG) monitoring, as well as manually annotated clinical interventions across nine pre-specified categories.

Patients were excluded if ICP monitoring was performed using exclusively an external ventricular drain (EVD), as reliable signal acquisition is only possible when the drain is temporarily closed.

Ethics

All data used in this study were collected under the ethical framework of the CENTER-TBI study. The CENTER-TBI study (EC grant 602150) has been conducted in accord with all relevant laws of the EU, if directly applicable or of direct effect, and all relevant laws of the country where the recruiting sites were located, including, but not limited to, the relevant privacy and data protection laws and regulations (the "Privacy Law"), the relevant laws and regulations on the use of human materials, and all relevant guidance relating to clinical studies from time to time in force, including, but not limited to, the ICH Harmonized Tripartite Guideline for Good Clinical Practice

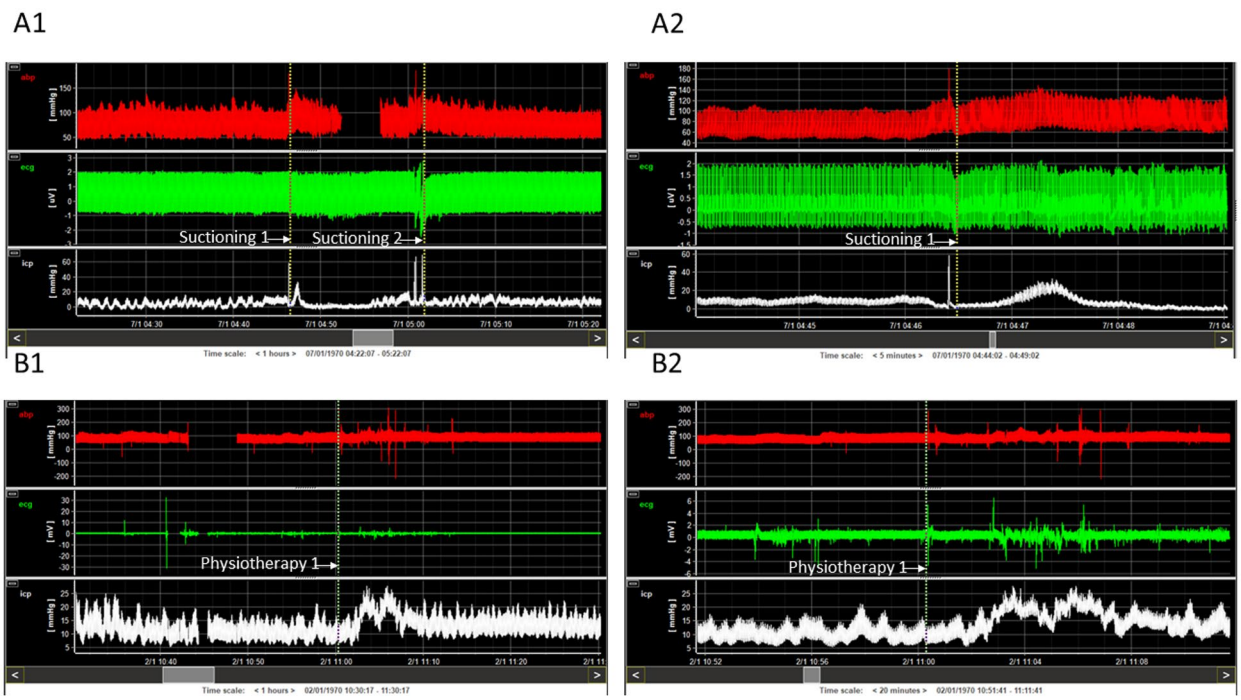

Fig. 2 Representative high-frequency signatures used for Step 1 visual screening. (**A1**–**A2**) Suctioning examples. **A1** A 1-h window (−30 to +30 min) centred on the annotated time stamps, in vertical dashed lines, illustrating two suctioning events (Suctioning 1–2). **A2** A magnified view of Suctioning 1 (5-min window), highlighting a brief, temporally localised perturbation across multiple channels and the associated short-lived physiological response (e.g. transient increase in ABP and ICP). (**B1**–**B2**) Physiotherapy example. **B1** A 1-h window (−30 to 30 min) centred on Physiotherapy 1. **B2** A magnified view of the same event (20-min window), illustrating more prominent manipulation-related artefacts that are often longer in duration. Traces include ABP (red), ECG (green) and ICP (white). *Y*-axis scales are not standardised across panels, as applied during screening. Signal gaps reflect segments removed during CENTER-TBI data curation when values, or their variability, were deemed non-physiological. ABP, arterial blood pressure; ECG, electrocardiogram; ICP, intracranial pressure

(ICH GCP; CPMP/ICH/135/95) and the World Medical Association Declaration of Helsinki entitled "Ethical Principles for Medical Research Involving Human Subjects." Informed consent by the patients and/or the legal representative/next of kin was obtained, according to the local legislations, for all patients recruited in the Core Dataset of CENTER-TBI and documented in the electronic case report form (e-CRF). Ethical approval was obtained for each recruiting site. The list of sites, ethical committees, approval numbers and approval dates can be found on the following website: https://www.center-tbi.eu/project/ethical-approval.

Data Collection and Monitoring Infrastructure

Data for the CENTER-TBI study has been collected through the Quesgen e-CRF (Quesgen Systems Inc, USA), hosted on the International Neuroinformatics Coordinating Facility (INCF) platform and Leiden University Medical Center (LUMC) Opal platform and extracted via the INCF Neurobot tool (INCF, Sweden)/LUMC Opal platform (LUMC, the Netherlands) Version 3.1 of the CENTER-TBI dataset was used in this manuscript.

For patient monitoring and data collection in the high-resolution repository, the Intensive Care Monitoring (ICM+) platform software (Cambridge Enterprise Ltd., Cambridge, UK) and/or Moberg Central Nervous System (CNS) Monitor (Moberg Research Inc., now part of Natus Medical, Middleton, WI, USA) were used.

