Is It Time to Change Our Reference Curve for Femur Length? Using the Z-Score to Select the Best Chart in a Chinese Population

Boya Li1, Huixia Yang1*, Yumei Wei1, Rina Su1, Chen Wang1, Wenying Meng2, Yongqing Wang3, Lixin Shang4, Zhenyu Cai5, Liping Ji6, Yunfeng Wang7, Ying Sun8, Jiaxiu Liu9, Li Wei10, Yufeng Sun11, Xueying Zhang12, Tianxia Luo13, Haixia Chen14, Lijun Yu15

1 Department of Obstetrics and Gynecology, Peking University First Hospital, Beijing, China, 2 Tongzhou Maternal and Child Health Hospital of Beijing, Beijing, China, 3 Peking University Third Hospital, Beijing, China, 4 General Hospital of Beijing Military Region, Beijing, China, 5 Center Hospital of Aviation Industry, Beijing, China, 6 Pinggu Maternal and Child Health Hospital of Beijing, Beijing, China, 7 Beijing Hospital of Miyun City, Beijing, China, 8 Navy General Hospital, Beijing, China, 9 Beijing Daxing District Hongxing Hospital, Beijing, China, 10 Beijing Chui Yang Liu Hospital, Beijing, China, 11 Peking University Shougang Hospital, Beijing, China, 12 Combined with Traditional Chinese and Western Medicine Hospital of Beijing City, Beijing, China, 13 Beijing No.6 Hospital, Beijing, China, 14 Beijing Changping Hospital of Traditional Chinese Medicine, Beijing, China, 15 General Hospital of Jingmei Group, Beijing, China

* yanghuixia@bjmu.edu.cn

Abstract

Objective

To use Z-scores to compare different charts of femur length (FL) applied to our population with the aim of identifying the most appropriate chart.

Methods

A retrospective study was conducted in Beijing. Fifteen hospitals in Beijing were chosen as clusters using a systemic cluster sampling method, in which 15,194 pregnant women delivered from June 20th to November 30th, 2013. The measurements of FL in the second and third trimester were recorded, as well as the last measurement obtained before delivery. Based on the inclusion and exclusion criteria, we identified FL measurements from 19996 ultrasonograms from 7194 patients between 11 and 42 weeks gestation. The FL data were then transformed into Z-scores that were calculated using three series of reference equations obtained from three reports: Leung TN, Pang MW et al (2008); Chitty LS, Altman DG et al (1994); and Papageorghiou AT et al (2014). Each Z-score distribution was presented as the mean and standard deviation (SD). Skewness and kurtosis were compared with the standard normal distribution using the Kolmogorov-Smirnov test. The histogram of their distributions was superimposed on the non-skewed standard normal curve (mean = 0, SD = 1) to provide a direct visual impression. Finally, the sensitivity and specificity of each reference chart for identifying fetuses <5th or >95th percentile (based on the observed distribution of Z-scores) were calculated. The Youden index was also listed. A scatter diagram with the...
5th, 50th, and 95th percentile curves calculated from and superimposed on each reference chart was presented to provide a visual impression.

Results

The three Z-score distribution curves appeared to be normal, but none of them matched the expected standard normal distribution. In our study, the Papageorghiou reference curve provided the best results, with a sensitivity of 100% for identifying fetuses with measurements < 5th and > 95th percentile, and specificities of 99.9% and 81.5%, respectively.

Conclusions

It is important to choose an appropriate reference curve when defining what is normal. The Papageorghiou reference curve for FL seems to be the best fit for our population. Perhaps it is time to change our reference curve for femur length.

Introduction

The widespread use of ultrasound allows the measurement of fetal biometry and the estimation of fetal growth, thus making is possible to identify abnormal fetal growth patterns antenatally.

Of all routine ultrasound measurements, femur length (FL) is unique. It is not only a parameter that can assess fetal size but can also alert clinicians to the possible presence of fetal chromosomal abnormalities, intrauterine growth restriction and fetal malformations, particularly skeletal dysplasia, when it is below the expected range (5th percentile).

