Abstract

Background: The validated Predicting Abusive Head Trauma (PredAHT) tool estimates the probability of abusive head trauma (AHT) in children <3 years old with intracranial injury. **Objective:** To explore the impact of PredAHT on clinicians' AHT probability estimates and child protection (CP) actions, and assess inter-rater agreement between their estimates and between their CP actions, before and after PredAHT. **Participants and Setting:** Twenty-nine clinicians from different specialties, at teaching and community hospitals. **Methods:** Clinicians estimated the probability of AHT and indicated their CP actions in six clinical vignettes. One vignette described a child with AHT, another described a child with non-AHT, and four represented "gray" cases, where the diagnosis was uncertain. Clinicians calculated the PredAHT score, and reported whether this altered their estimate/actions. The 'think-aloud' method was used to capture the reasoning behind their responses. Analysis included linear modelling, linear mixed-effects modelling, chi-square tests, Fisher's exact tests, intraclass correlation, Gwet's AC_1 coefficient and thematic analysis. **Results:** Overall, PredAHT significantly influenced clinicians' probability estimates in all vignettes ($p<0.001$), although the impact on individual clinicians varied. However, the influence of PredAHT on clinicians' CP actions was limited; after using PredAHT, 9/29 clinicians changed their CP actions in only 11/174 instances. Clinicians' AHT probability estimates and CP actions varied somewhat both before and after PredAHT. Qualitative data suggested that PredAHT may increase clinicians' confidence in their decisions when considered alongside other associated clinical, historical and social factors. **Conclusions:** PredAHT significantly influenced clinicians' AHT probability estimates, but had minimal impact on their CP actions.

Keywords: Abusive head trauma, Child physical abuse, Clinical prediction tool, Child protection

Introduction

It is the responsibility of all clinicians to act upon suspicions of abusive head trauma (AHT), to investigate cases fully, and where necessary to refer cases to children's services. Clinicians from a range of pediatric specialties must piece together all available information and make a decision, based upon the balance of probabilities, about the likelihood of AHT (Colbourne, 2015). Ultimately, distinguishing between AHT and non-abusive head trauma (nAHT) relies on a forensic assessment of the clinical and investigation findings in the context of the history given, and a thorough consideration of the differential diagnoses, and requires a multidisciplinary team approach.

In the United Kingdom (UK), Child Abuse Pediatrics is not a clinical subspecialty as in the United States (US). Most cases of suspected AHT are referred to a community pediatrician for expert child protection (CP) advice. These are doctors who have specialist training in CP and safeguarding. Numerous studies have demonstrated variability in clinicians' confidence and experience in identifying child abuse; their perceived likelihood of abuse; the investigations and evaluation strategies used; and diagnostic decisions made (e.g. Anderst, Nielsen-Parker, Moffatt, Frazier & Kennedy, 2016; Flaherty et al., 2006; Laskey, Sheridan & Hymel, 2007; Lindberg, Lindsell & Shapiro, 2008; Wood et al., 2010).

The Predicting Abusive Head Trauma (PredAHT) clinical prediction tool (CPT) was developed to assist clinicians in deciding which children <3 years old with intracranial injury (ICI), require additional specialist clinical, multidisciplinary and multiagency investigations for possible AHT (Cowley, Morris, Maguire, Farewell & Kemp, 2015; Maguire, Kemp, Lumb & Farewell, 2011). It is intended for use by any clinician involved in the evaluation of children where AHT may be considered within the differential diagnosis, alongside their clinical judgment and in combination with all other information about each case. The derivation study gave predicted probabilities of AHT for 64 possible combinations of the presence or absence of six clinical features, detailed in Table 1 (Maguire et al., 2011). In an external validation study the sensitivity of PredAHT was 72.3% and the specificity was 85.7% using a 50% probability cut-off (Cowley et al., 2015).

More recently we developed a computerized version of PredAHT (Cowley et al., 2018). In summary, we used our derivation dataset (Maguire et al., 2011) to estimate the probability of AHT when one or more features were unknown using multiple imputation by chained equations (van Buuren & Groothuis-Oudshoorn, 2011). We previously used this technique in our validation study as it was found to be the best available approach in comparison to alternative imputation methods (Cowley et al., 2015). We then calculated likelihood ratios for each combination of features, to allow clinicians to incorporate their own prior probability of AHT based on factors unaccounted for by PredAHT e.g. purported history, clinical presentation or psychosocial features. The "baseline" prior probability is 34%, which is simply the prevalence of AHT in the data used to derive the tool. PredAHT thus provides predicted probabilities and likelihood ratios for all 729 potential combinations of the six clinical features, depending on whether each is present, absent *or unknown*. PredAHT could therefore contribute to decision-making at multiple points along the assessment pathway, according to the extent of information available about each of the six features.

There are three main stages to the development of CPTs; derivation, validation, and impact analysis to determine their impact on clinician behavior and patient care (McGinn et al., 2000). In addition, it is recommended that preparatory work is undertaken prior to a formal experimental impact analysis study, to assess the acceptability of the tool and the feasibility of conducting such a study in clinical practice (Wallace et al., 2011). A recent qualitative study concluded that PredAHT was acceptable to a range of CP professionals and could potentially increase professionals' confidence in their decision-making (Cowley et al., 2018). The current study explores the potential impact of PredAHT on clinicians' judgments and decision-making in simulated clinical scenarios.

Experimental vignette methodologies are ideal for analyzing medical decisions that necessitate judgment around sensitive topics and under conditions of uncertainty (Aguinis & Bradley, 2014; Evans et al., 2015). Fictitious yet plausible vignettes, designed through systematic manipulation and control of variables, allow researchers to measure multiple predictors of clinician behavior, maximizing internal and external validity (Aguinis & Bradley, 2014; Evans et al., 2015), and to assess the quality of clinical practice in complex medical situations (Peabody et al., 2004; Rousseau, Rozenberg & Ravaud, 2015). Using six clinical vignettes, this study aimed to explore the impact of PredAHT on clinicians' probability estimates of AHT, and their proposed CP actions, assessing the rationale behind their responses, and the degree of agreement between clinicians' judgments both before, and after, using PredAHT.

Methods

This was a vignette-based cross-sectional survey of clinicians involved in the assessment of young children with ICI, where AHT is amongst the differential diagnosis. The concurrent 'think-aloud' method was used to capture participants' rationale for their responses to the vignette questions. This method instructs participants to articulate their thoughts and feelings as they perform a task, and is based on the assumption that an individual's cognitive processes are directly accessible as verbal data (Ericsson & Simon, 1999). It is often used to study clinicians' diagnostic reasoning and clinical decision-making alongside vignettes (e.g. Skånér, Backlund, Montgomery, Bring & Strender, 2005; Thackray & Roberts, 2017). The study therefore adopted a convergent mixed methods approach, using qualitative methods to gain a comprehensive understanding of the quantitative results

(Creswell, 2013). The study received ethical approval from the Cardiff University School of Medicine Research Ethics Committee (Ref: 15/35).

Participant recruitment

Participants were recruited via email using purposive and snowball sampling. We targeted clinicians across south west United Kingdom (UK) with experience evaluating young children with ICI where AHT is a possible cause. A list of potential participants was identified through personal contacts of the research team who were sent an information sheet to explain the study and asked to suggest clinicians who were eligible to take part. A random selection of 40 clinicians from this list with different levels of CP experience and seniority were then invited to participate (Figure 1). In this study the term “clinician” refers to medical doctors and specialist nurses, who were sampled from three teaching hospitals and two district general (community) hospitals across a range of specialties including pediatrics, radiology and neurosurgery.

Vignette design

Six clinical vignettes were designed according to methodological recommendations and best practices described in the literature and reported in Appendix 1 (Aguinis & Bradley, 2014; Evans et al., 2015). All described children <3 years old with ICI evident on neuroimaging. Table 2 lists the key features of each vignette; Supplementary Table 1 includes the full vignettes as presented to clinicians. Each vignette was comprised of two sections. Section one included the child’s age, gender, any history of trauma or social history, and the characteristics of the ICI. Section two included the clinical information required to complete PredAHT, namely; whether the six clinical features were present, absent or unknown.

Two vignettes were based on real cases. One described a child with confirmed AHT (“V1:AHT”), the other a child with confirmed nAHT (“V2:nAHT”). Demographic details were altered to protect the identity of the children. We hypothesized that PredAHT would have the greatest impact on decision-making in “gray” cases, where there is uncertainty surrounding the diagnosis (Chaiyachati, Asnes, Moles, Schaeffer & Leventhal, 2016). The remaining four vignettes were designed to represent such cases. We created two gray cases (“V3:AHT*” and “V4:nAHT*”) by altering elements of the history and social history from “V1:AHT” and “V2:nAHT” but keeping the clinical features the same. Similar approaches have been taken in previous studies (Anderst et al., 2016; Laskey et al., 2007). In “V3:AHT*” the child is older than in “V1:AHT”, and it is developmentally plausible that a short fall occurred. The incident was unwitnessed, and the clinical features and severity of the injuries appear discordant with the mechanism of injury (Jenny, 2014; Maguire et al., 2013; Sturm, Knecht, Landau & Menke, 2009). In “V4:nAHT*”, there are inconsistencies within the history, a delay in presentation, plus social concerns within the family that may increase suspicion of AHT in comparison to “V2:nAHT”. Two further gray cases (“V5:ICI-only” and “V6:missing”) were developed around one of the most challenging clinical scenarios whereby a baby has ICI with no additional clinical features suggestive of abuse. “V6:missing” is almost identical to “V5:ICI-only”, but neither skeletal radiology nor ophthalmology examination were undertaken. This vignette was created to explore the effects of missing data and the imputation feature of PredAHT.

Data collection

Written informed consent was obtained from all participants. The researcher explained how PredAHT was developed and validated, and described its various features and intended purpose, to each participant. Participants completed the six vignettes in a random sequence, to account for possible order effects. The data collection procedure is outlined in Figure 2 and took approximately 45 minutes. Participants first estimated their own prior probability of AHT for each vignette based on the information given in section 1. They then estimated their Time 1 probability of AHT and Time 1 proposed CP action for each vignette, based on further information given in section 2. The PredAHT score was then calculated for each vignette using the clinicians' prior probabilities, and the clinical details in section 2. Finally, participants estimated their Time 2 probability of AHT and Time 2 proposed CP action for each vignette, after seeing the PredAHT score. CP actions were aligned with three categories of concern (Table 3), as per National Institute for Health and Care Excellence child maltreatment guidelines (National Collaborating Centre for Women's and Children's Health, 2009). Text boxes were included for comments and participants were asked to verbalize their thought processes when deciding on their estimated probabilities of AHT and their proposed CP actions. If participants paused for longer than a few seconds, the researcher reminded them to keep thinking aloud. Otherwise, all interaction between the participants and researcher was minimized so as not to interrupt the participants' flow of thoughts. This enabled participants' verbalizations to be transcribed by the researcher in real time.

Quantitative analysis

Statistical analyses were performed using R software version 3.2.3 (R Core Team, 2015); $p<0.05$ was considered statistically significant. We used linear modelling and linear mixed effects modelling to analyze the impact of PredAHT on clinicians' probability estimates of AHT, and chi-square tests and Fisher's exact tests to analyze the impact of PredAHT on clinicians' proposed CP actions. We used the intraclass correlation coefficient (ICC) to assess inter-rater reliability between clinicians' probability estimates of AHT, and

Gwet’s AC$_1$ coefficient and jackknifing to assess inter-rater reliability between clinicians’ proposed CP actions. Further detail is provided in Appendix 2.

