Published in final edited form as:
J Pediatr. 2018 July ; 198: 144–150.e4. doi:10.1016/j.jpeds.2018.01.033.

Classifying Injuries in Young Children as Abusive or Accidental: Reliability and Accuracy of an Expert Panel Approach

Douglas J. Lorenz, PhD[a], **Mary Clyde Pierce, MD**[b,c], **Kim Kaczor, MS**[b], **Rachel P. Berger, MD, MPH**[d], **Gina Bertocci, PhD, PE**[e], **Bruce E. Herman, MD**[f], **Sandra Herr, MD**[g], **Kent P. Hymel, MD**[h], **Carole Jenny, MD, MBA**[i], **John M. Leventhal, MD**[j], **Karen Sheehan, MD, MPH**[b,c], and **Noel Zuckerbraun, MD, MPH**[d]

[a]Department of Bioinformatics and Biostatistics, School of Public Health and Information Sciences, University of Louisville, 485 E. Gray St., Louisville, KY 40202

[b]Division of Emergency Medicine, Ann & Robert H. Lurie Children's Hospital of Chicago, 225 E. Chicago Ave., Box 62, Chicago IL 60611

[c]Department of Pediatrics, Northwestern University Feinberg School of Medicine, Chicago, IL 60611

[d]Department of Pediatrics, University of Pittsburgh, Children's Hospital of Pittsburgh of UPMC, 4401 Penn Avenue, Pittsburgh, PA 15224

[e]Department of Bioengineering, J.B. Speed School of Engineering, University of Louisville, 500 S. Preston St., Louisville, KY 40202

[f]Department of Pediatrics, University of Utah School of Medicine, 81 N. Mario Capecchi Dr. Salt Lake City, UT 84113

[g]Division of Pediatric Emergency Medicine, University of Louisville, 571 S. Floyd St, Suite 300, Louiville KY 40202

[h]Department of Pediatrics, Division of Child Abuse Pediatrics, 500 University Drive, P.O. Box 850, Hershey, PA 17033-0850

[i]Department of Pediatrics, University of Washington, Seattle Children's Hospital, M/S M2-10, 4800 Sand Point Way NE, Seattle, WA 98105

[j]Department of Pediatrics, Yale School of Medicine, 333 Cedar St. New Haven, CT 06520

Abstract

Objective—To assess interrater reliability and accuracy of an expert panel in classifying injuries of patients as abusive or accidental based on comprehensive case information.

Corresponding author: Douglas J. Lorenz, Department of Bioinformatics and Biostatistics, School of Public Health and Information Science, University of Louisville, 485. E. Gray St., Louisville, KY 40202, djlore01@louisville.edu, 1-502-852-3635.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

None of the other authors has potential conflicts of interest.

Study design—Data came from a prospective, observational, multi-center study investigating bruising characteristics of children younger than 4 years. We enrolled 2166 patients with broad ranges of illnesses and injuries to one of 5 pediatric emergency departments (PED) in whom bruises were identified during examination. We collected comprehensive data regarding current and past injuries and illnesses, and provided de-identified, standardized case information to a 9-member multi-disciplinary panel of experts with extensive experience in pediatric injury. Each panelist classified cases using a 5-level ordinal scale ranging from Definite Abuse to Definite Accident. Panelists also assessed whether report to child protective services (CPS) was warranted. We calculated reliability coefficients for likelihood of abuse and decision to report to CPS.

Results—The interrater reliability of the panelists was high. The Kendall coefficient (95% CI) for the likelihood of abuse was 0.89 (0.87, 0.91) and the kappa coefficient for the decision to report to CPS was 0.91 (0.87, 0.94). Reliability of pairs and subgroups of panelists were similarly high. A panel composite classification was nearly perfectly accurate in a subset of cases having definitive, corroborated injury status.

Conclusions—A panel of experts with different backgrounds but common expertise in pediatric injury is a reliable and accurate criterion standard for classifying pediatric injuries as abusive or accidental in a sample of children presenting to a PED.

Keywords

child abuse; criterion standard classification

Victims of child abuse are at high risk of future abuse and death.[1,2] Decision rules for identifying abusive injuries are valuable for settings such as pediatric emergency departments (PED). These rules require thorough evaluation to ensure both classification accuracy and agreement among users. A significant challenge in such development and evaluation is that the true nature of the injury, abusive or accidental, often is not definitively known. Therefore, clinicians and researchers alike must depend on empirical and/or corroborative[3] evidence to classify injuries.[4] The availability of corroborative information, such as a confession, is uncommon and the absence of a criterion standard for classifying injuries as abusive or accidental can adversely affect research in child abuse.

A potential approach to establishing a criterion standard classification of injuries is the opinion of a panel of pediatric injury experts—pediatric emergency medicine physicians, child abuse pediatricians, and researchers in the biomechanics of pediatric injury, —wherein classification is based on empirical evidence including consistency of history and injury compatibility. Expert panels have been widely used for the classification of uncertain outcomes in diverse clinical settings[5–8]. In child abuse research, expert panels have previously been used to evaluate the likelihood of abuse in children with fractures,[9–11] develop guidelines for ordering skeletal surveys,[12] examine differences of opinion among clinicians using different abuse rating scales,[13] and evaluate the impact of structured decision-making tools on child maltreatment decisions made by child abuse practitioners.[14]

The purposes of this study were to examine the interrater reliability of an expert panel with regard to classifying pediatric injuries as abusive or accidental, and to demonstrate the accuracy of the approach in cases with definitive corroborative information. We hypothesized that the expert panel would exhibit substantial interrater reliability and accuracy in the classification of pediatric injuries as abusive or accidental.