However, a short FL does not always indicate abnormal fetal growth. In our clinical practice, we have found that a number of fetuses with a “short femur length” were very healthy. This may, in part, be because Down’s syndrome screening has become a routine risk assessment for aneuploidy in China, and women with a high risk of aneuploidy are offered amniocentesis. Most women choose to terminate their pregnancy when an aneuploidy such as trisomy 21, 18 or 13 is diagnosed. Nevertheless, our popular fetal charts should also be reexamined, and the following questions, addressed: are these charts of high quality in terms of both design and statistical methodology? Are they applicable in a Chinese population?

In 2014, the Fetal Growth Longitudinal Study of the INTERGROWTH-21st Project, a multi-center, population-based longitudinal study, published their data and recommended international fetal growth standards for the clinical interpretation of routine ultrasound measurements and for comparisons across populations [1]. Because this study involved a large population and was of high quality, it raised the following question with respect to the older charts: is it time for a change?

As mentioned by McCarthy EA et al, “Inconsistent chart use and overestimation of fetal smallness can result in cynicism, confusion, and anxiety for pregnant women and their caregivers at all stages of pregnancy” [2]. In this article, we focused on FL, using Z-scores that integrate the measurement itself, the mean and the SD into a single value [3] to compare different charts of FL in our population to identify the most appropriate chart.

We did not include measurements of abdominal circumference, head circumference and estimated fetal weight in our analysis.
Methods

Study design and participants

A retrospective study was conducted in Beijing. Fifteen hospitals in Beijing were chosen as clusters using a systemic cluster sampling method, in which 15,194 pregnant women delivered from June 20th to November 30th, 2013. The questionnaire was designed to obtain information by interviewing all patients and reviewing their medical records. We had access to identifying information during and after data collection. These hospital units managed both low- and high-risk obstetric populations. More than 99.9% of the parturients were Chinese.

FL was measured in a plane where the full femoral diaphysis was seen almost parallel to the transducer and the measurement was made from one end of the diaphysis to another. In the third trimester, particular care was taken not to include the distal femoral epiphysis in the measurement. FL data from the second and third trimesters were recorded, as well as the last measurement before delivery. Thus, we excluded the possibility of overestimated measurements due to excessive performance of ultrasounds by physicians who may have had certain concerns about a pregnancy. In some cases, such as preterm labor, we were unable to obtain three FL recordings.

Ethics Statement

The study was reviewed and approved by the Institutional Review Board of the First Hospital, Peking University (Reference number: 2013[572]). All participants provided written informed consent, and the Ethics Committee approved the consent procedure.

Exclusion criteria

Because most reference curves were developed based on "normal" pregnancies, we excluded women who were at high risk for pregnancy complications, as well as those whose gestational age may have been inaccurate. The exclusion criteria were as follows:

1. Non-Chinese ethnicity.
2. Women with no fetal crown—rump length (CRL) measurements between 6 weeks and 0 days and 14 weeks and 0 days and in whom the difference in gestational age based on the last menstrual period and the fetal crown—rump length (CRL) as measured via ultrasound between 6 weeks and 0 days and 14 weeks and 0 days after the LMP was 7 days or more, using the formula described by Robinson and Fleming [4].
3. \( \text{GA(days) = 8.052} \times \sqrt{\text{CRL} \times 1.037} + 23.73 \) Women with twin and multiple gestation.
4. Women with disorders that may affect fetal growth: pre-gestational diabetes mellitus and cardiovascular disease (pre-existing hypertension, heart failure, coronary heart disease, arrhythmia, valvular heart disease).
5. Women with severe pregnancy complications: preeclampsia, eclampsia, HELLP syndrome.
6. Women with abnormal fetal outcomes: fetal malformations (congenital malformations diagnosed by ultrasound during pregnancy or at birth by clinical examination), fetal chromosomal abnormalities, evidence of fetal viral infection (cytomegalovirus infection), fetal death and stillbirth.
Method of dating pregnancy

Gestational age was dated according to last menstrual period precisely to the day and checked by the CRL measurement.