Qualitative analysis

Participants’ verbal data and free text comments were classified into themes using thematic analysis (Braun & Clarke, 2006). Thematic analysis has been used to analyze ‘think-aloud’ data in a number of studies (e.g. Thackray & Roberts, 2017). Analysis entailed grouping codes into categories, and further arranging categories under overarching themes. This involved six phases including 1) familiarization with the data 2) generating initial codes 3) searching for themes 4) reviewing themes 5) defining and naming themes and 6) writing up the results (Braun & Clarke, 2006). To enhance the trustworthiness and rigor of the thematic analysis, a purposeful approach was adopted (Nowell, Norris, White, & Moules, 2017). The first author developed an analytic framework that was amended as new data were collected; all categories and their definitions are detailed in the framework (Appendix 3). Findings were discussed at research team meetings; disagreements regarding data interpretation were resolved by consensus. In the interests of reflexivity, the researcher considered how her own values and assumptions as a student involved in developing PredAHT might influence the interpretation of the findings.

Quantitative Results

Response rates and participant demographics

All vignettes were completed by 29 clinicians in a fully-crossed design between April–September 2016. Twenty-four of the clinicians also took part in a qualitative study on the acceptability of PredAHT (Cowley et al., 2018). Response rates are shown in Figure 1. Participant demographics are reported in Table 4.

Descriptive statistics

There were no missing data and no obvious order effects. Table 5 shows clinicians' mean probability estimates of AHT for each vignette.

Impact of PredAHT on clinicians' probability estimates of AHT

The PredAHT score significantly influenced clinicians' AHT probability estimates in all vignettes (p<.001). Figure 3 shows the estimated linear model slope coefficients $\hat{\beta}$ and 95% confidence intervals for each vignette. Higher slope coefficients $\hat{\beta}$ indicate a greater impact of PredAHT on clinicians' AHT probability estimates. PredAHT had the greatest impact on clinicians' probability estimates of AHT in "V3:AHT*" and the least impact in "V5:ICI-only". Mixed modelling revealed a significant impact of PredAHT on clinicians' probability estimates of AHT overall across vignettes ($\hat{\beta}$ = 0.35, SE = 0.07, p<.001, 95%CI 0.21–0.50). PredAHT appeared most influential for those based at teaching hospitals, for those other than general or community pediatricians, for younger clinicians, for clinicians with the least CP experience and no formal training in pediatric head injuries, and for trainee doctors, however these differences were not statistically significant. Variation in the slope coefficients $\hat{\beta}$ was greater between clinicians than between vignettes (Supplementary Figure 1). This means that the impact of PredAHT was reasonably consistent across vignettes, but varied between individual clinicians.

Impact of PredAHT on clinicians' proposed CP actions

The majority of clinicians would have referred to children's social care at both Time 1 and Time 2 in all cases except "V2:nAHT", where most clinicians elected to investigate further (Figure 4). However, 9/29 (31%) clinicians changed their proposed CP action in 11/174 (6%) instances after using PredAHT (Supplementary Figure 2). Chi-square and

Fisher’s exact tests revealed no significant associations between a change in action and any demographic variables (age, specialty, hospital type, years of CP experience, pediatric head injury training, seniority). In four instances where their probability of AHT increased after using PredAHT, clinicians escalated their proposed CP actions. In four instances where their probability of AHT decreased after using PredAHT, clinicians downgraded their proposed CP actions. Three clinicians changed their proposed CP actions despite not altering their own probability estimate of AHT after using PredAHT.

Inter-rater reliability of clinicians’ probability estimates of AHT

Inter-rater agreement of clinicians’ prior and Time 1 probabilities was “fair” according to published guidelines (Cicchetti, 1994); prior ICC 0.55 (95%CI 0.31−0.88); Time 1 ICC 0.59 (95%CI 0.35−0.90). Agreement at Time 2 increased to “good”; ICC 0.66 (95%CI 0.42−0.92). However, this difference did not reach statistical significance.

Inter-rater reliability of clinicians’ proposed CP actions

Inter-rater agreement of their CP actions was “fair” under Gwet’s model (Gwet, 2012); Time 1 AC_1 coefficient 0.59 (95%CI 0.33–0.85); Time 2 AC_1 coefficient 0.61 (95%CI 0.32–0.90).

Impact of clinicians’ prior probabilities on the PredAHT score

Supplementary Figure 3 compares the PredAHT score, given a baseline prior probability of 0.34, with the scores obtained when clinicians’ prior probabilities were incorporated. Scores incorporating clinicians’ prior probabilities were similar to what would be obtained given the baseline prior for “V1:AHT”, “V2:nAHT” and “V3:AHT*”. However, PredAHT scores with clinicians' priors were higher than PredAHT scores using the baseline prior for "V4:nAHT*", "V5:ICI-only" and "V6:missing".

Qualitative Results

Four overarching themes were identified: clinicians' rationale for their responses, evaluations of PredAHT, interpretations of probabilities, and comments on the vignettes. Data are presented using quotations, selected as examples of the themes that were generated from the data.

Rationale for responses

Clinicians' comments confirmed that they found the "gray" cases difficult to classify: *"I find these really difficult, the 3-month-old rolling off the sofa." Clinician 13, "V5:ICI-only" & "V6:missing".*

For most clinicians, the presence of a concerning social history increased suspicion of AHT in "V4:nAHT*" as compared to "V2:nAHT". This explains why participants' estimated probabilities of AHT were higher on average for "V4:nAHT*" than for "V2:nAHT" even though the clinical features were the same: *"It shows what informs you in these cases, because for me the social services involvement and domestic violence are important." Clinician 20, "V4:nAHT*".* However, some clinicians placed more weight on the lack of additional clinical features concerning for AHT: *"The lack of clinical features is more important to me than the history here." Clinician 25, "V4:nAHT*".*

Almost all of the clinicians were highly suspicious of AHT in "V3:AHT*" due to the concerning clinical features, although the history was potentially less concerning than in "V1:AHT" (no history of trauma). Clinicians stated that the history did not match the severity of the injuries sustained: *"I am not happy with the history as 14-month old children fall a lot and don't get subdural hemorrhages." Clinician 8, "V3:AHT*".*

Clinicians gave reasons as to why they disagreed with the tool and confirmed why PredAHT had the lowest impact on their probability estimates for "V5:ICI-only" and "V6:missing": *"This is where the tool takes away some of the subtlety in the history, this is where I would say I don't care what it says." Clinician 12, "V5:ICI-only".* Although clinicians were informed about the imputation strategy built into PredAHT to account for missing investigations, they were reluctant to change their probability estimates from Time 1 to Time 2 in "V6:missing" because they didn't have the full clinical picture, and stated that PredAHT might act as a prompt for ordering further investigations: *"That's the reason for doing the whole package isn't it because if these things are absent it brings you right down again." Clinician 10, "V6:missing".*

Clinicians' reasons for their estimated probabilities of AHT and proposed CP actions included knowledge of the clinical features indicative of AHT and non-AHT. Clinical knowledge sometimes increased clinicians' suspicions of AHT: *"Retinal hemorrhages would increase my suspicion." Clinician 16, "V1:AHT".* Sometimes it decreased their suspicions: *"I would not be too concerned as the chair is very high, it is a linear undisplaced skull fracture and that type of floor is quite a hard floor." Clinician 16, "V2:nAHT".* Other times it contributed to their uncertainty about a case: *"Left parietal skull fracture, the most common skull fracture in both abused and non-abused children." Clinician 21, "V2:nAHT".* Lack of clinical knowledge also contributed to uncertainty in estimating the probability of AHT: *"See this is going into detail about the eye findings some of which I don't know the significance of." Clinician 15, "V1:AHT".*

Clinicians considered the age and developmental stage of the child when estimating the probability of AHT: *“A three-month old can’t roll so the history is immediately suspicious.” Clinician 9, “V5:ICI-only” & “V6:missing”.*

An important factor influencing some clinician’s probability estimates of AHT was a consistent history. However this was less important when there were other concerning features present: *“Even though the story is consistent the history is still dodgy and the neuroimaging features are suspicious.” Clinician 25, “V3:AHT*”.*

When completing the vignettes, clinicians deliberated over the purported mechanism of injury and whether this was consistent with the features observed: *“I would be worried that there’s no bruising because that means there’s no impact.” Clinician 21, “V3:AHT*”.*

Evaluations of PredAHT

Participants talked about the potential benefits of PredAHT while completing the vignettes. Overall, 27/29 clinicians would find PredAHT useful in their practice: *“This would undoubtedly be extremely useful.” Clinician 6.* However two clinicians were unsure: *“I think this would be more useful for older children but I’m not sure it actually adds much.” Clinician 15.* Clinicians would find PredAHT useful as they do not usually think in terms of probability when assessing risk: *“I never give percentages, even in court I would say that we don’t talk in those terms, and that’s why I think the tool is going to be helpful.” Clinician 5.* Many clinicians felt reassured by PredAHT, and reported that it gave them more confidence in their decisions, even if they did not change their CP actions based on the score (Table 6).

Participants also discussed the potential risks of PredAHT. Some thought that variables relating to the history should be included in the tool: *“There’s no factor for the lack of history is there which is key isn’t it?” Clinician 3.* Others felt that the tool cannot account for the subtleties that are often present in individual cases, or that since it cannot account for all possible indicators for abuse, a low score may provide false reassurance: *"The cheek bruising is really worrying, it shows that PredAHT can't take into account nuances with just yes and no answers." Clinician 9, "V2:nAHT"*. Some participants also discussed at length the need to understand how PredAHT works, and the importance of critically appraising the quality of the data that it is based on: *"We would need to know where the figures in the tool came from, and to make sure they are correct." Clinician 22*.

