International Journal of Legal Medicine
https://doi.org/10.1007/s00414-026-03831-z

ORIGINAL ARTICLE

Medical findings in young patients with head trauma: How strong is the evidence for abuse?

Kent P. Hymel[1] · **Marjan Sjerps**[2] · **Peter Vergeer**[2]

Received: 12 September 2025 / Accepted: 29 April 2026

Abstract

Background and objective A Likelihood ratio (LR) is a numerical measure of evidential value. Our objective was to use LRs to express the patient-specific evidential values of medical findings that best differentiate abusive versus non-abusive head trauma (AHT). We hypothesized that the evidential values of patients' AHT-related medical findings would be highly variable.

Methods We analyzed existing, uniform, prospective, de-identified data regarding 973 acutely head-injured children < 3 years hospitalized for intensive care across 18 sites between 2011 and 2021; applied two different proxies for AHT and non-AHT ground truth; trained and validated statistical models that differentiate AHT versus non-AHT; and analyzed patient-specific LRs in a $\log_{10}$ (LLR) format that facilitated assessment of evidential values (where LLR values > 0 and < 0 supported hypotheses of AHT and non-AHT, respectively).

Results The two best performing statistical models revealed evidential (LLR) values for patient-specific, AHT-related medical findings that varied from modest (-1 to + 1) to relatively large (-2.5 to -1 and + 1 to + 3.5), and values that were misleading (AHT patients with LLR values indicative of non-AHT, and vice versa). A few non-AHT patients presented with misleading evidence that was moderately strong, with LLRs approaching + 3.

Conclusions Reasonable medical certainty of AHT and non-AHT can be enhanced or limited by the highly variable evidential values of patients' most discriminating medical findings. Physicians can use available LRs to inform their AHT-related diagnostic reasoning, opinions, and testimony.

Keywords Abusive head trauma · Child physical abuse · Medical findings · Evidential value · Likelihood ratios

Introduction

Physicians' decisions to confirm or exclude abusive head trauma (AHT) can be difficult. There is no diagnostic gold standard; presenting signs and symptoms are frequently nonspecific; witnessed or admitted acts of abuse are rare; caregivers may deny, minimize, change, or fabricate their accounts of the injury event; visible signs of head trauma and/or abuse may be absent; patients are frequently unable to describe their traumas; and perceptions of psychosocial risk can distort clinical objectivity.

Physicians' decisions to confirm or exclude AHT can also be highly consequential. A flawed decision to exclude AHT places victims at substantial risk for additional abuse when returned to their abusive caregiver(s) [1, 2]. Conversely, a flawed decision to confirm AHT can strain the doctor-parent relationship, prolong hospitalizations, increase costs, expose the child to additional risks (e.g., sedation, radiation, false positive results), and trigger the child's placement in foster care. In the extreme, caregivers falsely accused of AHT can lose parental rights and/or face criminal prosecution.

Decisions to suspend or restore parental rights and to criminally convict or exonerate suspected perpetrators of abuse are debated in legal settings, where judges and juries often assign considerable weight to the opinions of physicians. Physicians who testify in cases of suspected AHT bear a clear responsibility to assign and communicate an accurate assessment of the evidential value of their medical

✉ Marjan Sjerps
m.sjerps@nfi.nl

[1] Penn State College of Medicine, Hershey, PA, USA

[2] Netherlands Forensic Institute, Laan Van Ypenburg 6, 2497 GB The Hague, The Netherlands

Published online: 19 May 2026

findings. Studies that quantify the evidential value of AHT-related medical findings are very limited [3, 4].

A likelihood ratio (LR) is a measure of evidential value, often reported in forensic science [5]. In the context of this paper, it is the ratio of two probabilities: (1) the probability of specific clinical findings given that the head-injured child was abused, and (2) the probability of the same clinical findings given that the head-injured child was not abused. The LR can thus be interpreted as a numerical degree of support for the hypothesis of AHT versus the hypothesis of non-AHT.

There are many statistical models that can be used to calculate patient-specific LRs based on a set of medical observations. One can use a Bayesian network, random forest, or logistic regression model. The latter is well-known in the medical domain and has been used before to model AHT [3], so it may be easier for physicians to understand. Other important modelling decisions include dealing with missing observations and dealing with observations that have more than two outcome categories (multinomial variables).

Previous validation studies for AHT statistical models [6, 7] assessed how well the LR values discriminated between abused and non-abused children. However, it is also important to study calibration [8], because the evidential value of medical findings should not be over- or under-estimated. Studying both discrimination and calibration requires different plots and metrics than used previously [9].

We analyzed an existing, prospective, de-identified, clinical data set (N = 973) to train and test statistical models to differentiate AHT versus non-AHT. Results were compared with the PediBIRN-7 prediction rule [3, 6], which handled missing observations and multinomial variables differently. The comparison was based on an extended test dataset (N = 486) and assessed both discrimination and calibration. Our objective was to construct an explainable model for calculating patient-specific evidential values of medical findings that best differentiate AHT versus non-AHT, while not over- or under-estimating that evidential value.

We hypothesized that the evidential values of patients' AHT-related medical findings would be highly variable.

