

## Validade do exame clínico do recém-nascido para a estimação da idade gestacional: uma comparação do escore New Ballard com a data da última menstruação e ultra-sonografia

Validity of neonatal clinical assessment for estimation of gestational age: comparison of New Ballard score with date of last menstrual period and ultrasonography

Claudia Leite Moraes <sup>1,2</sup>  
Michael E. Reichenheim <sup>1</sup>

<sup>1</sup> Núcleo de Pesquisa das Violências, Departamento de Epidemiologia, Instituto de Medicina Social, Universidade do Estado do Rio de Janeiro.  
Rua São Francisco Xavier 524, 7º andar,  
Rio de Janeiro, RJ 20559-900, Brasil.  
clmoraes@ims.uerj.br  
michael@ims.uerj.br

<sup>2</sup> Departamento de Endemias Samuel Pessoa, Escola Nacional de Saúde Pública, Fundação Oswaldo Cruz.  
Rua Leopoldo Bulhões 1480, Mangueiras,  
Rio de Janeiro, RJ 21045-900, Brasil.

**Abstract** *The objective of this article is to evaluate the accuracy of the New Ballard (NB) score in detecting prematurity. Date of last menstrual period (LMP) and ultrasonography (USG) are used as reference procedures. First, an assessment of inter-observer reliability of the NB score measurements and the USG reports was carried out. Regarding the NB scores, the kappa coefficient and the Intraclass Correlation Coefficient were, respectively, 0.74 (CI 95% 0.49-0.99) and 0.88 (CI 95% 0.78-0.93). For the USG analysis an Intraclass Correlation Coefficient of 0.96 (CI 95% 0.91-0.98) was found. Turning to validity, sensitivity of the NB score was below 0.70 in all but one comparison, irrespective of which procedure was taken as reference. Although specificity was consistently above 0.90, results suggest that the diagnostic ability of the NB score to detect prematurity is far from ideal. The proportion of false negatives may be a warning against its use alone, both in clinical practice and in epidemiological research contexts. Rather, the procedure should be recommended mainly as an accessory diagnostic tool.*

**Key words** *Gestational Age; Premature Infant; Diagnostic Test*

**Resumo** *O objetivo deste artigo é avaliar a acurácia do escore New Ballard (NB) para o diagnóstico de prematuridade. A data da última menstruação e a ultra-sonografia obstétrica foram utilizadas como procedimentos de referência. Realizou-se um estudo de confiabilidade interobservador da aplicação do escore NB e dos laudos ultra-sonográficos. O coeficiente kappa e o coeficiente de correlação intraclasses foram 0,74 (IC 95%: 0,49-0,99) e 0,88 (IC 95%: 0,78-0,93), respectivamente. Na avaliação das ultra-sonografias, foi encontrado um coeficiente de correlação intraclasses de 0,96 (IC 95%: 0,91-0,98). Em relação à validade, as sensibilidades do escore NB foram inferiores a 0,70 em cinco das seis comparações realizadas, independentemente do procedimento utilizado como referência. Apesar de a especificidade ter sido superior a 0,90, estes resultados sugerem que a capacidade diagnóstica do escore para a identificação da prematuridade está longe do ideal. A grande proporção de falsos negativos questiona o seu uso de forma isolada, tanto na prática clínica, como nos estudos epidemiológicos. Logo, o procedimento deveria ser recomendado apenas como ferramenta acessória na avaliação da idade gestacional.*

**Palavras-chave** *Idade Gestacional; Prematuro; Teste Diagnóstico*

## Introdução

Apesar dos avanços alcançados pelos programas de investigação sobre os fatores de risco da prematuridade, a maioria dos nascimentos prematuros ainda não tem fisiopatogênese conhecida (Kramer, 1987; Kramer et al., 1988; Mavalankar et al., 1992; Fortier et al., 1995; Hedegaard et al., 1996; Lang et al., 1996). Dentre as principais dificuldades dessa linha de pesquisa, destaca-se a falta de consenso com relação aos indicadores e medidas usados para aferir alguns de seus potenciais fatores de risco e à estratégia padrão utilizada para seu diagnóstico (Dunn, 1985; Henriksen et al., 1995).

Classicamente, a duração da gestação tem sido estimada considerando-se a data da última menstruação (DUM) sempre que a mãe se recorde da mesma. Porém, alguns autores questionam a acurácia desse indicador (Kramer et al., 1988; Goldemberg et al., 1989; Berg & Bracken, 1992; Reuss et al., 1995). A possibilidade de o sangramento no primeiro trimestre ser confundido com a última menstruação e a ocorrência de ciclos longos e/ou irregulares, ampliando as diferenças entre idade da gestação e idade da concepção, são alguns dos motivos que dão base a esse questionamento. Ainda assim, a Organização Mundial de Saúde (OMS) recomenda a data da última menstruação como o método de escolha para o cálculo da idade gestacional, tanto para fins clínicos, como para as pesquisas na área (Alexander et al., 1990).

Nos últimos anos, observa-se uma crescente indicação e utilização da ultra-sonografia (USG) para a estimativa da data provável do parto e, conseqüentemente, da idade gestacional do recém-nascido. Apesar dessa tendência, tanto o *American College of Obstetricians and Gynecologists*, como o *National Institute of Health Task Force on Diagnostic Ultrasound Imaging in Pregnancy* ainda evitam sua recomendação na rotina do pré-natal. Tal ressalva é conseqüência da escassez de evidências que indiquem uma melhora da saúde materno-infantil após sua introdução como exame de rotina (Berg & Bracken, 1992; ACOG, 1993, *apud* Reuss et al., 1995). O elevado custo também tem sido um empecilho para sua utilização em grande escala, principalmente em países em desenvolvimento.

Em situações nas quais a validade das informações sobre a DUM seja questionável e a mãe não tenha realizado USG no primeiro ou segundo trimestre da gestação, a idade gestacional pode ser estimada por meio da observação de características físicas e neurológicas do

recém-nascido em seus primeiros dias de vida (Dubowitz et al., 1970; Capurro et al., 1978; Ballard et al., 1979; Ballard et al., 1991). No entanto, a utilização do exame clínico da criança para estimativa da idade gestacional também tem sido alvo de críticas ao longo da última década. A inadequação da maioria dos escores propostos para a avaliação de crianças em idades gestacionais extremas e/ou com baixo peso ao nascer é apontada em diversos estudos (Vogt et al., 1981; Spinnato et al., 1984; Constantine et al., 1987; Alexander et al., 1992; Gagliardi et al., 1992).