Choosing the reference curve

C. Ioannou et al [5] summarized and evaluated 83 studies of fetal biometry before 2012. This study provided basic insight into the ultrasound size charts that had been developed according to different populations worldwide, as well as their quality based on the study design, statistical analysis and reporting methods. In this systematic review, three publications from China were included: Lei H et al [6] did not provide the equation for the mean and SD and was thus excluded from our study; Pang MW et al [7] customized the fetal biometric charts not only according to gestational age but also according to variables such as maternal and pregnancy characteristics, including booking weight and height, age, parity and fetal sex. This study was excluded because the gestational weeks were between 24 and 40 weeks. We included the charts and reference equations reported in the Hong Kong Chinese population study by Leung TN, Pang MW et al [8], whose study design, statistical analysis and reporting methods were of high quality. The reference curve developed by Chitty LS, Altman DG et al [9] was also included in our study because of its high quality despite its early publication and narrow population (UK only). In addition, the latter reference is a classic chart that has been widely used.

Lastly, the reference curve published by the Fetal Growth Longitudinal Study of the INTERGROWTH-21st Project [1], a multi-center, population-based longitudinal study (that included a center in China), was included and evaluated for its suitability. This reference is the latest and newest study evaluating fetal charts that employed a scientific design, strict quality control and rigorous statistical analysis and has garnered widespread attention since its publication.

Table 1 shows the information of the three reference curves and their equations for the mean and SD of FL.

Table 1. Reference Equations for the mean and SD of FL.

<table>
<thead>
<tr>
<th>Author and publication year</th>
<th>Population</th>
<th>Reference Equations for the mean (in mm) of FL according to exact gestational age (in weeks). GA = exact gestational age</th>
<th>Reference Equations for the SD (in mm) of FL according to exact gestational age (in weeks). GA = exact gestational age</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leung TN, Pang MW, et al(2008)</td>
<td>China (Hong Kong)</td>
<td>[ FL_{\text{mean}} = 10 \times (-4.445082 + 0.492073 \times GA - 0.0067 \times GA^2 + 0.000042 \times GA^3) ]</td>
<td>[ SD = 12.53 \times (0.103184 + 0.002539 \times GA) ]</td>
</tr>
<tr>
<td>Chitty LS, Altman DG, et al(1994)</td>
<td>UK</td>
<td>[ FL_{\text{mean}} = -32.43 + 3.416 \times GA - 0.0004791 \times GA^2 ]</td>
<td>[ SD = 1.060 + 0.05833 \times GA ]</td>
</tr>
<tr>
<td>Papageorghiou, AT, et al(2014)</td>
<td>Pelotas, Brazil; Turin, Italy; Muscat, Oman; Oxford, UK; Seattle WA, USA; Shunyi County in Beijing, China; the central area of Nagpur, India; and the Parklands suburb of Nairobi, Kenya</td>
<td>[ FL_{\text{mean}} = -39.9616 + 4.32298 \times GA - 0.0380156 \times GA^2 ]</td>
<td>[ SD = \exp (0.605843 - 42.0014 \times GA^2 + 0.00000917972 \times GA^3) ]</td>
</tr>
</tbody>
</table>

doi:10.1371/journal.pone.0159733.t001

Statistical analysis

FL measurements from 19996 ultrasounds from 7194 patients between 11 and 42 weeks gestation were analyzed. The FL data were then transformed into Z-scores, calculated using three series of reference equations described in three studies: Leung TN, Pang MW et al (2008) [8]; Chitty LS, Altman DG et al (1994) [9]; and Papageorghiou, AT et al (2014) [1].
Statistical analysis was performed using SPSS version 18.0. Z-scores were calculated according to gestational age using the following formula [3]:

\[ Z \text{-score} = \frac{\text{observed FL} - \text{expected FL mean}}{\text{SD mean}} \]

The observed FL is the value obtained from the measurements, the expected FL mean is the value for our population calculated from the reference equations at this gestational age, and the SD mean is the SD associated with the mean value calculated from the reference equations at the same gestational age from our population [3]. According to the definition [3], Z-scores should follow a non-skewed standard normal distribution with a mean of 0 and an SD of 1 if the measurements are consistent with the reference equations used to calculate them. By definition [3], in a standard normal distribution, the -1 SD to +1 SD interval includes 68% of the population and the +2 SD interval includes 95% of the population, with the 5th percentile corresponding to -1.645 SD and the 95th percentile corresponding to +1.645 SD.