Methods

Overview

The existing, de-identified data used in this analysis were captured by Pediatric Brain Injury Research Network (PediBIRN) investigators in three, sequential, multicenter studies conducted across 18 North American pediatric intensive care units between February 2011 and March 2021 [7, 10, 11]. All three studies used the same inclusion and exclusion criteria and patient-related data forms. The Institutional Review Board at Penn State Health Hershey Medical Center determined that this secondary analysis was not human subject research.

Eligible patients were children under 3 years of age hospitalized for intensive care of acutely symptomatic, closed, traumatic, cranial or intracranial injuries confirmed on initial neuroimaging. Patients with preexisting brain abnormalities and patients whose head injuries resulted from collisions involving motor vehicle(s) were excluded. In all three studies, prospective study design facilitated capture of complete required data.

The data set

The aggregate PediBIRN data set contains uniform, clinical, historical, and radiological data regarding 987 patients. Fourteen patients with obvious data discrepancies were excluded, leaving 973 patients for analysis. The medical variables used in analysis are listed in Table 1. Because physicians were free to launch or forgo abuse evaluations based on their patient-specific assessments of abuse probability and cost versus benefit, some patients did not undergo skeletal survey and/or retinal examination. The data set includes variables selected previously for inclusion in the 'PediBIRN-7'—a validated, seven variable, clinical prediction rule that incorporates the positive or negative predictive contributions of patients' completed abuse evaluations to estimate AHT probability after abuse evaluation [3, 6].

Multinomial variables

Some variables (Table 1) had more than two outcome categories. We chose to deal with these variables such that the resulting model would incorporate all possible outcome categories explicitly. This differs from the PediBIRN-7 methodology, which recoded several multinomial outcome categories into yes/no variables. This recoding caused a potential loss of information, possibly resulting in less discriminating LRs.

Missing observations

Because physicians did not order every available AHT test for each head-injured child, the dataset has missing observations for variables describing the outcome of such tests. These missing observations were handled by imputation to derive the PediBIRN-7 prediction rule [3]. We elected instead to deal with missing test results by adding 'not ordered/completed' as a possible outcome category. Doing so allowed us to derive the evidential value of the 'observation' that the test was not ordered or completed. Both

Table 1 Variables included in the aggregate data set (N = 973) and considered for inclusion in statistical models

Description	Data type
Age at time of hospital admission (months)	Number
Any acute respiratory compromise[1]	Dichotomous
Any seizure(s)	Dichotomous
Any acute encephalopathy[2]	Multicategory
No acute encephalopathy observed	
Acute encephalopathy, resolved prior to admission	
Acute encephalopathy, resolved in <24 hours	
Acute encephalopathy, lasting >24 hours, without deterioration	
Acute encephalopathy, lasting >24 hours, with deterioration	
Any craniofacial bruising, abrasion(s), subgaleal hematoma(s), or cephalohematoma(s)	Dichotomous
Any bruising of the torso, ear(s), or neck	Dichotomous
Any skull fracture(s)	Multicategory
No skull fracture(s) observed	
An isolated, unilateral, nondiastatic, linear, parietal skull fracture	
Any skull fracture(s) other than an isolated, unilateral, non-diastatic, linear, parietal fracture	
Any epidural hemorrhage(s)	Dichotomous
Any subdural hemorrhage(s) or collection(s)	Multicategory
No subdural hemorrhage(s) or collection(s) observed	
Subdural hemorrhage(s) or collection(s) that are unilateral	
Subdural hemorrhage(s) or collection(s) that are bilateral	
Subdural hemorrhage(s) or collection(s) that are interhemispheric	
Any subarachnoid hemorrhage(s)	Dichotomous
Any brain contusion(s), laceration(s), or hemorrhage(s)	Multicategory
No brain contusion(s), laceration(s), or hemorrhage(s) observed	
Focal, cortical, brain contusion(s), laceration(s), or hemorrhage(s)	
Brain contusion(s), laceration(s), or hemorrhage(s) involving the subcortical (or deeper) brain	
Brain contusion(s), laceration(s), or hemorrhage(s) characterized as diffuse axonal injury	
Any brain hypoxia, ischemia, or swelling	Multicategory
No brain hypoxia, ischemia, or swelling observed	
Brain hypoxia, ischemia, or swelling that is unilateral and limited to the cortical brain	
Brain hypoxia, ischemia, or swelling that is bilateral or involves the subcortical (or deeper) brain	
AST and/or ALT >80 IU/L any time after hospital admission	Dichotomous
Skeletal survey	Multicategory
Skeletal survey not ordered or not completed	
Skeletal survey completed	
Skeletal survey completed and revealed fracture(s) moderately or highly specific for abuse[3]	
Skeletal survey completed and failed to reveal fracture(s) moderately or highly specific for abuse[3]	
Retinal exam	Multicategory
Retinal exam by an ophthalmologist not ordered or not completed	
Retinal exam by an ophthalmologist completed	
Retinal exam by an ophthalmologist completed and revealed retinoschisis	
Retinal exam by an ophthalmologist completed and revealed dense, extensive, retinal hemorrhages extending to the ora serrata	
Retinal exam by an ophthalmologist completed and failed to reveal retinoschisis or dense, extensive, retinal hemorrhages extending to the ora serrata	

Abbreviations: AHT=abusive head trauma, ALT=alanine transaminase, AST=aspartate transaminase, IU/L=international units per liter, PediBIRN=pediatric brain injury research network

[1]Defined a priori as "infrequent or labored respirations, apnea, or any requirement for intubation or assisted ventilation"

[2]Defined a priori as "an initial, clear, impairment or loss of consciousness at the scene of injury, during transport, in the Emergency Department, or prior to hospital admission"

[3]Including classic metaphyseal lesion fractures or epiphyseal separations, rib fractures, fractures of the scapula or sternum, fractures of digits, vertebral body fractures, dislocations or fractures of spinous processes

options for dealing with missing data—imputation and the use of multinomial variables—introduce confounding with other observations that cannot be avoided.

