Sources of Imprecision in Laboratories Screening for Congenital Hypothyroidism: Analysis of Nine Years of Performance Data

Sandra E. Jewell, William J. Slazzyk, S. Jay Smith, and W. Harry Hannon

We evaluated sources of imprecision in screening assays for congenital hypothyroidism performed between 1979 and 1987 by more than 45 laboratories participating in the national Neonatal Screening Standardization Program offered by the Centers for Disease Control. Reported concentrations were available from approximately 70,000 enriched dried-blood spot samples for thyroxin and thyrotropin, the primary and secondary screening assays, respectively. The most important sources of variation in assayed concentrations are within-laboratory imprecision and the methods or kits used. The importance of each of these sources depends on the analyte and its concentration. In most cases, the greatest source of measurement variation is attributable to within-laboratory imprecision (coefficient of variation 13% to 26%); however, kits are the largest source of variation at the lower concentrations of thyroxin. For both analytes, the greatest relative imprecision occurred at the low concentration, which was within the critical range of the cutoff values for detection of presumptive positive specimens. Imprecision in this range increases the risk of misclassifications. Minimizing within-laboratory imprecision and differences among kits seems to be the best way to improve overall efficiency of screening laboratories.

Additional Keyphrases: neonatal screening · heritable disorders · monitoring assay stability · blood-spot samples · interlaboratory performance · "kit" methods · intralaboratory performance

Congenital hypothyroidism is a metabolic disorder with a nationwide incidence of 1 in 4500 newborns (1). It causes irreversible brain damage if the disease is not diagnosed and the infant treated soon after birth (2, 3). In the U.S., neonatal screening programs for congenital hypothyroidism are mandatory in most states (4).

In 1979, the Centers for Disease Control (CDC) instituted an external quality-assurance program, with co-sponsorship by the Health Resources and Services Administration, designed to assist both the laboratories that are operating screening programs for congenital hypothyroidism and the manufacturers of diagnostic products used in screening. The voluntary program offered each laboratory an opportunity to monitor the long-term stability of its assays for thyroxin (T4) and thyrotropin (thyroid-stimulating hormone, TSH) by providing a six-month supply of each lot of dried blood-spot quality control reference materials, a data report sheet for the first 10 analytical runs, and then providing the statistical analysis of the results derived. These external quality control materials allowed the laboratories to maintain the high-volume screening workload and rapid turnaround of data with a minimum of analytical downtime owing to changes in routine quality control specimens, screening assays, reagents, and personnel. Between 1979 and 1987, more than 45 laboratories—including CDC, state, private, and manufacturer—participated in the program and assayed more than 70,000 reference specimens for T4 and TSH. The laboratories most actively involved in the screening program were state public-health laboratories, and they produced about three-quarters of the data.

Imprecision in the screening process can produce either falsely positive values—which reduce the cost-effectiveness of screening, increase and misdirect the workload for follow-up activities, and produce anxiety for affected parents—or falsely negative values—which result in the failure to detect the condition and have potential legal ramifications. The purpose of our analyses was to determine the quality of laboratory performance and the sources and relative importance of the factors influencing analytical variability in the screening process to better direct efforts for laboratory improvement.

Methods and Materials

Materials

A quality control production lot consisted of specimen sheets of whole-blood spots on filter paper. The blood spots were extensively evaluated for stability and homogeneity (5, 6). Each specimen sheet contained five rows each of three concentrations of T4 or TSH and was numbered with a six-digit series. A random number table was used to select sheets from the total production lot to send to each laboratory. The T4 specimens were divided among 33 production lots at three concentrations—low, medium, and high. For TSH, there were 24 production lots at three target concentrations. Participants received one sheet from the previous lot with each new specimen set, to provide analytical continuity. The blood-spot sheets were packaged for distribution in plastic bags containing desiccant and mailed semianually to participating laboratories. Specimens were to be refrigerated on arrival. These shipments included instructions for analysis of each assay, along with the data-collection forms, which were to be completed and returned to CDC. The laboratories were asked to document the methods used in their assays. After receiving the completed form, CDC statistically analyzed the data, and results were then returned to the participant laboratories.