Na tentativa de contornar essas limitações, Ballard et al. (1991) propuseram o escore *New Ballard* (NB) como uma alternativa aos anteriores. A possibilidade de avaliar prematuros extremos, assim como a relativa simplicidade de execução do novo escore foram os principais motivos que incentivaram uma crescente utilização deste em diversas partes do mundo. No Brasil, essas características do NB foram responsáveis por sua recente recomendação pela Sociedade Brasileira de Pediatria para avaliação neonatal e por sua crescente utilização nas instituições de referência como procedimento de escolha para a avaliação pós-natal da idade gestacional do recém-nascido.

Porém, apesar disso, observa-se uma escassez de estudos que discutam a validade desse escore. Uma ampla revisão bibliográfica identificou apenas quatro estudos com esse propósito. A investigação que propôs o escore precursor do NB analisou a validade da primeira versão do escore, comparando-a com um escore clínico anterior e não com procedimentos considerados de referência, a saber, DUM e USG (Ballard et al., 1979).

Problemas também foram evidenciados no estudo seguinte, responsável pela divulgação do escore NB (Ballard et al., 1991). Nessa pesquisa, os autores utilizaram estimadores de concordância que não são habitualmente empregados para estudos de validade, especificamente, o coeficiente de correlação de Pearson e medidas de concordância simples que desconsideram a possibilidade do acaso (Bland & Altman, 1986).

Os outros dois estudos apenas avaliam o escore precursor do NB, sendo um deles restrito à avaliação de recém-nascidos com baixo peso ao nascer (Constantine et al., 1987; Alexander et al., 1992). A relativa incipiência desse programa de investigação é, possivelmente, responsável pela existência de muitas dúvidas a respeito da validade do escore. O desempenho deste no diagnóstico das diferentes faixas de idade gestacional, o melhor momento para

a realização do exame, as possíveis influências de determinadas patologias maternas gestacionais e o papel do peso ao nascimento na acurácia de suas estimativas são algumas das questões que permanecem em aberto.

Diante da insuficiência de informações e, considerando a tendência atual do progressivo uso do escore NB em nosso país, optou-se por discutir novamente alguns aspectos da validade concorrente do procedimento, desta feita, utilizando métodos recomendados pela literatura.

O objetivo deste artigo é avaliar a acurácia do escore NB para o diagnóstico de prematuridade, discutindo as repercussões dos achados, tanto no contexto clínico, como no epidemiológico. Dada a importância de identificar problemas nos processos de aferição da pesquisa, foi realizada uma avaliação da confiabilidade interobservador da aplicação do escore e uma avaliação da qualidade dos laudos dos exames ultra-sonográficos no início do período de coleta de dados do estudo principal.

## Método

### Questões teórico-conceituais e definição de variáveis e indicadores

O escore NB foi proposto em 1991, a partir da ampliação de escore anterior (Ballard et al., 1979), visando à incorporação de características de recém-nascidos com idade gestacional inferior a 28 semanas. Como uma alternativa simplificada ao escore de Dubowitz (Dubowitz et al., 1970), o NB é formado por 12 itens que avaliam características somáticas e neurológicas da criança em seus primeiros dias de vida. O componente somático envolve a observação da pele, lanugo, superfície plantar, auréolas mamárias, olhos/orelhas e órgão genital. No componente neurológico, examina-se o recém-nascido quanto à postura, ângulo do punho, recuo do braço, ângulo poplíteo, flexibilidade da articulação acrômio-clavicular (sinal do cachecol) e grau de resistência à extensão da musculatura de membros inferiores (manobra calcanhar-orelha). A idade gestacional é subsequentemente calculada com base na conversão do total de pontos atribuídos aos itens, em semanas de gestação, segundo um sistema de correspondência pré-estabelecido (Ballard et al., 1991).

A estimativa da idade gestacional através da DUM é a forma mais simples e antiga de se calcular a duração da gestação. Através da regra de Naegele, subtraem-se três meses e somam-se sete dias ao primeiro dia da última

menstruação para o cálculo da data provável do parto. Neste estudo, a duração da gestação pela DUM foi calculada valendo-se do algoritmo  $(DN - M)/7$ , sendo DN a data de nascimento da criança e M o primeiro dia da última menstruação. Como recomenda a Organização Mundial de Saúde, apenas as semanas completas foram consideradas (OMS, 1995).

As comparações que utilizaram informações sobre a DUM consideraram tanto informações obtidas através da revisão do prontuário médico ( $DUM_p$ ), quanto as obtidas mediante entrevista com a mãe ( $DUM_E$ ). A análise de dados também contemplou a apreciação de um grupo particular de mães que apresentou diferença entre a idade gestacional estimada pela DUM e pela USG menor ou igual a duas semanas. Essas mulheres foram qualificadas como tendo a DUM confirmada pela USG ( $DUM_{conf1}$  e  $DUM_{conf2}$  se envolvendo, respectivamente, a DUM oriunda da entrevista e a do prontuário).

As mães que apresentaram ciclos regulares, com período de duração entre 25 e 31 dias, e que não relataram modificação das características da última menstruação com relação às anteriores, foram denominadas mães regulares ( $DUM_{ER}$  e  $DUM_{PR}$  se relacionada, respectivamente, à DUM da entrevista e à do prontuário). Esse recorte visou acompanhar a literatura que aventa a possibilidade de problemas de estimativa da idade gestacional pela DUM (Kramer et al., 1988; Alexander et al., 1990; Berg, 1991; Berg & Bracken, 1992). A retirada de mulheres com ciclos longos, irregulares e com sangramento no primeiro trimestre de gestação evitaria um erro de estimativa.

A idade gestacional estimada a partir da USG se baseou nos laudos de exames realizados durante o pré-natal. Atendendo às recomendações da literatura de referência, apenas os exames realizados entre a 7ª e a 18ª semana de gestação foram utilizados no estudo (Campbell et al., 1985; Rossavik & Fishburne, 1989; Berg & Bracken, 1992). A delimitação do período da gestação para a realização da USG leva em conta que a precisão da estimativa da idade gestacional diminui à medida que a gravidez avança.

Um outro subgrupo contemplado na análise refere-se às mulheres com convergência entre as idades gestacionais estimadas pelos diferentes procedimentos. Foram qualificadas como ultra-sonografias-padrão ( $USG_{padrão}$ ) aquelas que apresentavam diferenças iguais ou inferiores a duas semanas entre as estimativas da USG e da DUM (prontuário e entrevista).

