

education in outpatient fluid therapy should be recognized by providers of child health care.

References

1. Partridge JC, Payne ML, Leisgang JJ, et al: Water intoxication secondary to feeding mismanagement. *Am J Dis Child* 135:38, 1981
2. Tejani A, Dobias B, Mahadevan R: Osmolar relation-

ships in infantile dehydration. *Am J Dis Child* 135:1000, 1981

3. Nickman SL, Buckler JM, Weiner LB: Further experiences with water intoxication. *Pediatrics* 41:149, 1968

4. Crumpacker RW, Kriel RL: Voluntary water intoxication in normal infants. *Neurology* 23:1251, 1973

5. Dugan S, Holliday MA: Water intoxication in two infants following voluntary ingestion of excessive fluids. *Pediatrics* 39:418, 1967

6. Linshaw MA, Hipp T, Gruskin A: Infantile psychogenic water drinking. *J Pediatr* 85:520, 1974

Assessing the Reliability of Data From Patient Medical Records

Daniel C. Cherkin, PhD, William R. Phillips, MD, MPH, and William R. Gillanders, MD
Seattle and White Salmon, Washington

Research performed in the community practice setting and using medical records, although highly appropriate for the study of common clinical conditions, involves a unique set of methodologic problems. The credibility of research conducted in this setting will depend in large part upon how effectively the investigators deal with these problems. One common concern in practice-based research is data reliability.

Reliability is a property of measurement that refers to the replicability or stability of measurement results. A measure is reliable if, when repeatedly applied to an unchanging object, it yields the same result. Investigators often neglect any formal assessment of the reliability of their measures and in effect implicitly assume them to be perfectly reliable. If the reliability of a measure is not assessed, conclusions drawn from the research are subject to criticism.

This communication describes an assessment of the reliability of data abstracted from hospital records used in a study describing the obstetrical

experience of a rural family practice. The focus here is on how reliably data were abstracted from medical records. Other types of reliability, such as that concerning the initial recording of data in the chart, though important, are beyond the scope of this study. Others have discussed problems with the reliability of morbidity data and with encounter-based data collection in more general terms.^{1,2}

Methods

In a study of the obstetrical experience of residency-trained family physicians in rural practice, data were abstracted from the hospital charts of 709 deliveries occurring between 1976 and 1980. The data abstraction form covered 64 items including demographic characteristics of the gravida, pregnancy history, prenatal complications, intrapartum complications and interventions, postpartum complications and interventions, and characteristics and complications of the newborn. All data abstraction was done by a single accredited records technician.

A random sample of 50 of the 709 hospital

From the Department of Family Medicine, School of Medicine, University of Washington, Seattle, Washington. Requests for reprints should be addressed to Dr. Daniel C. Cherkin, Department of Family Medicine, JD-13, School of Medicine, University of Washington, Seattle, WA 98195.

Continued on page 939

Table 1. Reliability Problems in Obstetric Care Variables Abstracted From Hospital Records

Variable	Agreement Rate* %	Mean Value**	Comments
First stage of labor (hours)	60	8.9	82% within 1 hour 94% within 2 hours
Second stage of labor (minutes)	62	45.9	72% within 5 minutes 82% within 15 minutes 96% within 30 minutes
Admission hematocrit (%)	68	36.9	78% within 2 points 12% off by 4 or more points
Admission weight (lb)	82	163.4	98% within 5 pounds
Childbirth education classes (yes, no)	82	—	See text
Weeks of gestation	84	40.2	96% within 1 week 98% within 2 weeks

*Agreement of values obtained at second abstracting with those obtained at original abstracting for sample of 50 hospital records
**Mean values calculated from data on all 709 deliveries

charts was selected and the records technician was asked to re-abstract the data from these records on the same type of forms used for the initial chart review. At the time of the initial abstracting process the records technician did not know a reliability study was planned. Several weeks elapsed between the first and second abstracting reviews. When performing the second abstracting review, the records technician did not have access to the original abstracts. The original and re-abstracted data for the 50 deliveries were then compared for discrepancies. For simplicity, reliability is measured by the percentage agreement between data collected at the first and second data abstracting reviews. More sophisticated measures such as Cohen's kappa, which measures percentage agreement after chance agreement has been removed, may be more appropriate under certain circumstances.³

Results

Of the 3,200 pairs of values compared for reliability (64 variables \times 50 deliveries), 3,071 (96.0 percent) were in perfect agreement. Thirty-three variables (51.6 percent) had no discrepancies whatsoever, and 58 variables (90.1 percent) agreed in over 90 percent of cases.