Statistical Methods

Participating laboratories were assigned an identification number that designated their type—either state, manufacturer, or private. The data for production lots of the same enriched concentration were aggregated into "pools" according to analyte. A preliminary histogram analysis
was done on each pool that identified outlying values. To remove the extreme effects of these few points on estimated statistics, we deleted measurements exceeding the 99th percentile.

The enrichment preparation of the whole-blood pools with analytes for the dried-blood spot materials was performed, with an identical protocol for each new lot, semi-annually over a nine-year period (6). To eliminate the confounding effect that among-lot concentration differences would have introduced, we standardized all data so that individual lot means equaled the overall consensus mean for each analyte and concentration. This adjustment allowed our analysis to focus on laboratory sources of imprecision.

Analysis of variance was used to quantify and categorize variability. Variability was also expressed in terms of coefficients of variation (CVs), percentages of total variance, and standardized differences from the mean value. "Percentage of total variance" refers to the amount of variation attributable to each component factor in the analysis. Within-laboratory imprecision, the term used here, is also known as "replication error." Throughout, we refer to the screening assay methods used by the laboratories as "kits." The kits included in this analysis were made either in-house or by Becton Dickinson Immunodiagnostics (Orangeburg, NY 10962), Diagnostic Products Corp. (Los Angeles, CA 90045), Kallestad Laboratories—only T₄—(Chaska, MN 55318), Meloy Laboratories, Inc. (Springfield, VA 22151), Micromedic System (Horsham, PA 19044), Neometrics, Inc. (E. Northport, NY 11731), Nichols Institute (San Juan Capistrano, CA 92675), or Nuclear Medical Systems—only TSH—(now Biomerica, Newport Beach, CA 92663). In Figure 3, the character symbols denoting kit manufacturers were assigned alphabetically according to the name of the manufacturer. For both analytes, the symbol "C" represents the in-house method.

Standardized deviations were used to compare lot and kit means within a given concentration and to display them in a uniform format. Because the supplements and consensus means varied slightly from lot to lot, absolute values could not be used for this purpose. Standardized deviations were calculated by first determining the overall mean value and standard deviation for the analyte. Then, the overall mean value was subtracted from each value for the analyte and divided by the overall standard deviation. The results expressed the standard deviations of the individual values from a mean value of zero. This is referred to as "standardizing to zero." A standardized deviation of zero would indicate that the value of the individual observation was the same as the overall mean value for the variable analyzed. A standardized deviation value of 1 would show that the individual observation was one standard deviation away from the overall mean value.

Results

We initially considered five factors as possible sources of variation in laboratory measurements: (a) frequency of participation in the program; (b) the type of laboratory performing the assay; (c) within-laboratory imprecision; (d) imprecision among laboratories; and (e) the kits used. To examine whether a laboratory's frequency of participation contributed significantly to its performance, the laboratories were divided into those that sent results annually during the period 1979–1987 and those that did not. We found no statistically significant differences (P > 0.05) attributable to the frequency of laboratory participation and concluded that this factor was not an important discriminating variable in the precision of results.

Similarly, when results were analyzed according to the type of laboratory—state, manufacturer, or private—performing the assay, state laboratories gave the highest CVs at the low and medium concentrations of both T₄ and TSH, but there were no statistically important differences in mean values. We concluded that data from different laboratories could be analyzed as a single factor without categorizing by type.