Methods

Data came from patients enrolled in a prospective, observational, multi-center study investigating the bruising characteristics of young children and the psychosocial characteristics of their families. Eligible children were less than 4 years of age, presented to a PED participating in the bruising study with any chief complaint, and had bruising identified by a previously described[15,16] structured skin examination. Excluded children were patients with known coagulation abnormalities, severe neurologic impairments, severe extensive skin disorders, and motor vehicle crash victims. The study was conducted at five children's hospitals, each associated with a large, urban, academically-affiliated tertiary care hospital: Ann & Robert H. Lurie Children's Hospital of Chicago, University of Chicago Medicine Comer Children's Hospital, Cincinnati Children's Hospital Medical Center, Rady Children's Hospital, and Norton Children's Hospital. The sites cumulatively evaluate over 70,000 children annually in the target age range. Study investigators enrolled patients from participating sites by informed parental consent unless the team providing treatment at the participating PED obtained a child abuse consultation, in which case waivers of authorization were allowed. Institutional Review Board (IRB) approval was obtained at each site.

Expert Panel and Data Provided for Case Assessment

The expert panel included 9 members: 4 child abuse pediatricians, 4 pediatric emergency medicine physicians, and a bioengineer with expertise in pediatric injury. Panelists had 14 to 39 years of experience in their respective fields.

Each panelist received de-identified case information in a standardized electronic format, including data from the patient's current visit to a participating PED as well as available data from any previous ED visits for any chief complaint. The following case information was included: patient's age and sex; all notes related to the patient's reason for visit and history, taken verbatim from the medical documentation after redaction of identifiers; additional detailed historical data regarding the injury event; photographs of the skin injuries; and diagnostic imaging that identified any internal injuries (fractures, brain hemorrhages, chest or abdominal injuries) along with the official radiologist's report. Each panelist independently reviewed the case information and was required to answer a structured series of questions regarding history consistency, injury compatibility, and other case characteristics (Appendix; available at www.jpeds.com). We blinded panelists to all case data on psychosocial risk factors, including history of domestic violence, substance abuse, criminal activity, etc. Each panelist independently rated the likelihood of abuse on an ordinal scale with 5 levels: "Definite Abuse," "Likely Abuse," "Indeterminate," "Likely Accident," and "Definite Accident." Panelists were given no guidance in distinguishing "definite" from "likely," and exercised full and free judgment in their selections. Additionally, each panelist provided a yes or no answer to the question: "Is a report to state child protective services (CPS) indicated?"

Enrollment and case categorization occurred from December 2011 through March 2016. Members of the expert panel evaluated and classified 2166 cases. We randomly selected a subset of 201 test cases to be reviewed by all 9 panelists, representing roughly 10% of the targeted enrollment of 2000 cases. As it was not feasible for each panelist to review all 2166 cases, we randomly assigned the remaining cases so that at least two panelists independently reviewed each case. Panelists were not given discretion over which cases to review. All panelists provided ratings for the cases they were assigned with no refusals.

Of the 2166 cases, we found that 584 could be more definitively classified as abuse or accident based on additional corroborative information obtained after the visit to the PED.[4] Corroborated cases were defined as those in which there was video capture of the event, a confirmed public event, a third party account such as confirmatory documentation from a licensed daycare, a confession of abuse, criminal conviction of abuse, injury from a confirmed domestic violence conflict, or concurrent sibling injury with confession and/or conviction of abuse. Two investigators not on the expert panel coded the presence or absence of corroborative criteria for each case. Corroborative information was in some cases available to panelists through the history of injury, but panelists were not aware of the criteria used to define corroborated cases. For example, if a child fell from a swing set in a confirmed public event, the panelists reviewing the case would know from the injury history that the injury occurred in a public setting, but was not aware that this information qualified the case as a corroborated accident. We used these corroborated case classifications to determine the accuracy of the classifications provided by the expert panel. Figure 1 (available at www.jpeds.com) provides a diagram of the different subsets of cases and how they were use in analyses of reliability and accuracy.

Statistical Analyses

We summarized panelists' classifications of the likelihood of abuse and reports to CPS with counts and percentages. We assessed the interrater reliability of the expert panel for the 5-level likelihood of abuse by calculating Kendall's coefficient of concordance, a measure of reliability for ordinally scaled variables. We calculated Fleiss' kappa coefficient to assess reliability for the report to CPS, which is appropriate for dichotomous variables. To account for differences in panelists' personal inclinations to use the "definite abuse" and "definite accident" categories, we also calculated reliability statistics for derived three-level and binary versions of the 5-level classification (Figure 2; available at www.jpeds.com). The 3-level version grouped definite and likely accidents and definite and likely abuse cases into single "accident" and "abuse" categories and preserved the indeterminate category. For this derived 3-level classification, we calculated the Kendall coefficient to evaluate interrater reliability. The binary version was created from the three-level classification by reclassifying indeterminate responses according to the answer to the report to CPS question – those answered with "Yes" were classified as abuse and those with "No" as accident. We calculated Fleiss' kappa to assess interrater reliability. The original 5-level ordinal classification provided panelists the freedom to classify injuries on a scale graded by the certainty of the classification. The statistically derived 3-level and binary versions removed the degree of certainty (definite vs likely) of the classification. The 3-level classification kept a measure of uncertainty by retaining the indeterminate category, and the binary classification removed it.

We calculated reliability coefficients on the subset of 201 test cases for all 9 panelists. We also calculated coefficients within and between groups of panelists defined by area of expertise—pediatric emergency medicine physicians (4 panelists), child abuse pediatricians (4 panelists), and the bioengineer—to evaluate agreement among panelists of similar and different backgrounds. Additionally, we calculated reliability coefficients for the 36 possible pairings of individual panelists to examine agreement for pairs of panelists. The pairwise analyses were conducted on all cases commonly reviewed by each pairing of panelists.

To evaluate the accuracy of the expert panel, we descriptively analyzed the agreement between panelists' binary classifications and corroborative classifications of abuse or accident with counts and percentages. For each case, we defined a panel composite classification to be abuse if the majority of panelists reviewing a case gave classifications of abuse, accident if the majority issued accident classifications, and "unclassified" otherwise. We analyzed agreement between the panel composite classification and the corroborative classification with counts and percentages as well.

