Observation Scales to Identify Serious Illness in Febrile Children

Paul L. McCarthy, Michael R. Sharpe, Sydney Z. Spiesel, Thomas F. Dolan, Brian W. Forsyth, Thomas G. DeWitt, Howard D. Fink, Michael A. Baron and Domenic V. Cicchetti

Pediatrics 1982;70;802-809

Observation Scales to Identify Serious Illness in Febrile Children

Paul L. McCarthy, MD, Michael R. Sharpe, MD, Sydney Z. Spiesel, MD, PhD, Thomas F. Dolan, MD, Brian W. Forsyth, MD, Thomas G. DeWitt, MD, Howard D. Fink, MD, Michael A. Baron, MD, and Domenic V. Cicchetti, PhD

From the Yale University School of Medicine, Yale-New Haven Hospital, and Veterans Administration Medical Center, West Haven, Connecticut

ABSTRACT. The pediatrician makes a judgment of the degree of illness (toxicity) of a febrile child based on observation prior to history and physical examination. In order to define valid and reliable observation data for that judgment, data from two previous studies were used to construct three-point scales of 14 observation items correlated with serious illness in those reports. Between Nov 1, 1980 and March 1, 1981, these 14 scaled items were scored simultaneously by attending physicians, residents, and nurses prior to history and physical examination on 312 febrile children aged ≤24 months seen consecutively in our Primary Care Center-Emergency Room and in one private practice. Of these 312 children, 37 had serious illness. Multiple regression analysis based on patients seen by at least one attending physician in the Primary Care Center revealed six items (quality of cry, reaction to parents, state variation, color, state of hydration, and response to social overtures) that were significant and independent predictors of serious illness (multiple R = 0.63). The observed agreement for scoring these six items between two attending physicians who saw one third of the patients ranged from 88% to 97%. The chance corrected agreement levels (κw) for these six items were, with one exception, clinically significant (κw = .47 to .73). A discriminant function analysis revealed that these six items when used together had a specificity of 88% and a sensitivity of 77% for serious illness. Individual scores for each of the six key items were added to yield a total score for each patient. Only 2.7% of patients with a score ≤10 had a serious illness; 92.3% with a score ≥16 had a serious illness. The sensitivity of the six-item model for serious illness when combined with history and physical examination was 92%. In the population studied, this predictive model, when used prior to history and physical examination, was reliable, predictive, specific, and sensitive for serious illness in febrile children. It was most sensitive when combined with history and physical examination. The model will need to be validated on a new population of patients. *Pediatrics* 70:802–809, 1982; *clinical judgment, diagnosis, fever, toxicity.*

Received for publication Nov 9, 1981; accepted April 7, 1982.
Reprint requests to (P.L.M.) Primary Care Center, Yale-New Haven Hospital, 789 Howard Ave, New Haven, CT 06504.

During physical examination the young child with fever and a serious illness may not manifest classic findings suggestive of that illness, eg, the child <16 months old with bacterial meningitis may not have meningeal signs.[1] Consequently, the pediatrician relies on clues gained prior to physical examination that might be indicators of the presence of a serious illness. Several recent studies[2,3] have recognized the importance of these clues (sometimes referred to as "toxicity") in identifying young children with serious illnesses but these clues have not been characterized precisely.

In an initial attempt to define these clues we asked practicing pediatricians to list key history and observation items on which they rely to judge the degree of illness of a febrile child prior to physical examination.[4] The five history and eight observation items they listed were vague, eg, playfulness, alertness. Scales of these items were constructed. Because the points of the scales were poorly defined, observer agreement in scoring the items prior to physical examination on 219 febrile patients was low. Pediatricians were able, however, to identify 57% of children with serious illnesses prior to physical examination and relied more on observation than history items. The specificity of the judgment was 76% and the predictive value 20%.