High-frequency signals—including ICP, ABP, ECG and peripheral oxygen saturation (SpO_2) and, where possible, temperature (core, external or intracranial), brain tissue oxygen tension ($PbtO_2$), central venous pressure (CVP) and end-tidal carbon dioxide ($EtCO_2$)—were continuously recorded throughout the ICU stay. All signals were sampled at ≥ 100 Hz using either ICM+ software or the Moberg CNS Monitor. Details on signal acquisition have been previously described by Zeiler et al. [10]. To promote consistent data entry across centres, a minimum target of seven annotations per patient was

recommended as part of the data collection guidelines and used as an incentive for data completeness, rather than as a strict inclusion criterion. The implementation was left to the discretion of each centre. All the centres, except for one, were equipped with a touch-screen laptop running ICM+, which either served as the primary data collection solution or as a downstream processor when CNS Monitor was used for that purpose. This set-up ensured a consistent and standardised platform for event entry, using a purpose-designed annotation form displayed within the ICM+ screen (Fig. 1). Nine categories of events were available for annotation, including inotropes/vasoconstrictors increase or decrease, fluid bolus, osmotherapy (mannitol or hypertonic saline), sedation bolus/analgesia, sedation/analgesia increase or decrease, suctioning/airway procedures and patient movement, turning or washing. Throughout this manuscript, the term "physiotherapy" is used to refer collectively to this group of patient handling and mobilisation activities. In the design process of this form, a balance was sought between what the bedside staff was being asked to do and the expectation of adherence to the data annotation requirements according to the principle "ask for too little or too much and end up with too little regardless". Thus, the nine categories were carefully selected to ensure a sufficient number of meaningful annotations with minimal effort required from the bedside staff. For the same reason, the specific amount of medication/fluids administered was not asked for, but merely the action of medication changes was annotated. Each annotation was time-stamped and stored alongside the physiological recordings. A typical example of an annotated recording is shown in Supplementary Fig. S1.

Data (Pre-)Processing

Data pre-processing followed a previously described procedure [10]. In the data curation phase of the CENTER-TBI project, the collected annotated monitoring data were compressed into HDF5 after a preliminary manual artefact removal process. This step involved visual inspection and exclusion of periods where the values of individual variables did not represent valid physiological measurements (e.g. arterial line failures). In addition, an automated artefact removal algorithm was applied for the ABP and ICP signals on the basis of predefined thresholds. For ABP, only values within 0–300 mm Hg with a minimum pulse amplitude of 20 mm Hg were considered valid. For ICP, the valid range was −30 to 200 mm Hg, with a minimum pulse amplitude threshold of 0.04 mm Hg [11]. All signals were then downsampled to 1-min resolution.

At the annotation level, interventions recorded outside the periods represented in the available high-resolution monitoring files were excluded, as they could not be visualised or verified against physiological data. This occurred because, during central data curation prior to HDF5 compression, certain data segments were removed—for instance, sections affected by artefactual signals, changes in the ICP monitoring source or temporary data loss due to cable disconnection or network interruptions. A descriptive analysis was then performed to characterise the distribution of the retained annotations.

Definition of Annotation Plausibility

Annotation plausibility was defined as the temporal and physiological consistency between a recorded annotation time stamp and an expected signal change attributable to the annotated intervention, assessed within a prespecified tolerance window around the time stamp. Operational definitions used for visual screening are provided in the Supplementary Material.

Step 1: Visual Inspection Approach

To assess the plausibility of the annotations, each patient's high-frequency data file was visually inspected by one reviewer (S.T.). The assessment focused on suctioning and physiotherapy annotations. A temporal imprecision of up to ±20 min was tolerated. Files in which suctioning and physiotherapy annotations consistently showed plausible physiological signature were classified *high evidence* (representative examples are shown in Fig. 2), files in which plausible annotations were rare or absent were classified as *low evidence*, and *moderate evidence* otherwise. Patient files with fewer than ten total clinical annotations (across all nine annotation categories) were also classified as *low evidence*. This cut-off was pragmatic and informed by the CENTER-TBI guidance recommending a minimum target of seven annotations per patient.

All files subsequently underwent an independent review by a second reviewer (J.T.) using a complementary assessment strategy focused on identifying recordings with insufficient evidence supporting annotation plausibility. Discrepant classifications were resolved through consensus adjudication or, when necessary, by discussion with a third reviewer (P.S.). In addition, an independent event-level review of suctioning annotations within the high-evidence files was performed by a fourth reviewer (R.T.). Detailed description of the methods is available in the Supplementary Material. The operational definitions of the three evidence categories, together with the corresponding review process, are summarised in Supplementary Table S1.

This classification was used to stratify the cohort prior to the application of the automated analysis; however, the event-level algorithm was applied to all

osmotherapy annotations independently of the Step 1 file-level categories.

Step 2: Automated Classification of Osmotherapy Annotations

Data from the ICM+ [12] software were automatically exported in CSV format on the basis of the events mark-up, including up to 30 min before and 90 min after each osmotherapy event. Data were summarised, with the mean, minimum, maximum and standard deviation computed for each variable. In addition, ICP features were derived, including the percentage of time with ICP > 20 mm Hg, and an indicator of sustained intracranial hypertension, defined as ICP > 20 mm Hg for at least 5 min consecutively [13]. A 1-min tolerance buffer was applied to account for transient drops below the threshold. Subsequent data processing and statistical analyses were performed in R (version 4.4.2; R Foundation for Statistical Computing, Vienna, Austria).

An ICP-based classification algorithm was developed to evaluate the plausibility of osmotherapy annotations. The ICP threshold of 20 mm Hg was chosen to reflect contemporary clinical practice and guideline recommendations during the start of the CENTER-TBI recruitment period [14]. To account for temporal imprecision, a 40-min baseline window (−30 to +10 min around the time stamp) was analysed and divided into four consecutive 10-min segments. Each segment was assessed for sustained intracranial hypertension (ICP ≥ 20 mm Hg for at least 5 min). If multiple segments met this criterion, the baseline was defined using a segment with the greatest proportion of time spent above the ICP threshold; ties were resolved by selecting the segment with the highest mean ICP. Osmotherapy annotations without any valid baseline were classified as *rejected*.

Step 3: Post-intervention Classification

For osmotherapy annotations with a valid baseline, the post-intervention window (+10 to +90 min) was analysed to determine whether the intervention produced a relevant physiological effect. They were categorised as:

- *Effective*: absence of any episode of sustained intracranial hypertension (ICP > 20 mm Hg for ≥ 5 min consecutively) within 90 min after the intervention, and/or a mean ICP decrease of ≥ 10 mm Hg compared with baseline [13].
- *Ineffective*: one or more episodes of sustained intracranial hypertension (ICP > 20 mm Hg for ≥ 5 min) occurred within 90 min post-intervention, and the mean ICP reduction was < 10 mm Hg.