### **Desenho do estudo, coleta e análise de dados**

A coleta das informações de interesse foi realizada, pelo primeiro autor, entre dezembro de 1998 e março de 1999 na maternidade do Instituto Fernandes Figueira (IFF/Fiocruz), localizada na cidade do Rio de Janeiro. A coleta de dados sucedeu a uma etapa de treinamento e pré-teste realizados no serviço. Para a obtenção de informações do prontuário e entrevista com a mãe, foram utilizados três formulários estruturados, particularmente elaborados para a pesquisa. O primeiro contemplou informações relacionadas ao pré-natal obtidas mediante consulta ao prontuário médico da parturiente. No segundo, anotaram-se dados sobre a gestação coletados diretamente com as mães. No último, registraram-se informações a respeito tanto do parto e do recém-nascido contidas no prontuário neonatal, quanto da avaliação clínica de maturidade do recém-nascido.

- **Estudo de confiabilidade do processo de aplicação do escore New Ballard**

Essa etapa envolveu a avaliação de 52 crianças (cerca de 25% da amostra) por duas examinadoras. Os exames foram realizados por CLM e replicados por uma neonatologista do IFF com vasta experiência na utilização do escore. Os dois exames foram realizados de forma independente e mascarada em recém-nascidos com menos de 72 horas de vida, com intervalo máximo de quatro horas entre as duas aferições.

Foram analisadas a confiabilidade de cada item do escore NB, de cada um dos componentes do mesmo (somático e neurológico) e do escore total. Além disso, foi analisada a confiabilidade entre as duas examinadoras com relação ao diagnóstico de prematuridade.

- **Avaliação da qualidade da idade gestacional baseada na ultra-sonografia**

Três estratégias foram utilizadas visando avaliar a qualidade dos laudos das USGs. Na primeira, avaliou-se a adequação da escolha das estruturas fetais mensuradas para fundamentar a estimativa da idade gestacional, segundo as recomendações de Hadlock et al. (1984), Seeds (1996) e Wagner & Calhoun (1998). A segunda estratégia consistiu, de forma mascarada e independente, em avaliar a equivalência dos diagnósticos das idades gestacionais oriundos dos laudos originais e dos de um examinador externo com reconhecida experiência na realização do exame. Para essa análise, todos

os laudos disponíveis (cerca de 80% do total de exames) foram refeitos utilizando-se as tabelas de Hadlock et al. (1984), Merz et al. (1987) e Merz (1988). Ambas as avaliações contemplaram, em separado, os exames realizados no IFF e os dos outros serviços. Por fim, avaliaram-se os estimadores de acurácia do escore NB, contrastando-se os resultados obtidos por meio da utilização, como padrão-ouro, das USGs realizadas no IFF (putativamente consideradas de bom padrão) e as realizadas em outros serviços.

- **Avaliação da acurácia do escore NB para o diagnóstico de prematuridade**

Estudou-se a validade concorrente do escore NB (Streiner & Norman, 1995). Em virtude da falta de consenso na literatura com relação ao procedimento de referência (padrão-ouro) para a estimativa da idade gestacional, avaliou-se, inicialmente, o grau de concordância entre as estimadas pela USG e DUM de todas as mulheres, bem como apenas das mães regulares. Da mesma forma, estimou-se a concordância entre esses procedimentos quanto ao diagnóstico de prematuridade. Uma excelente concordância nos dados permitiria optar por apenas um dos procedimentos como referência. Como isso não foi observado (ver resultados), decidiu-se pela comparação do NB com mais de um procedimento.

A análise que assumiu a DUM como padrão-ouro considerou todas as mulheres com a informação ( $DUM_E$  e  $DUM_P$ ), bem como as que tinham sua DUM confirmada pela USG ( $DUM_{conf1}$  e  $DUM_{conf2}$ ). Quando a USG foi utilizada como padrão-ouro, os indicadores foram estimados segundo o local de realização do exame. Além disso, avaliou-se a acurácia da escala considerando-se apenas as mães que apresentavam convergência entre as idades gestacionais estimadas pela DUM e USG ( $USG_{padrão}$ ).

Excluíram-se os recém-nascidos com malformações congênicas ou patologias graves que impossibilitassem a aplicação do escore, e/ou com mais de 72 horas de vida no momento do exame clínico, e/ou quando não havia informação sobre a DUM e USG. Foram também excluídos três recém-nascidos em razão da incompatibilidade entre as idades gestacionais e os respectivos quadros clínicos. O tamanho amostral efetivo para a estimativa da acurácia do escore variou conforme a estratégia utilizada como padrão-ouro (ver tabelas na seção de resultados).

## • Análise de dados

Para a avaliação da confiabilidade interobservador da escala NB e dos laudos ultra-sonográficos, utilizou-se o coeficiente de correlação intraclasse (*Intraclass Correlation Coefficient – ICC*). Para o cálculo do ICC foi utilizado o escore  $\alpha$  e o modelo que assume o caráter não aleatório dos examinadores (*two-way mixed effect model*) (Bartko, 1966, 1976; Shrout & Fleiss, 1979). O coeficiente *kappa* (Cohen, 1960; Fleiss et al., 1969) foi utilizado para a avaliação da equivalência entre as examinadoras, a respeito do diagnóstico de prematuridade. Para efeito de interpretação, usou-se a classificação sugerida por Landis & Koch (1977) que julga uma concordância excelente, se o *kappa* for acima de 0,75, aceitável, se entre 0,40 e 0,75, e inadequada, se abaixo de 0,40.

A acurácia do escore foi estimada por intermédio dos indicadores de sensibilidade e especificidade. A sensibilidade é definida como a capacidade de o teste diagnóstico identificar os verdadeiros positivos, e a especificidade, como a capacidade de o teste diagnóstico identificar os verdadeiros negativos (Fletcher et al., 1996). Os intervalos de confiança (IC) consideraram um nível de confiança de 95%.

## Resultados

### Descrição da população estudada

A média de horas de vida no momento da realização do exame foi de trinta horas, com 80% das crianças encontrando-se entre a 7ª e a 48ª hora após o nascimento. Cerca de metade dos partos foram normais (51%), e 27% das mães apresentaram alguma intercorrência na gravi-

dez potencialmente modificadora do grau de maturação do recém-nascido. Das mães que participaram do estudo, 63% haviam realizado USG no período gestacional preconizado, 75% se recordava da DUM no momento da entrevista e tinham a mesma registrada no prontuário. Das mães que informaram a DUM, 59% puderam ser caracterizadas como mães regulares.

Quanto aos recém-nascidos, aproximadamente um quarto (24,5%) dos que tinham seus pesos ao nascimento registrados no prontuário (90%) apresentava baixo peso ao nascer (peso ao nascimento < 2.500g). A prematuridade foi detectada em 25% e 35%, segundo a DUM e a USG, respectivamente.

