ARTICLE

Variability in Expert Assessments of Child Physical Abuse Likelihood

Daniel Martin Lindberg, MD[a], Christopher John Lindsell, PhD[b], Robert Allan Shapiro, MD[c]

[a]Department of Emergency Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts; [b]Department of Emergency Medicine, University Hospital–Cincinnati, and [c]Department of Pediatrics, Mayerson Center for Safe and Healthy Children, Cincinnati Children's Hospital Medical Center, University of Cincinnati, Cincinnati, Ohio

The authors have indicated they have no financial relationships relevant to this article to disclose.

What's Known on This Subject

Determinations of child physical abuse are contentious and often controversial. Several scales have been proposed to categorize the degree of certainty in child abuse diagnoses.

What This Study Adds

Even among child abuse experts, assessments of abuse likelihood often show broad variability. One scale showed a moderate decrease in this variability. Experts showed moderate agreement about what constitutes "reasonable suspicion" of abuse.

ABSTRACT

OBJECTIVES. In the absence of a gold standard, clinicians and researchers often categorize their opinions of the likelihood of inflicted injury using several ordinal scales. The objective of this protocol was to determine the reliability of expert ratings using several of these scales.

METHODS. Participants were pediatricians with substantial academic and clinical activity in the evaluation of children with concerns for physical abuse. The facts from several cases that were referred to 1 hospital's child abuse team were abstracted and recorded as in a multidisciplinary team conference. Participants viewed the recording and rated each case using several scales of child abuse likelihood.

RESULTS. Participants ($n = 22$) showed broad variability for most cases on all scales. Variability was lowest for cases with the highest aggregate concern for abuse. One scale that included examples of cases fitting each category and standard reporting language to summarize results showed a modest (18%–23%) decrease in variability among participants. The interpretation of the categories used by the scales was more consistent. Cases were rarely rated as "definite abuse" when likelihood was estimated at ≤95%. Only 7 of 156 cases rated ≤15% likelihood were rated as "no reasonable concern for abuse." Only 9 of 858 cases rated ≥35% likelihood were rated as "reasonable concern for abuse."

CONCLUSIONS. Assessments of child abuse likelihood often show broad variability between experts. Although a rating scale with patient examples and standard reporting language may decrease variability, clinicians and researchers should be cautious when interpreting abuse likelihood assessments from a single expert. These data support the peer-review or multidisciplinary team approach to child abuse assessments.

www.pediatrics.org/cgi/doi/10.1542/peds.2007-2485

doi:10.1542/peds.2007-2485

Key Words
child abuse, interobserver variation, diagnosis

Accepted for publication Sep 26, 2007

Address correspondence to Daniel Martin Lindberg, MD, 75 Francis St, Neville House, Department of Emergency Medicine, Brigham and Women's Hospital, Boston, MA 02115. E-mail: dlindberg@partners.org

PHYSICIANS MAKE DIAGNOSES with varying degrees of certainty. When the consequences of an incorrect diagnosis are serious, diagnostic certainty should be maximized; however, there are situations (eg, potential child abuse) in which the risks are great, yet no gold standard test exists to maximize certainty. To avoid the serious consequences of incorrectly diagnosing or excluding child physical abuse, clinicians must understand and communicate accurately the degree of diagnostic certainty to patients, families, prosecutors, jurors, law enforcement, and social services, who in turn must be able to recognize when an acceptable level of diagnostic certainty exists and when it does not.

In the absence of a gold standard for the diagnosis of child physical abuse, several scales have been devised to facilitate categorizing the certainty of abuse,[1–5] but none has demonstrated consistent agreement among multiple raters. In 1 study, a collection of fictional vignettes of children with head injury, written to approximate the clinical scenarios encountered by child abuse professionals, resulted in widely divergent opinions of the likelihood of abuse when rated by pediatricians and pathologists with experience in evaluating cases of possible child physical abuse.[6] Divergent opinions could have been the result of either different conceptions of abuse likelihood or different understandings of the categories in the scale that was tested. For example, whereas 1 expert may consider a child with a 95% likelihood of abuse in the "presumptive inflicted" category, another may consider that sufficient to fit into the category of "definitive inflicted." These 2 diagnoses, although similar, could result in significantly different

investigative, social, and legal consequences. Furthermore, although many statutes require a physician to report a case to children's services when there is a "reasonable suspicion of inflicted injury," the standards for determining reasonable suspicion are not further defined and have been interpreted differently by different clinicians.[7,8]

METHODS

Study Design

Study participants viewed prerecorded video case vignettes and rated them using several scales developed for evaluating the likelihood of child physical abuse. This study was approved by the institutional review board of the Cincinnati Children's Hospital Medical Center.

Recruitment of Participants

Participants were recruited from the Helfer Society, an international honor society of physicians with substantial professional involvement and expertise in the field of child abuse pediatrics. The group is composed of nearly 200 physicians, predominantly pediatricians. Participants were recruited by distribution of paper fliers at a national meeting of the society and by distribution of a request to participate to the group's electronic discussion group. Requests to participate were sent by e-mail 4 times at monthly intervals. Invitations included a brief description of the study and a statement of the expected time to complete the assessments. For maximization of enrollment, participants were offered a chance to enter a drawing for an Apple iPod and were told that they would receive a summary of the aggregate results. It was also stated that responses would not be identifiable to investigators and that participants would select their own unique identifiers to facilitate anonymous communication. Recruitment began on November 11, 2006, and was closed on April 15, 2007.