Patient labeling

To label patients as AHT, non-AHT, or indeterminate in the absence of gold standards, ground truth was estimated in two ways: (1) applying the a priori definitional criteria (Table 2) used to derive and validate the PediBIRN-7 prediction rule [3, 6] and (2) applying physicians' final consensus diagnoses.

Train-test design

The data set (N = 973) was split into two parts. The first was used to train logistic regression models to differentiate AHT versus non-AHT. The second part was used to test the models. The training data set included data captured during the network's first two multicenter studies (n = 487) [7, 10]. Equivalent data captured during the third multicenter study (n = 486) [11] were allocated to the test data set.

Variable selection

Two logistic regression models were trained and tested: one applying definitional criteria (Table 2) and one applying physicians' final consensus diagnoses as proxies for AHT and non-AHT ground truth. To train the models, stepwise binary logistic regression with forward selection of variables was performed in IBM SPSS version 27. Patients labeled as indeterminate were not used to train the models, leaving n = 462 and n = 444 patients in the training data set, respectively. To minimize circularity, medical variables included in definitional criteria were excluded from those considered for model inclusion via stepwise logistic regression.

From inferred probabilities to LRs

The trained logistic regression models yielded posterior probabilities for AHT and non-AHT. Likelihood ratios were calculated by dividing the posterior odds by the prior odds, based on the observed frequencies of AHT and non-AHT in the training data set.

Patient-specific LRs

Patient-specific LRs were calculated for every patient in the test data set, based on each patient's specific combination of the positive and negative medical findings selected for inclusion in each model. To facilitate visual and clinical characterization of test patients' aggregate results, the LRs were analyzed as $\log_{10}$ LR values (LLR), which reflect evidential value along a continuous numerical scale, where: (1) 0 represents neutral evidence (the findings do not discriminate between AHT and non-AHT), (2) values > 0 support the hypothesis of AHT, and (3) values < 0 support the hypothesis of non-AHT. The more the LLR deviated from zero, the stronger the evidence and the greater the degree of support. To clarify, LLR values of + 1, + 2, and + 3 describe medical findings observed 10, 100, and 1000 times more frequently

Table 2 The PediBIRN network's a priori definitional criteria for abusive, non-abusive, and indeterminate head trauma, used to derive and validate the PediBIRN-7

A patient's head trauma was classified as abusive IF…
• The primary caregiver[1] admitted abusive acts
• Abusive acts by the primary caregiver[1] were witnessed by an unbiased, independent observer
• The primary caregiver[1] specifically denied any head trauma, even though the pre-ambulatory child in his or her care became acutely, clearly and persistently ill with clinical signs subsequently linked to traumatic cranial injuries visible on CT or MRI
• The primary caregiver's[1] account of the child's head injury event was clearly historically inconsistent with repetition over time
• The primary caregiver's[1] account of the child's head injury event was clearly developmentally inconsistent with child's known (or expected) gross motor skills
• Abuse evaluations revealed patterned bruising or dry contact burns, hot water immersion burns, or CT-confirmed intra-abdominal injury
A patient's head trauma was classified as non-abusive IF…
• The child's head injury event was witnessed by an unbiased, independent observer who described the event as accidental (non-abusive), OR…
• The primary caregiver[1] provided an account of the child's head injury event that was both historically consistent with repetition over time and developmentally consistent with the child's known (or expected) gross motor skills…AND…abuse evaluation failed to reveal patterned bruising or dry contact burns, hot water immersion burns, or CT-confirmed intra-abdominal injury
All remaining patients' head traumas were classified as indeterminate

Abbreviations: CT = computed tomography, MRI = magnetic resonance imaging, PediBIRN = pediatric brain injury research network

[1]Defined a priori as the person responsible for the child when he or she was acutely head injured or first became clearly and persistently ill with clinical signs subsequently linked to traumatic cranial injuries visible on CT or MRI

This table was published in Child Abuse & Neglect, Vol 88, Hymel KP, Wang M, Chinchilli VM, et al. Estimating the probability of abusive head trauma after abuse evaluation, pp 266—74, 2019, Copyright Elsevier, adapted and reprinted with permission from Elsevier

in patients labeled as AHT. Conversely, LLR values of −1, −2, and −3 describe medical findings observed 10, 100, and 1000 times more frequently in patients labeled as non-AHT.

Model performance

To assess and compare model performance, we plotted the LLR values in overlay-histograms showing both AHT and non-AHT labeled patients in the test data set. We also applied the cost log LR (Cllr), a metric that uses a logarithmic penalty function to place a cost on a LR, such that the cost increases when support for the ground truth hypothesis decreases. Thus, the Cllr is an addition of a cost due to lack of discrimination and a cost due to lack of calibration [9]. The Cllr was defined as the average cost for the LRs in the test data set. The lower its value, the better the model, where: (1) the value 0 indicates a perfect LR model, (2) values between 0 and 1 indicate an informative but imperfect LR model, and (3) values > 1 indicate a useless model.

