Acta Obstetricia et Gynecologica. 2006; 85: 1276–1279

Taylor & Francis
Taylor & Francis Group

LETTER TO THE EDITOR

Regarding "Term prediction with ultrasound: evaluation of a new dating curve for biparietal diameter"

Sir,

In their article "Term prediction with ultrasound: evaluation of a new dating curve for biparietal diameter" (1), Backe and Nakling present a validation of a recently published dating method, "Terminhjulet" (Method B) (2), and an old method, "Snurra" (Method A) (3); the latter has been in use since 1984 in all Norwegian departments to assess gestational age and predict the expected day of delivery.

The authors compare the two methods by computing the *mean* difference between the observed and expected day of delivery, which is presented in their Table II, and in Figure 2 for individual biparietal diameter (BPD) values. This shows "Terminhjulet" as having a mean residual of −0.7 days and "Snurra" −3.5 days. The authors conclude that "the underestimation of fetal age by the BPD dating curves used in Norway for the last 20 years may lead to wrong clinical decisions, and the new reference values should be used". Their conclusion is incorrect and is caused by wrongful use of the *mean* as a statistical tool to evaluate the data. Additionally, several other factors deserve comments.

The distribution of the duration of the pregnancy for the human fetus is highly skewed with a long left tail of preterm births, mainly caused by pathology. When evaluating a system that is designed to predict term in normal pregnancies, as these two methods are, one would prefer to exclude the pathological cases. However, because there are no good, independent measures of such pathology, the cases cannot easily be identified and excluded from the evaluation. One must consequently choose measures of performance that are reasonably insensitive to abnormal cases or outliers (gross errors in data). The mean is highly sensitive to observations in the tails of the distribution, and particularly so for skewed distributions. The pathological, early births will draw the "true" mean residual in a negative direction, yet their presence has no relevance for the predictive capacity of the method evaluated. Analyses we have done on a dataset of approximately 50,000 ultrasound scans from Trondheim, Norway, show that the leftmost 6% of the residual distribution account for a change in the mean of 2 days. In addition, residuals are also shifted in a negative direction by the inductions for post-term, which artificially shortens the duration of the pregnancy. The authors' unjustified exclusion of all post-term inductions adds to the problem rather than diminishing it. Thus, a mean residual as close as possible to zero does not constitute a proof of soundness. On the contrary, it indicates that the method predicts a too early term.

The *median* is a robust parameter for the evaluation of skewed distribution such as the birth distribution (4,5). It is far less sensitive to the pathological processes at the extreme range of the curve, such as the pathological preterm births. Indeed, also the inductions post-term can be managed statistically using the median, without introducing a bias. In their article, the authors should have focused on the median as the measure of goodness. In fact, the authors present only the overall median residual, and when computing it they have not used an appropriate method for rounded data, making their median comparison imprecise and possibly biased. The median should have been presented with decimals. Finally, the use of mean values in the important Figure 2 in Backe and Nakling's paper gives a completely misleading impression.

As an example of the difficulties in interpreting the mean, the authors' statement about inductions in the first paragraph of the discussion is incorrect: "This selective exclusion of cases with long duration will bias the comparison in favor of method A. Despite the inherent bias, method B has a significantly smaller mean prediction error than method A . . .". In fact, the opposite is more likely: from the medians given in Table II it can be seen that *inclusion* of the post-term inductions with their large positive values would improve the median of "Snurra" (method A) compared to the median of "Terminhjulet" (method B). It is not possible to know the precise effect on the means, but it is likely to be in the same direction. The exact values of the post-term inductions will not influence the median, in contrast to the mean, and

(Received 25 April 2006; accepted 24 April 2006)

ISSN 0001-6349 print/ISSN 1600-0412 online
DOI: 10.1080/00016340600839668

the post-term inductions must be included to avoid a negative bias in the results.

To add to the confusion, only pregnancies with "reliable" last menstrual period (LMP) have been selected to be evaluated in Backe and Nakling's paper. While this seems reasonable when used for LMP dating, it is most important to evaluate ultrasound dating precisely when LMP is *not* reliable, because in a population setting ultrasound is, indeed, used on almost all pregnancies.

