

Accuracy of Clinical Judgment in Neonatal Jaundice

Virginia A. Moyer, MD, MPH; Chul Ahn, PhD; Stephanie Sneed, BS

Context: Recommendations for management of jaundice in newborns presume that jaundice is a reliable clinical finding and that the pattern and intensity of jaundice reflects the degree of elevation of the serum bilirubin level.

Objectives: To determine whether experienced observers agree in describing the extent of jaundice and to evaluate the reliability of visual assessment as an indication for the measurement of serum bilirubin levels.

Design: Comparison of independent judgments of the extent of jaundice between examiners and with actual serum bilirubin measurements.

Setting: Well-newborn nursery in an urban public hospital.

Participants: A convenience sample of 122 healthy term newborns whose bilirubin concentration was measured

in the course of standard newborn care. Observers were experienced pediatric nurse practitioners, pediatric house staff, and pediatric attending physicians.

Results: Agreement was moderately good for whether an infant's skin was darkly pigmented ($\kappa = 0.56$). However, agreement between observers regarding the presence of jaundice at each specific body site was poor (0%-23% agreement beyond chance); correlation between estimated bilirubin concentrations was similarly poor (Pearson correlation coefficient, 0.37). Correlation between estimated and actual bilirubin values was slightly better (Pearson correlation coefficient, 0.43-0.54).

Conclusions: Clinical examination with visual assessment for jaundice in newborns is neither reliable nor accurate. The decision to perform serum bilirubin testing should be based on additional factors.

Arch Pediatr Adolesc Med. 2000;154:391-394

Editor's Note: The take-home message for me is that no jaundice below the nipple line equals no bilirubin test, unless there's some other indication.

Catherine D. DeAngelis, MD

From the Departments of Pediatrics (Dr Moyer) and Internal Medicine (Drs Moyer and Ahn), University of Texas–Houston Health Science Center. Ms Sneed is a student at the University of Texas–Houston.

CLINICIANS ASSUME that jaundice is a reliable clinical finding among examiners and that its pattern and intensity in newborns reflect the degree of elevation of the serum bilirubin concentration. Decisions regarding the need for bilirubin testing in newborns are generally based on these assumptions, but confirmatory data are limited. Davidson et al,¹ in 1941, noted wide variability among infants in the correlation between visible jaundice and serum bilirubin level. The cephalocaudal progression of dermal icterus has long been noted,² but it was not until 1969 that Kra-

mer³ examined the correlation of clinical jaundice with serum bilirubin measurement. In that study, a single observer noted the presence or absence of jaundice in each of 5 "dermal zones" (progressing cephalocaudally) and found a correlation between the two, but with a wide range of bilirubin concentrations for jaundice in each of the dermal zones. Ebbesen,⁴ examining infants in daylight rather than in fluorescent light, found a similar correlation between the extent of dermal icterus and the serum bilirubin value but again noted a wide range of bilirubin values for each dermal zone. He found that no infant whose jaundice did not progress below the knees (his region 3) had a bilirubin concentration greater than 188 $\mu\text{mol/L}$ ($>11 \text{ mg/dL}$). Madlon-Kay⁵ evaluated the accuracy and agreement of physicians, nurses, and parents in detecting jaundice in newborns by asking each observer to make a single subjective judgment of whether an infant was or was not jaun-

PARTICIPANTS AND METHODS

This observational study took place in the well-newborn nursery of an urban public hospital and used a convenience sample of generally healthy infants who had serum bilirubin concentration measured for any reason. Bilirubin was measured if the infant appeared icteric or if the mother was Rh-negative or had a positive maternal Coombs test result (whether the infant was clinically jaundiced or not). To obviate observer bias, infants with previous bilirubin determination and infants receiving phototherapy were excluded. Observers consisted of experienced pediatric residents, pediatric nurse practitioners, and pediatric attending physicians, some of whom were involved in the clinical care of the infants studied. For each infant, the observer recorded his or her status as resident, nurse practitioner, or attending physician, but observers were not individually identified.

