Audiometric Accuracy of the Click ABR in Infants at Risk for Hearing Loss

Martyn L. Hyde
Krista Riko
Kathy Malizia

Abstract
The auditory brainstem response (ABR) to clicks is widely used for early detection of hearing loss in the child at risk for hearing dysfunction, but there is a lack of direct, large-sample estimates of test accuracy.

In this report, results and preliminary analyses are presented that relate click ABR thresholds obtained at 3 to 12 months corrected age to detailed follow-up behavioral pure-tone audiometry at 3 to 8 years of age, for 1,367 ears in 713 children at risk for hearing loss.

The data are analyzed in terms of conventional 2x2 decision matrices and associated parameters, using dichotomous (binary) measures of hearing loss and ABR test outcome. The accuracy of the ABR appears to depend strongly on the precise criteria that are chosen to define both hearing loss and ABR outcome.

ABR accuracy is excellent for detecting average sensorineural hearing loss at 2 and 4 kHz in excess of 30 dB, and the overall results for a wide range of hearing loss and ABR abnormality criteria can be conveniently summarized in terms of relative operating characteristics (ROCs).

Key Words: Auditory brainstem response (ABR), accuracy, validity, infant hearing screening, relative operating characteristic

It is widely believed that early detection and management of hearing loss in the infant will improve the development of communicative skills. There is a lack of formal, direct scientific evidence that this is indeed so, but such evidence will be difficult to obtain. The hypothesis has much face validity, and it seems evident that accurate information about hearing must at least facilitate the early management of the multiply-handicapped child.

In 1982, the U.S. Joint Committee on Infant Hearing recommended that infants at risk for hearing loss should be screened by 3 months of age and that intervention should be started by 6 months of age, wherever possible. The Joint Committee did not specify any particular method of testing hearing, but a consensus is now emerging that behavioral observation audiometry and, more generally, any methods based on scoring of reflex responses to loud sounds are not sufficiently accurate, especially if it is desired to detect mild or moderate hearing losses. In 1987, the U.S. Committee on Hearing, Bioacoustics and Biomechanics (CHABA) concluded that the auditory brainstem response (ABR) was the most objective available tool. In 1989, the ASHA Committee on Infant Hearing recommended audiometric screening using the ABR.

During the last decade, there have been many primary reports concerning the use of the ABR in infant hearing testing (Alberti et al, 1985; Swigonski et al, 1987), and there has been a proliferation of clinical programs for early detection of hearing loss, based on ABR techniques. Most commonly, a click stimulus is used to elicit the ABR. Thus, the click ABR has emerged as a commonly used tool for screening or for threshold estimation in infants.

Despite the popularity of the ABR, there is a need for more data concerning the audiometric accuracy of this tool in the high-risk infant
population. While there have been several reports that address follow-up audiometry in ABR screening failures, a comprehensive review by Murray et al (1985) emphasized the need to verify ABR test outcomes not only in those who fail an ABR screening test, but also in those who pass. This was endorsed by CHABA in 1987, noting a need for statistically adequate studies of ABR test accuracy. This need continues to exist. The practice of validating neonatal or infant ABR data using a subsequent ABR test as the “gold standard” addresses the reproducibility of ABR results, but not their accuracy as estimators of true hearing thresholds.

At present, it is probably necessary to use reliable behavioral audiometric results as the gold standard, despite the problems that arise from the time interval between the ABR testing and the point at which behavioral testing becomes sufficiently reliable to serve as a standard. To accommodate the Joint Committee guidelines it is necessary to test below 6 months of age, but is testing best done in the neonate prior to discharge, or as late as possible, consistent with the guidelines? Perhaps the most important factor is access to the patient, which is potentially guaranteed if testing is performed prior to discharge, whereas the parents may not comply with instructions to return for postdischarge testing. A question here is whether parents who will not bring the child back for later evaluation would comply in any useful way with management strategies based on a predischARGE test. Clearly, this is not a straightforward question, and local conditions may dictate local solutions.

The practice at Mount Sinai Hospital, Toronto, since 1984 has been to test at-risk infants as outpatients at about 4 months corrected age, that is, at about 56 weeks postconception. One of the reasons for this was a finding from an earlier research study (Hyde et al, 1984) that ABR results differed considerably in the neonatal period and at about 4 months of age. The later results are certainly more relevant to any subsequent intervention, and this view has been supported by Swigonski et al (1987). Thus, ABR results at 3 to 6 months of age are of particular importance; the ABR testing to be evaluated in this report was conducted in the first year of life, but never at less than 3 months corrected age.