A subsequent study of 262 febrile children was done to define more precisely the history and observation items that are used to judge degree of illness prior to physical examination.[5] Multiple observers, including attending physicians, pediatric nurses, and pediatric residents simultaneously eval-

uated each child prior to physical examination and listed specific history and observation items on a blank, lined form. Observers scored each item as indicating normality, or mild, moderate, or severe impairment. There were 139 different history items and 186 different observation items noted. All observers relied more on observation than history to judge the degree of illness. Participants observed not only traditional clinical items (state of hydration, respiratory pattern, color) but also, and more frequently, the child's response to stimuli given by the parent or the observer.

In the second study, vs the initial report, pediatricians were able to identify 70% of children with serious illnesses prior to physical examination. The specificity and predictive value of that judgment were 79% and 29%, respectively.

In the present report, we selected observation items associated with serious illnesses in our previous study[5] and then constructed well-defined scales for each item based on data in that study. The purpose of this report was to identify those observation items that could be used to identify, reliably and validly, serious illnesses in children with fever.

MATERIALS AND METHODS

From Nov 1, 1980 to March 1, 1981, consecutive children aged ≤24 months with fever ≥38.3 C (101.0 F) were evaluated. The patients were seen in the Yale-New Haven Hospital Primary Care Center-Emergency Room (PCC) or in one private practice in Milford, CT. In the PCC, the child was observed by one or two attending physicians, a resident, and a nurse prior to history and physical examination and before antipyretics were given. In order to evaluate interobserver reliability, the same two attending physicians (P.L.M., M.R.S.) saw one third of the patients in the PCC. In the private practice the patients were seen by a single observer. Both partners in the practice participated in the study.

We reviewed our previous report[5] to identify: (1) observation data that were used to describe the 18 patients with serious illnesses; and (2) observation data that when used were scored as indicating impairment even though the patient did not have a serious illness; for example, there were seven items that observers used to describe impaired eye appearance and/or function and that were given mean scores of ≥3 (1 = normal to 4 = severe impairment). In total, 63/189 observation items had mean scores of ≥3.0.

This review disclosed that all of the data describing seriously ill children or impairment could be categorized into one of 14 areas: color, hydration, respirations, movement, eye appearance, quality of cry, reaction to parents' stimulation, reaction to observer's stimulation, state variation, response to noise, response to visual stimulation, response to social overtures, reaching or grasping for a presented object, and playing with a presented object. The observation data identified in the review were next used to construct the scale points for these 14 areas. There was sufficient data to define only 2 degrees of impairment, moderate and severe. For example, the scale point for moderate impairment for "reaction to parents' stimulation" was defined by the following observation data: "cries off and on"; the scale point for severe impairment was "continual cry or hardly responds." The choice of data for the moderate or severe scale point was based on the mean score given that data in our previous study (3 = moderate, 4 = severe). As the review of the data up to this point was focused on impairment, it was necessary to review our previous report again in order to define the normal scale point for each of the 14 areas. For example, "cries briefly then stops," or "content and not crying," was the most common way in which a normal (1 = normal) reaction to parents' stimulation was described in that report.

Each area or item was, therefore, initially given a three-point scale. Whether these three-point scales defined sufficiently the manner in which a pediatrician evaluates a child for that item was not certain. That is, were there other data that the pediatrician would wish to include in evaluating a child for "reaction to parents' stimulation" at the normal, or moderate, or severe impairment points? Additionally, were there data that were not included on the scales but that would indicate impairment somewhere between normal and moderate or between moderate and severe? Because of these considerations, a five-point scale of each observation item was constructed. The normal, moderate impairment, and severe impairment points were defined by data as noted above. Space was also given at these points in which an observer could write additional data. Space was also given to write in data between normal-moderate and moderate-severe. Thus the scale for each of the 14 items was a five-point scale (Table 1).

After the scales for these 14 items were constructed, they were reviewed by selected private practitioners and attending pediatricians. The changes suggested were minor.