For annotations without a valid baseline (*rejected* events), ICP trajectories in the post-intervention window were also examined to confirm that values remained below the treatment threshold.

Statistical Analysis

Data are reported as frequencies and percentages. Annotation counts and their distributions were quantified by category and examined across patients, over time (ICU days and time of day), and in relation to the 6-month Extended Glasgow Outcome Scale (GOSE) score.

Pearson's Chi-squared test was used to assess whether the distribution of osmotherapy event classifications (*effective, ineffective, rejected*; Steps 2–3) differed according to the file-level evidence classification obtained from visual screening (*high, moderate, low evidence*; Step 1). Agreement for the identification of low-evidence recordings in the independent file-level review was quantified using Cohen's kappa. Spearman's rank correlation test was used to evaluate the association between the total number of annotations per patient and functional outcome (GOSE score at 6 months). Sensitivity analyses were performed using alternative ICP thresholds of 15 and 25 mm Hg. Additional descriptive analyses, including comparisons between patients monitored with intraparenchymal and EVD, analyses of recorded co-interventions and haemodynamic trends in effective and ineffective events were reported in the Supplementary Material. Two-sided p-values < 0.05 were considered statistically significant.

Results

Study Cohort

Of the 277 patients included in the CENTER-TBI high-resolution database, 37 were excluded due to the use of EVD. Of the remaining 240 patients, 210 had at least one annotated intervention. Five of these patients were further excluded because they did not have the minimal required signals (ICP, ABP and ECG), resulting in a final cohort of 205 patients. Osmotherapy annotations were present in 76 of these 205 patients (37.1%).

Clinical characteristics of the included intraparenchymal ICP cohort and the excluded EVD cohort are compared in Supplementary Table S2, showing broadly comparable baseline demographic, injury severity and outcome characteristics.

Distribution of Annotations

A total of 16,023 annotations were originally recorded in the 205 included patients. Most annotations belonged to the nine predefined intervention categories established in the CENTER-TBI dataset. However, 555 fell outside the periods covered by available physiological monitoring,

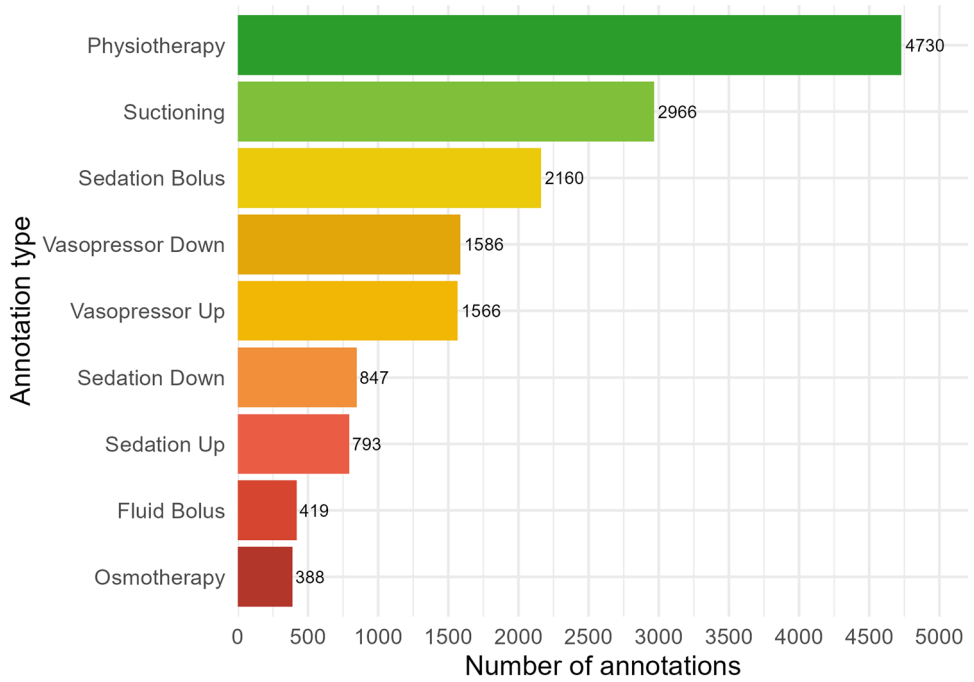

Fig. 3 Distribution of annotated interventions in the CENTER-TBI high-resolution dataset (n = 15,455). The bar chart shows the total number of annotations by type. CENTER-TBI, Collaborative European NeuroTrauma Effectiveness Research in Traumatic Brain Injury

and 13 belonged to additional categories ("EVD open" or "craniectomy") that were not included in the present analysis. After these exclusions, 15,455 annotations were analysed.

Physiotherapy (n = 4730, 30.6%) and suctioning (n = 2966, 19.2%) were the most frequent events, while osmotherapy (n = 388, 2.5%) was the least frequent (Fig. 3). The number of annotations per patient was highly variable (Supplementary Fig. S4). The mean number of annotations per patient per day was highest during the first 24 h and decreased gradually thereafter, remaining above 10 per patient per day during the first 120 h of monitoring (Supplementary Fig. S5). Among osmotherapy annotations, the proportion of rejected (classified in Step 2) was higher during the first ICU days, when more annotations were available, and tended to be lower in later epochs (Supplementary Fig. S6). The hourly distribution was relatively constant across the 24-h period, except for annotations indicating a reduction in sedation, which peaked in the morning (Supplementary Fig. S7).

Total annotation count was not correlated with 6-month GOSE score in the subset with available outcome data (n = 156; Spearman's $\rho = 0.09$, $p = 0.27$) Annotation density, defined as total number of annotations per patient per day of monitoring, also showed no significant association with 6-month GOSE score (Spearman's $\rho = 0.14$, $p = 0.08$; Supplementary Fig. S8).