Nonparametric 95% confidence intervals for reliability coefficients were calculated by bootstrap over 10,000 Monte Carlo loops. Analyses were conducted in the open source R software environment.[17]

Results

The 2166 cases included children of average age 2.1 years, with ages ranging from 4 days to just under 4 years. The majority of children were male (1299, 60%), white (1789, 83%), and of non-Hispanic ethnicity (1489, 69%). 51% (1104) had government insurance, and 45% (974) private insurance. The stated reason for seeking care was medical in nature (e.g., fever, seizure) in 991 cases (46%), injury/trauma evaluation (e.g., fall from bed) in 958 cases (44%), and abuse evaluation referrals in 217 cases (10%).

Panelists classified between 67% and 76% of cases as likely or definite accident (Table 1), 18% to 26% of their cases as likely or definite abuse, and 3% to 10% as indeterminate. There were significant differences in the use of the "definite" categories among panelists ($p < .001$). The percentage of cases reported to CPS was consistent among panelists, ranging from 24% to 29% ($p = .63$).

Reliability coefficients for the full panel and subgroups of panelists were high for each version of the abuse likelihood classification as well as the report to CPS (Table 2), reflecting substantial agreement regardless of the background of the experts. All coefficients were 0.89 or greater, and all 95% confidence interval lower bounds were 0.85 or greater.

For the full panel, unanimous agreement on the 5-level ordinal classification occurred in 24 cases (12%), 21 rated as definite accident and 3 as definite abuse. There were 16 (8%)

abuse-accident disagreements, wherein a case was rated as abuse (definite or likely) by at least one panelist and as an accident (definite or likely) by at least one other panelist. Unanimous agreement on the 3-level classification occurred in 144 cases (72%), 86 accident and 58 abuse. For the binary classification, 178 cases (89%) exhibited unanimity, 105 accident and 73 abuse. There were 178 (89%) cases with unanimity for the decision to report to CPS, 105 "no" responses and 73 "yes."

Of the 2166 cases, 1436 (66%) were reviewed by 2 panelists, 365 (17%) by 3, and 365 (17%) by 4 or more. Figure 3 plots reliability coefficients and 95% confidence intervals for pairs of panelists. Kendall coefficients for the 5-level classification all exceeded 0.89, and kappa coefficients for the report to child protective services exceeded 0.85, indicating high reliability comparable with that shown by the full panel and panelist specialty groups. There were no systematic differences in reliability coefficients between pairs of panelists from the same background and pairs of different backgrounds. Results for the three-level and binary versions showed similarly high reliability (Figure 4; available at www.jpeds.com).

There was unanimous agreement – defined as identical abuse or accident classifications on the 5-level classification from all panelists reviewing the case – in 852 cases (39%). An additional 1048 cases (48%) exhibited partial agreement, in which all panelists provided classifications of abuse (definite or likely) or accident (definite or likely). There were 44 instances (2%) of abuse-accident disagreement, none of which were definite abuse-definite accident disagreements. In the remaining 222 (10%) cases, all panelists gave classifications of indeterminate, indeterminate and accident, or indeterminate and abuse.

Overall, 584 cases met one or more of the corroborative criteria, 148 corroborated abuse cases and 436 corroborated accidents. Individual panelists accurately classified between 95% and 98% of these corroborated cases of abuse and between 99% and 100% of corroborated accidents (Table III). The panel composite classification was in nearly perfect agreement with the corroborative classification.

Discussion

The interrater reliability of our full, 9-member expert panel was nearly perfect for ordinally-scaled assessments of the likelihood of abuse and three-level and binary classifications derived from these ordinal assessments. The panel exhibited nearly perfect interrater reliability in decisions to report cases to CPS. Reliability was high regardless of the panelists' background training, as subsets and pairs of panelists of similar and different backgrounds exhibited consistently high reliability coefficients. The panel was nearly perfectly accurate in classifying a subset of cases for which corroborative information allowed a definitive classification of abuse or accident.

As in other studies of child abuse[9–11,13,19], we used an ordinal scale measuring the likelihood of abuse, allowing some degree of uncertainty in the classifications made by panelists. Such a scale more appropriately parallels clinical decision making than rigid abuse-accident classifications, and we have demonstrated high interrater reliability under such uncertainty. Differences among panelists in their ordinal classifications may have been in part a function of differences in panelists' inclinations to rate cases as "definite" rather than "likely." The 3-level and binary classifications removed these panelist propensity effects. The high reliability of the panel for these derived versions showed that panelists agreed with respect to the overall conclusion of abuse or accident. The high reliability of the decision to report to CPS is noteworthy in that panelists were tasked with making a hypothetically actionable decision regarding their assessment of abuse. In a sense, the reliability of the panel for this decision represents agreement "in action" for cases of potential child abuse.

Other studies related to abusive trauma have utilized an expert panel approach.[8–13,18, 19] Most identified cases from an abuse consultation database or from medical record review, targeted a specific injury type (e.g., fracture), were retrospective in scope, created vignettes from a case series, or used hypothetical cases based on common injury types or scenarios. Most studies did not report reliability statistics. Two studies that focused on expert panel ratings of potential abuse highlighted the variability among panelist ratings. A study by Lindberg, Lindsell, and Shapiro[13] examined data gathered from true cases of varying types of physical abuse and noted variability among ratings from panelists within the same profession. They studied hypothetical cases of head injury and noted variability in ratings between two different types of specialists, child abuse pediatricians and pathologists.(18) Even though variability among professionals was noted, neither study reported reliability statistics. Although our study is difficult to compare with these studies, we shared similar goals in the need to evaluate the reliability of expert panel classification. Sittig et al conducted a study evaluating the diagnostic accuracy of a proposed screening tool for child abuse for use in the ED setting.[20] Their panel of 3 members with child abuse expertise scored the likelihood of abuse for 720 cases on a visual analog scale (0–100), and demonstrated strong reliability (intraclass correlation = 0.82 [0.80, 0.84]). We used a different statistical approach for reliability analysis, given the ordinal nature of our abuse likelihood measurement, but had similarly strong results. Our examination of interrater reliability in subgroups and pairs of panelists was unique, necessitated in part by the infeasibility of having the entire 9 member expert panel review all 2166 cases.