To provide a basis for comparison, LR, LLR, and Cllr values were also calculated using the PediBIRN-7 prediction rule, applying the same two proxies for AHT and non-AHT ground truth. To preserve focus, the assessment of evidential values was limited to patient-specific LLRs calculated using the two best-performing statistical models (i.e., those with lowest Cllr values, applying definitional criteria and physicians' final consensus diagnoses as proxies for AHT and non-AHT ground truth, respectively).

Results

Patient labeling

The distribution of study patients (N = 973) labeled as AHT, non-AHT, and indeterminate is presented in Table 3. The PediBIRN network's definitional criteria (Table 2) proved to be the more conservative proxy, sorting fewer patients as AHT than physicians.

Table 3 The distribution of study patients into cohorts

Applying…	**Physicians' final consensus diagnoses**			
Definitional criteria[1]	*AHT*	*Indeterminate*	*Non-AHT*	**Totals**
AHT	337	11	22	**370**
Indeterminate	63	7	9	**79**
Non-AHT	96	36	392	**524**
Totals	**496**	**54**	**423**	**973**

Abbreviations: AHT=abusive head trauma, PediBIRN=pediatric brain injury research network

[1] See Table 2

Variable selection and model performance

Applying definitional criteria, the stepwise logistic regression model performed best, yielding a Cllr of 0.791. Applying physicians' final consensus diagnoses, the PediBIRN-7 performed best with a Cllr of 0.403. The medical findings selected for inclusion in these two best-performing models revealed similarities and overlap (Table 4).

Patient-specific LLRs

Figures 1 and 2 present histograms of patient-specific LLR values for the test data set, calculated using these two best performing statistical models. In both figures, a fraction of patients labeled as AHT had misleading LLRs < 0, and a fraction of patients labeled as non-AHT had misleading LLRs > 0. In Fig. 1, the histogram based on the stepwise logistic regression model and definitional criteria revealed a large overlap in LLR values for AHT and non-AHT patients—a result typical of LR models where the discrimination power of the variables is modest. In Fig. 2, the histograms based on the PediBIRN-7 prediction rule and physicians' final consensus diagnoses revealed better separation than in Fig. 1, and the fraction of patients with misleading evidence was smaller. The evidential (LLR) values of test patients AHT-related medical findings varied widely—from modest (between −1 to +1) to moderately strong (from −2.5 to −1 and from +1 to +3.5).

Medical findings that best differentiated physician diagnoses of AHT versus non-AHT

The five highest positive and five lowest negative LLR values resulting from application of the PediBIRN-7 prediction rule and physicians' final consensus diagnoses are presented in Table 5. The five highest—supporting the hypothesis of AHT—occurred in test patients who shared four medical findings: (1) bilateral or interhemispheric subdural hemorrhages or collections; (2) bruising of the torso, ears, or neck; (3) skeletal survey that revealed fractures moderately or highly specific for abuse; and (4) retinal exam findings of retinoschisis or dense, extensive retinal hemorrhages. The four highest LLR values occurred in patients who also presented with brain hypoxia, ischemia, or swelling. Conversely, the four and five lowest negative LLR values—supporting the hypothesis of non-AHT—occurred in test patients lacking these same findings.

Table 4 Variables selected for inclusion in the two best performing statistical models

Variable	Stepwise logistic regression model, with forward selection of variables, applying definitional criteria	The PediBIRN 7-variable rule, applying physicians' final consensus diagnoses
Age at the time of hospital admission (months)	✓	
Any acute respiratory compromise[1]		✓
Any acute encephalopathy[2]	✓	
Any bruising of the torso, ear(s), or neck	✓	✓
Any skull fracture(s)	✓	
Skull fracture(s) other than an isolated, unilateral, nondiastatic, linear, parietal skull fracture		✓
Any subdural hemorrhage(s) or collection(s)	✓	
Subdural hemorrhage(s) or collection(s) that are bilateral OR interhemispheric		✓
Any brain hypoxia, ischemia, or swelling	✓	✓
Skeletal survey ordered and completed	✓	
Skeletal survey completed and revealed fracture(s) moderately or highly specific for abuse[3]	✓	✓
Retinal exam by an ophthalmologist completed and revealed retinoschisis OR dense, extensive, retinal hemorrhages extending to the ora serrata		✓

[1]Defined a priori as "infrequent or labored respirations, apnea, or any requirement for intubation or assisted ventilation"

[2]Defined a priori as "an initial, clear, impairment or loss of consciousness at the scene of injury, during transport, in the Emergency Department, or prior to hospital admission"

[3]Including classic metaphyseal lesion fractures or epiphyseal separations, rib fractures, fractures of the scapula or sternum, fractures of digits, vertebral body fractures, dislocations or fractures of spinous processes