The authors cite the paper by Kiserud and Rasmussen (6) to support their conclusion that "Snurra" is biased for low values of the BPD. They fail to mention, however, that the same paper shows that the *median* prediction error of "Snurra" is zero in the BPD range 45–48 mm, and barely above 1 day in the range 41–50 mm, based on a study of 8,029 pregnancies in Bergen, Norway.

The authors generalize and state that "It is likely that some of the older dating formulas were developed with inadequate statistical methods." The author is led to think that this was also the case for "Snurra". Interestingly, however, "Terminhjulet" uses essentially the same statistical method as was used for "Snurra", namely polynomial regressions. This is in spite of the fact that more appropriate methods are available today, such as nonlinear quantile regressions (7). It should be noted, however, that 25 years ago, when the first prediction methods were developed, reference materials did not usually include observations from early in the pregnancy, and were therefore not designed nor intended to be used in this range. "Snurra" *has* a problem in the lower range, as has been previously stated (8), but not to the extent claimed (1,9). For practical use, "Snurra" has been defined to be applied for BPD in the range between 38 and 60 mm and it works well in this range (5,8).

We are surprised that the authors are willing to accept the high post-term rate that would be the consequence of "Terminhjulet", namely a jump from 6.1% based on "Snurra" and today's clinical practice, to 12.0%, even though the population itself would remain unchanged. Needless to say, this would lead to a dramatic change in clinical practice with unforeseen consequences.

It should be added that this difference between methods is not due to post-term inductions being based on "Snurra". As would be expected, almost all post-term inductions take place at 294 days or later, according to "Snurra". If they were included, they would be defined as post-term by "Terminhjulet" as well. If they are excluded, as they are in this study, they reduce post-term percentages with an equal amount for both methods.

Related to the above, if "Terminhjulet" were to be used, 83.6% of the births would fall within days 259 and 293. With "Snurra", however, as many as 88.5% of the fetuses are born within days 259 and 293. We are surprised that the authors have not commented on this fact. "Terminhjulet" performs almost as poorly at the LMP method (80.5% within days 259 and 293) on the pregnant population from Oppland County in Norway.

The final statement by the authors, that "the underestimation of fetal age by the BPD dating curves used in Norway for the last 20 years may lead to wrong clinical decisions" is incorrect and based upon the inappropriate use of the mean as a statistical tool to evaluate the data, and not upon the method itself. It is a serious allegation based on wrongful use of statistics and should be retracted.

"Snurra" has managed to standardize ultrasound pregnancy dating in all of Norway over the last 20 years. A unified treatment regime *is* important and any method aiming at succeeding "Snurra" should involve a significant improvement both of the statistical methods used and of the size and representativeness of the reference material. We agree that any prediction method should be open for evaluation, but there is an equal burden of responsibility to use appropriate methods in the evaluation as there is in the construction of the prediction method itself.

The traditional polynomial regressions used in "Snurra" and "Terminhjulet" are becoming obsolete in constructing models in obstetrics. New methods using the large population-based data now available must form the future tools to predict gestational age and expected day of delivery. Such methods have recently been developed (10).

References

1. Backe B, Nakling J. Term prediction with ultrasound: evaluation of a new dating curve for biparietal diameter measurements. Acta Obstet Gynecol Scand. 2006;85:156–9.
2. Johnsen SL, Rasmussen S, Sollien R, Kiserud T. Fetal age assessment based on ultrasound head biometry and the effect of maternal and fetal factors. Acta Obstet Gynecol Scand. 2004;83:716–23.
3. Eik-Nes SH, Grøttum P. Graviditetskalenderen "Snurra". Drammen, Norway: Scan-Med AS; 1984.
4. Hampel FR, Ronchetti EM, Rousseeuw PJ, Stahel WA. Robust Statistics: The Approach Based on Influence Functions. New York: Wiley; 1986.
5. Tunón K, Eik-Nes SH, Grøttum P. A comparison between ultrasound and a reliable last menstrual period as predictors of the day of delivery in 15,000 examinations. Ultrasound Obstet Gynecol. 1996;8:178–85.
6. Kiserud T, Rasmussen S. Terminbestemmelse ved hjelp av ultralyd – kan metoden bli bedre? Tidskr Nor Lægeforen. 1999;119:4331–4.