It is worth noting that the applied ranges of gestational weeks for the three reference equations were different. Before analysis, the FL measured outside the application was removed as long as the measurement was more than 5 SDs (because these were regarded as implausible on the basis of gestational age distribution from all of the sites [1]).

Each Z-score distribution was expressed as the mean and SD, as well as skewness and kurtosis, which were compared with the standard normal distribution using the Kolmogorov-Smirnov test. The histogram of their distributions was superimposed on the non-skewed standard normal curve (mean = 0, SD = 1) to provide a direct visual impression.

Finally, the sensitivity and specificity of each reference chart for identifying fetuses <5th or >95th percentile (based on the observed distribution of Z-scores) were then calculated. The Youden index (YI = sensitivity + specificity -1) was also listed. A scatter diagram with the 5th, 50th, and 95th percentile curves calculated from each reference chart that was superimposed on it was presented to provide a visual impression.

## Results

### Baseline demographic characteristics

Table 2 shows the baseline demographic characteristics for the enrolled population of our study. The median age of the mothers was 28.8 years. The average maternal weight before pregnancy were 56.8 kg. The mean maternal height ± SD were 162.5 ± 4.8mm. The mean gestational age of delivery was 39.5 (range, 28–42) weeks. Six thousand two hundred and eighty-four subjects (87.4%) were nulliparous. Six thousand and twenty-one (96.2%) delivered at term, two hundred and fifty-eight (3.6%) delivered preterm (<37 weeks) and fifteen (0.2%) delivered postterm (≥42 weeks). The mean birth weight ± SD was 3374± 431.4 g. And the numbers of valid observations for FL at each gestational week are shown in Table 3.

<table>
<thead>
<tr>
<th>Baseline characteristics</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maternal age, years</td>
<td>28.8</td>
<td>3.9</td>
</tr>
<tr>
<td>Gestational age of delivery, weeks</td>
<td>39.6</td>
<td>1.4</td>
</tr>
<tr>
<td>Maternal weight before pregnancy, kg</td>
<td>56.8</td>
<td>8.8</td>
</tr>
<tr>
<td>Maternal height, cm</td>
<td>162.5</td>
<td>4.8</td>
</tr>
<tr>
<td>Maternal body-mass index before pregnancy, kg/m²</td>
<td>21.5</td>
<td>3.3</td>
</tr>
<tr>
<td>Weight of new-bore, g</td>
<td>3374.8</td>
<td>431.4</td>
</tr>
</tbody>
</table>

Table 2. The demographic characteristics of pregnant women enrolled in this study.
Do they match the standard normal distribution?

The Z-score distribution curves of the measurements appeared to be normal (Fig 1), but none of them exactly matched the expected standard normal distribution. Table 4 shows the Z-score distribution, which was expressed as the mean and SD, as well as skewness and kurtosis, and the outcome compared with the standard normal distribution using the Kolmogorov-Smirnov test.

A total of 18,536 measurements between 12 and 40 gestational weeks were transformed into Z-scores using the reference equations from Leung TN, Pang MW et al. The mean value of the Z-score was 0.8651, and the SD was 1.09136. Skewness and kurtosis were -0.053 and -0.545, respectively, both less than 1. However, when divided by the standard error (SE), the results

Table 3. The numbers of valid observations for femur length (FL) at each gestational week.