Case Vignettes

Fifty-five cases were abstracted from clinical consultations with children who were younger than 24 months and had been referred to the child protection team of a large, tertiary-care, academic pediatric medical center for concern of physical abuse. Cases were selected for presentation at random from among all records in the center's electronic database. Once selected and abstracted, the cases were put into random order and were presented to participants. Each participant was asked to rate each case. Cases were neither selected nor excluded because of the degree of certainty of the treating clinicians or other clinical factors.

The recorded vignettes included the child's age, the reason given for the child's presentation, any given history of trauma, the list of identified injuries, and results of diagnostic testing cited in the assessment of the original consulting team. For the sake of brevity, results of normal diagnostic studies (eg, coagulation studies, genetic testing for osteogenesis imperfecta) were not routinely given unless mentioned in the summary assessment of the original consulting team. Participants were instructed to assume that unmentioned diagnostic studies had been performed and were normal. Radiographs and de-identified photographs of significant injuries or mimics (eg, potential fractures, intracranial injuries, retinal hemorrhages, bruising) were included in case presentations. Demographic data such as race, ethnicity, and socioeconomic status were not included in the vignettes. The recorded vignettes did not include impressions of caregivers' demeanor, affect, or interaction with the child. Cases included all information available at the time the child abuse team made its final impression. Cases sometimes contained data about past or future incidents of abuse, which were investigated separately. Although information about future abuse is not available in clinical practice, child abuse consultants are sometimes asked to opine in legal settings as to the likely cause of previous injuries in children who are subsequently identified as having been abused. Participants were asked to consider this information but to restrict their ratings to the likelihood of an inflicted injury in the index episode. For example, it might be reasonable to consider a given injury more suspicious if a child had been injured by the same caregiver at another time, but other abuse would not automatically mean that a child was abused in the presenting episode.

Cases were initially presented to a pilot group of 5 child abuse experts via a synchronized audio and Internet conference (WebEx, Santa Clara, CA). Cases were presented in a manner similar to a multidisciplinary conference, but participants did not discuss their opinions with each other. Participants were allowed to ask questions but were asked to avoid revealing their own impressions of any given case as much as possible. Case presentations were video and audio recorded and may be viewed in their entirety at http://seraph.cchmc.org/MediasiteEX/Viewer/?peid=d2b185bc-aa59-47b6-8244-01c7c5f19156. Subsequent participants were able to view the recorded conference, including the questions asked by the participating panel, but were not able to ask additional questions. Rating was undertaken in a manner similar to the original conference. Responses of the 5 pilot participants were included with those of future participants and analyzed together. Participants faxed anonymous completed rating sheets to the principal investigator.

Rating Scales

Each participant was asked to rate the likelihood of child physical abuse in each case using each of 3 scales and by percentage likelihood. Scales were adapted from those used clinically in a variety of settings and from the format of several scales used to dichotomize cases for research. Scale A (Table 1), used by the clinical team that initially evaluated the patients, allows the consulting physician to choose a summary statement based on their ultimate impression of the case without restriction.

Scale B (Table 2) simplifies gradations of suspicion to the decision of whether to report to children's services. Although cases that are referred for child abuse consultation are often reported to children's services for issues of social support or poor supervision, participants were asked to restrict their answers to the decision to report

TABLE 1 **Scale A**

1. Not concerning for inflicted injury
2. Nonspecific/possible inflicted injury
3. Concerning/probable inflicted injury
4. Definite inflicted injury
5. Unable to determine

TABLE 2 **Scale B**

Rating	Action
1. No reasonable concern for abuse	Do not report to children's services
2. Unable to determine	Need more information
3. Reasonable concern for abuse	Report to children's services

for suspicion of inflicted injury alone; that is, a child who had a clearly accidental injury and may normally be referred to children's services for food stamps or supervision issues would still be rated as "no reasonable concern for abuse" according to this scale.

Scale C (Table 3) emulated the approach of several published scales[2–5] that use several ordered, criteria-defined categories. Each category was linked to examples of patients that would fit the category. Examples were based on the authors' understanding of the current literature but were meant to be guidelines rather than restrictions; that is, although patterned bruising was listed as an example of "definite abuse," a participant was encouraged to describe a child with patterned bruising using another category if they believed that the child had not definitely been abused. Terms such as "patterned bruising" and "cruising child" were not further defined for participants.

Finally, participants were asked to rate each case using a percentage: "Assume 100 children presented with the given set of facts. To the best of your ability, please estimate how many would have been victims of inflicted injury."

Participants also answered 3 final demographic and summary questions. First, participants were given a 5-point Likert scale to rate their agreement with the following sentence (5 = strongly agree): "I was given enough information to come to a reasonable conclusion." Participants were also asked for how many years they had been providing physical abuse consultation and for their average monthly census, or approximately how many consultations for physical abuse they participated in monthly (0, 1–2, 2–5, 5–10, or >10).

