Interobserver Agreement in Assessment of Clinical Variables in Children with Blunt Head Trauma

Marc H. Gorelick, MD, MSCE, Shireen M. Atabaki, MD, MPH, John Hoyle, MD, Peter S. Dayan, MD, MSc, James F. Holmes, MD, MPH, Richard Holubkov, PhD, David Monroe, MD, James M. Callahan, MD, Nathan Kuppermann, MD, MPH, for the Pediatric Emergency Care Applied Research Network (PECARN)

Abstract

**Objectives:** To be useful in development of clinical decision rules, clinical variables must demonstrate acceptable agreement when assessed by different observers. The objective was to determine the inter-observer agreement in the assessment of historical and physical examination findings of children undergoing emergency department (ED) evaluation for blunt head trauma.

**Methods:** This was a prospective cohort study of children younger than 18 years evaluated for blunt head trauma at one of 25 EDs in the Pediatric Emergency Care Applied Research Network (PECARN). Patients were excluded if injury occurred more than 24 hours prior to evaluation, if neuroimaging was obtained at another hospital prior to evaluation, or if the patient had a clinically trivial mechanism of injury. Two clinicians independently completed a standardized clinical assessment on a templated data form. Assessments were performed within 60 minutes of each other and prior to clinician review of any neuroimaging (if obtained). Agreement between the two observers beyond that expected by chance was calculated for each clinical variable, using the kappa ($\kappa$) statistic for categorical variables and weighted kappa for ordinal variables. Variables with a lower 95% confidence limit (LCL) of $\kappa > 0.4$ were considered to have acceptable agreement.

**Results:** Fifteen-hundred pairs of observations were obtained. Acceptable agreement was achieved in 27 of the 32 variables studied (84%). Mechanism of injury (low, medium, or high risk) had $\kappa = 0.83$. For subjective symptoms, kappa ranged from 0.47 (dizziness) to 0.93 (frequency of vomiting); all had 95% LCL > 0.4. Of the physical examination findings, kappa ranged from 0.22 (agitated) to 0.89 (Glasgow Coma Scale [GCS] score). The 95% LCL for kappa was <0.4 for four individual signs of altered mental status and for quality (i.e., boggy or firm) of scalp hematoma if present.

**Conclusions:** Both subjective and objective clinical variables in children with blunt head trauma can be assessed by different observers with acceptable agreement, making these variables suitable candidates for clinical decision rules.
Injuries are the leading cause of death in children older than 12 months, and traumatic brain injury (TBI) is the most common cause of mortality and morbidity among injured children. Computed tomography (CT) scanning is a widely accessible and sensitive tool for the detection of TBI. More than 260,000 CT scans are performed for evaluation of blunt head trauma in children annually in U.S. emergency departments (EDs), with a great deal of practice variation. A clinical decision rule in the setting of pediatric blunt head trauma could reduce both variation in practice and unnecessary CT scans. Such decision rules have helped optimize testing in the setting of adult head trauma, ankle injuries, and febrile illness.

Variables from the history and physical examination incorporated into clinical decision rules must be both reproducible and reliable. This is of particular importance in the pediatric population, where reliability of variables such as amnesia, headache, and dizziness depend on the child’s age and developmental level. The goal of our study was to determine interobserver agreement in the assessment of historical and clinical examination findings in children undergoing ED evaluation for blunt head trauma.

METHODS

Study Design
This was a multicenter cross-sectional study, conducted as part of a larger prospective cohort study to derive and validate a neuroimaging decision rule for children with blunt head trauma. The study was approved by the Human Subjects Review Boards of all participating institutions. Written or verbal informed consent for participation was obtained by clinicians or research staff at those institutions where it was required by the local institutional review board.