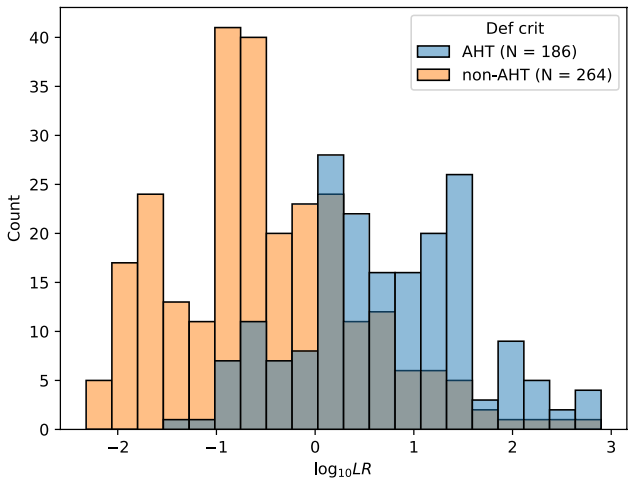

Fig. 1 The LLRs for AHT and non-AHT patients are mainly > 0 and < 0 respectively, as desired. 18% of AHT patients revealed misleading LLRs < 0, and 27% of non-AHT patients revealed misleading LLRs > 0. For some patients, this misleading evidence was moderately strong, with LLRs approaching 3

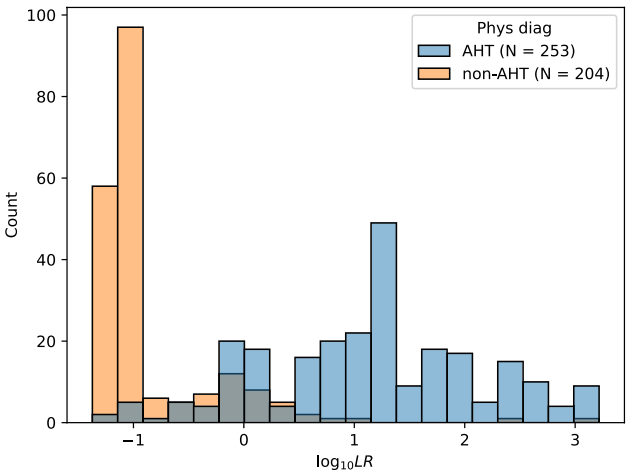

Fig. 2 Compared to Fig. 1, the two distributions show better separation and smaller fractions of patients with misleading evidence. Among AHT patients, 15% have misleading LLRs < 0. Among non-AHT patients, 9% have misleading LLRs > 0. Two non-AHT patients revealed quite large, strongly misleading LLR values of 2.5 and 3.2

Discussion

Prior work regarding the diagnosis of AHT and non-AHT

Across multiple decades, patient populations, and health care systems, clinician investigators applying divergent methodologies have observed the same significant differences reported to differentiate AHT and non-AHT [3, 4, 12–27]. The same differences were observed in two studies that compared cases of witnessed/admitted AHT versus witnessed non-AHT [19, 27]. An equivalent comparison of witnessed/admitted AHT versus physician diagnosed AHT not witnessed or admitted revealed a complete absence of significant differences, leading the authors to opine that participating physicians applied their knowledge of these reported injury patterns 'with reasonable and expected rigor' [19].

Table 5 Patient-specific LLR values applying the PediBIRN-7's prediction variables and physicians' final consensus diagnoses as proxies for AHT and non-AHT ground truth

Any acute respiratory compromise[1]	Any bruising of the torso, ear(s), or neck	Subdural hemorrhage(s) or collection(s) that are bilateral or interhemispheric	Any skull fracture(s) other than an isolated, unilateral, nondiastatic, linear, parietal skull fracture	Skeletal survey that revealed fracture(s) moderately or highly specific for abuse[2]	Retinal exam that revealed retinoschisis OR dense, extensive, retinal hemorrhages	Any brain hypoxia, ischemia, or swelling	LLR
✓	✓	✓	×	✓	✓	✓	3.41
×	✓	✓	×	✓	✓	✓	3.23
✓	✓	✓	✓	✓	✓	✓	3.18
×	✓	✓	✓	✓	✓	✓	3.00
✓	✓	✓	×	✓	✓	×	2.92
×	×	×	✓	×	×	✓	−0.69
✓	×	×	×	×	×	×	−0.76
×	×	×	×	×	×	×	−0.94
✓	×	×	✓	×	×	×	−0.99
×	×	×	✓	×	×	×	−1.18

Abbreviations: LLR = $\log_{10}$ likelihood ratio, PediBIRN = pediatric brain injury research network

[1]Defined a priori as "infrequent or labored respirations, apnea, or any requirement for intubation or assisted ventilation"

[2]Including classic metaphyseal lesion fractures or epiphyseal separations, rib fractures, fractures of the scapula or sternum, fractures of digits, vertebral body fractures, dislocations or fractures of spinous processes

Though promising, victims of AHT have been misdiagnosed and reinjured [1, 2], and physicians' decisions to evaluate and report suspected AHT have demonstrated disparities and apparent bias [28]. To address these challenges, investigators have derived, validated, and implemented evidence-based AHT screening and diagnostic tools [3, 4, 6, 7, 10, 11, 29, 30]. Like the statistical models presented in this study, two of these prediction tools facilitate estimation of AHT probability based on specific combinations of predictor variables [3, 4].

Similar studies

Maguire et al. analyzed aggregate data from six studies to create the PredAHT prediction tool for AHT [4]. Their report includes actual and imputed estimates of AHT probability for all combinations of the PredAHT's six clinical variables. The authors observed that "no set of clinical features was unique to AHT or non-AHT" and asserted that "features must be considered in the context of all other medical and social aspects of the case in question."