Six variables, however, were found to be discrepant in over 15 percent of comparisons and merit special attention (Table 1). The poorest agreement was for the measures of duration of the first and second stages of labor. Abstracting discrepancies were noted in more than one third of these cases. Discrepancies in the two recordings of the duration of the first stage of labor exceeded 10 percent of the mean value of this variable in almost one fifth of the cases. Discrepancies in measurements of the second stage of labor exceeded 10 percent of the mean value in more than one fourth of the cases. Prepartum hematocrit level was also found to be relatively unreliable, differing in one third of the cases with a discrepancy of greater than three percentage points (about 10 percent of the mean value) in 12 percent of the cases.

Admission weight, history of childbirth education classes, and weeks of gestation also presented some reliability problems, being discrepant in 16 to 18 percent of comparisons. Measures of admission weight and weeks of gestation, however, rarely differed from their respective mean values by more than 3 percent. Almost all of the discrepancies identified for history of childbirth education classes were attributable to cases being coded

as "no" on the first abstracting review and "unknown" on the second review.

Discussion

The investigators' expectations that these obstetric data had been reliably abstracted from hospital charts were not totally substantiated. Most variables were reliably recorded, but a few problem areas were identified. Two main factors seemed to be responsible for the major reliability problems: (1) difficulties in calculating certain measures from data available in the chart, and (2) the existence in the chart of more than one value for the same variable. The data for length of the first and second stages of labor exemplify the first type of difficulty. The actual duration of these stages was not recorded in the chart and had to be calculated from time notations that could not always be clearly related to the actual beginning and ending of each stage of labor.* The second problem, that of multiple values of the same variable in the chart, was responsible for the discrepancies identified in prepartum hematocrit level. The intent was to abstract the hematocrit recorded at the time of hospital admission, but this information was not always available, and hematocrit measurements from the third trimester were sometimes substituted.

Two other problems may have contributed to the discrepancies identified. First, information regarding whether a woman had taken childbirth education classes was not consistently recorded and, when it was recorded, could be found in several places in the chart. As a result, history of childbirth education classes would sometimes be coded as missing when in fact the information was available elsewhere in the chart. Finally, a small and unquantifiable amount of transcription error undoubtedly occurred. Without going back to the original charts, however, it is impossible to know how many of the discrepancies can be attributed to this problem. It is also not clear from this study whether, in the cases where abstracting discrepancies were identified, the data from the first review were more or less accurate than those abstracted at the second review.

*Since the time of onset of first-stage labor is subjective and reported by the patient, estimates of its duration are particularly prone to reliability problems.

Although the obstetric study data abstraction protocol was successful for most of the study variables, it was inadequate for the length of labor measures and for admission hematocrit levels. In those cases for which the length of labor variables were found to be discrepant, neither the first nor the second recorded value was consistently higher, and thus the estimates of the mean duration were not systematically biased. However, correlations between these and other variables could be obscured as a result of the introduction of what amounts to random error. The admission hematocrit value, on the other hand, was found to be higher on the initial review in 13 out of the 15 cases of discrepancy, and if for some reason the figures from the second review were closer to the true value, the original data would have yielded a biased high estimate of admission hematocrit levels.

This reliability assessment procedure (which increased the total cost of the study by only \$150 in chart abstraction time) allowed investigators to document the reliability of their data and to identify variables that required caution in analysis and discussion. The major lesson learned from this assessment was the importance of complete and precise specification, before data collection begins, of how data are to be abstracted from the patient records. Since unanticipated problems often arise, however, it would also be wise for the investigator to evaluate the reliability of the data recorded in the first 30 or so charts and to provide corrective feedback to the records technician. This evaluation could be accomplished by comparing data abstracted independently by both the records technician and the investigator.

As family medicine research continues to improve in quality, its contributions to knowledge will be more highly valued both within and outside of the discipline. Documenting the reliability of the data is just one of many ways in which the credibility of family practice research can be enhanced.

Acknowledgment

This study was supported by a grant from the American Academy of Family Physicians.

References

1. Anderson JE: Reliability of morbidity data in family practice. *J Fam Pract* 10:677, 1980
2. Candib LM: Stating the obvious: Problems with encounter-based data collection. *Fam Med* 15:59, 1983
3. Cohen J: A coefficient of agreement for nominal scales. *Educ Psychol Meas* 20:37, 1970