### Confiabilidade do processo de aplicação do escore New Ballard e laudos ultra-sonográficos

A Tabela 1 apresenta a confiabilidade interobservador de aplicação do escore *New Ballard*. Considerando-se todas as replicações para avaliar a equivalência entre as examinadoras quanto ao diagnóstico de prematuridade, obteve-se um coeficiente *kappa* de 0,74 (IC 95% 0,49-0,99). À análise usando a variável de forma contínua, encontra-se um ICC de 0,88 (IC 95% 0,78-0,93). Observa-se, consistentemente, uma maior confiabilidade do componente somático em relação ao neurológico. Uma exploração direta dos formulários mostrou que o quarto e o sexto itens do componente neurológico (avaliação do ângulo poplíteo e manobra calcânhar-orelha) apresentavam maior divergência de pontuação entre as examinadoras, colaborando para a sua menor confiabilidade. Os achados também sugerem um aumento da confiabilidade com o decorrer do trabalho de campo, quando se comparam as estimações

Tabela 1

Confiabilidade inter-observador do escore *New Ballard* (NB) e dos laudos ultra-sonográficos.

Indicador de confiabilidade	NB <sub>1</sub> (n = 26)			NB <sub>2</sub> (n = 52)			Laudos da USG (n = 81) total
	escore total	componente somático	componente neurológico	escore total	componente somático	componente neurológico	
Kappa (IC 95%)	0,43 (0,11-0,75)	– –	– –	0,74 (0,49-0,99)	– –	– –	– –
ICC (IC 95%)	0,77 (0,37-0,91)	0,79 (0,59-0,91)	0,67 (0,15-0,87)	0,88 (0,78-0,93)	0,89 (0,82-0,93)	0,71 (0,40-0,85)	0,96 (0,91-0,98)

\*NB<sub>1</sub>: Confiabilidade analisando as primeiras 31 replicações.

\*\*NB<sub>2</sub>: Confiabilidade analisando todas as replicações.

encontradas na metade do processo de replicação com as totais.

A confiabilidade dos laudos ultra-sonográficos também pode ser apreciada na Tabela 1. Ao se analisarem os exames em conjunto, o ICC foi 0,96 (0,91-0,98). Uma análise da equivalência entre os laudos segundo o local de realização do exame, não revelou uma diferença estatisticamente significativa entre os ICCs (dados não mostrados).

Observou-se que 93% das USGs realizadas no IFF e 90% das realizadas em outros serviços utilizaram estruturas fetais recomendadas pela literatura de referência em suas estimativas de idade gestacional. Oito exames não se adequavam a essas recomendações, uma vez que se baseavam em apenas uma estrutura fetal, encontrando-se a gestação após a 12ª semana.

#### **Concordância entre as idades gestacionais estimadas pelos diferentes procedimentos utilizados como padrão-ouro**

Conforme apresentado na Tabela 2, o coeficiente *kappa* se manteve entre 0,60 e 0,65 em todas as comparações que envolveram o diagnóstico de prematuridade pela USG e DUM. Nota-se um substancial aumento do coeficiente quando se avalia a concordância entre o diagnóstico baseado na DUM registrada no prontuário materno e o da entrevista com a mãe.

A concordância entre as idades gestacionais estimadas pela USG e DUM através do ICC é bem semelhante em todas as situações avaliadas, encontrando-se coeficientes entre 0,7 e 0,8. Ao se compararem apenas as idades gestacionais estimadas pelas duas fontes de infor-

mação sobre a DUM, observam-se valores mais altos, acima de 0,9. Uma análise contemplando somente o subgrupo das mães regulares não revelou alterações significativas das estimativas.

#### **Capacidade do escore New Ballard para o diagnóstico de prematuridade**

Na Tabela 3, encontram-se os indicadores de acurácia da escala NB para o diagnóstico de prematuridade. A sensibilidade variou entre 0,50 e 0,71. De modo geral, a capacidade de identificação de recém-nascidos prematuros foi menor quando se considerou a USG como padrão-ouro. A menor foi encontrada quando se utilizou como padrão a USG de pacientes que apresentavam convergência entre as idades gestacionais estimadas pela USG, DUM da entrevista e do prontuário (USG<sub>padrão</sub>). Os valores mais altos foram obtidos quando foram consideradas como referência as DUMs de mães que apresentavam convergência apenas entre uma das DUMs e a USG.

De forma diferente se comportou a especificidade do escore. Praticamente não há variação da medida de acordo com padrão-ouro estabelecido. Observa-se que houve uma oscilação entre 0,91 e 0,97 conforme o procedimento utilizado.

A Figura 1 apresenta o comportamento das diferenças entre a idade gestacional estimada pelo escore NB, DUM<sub>E</sub> e USG. Em geral, as estimadas pelo escore NB são maiores do que as estimadas pelos outros procedimentos. Porém, observa-se que, na maior parte dos contrastes, essa diferença não ultrapassa duas semanas de gestação. Ademais, nota-se que a diferença de-

Tabela 2

Equivalência entre as diferentes estratégias utilizadas como padrão-ouro para a estimativa da idade gestacional.

Indicador de confiabilidade	USG vs. DUM <sub>P</sub> (n = 93)	USG vs. DUM <sub>E</sub> (n = 91)	USG vs. DUM <sub>ER</sub> (n = 55)	USG vs. DUM <sub>PR</sub> (n = 57)	DUM <sub>P</sub> vs. DUM <sub>E</sub> (n = 132)	DUM <sub>P</sub> vs. DUM <sub>ER</sub> (n = 80)
Kappa (IC 95%)	0,65 (0,45-0,84)	0,65 (0,45-0,84)	0,63 (0,38-0,88)	0,60 (0,35-0,85)	0,95 (0,75-1,00)	1,0 (0,79-1,00)
ICC (IC 95%)	0,78 (0,68-0,85)	0,75 (0,64-0,83)	0,76 (0,62-0,85)	0,75 (0,61-0,84)	0,90 (0,86-0,93)	0,93 (0,89-0,95)

USG: ultra-sonografia.

DUM<sub>P</sub>: DUM retirada do prontuário.

DUM<sub>E</sub>: DUM obtida mediante entrevista com a mãe nas primeiras 72 horas após o parto.

DUM<sub>ER</sub>: DUM obtida mediante entrevista com as mães regulares nas primeiras 72 horas após o parto.

DUM<sub>PR</sub>: DUM retirada do prontuário das mães regulares.

Tabela 3

Validade do escore New Ballard (NB) para o diagnóstico de prematuridade de acordo com os diferentes padrões-ouro.