An important feature distinguishing our study from others was our ability to evaluate the accuracy of the expert panel using corroborative case information. Even though demonstrating interrater reliability is an important step in evaluating an expert panel, it is possible that the panel can be reliable but reach consistently incorrect conclusions. By demonstrating that our panelists were nearly perfectly accurate individually and in composite, we have mitigated such concerns.

Panelists provided classifications after reviewing case information in a standardized format and after answering a structured series of questions gauging history consistency, injury compatibility, and other case characteristics. Our results indicate that an expert panel is reliable and accurate in classifying abuse, but it is important to keep in mind that panelist expertise in child abuse and pediatric injury and structured reviews of case information were key features of our approach. Keenan et al advocates for much the same, indicating that implementing simple, structured tools like checklists can move child abuse pediatricians away from intuitive thinking to encourage more analytical thinking, in turn decreasing disparity of opinion among experts.[21] It is entirely possible that a panel composed of relatively inexperienced practitioners or those reviewing unstructured case information would fail to exhibit the reliability and accuracy exhibited by our panel of experts. Further research may study the performance of a panel with less experience in child abuse and pediatric injury, compare the accuracy and reliability of the inexperienced and experienced panels, and explore whether panel performance can be improved with rigorous education and training.

It is also noteworthy that we excluded psychosocial data from the case information given to panelists. Although such factors are critical when evaluating a child's overall safety, we did not want to risk influencing panelists' responses to specific questions regarding the nature of the injury. It is not known whether having such information would have impacted panelists' clinical decision outcomes, but our study goal was for panelists to classify the nature of the injury independent of the characteristics of the environment. The study by Keenan, et al[21] indicated that social information can be heavily influential in the diagnosis of child abuse under medical uncertainty. Further, they found that social information appeared to increase confidence in diagnoses but may have reduced reliance on more deliberative diagnostic processes (such as those based on medical information, case history, etc.). This, in turn, can lead to decreased agreement in diagnosis among child abuse pediatricians and may bias the decision-making process in cases of potential abuse.

Our study had limitations. In our analyses, we considered classifications of indeterminate to be ordinally intermediate to abuse and accident, i.e. "in the middle" of the spectrum from definite abuse to definite accident. An alternative approach would treat indeterminate classifications as completely outside this spectrum, much like missing data. Such an approach would require the development and use of missing data methods, like reweighting, [22–24] for reliability coefficients. Given the strong reliability of the panel and the relative infrequency of indeterminate classifications, it is not likely that such methods would produce results that substantially differed from our results. The use of the Kendall coefficient as the measure of reliability for the three-level classification can be questioned due to its highly discrete scale. We chose the Kendall coefficient to avoid the arbitrary assignment of weights required to calculate a weighted kappa coefficient. We evaluated the accuracy of the expert panel based upon the subset of cases with corroborative information. It is possible that these cases may have been among the easiest for panelists to classify, particularly in cases in which the injury history contained ostensibly corroborative information, potentially inflating the reported accuracy of the panel in classifying abuse. Further, the corroborated cases may not have been representative of the general population of bruised children presenting to pediatric emergency departments.

An important application of the expert panel approach is to serve as a criterion standard in the evaluation of clinical decision rules for the detection of child abuse. Very few criterion standards are perfect, and this is particularly true for those that involve human judgment. Our expert panel approach, although reliable and accurate, is no exception. Statistical methods exist for evaluating new clinical decision rules when criterion standards are imperfect.[25–29] A common feature of these methods is that the criterion standard imprecision propagates to the evaluation of a new decision rule. Specifically, criterion standards with higher error rates introduce more variability into the evaluation of a new decision rule, complicating inference about its diagnostic properties, such as sensitivity and specificity. Therefore, establishing a criterion standard with minimal classification error and imprecision is an important step in the process of evaluating new decision rules. We have demonstrated the reliability and accuracy of the expert panel approach and suggest that it provides a suitable criterion standard for the future evaluation of clinical decision rules for identifying child abuse.

Acknowledgments

We acknowledge the research enrollment team at Ann & Robert H. Lurie Children's Hospital of Chicago (Amanda Fingarson, DO; Emalee Flaherty, MD; Janet Flores, BS; Kristine Fortin, MD; Shelia Hickey, MSW, MJ; Kelly Liker, MD; Suzanne M. Schmidt, MD; Tara Shenoy, BA; Laura Turner Martinez, BS), Cincinnati Children's Hospital Medical Center (Berkeley Bennett, MD, MS; Kathi Makoroff, MD, MEd; Gabriel Meyers, BS), Rady Children's Hospital (Julia Magana, MD), University of Chicago Medicine Comer Children's Hospital (Shannon Staley, MD; Veena Ramaiah, MD), and Norton Children's Hospital (Melissa Currie, MD, Kendra Sikes, MS) for their extensive efforts in collecting data.

Supported by the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development (R01HD060997 to M.P.) and by The Grainger Foundation. B.H. received payment for expert witness record review and testimony related to children with concern for abuse. G.B. received payment for expert testimony in cases of pediatric injury. The Yale Department of Pediatrics received payment for J.L.'s expert testimony related to children with concern for abuse. University of Pittsburgh Physicians (UPP) received payment for R.B.'s expert testimony related to children with concern for abuse.