MATERIAL AND METHODS

Subjects

Subjects were drawn from Mount Sinai Hospital (MSH) and the Hospital for Sick Children (HSC), which are adjacent teaching hospitals in the Toronto downtown core. MSH is a general hospital with a program for early detection of hearing loss that is based on ABR and a risk register similar to that recommended by the Joint Committee. The prevalence of ABR screening failure at 30 dB nHL is typically 8 percent. HSC, in contrast, is a tertiary center with a much higher prevalence of ABR abnormality, typically 60 percent.

The study sample comprised 1,065 at-risk infants from MSH, a complete at-risk sample over a 3-year period, targeted without regard to whether or not an ABR had been performed. To this was added a random sample of 135 children from HSC who had had an ABR test in the first year of life. Thus, a total of 1,200 children were targeted for follow-up audiometric evaluation.

ABR Protocol

At MSH, all ABR testing was carried out in an audiometric soundroom, using a Nicolet MED-80 system. Silver chloride disk electrodes were attached at the high midline forehead and on both mastoids. Recording bandwidth was 150 to 3000 Hz (Butterworth, 12 dB/octave), with a 25.6 ms window. Two averages of at least 2000 stimuli each were accumulated, per stimulus condition. Thresholds were obtained for 2-1-2
ms trapezoidal notch masked tonepips at 500 Hz and 4000 Hz, but this report deals exclusively with click results.

At HSC, the equipment and protocol were similar, except that general anesthesia was applied where necessary in order to obtain satisfactory EEG conditions, and only clicks were used.

The criterion for abnormality of an ABR record was presence or absence of wave V or its ensuing negative wave. All ABR thresholds were estimated by an experienced observer (MLH). ABR threshold was defined as the lowest click level at which a clear and reproducible waveform was judged to be present. Ears for which the level of spontaneous electromyogenic artifact was high were rejected from the analysis.

Follow-Up Protocol

Targeted subjects were recalled at age 3 to 8 years. All testing was carried out by two experienced audiologists in an audiometric sound room. The testers had no knowledge of the ABR results in infancy. The protocol included conventional pure-tone air and bone conduction thresholds, speech recognition testing, acoustic impedance and reflex tests, and a language screening test, the TACL-R (Carrow-Woolfolk, 1985). Also, a questionnaire covering demographic factors and pertinent history was administered to parents. This report deals exclusively with the pure-tone audiometric data.

Data Management

All data from the risk assessment and ABR testing in infancy, and the follow-up results, were recorded in a relational database management system (Oracle RDBMS). Analysis was via Oracle SQL or SPSS-X (DEC Microvax II, VMS).

RESULTS

Of the targeted group of 1200 children, reliable behavioral pure-tone audiometry under earphones was subsequently obtained in 865 children (72 percent); of these, 713 (82 percent) had had an ABR test in the first year. The results to be presented here are based on 1,367 ears that yielded reliable click ABR thresholds and follow-up pure-tone thresholds. Of these 713 children, 397 were male and 316 were female. The mean age at follow-up was 3.9 years, with an SD of 0.9 and range of 3.0 to 8.0; 99 percent were tested at 6 years or less.

Figure 1 shows a scatterplot of the click ABR threshold obtained in infancy, on the ordinate, versus the average pure-tone threshold at 2 and 4 kHz, on the abscissa. While there are many possible audiometric measures that might be chosen to summarize hearing loss severity, this measure was noted by Gorga et al (1985) to be one of the best correlates of the click ABR threshold, when obtained concurrently. It can be argued that the performance of the click ABR should be assessed in relation to audiometric measures that reflect appropriately the acoustic and electrophysiologic attributes of the click.

The use of ears, as opposed to individuals, as data points can lead to incorrect confidence interval width for estimates of means, because the two ears of any given individual rarely yield statistically independent data. However, the use of exclusively right ear or left ear data is excessively conservative for many purposes, and discards information. No confidence intervals are given here, but they may be estimated conservatively by halving the sample size. There were no statistically significant differences between ears, as reflected in chi-square tests at the 0.05 level of significance.