Indicador de acurácia do escore	NB vs. DUM <sub>p</sub> (n = 140)	NB vs. DUM <sub>E</sub> (n = 146)	NB vs. DUM <sub>conf1</sub> (n = 78)	NB vs. DUM <sub>conf2</sub> (n = 80)	NB vs. USG (n = 116)	NB vs. USG <sub>padrão</sub> (n = 72)
Sensibilidade (IC 95%)	0,68 (0,49-0,82)	0,67 (0,49-0,81)	0,69 (0,41-0,88)	0,71 (0,44-0,89)	0,57 (0,41-0,73)	0,50 (0,29-0,71)
Especificidade (IC 95%)	0,92 (0,85-0,96)	0,91 (0,83-0,95)	0,95 (0,85-0,99)	0,95 (0,86-0,99)	0,97 (0,90-0,99)	0,9 (0,85-0,99)

USG: ultra-sonografia.

DUM<sub>p</sub>: DUM retirada do prontuário.

DUM<sub>E</sub>: DUM obtida mediante entrevista com a mãe nas primeiras 72 horas após o parto.

DUM<sub>conf1</sub>: DUM de mães que têm a diferença entre as idades gestacionais estimadas pela DUM da entrevista e a USG  $\leq 2$  semanas.

DUM<sub>conf2</sub>: DUM de mães que têm a diferença entre as idades gestacionais estimadas pela DUM do prontuário e a USG  $\leq 2$  semanas.

USG<sub>padrão</sub>: idade gestacional estimada pela USG quando as diferenças entre esta, a DUM da entrevista e a do prontuário são  $\leq 2$  semanas.

crece conforme aumenta a idade gestacional, sendo positiva em crianças com idade inferior a quarenta semanas, e passando a ser negativa a partir daí. Tal comportamento é mais evidente quando são consideradas as diferenças entre o escore NB e a DUM<sub>E</sub>. O mesmo foi observado quando se considerou a DUM<sub>p</sub> (dados não mostrados).

## Discussão

### Confiabilidade da escala NB e laudos ultra-sonográficos

De forma geral, os resultados encontrados no estudo da confiabilidade do escore NB sugerem uma boa equivalência entre as idades gestacionais estimadas pelas duas examinadoras. Observa-se uma maior confiabilidade entre as aferições do componente somático em relação ao neurológico, bem como em relação ao escore total. Esses achados também são relatados em estudos anteriores e fundamentam a proposta de alguns autores de utilizar exclusivamente o escore somático para estimar a idade gestacional de recém-nascidos com baixo peso ao nascer (Finnstrom, 1972, 1977; Constantine et al., 1987). A maior confiabilidade do componente somático talvez se justifique pela maior objetividade dos itens que o compõem, por sua relativa facilidade de observação, sua menor relação com o grau de atividade da criança e com determinadas patologias maternas durante a gestação (Gould et al., 1977; Amiel-Tison, 1980).

Vale a pena enfatizar que os resultados da avaliação de confiabilidade do escore não devem ser indiscriminadamente extrapolados para outros contextos. A confiabilidade do processo de mensuração é determinada por fatores que envolvem não só o instrumento e o que está sendo mensurado, mas os próprios examinadores e as circunstâncias das aferições. Isso, em princípio, torna qualquer avaliação particular à situação analisada. É interessante destacar que, por se tratar de uma pesquisa e envolver somente examinadores submetidos a um rigoroso treinamento prévio e à vigilância sistemática, possivelmente a confiabilidade foi maior do que a que seria encontrada em situações de rotina que, em geral, não envolvem o mesmo rigor.

Conforme esperado, houve uma melhora da confiabilidade com o decorrer do trabalho de campo. O aumento dos coeficientes explica-se pelo aprimoramento da forma de abordagem dos pacientes com o andamento do estudo. Além disso, deve ter contribuído o ativo treinamento realizado durante o período de coleta de dados. Ressalta-se que, apesar de a confiabilidade ter sido um pouco menor durante o período inicial do estudo, optou-se por não retirar esse grupo de recém-nascidos do restante da análise, uma vez que sua influência nos resultados foi discreta, sua manutenção sobrepunhando amplamente a perda de eficiência que decorreria da sua retirada. Aceita a confiabilidade do processo, o restante dos exames pôde ser realizado, exclusivamente, pela primeira autora.

A confiabilidade dos laudos dos exames ultra-sonográficos envolve três principais etapas.

Figura 1

Relação entre a diferença da idade gestacional estimada pelo escore NB e as idades gestacionais estimadas pela USG e DUM, considerando apenas as mães que apresentaram convergência entre as idades gestacionais estimadas pela DUM do prontuário, DUM da entrevista e a USG.

Figura 1a

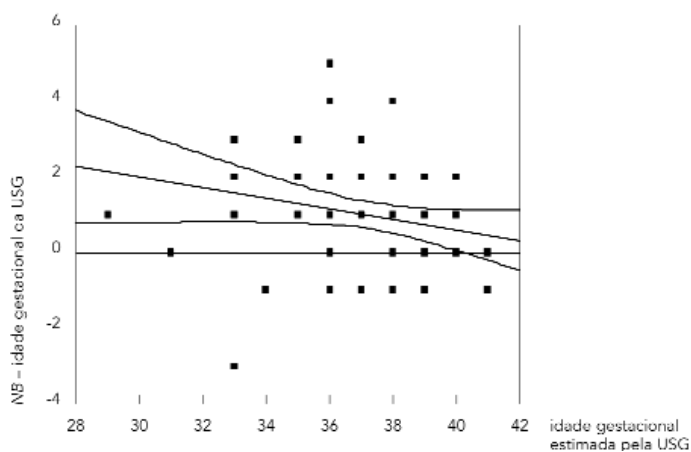
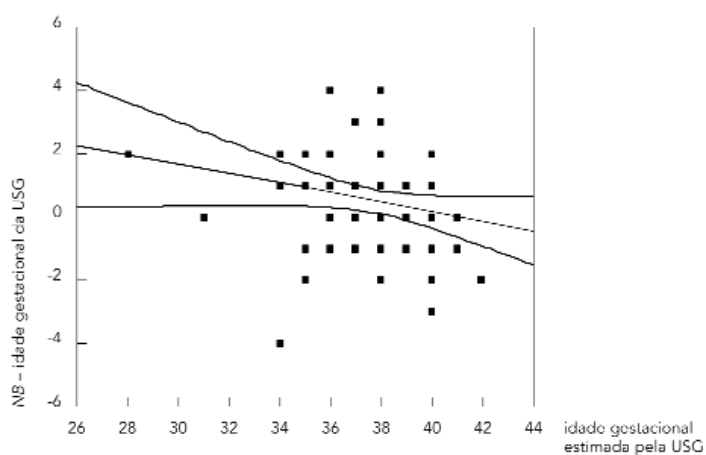


Figura 1b



A primeira diz respeito à identificação das estruturas fetais a serem mensuradas, a segunda é a própria mensuração das estruturas, e a terceira concerne à comparação dessas medidas com as esperadas para cada idade gestacional, de acordo com curvas-padrão. Neste estudo, em virtude da utilização de USGs realizadas em momentos anteriores à investigação, somente foi possível avaliar a primeira e a terceira etapas. Apesar disso, a excelente concordância observada entre os laudos originais e os do examinador externo foi considerada um indicador indireto da qualidade do processo de mensuração. A adequação das estruturas fetais mensuradas e a ausência de variação dos diferentes resultados da pesquisa segundo o local de realização do exame também são marcos para considerar satisfatória a qualidade das USGs.