Step 1: Visual Inspection Classification

Based on the visual inspection of physiotherapy and suctioning events, 123 patient files (60.0%) were classified as *high evidence*, 62 (30.2%) as *moderate evidence* and 20 (9.8%) as *low evidence*. The 388 osmotherapy events were distributed as follows: 161 events (41.5%) in *high-evidence* files, 145 (37.4%) in *moderate-evidence* files and 82 (21.1%) in *low-evidence* files. The complete lists of files assigned to each category are provided in Supplementary Table S3–5.

Agreement Between Reviewers and Adjudication

Prior to adjudication, agreement between reviewers for the identification of low-evidence files versus all other files was high (observed agreement 94.6%), with a Cohen's κ of 0.69 indicating substantial agreement after accounting for chance. Initial agreement between reviewers was observed for 194 of 205 recordings (94.6%), whereas 11 recordings received discrepant classifications. Of these, only two required adjudication by a third reviewer (P.S.). Following the consensus review, only 7 of 205 recordings (3.4%) were reclassified between the low- and moderate-evidence categories. For recordings classified as low

evidence, the principal reasons underlying the classification are reported in the Supplementary Table S6.

Event-Level Assessment within the High-Evidence Files

Of the 1769 suctioning annotations reviewed within the 123 files classified as high evidence, 1615 (91.3%) were considered physiologically plausible and 154 (8.7%) implausible.

Step 2: Baseline Classification of Osmotherapy Annotations

Of the 388 osmotherapy events, 248 (63.9%) met the baseline criterion of sustained intracranial hypertension and were included in the automated evaluation. The remaining 140 (36.1%) were rejected as invalid events. *Rejected* events showed ICP values stably below 20 mm Hg in both the pre- and post-annotation periods (Fig. 4).

Among valid events, the baseline period coincided with the 10 min immediately preceding the annotation in 35.9% of cases. In almost half of the cases (49.2%), however, the baseline was identified earlier (10–30 min before the annotation). In the remaining 14.9% of cases, the baseline occurred within 10 min after the annotation.

Step 3: Post-intervention Classification

For the 248 osmotherapy events with a valid baseline, the ICP trend in the post-annotation window (+10 to +90 min) was analysed. Events were categorised as *effective* (168, 67.7%) if ICP decreased by ≥ 10 mm Hg or stayed consistently ≤ 20 mm Hg, or *ineffective* (80, 32.3%) if ICP remained elevated without significant reduction (Fig. 4). Additional analyses of recorded co-interventions and mean ABP trends did not suggest systematic differences between effective and ineffective events (Supplementary Figs. S9–10).

Association Between Visual and Automated Annotation Classification

When osmotherapy events were stratified according to the file-level evidence classification obtained by visual screening, the proportion of *rejected* events was 18.0% among events from *high-evidence* files, 44.1% among events from *moderate-evidence* files and 57.3% from *low-evidence* files ($\chi^2 = 44.96$, $p < 0.0001$) (Fig. 5).

Sensitivity analyses using alternative baseline ICP thresholds of 15 mm Hg and 25 mm Hg yielded similar findings. The association between file-level evidence classification and physiology-based event classification remained statistically significant across all thresholds examined (ICP ≥ 15 mm Hg: $\chi^2 = 25.68$, $p < 0.001$; ICP ≥ 25 mm Hg: $\chi^2 = 53.93$, $p < 0.001$). The corresponding analyses are shown in Supplementary Figs. S11–S12.

Discussion

In this study, we developed and applied a three-step methodological framework to evaluate the plausibility of time-stamped clinical annotations in the CENTER-TBI high-resolution dataset, using osmotherapy as a

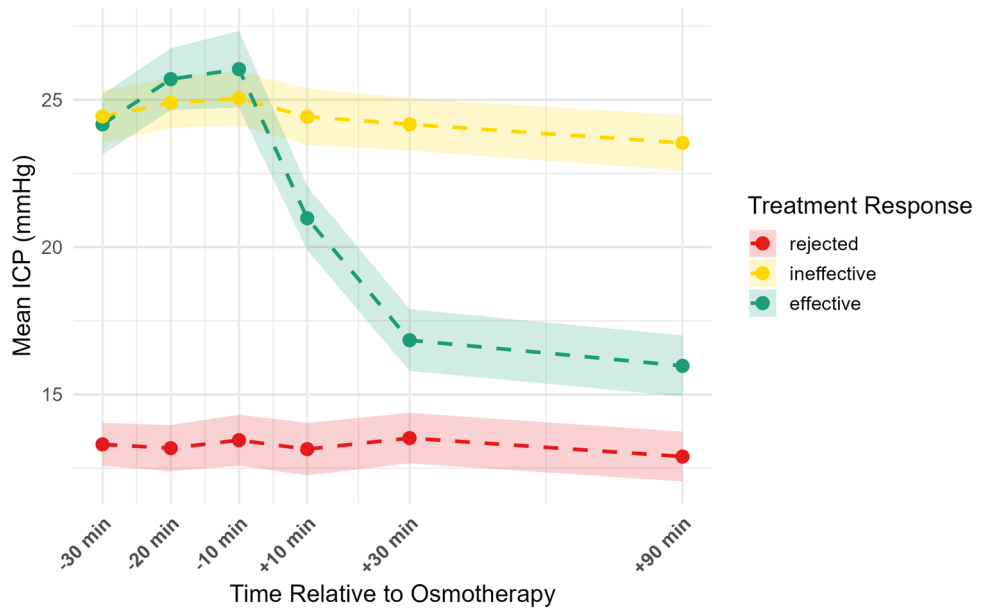

Fig. 4 Mean ICP trend before and after osmotherapy, stratified by algorithmic classification (effective, ineffective, rejected). Shaded bands indicate 95% confidence intervals. ICP, intracranial pressure