Interpretation of probabilities

Participants talked about their probability thresholds for investigation and referral as justification for their proposed CP actions. One clinician would refer all cases she considered to have a 50% risk or greater of AHT to social services, but would investigate cases she thought carried a lower probability of AHT: *"All that matters for referral is whether it's over 50% or not." Clinician 3.*

Many clinicians were interested in exploring the estimated post-test probabilities that PredAHT provided based on different prior probabilities. Some were shocked by the impact the prior probability had on the PredAHT score: *"I'm shocked by how much my prior probabilities have affected the scores. This makes me think I might be too hawkish about abuse." Clinician 26.* However other clinicians justified their high estimated prior probabilities due to the neuroimaging features in the vignettes: *"I can only say a 90% prior probability for all of these vignettes because if there is a subdural hemorrhage, to me that's a really high probability." Clinician 5*. Some questioned how they might estimate their prior probability in practice and mentioned that in reality some of the clinical features included in PredAHT may be incorporated in their prior probability estimates: *"That's interesting then to see how my gut feeling is coming in. Really I'm estimating the prior probability without* *knowing all the information. What are we taking into account with our prior probability in practice and what is our evidence for that?" Clinician 10.*

Comments on vignettes

Comments on the details of the vignettes themselves revealed important information about clinicians' behavior when assessing suspected AHT. Some questioned why certain investigations were or were not performed e.g. why a skeletal survey was not performed in "V1:AHT" and "V3:AHT*: *"You would still need to do a skeletal survey even if the probability is already high." Clinician 14, "V1:AHT".* Other asked why a skeletal survey and ophthalmology exam *were* ordered in "V2:nAHT": *"I'm not sure I would have done any of these tests in this case!" Clinician 15, "V2:nAHT".*

Some clinicians talked about additional investigations they would perform: *"I don't know why you keep missing the bloods out!" Clinician 16.* Similarly, many participants reported needing more detail on the history in order to make more informed probability estimates or CP decisions: *"The problem is you would want so much more information. I would assess if they could roll in the department." Clinician 19, "V6:missing".* In addition some participants wanted more detail on the clinical features in order to assess whether the mechanism was plausible: *"What side is the cheek bruising and is the bruising to the scalp the same side as the head injury?" Clinician 19, "V2:nAHT"*. The majority of clinicians were concerned about the cheek bruising in "V2:nAHT" and "V4:nAHT*", and wanted more information about the pattern and mechanism of the bruising. This explains why some participants' estimated probabilities of AHT were high for "V2:nAHT", despite the fact that this vignette represented a confirmed case of non-AHT: *"It would depend on the pattern of bruising to the cheeks." Clinician 4, "V2:nAHT"*

Finally, while considering the probability of AHT, clinicians discussed a variety of possible differential diagnoses, not detailed in the vignettes, that they would rule out in practice: *"He wouldn't have hit his head that badly just falling on a floor, unless he has got some bleeding disorder or something." Clinician 10, "V3:AHT*"*

Discussion

In this vignette study, statistical modelling demonstrated that PredAHT significantly influenced clinicians' AHT probability estimates in all vignettes. Interestingly however, clinicians' proposed CP actions were only influenced by PredAHT in a minority of cases, and PredAHT did not significantly improve the overall agreement between clinicians' AHT probability estimates or their proposed CP actions. Despite this, the 'think-aloud' data showed that 27/29 clinicians would find PredAHT useful in their practice, and that it provided them with greater confidence in their decisions in the vignette cases, confirming the findings of the recent qualitative study on the acceptability of PredAHT (Cowley et al., 2018). However, it was evident that clinicians were influenced by a variety of social, historical and clinical factors in each case, emphasizing the need to consider the PredAHT probabilities in the context of these associated factors.

PredAHT had the greatest impact in "V3:AHT*" and "V1:AHT". This suggests that PredAHT may act to increase clinicians' suspicions when there are several clinical features indicative of AHT and that it may help clinicians to remain objective during their assessment of a young child with ICI. PredAHT had the least impact in "V5:ICI-only", where the history and presentation was concerning, but due to the absence of any additional clinical features, the PredAHT score was low (3.7% at baseline). Reassuringly, this suggests that clinicians were not simply following PredAHT, but were considering factors that it cannot account for. Similarly, a number of clinicians reported disregarding the low PredAHT score (14.2% at baseline) for “V2:nAHT”, due to concerns about the cheek bruising, which is a recognized indicator of physical abuse (Kemp, Maguire, Nuttall, Collins & Dunstan, 2014). Even those who felt reassured by the score would have requested further information about this feature.

Despite being aware of the imputation strategy built into PredAHT to account for missing data, the tool had minimal impact on clinicians’ probability estimates in “V6:missing” . This highlights the importance of obtaining an ophthalmology exam and skeletal survey whenever AHT is suspected, in line with international recommendations (The Royal College of Ophthalmologists & the Royal College of Pediatrics & Child Health, 2013; The Royal College of Radiologists & the Royal College of Pediatrics & Child Health, 2008). Qualitative analysis suggested that PredAHT may help to standardize investigations in suspected AHT by highlighting the clinical significance of fractures and retinal hemorrhages, and the influence these features, if known, would have on the PredAHT score.

Although PredAHT significantly influenced clinicians’ probability estimates of AHT, there were only 11/174 instances where clinicians changed their proposed CP action after seeing the score. For analysis purposes, we collapsed the categories of CP action, however some clinicians who elected to investigate further would have conducted additional investigations after seeing the PredAHT score. With the exception of “V2:nAHT”, clinicians mean probability estimates of AHT exceeded 50% at all time points, and many clinicians did not change their actions as they had already elected to investigate/refer to children’s services at Time 1.

It was evident that probabilities are interpreted differently by different people, and clinicians have different thresholds on which they act. There is little professional agreement as to what equates to a “reasonable suspicion” of abuse, varying in one study from a probability of 10%–35%, 40%–50% or 60%–70% and for a smaller group to >75% (Levi & Brown, 2005). In another study, 51% of participants defined the term “reasonable medical certainty" in the context of child abuse as ≥90% probability, 30% defined it as ≤50% probability and 2% used a definition of ≤25% probability (Dias, Boehmer, Johnston-Walsh & Levi, 2015). Furthermore, Flaherty et al. (2008) found that clinicians only reported 73% of the children that they thought were likely or very likely abused to children's services, and only 24% of children that they thought were possibly abused. Other studies have found that improving clinicians' judgments of disease probability does not necessarily change or improve their treatment decisions and may have an unpredictable effect on clinicians' behavior; one possibility is that clinicians' CP actions in this study were not based on probabilistic thresholds (Poses, Cebul & Wigton, 1995). This is consistent with the observation that some clinicians changed their CP actions after seeing the PredAHT score, but not their probability estimate of AHT. Alternatively, this finding could suggest that PredAHT may help to reduce the uncertainty around clinicians' point estimates of the probability of AHT, and give them more confidence in their decisions; this was confirmed by the qualitative data, where many clinicians stated that they felt reassured by the tool even if they did not change their proposed CP action.

This study found that clinicians' AHT probability estimates for each vignette varied somewhat. This finding is consistent with other vignette studies evaluating the likelihood of abuse amongst clinicians. One such study asked US pediatricians to rate 16 cases of pediatric traumatic brain injury on a seven point scale ranging from definitive unintentional injury to definitive inflicted injury, and found they were unable to agree on the cause of the injuries in half of the scenarios (Laskey et al., 2007). Lindberg et al. (2008) found extensive variability between experienced CP pediatricians when estimating the likelihood of abuse in video vignettes of cases referred to a hospital child abuse team, using three rating scales and a percentage probability.

The PredAHT score with clinicians' priors was higher than the baseline score for "V4:nAHT*", "V5:ICI-only" and "V6:missing". Allowing clinicians to incorporate their prior probabilities of AHT enables them to take into account factors that PredAHT does not. Although higher prior probabilities may lead to higher PredAHT scores in some cases, this should prompt further investigation and may help to circumvent the possibility of false reassurance provided by a low score. However, it is important that clinicians' prior probabilities are evidence-based, to minimize the possibility of false accusations of abuse. Some clinicians were alarmed by the impact their prior probability of AHT had on the PredAHT score and questioned how they would estimate a prior probability in practice. Taken together, these results reinforce findings from a qualitative study on the acceptability of PredAHT (Cowley et al., 2018), that any training on PredAHT would need to incorporate guidance on estimating a prior probability of AHT.

The actual impact of PredAHT on clinicians' probability estimates of AHT and subsequent CP actions is likely to differ in clinical practice (Reilly & Evans, 2006). It is not yet known whether clinicians will use PredAHT, whether they will use it accurately, or what actions they may take in practice based on specific probability scores. Importantly, PredAHT was designed as an assistive tool, and the qualitative analysis highlighted that whilst clinicians felt that it would be useful in practice to support their decision-making, they also confirmed the importance and value of further essential information. PredAHT is *not* a diagnostic tool, and unlike a directive tool, PredAHT does not recommend a direct course of action based on the results. A directive, validated, highly sensitive 4-variable screening tool for AHT has been developed for use in the pediatric intensive care unit (PICU), to minimize missed cases of AHT and exclude AHT when negative (Hymel et al., 2014). When one or more of four clinical or neuroradiological variables are present in an acutely head-injured infant or young child, a thorough abuse evaluation is recommended. A recent potential impact study of this tool suggested that it may improve the identification of AHT in the intensive care setting (Hymel et al., 2015). However, in order to determine whether PredAHT or the 4-variable PICU tool can change clinician behavior for the better, and to determine their impact on relevant outcomes, formal impact analysis studies are required for both.

Strengths and Limitations

A strength of this study is the use of mixed methods; asking clinicians to articulate the reasoning behind their responses to the vignettes allowed for a meaningful interpretation of the quantitative data. Another strength is that the experimental control afforded by vignette studies permits assessment of the vignette factors' causal effect on the dependent variable. This enhances internal validity compared to traditional surveys (Aguinis & Bradley, 2014; Evans et al., 2015; Steiner, Atzmüller, & Su, 2016). The 'think-aloud' data provided additional evidence of internal validity, because clinicians confirmed that their probability estimates differed as a result of the factors manipulated in the vignettes. Since vignettes differ from real life situations, vignette studies are often criticised due to potential limitations in external validity (Steiner et al., 2016). Clinicians' responses may have been decontextualized from the types of responses they may have made in highly pressured or difficult real life situations, where decision-making does not just depend on a rational analysis of the features of a case. We used only six vignettes, yet there are many more scenarios in which PredAHT could be applied. Most participants were consultants, and half were community pediatricians with considerable CP experience; results may have been different amongst trainee doctors or other specialties involved in the assessment of suspected AHT. To maximize external validity, participants were randomly sampled from a larger pool of potential participants, which extends external validity at least to the target population of clinicians involved in suspected AHT cases (Steiner et al., 2016). The selectivity of vignettes and how this is interpreted by participants can generate valuable data in itself; in this study, clinicians' comments about clinical investigations, elements of the history, or differential diagnoses not detailed in the vignettes revealed insights about the factors influencing their judgments and decision-making in suspected AHT cases.

One possible limitation is that the order of the information presented in the vignettes may not have reflected clinical reality. For example, in practice it is likely that clinicians would have the information regarding apnea, seizures, and head/neck bruising prior to a child undergoing neuroimaging to look for possible ICI, and they may not gather information regarding the social history until later on in the assessment process. The information was presented as such because clinicians' estimated prior probability of AHT should not be based on the clinical features included in PredAHT but on the other features of a case that PredAHT cannot account for. The qualitative results revealed that in reality, it may be difficult for clinicians to estimate a prior probability of AHT excluding the clinical features in PredAHT once the presence or absence of these are already known.

While the impact of PredAHT differed by clinician demographic variables, these findings were not statistically significant; a larger study would be required to further examine these observed trends. Measures were taken to reduce potential subjectivity of qualitative data analysis and bias by involving the research team in data analysis and encouraging researcher reflexivity. Although participants were randomly sampled from a larger list of possible participants, such a sample is not as representative of the population as a probability random sample (Palinkas et al., 2015). Therefore there may be some degree of underestimation of the population variance and overstatement of statistical significance. We do not interpret our p values literally but treat them as a guide for further exploration.

Conclusion

This study has demonstrated that PredAHT had a significant impact on clinicians' AHT probability estimates, showing that clinicians are willing to alter their own probability estimate of AHT when exposed to a validated CPT. However, clinicians' proposed CP actions were only influenced by the tool in a minority of cases. Additional research is required to assess the actual impact of PredAHT in clinical practice.

References

Aguinis, H., & Bradley, K.J. (2014). Best practice recommendations for designing and implementing experimental vignette methodology studies. *Organizational Research Methods, 17*(4), 351–371.

Anderst, J., Nielsen-Parker, M., Moffatt, M., Frazier, T., & Kennedy, C. (2016). Using simulation to identify sources of medical diagnostic error in child physical abuse. *Child Abuse & Neglect, 52*, 62–69.