Statistical Analysis

For determination of the magnitude of interrater variability (reproducibility), the mean, minimum, maximum, and SD across participants were determined for each case using each scale. In addition, for each scale, the mean rating across cases was computed for each participant. Scale A was not truly ordinal; the value 5 (unable to determine whether abuse occurred) could not be ordered within the remaining available responses. Because this was used infrequently (4 participants used this determination at least once, with a total of 15 responses being 5), these responses were not included for numeric analysis.

Between-scale correlation was determined using Pearson's correlation coefficient. Because a correlation does not provide information on magnitude differences between scales (ie, is 1 scale likely to result in a lower estimated probability of abuse than another, even if correlated), between-scale differences were also evaluated. Responses were normalized to a 0 to 100 scale. For a given response x, the transformed response x^* was related by the formula $x^* = (x - 1)(100/n - 1)$, where n is the number of possible responses for the scale. For example, responses of 1 (not inflicted injury), 5 (very concerning for inflicted injury), and 7 (definite inflicted injury) on scale C would have been transformed to 0, 66.7, and 100, respectively. The mean and SD rating across participants were computed for each case using each scale. These means and SDs were compared between scales using a repeated measures analysis of variance with posthoc paired-samples t test. Analyses were conducted using SPSS 14.0 (SPSS Inc, Chicago, IL) and Microsoft Excel (Microsoft Corp, Redmond, WA).

RESULTS

Twenty-two participants completed the exercise and submitted answers. Two participants did not complete the percentage likelihood scale. Table 4 shows the self-reported characteristics of the participants. As a group, there was moderate agreement with the statement, "I was given enough information to come to a reasonable conclusion." There was no correlation between agreement with this statement and subsequent ratings. Participants had a range of experience and clinical census, but none reported participating in fewer than 2 to 5 physical abuse consultations monthly.

To determine whether participants were consistently more or less concerned for abuse relative to their peers, we ranked each participant's mean rating for each scale across all cases (Table 5). Participants with the lowest ranks had the lowest ratings (ie, lowest mean concern for abuse), whereas higher numbered ranks represent more concern for abuse. Whereas some participants were consistently ranked as more or less conservative on all scales, most had ranks that differed by the scale used. This method does not take into account the magnitude of differences between ranks. Ranking did not correlate with the monthly census reported by participants ($r = 0.253$, $P = .327$) or the number of years of experience ($r = 0.141$, $P = .588$). Because some participants may have an especially long or short tenure of experience relative to other Helfer Society members, experience is reported dichotomously to protect participant confidentiality.

Randomly selected cases nevertheless covered a broad spectrum of aggregate abuse likelihood as shown by the mean percentage likelihood responses from all participants (Fig 1); however, individual percentage ratings tended to gravitate toward the poles with nearly half of estimates outside the range between 10% and 90%

TABLE 3 **Scale C**

Rating	Criteria	Explanation
1. Definitely not inflicted injury	Significant, independently verifiable mechanism (MVC, pedestrian struck) Disinterested witness (police, ambulance, video documentation) Mimic (Mongolian spot, hemangioma)	Although no evaluation can completely exclude abuse, our evaluation has not raised a reasonable suspicion of abuse. The injuries or findings that we have described could reasonably be explained by accidental or benign events. Please do not hesitate to renew discussion if circumstances change. (ratings 1 and 2)
2. No concern for inflicted injury	Mechanism explains all injuries, consistent history	
3. Mildly concerning for inflicted injury	Somewhat concerning injuries with no offered history (multiple, nonpatterned bruises in a cruising child without bleeding diathesis, unexplained humerus fracture in 10-mo-old) Otherwise unconcerning injury with past suspicious injury and same caregiver	
4. Intermediately concerning for inflicted injury	Insufficient information to offer opinion Sequence of events clear, but uncertain whether they constitute abuse Necessary laboratory tests/consultation pending Concerning injury in the setting of bone fragility/bleeding diathesis	
5. Very concerning for inflicted injury	Given history unlikely to produce documented injuries Concerning injury with no history of trauma (4-mo-old with unexplained femur fracture)	
6. Substantial evidence of inflicted injury	Severe injury with no offered history in a child incapable of inflicting the injury on himself or herself History inconsistent with identified injuries Serious injuring with changing history or history inconsistent between caregivers Inappropriate delay in seeking care Multiple severe injuries of different age without plausible explanation	
7. Definite inflicted injury	Pattern bruises/burns Unexplained posterior rib fractures, metaphyseal fractures, characteristic retinal hemorrhages Highly suspicious injury (liver laceration, burn, pinna bruising, unexplained fracture) with definite subsequent abuse Reliable eyewitness of abuse Suspicious injury and concurrently abused sibling Obvious injury with significant, unexplained delay in seeking care (serious burn, unresponsive child, apparent prolonged seizure)	To a reasonable degree of medical certainty, the injuries/findings that we have described cannot plausibly be explained by accidental injury, preexisting medical illness, reasonable discipline, or benign events. (ratings 6 and 7)

MVC indicates motor vehicle crash.