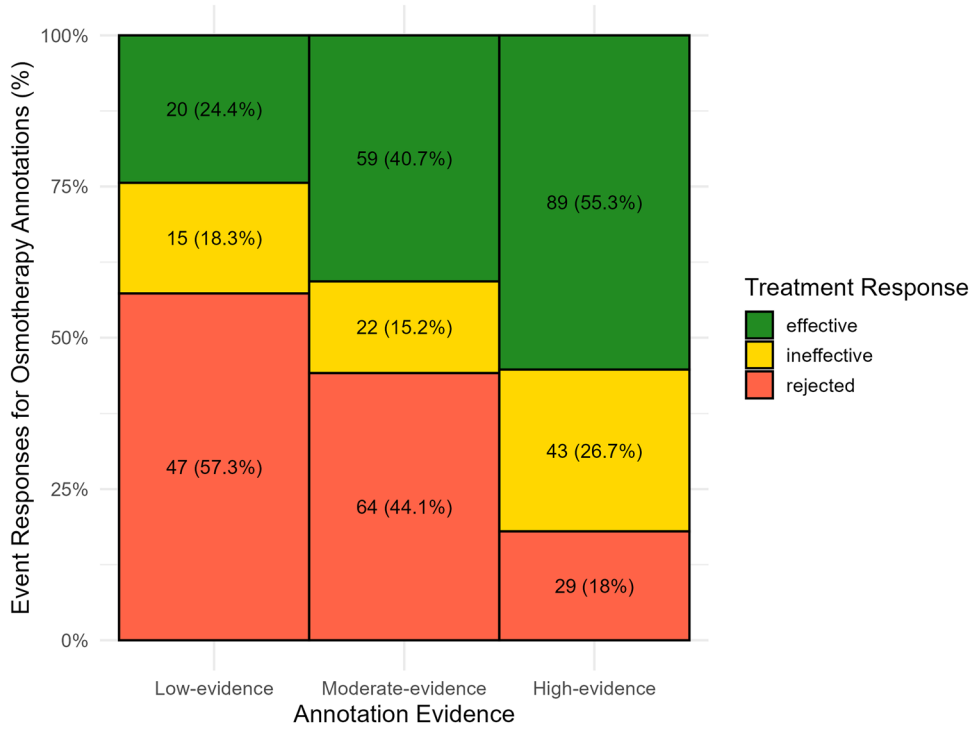

Fig. 5 Distribution of event responses for osmotherapy annotations classification (*effective, ineffective, rejected,* Steps 2–3) according to the level of annotation evidence assigned to each file during visual screening (Step 1)

representative intervention. The framework integrates a structured visual inspection of reference events with an automated, physiology-based classification to assess the internal consistency and physiological plausibility of annotations in the absence of an external ground truth. The agreement between the two independent strategies supports the complementary role of visual and automated approaches: visual inspection provides a qualitative assessment of annotation practices at the patient level, while automated classification offers an objective, physiology-based validation at the event level. Together, they provide complementary perspectives that support a more consistent and physiologically grounded interpretation of clinical annotations.

The relevance of clinical annotations lies in their role as contextual information linking physiological data to clinical meaning, which is essential for correctly interpreting high-resolution monitoring data [15]. Accurate and temporally aligned annotations provide this necessary context, allowing physiological trends to be understood in relation to clinical interventions rather than as isolated signal fluctuations. Supplementary Fig. S13 illustrates this concept, showing how clinical annotations clarify the meaning of abrupt variations in physiological signals.

However, despite their critical importance for data interpretation, the reliability of such annotations remains a major challenge [15]. Wac et al. [16] highlighted the lack of temporal precision, consistency and standardisation in manually annotated interventions, which can compromise downstream analyses and predictive modelling. Further work by the same group identified cognitive overload, fragmented workflows, unclear task responsibilities and insufficient training as key factors contributing to annotation inaccuracy [3]. Our descriptive analysis reflected these concerns, showing substantial variability in annotation practices. This heterogeneity reinforces the need for systematic evidence assessment of any manual annotations and for further efforts to create automated annotation systems.

Our approach was designed to accommodate temporal imprecision and timing errors, which are frequently observed in clinical annotations [4]. We allowed a temporal tolerance of up to ±20 min during visual screening and considered multiple baseline windows in the automated algorithm. This pragmatic strategy proved necessary, as our findings showed that, in the majority of the valid osmotherapy events, the selected baseline segment did not immediately precede the annotation time stamp but was located earlier or, in some cases, even after it.

This approach helped prevent misclassification due to inaccuracies resulting from post-factum (retrospective) annotations that were allowed in the CENTER-TBI high-resolution ICU data collection protocols to minimise additional burden imposed on bedside staff during demanding periods.

Suctioning and physiotherapy were selected as reference annotations due to their frequent occurrence and well-characterised physiological signatures. This made them suitable for a file-level visual screen of whether annotations were consistent with an expected physiological signature within the predefined tolerance window. Because these events provide repeated opportunities for verification, their consistency can be taken as evidence of an overall degree of compliance with the annotation protocol by the bedside staff. Although, in principle, annotation practices could vary across nursing shifts, in our visual inspection, we did not observe systematic differences when comparing annotations recorded at different times of day. Moreover, shift structures were not standardised across centres, making shift-level evaluation difficult to interpret. For these reasons, the annotations were assessed at the patient-file level, which provided a pragmatic and generalisable summary of within-file annotation consistency. The substantial agreement observed between reviewers suggests that recordings with insufficient evidence supporting annotation plausibility can be identified consistently despite the use of different assessment approaches. Furthermore, the high proportion of physiologically plausible annotations observed during the independent event-level review supports the interpretation that recordings classified as high-evidence predominantly contained valid clinical annotations. Although time-consuming, this approach allows for the injection of a degree of trust in interventions where algorithmic validation is not feasible, such as those without clearly interpretable baseline and post-intervention periods.

Automated artefact detection tools could perhaps assist in the process of validating such annotations, which are clearly reflected in physiological signal distortions, but current implementations struggle to reliably detect patient handling interventions due to physiological complexity [17]. The relatively high rate of automated misclassification in this case would translate to an artificial, significant decrease in the overall annotation quality metric, leading to an overly reduced size of the dataset selected for further analysis. These limitations supported our decision to use a structured visual assessment tailored to the specific characteristics of the data.

