The Magnitude and Challenge of False-Positive Newborn Screening Test Results

Charles Kwon, MD; Philip M. Farrell, MD, PhD

**Objectives:** This study examined for the first time to our knowledge the national data available from newborn screening programs in the United States and determined the salient characteristics of various screening tests for 3 hereditary metabolic disorders and 2 congenital endocrinopathies with emphasis on positive predictive values (PPVs) to delineate the magnitude of false-positive results.

**Methods:** Reports published by the Council of Regional Networks for Genetic Services for 1990 through 1994 were examined carefully, paying particular attention to phenylketonuria, galactosemia, biotinidase deficiency, congenital hypothyroidism, and congenital adrenal hyperplasia (CAH). Because of recent improvements in data collecting, reporting, and tabulating, we used data from 1993 and 1994 to determine the apparent sensitivity, specificity, relative incidence rates, and PPVs for the 5 disorders. For biotinidase deficiency and CAH, we also calculated relative incidence rates and PPVs for 1991 and 1992.

**Results:** Our analyses revealed the following best estimates for the relative incidence rates of 5 disorders: phenylketonuria, 1:14000; galactosemia, 1:59000; biotinidase deficiency, 1:80000; congenital hypothyroidism, 1:3300; and CAH, 1:20000. An apparent sensitivity of 100% has been reported by the various states for most of the disorders, and specificity levels are all above 99%. The PPVs, however, range from 0.5% to 6.0%. Consequently, on average, there are more than 50 false-positive results for every true-positive result identified through newborn screening in the United States.

**Conclusions:** The magnitude of false-positive results generated in newborn screening programs, particularly for congenital endocrinopathies, presents a great challenge for future improvement of this important public health program. Attention must be given to improved laboratory tests, use of more specific markers, and better risk communication for families of patients with false-positive test results.


**NEWBORN screening for hereditary metabolic diseases and congenital hypothyroidism has been widely accepted as a successful public health measure, aiding in early diagnosis and treatment. Originating in the 1960s with the phenylketonuria (PKU) test developed by Robert Guthrie,1 newborn screening today involves testing for up to 8 different disorders in some states.2 With improving laboratory technologies3 and the pending completion of the Human Genome Project, many new tests can be expected.4 For each screening test, its sensitivity—reflecting the probability that a person with disease will have a positive test result—is of greatest concern. For a screening test to be fully effective, its sensitivity should ideally approach 100%. The reasons for this are quite obvious; to assure that health care professionals and patients can depend on the newborn screening test result, it must reliably detect almost every case of the disease. Maintaining high sensitivity also affects how patients and their parents are counseled before and after test results. It should be noted, however, that owing to the element of human error and the potential for biologic variabilities, no screening test sensitivity can truly achieve 100% over long time periods.5,6 Nevertheless, it is an assumed and expected standard.

On the other hand, in efforts to detect all cases of disease, the occurrence of false-positive test results in newborn screening programs has been almost completely ignored. These misleading outcomes can be expressed statistically either by test specificity, reflecting the probability that a person without disease will have a negative test result, or by positive predictive value (PPV), the proportion of persons with positive test results who actually have the disease. Although
METHODS

This study was possible because of the cooperative efforts of the Council of Regional Networks for Genetic Services (CORN) and the Association of Public Health Laboratories who, under the leadership of Brad Therrell, PhD (Texas Department of Health, Austin), collect information on newborn screening results from all regions of the United States and publish annual tabulated data. The results from each state and/or region are reviewed centrally and published a few years after the annual screening period by CORN, an important organization that sponsors the data collection, tabulation, and publication processes with grant support from the Maternal and Child Health Program (project No. 5MCJ-13106-04-0) of the US Department of Health and Human Services, Washington, DC. The most recent report available when we performed our statistical analyses was for 1994 (published in January 1999). After reviewing all the CORN reports from this past decade, we concluded that the most recent 2 are more complete than those of 1990 through 1992. This conclusion was based on examining data quality, timing and completeness of collecting and reporting results, and the scope of tabulations from 1990 through 1994. The CORN report for 1990 was the only one we rejected completely because of variable reporting periods (ie, some states actually reported 1989 data), numerous approximations, and limited information on some conditions. On the other hand, the CORN reports for 1991 and 1992 were much improved with better definitions of the diseases screened in the US newborn population, fewer approximations, and collecting and reporting that was timely and more complete. Examination of the 1993 and 1994 reports revealed that they were of similar quality and contained the best data collection and reporting that we found available. The 1993 and 1994 CORN reports also feature uniformity of data presentation that facilitated our analyses.