Braun, V, & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101.

Chaiyachati, B.H., Asnes, A.G., Moles, R.L., Schaeffer, P., & Leventhal, J.M. (2016). Gray cases of child abuse: Investigating factors associated with uncertainty. *Child Abuse & Neglect, 51,* 87–92.

Cicchetti, D.V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*(4), 284–290.

Colbourne, M. (2015). Abusive head trauma: Evolution of a diagnosis. *BCMJ, 57,* 331–335.

Cowley, L.E., Farewell, D.M., & Kemp, A.M. (2018). Dataset on clinicians' probability estimates of abusive head trauma and proposed child protection actions in six clinical vignettes, before and after exposure to the validated Predicting Abusive Head Trauma (PredAHT) clinical prediction tool. *Data In Brief.*

Cowley, L.E., Maguire, S., Farewell, D.M., Quinn-Scoggins, H.D., Flynn, M.O., & Kemp, A.M. (2018). Acceptability of the Predicting Abusive Head Trauma (PredAHT) clinical prediction tool: A qualitative study with child protection professionals. *Child Abuse & Neglect, 81*, 192–205.

Cowley, L.E., Morris, C.B., Maguire, S.A., Farewell, D.M., & Kemp, A.M. (2015). Validation of a prediction tool for abusive head trauma. *Pediatrics*, *136*(2), 290–298.

Creswell, J.W. (2013). *Research design: Qualitative, quantitative, and mixed methods approaches.* Thousand Oaks, CA: Sage Publications, Inc.

Dias, M.S., Boehmer, S., Johnston-Walsh, L., & Levi, B.H. (2015). Defining 'reasonable medical certainty' in court: What does it mean to medical experts in child abuse cases*? Child Abuse & Neglect, 50*, 218–227.

Ericsson, K.A., & Simon, H.A. (1999). *Protocol analysis: Verbal reports as data.* Cambridge, MA: MIT Press.

Evans, S.C., Roberts, M.C., Keeley, J.W., Blossom, J.B., Amaro, C.M., Garcia, A.M,…Reed, G.M. (2015). Vignette methodologies for studying clinicians' decision-making: Validity, utility, and application in ICD-11 field studies. *International Journal of Clinical & Health Psychology, 15*(2), 160–70.

Flaherty, E.G., Sege, R.D., Griffith, J., Price, L.L., Wasserman, R., Slora, E.,…Binns, H.J. (2008). From suspicion of physical child abuse to reporting: Primary care clinician decision-making. *Pediatrics, 122*(3), 611–619.

Flaherty, E.G., Sege, R., Price, L.L., Christoffel, K.K., Norton, D.P., & O'Connor, K.G. (2006). Pediatrician characteristics associated with child abuse identification and reporting: Results from a national survey of pediatricians. *Child Maltreatment, 11*(4), 361–369.

Gwet, K.L. (2012). Benchmarking the agreement coefficient. *In:* Gwet, K.L. (Ed.) *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among multiple raters* (121–147). Gaithersburg, MD: Advanced Analytics, LLC.

Hymel, K.P., Armijo-Garcia, V., Foster, R., Frazier, T.N., Stoiko, M., Christie, L.M.,…Wang, M. (2014). Validation of a clinical prediction rule for pediatric abusive head trauma. *Pediatrics*, *134*(6), e1537–1544.

Hymel, K.P., Herman, B.E., Narang, SK., Graf, J.M., Frazier, T.N., Stoiko, M.,…Wang, M. (2015). Potential impact of a validated screening tool for pediatric abusive head trauma. *Journal of Pediatrics, 167*(6), 1375–1381.e1.

Jenny, C. (2014). Alternate theories of causation in abusive head trauma: What the science tells us. *Pediatric Radiology, 44*(Suppl 4), S543–S547.

Kemp, A.M., Maguire, S.A., Nuttall, D., Collins, P., & Dunstan, F. (2014). Bruising in children who are assessed for suspected physical abuse. *Archives of Disease in Childhood*, *99*(2), 108–113.

Laskey, A.L., Sheridan, M.J., & Hymel, K.P. (2007). Physicians' initial forensic impressions of hypothetical cases of pediatric traumatic brain injury. *Child Abuse & Neglect, 31*(4), 329–342.

Levi, B.H., & Brown, G. (2005). Reasonable suspicion: A study of Pennsylvania pediatricians regarding child abuse. *Pediatrics, 116*(1), e5–e12.

Lindberg, D.M., Lindsell, C.J., & Shapiro, R.A. (2008). Variability in expert assessment of child physical abuse likelihood. *Pediatrics, 121*(4), e945–e953.

Maguire, S.A., Kemp, A.M., Lumb, R.C., & Farewell, D.M. (2011). Estimating the probability of abusive head trauma: A pooled analysis. *Pediatrics, 128*(3), e550–e564.

Maguire, S.A., Lumb, R.C., Kemp, A.M., Moynihan, S., Bunting, H.J., Watts, P.O., & Adams, G.G. (2013). A systematic review of the differential diagnosis of retinal haemorrhages in children with clinical features associated with child abuse. *Child Abuse Review*, *22*(1), 29–43.

McGinn, T.G., Guyatt, G.H., Wyer, P.C., Naylor, C.D., Stiell, I.G., & Richardson, W.S. (2000). Users' guides to the medical literature: XXII: How to use articles about clinical decision rules. Evidence-Based Medicine Working Group. *JAMA*, *284*(1), 79–84.

National Collaborating Centre for Women's and Children's Health (UK). *When to suspect child maltreatment*. NICE Clinical Guidelines, No. 89. London, RCOG Press; July 2009.

Nowell, L.S., Norris, J.M., White, D.E., & Moules, N.J. (2017). Thematic analysis: Striving to meet the trustworthiness criteria. *International Journal of Qualitative Methods, 16,* 1–13.

Palinkas, L.A., Horwitz, S.M., Green, C.A., Wisdom, J.P., Duan, N., & Hoagwood, K. (2015). Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. *Administration & Policy in Mental Health, 42*(5), 533–544.

Peabody, J.W., Luck, J., Glassman, P., Jain, S., Hansen, J., Spell, M., & Lee, M. (2004). Measuring the quality of physician practice by using clinical vignettes: A prospective validation study. *Annals of Internal Medicine, 141*(10), 771–780.

Poses, R.M., Cebul, R.D., & Wigton, R.S. (1995). You can lead a horse to water–improving physicians' knowledge of probabilities may not affect their decisions. *Medical Decision Making, 15*(1), 65–75.

R Core Team. (2015). R: A language and environment for statistical computing (Version 3.2.3). R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/

Reilly, B.M., & Evans, A.T. (2006). Translating clinical research into clinical practice: Impact of using prediction rules to make decisions. *Annals of Internal Medicine, 144*(3), 201–209.

Rousseau, A., Rozenberg, P., & Ravaud, P. (2015). Assessing complex emergency management with clinical case-vignettes: A validation study. *PLOS ONE, 10*(9), e0138663.

Skånér, Y., Backlund, L., Montgomery, H., Bring, J., & Strender, L.E. (2005). General practitioners' reasoning when considering the diagnosis heart failure: A think-aloud study. *BMC Family Practice, 6,* 4.

Steiner, P.M., Atzmüller, C., & Su, D. (2016). Designing valid and reliable vignette experiments for survey research: A case study on the fair gender income gap. *Journal of Methods & Measurement in the Social Sciences,* 7(2): 52–94.

Sturm, V., Knecht, P.B., Landau, K., & Menke, M.N. (2009). Rare retinal haemorrhages in translational accidental head trauma in children. *Eye, 23*(7), 1535–1541.

Thackray, D. & Roberts, L. (2017). Exploring the clinical decision-making used by experienced cardiorespiratory physiotherapists: A mixed method qualitative design of simulation, video recording and think aloud techniques. *Nurse Education Today, 49,* 96–105.

The Royal College of Radiologists & the Royal College of Paediatrics & Child Health. *Standards for radiological investigations of suspected non-accidental injury*. London, RCR & RCPCH; March 2008.

The Royal College of Ophthalmologists & the Royal College of Paediatrics & Child Health. *Abusive head trauma and the eye in infancy*. London, RCO & RCPCH; 2013.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). MICE: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45*(3), 1–67.

Wallace, E., Smith, S.M., Perera-Salazar, R., Vaucher, P., McCowan, C., Collins, G.,…Fahey, T. (2011). Framework for the impact analysis and implementation of Clinical Prediction Rules (CPRs). *BMC Medical Informatics & Decision Making, 11*, 62.

Wood, J.N., Hall, M., Schilling, S., Keren, R., Mitra, N., & Rubin, D.M. (2010). Disparities in the evaluation and diagnosis of abuse among infants with traumatic brain injury. *Pediatrics*, *126*(3): 408–414.

Table 1. The six features included in the Predicting Abusive Head Trauma clinical prediction tool

Feature	Description
Head or neck bruising	Any documented bruising to head or neck
Seizures	Any documented seizures from a single seizure to status epilepticus
Apnea	Any apnea documented in the initial history or during inpatient stay
Rib fracture	Any rib fracture documented after appropriate radiologic imaging
Long-bone fracture	Any long-bone fracture documented after appropriate radiologic imaging
Retinal hemorrhage	Any retinal hemorrhage documented after indirect ophthalmologic examination by a pediatric ophthalmologist

Published previously in Pediatrics (Cowley et al., 2015) and reproduced with permission.

Table 2. Key features of each of the six clinical vignettes

	Information given in Section 1		Information given in Section 2						PredAHT Results	
Vignette	**Presentation, History and Social History**	**CT Scan Results**	**B**	**A**	**S**	**RF**	**LBF**	**RH**	**PredAHT Probability[a]**	**PredAHT Likelihood ratio**
1:AHT	3 months old Lethargy, vomiting No history of trauma	HII affecting both cerebral hemispheres, brainstem and thalami Hyperdense SDH at the vertex	No	Yes	Yes	?	?	Yes	98.4%	118.79
2:nAHT	23 months old No delay in presentation Fall from a chair at a height of 1.5 metres onto a tiled floor Consistent history between parents and over time	Frontal lobe hyperdense SDH Linear, undisplaced skull fracture of left frontal parietal bone	Yes	No	No	No	No	No	14.2%	0.32

3:AHT*	14 months old Lethargy, vomiting No delay in presentation Unwitnessed short fall onto wooden floor Consistent history over time	HII affecting both cerebral hemispheres, brainstem and thalami Hyperdense SDH at the vertex	No	Yes	Yes	?	?	Yes	98.4%	118.79
4:nAHT*	23 months old Six hour delay in presentation to the hospital Initially no history of trauma Possible fall from a chair at a height of 1.5 metres onto a tiled floor Domestic violence concerns Previous children's services involvement	Frontal lobe hyperdense SDH Linear, undisplaced skull fracture of left frontal parietal bone	Yes	No	No	No	No	No	14.2%	0.32
5:ICI-only	3 months old Lethargy, vomiting Rolled off the sofa onto the floor	Multiple small bilateral SDHs	No	No	No	No	No	No	3.7%	0.08
6:missing	3 months old Lethargy, vomiting Rolled off the sofa onto the floor	Multiple small bilateral SDHs	No	No	No	?	?	?	10.4%	0.22

B, head/neck bruising; A, apnea; S, seizures; RF, rib fractures; LBF, long-bone fractures; RH, retinal hemorrhages, PredAHT, Predicting Abusive Head Trauma tool; HII, hypoxic ischemic injury; SDH, subdural hemorrhage