(Fig 2). Similar data for each ordinal scale are shown in Fig 3.

Correlation between scales was very high (Table 6), with the exception of scale B. Because scale B has only 3 categories and most cases were considered to pass threshold for reporting, it is not surprising that correlation was low with this scale. Interrater variability within each scale was compared between scales. Table 7 shows the comparison of the means and SDs across participants for each case. Scale C showed significantly less variability (ie, a smaller SD across participants) compared with all other scales.

All scales demonstrated broad variability across participants for most cases (Fig 4). Whereas some cases with the highest aggregate concern for abuse showed substantial agreement, most cases generated responses from both ends of the spectrum, with some experts nearly sure that the child had been abused and others as sure that the child had not. Interrater variability was lowest in cases with the highest aggregate concern for abuse and highest for cases with intermediate concern.

Although participants showed broad variability in their ratings of most cases, they seemed to share a consistent understanding of some of the categories for each scale. Participants rarely described cases as "definite abuse" when their concern for inflicted injury was $<95\%$ using either scale A or scale C. Using responses on scale B as a proxy for the decision to report, respondents were unlikely to report when there was $<15\%$ concern for abuse but very likely to report for cases for which they estimated the likelihood of abuse at $>30\%$. Of 156 responses rated as $\leq 15\%$ likelihood of abuse, only 7

TABLE 4 **Self-reported Characteristics of Participants**

Parameter	Value
Enough information	
5: Strongly agree	5
4.5	1
4: Agree	11
3: Don't know/no opinion	3
2: Disagree	1
No response	1
Monthly census	
2–5	6
5–10	5
>10	11
Years of experience	
Median	10
Range	2.5 to >30

Participants answered 3 questions on completing the exercise. By using a Likert scale, they rated their agreement with the statement, "I was given enough information to come to a reasonable conclusion." Monthly census was defined as the average number of consultations for physical abuse in which the participant was involved.

were considered "reasonable concern for abuse." Of 858 cases rated as ≥30% likelihood of abuse, only 9 were also described as "no reasonable concern for abuse."

DISCUSSION

The cases we reviewed had, in aggregate, a spectrum of concern for abuse; however, cases with an intermediate aggregate concern for abuse were often the result of the averaging of widely divergent ratings at either end of the spectrum rather than substantial agreement that the case was intermediately concerning. This propensity of participants to cluster their responses at the ends of the spectrum is understandable in view of the clinical task at hand. In the end, each child presented either was or was not the victim of an inflicted injury. Even if we were to accept that 64% of all head injuries in infants are the result of abuse,[9] there are no children who have been "64% abused." Neither can one reasonably make 64% of a report to children's services. Nevertheless, we believe that our vignettes represent a spectrum of abuse similar to what is encountered by clinicians, researchers, and those called on for expert testimony.

All scales demonstrated better agreement with respect to cases with very high concern for abuse than for cases with intermediate aggregate concern for abuse. Scale C demonstrated a modest but significant reduction in variability of assessments taken across all likelihoods of abuse. When compared with scale A and the percentage likelihood of abuse scale, scale C showed an absolute decrease in the average SD of 3.79% and 2.91%, respectively. This represents 23% and 18% of the average variability in each case.

It bears emphasis that this is a study of agreement and not of validity. Although perfect agreement using any scale might demonstrate that experts in the evaluation of child abuse may agree on the significance of a given finding, it would not by itself prove that any finding actually constitutes evidence of abuse.

Our results reflect a broad degree of variability in abuse diagnosis among clinicians with substantial clinical activity

TABLE 5 **Variability in Concern for Abuse Cases: Rankings of Participants**

Mean Rank	Scale A	Scale B	Scale C	%	Monthly Census	Experience
3[a]	2	1	3	5	>10	10+
5[a]	6	5	5	2	>10	10+
5	1	15	2	1	>10	<10
5	3	6	1	11	2–5	10+
7	9	2	12	3	2–5	10+
7	7	8	7		5–10	10+
9	4	14	10	6	2–5	<10
9	5	4	11	14	5–10	10+
9	10	10	13	4	>10	10+
10	8	3	8	20	>10	<10
11	13	9	15	8	2–5	10+
12	11	13	6	16	>10	10+
12	14	18	9	7	2–5	10+
13	15	11	16	10	>10	10+
14	12	17	4	22	5–10	<10
14	17	7	19		>10	<10
16	16	19	21	9	>10	<10
17	20	12	20	15	2–5	<10
18[a]	19	16	18	19	5–10	<10
19[a]	21	21	14	18	5–10	10+
20[a]	18	22	22	17	>10	10+
20[a]	22	20	17	21	>10	10+

Each participant was ranked for each scale according to his or her mean rating across cases. Higher ranks indicate higher average rating (ie, more concern for abuse). Although some participants consistently ranked as more or less concerned for abuse ([a]), the majority varied between scales. Neither monthly census (of physical abuse cases) nor years of experience correlated with ranking. Two participants did not complete the percentage scale.

in the evaluation of child abuse. This variability is consistent with the findings of other investigators. Laskey et al[6] surveyed a large number of pediatricians and pathologists using hypothetical case scenarios of infants with head injury. Even in cases with a strong degree of agreement that included witnessed shaking, some respondents listed cases as "undetermined" or "possible unintentional injury." Broad disagreement accompanied cases without a witnessed abusive incident, a confession, or severe injuries in the absence of an explanation. Even in cases with more agreement, there was substantial disagreement between whether a case was "definite" or "probable." Our protocol expands on this work by showing that similar disagreement occurs with different rating scales applied to the abstracted facts from real cases.