Analysis of variance results for the remaining factors—kits, among-laboratory imprecision, and within-laboratory imprecision—are reported in Figure 1. A and B. The results show that the relative importance of the sources of variability for reported concentrations of the two analytes was quite different. For example, the overall patterns show that a larger proportion of the variance was attributable to kits and among-laboratory imprecision and a smaller proportion to within-laboratory variability for T₄ than for TSH. This result may be due in part to the smaller mean concentration values for T₄ and the poor discriminating capabilities of some kits at very low concentrations. The greatest source of imprecision for TSH, within-laboratory replication, accounted for more than 50% of the total variance at each concentration. Kits accounted for the highest percentages of variance at the low and medium concentrations of T₄, 51% and 39%, respectively. Within-laboratory replication produced the greatest amount of total variance, 44%, at the high concentration of T₄. The proportion of total T₄ variation ascribable to kits declined.
as mean concentrations increased, from 51% at the lowest concentration to 29% at the highest. For TSH, the percentage of total variance accounted for by kits was relatively stable at 29% to 34% over the three concentrations. The percent of total variance attributable to among-laboratory imprecision was between 22% and 27% for T4 and between 15% and 20% for TSH. The greatest source of variation for both analytes was within-laboratory imprecision, except at the low and medium concentrations of T4, where most of the imprecision was attributable to kits.

The relative variability, indicated by the CVs given in Table 1, shows that within-laboratory replication had the highest CVs, 26% to 16%, among the three concentrations for TSH. Kits used in TSH assays produced CVs that declined from 20% to 12% with increasing concentration. Variation among laboratories was associated with the lowest CVs for TSH, between 15% at the low concentration and 10% at the high. For T4, kits had the highest CVs at the low concentration, 31%, and among-laboratories had the lowest CVs at the medium and high concentrations. At the high concentration of T4, within-laboratory imprecision accounted for the highest CVs.

When mean values for blood-spot production lots were standardized to zero and plotted in chronological order, the within-lot variability showed no marked improvement in precision over time for either analyte (Figure 2, A–F). The confidence limits were widest for the lowest concentrations but narrowed as concentrations increased for each analyte. At each concentration of T4, most of the lot mean values fell slightly below the mean value for the given concentration. This finding indicated that a source of systematic variability unrelated to lots affected the laboratory results. The mean values for TSH materials generally varied above and below the overall mean value at each concentration.

Table 2 shows the consensus mean values from the laboratories, the enriched values, and the assayed mean

<table>
<thead>
<tr>
<th>Analyte</th>
<th>T&lt;sub&gt;4&lt;/sub&gt;</th>
<th>TSH</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concentration</td>
<td>Low</td>
<td>Medium</td>
</tr>
<tr>
<td>Statistic</td>
<td>CV</td>
<td>df</td>
</tr>
<tr>
<td>Kits</td>
<td>31</td>
<td>79</td>
</tr>
<tr>
<td>Among-labs.</td>
<td>20</td>
<td>727</td>
</tr>
<tr>
<td>Within-labs.</td>
<td>23</td>
<td>14104</td>
</tr>
</tbody>
</table>

Fig. 2. Percentage bias from the aggregate mean with 95% confidence limits by lot
Enclosed rectangle, 50% of run means; horizontal line inside rectangle, lot mean; extending lines, minimum and maximum values. Lots are plotted in chronological order.
values from CDC for the blood spots at the three concentrations for T₄ and TSH. Total measured analyte concentrations should reflect the enriched concentration for each pool plus the endogenous concentrations, which by analytical measurement are usually very low: <8 μg/L for T₄ and 2.5 micro-int. units/L for TSH. The consensus value for T₄ was below the expected (enriched-plus-endogenous) value at each concentration by 11% to 21%. For TSH, the consensus value was consistently 17% to 22% below the expected value. The mean assayed values (CDC) were always slightly greater than the expected values for both analytes.

Figure 3, A-D, shows the mean values of the kits expressed as standardized deviations from the overall mean values for the three concentrations for T₄ and TSH. Analytical kits, identified by Codes A through K, had a systematic influence on the results both in terms of mean values and variability. The nearly parallel lines for both mean values and CVs in these figures show that the bias introduced by methods is usually independent of concentration. For example, Kit A yielded a medium-range concentration and a high within-laboratory CV for T₄ at all concentrations, whereas Kit H consistently gave a low mean and a low CV. The kits that performed closest to the consensus mean for T₄ were Kit A at the low and medium concentrations and Kit F at the high concentration. Kit H consistently had the smallest imprecision. For TSH, Kit F produced a low mean and a high CV, and Kit K showed a high mean value, and Kit A a low CV for all concentrations. Kit H gave results for TSH that were close to the consensus mean at all concentrations and, along with Kit A, showed less imprecision than others. Few kit means or CVs fell outside of one standard deviation of the overall mean value and no results were greater than two standard deviations from the mean.