Study Setting and Population
This study was conducted at 25 EDs participating in the Pediatric Emergency Care Applied Research Network (PECARN), which includes both pediatric-specific and general EDs. All children younger than 18 years of age evaluated at one of the participating EDs for blunt head trauma between June 1, 2004, and May 1, 2006, were eligible for enrollment in the main derivation cohort study; a validation phase continued until September 23, 2006. Enrollment in our study occurred from June 1, 2004, to July 15, 2006, at which time the required sample size of 1,500 patients was achieved. The subjects in this study constitute a sample of patients enrolled in the main cohort study; this is a convenience sample of such patients who presented whenever two clinicians were available to perform independent clinical assessments. Patients with penetrating trauma, injury more than 24 hours prior to evaluation, or clinically insignificant injury were excluded. Clinically insignificant injury was defined a priori as both 1) minor mechanism of injury (fall from standing height or walking or running into a stationary object) and 2) no signs or symptoms of head injury or scalp laceration or abrasion alone.

Study Protocol
A clinician (faculty physician, pediatric emergency medicine fellow, resident, nurse practitioner, or physician assistant) performed a clinical evaluation and recorded results onto a structured case report form. If the primary evaluator was not a faculty or fellow physician, then the supervising faculty physician who examined the patient reviewed the data form, verified or corrected the recorded findings, and cosigned the form. A second faculty or fellow physician performed an independent evaluation and recorded the results separately, within 60 minutes of the first evaluation. Physician specialty (see Table 1) was determined based on the highest level of training. Clinical evaluations were performed prior to knowledge of the results of any imaging studies, if performed. Although site primary investigators were oriented to the case report form in training sessions prior to and during the study, and were provided a detailed study manual of operations, no standardized training of all clinicians evaluating patients in this study in eliciting clinical findings was done prior to or during the study.

Measurements
The case report form included information about clinical findings in several domains. In the analysis, some of these findings were categorized in alternative ways or combined with others to yield composite variables. The domains evaluated included mechanism of injury (evaluated as two different categorical variables), symptoms (7 findings, 11 variables), and physical examination (14 findings, 19 variables). Several findings (dizziness, headache, and amnesia) were not applicable to preverbal children and were only measured in verbal children for whom these questions were developmentally appropriate.

Data Analysis
Data analyses were performed using SAS 9.1.3 (SAS Institute, Cary, NC). Chance-adjusted agreement was calculated using the kappa statistic. Patients with missing values for a variable per either physician were excluded from the kappa analysis for that variable. For ordinal variables, the Fleiss-Cohen weighted kappa was used, with standard quadratic weights. The lower one-sided 95% confidence limit (LCL) was calculated using normal approximation methods. Agreement was considered acceptable if the LCL was greater than 0.4, indicating agreement that is at least fair according to
most existing guidelines. Our sample size was selected primarily on the basis of investigator consensus that a minimum of 20 subjects at each site would be desirable to maximize generalizability of the findings. In addition, we wanted to be able to demonstrate a LCL of the kappa of greater than 0.4 in a “worst-case” scenario, assuming a point estimate of the kappa of 0.5, for an uncommon finding (i.e., prevalence of 5%). This yielded a sample size estimate of 1,500 subjects. Finally, we also conducted an analysis in which we stratified subjects by the time between paired assessments (within 30 minutes vs. greater than 30 minutes).

RESULTS

A total of 1,500 subjects had paired independent observations by two clinicians, representing 3.7% of the 40,226 patients evaluated during the interobserver agreement enrollment period. The demographic and injury severity characteristics of the subjects in the interobserver agreement sample were similar to those in the full derivation cohort (Table 1). The following findings were not evaluated in patients in whom they were felt to be developmentally inappropriate: dizziness (not evaluated by one or both clinicians in 688 patients), amnesia (729 patients), and headache (375 patients). As these specific findings were not evaluated by either clinician for the vast majority of children younger than 2 years, the corresponding kappa statistics are reported only for patients 2 years or older. Repetitive speech was also assessed only for children older than 2 years, while bulging fontanelle was only assessed only for children younger than 1 year. Of our study population, 392 (26%) were younger than 2 years. The proportion of children with truly missing values from one or both raters (i.e., a variable was considered age appropriate but one or both ratings were missing) was generally low, ranging from 1% (scalp hematoma, mechanism of injury) to 23% (severity of headache).

The primary specialties of the clinicians performing the evaluations are shown in Table 2. The majority of clinicians performing the first evaluation were faculty physicians (43.6%) or fellows (18.0%); the remainder were resident physicians (35.6%) or allied health providers (2.9%) whose forms were reviewed and countersigned by an attending physician.