Paradise et al[10] found that physicians who rated themselves as "expert" in the evaluation of child sexual abuse often differed in their assessments of case vignettes with accompanying photographs of potential sexual abuse. Conversely, Roberts and Moran[11] showed high levels of agreement in the assessment of colposcopic photographs, but they compared the consensus assessments of 2 teams with 6 and 10 physicians, respectively. The individual impressions of each physician on each team were not reported.

This variability of abuse assessments has implications for the developing subspecialty of child abuse pediatrics. Because the vast majority of cases lack a gold standard, such as covert surveillance,[12] many studies of findings concerning for abuse use the opinion of the consulting

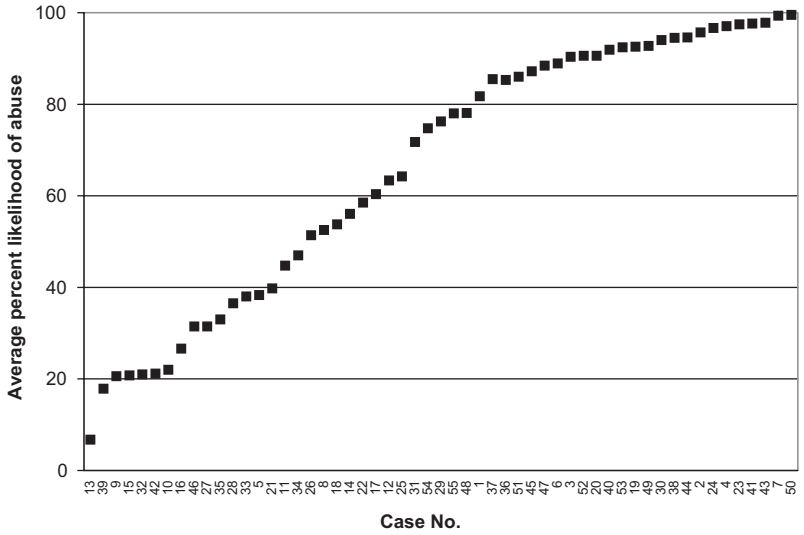

FIGURE 1
Average percentage likelihood of abuse across all participants for each case. Randomly selected cases had a spectrum of average likelihood for abuse.

medical team or of a panel of expert evaluators using a scale of abuse likelihood as the criterion standard for assigning cases to the status of inflicted or accidental injury.[2,3,13,14] Our data support previous suggestions that ratings of a single expert should be used cautiously in research.[6]

We do not believe that this variability is unique to assessments of potential physical abuse. Studies of diagnosis[15,16] and therapy[17–21] from other fields have shown broad variability in many conditions even in the setting of clear guidelines. It should be recognized, however, that variability in child abuse assessments might have greater impact than in other fields. In addition to the potential for serious morbidity and mortality in the case of continuing abuse, the social and legal implications of a diagnosis of child abuse can be profound, particularly because assessments of child abuse potential are received in a number of settings (social, legal) where the variability of medical practice may be less well understood.

Because there was a high degree of between-scale correlation and correlation with the percentage scale, our results suggest that different ratings on any scale do not result from different understandings of the categories used by the scales but rather from real differences of opinion about the determination of abuse likelihood. In contrast to previous work, our results suggest some agreement as to what constitutes a reasonable concern for abuse sufficient to report to children's services.[7] Although some respondents would report for less concern, there were few cases for which respondents would not report to children's services when they estimated more than a 30% risk for inflicted injury, even in the absence of concerns for neglect or social support. Similarly, few participants applied the label "definite abuse" when their estimate of the likelihood of abuse was <95%.

Our study has a number of limitations. As with any study using vignettes, it is unclear whether ratings would have been different if participants had performed their own primary evaluations. By using the abstracted facts of real cases, by including radiographs and images from the actual cases, and by allowing some participants to ask questions, we attempted to make our vignettes as