The 14 items were scored on consecutive febrile children without any communication between observers. Oral consent, which was recommended by the Human Investigation Committee at Yale-New Haven Hospital, was obtained from the parent but otherwise no information was sought from the child's caretaker. Some of the observation items

TABLE 1. Example of Observation Item and Five-Point Scale

Item	Normal 1	2	Moderate 3	4	Severe 5
Reaction to parents stimulation (hold, talk to, give bottle)	Cries briefly then stops OR Content and not crying Other data—	—	Cries off and on Other data—	—	Continual cry OR Hardly responds Other data—

required the observer or the parent to stimulate the child. The stimulation was defined on the data sheet and was delivered by different observers (usually an attending physician) at each encounter. Prior to a health professional's involvement in the study, the data sheets were reviewed with that person and questions were resolved.

Scoring took place with the child seated on the parent's lap. Items that required minimal or no observer interaction with the child were scored first (color, hydration, respirations, movement, and appearance of eyes). Items that required interaction with the child were then scored. As a first step, the parent was asked to comfort the child by holding, talking, or giving a bottle. Every attempt was made to place the child in a state of quiet wakefulness[6] prior to scoring the remaining items requiring stimulation: quality of cry, reaction to parent stimulation, reaction to observer stimulation, state variation, response to visual stimulation, response to noise stimulation, response to social overtures, reaching or grasping for a presented object, and playing with a presented object.

After the observation, the history and physical examination were performed by the resident and laboratory studies ordered at his or her discretion. The attending physician did not communicate with the house officer until the latter had made a decision about laboratory evaluation. The child was then admitted to the hospital or sent home with follow-up by that resident. Results of laboratory studies and the clinical course were reviewed by one of us (P.L.M.) within a week of the visit. Two months after completion of the study, the hospital PCC charts of all patients were reviewed to identify any additional laboratory or follow-up clinical information related to the acute febrile episode. One of us (P.L.M.) also communicated with the house staff of the other pediatric facility in New Haven to make certain that, if patients in the study did go to the other facility during their illness, we were aware of the results of that evaluation.

Serious illnesses were defined in one of several ways: (1) bacterial pathogens were isolated on cultures of blood, CSF, urine, stool, joint fluid, or deep soft tissue aspirates; (2) abnormalities of electrolytes (hypernatremia, acidosis), chest roentgenograms (infiltrates) blood gases (hypoxia in bronchiolitis), or CSF (pleocytosis) were documented.

Three patients who did not meet the above criteria but who, because of bronchiolitis, required prolonged hospitalization, intravenous hydration, and pulmonary toilet, were considered seriously ill.

The correlation between the independent (14 observation items) variables and the dependent variables (presence or absence of serious illness) was examined using the pearsonian r. Inasmuch as the independent variable was ordinal and the dependent variable existential, τ-C was used to corroborate the relation between independent and dependent variables indicated by the pearsonian r.[7]

The observation item with the highest pearsonian r was used as a first step in a multiple regression analysis. The multiple R value and the multiple R^2 were determined for each step in the analysis. The regression was terminated when the inclusion of the next observation item added <1% to the multiple R^2.

Three items, reaction to observer stimulation, response to noise stimulation, and reaching or grasping for presented object were deleted in the multiple regression analysis because of high intercorrelation with the observation items reaction to parents' stimulation, response to visual stimulation, and playing with presented object (intercorrelation r = .81, .67 and .61, respectively). The latter three items, with higher pearsonian r correlations to serious illnesses and/or higher interrater reliability than the former three items, were retained in the multiple regression analysis.

Once the key observation items were identified and appropriate weights assigned to each item from the regression analysis, then the specificity, sensitivity, and predictive value for serious illnesses of these key items functioning together (the predictive model) were studied. Because there were two groups in the dependent variable, ie, those who were and those who were not seriously ill, the method of selection of and the weights assigned to key observation items were identical for regression analysis and discriminant function analysis.