Appendix: Standardized questions each panelist was required to answer when reviewing cases

History Quality – Panelists select one of the following:

- History was consistent and thorough
- History has some inconsistencies, changes, or was somewhat vague
- History was inconsistent, changing, vague or absent
- History includes an explicit statement of physical assault

Injury Compatibility – Panelists select one of the following:

- Injury characteristics are accounted for by the history or reflect normal childhood activity
- Injury characteristics are not likely accounted for by the history or have some concerning features
- Injury characteristics are not accounted for by the history
- No history of injury provided and injury does not reflect normal childhood activity
- Injury characteristics are accounted for by the history but the history includes an explicit statement of physical assault

If injury is blamed on actions by a child (patient, sibling, other) – Panelists select one of the following ONLY IF injury is blamed on actions by a child:

- Child is developmentally capable of carrying out the described action(s)
- Unlikely that child is developmentally capable of carrying out the described action(s)
- Child is developmentally incapable of carrying out the described action(s)
- N/A; story does not involve a child's action(s)

Patient's behavior and actions after the event – Panelists select one of the following:

- Child's described behavior and actions after the incident/injury was consistent with the pathophysiological constraints and/or pain produced by the injury
- Child's described behavior and actions after the incident/injury may not be consistent with the pathophysiological constraints and/or pain produced by the injury
- Child's described behavior and actions after the incident/injury was inconsistent with the pathophysiological constraints and/or pain produced by the injury
- NA; no specific event and/or information provided

Other acute injuries present (excluding bruises/abrasions/lacerations)

- Are there any other acute injuries present (excluding bruises/abrasions/lacerations)? Yes/No
- *If yes, then select one of the following…*
  - Are multiple (more than one) acute injuries present (excluding bruises/abrasions/lacerations)?
  - Other acute injury/injuries explained by history
  - Other acute injury/injuries not explained by history or history includes statement of assault
  - Other acute injury/injuries are not likely accounted for by the history or have some concerning features
  - Insufficient Information

Healing injuries present (excluding bruises)

- Are healing injuries present (excluding bruises)? Yes/No
- *If yes, then select one of the following…*
  - Healing injuries; explained by history
  - Healing injuries; not explained by history or history includes explicit statement of physical assault
  - Healing injuries; not likely accounted for by the history or have some concerning features
  - Insufficient Information

Bruises Previously Documented

- Were bruises previously documented or observed on this child? Yes/No
- *If yes, then select one of the following…*
    - Prior bruises explained by history, consistent with ambulatory state, and no recognizable pattern present
    - Prior bruises; not explained by history, inconsistent with ambulatory state, or recognizable pattern present or history includes explicit statement of physical assault
    - Insufficient Information
    - Prior bruises not likely accounted for by the history or have some concerning features

Prior Traumas or Medical Events

- Are there any prior traumas or medical events that likely reflect or are suspicious of abuse? Yes*/No*
- Was abuse likely missed by the medical community? Yes/No

Timing of when care was sought – Panelists select one of the following:

- Sought care immediately or soon after the incident or trauma
- Delay in seeking treatment, but signs or symptoms of injury or medical condition are subtle
- Delay in seeking treatment in serious injury or medical condition
- Delay in seeking treatment in severe or life-threatening injury or medical condition

Panelists may also choose to select one or both of the following:

- Medical attention sought because person other than caregiver(s) at the time of incident has concerns
- Caregiver with the child at the onset of signs, symptoms, or trauma dissuades others from seeking medical care

Case Comments – Free response by panelists

Indicated Studies – Panelists select any that apply:

- None
- Social work consult
- Skeletal survey
- Follow up skeletal survey
- Brain imaging

- Abdominal imaging
- Trauma labs
- Eye exam

Conclusions & Actions Based on Medical and Historical Data - Panelists select one of the following:

- Clinically determined to be consistent with definite abuse/inflicted trauma
- Clinically determined to be consistent with likely abuse/inflicted trauma
- Indeterminate
- Clinically determined to be consistent with likely accident
- Clinically determined to be consistent with definite accident

Report to Child Protective Services:

- Is a report to child protective services indicated? Yes/No

Strength of the Evidence - Panelists select one of the following:

- Evidence supports testimony in a court of law that the case is most consistent with accidental injury and plausible
- If trauma history provided, evidence supports testimony in a court of law that the case is inconsistent with the history of trauma provided and/or is not plausible and is most consistent with inflicted trauma
- Evidence supports testimony in a court of law that the case is most consistent with inflicted trauma
- Evidence supports testimony in a court of law that the injury is consistent with the reported assault
- Further evidence/investigation is needed in order to testify

Abbreviations

PED	pediatric emergency department
CAP	child abuse pediatrician
PEM	pediatric emergency medicine
CPS	child protective services

References

1. U.S. Department of Health and Human Services, Administration for Children and Families, Administration on Children, Youth and Families, Children's Bureau. Child Maltreatment 2015. 2017. Available at: http://www.acf.hhs.gov/programs/cb/resource/child-maltreatment-2015. Accessed February 6, 2017