Using as our primary source the data from the National Newborn Screening Reports of 1993 and 1994, we totaled the number of infants screened, the number of positive test results, and the number of confirmed disease cases nationally for 5 diseases: PKU, galactosemia, and biotinidase deficiency as examples of biochemical disorders, and congenital hypothyroidism and congenital adrenal hyperplasia (CAH) as examples of inherited endocrinopathies.

These disorders and sickle cell disease are the congenital abnormalities most commonly screened in the United States. All states currently screen newborns for PKU and congenital hypothyroidism, while 48 states currently include galactosemia; 22, biotinidase deficiency; and 19, CAH in the battery of tests (as of 1998). Hemoglobinopathy screening is currently underway for the entire newborn populations of 39 states, but was excluded from this analysis.

Data available in the CORN reports are presented in tables listing definitions and quantitative information such as total births, “not normal” test results, number of confirmed (diagnosed) cases, and other variables such as laboratory methods, costs, etc. These figures enabled us to calculate apparent sensitivity, specificity, PPV, and relative incidence rate (defined as the number of infants born for each infant diagnosed with disease during a 1-year time period—reflecting the prevalence of these congenital disorders in an annualized newborn population). By using the reported results, we also determined if statistically significant differences occurred from year to year in the calculated PPV values and the 95% confidence intervals (CIs) for the relative incidence of disease. In some cases, the data reported by states were inconsistent or incomplete in the 1993 and 1994 CORN reports. Consequently, we contacted personnel at various state laboratories to clarify their experiences. Because the numbers could not always be reconciled or completed, some values that were obviously erroneous had to be eliminated in our calculations.

When our analyses were completed for 1993 and 1994 newborn screening data, we examined once again the 1991 and 1992 CORN reports to determine if the disease incidence and PPV calculations were consistent with findings earlier in the decade for PKU, galactosemia, and congenital hypothyroidism. In addition, we recalculated screening test outcomes for biotinidase deficiency and CAH using the larger database. Finally, to estimate as accurately as possible the total number of false-positive newborn screening test results per year in the United States, we also examined the 1990 through 1994 CORN reports for information on maple syrup urine disease and homocystinuria. Although roughly half of states screened for these diseases at the beginning of the decade, some have discontinued both tests, and the yield of confirmed cases through screening is minimal (eg, 5 infants with maple syrup urine disease in 1994 and only 1 with homocystinuria).

Recognizing that the magnitude of false-positive test results had not been determined in the United States, we analyzed recent data and calculated relevant test parameters. This article endeavors to show on a national level the surprisingly large ratio of false- to true-positive results in newborn screening tests and to present the best estimates obtained thus far of disease incidence in the United States for 5 disorders included in the panel of tests for many states.

RESULTS

Table 1 gives the results from the analyses of data on 3 biochemical disorders. In each case, some states did not report complete data (eg, the number of tests performed or the number of abnormal test results were omitted).
Their data were eliminated from the calculations of apparent sensitivity, specificity, and PPV, but were included in the calculation of relative disease incidence since the number of confirmed cases was reported. For PKU, data from Alabama, Delaware, Georgia, Maine, Nebraska, Ohio, and Wyoming in 1993 and Florida, Georgia, and Maine in 1994 were removed for the same reason. For galactosemia, the number of abnormal test results in Alabama was strikingly different from 1993 to 1994, decreasing from 21433 to 1053. Laboratory personnel in Alabama could not explain this variance; thus, their data were similarly eliminated. Also, data for Delaware, Georgia, Maine, Montana, and Ohio in 1993 and Alaska and Maine in 1994 were removed from our calculations owing to incomplete data.