The interobserver reliability of scoring the observation items was studied by using weighted kappa (κw), the statistic of choice when ordinal scales are

used.[8] Complete agreement was given a weight of 1 and complete disagreement a weight of 0; partial agreement weights of .67 and .33 were also assigned. Interobserver reliability was studied between the same two attending physicians (P.L.M., M.R.S.) who together saw one third of the patients in the PCC. A clinically significant level of agreement (κw) was defined as .50 or greater which follows standards suggested by statisticians in this field.[9] An excellent description of kappa in the medical literature is given by Koran.[10]

RESULTS

During a four-month period, Nov 1, 1980 to March 1, 1981, 312 consecutive febrile children seen in the PCC and in one private practice were enrolled in the study. Of 206 children seen in the PCC 33 had serious illnesses; 106 children, four of whom had serious illnesses, were seen in private practice.

A total of 557 observers saw the 206 patients in the PCC. At least one attending physician saw 193 of the patients. At least two attending physicians saw 113 patients. Pediatric house officers saw 152 children and pediatric nurses saw 84 patients.

The following aspects of performance by attending physicians in the PCC were studied: the use of defined and nondefined scale points, the relation of observation item scoring to the presence of serious illnesses, and interobserver agreement. Performance of attending physicians was studied because these physicians represented the most experienced observers in this study. Previous reports have noted that their judgments are more reliable and valid than the judgments of less experienced observers.[10] Attending physicians in the PCC used the defined scale points 96% of the time and nondefined points only 4% of the time. Consequently, in performing regression analysis, the only patients included were those who were seen by at least one attending physician in the PCC and for whom all observation items were scored by the attending physicians using defined scale points. These criteria eliminated 41 patients seen in the PCC, seven of whom had serious illnesses. Of the remaining 165 PCC patients, 26 had serious illnesses. The diagnoses in these 26 patients are shown in Table 2. Laboratory examinations were performed on these 165 patients as follows: 118 had a WBC count, 79 an ESR, 114 a blood culture, 72 a chest radiogram, 44 a lumbar puncture, 26 a urine culture, and 10 had serum electrolytes drawn.

The Pearson r correlation between scoring of attending physicians for the observation items and outcome (presence or absence of serious illness) was studied in these 165 patients. If two attending physicians saw the patient, the scores of the attending physician who had seen the most patients in the study were used. The item with the highest r correlation to outcome was quality of cry ($r = .494$) and this was used as the first step in the multiple regression analysis. The multiple regression analysis is shown in Table 3. There were six observation items that added 1% or more to the multiple R^2. If all 11 items (excluding the three highly intercorrelated items) were used in the regression analysis, the multiple R^2 was almost the same (42.2%) as when six items were used. The relation between scoring for observation items and the presence or absence of serious illness indicated by the pearsonian r was corroborated by τ-C.

A six-item predictive model, consisting of those key observation items identified in regression analysis and the weights given to those items in the analysis, was studied as a predictor of serious illnesses. The specificity, sensitivity, and positive predictive value of the model for serious illnesses were 88%, 77% and 56%, respectively. If a child was not impaired on these six items, then the probability of his having a serious illness was 4.7%. Height of fever was also entered into the regression analysis but did not add to the specificity, sensitivity, and pre-

TABLE 2. Diagnoses in 26 Children with Serious Illnesses Seen in Primary Care Center*

Diagnosis	No. of Children	Abnormal Test
Bacterial meningitis	2	CSF culture
Aseptic meningitis	1	CSF pleocytosis
Bacteremia	2	Blood culture
Pneumonia	7	Chest roentgenogram
Urinary tract infection	2	Urine culture
Septic arthritis	1	Joint fluid culture
Cellulitis/abscess	3	Deep soft tissue culture
Bronchiolitis/hypoxia	4	Blood gases
Bronchiolitis*	3	
Dehydration	1	Serum electrolytes
Total	26	

* See text for details.

TABLE 3. Stepwise Multiple Regression Analysis to Identify Observation Items Predictive of Serious Illness*

Observation Item	Multiple R Value	Multiple R^2 (%)	R^2 Change
Quality of cry	.494	24.4	
Reaction to parents' stimulation	.549	30.1	.057
State variation	.587	34.4	.043
Color	.609	37.1	.027
Hydration	.622	38.7	.016
Response to social overtures	.630	39.7	.010

* Based on 165 patients seen by at least one attending physician in primary care center.

dictive value of the six-item model. An 11-item predictive model, consisting of all 11 observation items used in the regression analysis and the weights given to those items in the analysis, was also studied as a predictor of serious illness. The specificity (90%), sensitivity (65%), and predictive value (55%) were not improved over the six-item model.