Kappa statistics are shown in Table 3 (historical variables) and Table 4 (physical examination findings). Overall, 27 (84%) of 32 variables considered individually or in combination met the criteria for acceptable agreement (i.e., kappa LCL of >0.4). The point estimate of the kappa was greater than 0.4 for all but three of the variables and 0.6 or greater for 19 of the 32 variables. All of the historical variables had acceptable agreement. Five of the physical examination findings, however, had unacceptable agreement. The presence or absence of scalp hematoma showed excellent agreement (κ = 0.65, LCL = 0.62), as did hematoma size and location (the latter being assessed only for patients with hematoma reported present). However, quality of scalp hematoma (firm vs. boggy) had unacceptable agreement (κ = 0.36, LCL = 0.28). Level of alertness, as indicated by the Glasgow Coma Scale (GCS) score, had acceptable agreement when considered as an ordinal variable, or dichotomized as 15 versus less than 15. We elicited four variables representing signs of altered mental status other than GCS. These included “agitated,” “slow to respond,” “sleepy,” and “repetitive speech.” All four individually had unacceptable agreement, with a LCL for κ < 0.4 for each. However, a composite variable for the presence or absence of any of these indicators of altered mental status had acceptable agreement (κ = 0.48, LCL = 0.42). Agreement was further improved when the definition of altered mental status was expanded to include these four specific indicators or GCS < 15 (κ = 0.53, LCL = 0.48).

When we analyzed interobserver agreement with subjects stratified by the time between paired assessments (within 30 minutes vs. greater than 30 minutes), we found that in general, agreement was slightly higher for assessments done within 30 minutes (data not shown). However, none of the five physical findings with a LCL for kappa below 0.4 in the full population had acceptable agreement when analysis was restricted to those reassessed within the shorter time frame.

DISCUSSION

In this large, multicenter study of blunt head trauma in children, agreement between observers was acceptable (kappa statistically significantly better than 0.4) for the large majority of findings elicited by history and physical

### Table 1

Characteristics of Supervising Physicians Performing Assessments on the Two Forms (n = 1,500)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pediatric emergency medicine</strong></td>
<td>724</td>
<td>159</td>
<td><strong>Emergency Medicine</strong></td>
<td>69</td>
<td>3</td>
<td>15</td>
</tr>
<tr>
<td>Pediatrics</td>
<td>200</td>
<td>137</td>
<td>Emergency Medicine</td>
<td>9</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>Emergency medicine</td>
<td>116</td>
<td>1</td>
<td>Emergency Medicine and Pediatrics</td>
<td>34</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Emergency medicine and pediatrics (dual certification)</td>
<td>6</td>
<td>0</td>
<td>Family medicine/Internal medicine/other</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Family medicine/internal medicine/other</td>
<td>9</td>
<td>3</td>
<td></td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
examination. The primary exception was in the area of altered mental status. Although agreement on specific individual aspects of altered mental status was poor, there was good agreement on the presence of any alteration in mental status. Our results suggest that the findings we studied can be elicited with sufficient reliability that they would be good candidates for inclusion in a clinical decision rule to stratify risk of injury or specific outcomes in children with blunt head trauma.

The generalizability of these results is enhanced by the diversity of settings and evaluators. The 25 study hospitals included both pediatric-specific and general EDs, located in urban, suburban, and rural settings. Similarly, we had a diverse group of observers with regard to their background and type of training, including fellows and residents in training, as well as allied health professionals. Although the majority of supervising physicians were pediatric emergency physicians, nearly one-half were trained in emergency medicine, general pediatrics, or another specialty without subspecialty training. Because nearly 90% of pediatric emergency visits in the United States are to general EDs,15 it is important that our results be applicable to the full spectrum of settings and providers who may be called upon to assess children with head injury and make decisions about neuroimaging and referral.

