Fetal monitoring at home in high-risk pregnancy. An integrated clinical and economic evaluation

Monincx, W.M.; Birnie, E.

Publication date
2000

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Chapter 3

Interobserver variability of the neurological optimality score

Wilma M. Monincx¹, Hetty Smolders-de Haas², Gouke J. Bonsel³, Hans A. Zondervan¹

¹Department of Obstetrics and Gynecology
²Department of Neonatology
³Department of Clinical Epidemiology and Biostatistics
Academic Medical Center Amsterdam, the Netherlands

Published in:
Abstract

Objective: To assess the interobserver reliability of the neurological optimality score.

Study design: The neurological optimality score of 21 full term healthy, neurologically normal newborn infants was determined by two well-trained observers.

Results: The interclass correlation coefficient was 0.31. Kappa for optimality (score of 58 or higher) was 0.19. A systematic difference of 1.3 points between the two observers was present.

Conclusions: The interobserver variability of the neurological optimality score of the newborn infant is substantial. Especially the subtle judgement of elicited responses as optimal or not-optimal proved to be critical in this concordance study.

A difference of at least two points in the score is considered as a valid endpoint for comparative studies. If two or more observers are involved in the neurological examination of the newborn infant in a study to assess influences on perinatal morbidity, frequent re-instruction sessions are recommended.
Introduction

The neurological examination of the newborn infant, as described by Prechtl, provides a detailed assessment of the integrity of the infant's nervous system. It is a rigidly standardized clinical procedure, which allows easy repetition to render reproducible results. The complete neurological Prechtl-examination requires approximately 250 different items to be scored according to well-defined criteria. At the end of the neurological Prechtl-examination a final neurological diagnosis can be reached. Clear-cut criteria for neurological syndromes as the apathy syndrome, the hemis syndromes or the hyperexcitability state are available. A somewhat more subjective clinical judgement is allowed alongside the strict neurological diagnosis from the complete neurological Prechtl-examination, classifying the child neurologically as “normal”, “suspect” or “abnormal”. The direct relation between the neurological condition of the infant in the first week after birth, the neurological diagnosis and the clinical judgement on the one hand and the subsequent neuromotor development on the other hand is not well established. Spontaneous recovery and uneventful further normal development occur frequently. Thus the ability to predict infant development from the neurological Prechtl-examination is not perfect.

In modern obstetrics results of care are measured in terms of perinatal morbidity. Formal scoring systems for handicaps, impairments and disabilities in later adult life are used to classify the late sequelae of perinatal events. The only valuable design to acquire this information is an extensive long-term follow-up of a child's development in neuromotor and cognitive function. The drawback is, that the results of yesterday's standard of perinatal care become available after a long time. Hence, the rapid introduction of new treatment modalities in obstetrics and neonatology can not be judged by the gold standard of long term outcome. Moreover, unrelated interval events (as trauma or disease) may hamper interpretation of the results. To overcome these problems, Prechtl postulated the “optimalit y concept”. A newborn in a neurologically optimal condition in the first week after birth has the best imaginable starting position for later development. An optimal neurological condition after birth, in this way, is considered to be the best proxy for long-term outcome. Apart from the definition of 'normal' response in the full neurological examination, Prechtl also defined the “optimal” response for a subset of 60 items. The summation of the number of optimal responses yields the neurological optimality score, which has been validated against long term outcome. An 'optimal' neurological condition provides the best prospects for an uneventful later development, whereas non-optimality is associated more frequently with later development of handicap, impairment or disability. It is obvious that neurological non-optimality is not the same as neurologically impaired, disabled or handicapped, but a proxy measure which is rapidly after birth available allowing for assessment of perinatal treatment policy.

The use of this neurological optimality score for the evaluation of the introduction of new perinatal technologies assumes that it is a reproducible measure. Where the
neurological examination of the newborn is strictly standardized and a definite neurological diagnosis and clinical judgement are well established to be reproducible, for the neurological optimality score no formal study of inter-rater variability has been reported sofar. In this paper we explore the interobserver reliability of the neurological optimality score.