The reliability of using the defined scale points for the 11 observation items included in the regression analysis was studied in the 68 children seen by the same two attending physicians in the PCC (P.L.M., M.R.S.). The results are shown in Table 4. The observed agreement between attending physicians in scoring all items was ≥80%. Agreement between attending physicians in scoring 5/6 items in the predictive model was in the clinically significant range. The κw for hydration, which was part of our predictive model, was 0.10. The observed agreement for scoring hydration was high (88%) but the chance expected agreement was also high (87%). The latter was due to the high use of the normal category (90%); that is, nearly all 68 children whose values were used to compute κw showed no impairment of hydration and the population allowed little use of the other scale points for hydration. κw for scoring hydration was, therefore, studied in 41 PCC patients seen by two attending physicians other than the P.L.M.-M.R.S. pairing. In these patients, κw for hydration was .55 and observed agreement was 93%.

The six items identified in this study as part of our predictive model and the three scale points for each item are shown in Table 5. A patient score was derived by summing the scores of the individual items. Of 165 patients 110 had a score of ≤10 and three of these (2.7%) had a serious illness; 13 patients had a score of ≥16 and 12 of these (92.3%) had a serious illness; and there were 42 patients with an intermediate score of 11 to 15 and 11 of these (26.2%) were seriously ill.

The performance of each of the 11 individual observation items as predictors of serious illnesses was also studied and compared with the perform-

TABLE 4. Agreement Data for 11 Observation Items Scored in 68 Children Seen by Same Two Attending Physicians in Primary Care Center

Observation Item	κw	Observed Agreement (%)	Change Expected Agreement (%)
Playing with object	.85	95	67
Movement	.79	94	72
Reaction to parent stimulation	.73*	92	69
Response to social overtures	.73*	90	64
Respirations	.58	82	56
Quality of cry	.56*	89	74
Color	.55*	97	93
Appearance of eyes	.50	80	59
State variation	.47*	95	91
Response to visual stimulation	.37	91	85
Hydration	.10†	88	87

* Item included in predictive model, $P \leq .0001$.
† Item included in predictive model, $P \leq .05$.

TABLE 5. Predictive Model: Six Observation Items and Their Scales

	Observation Item	1 Normal	3 Moderate Impairment	5 Severe Impairment
1	Quality of cry	Strong with normal tone OR Content and not crying	Whimpering OR Sobbing	Weak OR Moaning OR High pitched
2	Reaction to parent stimulation	Cries briefly then stops OR Content and not crying	Cries off and on	Continual cry OR Hardly responds
	State variation	If awake → stays awake OR If asleep and stimulated → wakes up quickly	Eyes close briefly → awake OR Awakes with prolonged stimulation	Falls to sleep OR Will not rouse
3	Color	Pink	Pale extremities OR Acrocyanosis	Pale OR Cyanotic OR Mottled OR Ashen
4	Hydration	Skin normal, eyes normal AND Mucous membranes moist	Skin, eyes-normal AND Mouth slightly dry	Skin doughy OR Tented AND Dry mucous membranes AND/OR Sunken eyes
5	Response (talk, smile) to social overtures	Smiles OR Alerts (≤2 mo)	Brief smile OR Alerts briefly (≤2 mo)	No smile Face anxious, dull, expressionless OR No alerting (≤2 mo)

ance of the six-item model based on regression analysis. None of the individual items performed as well as the predictive model. For example moderate or severe impairment scores for "appearance of eyes" had a specificity, sensitivity, and positive predictive value of 50%, 85%, and 24%, respectively. Moderate or severe impairment on "response to social overtures" had a specificity, sensitivity, and positive predictive value of 23%, 100%, and 20%, respectively.