[a] This was calculated using the "baseline" prior probability of 34%, the prevalence of abusive head trauma in the data used to derive the tool

Table 3. Possible child protection actions and associated categories of concern in line with National Institute for Health & Care Excellence (NICE) child maltreatment guidelines

Indicated child protection action		Category
No further child protection action		No concern (abuse excluded)
Investigate further:	Discuss with line manager Discuss with child protection colleague Gain collateral information from other agencies and health disciplines (e.g. health visitor) Order further investigations (please specify)	Concern (abuse considered)
Refer to children's services		Suspicion (abuse suspected)

Table 4. Demographics and characteristics of clinicians participating in the vignette study

Demographics / Characteristics	Community Paediatricians (N = 15)		General Paediatricians (N = 9)		Other Specialty (N = 5)	
	n	%	n	%	n	%
Gender						
Female	15	100	2	22.2	4	80
Male	0	0	7	77.8	1	20
Age group						
25–34	0	0	1	11.1	1	20
35–44	5	33.3	4	44.4	3	60
45–54	6	40	3	33.3	0	0
55–64	4	26.7	1	11.1	1	20
Ethnicity						
White British	12	80	6	66.7	4	80
White Other	2	13.3	1	11.1	1	20
Indian	1	6.7	2	22.2	0	0
Years in CP						
5–9	3	20	2	22.2	2	40
10–20	4	26.7	3	33.3	1	20
>20	8	53.3	4	44.4	2	40
CP training						
Yes	15	100	9	100	5	100
No	0	0	0	0	0	0
Paediatric HI training						
Yes	11	73.3	4	44.4	5	100
No	4	26.7	5	55.6	0	0
Hospital Type						
Teaching	11	73.3	5	55.6	5	100
District general	4	26.7	4	44.4	0	0
Seniority						

Consultant	8	53.3	9	100	3	60
Associate specialist	5	33.3	0	0	0	0
Trainee doctor	2	13.3	0	0	1	20
Senior staff nurse	0	0	0	0	1	20

CP = child protection, HI = head injuries

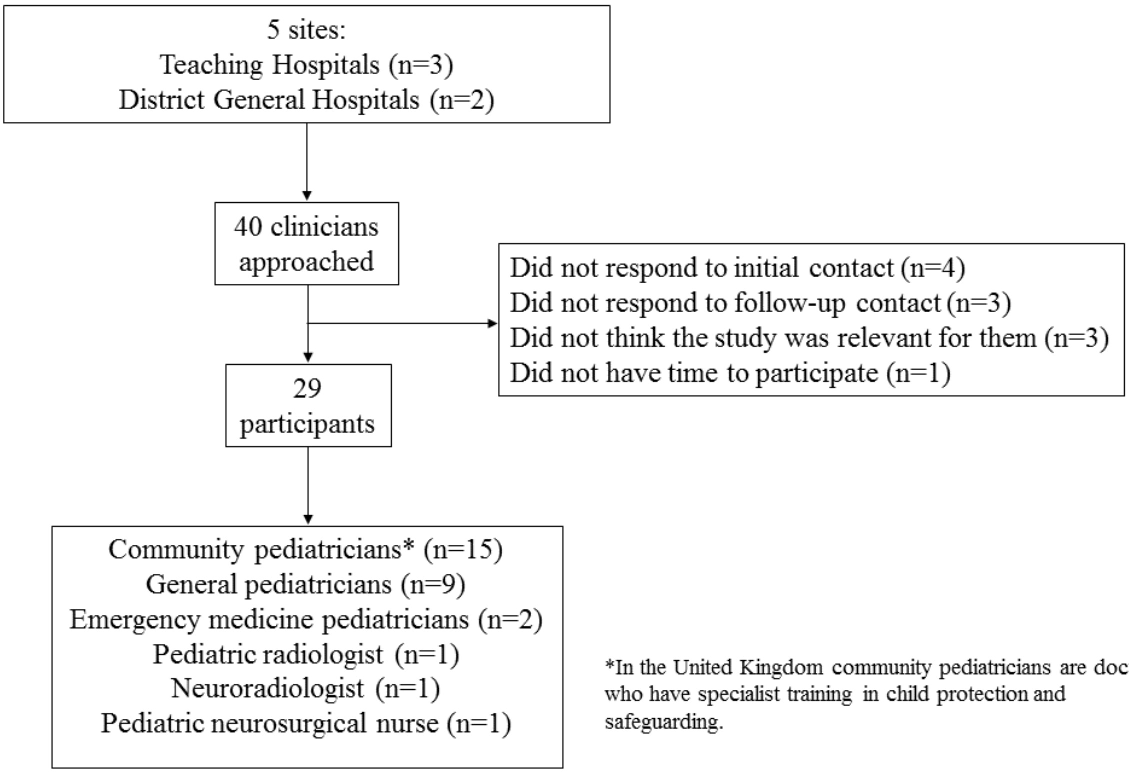
Table 5. Means, standard deviations, and minimum and maximum values of clinicians' probability estimates of AHT for each of the six vignettes

	Summary statistic	V1: AHT	V2: nAHT	V3: AHT*	V4: nAHT*	V5: ICI-only	V6: missing
Prior probability	Mean	80.28	32.45	72.34	64.34	78.28	77.93
	SD	(14.54)	(20.00)	(17.16)	(16.94)	(14.90)	(13.20)
	Min–Max	40–98	5–80	30–90	30–95	40–100	50–100
Time 1 probability	Mean	91.31	33.97	89.38	61.41	61.28	78.34
	SD	(9.38)	(21.97)	(12.02)	(19.82)	(24.61)	(13.25)
	Min–Max	60–100	5–90	50–100	30–99	10–100	50–95
Time 2 probability	Mean	95.06	26.72	95.61	54.36	54.55	72.00
	SD	(6.71)	(21.43)	(5.92)	(20.92)	(27.30)	(20.95)
	Min–Max	75–100	0–90	75–100	20–99	10–100	18–100

Table 6. Participants' reported that PredAHT increased their confidence in their decision-making in the vignette cases

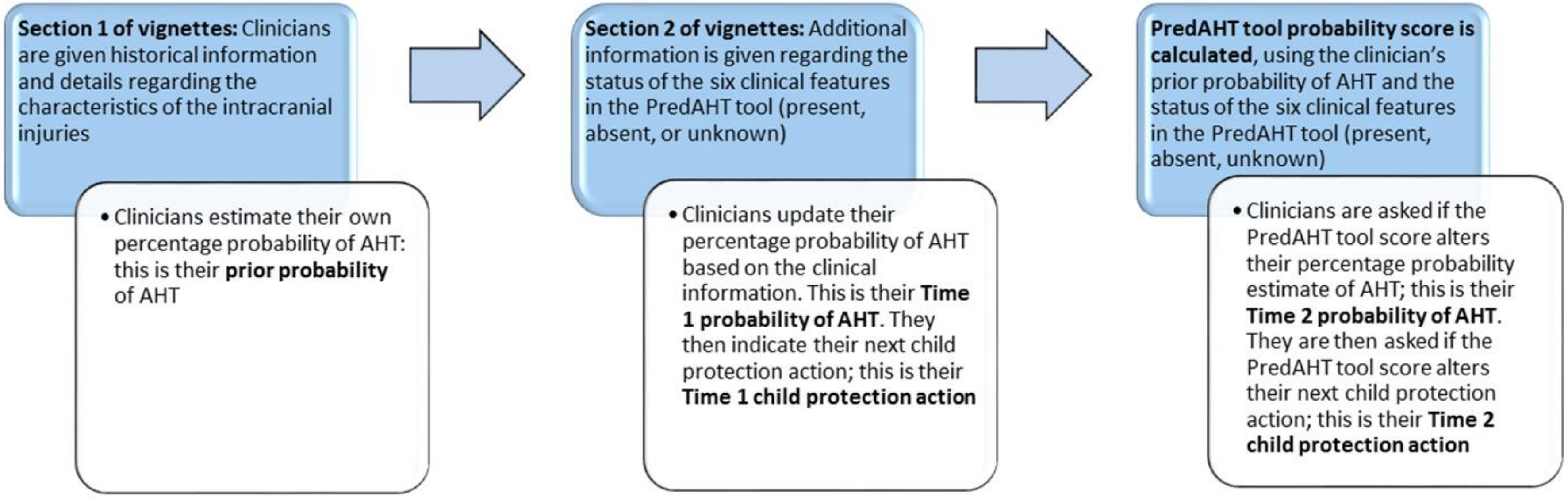
Clinician ID and specialty	**Vignette ID**	**Quote**
Clinician 16 Community pediatrician	V2:nAHT	*"I would still need more information about the cheek bruising but the low score (24%) would reassure me."*
Clinician 9 Community pediatrician	V3:AHT*	*"The history just doesn't fit with the level of trauma…the score helps to remind you that you are right to be concerned and helps you not to be too sensitive about the family."*
Clinician 10 Community pediatrician	V6:missing	*"The 7% would make me much more confident that this is an accident."*
Clinician 27 General pediatrician	V1:AHT	*"I think mostly where it helps is reassuring you."*
Clinician 16 Community pediatrician	V3:AHT*	*"My estimate is very close to PredAHT, so I wouldn't change my actions but my agreement with PredAHT would give me more confidence in expressing my opinion to multiagency colleagues."*
Clinician 17 Other specialty	V2:nAHT	*"Bruising to the cheeks made me worried but the tool would then reassure me to pull it back down."*
Clinician 25 General pediatrician	V3:AHT*	*"It would be helpful at the end to validate my opinion that probably it is abuse."*

Figure 1. Flowchart of clinicians participating in a vignette-based study investigating the potential impact of the Predicting Abusive Head Trauma clinical prediction tool


5 sites:
Teaching Hospitals (n=3)
District General Hospitals (n=2)
40 clinicians approached
Did not respond to initial contact (n=4)
Did not respond to follow-up contact (n=3)
Did not think the study was relevant for them (n=3)
Did not have time to participate (n=1)
29 participants
Community pediatricians* (n=15)
General pediatricians (n=9)
Emergency medicine pediatricians (n=2)
Pediatric radiologist (n=1)
Neuroradiologist (n=1)
Pediatric neurosurgical nurse (n=1)

*In the United Kingdom community pediatricians are doctors who have specialist training in child protection and safeguarding.