#### Concordância entre as idades gestacionais estimadas pelos diferentes procedimentos de referência

Observou-se uma excelente concordância entre a identificação de recém-nascidos prematuros baseada na DUM registrada no prontuário médico e aquela baseada no recordatório da mãe no momento da entrevista. Reforçando estudos anteriores, os achados sugerem a pertinência da utilização da mãe como fonte de informação sobre eventos relacionados à gravidez, sua vida reprodutiva e à saúde perinatal (Little, 1986; Olson et al., 1997; Yawn et al., 1998).

A concordância entre os diagnósticos de prematuridade obtidos por meio da DUM e USG é inferior à que se observa quando são comparadas as diferentes fontes de DUM. Todavia, em todas as comparações realizadas, os coeficientes *kappa* se mantêm acima de 0,60.

A semelhança encontrada entre os coeficientes obtidos ao se considerar o conjunto de mães e aos restritos às mães regulares diverge da literatura. Estudos anteriores têm apontado diferença relacionada à presença de ciclos menstruais longos e/ou irregulares e a ocorrência de sangramento no primeiro trimestre de gestação (Kramer et al., 1988; Berg & Bracken, 1992). Isso sugere que as diferenças entre as idades gestacionais estimadas pela USG e pela DUM envolvem questões ainda a esclarecer, merecendo maior exploração.

#### Acurácia do escore NB para o diagnóstico e prematuridade

Apesar de a sensibilidade do escore ter variado conforme o padrão-ouro utilizado, sua estimativa de ponto foi inferior a 0,70 em cinco das



seis comparações realizadas (Tabela 3). Observou-se, também, que todos os intervalos de confiança incluíram valores inferiores a 0,50. Esses achados questionam a adequação do escore NB como ferramenta diagnóstica isolada para a identificação de prematuridade.

As menores sensibilidades foram obtidas quando se considerou a USG como padrão-ouro. A avaliação que considerou como referência apenas as USGs de mães que apresentaram convergência entre as idades gestacionais estimadas pela DUM do prontuário, DUM da entrevista e a USG foi a que gerou a menor sensibilidade. Esse achado é importante, já que, ao se analisar apenas esse grupo de mães, está se utilizando um padrão-ouro próximo do ideal. A menor sensibilidade encontrada nessa situação sublinha a fragilidade do escore para o diagnóstico de crianças prematuras.

A variação das diferenças entre as idades gestacionais estimadas pelo escore NB e pelos demais procedimentos em diferentes faixas de maturidade do recém-nascido também merece alguma reflexão. Essa questão foi descrita em estudos anteriores (Alexander et al., 1990, 1992), e os resultados sugerem que a utilização do escore NB superestima a idade gestacional em recém-nascidos com menor duração da gestação e subestima a idade gestacional em recém-nascidos mais maduros. Esse comportamento é compatível com as baixas sensibilidades encontradas na avaliação da acurácia do procedimento no diagnóstico de prematuridade.

Os achados impõem que, no âmbito clínico, deva-se ter muita cautela com a utilização exclusiva do escore NB para estimativa da idade gestacional, na medida em que a falha no diagnóstico de prematuridade pode gerar diversos problemas na assistência neonatal e naquela prestada nos primeiros anos de vida da criança. O erro diagnóstico acarretaria uma conduta clínica que, ao desobrigar, indevidamente, uma série de cuidados neonatais e de puericultura, dificultaria a prevenção, detecção precoce e tratamento de problemas de saúde comuns em crianças prematuras.

Também no âmbito de investigações epidemiológicas, a baixa sensibilidade do escore NB pode trazer dificuldades. Em estudos interessados em estimar a incidência de prematuridade, por exemplo, a utilização isolada do escore como instrumento diagnóstico levaria a uma importante subestimação da magnitude do fenômeno.

A má classificação também traria problemas às investigações com objetivo de estudar os fatores de risco para o nascimento prematuro. Em estudos caso-controle, por exemplo, a

utilização exclusiva do escore NB para a identificação de prematuridade faria com que muitas crianças prematuras fossem selecionadas apenas para o grupo controle, já que o diagnóstico de prematuridade seria perdido. A repercussão disso seria o rompimento da estrutura de comparabilidade, necessária à validade de comparação (Reichenheim & Moraes, 1999).

A título de ilustração, utilizando-se como exemplo os dados desta investigação e considerando a USG<sub>padrão</sub> como referência, 22 recém-nascidos seriam, de fato, prematuros (verdadeiros positivos). Entretanto, com o diagnóstico de prematuridade baseado apenas no escore NB, 50% das crianças (11 recém-nascidos) seriam, indevidamente, consideradas como não prematuras, deixando de ser captadas para o grupo de casos. Se aleatórias, essas perdas, por si só, não trariam problemas de consistência, somente de eficiência. A dificuldade maior ocorreria se os casos fossem alocados no grupo-controle. Nessa situação, adviria um viés de seleção considerável. Em se assumindo o caráter não aleatório da má-classificação, uma alternativa para evitar o erro na estimativa das medidas de efeito de interesse seria a exclusão do grupo-controle de crianças que tivessem suas idades gestacionais estimadas, unicamente, pelo escore NB. Dessa forma, apenas os resultados positivos do escore, indicando prematuridade, seriam considerados.

Vale sublinhar que os resultados se referem somente aos dados agregados. No entanto, seria interessante que estudos futuros, além de servirem de instâncias de replicação e possível corroboração dos achados, contemplassem também análises de subgrupos. Alguns autores têm sugerido que a acurácia do escore NB possa depender de certas condições como, por exemplo, o número de horas de vida no momento do exame ou intercorrências na gravidez (Ballard et al., 1979; Constantine et al., 1987; Ballard et al., 1991; Gagliardi et al., 1992). Neste estudo, observou-se que cerca de um quarto das puérperas apresentavam algum agravo potencialmente modificador do grau de maturação do recém-nascido, não sendo implausível que modificações de efeito possam ter ocorrido. Infelizmente, o presente tamanho amostral não permitiu que maiores detalhes fossem contemplados. Como indicado na introdução, essas questões ainda merecem aprofundamento.