In order to validate our six-item predictive model, the following steps were taken:

1. The performance of the six-item model was studied in the 106 patients seen in private practice, only four of whom had serious illnesses. Because use of nondefined scale points eliminated two of these four seriously ill children, no meaningful application of the predictive model could be made to the private practice patients.

2. The original sample of 165 patients was divided into two groups (A and B) by use of a random-number table. On the basis of whether the first three digits produced an odd or even number, a subject was assigned to group A (N = 77) or group B (N = 88), respectively. There were 12 patients with serious illnesses in group A and 14 in group B. Next, a nonstep discriminant/regression analysis was performed on the six items, separately, for each group.[11] This produced multiple R values for groups A and B, respectively, of .78 and .58 as compared with .63 for the original sample of 165 patients. Next, the discriminant rule derived from group A was applied to each subject in group B; conversely, the discriminant rule derived from group B was applied to each subject in group A.[11] The resulting specificity, sensitivity, and positive predictive value were 83%, 83%, and 48%, respectively for group A and 88%, 64%, and 50%, respectively, for group B. These compare with the aforementioned corresponding values of 88%, 77%, and 56% for the full sample (N = 165).

DISCUSSION

The ability of the pediatrician to gain a sense of the well-being of the child by observing the child prior to the physical examination is critical in the evaluation of febrile children.[1-3] The young child with fever, even though more susceptible to diseases such as meningitis, bacteremia, and urinary tract infection, may not have the classic clinical findings (eg, meningismus, petechiae) suggestive of those diagnoses. Consequently, a series of studies was undertaken to define the data on which the judgment of degree of illness made prior to physical examination is based. Initially both history and observation data were studied in relation to this judgment; the studies revealed, however, that observation data were better correlated with this judgment.[4,5]

The present report, utilizing data from our previous studies and focusing on observation of febrile children, has identified six observation items with defined three-point scales for each item which predicted, validly and reliably, serious illness in our population of febrile children. The level of agreement (reliability) was far greater than seen in our previous report in which agreement for none of the 13 history or observation items reached clinical significance.[4] Definition of scale points and utilization of specific data, rather than vague terms, might explain improved agreement. In the previous study, one attending physician and one resident saw each patient. Using the same two experienced observers (attending physicians) could of itself improve agreement.[10] The specificity and sensitivity of the predictive model based on the six observation items was greater than the specificity and sensitivity noted in previous reports. This improved performance may be related to a more specific definition of scale points for the six items. This definition may have allowed observers to focus their attention on critical data. Additionally, the predictive value of the present model is far greater (56% vs 28%) than in previous reports. Predictive value is influenced by prevalence of disease[12] and, in the first two studies, the prevalence of serious illnesses (9.1% and 6.1%) was less than the prevalence of serious illnesses in the present study (26/165 patients used for the predictive model or 15.8%). Therefore, the considerable improvement in predictive value in this report vs previous reports may be related to differences in the populations studied.

The observation data identified as valid and reliable in the present report have much in common with the data identified in previous studies. In our first study[4] a panel of eight pediatricians recognized a priori that observation of hydration, color, and consolability (similar to reaction to parents' stimulation in the present study) were key data in the assessment of febrile children. Additionally, three and perhaps four of our six observation items relate to the child's response to stimuli (reaction to parents' stimulation, response to social overtures, state variation, quality of cry). In our second study, the manner in which the child responded to stimuli was critical observation data in judging the degree of illness of febrile children.[5] Nelson[13] defined a play score for acute pediatric illnesses, not necessarily febrile, and defined color as a critical predictive item. In addition, her "activity" scale was similar to our state variation and focused on level of consciousness. Scale points for our items and those of Nelson differ however. No reliability data were

presented in her study and the sample from which the predictive items were obtained were not exclusively febrile children. There are, therefore, common themes as well as important differences in the present study and previous reports.