At the time of an infant's initial bilirubin determination, 2 observers independently recorded their assessments. Infants were observed under bright indoor fluorescent lighting augmented by natural daylight from large windows in the nursery. Assessments were recorded for pre-specified parts of the body that reflected cephalocaudal progression and additional specific sites that were suggested by experienced pediatric faculty as being useful in assessing jaundice, such as the conjunctiva, the tip of the nose, and the palate. For each site, jaundice was subjectively

assessed as absent, slight, or obvious. To simulate usual conditions, observers were not given any further training regarding how to assess jaundice. Each observer also subjectively classified the infant's skin tone as "light" or "dark" to evaluate the effect of skin tone on the clinical judgment of jaundice in our ethnically diverse nursery population. Finally, each observer made a prediction (estimation) of the total serum bilirubin concentration based on the infant's clinical appearance. A serum bilirubin test was performed within 1 hour of the clinical assessments for each study patient. Clinical assessments were kept in sealed envelopes until after the bilirubin value was returned from the laboratory.

Analysis was done by weighted κ statistic for agreement between observers. The κ statistic measures the amount of agreement that exists between observers beyond that expected by chance alone. When the data measure something with more than 2 categories, such as progression and intensity of jaundice, then partial agreement can be measured using the weighted κ statistic, which gives some credit to observations that are "close" but not in perfect agreement. Weighted κ values were calculated for each part of the body. Pearson correlation coefficients were calculated for the bilirubin level predictions made by the 2 examiners and for the predictions and the actual bilirubin concentration. Statistical significance was set at $P < .05$.

This study was approved by the Committee for the Protection of Human Subjects of the University of Texas–Houston Health Science Center.

diced. She found a simple κ value of about 0.48 for this unitary judgment; the adjusted correlation coefficient for the physician estimates of the bilirubin value was 0.55. No study that we found examined interobserver variability in the visual assessment of the extent of jaundice in newborns.

The purpose of this study was to reevaluate the importance of clinical observation in the management of neonatal icterus. The specific objectives were to determine whether experienced observers agree in describing the extent of jaundice and to evaluate the reliability of visual assessment as an indication for the measurement serum bilirubin values.

RESULTS

A total of 122 healthy infants (66 boys) underwent serum bilirubin concentration measurements and examinations by 2 observers. Mean age of the infants was 2 days (range, 8 hours to 7 days); all weighed more than 2000 g and were older than 36 weeks' gestational age. Blood type and Coombs test results were unknown for most infants. Fifty-two percent of the observations were made by pediatric nurse practitioners, 45% were made by pediatric house officers, and 3% were made by attending pediatricians. Bilirubin values ranged from 7 to 286 $\mu\text{mol/L}$ (0.4–16.7 mg/dL) (mean, 178 $\mu\text{mol/L}$ [10.4 mg/dL]).

Although agreement was good for whether the infant's skin tone was light or dark (simple $\kappa = 0.56$), the weighted κ value for agreement between observers for jaundice at each level was at best only marginally greater

than chance alone (**Table**). For facial jaundice, for example, the weighted κ value was 0.16 (95% confidence interval, –0.02 to 0.34), indicating that agreement was only 16% greater than would have been expected by chance alone, a value that was not statistically different from 0. For jaundice extending down the trunk, the weighted κ value was 0.23 (95% confidence interval, 0.09 to 0.38), or 23% greater than chance alone would predict.

Despite an ethnically diverse population, 81% of our infants were judged by both observers not to be darkly pigmented. Because of the small number of darkly pigmented infants, we were not able to stratify infants by whether they were darkly pigmented. Infants were stratified into those with bilirubin concentrations of 205 $\mu\text{mol/L}$ or greater (≥ 12.0 mg/dL) and those with bilirubin concentrations below 205 $\mu\text{mol/L}$ (< 12.0 mg/dL) to determine whether the presence of higher levels of serum bilirubin altered the reliability of observations. In infants with higher bilirubin levels (≥ 205 $\mu\text{mol/L}$ [≥ 12.0 mg/dL]), weighted κ values ranged from –0.24 to 0.16, whereas in infants with lower bilirubin levels (< 205 $\mu\text{mol/L}$ [< 12.0 mg/dL]), weighted κ values ranged from 0.05 to 0.21, differences that were not statistically significant.