Our calculations revealed that the data available suggest that achievement of 100% apparent sensitivity and specificity greater than 99% has occurred for each of 3 hereditary metabolic disorders. Combining 1993 and 1994 data reveals an average relative incidence rate of 1:14200 for PKU and 1:63350 in 1993 and 1994. When biotinidase deficiency screening test results are associated with a much higher proportion of newborns who have true-positive test results compared with PKU and galactosemia. When comparing PPV between 1993 and 1994 for each disease, PKU showed a small statistical difference (P = .04), galactosemia showed no statistical difference (P = .70), but biotinidase deficiency showed significant statistical variance (P < .001). Because of the relatively small numbers for biotinidase deficiency screening, we again analyzed the 1991 and 1992 CORN reports and calculated PPV figures. In 1991 and 1992, biotinidase screening results produced PPVs of 3.7% and 9.1%, respectively. The average for 4 years was 6.4%. We attribute this variability to the relatively low number of abnormal test results and the relatively small number of cases in the United States with biotinidase screening included in their panel (only 13 states in 1991 and 20 in 1994). However, it is clear that despite this variability, the biotinidase test is 2-fold and more than 10-fold better than PKU and galactosemia, respectively, in reducing false-positive test results.

Table 2 gives our results on 2 endocrinopathies. Similar to that mentioned previously, in the case of congenital hypothyroidism, data for Alabama, Delaware, Georgia, Maine, Massachusetts, Montana, Tennessee, and

Table 1. National Newborn Screening Results for Biochemical Disorders

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive test results, No.</td>
<td>9232</td>
<td>9156</td>
<td>9221</td>
<td>10210</td>
<td>308</td>
<td>321</td>
</tr>
<tr>
<td>Confirmed cases, No.</td>
<td>245</td>
<td>269</td>
<td>53</td>
<td>54</td>
<td>21</td>
<td>19</td>
</tr>
<tr>
<td>Apparent sensitivity, %‡</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Specificity, %</td>
<td>99.7</td>
<td>99.8</td>
<td>99.7</td>
<td>99.7</td>
<td>99.9</td>
<td>99.9</td>
</tr>
<tr>
<td>Positive predictive value, %</td>
<td>2.65</td>
<td>3.16</td>
<td>0.57</td>
<td>0.53</td>
<td>6.80</td>
<td>5.90</td>
</tr>
<tr>
<td>Relative incidence rate §</td>
<td>1:15350</td>
<td>1:13050</td>
<td>1:54900</td>
<td>1:62800</td>
<td>1:59600</td>
<td>1:67100</td>
</tr>
<tr>
<td>(95% confidence interval)</td>
<td>(1:13700-1:17450)</td>
<td>(1:11700-1:14700)</td>
<td>(1:44200-1:72400)</td>
<td>(1:55500-1:72400)</td>
<td>(1:41300-1:104100)</td>
<td>(1:46100-1:123200)</td>
</tr>
</tbody>
</table>

*Includes classic phenylketonuria (PKU) and clinically significant variant PKU.
†Includes only classic galactosemia.
‡Relative incidence rate is defined as number of live births for each confirmed (diagnosed) case identified through newborn screening.

Table 2. National Newborn Screening Results for Endocrinopathies

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive test results, No.</td>
<td>64700</td>
<td>63035</td>
<td>11237</td>
<td>9410</td>
<td>9142</td>
<td>1203</td>
</tr>
<tr>
<td>Confirmed cases, No.</td>
<td>1142</td>
<td>1203</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Apparent sensitivity, %‡</td>
<td>98.3</td>
<td>98.5</td>
<td>99.0</td>
<td>99.2</td>
<td>1.77</td>
<td>1.91</td>
</tr>
<tr>
<td>Specificity, %</td>
<td>1.77</td>
<td>1.91</td>
<td>0.53</td>
<td>0.54</td>
<td>1.3300</td>
<td>1.3300</td>
</tr>
<tr>
<td>Positive predictive value, %</td>
<td>1.3300</td>
<td>1.3300</td>
<td>1.19200</td>
<td>1.25100</td>
<td>(1:3100-1:3500)</td>
<td>(1:3100-1:3500)</td>
</tr>
</tbody>
</table>

*Includes only primary congenital hypothyroidism.
†Includes classic congenital adrenal hyperplasia, salt wasting, and simple virilizing.
‡Relative incidence rate is defined as No. of live births for each confirmed (diagnosed) case identified through newborn screening.