Patients and methods

From a cohort of participants in a randomized controlled trial of domiciliary antenatal fetal monitoring versus clinical observation in high-risk pregnancies, we obtained informed consent from 21 parents to have their newborn infant neurologically examined twice. Twenty-one newborn infants were thus examined by two examiners (WMM and HS-deH) at separate times in the appropriate timeframe between 3 and 10 days post-partum, at corrected full term age. Both examiners received training in the neurological examination of the newborn at the department of Developmental Neurology at the University of Groningen by Prof. B.C. Touwen. For each neurological examination an appointment was made at the infants home, at the outpatient pediatric clinic or at the neonatology ward, depending on the place where the infant was. The examinations took place at least 6 hours apart, in similar conditions (e.g. time after feeding). The sequence of examiner 1 and 2 was randomly determined. A full neurological examination was performed and scored on the standard proforma sheet. A neurological diagnosis was made up and a clinical judgment was entered at the end of each examination. The optimality score was derived at a later stage from the records of the full neurological examination. The result of the examination of the first examiner was not available for the second examiner. The parents were instructed not to disclose any result of the first examination to the second examiner. Both examiners were instructed not to ask for the result of the first examination. The study was approved by the medical ethical committee of the hospital.

Statistical comparisons were made using the packages Epi Info (version 5) and SPSS using Pearson’s product moment correlation coefficient, interclass correlation coefficient, kappa-statistics and a graphical method as appropriate.

Kappa is a statistical measure to represent agreement beyond chance: it ranges from 0 to 1, where a value of 1 indicates perfect agreement and a value of 0 that the agreement is no better than chance. For clinical measures in general a kappa of 0.7 or higher is considered indicating excellent agreement, between 0.3 and 0.7 as fair, and below 0.3 as poor.
Results

The clinical characteristics of the 21 newborns are given in Table 1. It proved that both examiners classified all children neurologically normal. No definite neurological syndrome diagnosis was made in any of these children. Also the clinical judgement of both examiners for all 21 children was normal. None of the children was judged to be “abnormal” or “suspect”.

The neurological optimality score for these 21 newborn infants ranged between 53 and 60 for examiner A and between 55 and 60 for examiner B (Table 2). Neurological optimality scores of 58 and higher are considered to represent an optimal neurological condition. Examiner A rated 8 and examiner B rated 13 infants as optimal.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Median</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Birthweight (g)</td>
<td>3160</td>
<td>2230 – 4150</td>
</tr>
<tr>
<td>Gestational age at delivery (days)</td>
<td>279</td>
<td>258 – 302</td>
</tr>
<tr>
<td>Age at examination 1 (days)</td>
<td>6</td>
<td>4 – 49</td>
</tr>
<tr>
<td>Age at examination 2 (days)</td>
<td>8</td>
<td>5 – 50</td>
</tr>
<tr>
<td>% Caucasian</td>
<td>38</td>
<td></td>
</tr>
<tr>
<td>Infants with definite neurological syndrome</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Infants with clinical judgment “suspect” or “abnormal”</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Clinical characteristics of the 21 newborn infants

<table>
<thead>
<tr>
<th>Neurological optimality score*</th>
<th>Observer A number of infants</th>
<th>Observer B number of infants</th>
</tr>
</thead>
<tbody>
<tr>
<td>60</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>59</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>58</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>57</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>56</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>55</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>54</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>53</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>21</td>
<td>21</td>
</tr>
</tbody>
</table>

* A Neurological Optimality Score of 58 or higher is considered to be optimal.

Table 2. Frequency distribution of Prechtl Neurological Optimality Score of two observers (n=21 newborn infants)


**Figure 1.** Scatterplot of the neurological optimality score as given by two observers for 21 healthy, neurologically normal newborn infants. For each infant the optimality score of examiner 2 is given for the optimality score of examiner 1. Pearson’s product moment correlation is 0.40. The Interclass Correlation Coefficient is 0.31.

Figure 1 displays the individual scores for each newborn infant of both examiners. The Pearson’s correlation coefficient was 0.40. The interclass correlation coefficient was 0.31. The kappa for optimality (scores of 58 and higher) was 0.19.