The scatterplot in Figure 1 is presented in a manner that permits direct comparison with the standard 2x2 decision matrix or contingency table (Jerger, 1983; Sackett et al, 1985); the four cells of the table correspond directly to the four quadrants of the scatterplot delineated by the lines denoted as ‘disease criterion’ and ‘test criterion.’ Here, the target disease is hearing loss, with subjects denoted as disease-positive if they exceed the disease criterion, and disease-negative otherwise. This is based on the follow-up audiometry and is considered to represent the ‘true’ disease status. The subject tests positive for disease on the infant ABR test if the ABR threshold exceeds the test criterion. Any pair of disease and test criteria will generate a corresponding decision matrix. In principle, any disease and test criteria may be chosen, and the two criteria need not have the same numerical value. Indeed, because of the number of variables that affect the statistical relationship between the ABR threshold and the hearing loss measure (e.g., audiometric profile, EEG noise
Figure 1  A scatterplot of average behavioral pure-tone hearing threshold at 3 to 8 years (ordinate) versus the click ABR threshold in the first year, for 1,367 ears in 713 children at risk for hearing loss. The axes are such that overlaying disease (hearing loss) and test (ABR) abnormality criteria upon the scatterplot will generate directly the four cells of the standard 2x2 decision matrix. See text for explanation of "abnormal" cases, e.g., with "normal" abscissa values.

levels, ABR recognition criteria), there is no reason to assume that the two measures should intrinsically be the same.

The dichotomization of essentially continuous measures such as the ABR threshold or the average hearing level certainly results in a formal loss of information, but it permits the use of many powerful analysis techniques and is well suited to estimation of screening test performance. An alternative analytic approach is regression analysis, treating the ABR threshold as the independent variable (abscissa) that predicts the follow-up audiometric status; performance would be expressed in terms of bias and variability of the prediction. Here, the approach based on the decision matrix will be used.

In Figure 1, there is an obvious tendency for the ABR and behavioral thresholds to covary, clustering around the dotted diagonal that denotes equality. Furthermore, the audiologists' determination of the type of hearing loss at follow-up is also indicated by the symbol type. The term 'normal' denotes the fact that no behavioral threshold was greater than 20 dB, from 250 Hz to 8 kHz, which is a clinical criterion quite distinct from the outcome of applying a numerical criterion to the two-frequency average used in the figure. Thus, it is possible to have patients with clinical dysfunction at any point on the scatterplot.

When the criterion lines are overlaid on the scatterplot, it is immediately obvious that the criteria used to define the presence or absence of disease and pass or failure on the test under investigation (in this case, the infant ABR) will have a profound effect on the number of data points in the various quadrants of the scattergram. It follows that the decision matrix and its attendant parameters of test performance, such as the sensitivity (the conditional probability that the ABR test is positive for disease, given that disease is present), specificity (the conditional probability that the test is negative, given that disease is absent), and the likelihood ratios for the two test outcomes (for each out-
come, the ratio of its conditional probabilities when disease is present and absent) will depend on the disease and test criteria. Thus, some means of expressing the effect of criterion variation is required. One of the most elegant approaches is the Relative Operating Characteristic (ROC), known earlier as the receiver operating characteristic. Recently, Swets (1988) has emphasized the power and generality of the ROC approach to the analysis of test performance.

First, consider the generation of a single decision matrix. Suppose it is decided that average hearing losses greater than 40 dB, regardless of type of hearing loss, are the target disease, and it is required to evaluate a click ABR screening criterion level of 40 dB nHL (the equality is coincidental). The criteria would then have the positions shown in Figure 1. The 12 ears in the false-negative (lower left) quadrant, for example, are obvious on the scatterplot. The associated 2x2 table is shown in Table 1a; the sensitivity and specificity are 0.81 and 0.96, respectively, which is good performance.

To improve the sensitivity, the ABR abnormality criterion could be lowered, perhaps to 30 dB, and this would give the results shown in Table 1b. The sensitivity and specificity are now 0.84 and 0.91. Thus, lowering the ABR criterion reduces the false-negative rate from 0.19 to 0.16, but more than doubles the false-positive rate, from 0.04 to 0.09.

When defining the target hearing loss characteristics, there is more to consider than the nature of the hearing loss measure and the value of the abnormality criterion. For example, suppose it is desired to focus on sensorineural impairment. In terms of the scatterplot, this means that all unfilled circles in the two disease-positive quadrants are moved into the disease-negative column, becoming either false positives or true negatives. The revised scatterplot would yield the matrices shown in Table 1c and 1d, for 40 dB disease and 40 and 30 dB test criteria, respectively. Note that the focus on sensorineural hearing loss changes the decision matrices radically, particularly improving the sensitivity of the test.