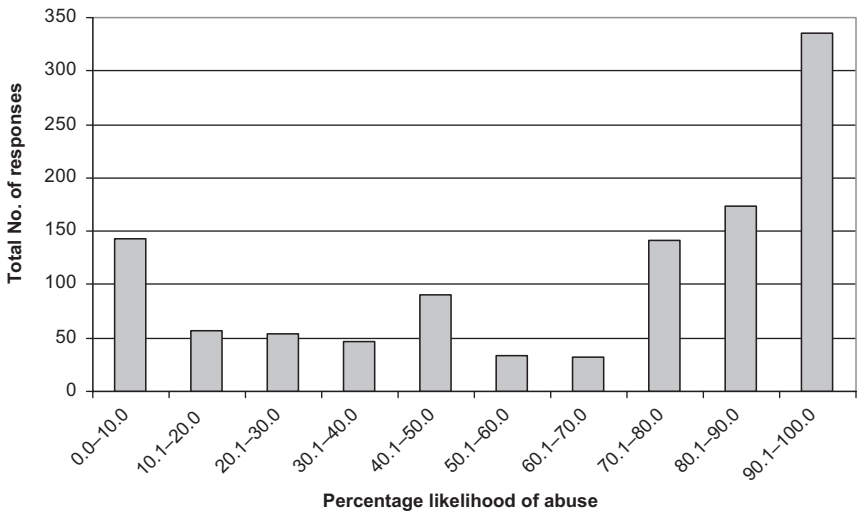

FIGURE 2
Total number of responses for each range of percentage likelihood of abuse. Participant responses clustered at the ends of the spectrum. Cases with intermediate average concern for abuse were likely to include divergent ratings rather than a consensus of intermediate ratings. The total number of responses was only 1100 because 2 participants did not complete the percentage scale.

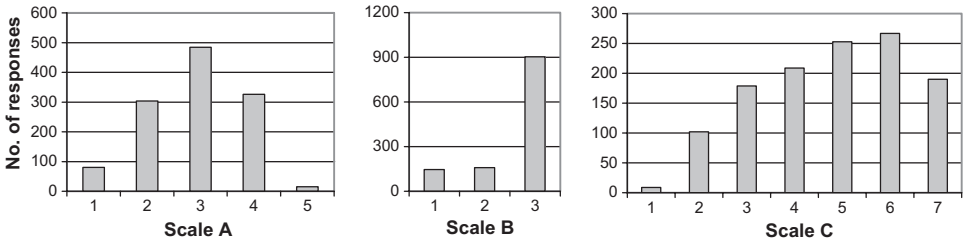

FIGURE 3
Number of responses for each rating-scale category. The majority of cases were at least reasonably concerning for abuse according to most participants.

realistic and complete as possible. Although we excluded demographic data and made no attempt to summarize the affect or reactions of patients or family, it is unclear whether these data would have increased or decreased variability. In addition, our admonition to assume that all testing not mentioned was normal could have increased consistency in cases for which different clinicians would have ordered different diagnostic workups. Our vignettes also included information from social service or legal evaluations only when this was available to the clinicians who had seen the case primarily. Although this information may have increased (or potentially decreased) the consistency of evaluations, such information is often unavailable when clinical, social, or legal conditions require a child abuse consultant to make an initial determination of the likelihood of abuse. Real-life decisions of child abuse likelihood are also the result of a deliberative process far more extensive than was practical for the participants in this study. It is possible that, even without consultation, more time for deliberation would have improved the consistency of child abuse ratings. Because of these limitations, it is difficult to apply these data to a situation in which 2 experts independently evaluate a child in a clinical setting; however, we believe that our vignettes presented similar data to a multidisciplinary case conference or those that might be available to an outside expert asked to comment on the likelihood of abuse for a social services evaluation or a legal proceeding.

Comparison between scales is also limited by the fact that scale B was fundamentally different from the other scales. Although we asked participants to discount concerns of potential neglect, the decision to report is only a part of the rating system inherent to the other scales. Because the minimal requirements for the highest designation was a "reasonable concern for inflicted injury," it is not surprising that there was limited correlation between the highest level of concern on scale B and the other scales that included more gradations of abuse likelihood.

TABLE 6 **Mean (SD) Correlation Coefficients Between Scales Across Individuals**

Parameter	Scale A	Scale B	Scale C
Scale B	0.72 (0.11)	—	—
Scale C	0.90 (0.05)	0.75 (0.09)	—
%	0.86 (0.07)	0.78 (0.10)	0.92 (0.04)

Note lower correlation for scale B, in which the majority of cases were rated at the highest value "reasonable concern for abuse."

Finally, our sample size was only roughly 10% (22 of nearly 200) of eligible participants. The small sample size demonstrates the potential for inclusion bias. Unlike previous studies,[6,10] our decision initially to solicit opinions only from experts with substantial clinical activity in child abuse evaluation limited our number of respondents. In addition, the time required to complete the exercise (nearly 2 hours) combined with the very modest compensation limited the number of participants. Although participants had a spectrum of experience and clinical activity, it is possible that responses could be different between our group and those without the time to participate.

Our data suggest a number of future directions for this work. We identified several cases for which there was broad agreement across participants. Additional evaluation should be undertaken to determine which case elements were associated with most agreement. Conversely, elements associated with most variability might provide direction for a new research agenda.

Many centers of excellence in child abuse evaluation use a team-based approach for evaluation of all child abuse cases. In our exercise, discussion between participants was prohibited. One study that compared consensus assessments of potential sexual abuse demonstrated very high agreement.[11] It is possible that agreement might be higher between independent teams that were able to review these cases and discuss opinions among themselves.