Applying different methods, PediBIRN investigators identified an optimal cluster of seven medical findings for inclusion in their 7-variable prediction tool (Tables 4 and 5) and published LRs and estimates of AHT probability for every combination of the seven variables observed in their study population [3]. Like Maquire et al. [4], the authors asserted that their patient-specific estimates of AHT probability "must be interpreted in the context of other relevant findings and data".

Other relevant findings and data

These include medical findings routinely captured and interpreted by physicians (e.g., the history of present illness, past and family medical history, the results of physical examination, specialty consultations, imaging studies, and tests designed to confirm or exclude medical mimics of head trauma) and non-medical findings routinely captured and interpreted by non-medical professionals (e.g., the results of family psychosocial risk assessment, forensic interviews, and scene investigations).

Added value of this study

Unlike the methods used to derive these two, prior, multivariable AHT prediction tools [3, 4], we measured the evidential value of missing variables (i.e., abuse evaluations "never ordered/completed") and assessed both model discrimination and calibration. Thus, the statistical models presented in this study represent an enhancement in evidence-based, multivariable approaches to AHT-related medical diagnosis. Perhaps more importantly, the conceptual novelty of this study lies in its focus on the LR as a measure of evidential value.

Our overall results

As presented in Figs. 1 and 2: (1) The evidential values of test patients' most discriminating medical findings varied

widely, from modest to moderately strong; (2) Calculated LLR values were sometimes misleading; and (3) For a small portion of test patients labeled as non-AHT, this misleading evidence was moderately strong.

The potential impacts of uncertainty regarding AHT and non-AHT ground truth

We suspect that uncertainty regarding AHT and non-AHT ground truth contributed significantly to the misleading evidence presented in Figs. 1 and 2. Attempting to deal with this uncertainty, we elected to apply two distinct proxies for AHT and non-AHT ground truth, believing that comparable results would convey a measure of validity. Our calculations of evidential value assume that these proxies are reasonably valid representations of actual truth. Each has its advantages and disadvantages [3].

Citing the rigor of their efforts to exclude alternate diagnoses and their freedom to consider additional medical and non-medical data, we suspect that physicians who evaluate cases of suspected AHT will defend the application of their final consensus diagnoses as a reasonable proxy for ground truth. The application of this proxy does have the potential to introduce variability (stemming from differences in physician training, experience, and/or bias) and embedded circular reasoning, resulting in diminished validity.

The application of a priori definitional criteria as proxies for AHT and non-AHT ground truth restricts the consideration of multiple variables otherwise available to diagnosing physicians. Properly designed, definitional criteria can reduce circularity. In this study, the medical variables considered for model inclusion (e.g., the results of retinal examinations and skeletal surveys) were excluded from the definitional criteria. Doing so increased reliance on historical criteria. We deemed this acceptable, as three of these historical criteria (Table 2) have demonstrated high specificity and positive predictive values for AHT [31].

Lacking definitive evidence of either proxy's validity, the misleading evidence revealed in Figs. 1 and 2 should be interpreted with a measure of caution. Stated differently, some patients labeled as AHT—with LLR values indicative of non-AHT—may not have been abused, and some patients labeled as non-AHT—with LLR values indicative of AHT—may have been abused.

The application of LRs in medical practice

We believe that LRs can inform AHT-related clinical judgement. Paraphrasing Hymel et al. [3]: "…physicians who find reasonable concordance between their final diagnostic impressions and a patient-specific LR will likely feel more confident that their impressions are valid. Physicians who instead find discordance may elect to press for further investigation, and to explore the possibilities of a false positive or negative result, an error in their clinical judgement, AHT masquerading as accidental trauma, and/or their own inexperience or implicit bias."

Unlike probability estimates, the calculation of AHT-relevant LRs does not depend on AHT prevalence (pre-test odds) in the study population. The formula for calculating patient-specific LRs using the stepwise logistic regression model and definitional criteria is available in Supplementary Materials. The PediBIRN-7's patient-specific LRs are available at www.pedibirn.com.

The Bayesian framework for forensic evidence interpretation

Many forensic scientists, especially those engaged in DNA analysis and interpretation, embrace the Bayesian framework for evidence interpretation [5, 32, 33]. This framework promotes a process in which experts use LRs to inform legal decision makers about the evidential value of findings within their specific domains of expertise. Evidential value is also determined by other factors, such as the relevance of the hypotheses considered and the credibility of the methods used.

The legal decision maker can then use this information to update their estimate of prior odds in any way that they see fit. Professional judges who are aware of Bayes rule (prior odds x LR = posterior odds) may use it consciously, but in practice the vast majority will process the information differently.

The application of LRs in legal settings

Although medical experts have been very slow to adopt this framework, we believe they will find the LR to be a highly informative, evidence-based measure of evidential strength. Physicians who opine and/or testify in legal settings can communicate evidential strength explicitly by reporting the LR. They can also do so implicitly by providing a carefully worded summary of all findings considered, and by explaining why specific combinations of these findings support their overall conclusions.

When medical experts choose to report a patient-specific LR, they should be aware that lay people may not interpret it as intended. For example, the notorious 'prosecutor's fallacy' is difficult to avoid [34]. Accordingly, they should take care that the numbers do not overwhelm the legal decision maker and remember to emphasize the need to consider all relevant findings as a whole. We provide an example of such testimony in Supplementary Materials.