Correlation between the 2 observers' predictions (estimations) of bilirubin values (Pearson correlation coefficient, 0.37) (**Figure 1**) was not as high as was correlation between the estimated bilirubin and the serum bilirubin levels (Pearson correlation coefficients, 0.43 and 0.54 for the 2 groups of observers) (**Figure 2**).

Agreement Between Observers for Jaundice in Infants

Site	Weighted κ	95% Confidence Interval
Face and neck	0.16	-0.02 to 0.34
Neck to nipple line	0.15	0.01 to 0.29
Nipple line to umbilicus	0.23	0.09 to 0.38
Umbilicus to groin	0.19	0.05 to 0.34
Upper legs	0.20	0.06 to 0.35
Lower legs	0.16	-0.01 to 0.34
Soles	-0.06	-0.20 to 0.08
Arms	0.02	-0.14 to 0.18
Palms	0.05	-0.16 to 0.25
Conjunctiva	0.10	-0.04 to 0.23
Tip of nose	0.08	-0.07 to 0.22
Palate	0.11	-0.03 to 0.25

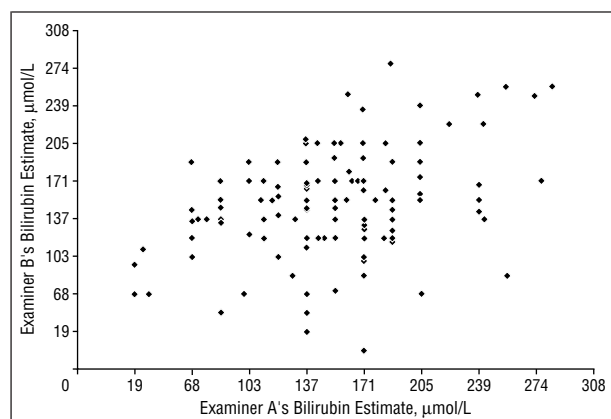


Figure 1. Comparison of estimated bilirubin values. To convert bilirubin from micromoles per liter to milligrams per deciliter, divide micromoles per liter by 17.1.

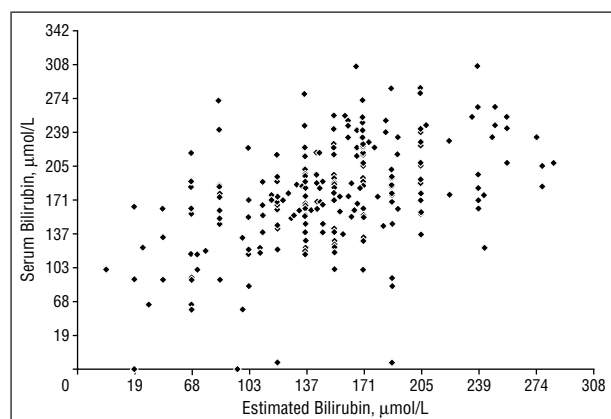


Figure 2. Comparison of estimated and serum bilirubin values. To convert bilirubin from micromoles per liter to milligrams per deciliter, divide micromoles per liter by 17.1.

A serum bilirubin value of 205 $\mu\text{mol/L}$ (12.0 mg/dL) has been used to distinguish physiologic from nonphysiologic jaundice. In our study, the presence of any visible jaundice extending to the lower chest (between the nipple line and the umbilicus) had the best combination of sensitivity and specificity for a bilirubin level of 205 $\mu\text{mol/L}$ (12.0 mg/dL). Nearly all of the infants with bilirubin values greater than 205 $\mu\text{mol/L}$ (>12.0 mg/dL) had jaundice noted at least to the lower chest.