For osmotherapy, annotation plausibility could also be assessed at the event level because treatment context can be defined using prespecified physiological criteria. This automated strategy is applicable to selected interventions such as osmotherapy or vasopressor administration, where physiological thresholds can be reasonably defined [14, 18]. Step 2 evaluates whether each osmotherapy annotation is temporally compatible with sustained intracranial hypertension. A conceptually similar approach was proposed by Bögli et al. [13], who evaluated the effectiveness of osmotherapy in paediatric TBI based on predefined ICP thresholds. Our framework conservatively excluded osmotherapy events solely on the basis of baseline plausibility, as it is impossible to distinguish between inaccurate annotation and a true non-responder on the basis of the treatment response only. This approach minimises the risk of discarding valid but ineffective treatments. Most events classified as *ineffective* came from *well-annotated* files and showed prolonged elevated ICP, suggesting that they were true *non-responders.* By contrast, *rejected* events typically had consistently low ICP values, as shown in Fig. 4, reinforcing that they represent implausible events. Within this framework, treatment response is therefore used to differentiate plausible non-responders from annotations without a physiologically credible treatment context, rather than to estimate the clinical efficacy of osmotherapy. However, substantial variation in treatment policies across CENTER-TBI centres cannot be excluded, and in some units, osmotherapy may have been used as a preventive measure rather than in response to intracranial hypertension, despite recommendations from the contemporaneous Brain Trauma Foundation Guidelines [14, 18].

To our knowledge, this is the first study to provide a systematic audit of time-stamped clinical annotations in the CENTER-TBI high-resolution dataset. Our findings support combining visual screening with physiological classification as a pragmatic strategy. Visual inspection identifies problematic data files, while automated classification offers an objective tool to assess the plausibility of individual annotations. The agreement between these two independent methods offers evidence of internal consistency.

While this study presents a practical framework for addressing the problem of inaccurate manual annotations of ICU high-resolution physiological recordings, several limitations should be acknowledged. First, although additional independent file-level and event-level reviews were performed, the three-level visual classification was not independently replicated in its entirety. In particular, the additional file-level review (reviewer 2) primarily focused on the identification of low-evidence

recordings, whereas the event-level review (reviewer 4) assessed the plausibility of annotations within high-evidence files. Consequently, the distinction between high- and moderate-evidence recordings retains an element of subjectivity and may depend on the interpretation of borderline cases. Occasional false positives cannot be entirely excluded, as concurrent procedures may generate similar artefacts within the temporal tolerance window. Second, the framework cannot detect missing annotations (interventions performed but never documented) and therefore cannot quantify under-documentation. Third the algorithmic definition of treatment response – based on a ≥ 10 mm Hg ICP reduction or ICP normalisation – was adapted from prior literature [13] and retains a degree of arbitrariness. Fourth, patients monitored exclusively with EVDs were excluded because ICP monitoring through an EVD is discontinuous during drain-open periods. Although recent work has shown that drain-state can be identified automatically from ICP waveforms, enabling selection of analysable epochs [19], implementing and validating such approaches was beyond the scope of the present study. This exclusion may have introduced selection bias, although clinical characteristics of patients with intraparenchymal and EVD monitoring were broadly comparable (Supplementary Table S2). Notably, the included cohort was not restricted to patients without CSF diversion, as intraparenchymal ICP monitoring could be combined with EVD treatment when clinically indicated. Fifth, concurrent adjustments in sedation, vasopressor support or patient positioning could have influenced ICP trends and therefore represent a potential source of confounding. However, additional analyses of recorded co-interventions and mean ABP trajectories did not suggest systematic differences between effective and ineffective osmotherapy events (Supplementary Figures S9–S10). Sixth, the decision to classify recordings with fewer than ten total annotations as *low-evidence* was pragmatic, but informed by CENTER-TBI guidance (minimum requirement of seven annotations per recording). Because a minimum count can be met even with clustered annotations (e.g. added to meet the minimum target), we used a slightly higher cut-off as a conservative buffer. Alternative thresholds could yield different file-level categorisations. Finally, the automated classification focused solely on osmotherapy events, and the framework was evaluated using a single dataset (CENTER-TBI). Future research is necessary to assess its generalisability across different datasets and patient populations.

It should be emphasised, however, that this framework represents a pragmatic solution to a broader structural problem that requires attention at the point of data collection. Prospective ICU data collection should aim for greater detail, including systematic, automated recording of medication type, dose and their precise timing, as well as all the nursing procedures. This is particularly important in the current context, where data-driven artificial intelligence (AI) algorithms are increasingly used to support clinical decision-making, including in ICU environments. Until such data collection practices become standard, manual annotations will remain the main source of intervention records, underscoring the need for strategies to assess them.

Conclusions

We proposed a pragmatic three-step framework to assess the physiological plausibility and internal consistency of time-stamped clinical annotations in high-resolution neurocritical care data. By combining visual inspection of reference annotations with an automated, physiology-based classification of osmotherapy events, the framework enables identification of annotations that are physiologically implausible. While our framework was tailored to osmotherapy, the same principles may be generalisable to other interventions with physiologically grounded criteria, thereby supporting more robust and interpretable event-based analyses.

Supplementary Information
The online version contains supplementary material available at https://doi.org/10.1007/s12028-026-02637-6.

Author details
[1] Department of Neurology, Kepler University Hospital, Johannes Kepler University Linz, Linz, Austria. [2] Clinical Research Institute for Neuroscience, Johannes Kepler University Linz, Linz, Austria. [3] Brain Physics Laboratory, Division of Neurosurgery, Department of Clinical Neurosciences, University of Cambridge, Cambridge, UK. [4] Department of Neurology and Neurocritical Care Unit, Clinical Neuroscience Center, University Hospital Zurich, University of Zurich, Zurich, Switzerland. [5] Neuroscience Intensive Care Unit, Department of Anesthesia and Critical Care, Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Milan, Italy. [6] Department of Pathophysiology and Transplantation, University of Milan, Milan, Italy.

Acknowledgments
The authors thank Rebecca Tesch, medical student at Johannes Kepler University (JKU), Linz, Austria, for performing the independent event-level review and contributing to the independent validation of the visual classification. Data used in preparation of this manuscript were obtained in the context of CENTER-TBI, a large collaborative project with the support of the European Union 7th Framework program (EC grant 602150). Additional funding was obtained from the Hannelore Kohl Stiftung (Germany), from OneMind (USA) and from Integra LifeSciences Corporation (USA).