A more complete picture of the way in which ABR test accuracy depends on the various criteria is shown in Figure 2. Here, the ordinate is the true-positive rate, or sensitivity, and the abscissa is the false-positive rate, namely the complement of the specificity. In this so-called

| Table 1 Examples of 2x2 Contingency Tables (Decision Matrices) Derivable From Figure 1 |
|------------------|------------------|
| (a) All Hearing Loss | (b) All Hearing Loss |
| Yes(D+) No(D-) | Yes(D+) No(D-) |
| ABR 40 dB nHL Fail(T+) 51 50 | ABR 30 dB nHL Fail(T+) 53 117 |
| Pass(T-) 12 1254 | Pass(T-) 10 1187 |
| Sensitivity 0.81 | Sensitivity: 0.84 |
| Specificity 0.96 | Specificity 0.91 |
| LR+ 21.1 LR- 0.20 | LR+ infinite LR- 0.0 |

<table>
<thead>
<tr>
<th>(c) SN Hearing Loss</th>
<th>(d) SN Hearing Loss</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes(D+) No(D-)</td>
<td>Yes(D+) No(D-)</td>
</tr>
<tr>
<td>ABR 40 dB nHL Fail(T+) 44 57</td>
<td>ABR 30 dB nHL Fail(T+) 45 125</td>
</tr>
<tr>
<td>Pass(T-) 1 1265</td>
<td>Pass(T-) 0 1197</td>
</tr>
<tr>
<td>Sensitivity 0.98</td>
<td>Sensitivity 1.0</td>
</tr>
<tr>
<td>Specificity 0.96</td>
<td>Specificity 0.91</td>
</tr>
<tr>
<td>LR+ 22.7 LR- 0.02</td>
<td>LR+ infinite LR- 0.0</td>
</tr>
</tbody>
</table>

The hearing loss criterion used here is that the average pure-tone threshold at 2 and 4 kHz is over 40 dB HL. The ABR criteria are a click ABR threshold of over 40 dB nHL or over 30 dB nHL. The sensitivity, specificity, and likelihood ratios for positive (fail) and negative (pass) ABR outcomes are also shown. LR+ = sensitivity/(1-specificity) and LR- = (1-sensitivity)/specificity. Post-test odds for disease = pretest odds x LR(test outcome).
parameter is essentially the distance between the means of the disease-negative and disease-positive distributions, expressed in terms of a number of standard deviations, and is a useful summary measure of test accuracy. A useless (random) test has a $d'$ of zero, whereas an excellent test might achieve a value of about 3.0 or more. See Swets (1988) for a more detailed discussion of ROC techniques and test accuracy measures.

In Figure 2 are plotted six ROCs, three for a disease definition that includes all types of hearing loss and three for a definition restricted to sensorineural hearing loss. In each case, the hearing loss severity criterion takes values of 20, 30, and 40 dB as indicated. For every ROC, the four points correspond to four values of the ABR abnormality criterion, namely 60, 50, 40, and 30 dB nHL, reading from left to right.

For every ROC in Figure 2, the trading relationship between sensitivity and specificity as the test criterion is changed is apparent. Accuracy improves as the hearing loss criterion increases, and the accuracy is generally higher for sensorineural hearing loss. All the ROCs show good linearity, and those for sensorineural loss tend to follow the normal, equal variance model most closely. For sensorineural loss of over 30 dB, averaged at 2 and 4 kHz, the ROC slope is close to unity and the $d'$ is about 3.5, which means that the infant ABR is an excellent test for hearing loss thus defined. Note that changing the ABR abnormality criterion alters the sensitivity and specificity according to the bilinear normal ROC rule, but this does not alter the global accuracy of the test, as expressed by the approximately constant $d'$ value of about 3.5.

For all hearing loss criteria other than those just noted, the ROC slope is very different from unity, so there is no unique value of $d'$ that globally summarizes the test performance over a range of test criterion values. Here, it is necessary to resort to other global measures of test accuracy, such as the area under the ROC curve, usually denoted as $A$ (Swets, 1988), or to deal with particular sensitivity-specificity pairs, in discussions of test accuracy.

**DISCUSSION**

The results and analyses presented in this report are preliminary, but may be of interest to those who are about to establish, or
are already engaged in, clinical programs for detection of hearing loss using ABR techniques. Insight into the relationships between sensitivity and specificity as a function of the test and disease criteria is an essential prerequisite for meaningful cost benefit analysis and resource allocation.

The data outlined here suggest that as a detector of sensorineural hearing loss with an average value of greater than 30 dB at 2 and 4 kHz, the click ABR test performed at between 3 months and 1 year of age is an excellent tool. By varying the ABR test failure criterion from 30 dB through 60 dB, the sensitivity can be adjusted from over 98 percent down to about 85 percent, with concomitant change in the false-positive rate from about 10 percent down to less than 1 percent. The choice of ABR criterion should depend on factors such as the prevalence of hearing loss in the tested population and the quantitative costs associated with test overhead and outcome error (Sackett et al, 1985).