To investigate further the nature of the difference between the observations of the two examiners, we plotted the differences between the two scores against the mean
INTEROBSERVER VARIABILITY OF THE NEUROLOGICAL OPTIMALITY SCORE

Legend:
MN: mean of the neurological optimality score of the two examiners for each infant
DIFF: difference between the optimality score of the two examiners for each infant
- circle: 1 newborn infant
- circle: 2 newborn infants
- circle: 3 newborn infants
- circle: 4 newborn infants

Figure 2. Difference per mean plot of the optimality score of two examiners for 21 healthy, neurologically normal newborn infants. For each newborn infant the difference between the scores of the two examiners is given for the mean of the scores of the two examiners.

value of the scores (Figure 2). It shows that observer A on average scored 1.333 points higher than observer B (S.D. 1.9). In five cases the difference was 4 points: for one infant both scores were non-optimal (53 and 57 respectively), for four other infants one examiner scored optimal (score of 60), whereas the other examiner classified the children as not optimal (score of 56). Review of the original examination sheets revealed no systematic explanation for these differences.
Of all 60 items of the neurological optimality score we found 26 variables that were scored as optimal by both examiners. These were items, generally based on observation, that were related to optimality of stability of states, position in rest and in supine and posture-variables. For 15 other items of the neurological optimality score examiner A scored all infants as optimal, where Examiner B scored 1 to 6 children as not optimal. These items were all responses that were elicited e.g.: sucking response, Bauer and Galant response. The amplitude of the Moro response was in 6 cases found not optimal by this examiner. Examiner B scored for another 7 items all children optimal, where examiner A scored 1 to 3 children as not optimal. These items were e.g.: tremor frequency of the Moro response and surprisingly, head circumference. For twelve items the infants were scored differently by both examiners, all but one of them reflexes e.g.: knee tendon reflex, rooting and stepping. Kappa values for theses individual items ranged from 0.08 for the optimality of the threshold for the Moro response, to 0.125 for the optimality of the lip reflex and to 0.5 for the glabella reflex, and 0.5 for the acoustical blink.

Discussion

The interobserver variability of the neurological optimality score in this series of 21 neurologically normal newborns proved to be of a level somewhat higher than in general might be expected for a rigidly standardized clinical examination. A kappa for optimality of 0.19 and an interclass correlation coefficient of 0.31 is rather disappointing. However, four important factors should be taken into account. Firstly, the neurological optimality score is a global derivative of the full neurological Prechtl-examination, which consists of formal scoring of about 250 variables and a continuous rather than a dichotomous total score. Secondly, our sample of 21 apparently neurologically normal newborn infants covers only the small upper part of the scale with consequent loss of discriminative power and an increased role of random error. Despite the small sample size of this study we do not expect the results to be much different in larger sample of similar children. The results, surely, may be expected to be better if this study was performed in a group covering the full range of possible optimality scores, i.e. also containing some neurologically not normal children. Thirdly, this add-on study tested agreement under practical field conditions. Finally, differences to some extent could be explained: the more manipulations are required, the lower the agreement. In adults a similar low agreement on reflexes has been reported. The subjective decision of an examiner to classify a certain elicited response, as firm or less firm remains the weak cornerstone of any clinical neurological examination.

These results do have implications for the use of the neurological optimality score as an instrument to assess effects of changes in policy in obstetrics and neonatology.

For evaluative studies on changes in obstetric or neonatal treatment policy we claim a difference of at least two points in the optimality score to be taken as a clinically
relevant outcome benefit. The interrater variability in our study showed a systematic deviation between two examiners of 1.3 points. When a smaller difference is observed, it might be explained by the variability in the endpoint assessment itself. Appointing only one examiner to assess all children could circumvent this problem. Then, intraobserver reliability must be high enough to justify this approach. We were not able to study this additionally in this context.

If two or more observers are involved in the neurological examination, with subjects judged by only one of them, as for example in large field studies, interobserver variance should be checked for frequently by re-instruction sessions. Use of a videoobservation of the neurological behavior of the newborn child may prove of more value as all observational items were rather consistent, whereas the more subtle judgment of an elicited response, as a tendon reflex, proved to be the most delicate in terms of agreement. Therefore, we suggest that in small studies the results of two or more examiners for all children should be taken into account or, although less easy to quantify, only the results of the full Prechtl examination.
References