The overall average incidence for 4 years (4902516 newborns tested) was 1 confirmed case of biotinidase deficiency for every 80369 live births.

Our PPV calculations given in Table 1 reveal that biotinidase deficiency screening test results are associated with a much higher proportion of newborns who have true-positive test results compared with PKU and galactosemia. When comparing PPV between 1993 and 1994 for each disease, PKU showed a small statistical difference (P = .04), galactosemia showed no statistical difference (P = .70), but biotinidase deficiency showed significant statistical variance (P < .001). Because of the relatively small numbers for biotinidase deficiency screening, we again analyzed the 1991 and 1992 CORN reports and calculated PPV figures. In 1991 and 1992, biotinidase screening results produced PPVs of 3.7% and 9.1%, respectively. The average for 4 years was 6.4%. We attribute this variability to both the relatively low number of abnormal test results and the relatively small number of cases in the United States with biotinidase screening included in their panel (only 13 states in 1991 and 20 in 1994). However, it is clear that despite this variability, the biotinidase test is 2-fold and more than 10-fold better than PKU and galactosemia, respectively, in reducing false-positive test results.
West Virginia in 1993 and Alabama and Maine in 1994 were removed owing to incomplete data. For CAH, data for Georgia in 1993 and Alabama and Georgia in 1994 were eliminated. Both congenital hypothyroidism and CAH produced specificities that were slightly lower than those of the biochemical disorders. Calculating disease incidence reveals that congenital hypothyroidism has the highest relative incidence rate in the newborn population of the United States, averaging 1:3300. The high precision and consistency of this test is also obvious from the relatively narrow 95% CI given in Table 2. Similar results were found in the 1992 CORN report, but 1991 reporting was too incomplete for accurate calculations. Although the estimated incidence of CAH was less precise, the average of 1:22,150 for 1993 and 1994 was similar to the values of 1:19,291 and 1:20,671 for 1991 and 1992, respectively. The overall average for 4 years (4233,799 newborns tested) was 1 confirmed case of classic, salt wasting, or simple virilizing CAH for every 20,355 live births, which is near the mid point of the 95% CIs given in Table 2. Results for CAH revealed the lowest PPV of 5 disorders studied, at 0.53% and 0.54% for 1993 and 1994, respectively. There was no statistical variance between the PPV of the 2 years for either congenital hypothyroidism or CAH.

The magnitude of the false-positive test results can perhaps best be appreciated by examining aggregated actual numbers given in Table 3. For 5 disorders, more than 90,000 false-positive test results occurred in each of 2 years examined in detail. The ratio of false- to true-positive test results is more than 50:1 in both 1993 and 1994. It is clear from the data given in Table 2 that screening for congenital endocrinopathies contributes the highest fraction of false-positive test results.

### Table 3. Total Number of False-Positive Test Results Compared With Newborns With Confirmed Phenylketonuria, Galactosemia, Biotinidase Deficiency, Congenital Hypothyroidism, and Congenital Adrenal Hyperplasia

<table>
<thead>
<tr>
<th>Result</th>
<th>1993</th>
<th>1994</th>
</tr>
</thead>
<tbody>
<tr>
<td>False-positive test results, No.</td>
<td>93171</td>
<td>90516</td>
</tr>
<tr>
<td>Confirmed cases, No.</td>
<td>1530</td>
<td>1816</td>
</tr>
<tr>
<td>Ratio</td>
<td>61:1</td>
<td>56:1</td>
</tr>
</tbody>
</table>

**COMMENT**

This is the first study to our knowledge that analyzes the results associated with newborn screening programs in the United States. As these programs attempt to detect every case of hereditary metabolic disorders and congenital endocrinopathies, they generate data useful for determining disease incidence rates on a national level and the characteristics of laboratory testing procedures. By examining CORN reports from 1990 through 1994, we were able to verify improvement of national data collection and to conclude that 1993 and 1994 results portray the experience in a reasonably accurate fashion. Satisfied with the consistency of the data available, we calculated disease incidence for 5 disorders. This enabled us to determine best estimates of relative incidence rates with more than 4 million births available for each determination. This gives us a high level of confidence in the following estimates. For PKU, we determined that the disease incidence nationwide is 1:14,000, with a predicted range of 1:12,000 to 1:17,000. These figures are in keeping with expectations from other studies.11,12 Our analyses concerning galactosemia reveal a relative incidence of 1:59,000, a figure close to an earlier determination of 1:62,000.13 Biotinidase deficiency is the disorder that is the most challenging for determining incidence; however, we were able to show that on average, 1 case should be expected for every 80,000 live births. The 95% CI, however, is so large (Table 1) that many state screening programs may experience years without any such diagnoses. Reviewing the 1993 and 1994 CORN reports reveals that 22 of 40 state-years resulted in no diagnoses of biotinidase deficiency through neonatal screening.