7. Yu K, Jones MC. Local linear quantile regression. J Am Stat Assoc. 1998;93:228–37.
8. Tunón K, Eik-Nes SH, Grøttum P. The impact of fetal, maternal and external factors on prediction of the day of delivery by the use of ultrasound. Ultrasound Obstet Gynecol. 1998;11:99–103.
9. Kiserud T. Svangerskapsvarighet, fosteralder og fødselsvekt. In: Bjugn R, Erichsen A, Vege Å, editors. Veileder ved obduksjon av foster og barn. Oslo: Den norske lægeforening; 2004. p. 47–57.
10. Eik-Nes SH, Blaas HG, Grøttum P, Gjessing H. Predicting remaining time of pregnancy: a new approach to the prediction of day of delivery. 15th World Congress on Ultrasound in Obstetrics and Gynecology. Book of Abstracts. Ultrasound Obstet Gynecol. 2005;26:341.

Sturla H. Eik-Nes
Per Grøttum
Håkon Gjessing

Address for correspondence
Sturla H. Eik-Nes
National Center for Fetal Medicine
St. Olav's Hospital
N-7006 Trondheim
Norway
E-mail: sturla.eik-nes@ntnu.no

REPLY

Sir,

We appreciate the interest in our recent paper (1) taken by Eik-Nes, Grøttum and Gjessing.

We investigated the precision of term prediction with ultrasound in pregnant women with regular periods. We compared two sets of normal values for BPD (biparietal diameter): the values issued by Eik-Nes and Grøttum (2) (method A) and the new values (3) developed by Johnsen, Rasmussen, and Kiserud (method B). We calculated the difference between observed and predicted day of delivery according to method A and method B. The mean, median, and standard deviations are reported as well as the percentage of preterm, term, and post-term deliveries. Also, the distribution curves are presented. Thus, we have used all the common approaches to describe the distribution of deliveries in a detailed and comprehensive manner.

We conclude that method B is better than method A (1). This is based on several findings: the mean and the median prediction errors for method B are closer to 0 and similar to the results of last menstrual period-based prediction. Also important is the visual observation that the distribution curve for method A is shifted to the left compared with the other two curves, which run fairly parallel to each other. These findings reflect a systematic underestimation of fetal age with method A in comparison with the other two methods. This systematic underestimation has been documented previously both by us (4) and by Eik-Nes' own group (5), and this error has also been commented on by others (6).

Another important argument is that method A performs less well than method B in early second trimester. Thus, our conclusion is based on a number of arguments and not solely on the observation that the mean prediction error of method B is closer to 0 than method A. We are well aware of the skewed distribution of human gestational length. This does not prohibit using the mean and median difference between expected and observed day of delivery to demonstrate the shift in the central distribution when different classification methods are applied on a cohort of women.

Eik-Nes, Grøttum, and Gjessing write that the validity of their standard values is limited to BPD within 38–60 mm. In our opinion, this is not commonly known. This problem with the standard ultrasound method was documented and discussed in 1999 (7) and led to the subsequent development of new normal values.

A fetus with BPD 29 mm is 92 days according to method A but 100 days according to method B (Table I). If BPD is 29–37 mm the difference in

Table I. Gestational length by biparietal diameter, for method A and method B.

	Gestational length (days)		
BPD (mm)	Method A	Method B	Difference (days)
29	92	100	8
30	94	102	8
31	96	104	8
32	99	105	6
33	101	107	6
34	103	109	6
35	105	111	6
36	108	113	5
37	110	115	5

(Received 24 May 2006; accepted 24 May 2006)

ISSN 0001-6349 print/ISSN 1600-0412 online
DOI: 10.1080/00016340600839718