CONCLUSION

When using the abstracted facts of physical abuse consultations, individual assessments of child abuse likelihood by physicians with substantial clinical activity in

TABLE 7 **Average Means and SD for Each Scale Across Participants**

Parameter	Mean (95% CI)
Means	
Scale A	62.50 (55.90–69.10)
Scale B	81.35 (74.52–88.17)
Scale C	63.05 (56.90–69.19)
%	66.22 (58.51–73.93)
SD	
Scale A	16.63 (15.27–17.98)
Scale B	17.51 (12.92–22.10)
Scale C	12.84 (11.79–13.89)
%	15.74 (13.31–18.18)

With the exception of scale B, scales had similar means. Average SDs were similar for all scales except scale C (versus scale A: $P < .0001$, versus scale B: $P = .0255$, versus %: $P = .0019$). CI indicates confidence interval.

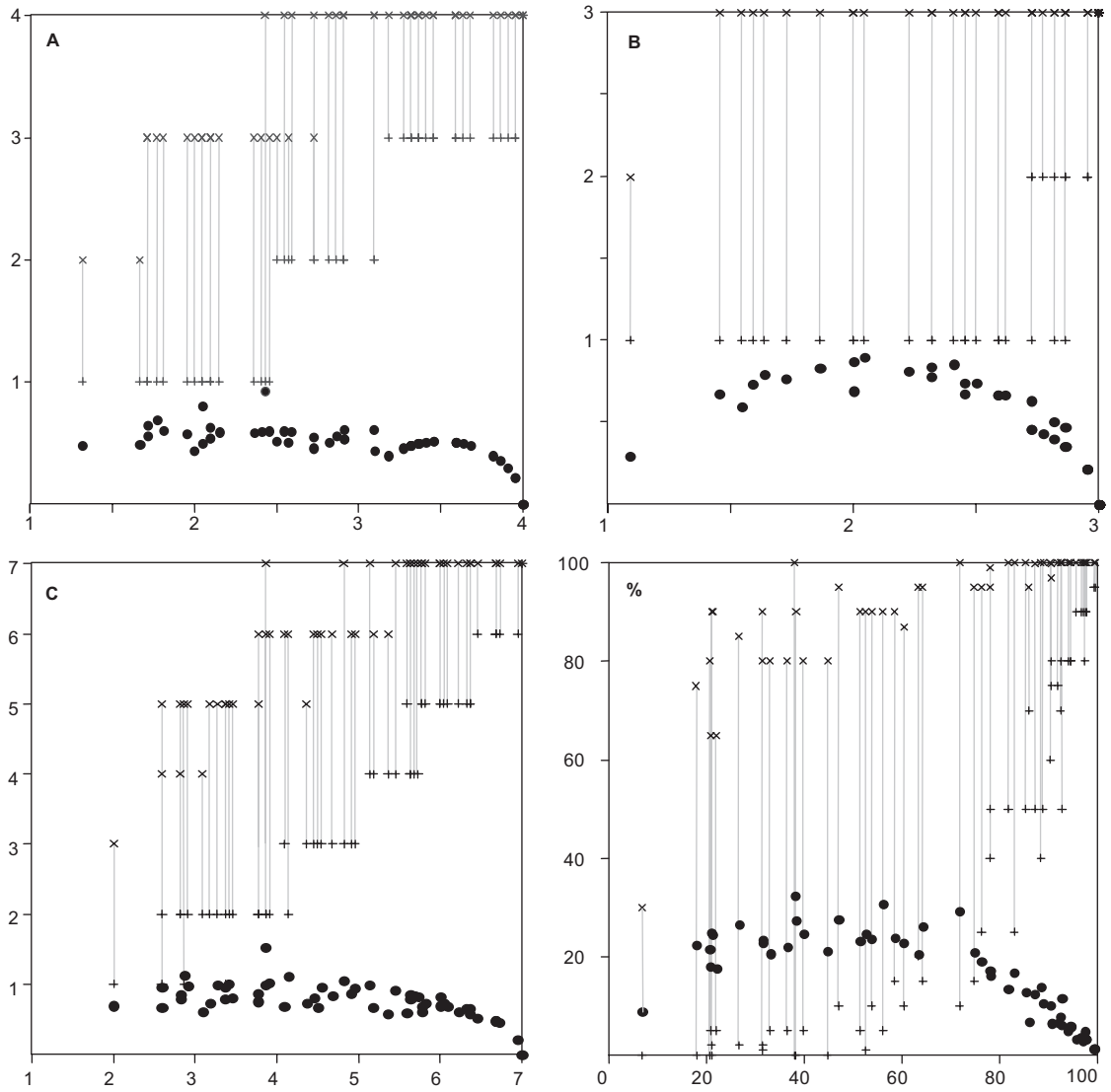

FIGURE 4

Relationship of variability to aggregate concern for abuse for each scale. For each case, the minimum (+), maximum (X), and SD (●) are plotted (y-axis) according to the average rating by all participants (x-axis). Lines connect corresponding minima and maxima. Scales with substantial disagreement will show more spread and higher SDs. Variability for all scales was highest among cases with moderate aggregate concern for abuse and lowest for cases with highest aggregate concern for abuse.

child abuse often show broad variability. Using a scale with example criteria and standard language for communication with social and legal colleagues may moderately improve consistency between clinicians. Although important, the medical determination of abuse is only 1 part of the broader determination of the best plan for any potential victim of abuse. Consultation with other physicians as well as experts from medical, social, criminal, and legal fields is to be recommended.