Congenital endocrinopathies are detected with more consistency than the 3 hereditary metabolic disorders. We were able to determine a precise relative incidence rate of congenital hypothyroidism of 1:3300, which is in keeping with the literature.3 With CAH our results consistently show an incidence of 1:20,000; thus, we feel that this figure should be regarded as currently the best estimate rather than the incidence of 1:11,000 to 1:15,000 reported previously.14,15 Determination of the characteristics of laboratory tests was our second objective. The usual determinants are sensitivity, specificity, PPV, and negative predictive value. The sensitivity and specificity are characteristics of the test per se, but disease prevalence influences the predictive values for both positive and negative test results.6 The results of our analyses reveal that PPV is more salient than test specificity and can be especially important in reaching judgments about the quantitative impact of false-positive newborn screening outcomes. The range of PPVs determined in this study was quite wide, ie, approximately 0.5% for galactosemia and CAH compared with 6.0% for biotinidase deficiency. For each disorder, however, it is obvious that newborn screening produces a relatively high number of false-positive results. This is best indicated by a combination of consistently low PPVs and the resultant ratio of false- to true-positive results. Values approximating 0.5% that are evident for both galactosemia and CAH translate into about 200 false-positive newborn test results for every confirmed case. Also, there has been little change in the PPVs associated with newborn screening results in the current decade. This indicates that there has been no improvement in the number of false-positive test results in newborn screening programs. As mentioned previously, laboratory costs and toxic effects secondary to psychological stress in the parents are primary reasons to address this issue of false-positive results.

Screening for congenital hypothyroidism has the highest absolute number of false-positive results. The cost of repeated testing is more than $2 million annually, based on information in the 1994 CORN report. Attempts have been made to decrease this number by using age-dependent criteria for thyrotropin levels.16 Similarly, several regions performing CAH screening have imple-
mented weight-dependent criteria. While these are potentially important enhancements, further refinements are necessary such as the use of molecular tests. Also, as new improvements in screening methods develop, it will be important to implement them nationwide. There is considerable variation in the types of assays, laboratory value cutoffs, and even the diseases screened from region to region, producing variable results.

Achieving uniformity nationwide in newborn screening programs would hasten improvements and also facilitate more complete data collection. In addition, since false-positive results will never be completely eliminated, greater care and consideration should be given to risk communication and counseling of parents before and after the screening results. This should be regarded as an obligation of newborn screening programs because of the potential psychosocial harm of false-positive test results. Unfortunately, genetic counseling is quite variable across the country.

Newborn screening tests have significantly enhanced the detection of congenital disorders of metabolism, and programs employing such methods have made extraordinary contributions to public health. The good of these programs is indisputable, but we must recognize false-positive test results as an adverse effect and devote more attention to addressing the morbidities associated with them, along with the opportunities to reduce the magnitude of the problem. It must also be remembered that the clinical practice associated with newborn screening should begin with, not end with, the screening test result. Effective communication needs to be part of optimized practice for families who encounter either false- or true-positive results.

Accepted for publication January 13, 2000.

This study was supported by grants DK 34108 and M01 RR03186 from the National Institutes of Health, Bethesda, Md, and a grant (A001-5-01) from the Cystic Fibrosis Foundation, Bethesda (Dr Farrell).

We thank Brad Therrell, PhD, and Gary Hoffman for help in obtaining and interpreting the data used in this project. We are also grateful to Lan Zeng, MS, for her invaluable assistance with statistical analyses.

Corresponding author: Philip M. Farrell, MD, PhD, University of Wisconsin Medical School, Room 1217 MSC, 1300 University Ave; Madison, WI 53706 (e-mail: pmfarrel@facstaff.wisc.edu).

REFERENCES