CENTER-TBI High Resolution Sub-Study Participants and Investigators: Audny Anke, Department of Physical Medicine and Rehabilitation, University Hospital Northern Norway; Ronny Beer, Raimund Helbok, Department of Neurology, Neurological Intensive Care Unit, Medical University of Innsbruck, Innsbruck, Austria; Bo-Michael Bellander, David Nelson, Department of Neurosurgery & Anesthesia & Intensive Care Medicine, Karolinska University Hospital, Stockholm, Sweden; Andras Buki, Department of Neurosurgery,

University of Pecs and MTA-PTE Clinical Neuroscience MR Research Group and Janos Szentagothai Research Centre, University of Pecs, Hungarian Brain Research Program, Pecs, Hungary; Giorgio Chevallard, Arturo Chieregato, NeuroIntensive Care, Niguarda Hospital, Milan, Italy; Giuseppe Citerio, NeuroIntensive Care Unit, Department of Anesthesia & Intensive Care, ASST di Monza, Monza, Italy and School of Medicine and Surgery, Università Milano Bicocca, Milano, Italy; Endre Czeiter, Department of Neurosurgery, University of Pecs and MTA-PTE Clinical Neuroscience MR Research Group and Janos Szentagothai Research Centre, University of Pecs, Hungarian Brain Research Program (Grant No. KTIA 13 NAP-A-II/8), Pecs, Hungary; Bart Depreitere, Department of Neurosurgery, University Hospitals Leuven, Leuven, Belgium; George Eapen†; Shirin Frisvold, Department of Anesthesiology and Intensive care, University Hospital Northern Norway, Tromso, Norway; Stefan Jankowski, Neurointensive Care, Sheffield Teaching Hospitals National Health Service (NHS) Foundation Trust, Sheffield, UK; Daniel Kondziella, Departments of Neurology, Clinical Neurophysiology and Neuroanesthesiology, Region Hovedstaden Rigshospitalet, Copenhagen, Denmark; Lars-Owe Koskinen Department of Clinical Neuroscience, Neurosurgery, Umea University, Umea, Sweden; Geert Meyfroidt, Intensive Care Medicine, University Hospitals Leuven, Leuven, Belgium; Kirsten Moeller, Department Neuroanesthesiology, Region Hovedstaden Rigshospitalet, Copenhagen, Denmark; Anna Piippo-Karjalainen, Rahul Raj, Helsinki University Central Hospital, Helsinki, Finland; Andreea Radoi, Juan Sahuquillo, Department of Neurosurgery, Vall d'Hebron University Hospital, Barcelona, Spain; Arminas Ragauskas, Saulius Rocka, Department of Neurosurgery, Kaunas University of Technology and Vilnius University, Vilnius, Lithuania; Jonathan Rhodes, Department of Anaesthesia, Critical Care & Pain Medicine NHS Lothian & University of Edinburg, Edinburgh, UK; Rolf Rossaint, Ana Stevanovic, Department of Anaesthesiology, University Hospital of Aachen, Aachen, Germany; Oliver Sakowitz, Klinik für Neurochirurgie, Klinikum Ludwigsburg, Ludwigsburg, Germany and Department of Neurosurgery, University Hospital Heidelberg, Heidelberg, Germany; Nina Sundström, Department of Radiation Sciences, Biomedical Engineering, Umea University, Umea, Sweden; Riikka Takala, Perioperative Services, Intensive Care Medicine and Pain Management, Turku University Central Hospital and University of Turku, Turku, Finland; Tomas Tamosuitis, Neuro-intensive Care Unit, Kaunas University of Health Sciences, Kaunas, Lithuania; Olli Tenovuo, Rehabilitation and Brain Trauma, Turku University Central Hospital and University of Turku, Turku, Finland; Peter Vajkoczy, Neurologie, Neurochirurgie und Psychiatrie, Charité—Universitätsmedizin Berlin, Berlin, Germany; Alessia Vargiolu, NeuroIntensive Care Unit, Department of Anesthesia & Intensive Care, ASST di Monza, Monza, Italy; Rimantas Vilcinis, Department of Neurosurgery, Kaunas University of Health Sciences, Kaunas, Lithuania; Stefan Wolf, Interdisciplinary Neuro Intensive Care Unit, Charité—Universitätsmedizin Berlin, Berlin, Germany; and Alexander Younsi, Department of Neurosurgery, University Hospital Heidelberg, Heidelberg, Germany.

SOPRANI collaborators: Alan Urban, Gabriel Montaldo and Eloïse Baud, Flemish Institute for Biotechnology (VIB), KU Leuven, Leuven, Belgium; Dick Moberg, Moberg Analytics, Inc. Philadelphia, USA; Geert Meyfroidt, Fadime Tokmak, Laboratory of Intensive Care Medicine, Department of Cellular and Molecular Medicine, KU Leuven, Leuven, Belgium; Bart Depreitere, Fabio Gonçalves and Elle Scheijen, Research Group Experimental Neurosurgery and Neuroanatomy, Department of Neurosciences, KU Leuven, Leuven, Belgium; Jens P. Dreier, Malaika Mohammad and Coline L. Lemale, Centre for Stroke Research Berlin, Charité—Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin and Berlin Institute of Health, Berlin, Germany; Samira Saadoun, Marios C. Papadopoulos and Chibuzor Love Ilochonwu, Department of Neurosurgery, St George's University Hospitals NHS Foundation Trust, London, UK; Peter Smielewski and Wenhao Xu, Brain Physics Laboratory, Department of Clinical Neurosciences, University of Cambridge, Cambridge, UK; Raimund Helbok and Sara Turella, Department of Neurology, Kepler University Hospital and Clinical Research Institute for Neuroscience, Johannes Kepler University Linz, Linz, Austria; Ewout W. Steyerberg, Nicole S. Erler, Eugenia Driusso and Aneeta Chacko, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, the Netherlands; Wilco Peul, Thomas van Essen, Jeroen van Dijck and Cansu Rehber, Department of Neurosurgery, University Neurosurgical Center Holland, Leiden University Medical Center, Haaglanden Medical Center, Haga Teaching Hospital, Leiden and The Hague, the Netherlands.

Author Contributions

ST, JT, EB, PS and RH contributed to the conception and design of the study. ST performed the data analysis. EB, SYB and IO contributed to data analysis. ST and JT contributed to data interpretation. EB, TZ, PS, BI and RH participated in data acquisition within the CENTER-TBI study. ST, JT, PS and RH drafted the manuscript. All authors critically revised the manuscript for important intellectual content and approved the final version.