ACKNOWLEDGMENTS

We thank Casey Williams and the digital classroom team of Cincinnati Children's Hospital Medical Center for coordinating, recording, and disseminating our vignette presentations.

REFERENCES

1. Feldman KW, Bethel R, Shugerman RP, Grossman DC, Grady MS, Ellenbogen RG. The cause of infant and toddler subdural hemorrhage: a prospective study. *Pediatrics.* 2001;108(3):636–646
2. 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;96(4 Pt 1):667–671
3. Leventhal JM, Thomas SA, Rosenfield NS, Markowitz RI. Fractures in young children: distinguishing child abuse from unintentional injuries. *Am J Dis Child.* 1993;147(1):87–92
4. Thomas SA, Rosenfield NS, Leventhal JM, Markowitz RI. Long-bone fractures in young children: distinguishing accidental injuries from child abuse. *Pediatrics.* 1991;88(3):471–476
5. Duhaime AC, Alario AJ, Lewander WJ, et al. Head injury in very young children: mechanisms, injury types, and ophthalmologic findings in 100 hospitalized patients younger than 2 years of age. *Pediatrics.* 1992;90(2 Pt 1):179–185

6. Laskey AL, Sheridan MJ, Hymel KP. Physicians' initial forensic impressions of hypothetical cases of pediatric traumatic brain injury. *Child Abuse Negl.* 2007;31(4):329–342
7. Levi BH, Brown G. Reasonable suspicion: a study of Pennsylvania pediatricians regarding child abuse. *Pediatrics.* 2005;116(1). Available at: www.pediatrics.org/cgi/content/full/116/1/e5
8. Levi BH, Brown G, Erb C. Reasonable suspicion: a pilot study of pediatric residents. *Child Abuse Negl.* 2006;30(4):345–356
9. Billmire M, Myers P. Serious head injury in infants: accident or abuse. *Pediatrics.* 1985;75(2):340–342
10. Paradise JE, Finkel MA, Beiser AS, Berenson AB, Greenberg DB, Winter MR. Assessments of girl's genital findings and the likelihood of sexual abuse: agreement among physicians self-rated as skilled. *Arch Pediatr Adolesc Med.* 1997;151(9):883–891
11. Roberts I, Moran K. Inter-rater reliability in the medical diagnosis of child sexual abuse. *J Paediatr Child Health.* 1995;31(4):290–291
12. Southall DP, Stebbens VA, Rees SV, Lang MH, Warner JO, Shinebourne EA. Apnoeic episodes induced by smothering: two cases identified by covert video surveillance. *Br Med J (Clin Res Ed).* 1987;294(6588):1637–1641
13. Falcone RA Jr, Brown RL, Garcia VF. Disparities in child abuse mortality are not explained by injury severity. *J Pediatr Surg.* 2007;42(6):1031–1037
14. Hymel KP, Makoroff KL, Laskey AL, Conaway MR, Blackman JA. Mechanisms, clinical presentations, injuries, and outcomes from inflicted versus noninflicted head trauma during infancy: results of a prospective, multicentered, comparative study. *Pediatrics.* 2007;119(5):922–929
15. Taggart NW, Haglund CM, Tester DJ, Ackerman MJ. Diagnostic miscues in congenital long-QT syndrome. *Circulation.* 2007;115(20):2613–2620
16. Elesber AA, Decker WW, Smars PA, Hodge DO, Shen WK. Impact of the application of the American College of Emergency Physicians recommendations for the admission of patients with syncope on a retrospectively studied population presenting to the emergency department. *Am Heart J.* 2005;149(5):826–831
17. Patel MR, Chen AY, Roe MT, et al. A comparison of acute coronary syndrome care at academic and nonacademic hospitals. *Am J Med.* 2007;120(1):40–46
18. Kostis WJ, Demissie K, Marcella SW, Shao YH, Wilson AC, Moreyra AE. Weekend versus weekday admission and mortality from myocardial infarction. *N Engl J Med.* 2007;356(11):1099–1109
19. Frampton AE, Eynon CA. High dose methylprednisolone in the immediate management of acute, blunt spinal cord injury: what is the current practice in emergency departments, spinal units, and neurosurgical units in the UK? *Emerg Med J.* 2006;23(7):550–553
20. Belliard G, Catez E, Charron C, et al. Efficacy of therapeutic hypothermia after out-of-hospital cardiac arrest due to ventricular fibrillation. *Resuscitation.* 2007;75(2):252–259
21. Wennberg JE, Fisher ES, Sharp SM, et al. *The Care of Patients With Severe Chronic Illness: The Dartmouth Atlas of Health Care 2006*. Available at: www.dartmouthatlas.org. Accessed September 21, 2007