Figure 2. Flowchart of data collection procedure



Section 1 of vignettes: Clinicians are given historical information and details regarding the characteristics of the intracranial injuries

- Clinicians estimate their own percentage probability of AHT: this is their **prior probability** of AHT

Section 2 of vignettes: Additional information is given regarding the status of the six clinical features in the PredAHT tool (present, absent, or unknown)

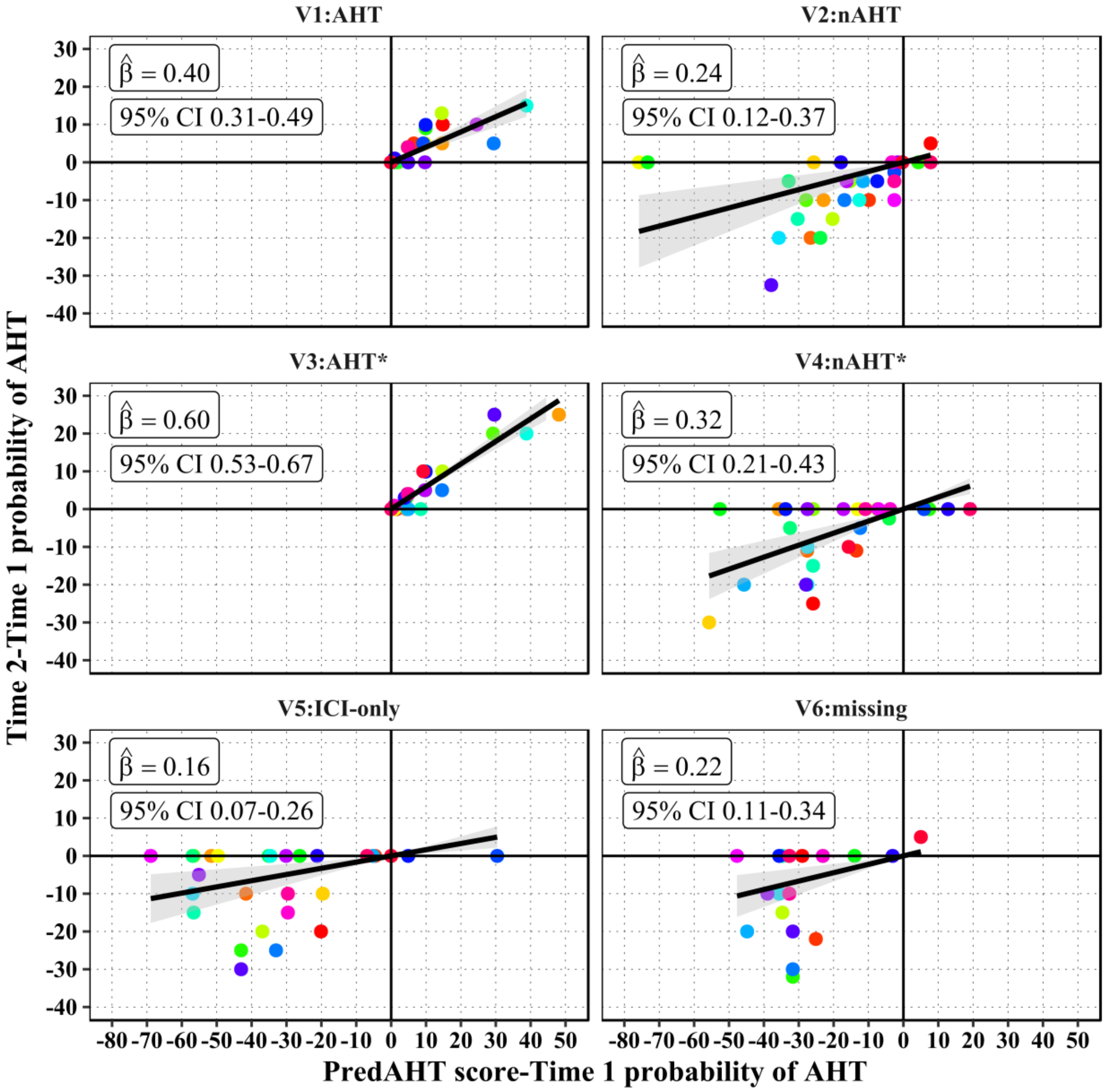
- Clinicians update their percentage probability of AHT based on the clinical information. This is their **Time 1 probability of AHT**. They then indicate their next child protection action; this is their **Time 1 child protection action**

PredAHT tool probability score is calculated, using the clinician's prior probability of AHT and the status of the six clinical features in the PredAHT tool (present, absent, unknown)

- Clinicians are asked if the PredAHT tool score alters their percentage probability estimate of AHT; this is their **Time 2 probability of AHT**. They are then asked if the PredAHT tool score alters their next child protection action; this is their **Time 2 child protection action**

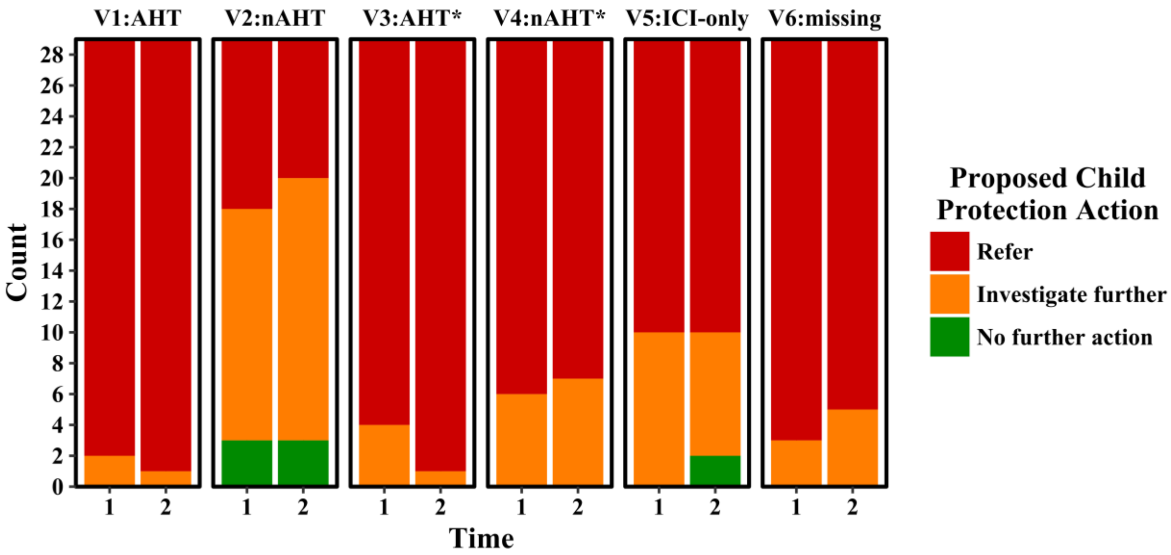
AHT = abusive head trauma, PredAHT = Predicting Abusive Head Trauma tool

Figure 3. The impact of the Predicting Abusive Head Trauma tool on clinicians' probability estimates of AHT for each of the six vignettes

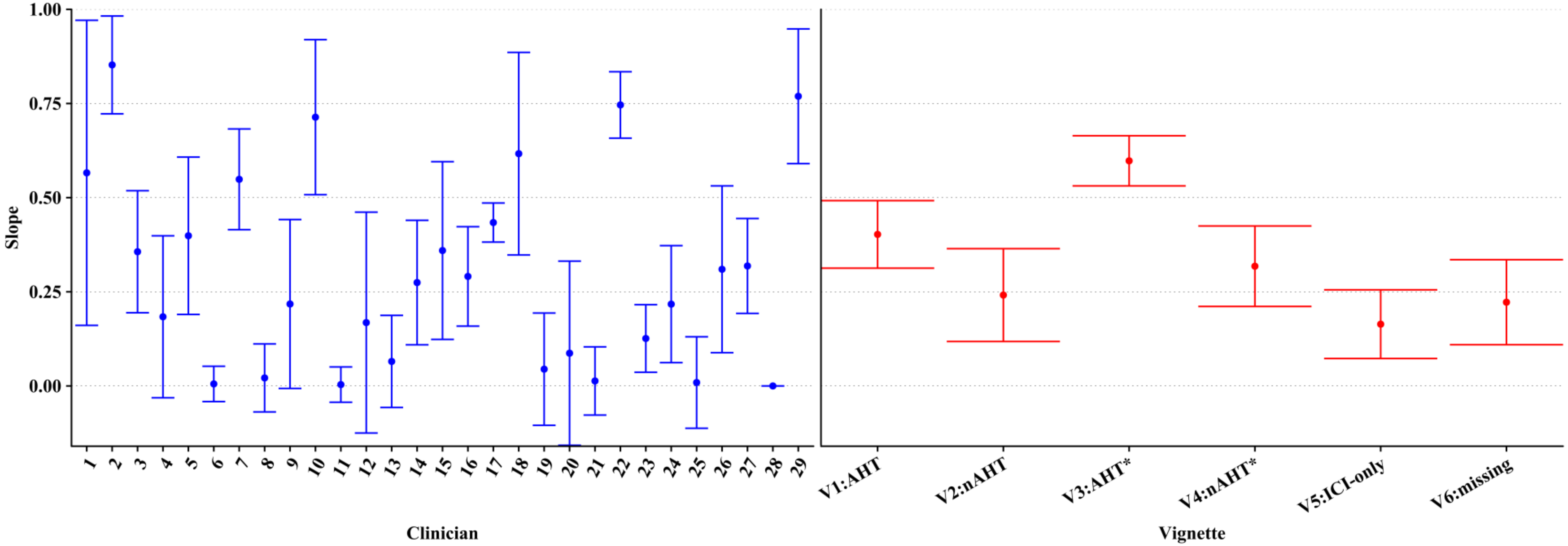


Colored dots represent different clinicians. Higher coefficients $\hat{\beta}$ indicate a greater impact of PredAHT on clinicians' probability estimates of AHT. Points at 0 on the x-axis indicate no difference between the clinicians' Time 1 probability estimate and the PredAHT score. Points at 0 on the y-axis indicate no change in clinicians' probability estimates of AHT from Time 1 to Time 2. Points greater than 0 on the y-axis indicate an increase in clinicians' probability estimates of AHT from Time 1 to Time 2. Points less than 0 on the y-axis indicate a decrease in clinicians' probability estimates of AHT from Time 1 to Time 2.

Figure 4. Clinicians' proposed Time 1 and Time 2 child protection actions for each of the six clinical vignettes

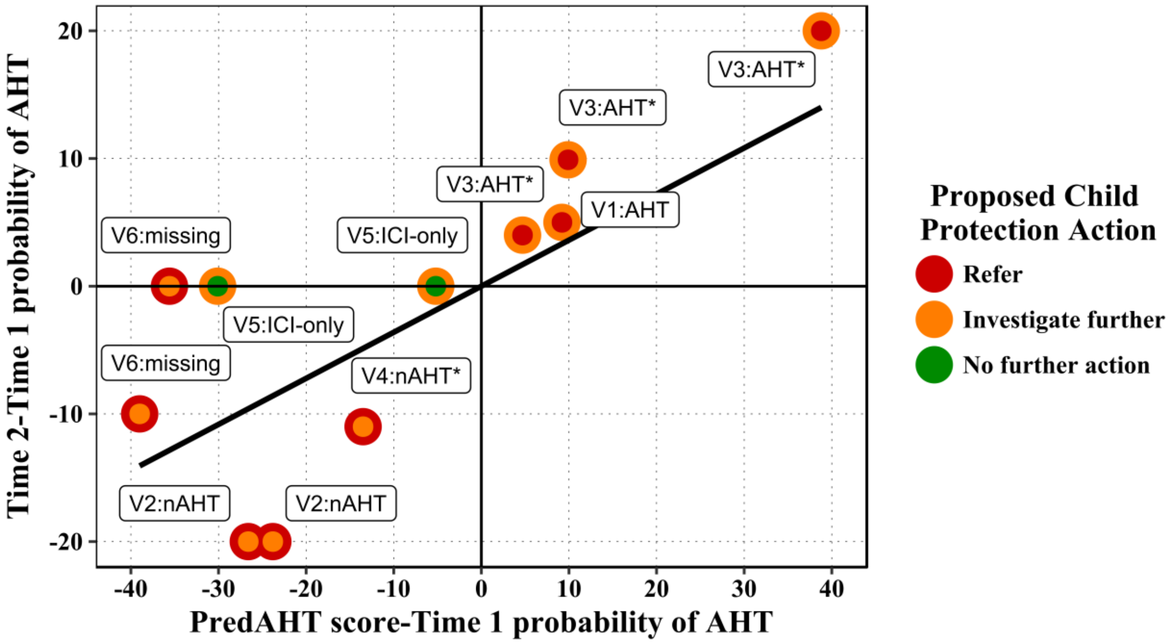


Supplementary Figure 1. Slope coefficients $\widehat{\beta}$ for each of the 29 participating clinicians relative to the slope coefficients $\widehat{\beta}$ for each of the six vignettes