2. Jenny C, Hymel KP, Ritzen A, Reinert SE, Hay TC. Analysis of missed cases of abusive head trauma. JAMA. 1999 Feb.281:621–626. [PubMed: 10029123]
3. Vinchon M, Foort-Dhellemmes S, Desurmont M, Delestret I. Confessed abuse versus witnessed accidents in infants: comparison of clinical, radiological, and ophthalmological data in corroborated cases. Childs Nerv Syst. 2010; 26:637–645. [PubMed: 19946688]
4. Louwers ECFM, Korfage IJ, Affourtit MJ, Scheewe DJ, van de Merwe MH, Vooijs-Moulaert AF, et al. Effects of systematic screening and detection of child abuse in emergency departments. Pediatrics. 2012; 130:457–464. [PubMed: 22926179]
5. Ponte C, Craven A, Robson J, Grayson PC, Suppiah R, Watts RA, et al. Review of the expert panel methodology in the diagnostic and classification criteria for vasculitis study: a pilot study. Rheumatology. 2014; 53:i15–i16.
6. Willinger M, James LS, Catz C. Defining the Sudden Infant Death Syndrome (SIDS): Deliberations of an Expert Panel Convened by the National Institute of Child Health and Human Development. Pediatric Pathology. 1991; 11:677–684. [PubMed: 1745639]
7. Dasgupta B, Salvarani C, Schirmer M, Crowson CS, Maradit-Kremers H, Hutchings A, members of the American College of Rheumatology Work Group for Development of Classification Criteria for PMR. Developing classification criteria for polymyalgia rheumatica: comparison of views from an expert panel and wider survey. The Journal of Rheumatology February. 2008; 35:270–277.
8. Posner K, Oquendo MA, Gould M, Stanley B, Davies M. Columbia Classification Algorithm of Suicide Assessment (C-CASA): Classification of Suicidal Events in the FDA's Pediatric Suicidal Risk Analysis of Antidepressants. Amer J Psych. 2007; 164:1035–43.
9. Leventhal JM, Thomas SA, Rosenfield NS, Markowitz RI. Fractures in young children: Distinguishing child abuse from unintentional injuries. Am J Dis Child. 1993 Jan.147:87–92. [PubMed: 8418609]
10. Thomas SA, Rosenfield NS, Leventhal JM, Markowitz RI. Long-bone fractures in young children: distinguishing accidental injuries from child abuse. Pediatrics. 1991 Sep.88:471–476. [PubMed: 1881725]
11. Strait RT, Siegel RM, Shapiro RA. Humeral fractures without obvious etiologies in children less than 3 years of age: when is it abuse? Pediatrics. 1995 Oct.96:667–671. [PubMed: 7567328]
12. Wood JN, Fakeye O, Mondestin V, Rubin DM, Localio R, Feudtner C. Development of hospital-based guidelines for skeletal survey in young children with bruises. Pediatrics. 2015 Jan.135:e313–e320.
13. Lindberg DM, Lindsell CJ, Shapiro RA. Variability in expert assessments of child physical abuse likelihood. Pediatrics. 2008 Apr.121:e945–e953. [PubMed: 18381522]
14. Bartelink C, van Yperen TA, ten Berge IJ, de Kwaadsteniet L, Witteman CLM. Agreement on child maltreatment decisions: A nonrandomized study on the effects of structured decision-making. Child & Youth Care Forum. 2014; 43:639–654.
15. Pierce MC, Magana JN, Kaczor K, Lorenz DJ, Meyers G, Bennett BL, et al. The Prevalence of Bruising Among Infants in Pediatric Emergency Departments. Ann Emerg Med. 2016; 67:1–8. [PubMed: 26233923]
16. Valley MA, Heard KJ, Ginde AA, Lezotte DC, Lowenstein SR. Observational studies of patients in the emergency department: a comparison of four sampling methods. Ann Emerg Med. 2012; 60:139–145. [PubMed: 22401950]
17. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing; Vienna, Austria: 2016. URL http://www.R-project.org/.
18. Laskey AL, Sheridan MJ, Hymel KP. Physicians' initial forensic impressions of hypothetical cases of pediatric traumatic brain injury. Child Abuse Negl. 2007; 31:329–342. [PubMed: 17408739]
19. Buesser KE, Leventhal JM, Gaither JR, Tate V, Cooperman DR, Moles RL, et al. Inter-rater reliability of physical abuse determinations in young children with fractures. Child Abuse Negl. 2017; 72:140–146. [PubMed: 28802910]
20. Sittig JS, Uiterwaal CSPM, Moons KGM, Russel IMB, Nievelstein RAJ, Nieuwenhuis EES, et al. Value of systematic detection of physical child abuse at emergency rooms: a cross-sectional diagnostic accuracy study. BMJ Open. 2016; 6:e010788.

21. Keenan HT, Cook LJ, Olson LM, Bardsley T, Campbell KA. Social Intuition and Social Information in Physical Child Abuse Evaluation and Diagnosis. Pediatrics. 2017; 140:e20171188. [PubMed: 29074609]

22. Lorenz DJ, Datta S, Harkema SJ. Marginal association measures for clustered data. Stat Med. Nov 30.2011 30:3181–91. [PubMed: 21953204]

23. Lorenz DJ, Levy S, Datta S. Inferring marginal association with paired and unpaired clustered data. Statistical Methods in Medical Research. 2016 Sep 20. pii:0962280216669184.

24. Chen Z, Xie Y. Marginal analysis of measurement agreement among multiple raters with non-ignorable missing ratings. Statistics and Its Interface. 2014; 7:113–120.

25. Walter SD, Irwig L, Glasziou PP. Meta-analysis of diagnostic tests with imperfect reference standards. J Clin Epidemiol. 1999 Oct.52:943–951. [PubMed: 10513757]

26. Chu H, Chen S, Louis TA. Random Effects Models in a Meta-Analysis of the Accuracy of Two Diagnostic Tests Without a Gold Standard. J Am Stat Assoc. 2009 Jun 1.104:512–523. [PubMed: 19562044]

27. Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. Journal of Clinical Epidemiology. 2005; 58:982–990. [PubMed: 16168343]

28. Sadatsafavi M, Shahidi N, Marra F, FitzGerald MJ, Elwood KR, Guo N, et al. A statistical method was used for the meta-analysis of tests for latent TB in the absence of a gold standard, combining random effect and latent-class methods to estimate test accuracy. Journal of Clinical Epidemiology. 2010; 63:257–269. [PubMed: 19692208]

29. Zhou, XH., Obuchowski, NA., Obuchowski, DM. Statistical Methods in Diagnostic Medicine. New York: John Wiley and Sons; 2002.