We support the inclusion of LRs in written and/or oral testimony by qualified medical experts who have completed

a thorough medical evaluation for abuse, considered all relevant findings and data, and worked rigorously to exclude alternate medical diagnoses. In support of this position, we challenge stakeholders to (1) remember that "a reasonable degree of medical certainty" is itself a statement of probability, (2) acknowledge that physicians' inherent biases, their desire to protect children or caregivers, and an adversarial legal system may lead them to overstate the certainty of their diagnostic conclusions, and (3) acknowledge that physicians providing expert testimony in an adversarial legal system may feel compelled to testify using definitive language (e.g., "My patient was abused"). The use of LRs could serve to support or temper an expert's statement.

Additional topics for discussion

Moving forward, we offer the following questions as topics for future discussion or debate: (1) Does "a reasonable degree of medical certainty of abuse" equal "My patient was abused"? (2) Absent gold standards, are definitive diagnoses of AHT and non-AHT possible? If so, when? (3) How can physicians balance their desire to protect children or caregivers and their commitment to professional objectivity? (4) Should expert testimony regarding LRs and evidential values be compelled? (5) Would doing so enhance evidential clarity and/or physicians' credibility? (6) Should expert testimony regarding AHT-related LRs be restricted? If so, when? and (7) Should physicians who opine and/or testify regarding AHT be compelled to reveal the degree to which they relied on non-medical data?

Standardization

The 2025 ISO—a summary of forensic standards published by the International Organization for Standardization—includes a discussion of LRs and multiple examples of their application, including an example of suspected AHT [32]. Although we believe the application of LRs can provide meaningful clarity to triers of fact, we remain hesitant to predict if or when their application will become standard medicolegal practice in cases of suspected AHT.

Study strengths

Analyses were based on uniform, prospective data captured across 18 sites [7, 10, 11]. Definitional criteria for AHT, non-AHT, and for positive skeletal surveys and retinal exams were defined a priori and designed to minimize circularity [7, 10, 11]. The medical findings included in the aggregate data set had previously demonstrated acceptable interrater reliability [10]. To calculate patient-specific LRs, we derived two statistical models and applied two different proxies for AHT and non-AHT ground truth. Models were tested and trained on separate, highly equivalent datasets.

Study limitations

The proxies for AHT and non-AHT ground truth applied in these analyses are likely imperfect. Variability in the number, timing, and modalities of neuroimaging may have impacted recognition of cranial findings. Some patients did not undergo skeletal survey and/or retinal exam, resulting in missing values. Our results may not be generalizable to non-intensive care settings.

Conclusion

Reasonable medical certainty of AHT and non-AHT can be enhanced or limited by the highly variable evidential values of patients' most discriminating medical findings. Physicians can use available LRs to inform their AHT-related diagnostic reasoning, opinions, and testimony. Doing so requires a thorough understanding of their strengths and limitations.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00414-026-03831-z.

Acknowledgements The authors wish to acknowledge and thank Dr. Wouter Karst for his contributions related to the conception of this study.

Authors' contributions Dr Hymel designed the data collection instruments, coordinated and supervised data collection, drafted the initial manuscript, and critically reviewed and revised the manuscript. Drs Sjerps and Vergeer conceptualized and designed the study, curated the data, designed and executed all analyses, and critically reviewed the manuscript for important intellectual content. All authors approved the final manuscript and agreed to be accountable for all aspects of the work.

Funding Dr. Hymel's work related to data acquisition was funded, in part, by the Eunice Kennedy Shriver National Institute of Child Health and Human Development (grant number P50HD089922). The National Institutes of Health had no role in the design or conduct of the study; the collection, management, analysis, or interpretation of the data; the preparation, review, or approval of the manuscript; or the decision to submit the manuscript for publication.

Declarations

Ethical approval The Institutional Review Board at Penn State Health Hershey Medical Center determined that this secondary analysis was not human subject research.

Conflict of interest The authors have no conflicts of interest relevant to this article to disclose.