There were 243 observations made of this area (one observer did not record an observation at this level on one infant), 69 in infants with bilirubin levels of 205 $\mu\text{mol/L}$ or greater (≥ 12.0 mg/dL), and 174 in infants with bilirubin levels less than 205 $\mu\text{mol/L}$ (<12.0 mg/dL). Two observers (in 2 different infants) did not note jaundice on the lower chest when the serum bilirubin level was greater than 205 $\mu\text{mol/L}$ (>12.0 mg/dL); the bilirubin level was 217 $\mu\text{mol/L}$ (12.7 mg/dL) in one infant and 279 $\mu\text{mol/L}$ (16.3 mg/dL) in the other. The second observer noted jaundice between the nipple line and the umbilicus for both of these infants. This high level of sensitivity (97%) suggests that infants whose jaundice does not extend below the nipple line are not likely to have bilirubin concentrations of 205 $\mu\text{mol/L}$ or greater (≥ 12.0 mg/dL). However, 81% of infants with bilirubin values less than 205 $\mu\text{mol/L}$ (<12.0 mg/dL) also had jaundice below the middle of the chest, so that this test is useful only in excluding high bilirubin values.

COMMENT

The American Academy of Pediatrics recommendations for management of hyperbilirubinemia presume that clinical examination will be sufficient for identification of infants who need serum bilirubin testing.⁶ To use the test in this fashion, it should be reliable and sensitive, with minimal interobserver variation, to ensure recognition of infants with significant elevation of serum bilirubin concentration. The existing data supporting the accuracy and reliability of clinical examination were not compelling, and reevaluation was therefore undertaken using the criteria of previous studies as well as some additional sites (such as the tip of the nose and the palate) reported anecdotally by experienced pediatricians. Because the ethnic background of the study population was diverse, the potential effect of skin tone on the clinical evaluation of jaundice was also assessed.

Under optimal conditions, each infant would be observed by an experienced clinician more than once over time, at the window (but not in direct sunshine) on a day with bright light. However, the conditions of this study, with variability in weather conditions (and, hence, in the intensity of natural light from the windows) and the use of single observations, reflect the limits of the real-world setting in which clinicians must practice, making choices such as whether to obtain laboratory tests.

Results of this study demonstrate that agreement between experienced examiners regarding the extent of jaundice in otherwise healthy newborns is not much better than would be predicted by chance. Our examiners were largely experienced pediatric nurse practitioners with several years of full-time experience in the well-newborn nursery. Despite this level of experience, clinical agreement was poor, with no "learning curve" evident during the course of the study. This is in contrast to the finding in this study of relatively good agreement between observers regarding whether an infant was or was not darkly pigmented. In addition to cephalocaudal progression of jaundice, we evaluated sites that were suggested by experienced pediatricians to be useful in predicting significant hyperbilirubinemia and found agreement about all

to be poor; for some (such as palms and soles), there was no agreement at all beyond that predicted by chance. The range of bilirubin values found in our patients did not include the very high values that are occasionally seen in infants after discharge from the hospital. The only consistent finding was that infants with no jaundice below the middle of the chest (nipple line) had bilirubin values less than 205 $\mu\text{mol/L}$ ($<12 \text{ mg/dL}$).

Possible explanations for our findings include the subjective nature of determining whether a child's skin color appears yellow, which may be analogous to subjective determinations of pallor or cyanosis. Margolis et al⁷ found the likelihood ratio of cyanosis in predicting oxygen saturation below 95% to be approximately 10, a value that was too low to allow this finding to be used alone to predict hypoxia. In our study, a finding of no jaundice below the nipple line had a likelihood ratio (sensitivity/[1 – specificity]) of about 1.2 for the prediction of bilirubin concentration above 205 $\mu\text{mol/L}$ ($>12.0 \text{ mg/dL}$), reflecting poor ability to predict high bilirubin values. It is not surprising that the extent of jaundice does a poor job of predicting serum bilirubin level because many other factors may affect skin color,⁸ and bilirubin deposition in the skin may vary from one infant to another. Also, perception of color may vary among individuals so that it may be unreasonable to expect 2 individuals to agree.⁹ There did not seem to be systematic variation between observers in assessment of jaundice, and no one observer was better than others in predicting the actual bilirubin levels from the clinical appearance. These results, and those of Davidson et al,¹ suggest that the variability between skin color and bilirubin level is peculiar to each infant rather than just observer dependent. This is supported by data collected by Tayaba et al¹⁰ using spectral reflectance measures: the instrument is more accurate when a baseline measurement of an infant's skin has been made before the onset of jaundice and this value can be factored into later measurements. The bilirubin metering systems that are based on measurement of infant skin color demonstrate high correlation between skin color but require baseline measurements in all infants to correct for underlying skin color.