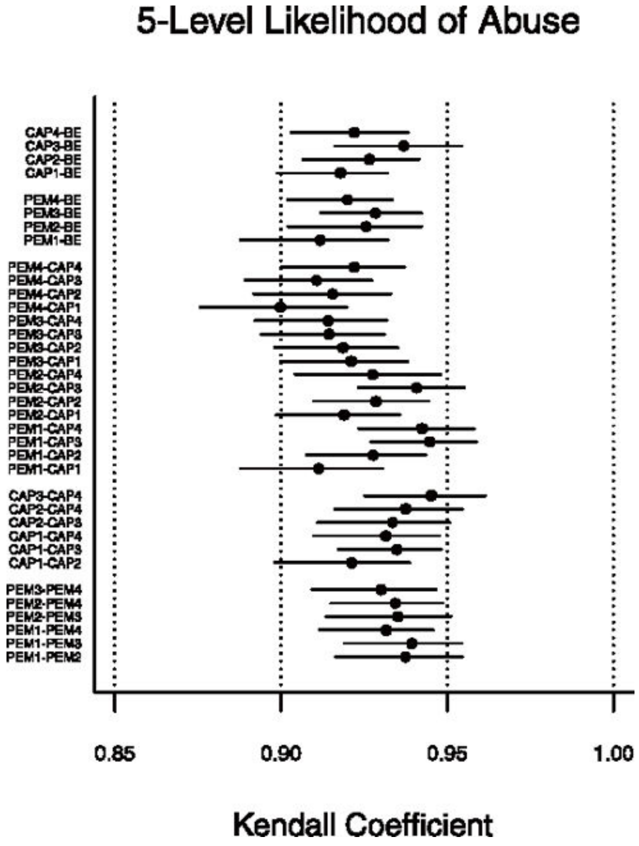

Figure 1.

Diagram of case subsets used for different reliability and accuracy analyses. 58 of the 201 randomly test cases had corroborative information (29 abuse, 29 accident).

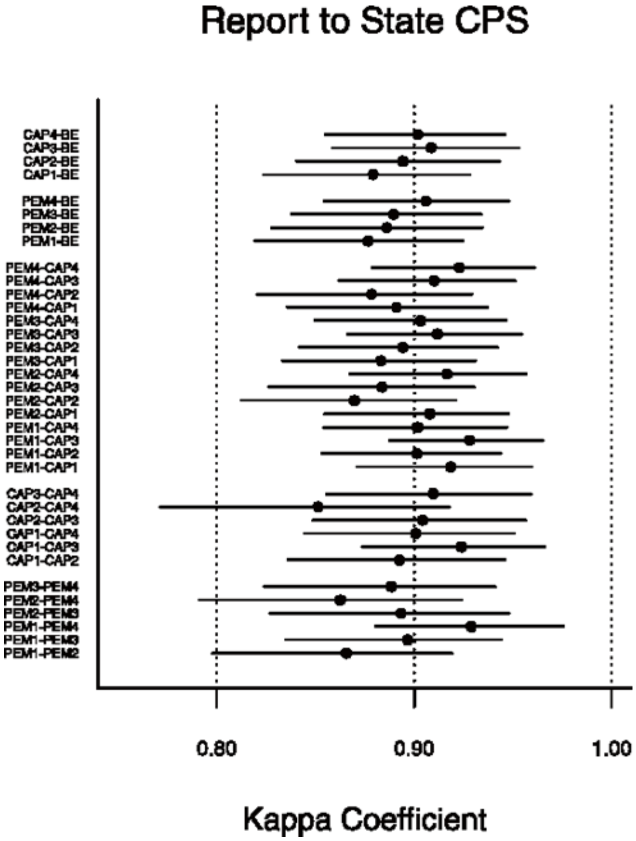

Figure 2.

Schematic diagram of the derivation of the three-level and binary versions of the ordinal likelihood of abuse.

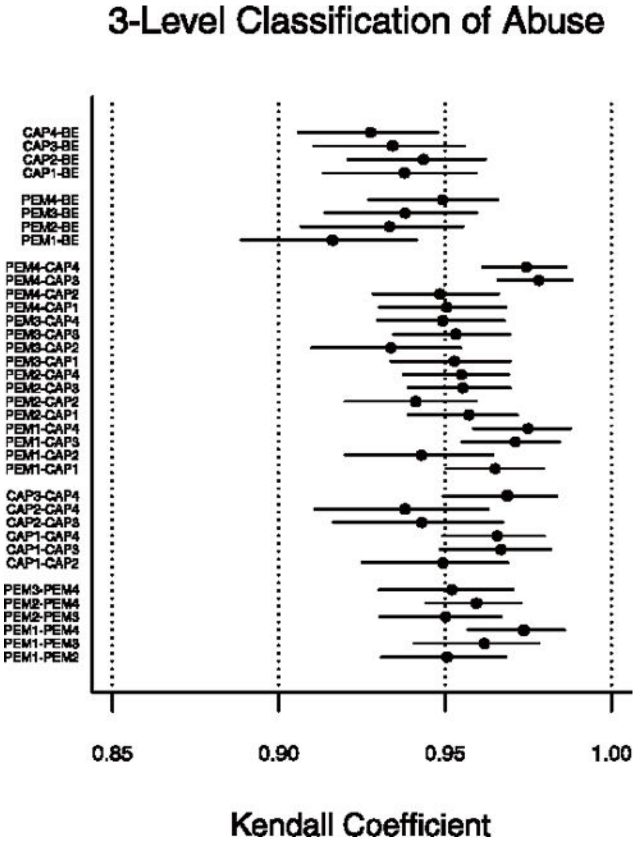

Figure 3.

Reliability coefficients for pairwise evaluation of expert panelists. Plotted points are Kendall coefficients for the ordinal likelihood of abuse classification and kappa coefficients for the report to child protective services. Horizontal lines represent 95% confidence intervals.