References

1. Jenny C, Hymel KP, Ritzen A, Reinert SE, Hay T (1999) Analysis of missed cases of abusive head trauma. JAMA 281:621–626
2. Letson MM, Cooper JN, Deans KJ et al (2016) Prior opportunities to identify abuse in children with abusive head trauma. Child Abuse Negl 60:36–45
3. Hymel KP, Wang M, Chinchilli VM et al (2019) Estimating the probability of abusive head trauma after abuse evaluation. Child Abuse Negl 88:266–274
4. Maguire SA, Kemp AM, Lumb RC, Farewell DM (2011) Estimating the probability of abusive head trauma. A pooled analysis. Pediatrics 128:e550-564
5. Aitken CCG, Barrett A, Berger CEH, et el. ENFSI guideline for evaluative reporting in forensic science: strengthening the evaluation of forensic results across Europe (STEOFRAE). The European Network of Forensic Science Institutes; 2015:16–18 (available from https://enfsi.eu/wp-content/uploads/2016/09/m1_guideline.pdf)
6. Hymel KP, Carroll CL, Frazier TN et al (2024) Validation of the PediBIRN-7 clinical prediction rule for pediatric abusive head trauma. Child Abuse Negl 152:106799
7. Hymel KP, Armijo-Garcia V, Foster R et al (2014) Validation of a clinical prediction rule for pediatric abusive head trauma. Pediatrics 134:e1537-1544
8. Dawid AP (1982) The well-calibrated Bayesian. J Am Stat Assoc 77:605–610
9. Ramos D, Meuwly D, Haraksim R, Berger CEH (2020) Validation of forensic automatic likelihood ratio methods. In: Banks DL, Kafadar K, Kaye DH, Tackett M (eds) Handbook of forensic statistics. CRC Press, pp 143–162
10. Hymel KP, Willson DF, Boos SC et al (2013) Derivation of a clinical prediction rule for pediatric abusive head trauma. Pediatr Crit Care Med 14:210–220
11. Hymel KP, Armijo-Garcia V, Musick M et al (2021) A cluster randomized trial to reduce missed abusive head trauma in pediatric intensive care settings. J Pediatr 236:260–268
12. Amagasa S, Matsui H, Tsuji S, Uematsu S, Moriya T, Kinoshita K (2018) Characteristics distinguishing abusive head trauma from accidental head trauma in infants with traumatic intracranial hemorrhage in Japan. Acute Med Surg 5:265–271
13. Bechtel KM, Stoessel K, Leventhal JM et al (2004) Characteristics that distinguish accidental from abusive injury in hospitalized young children with head trauma. Pediatrics 14:165–168
14. Binenbaum G, Mirza-George N, Christian CW, Forbes BJ (2009) Odds of abuse associated with retinal hemorrhages in children suspected of abuse. J Pediatr Ophthalmol Strabismus 13:268–272
15. Boos SC, Wang M, Karst WA, Hymel KP (2022) Traumatic head injury and the diagnosis of abuse: a cluster analysis. Pediatrics 149:e2021051742
16. Ewing-Cobbs L, Prasad M, Kramer L et al (2000) Acute neuroradiologic findings in young children with inflicted or noninflicted traumatic brain injury. Childs Nerv Syst 16:25–33
17. Feldman KW, Bethel R, Shugerman RP, Grossman DC, Grady MS, Ellenbogen RG (2001) The cause of infant and toddler subdural hemorrhage: a prospective study. Pediatrics 108:636–646
18. Hymel KP, Makoroff KL, Laskey AL, Conaway MR, Blackman JA (2007) Mechanisms, clinical presentations, injuries, and outcomes from inflicted versus noninflicted head trauma during infancy: results of a prospective, comparative, multicenter study. Pediatrics 119:922–929
19. Hymel KP, Boos SC, Armijo-Garcia V et al (2022) An analysis of physicians' diagnostic reasoning regarding pediatric abusive head trauma. Child Abuse Negl 129:105666
20. Kelly P, Simon J, Vincent AL, Reed P (2015) Abusive head trauma and accidental head trauma: a 20 year comparative study of referrals to a hospital child protection team. Arch Dis Child 100:1123–1130
21. Kemp AM, Jaspan T, Griffiths J et al (2011) Neuroimaging: what neuroradiological features distinguish abusive from non-abusive head trauma: a systematic review. Arch Dis Child 96:1103–1112
22. Maguire S, Pickerd N, Farewell D, Mann M, Tempest V, Kemp AM (2009) Which clinical features distinguish inflicted from non inflicted brain injury? A systematic review. Arch Dis Child 94:860–867
23. Narang S, Clarke J (2014) Abusive head trauma: past, present and future. J Child Neurol 29:1747–1756
24. Piteau SJ, Ward MG, Barrowman NJ, Plint AC (2012) Clinical and radiographic characteristics associated with abusive and nonabusive head trauma: a systematic review. Pediatrics 130:315–323
25. Reece RM, Sege R (2000) Childhood head injuries: accidental or inflicted? Arch Pediatr Adolesc Med 154:11–15
26. Vinchon M, deFoort-Dhellemmes S, Desurmont M, Dhellemmes P (2005) Accidental and nonaccidental head injury in infants: a prospective study. J Neurosurg 102:380–385
27. Vinchon M, deFoort-Dhellemmes S, Desurmont M, Delestret I (2010) Confessed abuse versus witnessed accidents in infants: comparison of clinical, radiological, and ophthalmological data in corroborated cases. Childs Nerv Syst 26:637–645
28. Hymel KP, Laskey AL, Crowell KR et al (2018) Racial/Ethnic disparities and bias in the evaluation and reporting of abusive head trauma. J Pediatr 198:137–143
29. Hymel KP, Karst W, Marinello M et al (2022) Screening for pediatric abusive head trauma: are three variables enough? Child Abuse Negl 125:105518
30. Berger RP, Fromkin J, Herman B et al (2016) Validation of the Pittsburgh infant brain injury score for abusive head trauma. Pediatrics 138(1):e20153756
31. Hymel KP, Lee G, Boos SC et al (2020) Estimating the relevance of historical red flags in the diagnosis of abusive head trauma. J Pediatr 218:178–183
32. International Organization for Standardization. (2025). *Forensic Sciences*. 21043–4:2025. Part 4: Interpretation. Annex C, example C3. Available from: ISO 21043–4:2025-Forensic sciences—Part 4: Interpretation. (visited March 26 2026) . https://www.iso.org/obp/ui/en/#iso:std:iso:21043:-4:ed-1:v1:en
33. Hicks T, Buckleton J, Castella V, Evett I, Jackson G (2022) A logical framework for forensic DNA interpretation. Genes 13:957
34. Thompson WC, Newman EJ (2015) Lay understanding of forensic sciences: evaluation of random match probabilities, likelihood ratios, and verbal equivalents. Law Hum Behav 39:332–349