Discussion

The precision of laboratory measurements is an important factor in public-health programs. In neonatal screening, we have found that the primary determinants of precision are within-laboratory variability and the differences among kits. Although our data indicate that a few kits give more-precise results for some concentrations than for others, the performance of most kits remained relatively stable for mean values and CVs at all concentrations for both T₄ and TSH. This consistency indicates that kits must be regarded as an independent source of systematic bias in the screening process.

The consensus mean values for the dried-blood spot materials were lower than the expected values for T₄ and TSH at all concentrations. The values for enriched samples given in Table 2 for the blood-spot pools do not include the endogenous concentrations in these pools. The mean assay values determined at CDC for blood-spot preparations indicate excellent analytical recoveries for the expected values, but they are consistently higher than the consensus values. The World Health Organization's International Reference Preparations (68/38 and 80/558) are used to enrich the blood pools for preparation of the TSH dried-blood spots. When calibrators traceable to this reference material are used, the recoveries obtained at CDC for the distributed blood spots are excellent. The differences between the consensus value and the expected values are wider across the concentrations for TSH than T₄, although an international reference calibration material is available only for TSH. The availability of the TSH reference material makes the bias difficult to explain. In a 1981 study (5) the TSH values for the low- and medium-concentration pools were found to be close to the expected value, but the high-concentration pool was about 10% above. This contrasts with our data and may indicate that there has been a change in the TSH calibration for kits.

An analysis of the ranges and standard deviations of laboratory values submitted over the nine-year period showed that there was no obvious improvement in precision at any concentration for either analyte as a function of time. The fact that the T₄ and TSH results are used for screening instead of diagnostic purposes may have obscured the need to improve analytical precision. For both analytes, the greatest relative imprecision occurs at the low concentration, which is in the critical range of the cutoff values for detection of presumptive positive specimens.

Although we have presented evidence of the degree of imprecision involved in the neonatal screening process, the impact of measurement error is difficult to quantify. Analytical methods did not, up to 1985, contribute directly to a missed case of congenital hypothyroidism (7, 8). Measure-
ment error for the primary screen, T₄, is handled on a daily basis by most screening laboratories, usually by setting floating cutoffs to include many false positives and to virtually preclude false-negative specimens. Therefore, the actual impact of T₄ imprecision is measured less in terms of false negatives than in the cost-effectiveness of each screening program.

Because TSH assays are used with a fixed cutoff, within-laboratory imprecision will have a direct impact on the number of assay results misclassified owing to this second-ary screen. For example, given the TSH average 26% CV at the low concentration and a presumptive positive decision level of 20 milli-int. units/L, a third of assays with a true value of 20 milli-int. units/L but can show TSH values as low as 25 (2). It follows, therefore, that imprecision in a screening laboratory can lead to missed cases, but more typically will mean needless anxiety for parents and that the follow-up staff is unnecessarily burdened with falsely positive values.

One of the 10-year public-health objectives issued in 1980 for the nation (9) stated: "By 1990, virtually all newborns should be screened for metabolic disorders for which effective and efficient tests and treatment are available (e.g., PKU and hypothyroidism)." Given the many sources of imprecision in the screening process, we believe that our data show generally good performance by the screening laboratories and assay kits. Multiple sources of variability notwithstanding, screening results are largely homogeneous; e.g., few results fell outside of one standard deviation of the mean concentration or of the mean CV value, whether the variability was quantified according to kits, among-laboratory or within-laboratory imprecision. Overall the data indicate that the 1990 objective is being met with effective screening tests, although efforts to improve precision will contribute to increased operational efficiency.

This study was supported in part by an interagency agreement with the Health Resources and Services Administration. We thank Linnard Taylor, Theodoria Prater, and Charlie Banks for preparing materials, and Roberta Jensen, Therese Rogers, Barbara Adam, and Dr. James E. Myrick for technical assistance.

References