Our data suggest that within the range of bilirubin values we studied, the relationship between clinical appearance of jaundice and serum bilirubin measurement is not consistent or precise enough to use for prediction of elevated bilirubin levels. The issue of primary importance is whether the clinical examination, with its poor reliability and only moderate correlation with serum bilirubin concentration, can be used as a clinical screening tool to decide when a serum bilirubin test should be ordered. In our study, finding no jaundice below the nipple line reliably predicted that an infant would have a bilirubin concentration below 205 $\mu\text{mol/L}$ ($<12.0 \text{ mg/dL}$),

but the converse was not the case, ie, finding jaundice below the nipple line did not reliably predict which infants would have a bilirubin level of 205 $\mu\text{mol/L}$ or higher ($\geq 12.0 \text{ mg/dL}$). Thus, infants with minimal clinical jaundice are unlikely to have bilirubin concentrations greater than 205 $\mu\text{mol/L}$ ($>12.0 \text{ mg/dL}$), a finding that might be useful in determining which infants certainly do not need serum bilirubin determinations but that will not reliably predict which infants do need bilirubin determinations.

In conclusion, clinical examination for infantile jaundice was not reliable, and prediction of serum bilirubin concentration using clinical examination alone had poor accuracy in our study. Predictions from anecdotally suggested "special sites" were even worse. Sensitivity of clinical observation for high bilirubin levels was higher than specificity, implying that it is more difficult to predict elevated than low bilirubin levels. Rather than basing need for bilirubin testing on clinical observation of jaundice, it should be based on risk factors for severe hyperbilirubinemia and recognition of the extreme rarity of illness caused by hyperbilirubinemia in otherwise healthy infants.

Accepted for publication September 22, 1999.

Presented at the Pediatric Academic Societies Annual Meeting, New Orleans, La, May 1-5, 1998.

Corresponding author: Virginia A. Moyer, MD, MPH, Department of Pediatrics, University of Texas–Houston Health Science Center, 5656 Kelley St, Houston, TX 77026 (e-mail: virginia.moyer@uth.tmc.edu).

REFERENCES

- Davidson LT, Merritt KK, Weech AA. Hyperbilirubinemia in the newborn. *AJDC*. 1941;61:958-980.
- Rolleston H. *Diseases of the Liver, Gall Bladder and Bile Ducts*. Philadelphia, Pa: WB Saunders Co; 1905.
- Kramer L. Advancement of dermal icterus in the jaundiced newborn. *AJDC*. 1969; 118:454-458.
- Ebbesen F. The relationship between the cephalo-pedal progress of clinical icterus and the serum bilirubin concentration in newborn infants without blood type sensitization. *Acta Obstet Gynecol Scand*. 1975;54:329-332.
- Madlon-Kay D. Recognition of the presence and severity of newborn jaundice by parents, nurses, physicians and icterometer. *Pediatrics [serial online]*. 1997; 100:E3.
- Provisional Committee for Quality Improvement and Subcommittee on Hyperbilirubinemia. Practice parameter: management of hyperbilirubinemia in the healthy term newborn. *Pediatrics*. 1994;94:558-565.
- Margolis P, Ferkol T, Marsocci S, et al. Accuracy of the clinical examination in detecting hypoxemia in infants with respiratory illness. *J Pediatr*. 1994;124:552-560.
- Knudsen A. Prediction, measurement and clinical significance of yellow skin colour in mature healthy neonates. *Dan Med Bull*. 1994;41:386-397.
- Hurlbert A. Visual perception: deciphering the colour code. *Nature*. 1991;349: 191-193.
- Tayaba R, Gribetz D, Gribetz I, Holzman I. Noninvasive estimation of serum bilirubin. *Pediatrics [serial online]*. 1998;102:E28.