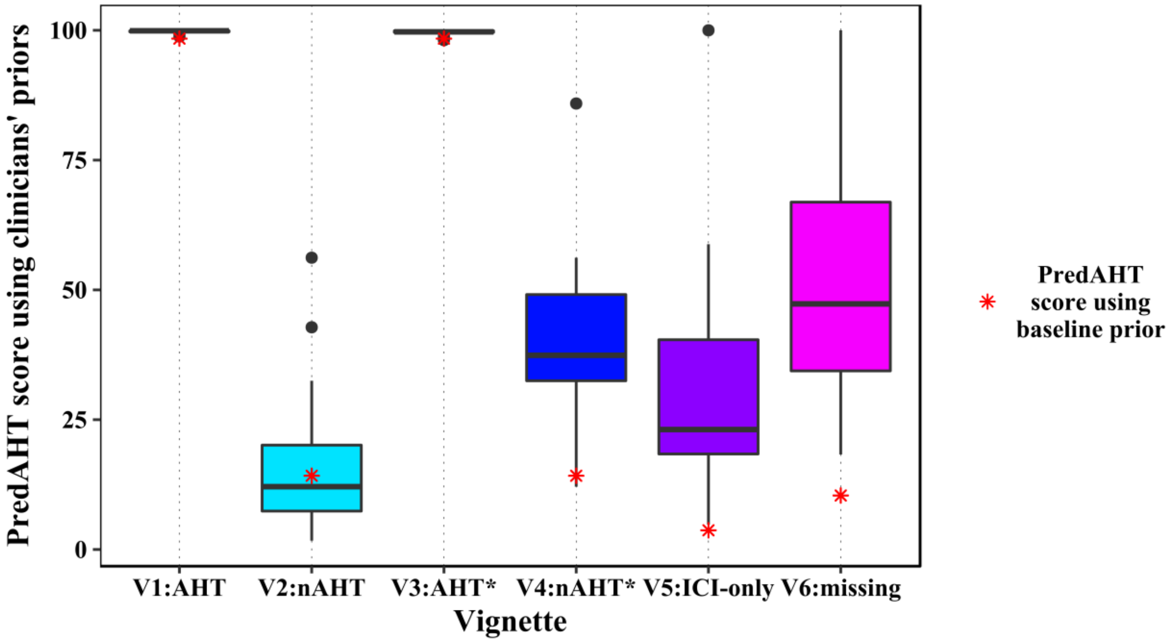
Error bars represent 95% confidence intervals. Higher coefficients $\hat{\beta}$ indicate a greater impact of the Predicting Abusive Head Trauma tool on clinicians' probability estimates of abusive head trauma.

Supplementary Figure 2. The impact of the Predicting Abusive Head Trauma (PredAHT) tool on clinicians' proposed child protection (CP) actions.



Seven clinicians changed their proposed CP action for one of the six vignettes. Two clinicians changed their proposed CP action for two of the six vignettes. Clinicians' Time 1 actions (before PredAHT) are indicated by the larger circle, and their Time 2 actions (after PredAHT) are indicated by the smaller circle. Points at 0 on the y-axis indicate that clinicians did not change their probability estimate of abusive head trauma (AHT) from Time 1 to Time 2, despite changing their CP action. Points greater than 0 on the y-axis indicate an increase in clinicians' probability estimates of AHT from Time 1 to Time 2. Points less than 0 on the y-axis indicate a decrease in clinicians' probability estimates of AHT from Time 1 to Time 2.

Supplementary Figure 3. Comparison of the Predicting Abusive Head Trauma (PredAHT) tool scores incorporating clinicians' priors with the PredAHT scores using the baseline prior, for each of the six vignettes.



PredAHT scores using clinicians' priors were higher than the PredAHT scores using the baseline prior in three of the six vignettes.

Supplementary Table 1. Clinical vignettes

V1:AHT	**Section 1:** A 3-month-old female infant presents to the hospital with lethargy and vomiting and no history of trauma. A CT scan of head reveals hypoxic ischaemic injury affecting both cerebral hemispheres, the brainstem and the thalami, and an acute (i.e. hyperdense) subdural haemorrhage at the vertex. **Section 2:** Apnoea and seizures are noted to be present, but there is no evidence of head or neck bruising. The ophthalmology exam reveals bilateral superficial and deep multi-layered retinal haemorrhages. Location: Zone 1 and outside Zone 1, posterior pole and periphery. Number: confluent. Additional features: macular detachment and retinal folds adjacent to the macular. A skeletal survey is not performed and so it is unknown whether the child has any rib or long-bone fractures.
V2:nAHT	**Section 1:** A 23-month-old female infant presents to the hospital immediately following a head trauma. Both parents, when interviewed separately, said that the child had fallen off a chair at a height of approximately 1.5 metres onto a tiled floor. Both parents' accounts remain consistent over time. Inflicted trauma is vehemently

	denied. A CT scan of head reveals a frontal lobe hyperdense subdural haemorrhage and a linear, undisplaced skull fracture of the left frontal parietal bone. **Section 2***:* Apnoea and seizures are noted to be absent. Bruising to the scalp and cheeks is noted. The ophthalmoscopy exam and skeletal survey are both negative.
V3:AHT*	**Section 1:** *A 14-month-old male infant* presents to the hospital with lethargy and vomiting. *His father states that he left the room momentarily and found him on the wooden floor after falling indoors. He brought him to the emergency department immediately and his story remains consistent over time. He denies inflicted trauma and states that the child has recently began to walk independently*. A CT scan of head reveals hypoxic ischaemic injury affecting both cerebral hemispheres, the brainstem and the thalami, and an acute (i.e. hyperdense) subdural haemorrhage at the vertex. **Section 2:** Apnoea and seizures are noted to be present, but there is no evidence of head or neck bruising. The ophthalmology exam reveals bilateral superficial and deep multi-layered retinal haemorrhages. Location: Zone 1 and outside Zone 1, posterior pole and periphery. Number: confluent. Additional features: macular detachment and retinal folds adjacent to the macular. A skeletal survey is not performed and so it is unknown whether the child has any rib or long-bone fractures.
V4:nAHT*	**Section 1:** A 23-month-old female infant *presents to the hospital with her mother. Initially no history of trauma is provided but following questioning the mother states that the child may have fallen* off a chair at a height of approximately 1.5 metres onto a tiled floor. *There are concerns about domestic violence within the family and there has been previous involvement with social services. It emerges that the incident occurred approximately six hours prior to presentation to the hospital.* A CT scan of head reveals a frontal lobe hyperdense subdural haemorrhage and a linear, undisplaced skull fracture of the left frontal parietal bone. **Section 2:** Apnoea and seizures are noted to be absent. Bruising to the scalp and cheeks is noted. The ophthalmoscopy exam and skeletal survey are both negative.
V5:ICI-only	**Section 1***:* A 3-month-old female infant presents to the hospital with lethargy and vomiting. The parents state that the baby rolled off the sofa onto the floor. A CT scan of head reveals multiple small bilateral subdural haemorrhages. **Section 2:** Apnoea and seizures are noted to be absent, and there is no evidence of head or neck bruising. The ophthalmology exam and skeletal survey are both negative.
V6:missing	**Section 1:** A 3-month-old female infant presents to the hospital with lethargy and vomiting. The parents state that the baby rolled off the sofa onto the floor. A CT scan of head reveals multiple small bilateral subdural haemorrhages. **Section 2:** Apnoea and seizures are noted to be absent, and there is no evidence of head or neck bruising. *An ophthalmology exam and a skeletal survey have not yet been performed, and so it is unknown whether the child has any rib or long-bone fractures, or retinal haemorrhages.*

Appendix 1. Methodological recommendations and best practices for designing vignette studies

Recommendations for vignette content

Developed from:

Evans, S.C., Roberts, M.C., Keeley, J.W., Blossom, J.B., Amaro, C.M., Garcia, A.M,…Reed, G.M. (2015). Vignette methodologies for studying clinicians' decision-making: Validity, utility, and application in ICD-11 field studies. *International Journal of Clinical & Health Psychology*, *15*(2), 160–70.

Recommendation number	Vignettes should	Reported:
1	Derive from the literature and/or clinical experience	Vignette design
2	Be clear, well-written and carefully edited	Vignette design; vignettes were reviewed by supervisory team and edited accordingly, and piloted before use. See vignettes, Supplementary Table 1
3	Not be longer than necessary (typically between 50 and 500 words)	See vignettes, Supplementary Table 1
4	Follow a narrative, story-like progression	See vignettes, Supplementary Table 1. Initial information presented in section 1 followed by additional clinical details in section 2
5	Follow a similar structure and style for all vignettes in the study	See vignettes, Supplementary Table 1. All vignettes followed a similar style and structure
6	Use present tense (past tense only for history and background information)	See vignettes, Supplementary Table 1. All written in present tense
7	Avoid placing the participant "in the vignette" (e.g. as first or third-person character)	See vignettes, Supplementary Table 1. Participants were not "placed in the vignette" but were asked to answer survey questions as they would in clinical practice
8	Balance gender and age across vignettes	See vignettes, Supplementary Table 1.
9	Be as neutral as possible with respect to cultural and socio-economic factors,	See vignettes, Supplementary Table 1.

	unless these are included among the experimental variables	Cultural and socio-economic factors were not included as variables nor mentioned
10	Resemble real people, not a personification of a list of symptoms or behaviours	See vignettes, Supplementary Table 1
11	Be relatable, relevant, and plausible to participants	Vignette design. The vignettes were piloted and were felt to be clear and to reflect plausible cases. This was further confirmed by the 'think-aloud' technique
12	Avoid "red herrings", misleading details, and bizarre content	See vignettes, Supplementary Table 1. There was no misleading or bizarre content, vignettes were designed to represent plausible cases
13	Highlight the key variables of interest, facilitating experimental effects	Changes to key variables were indicated in italics
14	Facilitate participant engagement and thinking by including vague or ambiguous elements	Four vignettes were designed as "grey" cases to introduce uncertainty into the decision and stimulate reasoning. This was confirmed by the 'think-aloud' technique
15	Cover all pertinent variables (or omit selected variables for specific purposes)	It was not possible to cover all pertinent variables, the omission of certain information led to useful insights in itself, as confirmed by the qualitative analysis of the 'think-aloud' data

Best practice recommendations for designing and implementing experimental vignette methodology studies

Developed from:

Aguinis, H., & Bradley, K.J. (2014). Best practice recommendations for designing and implementing experimental vignette methodology studies. *Organizational Research Methods, 17*(4), 351–371.