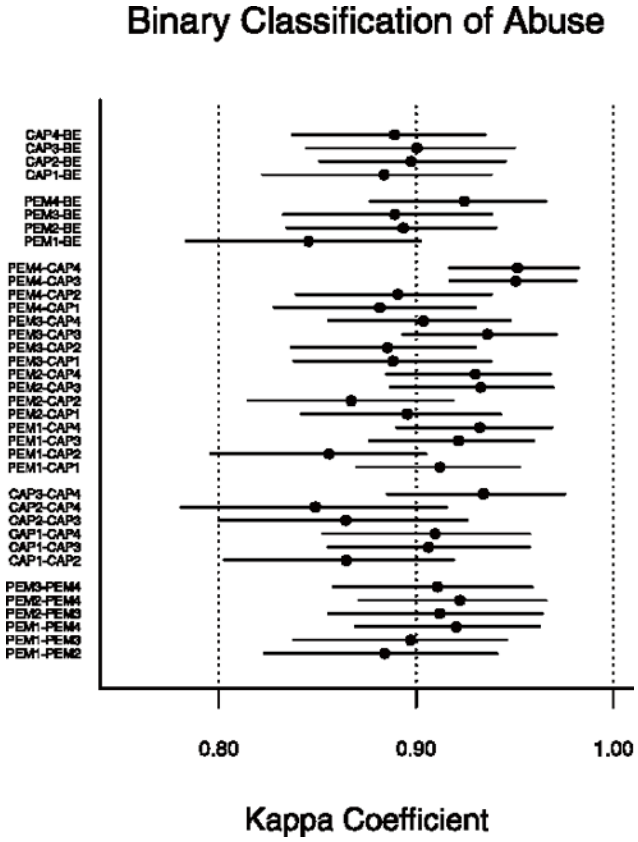

Figure 4.

Reliability coefficients for pairwise evaluation of expert panelists. Plotted points are Kendall coefficients for the three-level likelihood of abuse classification and kappa coefficients for the binary classification of abuse. Horizontal lines represent 95% confidence intervals.

Table 1

Summary of ratings of 2166 cases from 9 expert panelists. Values are counts and within-column percentages. Panelists are numbered within panelist groups. PEM = Pediatric Emergency Medicine, CAP = Child Abuse Pediatricians, BE = Bioengineer. There were 5 instances in which panelists did not provide an answer to the report to child protective services question (PEM2, CAP1, CAP2, BE).

Classification	**PEM1**	**PEM2**	**PEM3**	**PEM4**	**CAP1**	**CAP2**	**CAP3**	**CAP4**	**BE**
(Cases reviewed)	**(751)**	**(742)**	**(739)**	**(741)**	**(742)**	**(744)**	**(744)**	**(739)**	**(685)**
Definite Abuse	126 (17%)	117 (16%)	118 (16%)	77 (10%)	86 (12%)	64 (9%)	101 (14%)	91 (12%)	148 (22%)
Likely Abuse	38 (5%)	42 (6%)	47 (6%)	82 (11%)	62 (8%)	94 (13%)	42 (6%)	44 (6%)	29 (4%)
Indeterminate	22 (3%)	73 (10%)	44 (6%)	21 (3%)	55 (7%)	48 (6%)	47 (6%)	39 (5%)	46 (7%)
Likely Accident	62 (8%)	459 (62%)	360 (49%)	444 (60%)	367 (49%)	496 (67%)	69 (9%)	33 (4%)	75 (11%)
Definite Accident	503 (67%)	51 (7%)	170 (23%)	117 (16%)	172 (23%)	42 (6%)	485 (65%)	532 (72%)	387 (56%)
Report to Child	190/751	200/741	181/739	197/741	202/741	186/742	192/744	184/739	198/684
Protective Services	(25%)	(27%)	(24%)	(27%)	(27%)	(25%)	(26%)	(25%)	(29%)

Table 2

Reliability coefficients for 201 test cases for the full expert panel and subgroups of panelists. Values are Kendall coefficients (5- and 3-level classification) or kappa coefficients (Report to CPS, Binary classification) with 95% confidence intervals.

Panelist Subgroup	5-level classification	Report to CPS	3-level classification	Binary classification
Full Panel	0.89 (0.87, 0.91)	0.91 (0.86, 0.94)	0.92 (0.90, 0.94)	0.91 (0.87, 0.95)
PEM only	0.91 (0.88, 0.92)	0.90 (0.86, 0.95)	0.94 (0.91, 0.96)	0.92 (0.88, 0.96)
PEM + CAP	0.90 (0.87, 0.91)	0.91 (0.86, 0.94)	0.93 (0.91, 0.95)	0.92 (0.88, 0.95)
PEM + BE	0.90 (0.87, 0.92)	0.90 (0.85, 0.94)	0.92 (0.89, 0.94)	0.91 (0.87, 0.95)
CAP only	0.92 (0.90, 0.93)	0.90 (0.86, 0.94)	0.95 (0.92, 0.97)	0.91 (0.87, 0.95)
CAP + BE	0.91 (0.89, 0.93)	0.91 (0.87, 0.94)	0.93 (0.91, 0.96)	0.91 (0.87, 0.95)

Table 3

Summary of ratings of cases meeting corroborative criteria for abuse and accident from 9 expert panelists. Values are counts and within-column percentages. Denominators differ among panelists since corroborated cases were not systematically assigned to panelists.

Panelist	Corroborated Abuse	Corroborated Accident
PEM1	63/63 (100%)	134/135 (99%)
PEM2	65/65 (100%)	131/131 (100%)
PEM3	58/59 (98%)	134/135 (99%)
PEM4	61/64 (95%)	142/142 (100%)
CAP1	59/59 (100%)	134/135 (99%)
CAP2	66/66 (100%)	142/144 (99%)
CAP3	65/67 (97%)	134/134 (100%)
CAP4	55/56 (98%)	127/128 (99%)
BE	66/67 (99%)	118/119 (99%)
Panel Composite	147/148 (99%)	435/436 (99%)