Em que pese aos problemas já discutidos, em algumas situações o exame clínico do recém-nascido logo após seu nascimento é a única estratégia possível, tanto para pediatras, como para pesquisadores na área de saúde peri-

natal. Grande parcela das mulheres brasileiras ainda não tem acesso ao pré-natal ou o iniciam tardiamente, impossibilitando a realização da USG em períodos gestacionais nos quais sua capacidade diagnóstica é maior. Ademais, em algumas situações, a DUM é desconhecida. A utilização do exame clínico do recém-nascido, nessas circunstâncias, apresenta-se como a única ferramenta para a estimativa da idade gestacional da criança.

Se, de um lado, existem circunstâncias em que a utilização do escore é mandatória, por outro, as questões que emergem dos achados da pesquisa merecem ponderação. Ainda que necessitando de novas investigações, os resultados já servem, no mínimo, como um alerta às instâncias normativas e às instituições de ensino, pesquisa e assistência pediátricas: a baixa sensibilidade do escore NB para o diagnóstico de prematuridade, estimada utilizando-se vários procedimentos de referência, questiona a sua indicação como procedimento de escolha para estimativa da idade gestacional do recém-

nascido. Os resultados sugerem que o escore NB seja considerado apenas como uma ferramenta acessória para a identificação de crianças prematuras.

No momento em que se preconiza a utilização do escore em estudo para avaliação neonatal, é necessário qualificar a recomendação dele, desencorajando o seu uso exclusivo e enfatizando a necessidade de se obterem informações sobre a USG e/ou DUM para uma estimativa da idade gestacional mais acurada. Estes últimos parecem ainda ser os mais indicados, a despeito das limitações apontadas na introdução. Contudo, a escolha entre os dois também merece qualificação. Se a ampliação da cobertura da USG durante o pré-natal é ainda uma questão em debate em termos de custos e prioridades, este não é o caso da DUM. A recuperação desta informação, mediante revisão do prontuário e/ou com base na entrevista com a mãe, é uma providência simples, viável, podendo ser, incondicionalmente, incentivada.

## Agradecimentos

Michael E. Reichenheim foi apoiado pela FAPERJ (processos E-26/171.223/98 e E-26/150.893/99). Cláudia Leite Moraes foi apoiada pela ENSP-Fiocruz. Os autores agradecem à Profa. Maria Elisabeth Lopes Moreira, aos doutores Roberto Noya Galuzzo e Rita Rosana Teixeira Brás e a todos os membros do corpo clínico do Departamento de Neonatologia do Instituto Fernandes Figueira/FIOCRUZ, sem os quais não seria possível a realização da investigação. Agradecemos, ainda, a Alessandra Silva Dias, estagiária do Instituto de Medicina Social/UERJ, pela colaboração no processo de digitação de dados, e ao Prof. Eduardo Faerstein, pelos valiosos comentários feitos a respeito da primeira versão desse artigo.

## Referências

- ALEXANDER, G. R.; TOMPKINS, M. E. & CORNELLY, D. A., 1990. Gestational age reporting and preterm delivery. *Public Health Reports*, 105:267-275.
- ALEXANDER, G. R.; CAUNES, F.; HULSEY, T. C.; TOMPKINS, M. E. & ALLEN, A., 1992. Validity of postnatal assessments of gestational age: A comparison of the method of Ballard et al. and early ultrasonography. *American Journal of Obstetrics and Gynecology*, 166:891-895.
- AMIEL-TISON, C., 1980. Possible acceleration of neurological maturation following high-risk pregnancy. *American Journal of Obstetrics and Gynecology*, 138:303-306.
- BALLARD, J. L.; KHOURY, J. C.; WEDIG, K.; WANG, L.; EILERS-WALSMAN, B. L. & LIPP, R., 1991. *New Ballard score*, expanded to include extremely premature infants. *Journal of Pediatrics*, 119:417-423.