Item number	Guide questions/description	Reported:
Planning an EVM study		

Decision Point 1	Deciding whether EVM is a suitable approach	Introduction
Decision Point 2	Choosing the type of EVM	Paper people study
Decision Point 3	Choosing the type of research design	Within-person fully-crossed design where all clinicians completed all vignettes
Decision Point 4	Choosing the level of immersion	Written vignette only
Decision Point 5	Specifying the number and levels of the manipulated factors	Three factors each with two levels (Concerning history yes/no, concerning social history yes/no, missing data yes/no)
Decision Point 6	Choosing the number of vignettes	Six
Implementing an EVM study		
Decision Point 7	Specifying the sample and number of participants	Clinicians from a variety of specialities involved in suspected AHT cases. 40 were approached to take part
Decision Point 8	Choosing the setting and timing for administration	At the participants workplace in a single session
Decision Point 9	Choosing the best method for analysing the data	Linear models and linear mixed effects models
Reporting results of an EVM study		
Decision Point 10	Choosing how transparent to be in the final presentation of results and methodology	See methods and results. Full vignettes provided plus detailed description of their derivation, the analysis and the results

Appendix 2. Quantitative Analysis

Analysis focused on determining the impact of PredAHT on clinicians' probability estimates of AHT and their proposed CP actions, and assessing the degree of agreement between their probability estimates and between their proposed CP actions both before, and after, seeing the PredAHT score.

Exploratory data analysis

Exploratory data analysis was conducted through graphical displays, to determine plausible models for the data and examine relationships between variables.

Impact of PredAHT on clinicians' probability estimates of AHT

To assess the impact of PredAHT on clinicians' probability estimates of AHT, six linear models were fitted for each vignette, using the formula:

$$y_2 = y_1 + \beta\,(t - y_1) + \varepsilon$$

, where y_2 is the Time 2 probability estimate, y_1 is the Time 1 probability estimate, t is the PredAHT score with clinicians' priors, β is the slope, and ε is the error term. The slope β represents the average proportion of the distance between y_1 and t that clinicians move after seeing the PredAHT score. For example, if $y_1 = 50\%$, $t = 70\%$ and $y_2 = 60\%$, then $\beta = 0.5$. A slope of 0 indicates no difference between clinicians' Time 1 and Time 2 probability estimates on average (if $\beta = 0$, $y_2 = y_1 + \varepsilon$), while a slope of 1 means that clinicians Time 2 estimates are the same as the PredAHT score on average (if $\beta = 1$, $y_2 = t + \varepsilon$). The intercept was not included, as the expected value of the independent variable given that the dependent variable is 0, is 0 (Eisenhauer, 2003). In other words, if $y_1 - t = 0$, then $y_2 - y_1$ will also equal 0 on average.

Next we assessed the overall impact of PredAHT across vignettes. Due to the multilevel nature of the data (vignette level and clinician level), analyses that focus on both levels simultaneously must be used (Atzmüller & Steiner, 2010; Aguinis & Bradley, 2014). We fitted several linear mixed models, with random effects at the clinician and vignette levels. To examine the influence of demographics (hospital type, clinician specialty, clinician age, years of CP experience, pediatric head injury training, clinician seniority), we compared a reduced model with the fixed effect described in the formula above and the random effects, with six models allowing the average proportion moved to vary across the categorical demographic variables. The R package ‘lme4’ was used for mixed model fitting (Bates, Maechler, Bolker & Walker, 2015). Models were fitted using the maximum likelihood criterion, to enable comparison using likelihood ratio tests. Profile likelihood confidence intervals were computed for model parameters. The R package ‘multcomp’ was used to obtain p-values for fixed effects coefficients (Hothorn, Bretz & Westfall, 2008), and the ‘lsmeans’ package was used for pairwise comparisons between factor levels (Lenth, 2016).

Impact of PredAHT on clinicians' proposed CP actions

To analyze whether certain clinicians were more likely to change their child protection action after seeing the PredAHT score, the chi-square test of independence and Fisher's exact test were used to examine associations between categorical variables (change in CP action vs. hospital type, clinician specialty, clinician age, years of CP experience, pediatric head injury training, and clinician seniority). A change in child protection action was specified as one associated with an increase or decrease in the level of concern from Time 1 to Time 2 (see Table 3 in the manuscript).

Inter rater reliability of clinicians' probability estimates of AHT

Inter-rater reliability statistics were estimated to analyze the degree of agreement between clinicians in their three probability estimates across vignettes. The intra-class correlation (ICC) statistic was most suitable because the data are continuous, multiple clinicians participated, and the design is fully crossed, with all clinicians rating all vignettes (Hallgren, 2012). ICCs were obtained based on a 2-way random effects model with absolute agreement (ICC, 2), with single-measures ICCs reported. The R package 'psych' was used (Revelle, 2017).

Inter rater reliability of clinicians' proposed CP actions

Inter-rater agreement between clinicians' child protection actions at Time 1 and Time 2 was estimated using Gwet's AC_1 coefficient (Gwet, 2010). This method was chosen due to the paradoxes exhibited by the kappa statistic (Feinstein & Cicchetti, 1990). An alternative and more stable multiple-rater agreement coefficient, the AC_1, was proposed by Gwet (Gwet, 2008). Gwet's AC_1 has been proven to be robust to the "kappa paradox" and to demonstrate plausible values in line with observed percent agreement values (Gwet, 2002a; Gwet, 2002b; Wongpakaran et al., 2013; Walsh et al., 2014; Zec et al., 2017). We used jackknifing to estimate the variance due to the sampling of clinicians (Gwet, 2012). The jackknife is a resampling method particularly useful for variance estimation. In the simplest case, used here, jackknife resampling is accomplished by sequentially deleting single cases from the original sample (Friedl & Stampfer, 2014).

Impact of clinicians' prior probabilities on the PredAHT score

We compared the PredAHT score with and without clinicians' prior probabilities, to assess the impact of clinicians' priors for each vignette.

References

Aguinis, A., & Bradley, K.J. (2014). Best practice recommendations for designing and implementing experimental vignette methodology studies. *Organizational Research Methods,* 17(4): 351–371.

Atzmüller, C., & Steiner, P.M. (2010). Experimental vignette studies in survey research. *Methodology: European Journal of Research Methods for the Behavioral & Social Sciences,* 6(3): 128–138.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48.

Eisenhauer, J.G. (2003). Regression through the origin. *Teaching Statistics: An International Journal for Teachers,* 25(3): 76–80.

Feinstein, A.R., & Cicchetti, D.V. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology, 43*(6), 543–549.

Friedl, H., & Stampfer, E. (2014). Jackknife resampling. *In:* Balakrishnan, N., Colton, T., Everitt, B., Piegorsch, W., Ruggeri, F., & Teugels, J.L. (Eds.) *Wiley StatsRef: Statistics Reference Online*.

Gwet, K.L. (2002a). Kappa statistic is not satisfactory for assessing the extent of agreement between raters. *Statistical Methods for Inter-Rater Reliability Assessment,* 1, 1–6.

Gwet, K.L. (2002b). Inter-rater reliability: Dependency on trait prevalence and marginal homogeneity. *Statistical Methods for Inter-Rater Reliability Assessment,* 2, 1–9.

Gwet, K.L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical & Statistical Psychology,* 61(1): 29–48.

Gwet, K.L. (2010). Inter-rater reliability with R: R functions for calculating agreement coefficients. Retrieved from: http://www.agreestat.com/r_functions.html

Gwet, K.L. (2012). Agreement coefficients and statistical inference. *In:* Gwet, K.L. (Ed.) *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among multiple raters* (93-119). Gaithersburg, MD: Advanced Analytics, LLC.

Hallgren, K.A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials for Quantitative Methods in Psychology*, 8(1), 23–34.

Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal*, *50*(3), 346–363.

Lenth, R.V. (2016). Least-squares means: The R package lsmeans. *Journal of Statistical Software*, *69*(1), 1–33.

Revelle, W. (2017). psych: Procedures for personality and psychological research, Northwestern University, Evanston, Illinois, USA. https://CRAN.R-project.org/package=psych Version = 1.7.3.

Walsh, P., Thornton, J., Asato, J., Walker, N., McCoy, G., Baal, J.,..Banimahd, F. (2014). Approaches to describing inter-rater reliability of the overall clinical appearance of febrile infants and toddlers in the emergency department. *PeerJ*, *2*, e651.

Wongpakaran, N., Wongpakaran, T., Wedding, D., & Gwet, K.L. (2013). A comparison of Cohen’s Kappa and Gwet’s AC1 when calculating inter-rater reliability coefficients: A study conducted with personality disorder samples. *BMC Medical Research Methodology*, *13*, 61.

Zec, S., Soriani, N., Comoretto, R., & Baldi, I. (2017). High agreement and high prevalence: The paradox of Cohen's kappa. *The Open Nursing Journal*, *11*, 211–218.

Appendix 3. Analytic Framework

Theme	Category	Definitions
Rationale for responses	**Gray cases difficult**	Any comments about the ease or difficulty of estimating the probability of AHT or deciding on proposed child protection actions for the “gray” cases; any reasons why the “gray” cases were difficult to classify
	Impact of social history	Any comments about the impact the social history had on participants estimated probabilities or proposed child protection actions; any comparisons between “V2:nAHT” and “V4:nAHT*”
	History doesn’t match level of trauma	Any discussions about the impact the history had on participants estimated probabilities or proposed child protection actions in “V1:AHT” and “V3:AHT*”; any comparisons between “V1:AHT” and “V3:AHT*”
	Agreement/disagreement with tool	Any reasons why participants disagreed with the PredAHT score and therefore did not change their probability estimates or proposed child protection actions at Time 2. Any reasons why participants agreed with the PredAHT score and therefore did change their probability estimates or proposed child protection actions at Time 2
	Knowledge of clinical features	Any comments about the impact participants’ knowledge of the clinical features indicative of AHT and nAHT had on their probability estimates or proposed child protection actions
	Developmental stage	Any considerations about the child’s age and developmental stage when completing the vignettes
	Consistent history	Any discussions about the impact a consistent or inconsistent history had on participants probability estimates or proposed child protection actions
	Mechanism of injury	Any considerations about the proposed mechanism of injury and whether this was consistent with the clinical features and level of trauma observed
Evaluations of PredAHT	**Potential benefits**	Any discussions about whether PredAHT would be useful for participants in their clinical practice and why; comments about how PredAHT might help participants to quantify risk; comments about how PredAHT could reassure participants that their suspicions (or lack thereof) are justified and provide them with confidence in their opinions
	Potential risks	Any discussions about the potential risks or downsides of using PredAHT including comments about important features missing from PredAHT; comments about potential false reassurance from a low score; comments about how PredAHT cannot take into account all potential indicators of abuse or nuances in individual cases; comments about the need to understand and explain how PredAHT works, and appraise the quality of the underlying data and the accuracy of the scores
Interpretations of probabilities	**Threshold probability**	Any comments about participants’ accepted probability thresholds for investigation and referral of suspected AHT
	Impact of the prior probability	Any discussions about the impact participants estimated prior probabilities had on the post-test probability provided by PredAHT; any reasons participants gave for their prior probabilities; discussions about how participants would estimate a prior probability in practice and the information they would use to do this

Comments on details of the vignettes	**Investigations**	Any comments about why certain investigations were or were not performed; comments about additional investigations participants would order that are not detailed in the vignettes
	Detail of the history/clinical features	Any discussions about needing additional detail about the history or clinical findings in order to estimate the probability of AHT, including the age and pattern of clinical findings or more detail on the proposed mechanism of injury
	Differential diagnoses	Any comments about the differential diagnoses, not detailed in the vignettes, that participants would rule out in practice