In this study, the hearing status as determined at follow-up is used as a proxy for the true hearing status in the first year of life because there is no behavioral technique that has sufficient accuracy to constitute a gold standard at the time of the ABR test. The problem with a delay of several years is that the true hearing status may change. First, consider sensorineural hearing loss. In infants and young children, this will rarely fluctuate or remit, but certainly may be initiated or may progress in the period between the ABR and the follow-up audiometry. This will cause apparent false-negative ABR outcomes. From a regression standpoint, there will be negative bias in the ABR as a predictor of the true hearing level. In the scatterplot of Figure 1, there is one sensorineural loss point (at 55, 40) that is likely to be counted as false-negative. This arose from a child who was born at 1260 g after 30 weeks, with severe asphyxia and intraventricular hemorrhage. Also, at higher ABR thresholds there are several points for which the hearing loss is 20 dB or more greater than the ABR threshold. Because ABR thresholds are typically greater than concurrent behavioral thresholds, progression of hearing loss could be suspected for several of these points.

If sensorineural hearing loss rarely remits, resolving conductive impairment is by far the most probable cause of false-positive ABR findings.

If conductive hearing loss were to be encompassed in the disease definition, which is not unreasonable clinically, then there are much more serious difficulties inherent in the follow-up validation because of the greater probability of change in hearing. For example, even if the conductive hearing loss components in infancy and at follow-up were statistically unrelated, there would by chance be a certain proportion of test outcomes that are apparently, but not actually, correct. A more detailed analysis might take account of such factors, but the simplest approach, adopted here, is to focus on the data for sensorineural loss; in that regard, conductive pathology is a source of variation that can only degrade the apparent performance of the ABR test.

There are several techniques that might be used to improve the performance of the ABR. For example, it may be possible to identify conductive hearing loss on the basis of ABR wave latency increase, e.g., for wave I. If such a classifier performed well, then many of the false positives that limit the observed ABR performance might move to the true-negative region. It remains to be seen whether a latency rule can be derived that will produce a net performance enhancement.

In the bottom right-hand corner of the scatterplot of Figure 1, there are several points indicating the presence of sensorineural hearing loss. For example, hearing losses at low or high frequencies could and did occur, yet were not reflected in the behavioral outcome measure selected here, namely the 2 and 4 kHz average pure-tone hearing level. Because a hearing loss component in infancy at 8 kHz might indicate a progressive sensorineural disorder that should at least be flagged for monitoring, the 2 and 4 kHz average is imperfect as a clinical outcome measure. Indeed, there are many other measures that merit consideration in a more detailed analysis.

Furthermore, features of the audiometric contour may contribute in other ways to discrepancies between ABR and behavioral measures. For example, underestimation of average hearing loss by the ABR threshold may result from better hearing at frequencies outside the domain of the average but within the wide excitation bandwidth of the click (Kileny and Magathan, 1987).

When the ABR test is normal but there is sensorineural hearing loss at follow-up, progressive or adventitious dysfunction is one cause of apparent error, but the fault lies in the deferred gold standard. On the other hand, it is possible
to see such a result even though the hearing loss was present throughout, if the dysfunction is more rostral than the site of generation of ABR wave V (probably subcollicular). This is a genuine error and limitation on the part of the ABR test. It is not yet clear if cortical dysfunction, for example, might have contributed to any of the data reported here.

False-positive ABR findings, in the sense of false-positive detection of disease, could arise due to poor recording conditions or techniques, but records that showed poor reproducibility were excluded from this dataset. A more subtle possible cause is poor ABR development due to inadequate neural synchronization; this is of concern in those at risk for neurodevelopmental disorders, such as the graduates of an intensive care nursery.

In conclusion, the results presented here are encouraging for advocates of the ABR as a tool for early detection of hearing loss, but much further analysis of this dataset is required, and is in progress. This analysis includes examination of the contribution that tonepip ABR thresholds and ABR wave latency measurements can make to test accuracy, introduction of the risk factors as covariates and sample stratifiers, and exploration of relationships among risk factors, ABR findings, and language screening outcomes. Further studies in this area are required to delineate fully the confidence that should be placed in infant ABR outcomes, either as bases for intervention or as secondary standards by which newer techniques such as evoked oto-acoustic emission measurement may be evaluated.

REFERENCES