Source of Support

Open access funding provided by Johannes Kepler University Linz. This study was funded by the European Union's Marie Skłodowska-Curie Actions (MSCA) programme under the Smart Neuromonitoring to Support Precision Medicine in Acute Central Nervous Injury (SOPRANI) project (Grant Agreement No. 101119916).

Conflicts of Interest

ICM+ is software licensed by Cambridge Enterprise Ltd. Peter Smielewski has a financial interest in a part of the licensing fee. On behalf of all authors, the corresponding author states that all other authors report no disclosures relevant to this manuscript.

Ethical Approval/Informed Consent

This study involved human participants only. All data were collected under the ethical framework of the CENTER-TBI study (EC grant 602150), which received ethical approval from the institutional review boards of all participating centres. The study was conducted in accordance with the ethical standards of the 1964 Declaration of Helsinki and its later amendments. Written informed consent was obtained from patients or their legal representatives in accordance with local regulations. This secondary retrospective analysis did not require additional ethical approval. For this type of study, formal consent is not required.

Artificial Intelligence

During the preparation of this manuscript, the authors used ChatGPT (OpenAI) for English-language editing and to improve the readability of the text. The authors reviewed and edited the resulting text and take full responsibility for the final content of the manuscript.

Open Access

This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 11 November 2025 Accepted: 4 August 2026
Published online: 03 September 2026

References

1. Tas J, Beqiri E, Van Kaam RC, et al. Targeting autoregulation-guided cerebral perfusion pressure after traumatic brain injury (COGiTATE): a feasibility randomized controlled clinical trial. J Neurotrauma. 2021;38(20):2790–800. https://doi.org/10.1089/NEU.2021.0197.
2. Carra G, Güiza F, Piper I, et al. Development and external validation of a machine learning model for the early prediction of doses of harmful intracranial pressure in patients with severe traumatic brain injury. J Neurotrauma. 2023;40(5–6):514–22. https://doi.org/10.1089/NEU.2022.0251.

3. Wac M, Santos-Rodriguez R, McWilliams C, Bourdeaux C. Capturing requirements for a data annotation tool for intensive care: experimental user-centered design study. JMIR Hum Factors. 2025. https://doi.org/10.2196/56880.
4. Goodwin AJ, Eytan D, Dixon W, et al. Timing errors and temporal uncertainty in clinical databases—a narrative review. Front Digit Health. 2022. https://doi.org/10.3389/FDGTH.2022.932599.
5. Sylolypavan A, Sleeman D, Wu H, Sim M. The impact of inconsistent human annotations on AI driven clinical decision making. NPJ Digit Med. 2023. https://doi.org/10.1038/S41746-023-00773-3.
6. Adams RJ, Marlin BM. Learning time series detection models from temporally imprecise labels. Proc Mach Learn Res. 2017;54:157.
7. Brown SM, Sorensen J, Lanspa MJ, et al. Multi-complexity measures of heart rate variability and the effect of vasopressor titration: a prospective cohort study of patients with septic shock. BMC Infect Dis. 2016. https://doi.org/10.1186/S12879-016-1896-1.
8. Maas AIR, Ercole A, De Keyser V, Menon DK, Steyerberg EW. Opportunities and challenges in high-quality contemporary data collection in traumatic brain injury: the CENTER-TBI experience. Neurocrit Care. 2022;37(Suppl 2):192–201. https://doi.org/10.1007/S12028-022-01471-W.
9. Maas AIR, Menon DK, Steyerberg EW, et al. Collaborative European neurotrauma effectiveness research in traumatic brain injury (CENTER-TBI): a prospective longitudinal observational study. Neurosurgery. 2015;76(1):67–80. https://doi.org/10.1227/NEU.0000000000000575.
10. Zeiler FA, Ercole A, Beqiri E, et al. Association between cerebrovascular reactivity monitoring and mortality is preserved when adjusting for baseline admission characteristics in adult traumatic brain injury: a CENTER-TBI study. J Neurotrauma. 2020;37(10):1233. https://doi.org/10.1089/NEU.2019.6808.
11. Bögli SY, Cucciolini G, Cherchi MS, et al. Feasibility and safety of integrating extended TCD assessments in a full multimodal neuromonitoring protocol after traumatic brain injury. Ultrasound Med Biol. 2024;50(11):1704–15. https://doi.org/10.1016/J.ULTRASMEDBIO.2024.07.009.
12. Smielewski P, Beqiri E, Mataczynski C, et al. Advanced neuromonitoring powered by ICM+ and its place in the brand new AI world, reflections at the 20th anniversary boundary. Brain Spine. 2024;4:102835. https://doi.org/10.1016/J.BAS.2024.102835.
13. Bögli SY, Olakorede I, Smith CA, et al. The value of dynamic cerebral compliance monitoring after pediatric traumatic brain injury: a STAR-SHIP study sub-analysis. Crit Care. 2025. https://doi.org/10.1186/S13054-025-05403-W.
14. Brain Trauma Foundation. Guidelines for the management of severe TBI, 3rd edn. Accessed Oct 15, 2025. https://braintrauma.org/coma/guidelines/severe-tbi-3rd-edition
15. Foreman B, Lissak IA, Kamireddi N, Moberg D, Rosenthal ES. Challenges and opportunities in multimodal monitoring and data analytics in traumatic brain injury. Curr Neurol Neurosci Rep. 2021. https://doi.org/10.1007/S11910-021-01098-Y.
16. Wac M, Santos-Rodriguez R, McWilliams C, Bourdeaux C. Evaluation of a data annotation platform for large, time-series datasets in intensive care: mixed methods study. 2024; https://arxiv.org/pdf/2410.16959.
17. Georgatzis K, Lal P, Hawthorne C, et al. Artefact in physiological data collected from patients with brain injury: quantifying the problem and providing a solution using a factorial switching linear dynamical systems approach. Acta Neurochir Suppl (Wien). 2016;122:301–5. https://doi.org/10.1007/978-3-319-22533-3_60.
18. Carney N, Totten AM, Eilly C, et al. Guidelines for the management of severe traumatic brain injury, fourth edition. Neurosurgery. 2017;80(10):6–15.
19. Mathur R, Yellapantula S, Cheng L, et al. Classification of intracranial pressure epochs using a novel machine learning framework. NPJ Digit Med. 2025;8(1):201. https://doi.org/10.1038/S41746-025-01612-3.