- BALLARD, J. L.; NOVAK, K. K. & DRIVER, M., 1979. A simplified score for assessment of fetal maturation of newly born infants. *Journal of Pediatrics*, 95:769-774.
- BARTKO, J. J., 1966. The intraclass correlation coefficient as a measure of reliability. *Psychological Reports*, 19:3-11.
- BARTKO, J. J., 1976. On various intraclass correlation reliability coefficients. *Psychological Bulletin*, 83: 762-765.
- BERG, A. T. & BRACKEN, M. B., 1992. Measuring gestational age: An uncertain proposition. *British Journal of Obstetrics and Gynaecology*, 99:280-282.
- BERG, A. T., 1991. Menstrual cycle length and the calculation of gestational age. *American Journal of Epidemiology*, 133:585-589.
- BLAND, J. M. & ALTMAN, D. G., 1986. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*, 1:307-310.
- CAMPBELL, S.; WARSOFF, S. L.; LITTLE, D. & COOPER, D. J., 1985. Routine ultrasound screening for the prediction of gestational age. *Obstetrics & Gynecology*, 65:613-620.
- CAPURRO, H.; KONICHEZKY, S.; FONSECA, D. & CALDEYRO-BARCIA, R., 1978. A simplified method for diagnosis of gestational age in newborn infant. *Journal of Pediatrics*, 93:120-122.
- COHEN, J., 1960. A coefficient of agreement for nominal scale. *Educational and Psychological Necessities*, 20:37-46.
- CONSTANTINE, N. A.; KRAEMER, H. C.; KENDALL-TACKETT, K. A.; BENNETT, F. C.; TYSON, J. E. & GROSS, R., 1987. Use of physical and neurologic observations in assessment of gestational age in low birth weight infants. *Journal of Pediatrics*, 110:921-928.
- DUBOWITZ, L. M. S.; DUBOWITZ, V. & GOLDBERG, C., 1970. Clinical assessment of gestational age in the newborn infants. *Journal of Pediatrics*, 77:1-10.
- DUNN, P. M., 1985. The search for perinatal definitions and standards. *Acta Paediatrica Scandinavica*, 319(Sup.1):7-16.
- FINNSTROM, O., 1972. Studies on maturity in newborn infants. II. External characteristics. *Acta Paediatrica Scandinavica*, 61:24-32.
- FINNSTROM, O., 1977. Studies on maturity in newborn infants. IX. Further observations on the use of external characteristics in estimating gestational age. *Acta Paediatrica Scandinavica*, 66:601-604.
- FLEISS, J. L.; COHEN, J. & EVERITT, B. S., 1969. Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, 72:332-337.
- FLETCHER, R. H.; FLETCHER, S. W. & WAGNER, E. H., 1996. *Clinical Epidemiology: The Essentials*. 3rd Ed. Baltimore: Williams & Wilkins.
- FORTIER, I.; MARCOUX, S. & BRISSON, J., 1995. Maternal work during pregnancy and the risks of delivering a small-for-gestational-age or preterm infant. *Scandinavian Journal of Environmental Health*, 21:412-418.
- GAGLIARDI, L.; SCIMONE, F.; DEL PRETE, A.; PETECA, C.; STIVAL, G.; PASINETTI, G.; TEANI, M.; BIANCHI, P.; MARANI, M. & DUCA, P., 1992. Precision of gestational age assessment in neonate. *Acta Paediatrica*, 81:95-99.
- GOLDEMBERG, R. L.; DAVIS, R. O.; CUTTER, G. R.; HOFFMAN, M. A.; BRUMFIELD, C. G. & FOSTER, J. M., 1989. Prematurity, postdates and growth retardation: The influence of use of ultrasonography on reported gestational age. *American Journal of Obstetrics and Gynecology*, 160:462-470.
- GOULD, J. B.; GLUCK, L. & KULOVICH, M. V., 1977. The relationship between accelerated pulmonary maturity and accelerated neurological maturity in certain chronically stressed pregnancies. *American Journal of Obstetrics and Gynecology*, 127:181-186.
- HADLOCK, F. P.; DETER, R. L.; HARIST, R. B. & PARK, S. K., 1984. Estimating fetal age: Computed assisted analysis of multiple fetal growth parameters. *Radiology*, 152:497-501.
- HEDEGAARD, M.; HENRIKSEN, T. B.; SECHER, N. J.; HATCH, M. C. & SABROE, S., 1996. Do stressful life events affect duration of gestation and risk of preterm delivery? *Epidemiology*, 7:339-345.
- HENRIKSEN, T. B.; WILCOX, A. J.; HEDEGAARD, M. & SECHER, N. J., 1995. Bias in studies of preterm and postterm delivery due to ultrasound assessment of gestational age. *Epidemiology*, 6:533-537.
- KRAMER, M. S., 1987. Determinants of low birth weight: Methodological assessment and meta-analysis. *Bulletin of the World Health Organization*, 65:663-737.
- KRAMER, M. S.; McLEAN, F. H.; BOYD, M. E.; USHER, R. H., 1988. The validity of gestational age estimation by menstrual dating in term, preterm, and postterm gestations. *JAMA*, 260:3306-3308.
- LANDIS, J. R. & KOCH, G. G., 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33:159-174.
- LANG, J. M.; LIEBERMAN, E. & COHEN, A., 1996. A comparison of risk factors for preterm labor and term small-for-gestational-age birth. *Epidemiology*, 7:369-376.
- LITTLE, R. E., 1986. Birthweight and gestational age: Mother's estimates compared with state and hospital records. *American Journal of Public Health*, 76:1350-1351.
- MAVALANKAR, D. V.; GRAY, R. H. & TRIVEDI, C. R., 1992. Risk factors for preterm and term low birthweight in Ahmedabad, India. *International Journal of Epidemiology*, 21:263-272.
- MERZ, E., 1988. *Sonographische Diagnostik in Gynakologie und Geburtshilfe: Lehrbuch und atlas*. New York: Georg Thieme.
- MERZ, E.; KIM-KERN, M. S. & PEHL, S., 1987. Ultrasonic mensuration of fetal limb bones in the second and third trimesters. *Journal of Clinical Ultrasound*, 15:175-183.
- OLSON, J. E.; SHU, X. O.; ROSS, J. A.; PENDERGRASS, T. & ROBISON, L. L., 1997. Medical record validation of maternally reported birth characteristics and pregnancy-related events: A report from children's cancer group. *American Journal of Epidemiology*, 145:58-67.
- OMS (Organização Mundial da Saúde), 1995. *Classificação Estatística Internacional de Doenças e Problemas Relacionados à Saúde. Décima Revisão*. São Paulo: Centro Colaborador da OMS para Classificação de Doenças em Português, Faculdade de Saúde Pública, Universidade de São Paulo/EDUSP.

- REICHENHEIM, M. E. & MORAES, C. L., 1998. Alguns pilares para apreciação da validade em estudos epidemiológicos. *Revista Brasileira de Epidemiologia*, 1:131-148.
- REUSS, M. L.; HATCH, M. C. & SUSSE, M., 1995. Early ultrasound dating of pregnancy: Selection and measurement errors. *Journal of Clinical Epidemiology*, 48:667-674.
- REUSS, M. L.; HATCH, M. C. & SUSSE, M., 1995. Early ultrasound dating of pregnancy: Selection and measurement biases. *Journal of Clinical Epidemiology*, 48:667-674.
- ROSSAVICK, L. K. & FISHBURNE, J. I., 1989. Conceptional age, menstrual age, and ultrasound age: A second trimester comparison of pregnancies of known conceptional date with pregnancies dated from the last menstrual period. *Obstetrics & Gynecology*, 73:243-249.
- SEEDS, J. W., 1996. The routine or screening obstetrical ultrasound examination. *Clinical Obstetrics and Gynecology*, 39:814-830.
- SHROUT, P. E. & FLEISS, J. L., 1979. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86:420-428.
- SPINNATO, J. A.; SIBAI, B. M.; SHAVER, D. C. & ANDERSON, G. D., 1984. Inaccuracy of Dubowitz gestational age in low birth weight infants. *Obstetrics & Gynecology*, 63:491-495.
- STREINER, D. L. & NORMAN, G. R., 1995. *Health Measurement Scales: A Practical Guide to their Development and Use*. New York: Oxford University Press.
- VOGT, H.; HANSBERG, B.; FINNE, P. H. & STENSBERG, A., 1981. Clinical assessment of gestational age in the newborn infant. An evaluation of two methods. *Acta Paediatrica Scandinavica*, 70:669-672.
- WAGNER, R. K. & CALHOUN, B. C., 1998. The routine obstetric ultrasound examination. *Obstetrics and Gynecology Clinics of North America*, 25: 451-463.
- YAWN, B. P.; SUMAN, V. J. & JACOBSEN, S. J., 1998. Maternal recall of distant pregnancy events. *Journal of Clinical Epidemiology*, 51:399-405.