When a predictive model is condensed from a greater number of items, then it is necessary to test that model on a different population (transcription)[14,15] to ascertain whether the model is valid when applied to other patients. The performance of the predictive model could not be studied in the group of private practice patients because of the small number of ill children seen in that setting. The original 165 children seen in PCC (from whom the six-item model was derived) were therefore divided into two groups. The six-item predictive model, applied to each randomly formed subgroup, performed almost as well as it did on the complete sample of 165 subjects. However, given the problems associated with this and other methods (eg, jackknifing), which are applied to cross-validating within a single population, ie, successful replication caused by correlated subsamples,[11] we realize that a more definitive test of the validity of this predictive model will depend upon the results of its application to an independent sample of febrile children.

The sensitivity, specificity, and predictive value of the individual observation items were studied. None performed as well as the six-item predictive model. The use of scales for "response (talk, smile) to social overtures" gave interesting results, however. Impairment of this response was seen in 100% of febrile children with serious illnesses (sensitivity 100%). No child who smiled normally had a serious illness (negative predictive accuracy 100%). However, fewer than 1/5 children who did not smile normally had a serious illness. Thus, impairment of smiling is commonly seen in febrile children who are not seriously ill.

The data in this study and previous studies indicate that the clinical evaluation of febrile children is a complex series of steps based on appreciation of risk factors (age, temperature) and then on observation of the child, history, and physical examination. This series of steps represents a rich data base that allows the pediatrician to separate those children requiring laboratory evaluation from those who have minor illnesses. Appreciation of age and temperature risk factors alone may indicate the need for laboratory evaluation. Examples of such patients would be the 1-month-old infant with fever (age as a risk factor) or the 18-month-old child with 41 C (105.8 F) fever (temperature as a risk factor).[16–18] Observation of the child who has no age or temperature risk factors may indicate no impairment of the observation items mentioned in our model. The history on this same child, however, may indicate bloody diarrhea or crying when urinating. These data outweigh benign observation data and lack of age and temperature risk factors and indicate the need for laboratory evaluation. Or, there may be no signs of significant impairment by observation or history and age or temperature risk factors may be absent; physical examination may, however, detect the presence of pulmonary rales. This finding would outweigh the previously benign data and indicate the need for further evaluation. It is by balancing the varied sources of information that the pediatrician arrives at a plan. Certainly, if the child appears well by observation, has no clues indicating serious illness by history and physical examination and no age or temperature risk factors, then the risk of serious illness is low. In our study, 20/26 seriously ill patients were identified by observation. Of the remaining six patients, four had findings by physical examination suggesting the presence of a serious illness: one child with periorbital cellulitis had a swollen red eye and three children with pneumonia or bronchiolitis had signs of respiratory distress. The sensitivity of observation, history, and physical examination for serious illnesses was thus 92% (24/26 patients). Therefore, all pertinent data must be gathered at each step of the evaluation in order to identify febrile children with serious illnesses. The predictive model is of value in focusing the pediatrician's attention on critical data to be obtained by observation.

IMPLICATIONS AND SPECULATION

Each step in the clinical evaluation of febrile children (observation, history, and physical examination) potentially may generate a hypothesis about whether the child is seriously ill or well. Observation is, most often, the initial hypothesis-generating maneuver. If a child acts normally, then a hypothesis of wellness is generated and, unless specific findings arise during history and physical examination, the hypothesis is accepted. Two of three children in our study (110/165) appeared well. Consequently, the hypothesis of well child by observation is the most frequent judgment made by pediatricians about febrile children. Only 2.7% of these well-appearing children had a serious illness. If a child appears ill, then the hypothesis is generated that serious illness is present and history and physical examination data are sought to buttress this hypothesis. Although children who appear seriously ill are a less common clinical problem than the well-appearing child, the hypothesis generated that they are seriously ill is highly accurate. Of the 13 children who appeared most ill in our study, 12 had a serious illness. Perhaps the greatest challenge

to the pediatrician is the approximately one in four children (42/165 in our study) who by observation, occupy a middle ground. From our data, approximately 25% of these febrile children will have a serious illness. Inasmuch as observation generates an equivocal hypothesis about wellness (or illness), hypothesis-generating clues must be sought in the history, physical examination, and, frequently, screening laboratory information.