Although there are limited data on interobserver agreement in the assessment of signs and symptoms of head injury in children, our results are generally consistent with those of others. Altered mental status, specifically as assessed by the GCS and other rating scales, has been most extensively studied, although much of the published work is in adult patients,16,17 or in patients with conditions other than acute head injury.18,19 In a study of children with minor head injury (defined as GCS 13–15) at nine Canadian pediatric teaching hospitals, Osmond et al.20 found moderate

### Table 2

<table>
<thead>
<tr>
<th>Characteristics of Patients in the Interobserver Agreement Study and the Full Main Cohort Study</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean age (yr)</td>
</tr>
<tr>
<td>Gender (% male)</td>
</tr>
<tr>
<td>Race (% white)</td>
</tr>
<tr>
<td>Disposition (% discharged home from ED)</td>
</tr>
<tr>
<td>Injury on cranial CT (of those with CT scan performed, % with intracranial injuries)</td>
</tr>
</tbody>
</table>

CT = computed tomography; ED = emergency department.

### Table 3

<table>
<thead>
<tr>
<th>Interobserver Agreement for Historical Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>All subjects (n = 1,500)</td>
</tr>
<tr>
<td>Mechanism of injury (low, medium, high risk*)</td>
</tr>
<tr>
<td>Mechanism of injury (low/medium vs. high risk)</td>
</tr>
<tr>
<td>Dizziness</td>
</tr>
<tr>
<td>Amnesia for event</td>
</tr>
<tr>
<td>Any loss of consciousness (LOC)</td>
</tr>
<tr>
<td>Yes vs. no</td>
</tr>
<tr>
<td>Yes/suspected vs. no</td>
</tr>
<tr>
<td>LOC duration† (none, &lt;5 seconds, 5 seconds to &lt;1 minute, 1–5 minutes, &gt;5 minutes)</td>
</tr>
<tr>
<td>Seizure</td>
</tr>
<tr>
<td>Acting normal according to the parent</td>
</tr>
<tr>
<td>Headache severity† (none, mild, moderate, severe)</td>
</tr>
<tr>
<td>Vomiting frequency† (none, once, twice, &gt;2 times)</td>
</tr>
<tr>
<td>—</td>
</tr>
<tr>
<td>84% none, 6% once, 3% twice, 7% &gt;twice</td>
</tr>
</tbody>
</table>

*High risk: motor vehicle crash involving ejection, rollover, death of other occupant; pedestrian or unhelmeted bicyclist struck by motorized vehicle; or fall > 5 feet; low risk: fall from ground level or walking/running into stationary object; all other mechanisms are medium risk.
†Fleiss-Cohen weighted kappa.
LCL = lower confidence limit.
agreement in initial GCS, with $\kappa = 0.54$. Holmes et al.\textsuperscript{21} studying children with head injury of any severity at a single ED, found excellent agreement on GCS, with weighted $\kappa = 0.77$ among patients 2 years of age and younger, and $\kappa = 0.91$ for older children. Osmond et al.\textsuperscript{20} found "moderate" agreement ($\kappa$ between 0.4 and 0.6) for 3 of 12 findings evaluated and substantial agreement ($\kappa > 0.6$) for 8 of the 12 findings. Palchak et al.\textsuperscript{22} studying 109 children with head injury at a single general emergency department, found acceptable agreement ($\kappa > 0.5$) for all 10 findings they evaluated. Our study confirms these previous results, but also extends them to a larger number of patients and to a greater variety of settings and observers.

**LIMITATIONS**

Although our study took place at 25 EDs representing a spectrum of settings and types, most are teaching centers. Our results may not be applicable to smaller community EDs where clinicians may have less exposure to children with head injury. However, our clinician evaluators represent a broad diversity of clinical training, which is likely to be seen in community EDs. In addition, despite the large sample size, some of the findings (e.g., bulging fontanelle, palpable fracture) had a low prevalence, leading to relatively wide confidence intervals and making these variables difficult to assess statistically. Finally, although missing values were generally uncommon, since patients for whom one or both physicians were unable to assess a finding were excluded from analyses for that finding, results for a few findings with higher proportions of missing values could be subject to bias if the frequency of being missing is not random.

**CONCLUSIONS**

There is acceptable interobserver agreement, among a diverse group of clinicians, in the assessment of most historical and physical examination findings in children with blunt head trauma. Our results suggest that these findings can be obtained reliably and therefore could be considered for incorporation in a decision rule to identify children with blunt head trauma at risk for intracranial injury.

**References**


* deceased