Observation, as the initial hypothesis-generating maneuver in evaluating febrile children, establishes the prior probability of disease and allows the pediatrician to interpret further clinical and laboratory data in light of that probability.[19] As such, it is a critical part of the diagnostic process.

ACKNOWLEDGMENTS

This work was supported in part by a General Pediatric Academic Development Award from the Robert Wood Johnson Foundation.

We thank Howard Pearson, MD, and James Jekel, MD, for their support; William Kessen, PhD, for his suggestions about the formation of scales; Lorraine Gambino for her statistical analyses of the data; the pediatric housestaff for their cooperation; Lucille Tomasso, RN, and NoElla Appi, RN, for their participation; and Barbara Serphillips for typing the manuscript.

REFERENCES

1. McCarthy PL: What tests are indicated for the child less than two with fever. *Pediatr Rev* 1:51, 1979
2. Wright PF, Thompson J, McKee KT Jr, et al: Patterns of illness in the highly febrile young child: Epidemiologic, clinical, and laboratory correlates. *Pediatrics* 67:694, 1981
3. Waskerwitz S, Berkelhamer JE: Outpatient bacteremia: Clinical findings in children under two with initial temperature of 39.5°C. *J Pediatr* 99:231, 1981
4. McCarthy PL, Jekel JF, Stashwick CA, et al: History and observation variables in assessing febrile children. *Pediatrics* 65:1090, 1980
5. McCarthy PL, Jekel JF, Stashwick CA, et al: Further definition of history and observation variables in assessing febrile children. *Pediatrics* 67:687, 1981
6. Beintema DJ: A neurological study of newborn infants. *Clin Dev Med* 28:27, 1968
7. Kendall MG: *Rank Correlation Methods.* New York, Hafner Press, 1962
8. Cicchetti DV: Assessing inter-rater reliability for rating scales: Resolving some basic issues. *Br J Psychiatry* 129:452, 1976
9. Cicchetti DV, Sparrow SS: Developing criteria for establishing inter-rater reliability of specific items: Applications to assessment of adaptive behavior. *Am J Ment Retardation* 86:127, 1981
10. Koran LM: The reliability of clinical methods, data and judgments (parts I and II). *N Engl J Med* 293:642, 695, 1975
11. Peduzzi PN, Detre KM, Chan YK, et al: Validation of a risk function to predict mortality in a VA population with coronary artery disease. *Controlled Clin Trials*, in press 1982
12. Jekel JF, Greenberg RA, Drake BM: Influence of prevalence of infection on tuberculin skin testing program. *Pub Health Rep* 84:883, 1969
13. Nelson KE: An index of severity for acute pediatric illness. *Am J Public Health* 70:804, 1980
14. McNeil BJ, Hanley JA: Statistical approaches to clinical predictions. *N Engl J Med* 304:1292, 1981
15. Dolgin SM, Schwartz JS, Kressel HY, et al: Identification of patients with cholesterol or pigment gallstones by discriminate analysis of radiographic features. *N Engl J Med* 304:808, 1981
16. Robert K, Borzy M: Fever in the first eight weeks of life. *John Hopkins Med J* 141:9, 1977
17. McGowan JE, Bratton L, Klein JO, et al. Bacteremia in febrile children seen in a "walk-in" pediatric clinic. *N Engl J Med* 288:1309, 1973
18. McCarthy PL, Dolan TF: Hyperpyrexia in children: Eight year emergency room experience. *Am J Dis Child* 130:849, 1976
19. Khan MA, Khan MA: Diagnostic value of HLA-B27 testing in ankylosing spondylitis and Reiter's syndrome. *Ann Intern Med* 96:70, 1982

Observation Scales to Identify Serious Illness in Febrile Children

Paul L. McCarthy, Michael R. Sharpe, Sydney Z. Spiesel, Thomas F. Dolan, Brian W. Forsyth, Thomas G. DeWitt, Howard D. Fink, Michael A. Baron and Domenic V. Cicchetti

Pediatrics 1982;70;802-809