<table>
<thead>
<tr>
<th>Gestational age</th>
<th>Number of observation</th>
<th>Gestational age</th>
<th>Number of observation</th>
<th>Gestational age</th>
<th>Number of observation</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>1</td>
<td>22</td>
<td>3109</td>
<td>33</td>
<td>285</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>23</td>
<td>2270</td>
<td>34</td>
<td>183</td>
</tr>
<tr>
<td>13</td>
<td>2</td>
<td>24</td>
<td>666</td>
<td>35</td>
<td>300</td>
</tr>
<tr>
<td>14</td>
<td>5</td>
<td>25</td>
<td>218</td>
<td>36</td>
<td>1013</td>
</tr>
<tr>
<td>15</td>
<td>6</td>
<td>26</td>
<td>88</td>
<td>37</td>
<td>1727</td>
</tr>
<tr>
<td>16</td>
<td>23</td>
<td>27</td>
<td>91</td>
<td>38</td>
<td>1077</td>
</tr>
<tr>
<td>17</td>
<td>17</td>
<td>28</td>
<td>243</td>
<td>39</td>
<td>1472</td>
</tr>
<tr>
<td>18</td>
<td>15</td>
<td>29</td>
<td>702</td>
<td>40</td>
<td>1254</td>
</tr>
<tr>
<td>19</td>
<td>23</td>
<td>30</td>
<td>1925</td>
<td>41</td>
<td>139</td>
</tr>
<tr>
<td>20</td>
<td>105</td>
<td>31</td>
<td>1426</td>
<td>42</td>
<td>1</td>
</tr>
<tr>
<td>21</td>
<td>519</td>
<td>32</td>
<td>1090</td>
<td>Total</td>
<td>19996</td>
</tr>
</tbody>
</table>

doi:10.1371/journal.pone.0159733.t003


doi:10.1371/journal.pone.0159733.g001
were -2.94 and 15.1, both absolute values >2, thus refuting the normal distribution hypothesis. The result of the Kolmogorov-Smirnov test also confirmed that the Z-score distribution refuted the normal distribution hypothesis (Z = 1.532, P = 0.018). In the histogram of Z-score distributions with a centered and superimposed standard normal reference curve (Fig 1a), the histogram of Z-scores calculated using the Leung TN, Pang MW et al equations was clearly skewed to the right.

Using the reference equations from Chitty LS, Altman DG et al, a total of 19,588 measurements between 12–42 gestational weeks were transformed into Z-scores. The mean value of the Z-scores was -0.1322, and the SD was 0.8719. Skewness and kurtosis were 0.001 and 0.911, respectively, both less than 1. However, when divided by the standard error (SE), the results were 0.056 and 2.6, the absolute value of the latter was >2, thus refuting the normal distribution hypothesis. The result of the Kolmogorov-Smirnov test also confirmed this conclusion (Z = 1.989, P = 0.001). The histogram of Z-score distributions calculated using this equation (Fig 1b) seemed to be narrower compared with the standard normal reference curve.

Finally, we used the reference equations provided by Papageorghiou AT et al. A total of 19,515 measurements between 14–42 gestational weeks were transformed into Z-scores. The mean value of the Z-scores was 0.5845, and the SD was 1.3923. The skewness and kurtosis were 0.043 and 0.029, respectively, both less than 1. However, when divided by the standard error (SE), the results were 2.38 and 0.83, and the absolute value of the former was >2, thus refuting the normal distribution hypothesis. The result of the Kolmogorov-Smirnov test confirmed the same hypothesis (Z = 1.755, P = 0.004). In the histogram of Z-score distributions with a centered and superimposed standard normal reference curve (Fig 1c), the histogram of Z-scores calculated using the equations from Papageorghiou AT et al (2014) seemed to be slightly wider and lower.

Are they effective at identifying measurements <5th or >95th percentile?

From the scatter diagram of the 5th, 50th, and 95th percentile curves calculated from each superimposed reference chart (Fig 2), we were able to obtain a rough direct impression. The overall results for the classification of the fetuses using the 5th and 95th percentiles from each of the three reference curves for each parameter are shown in Tables 5 and 6 (see Tables 5 and 6).

When using the reference equations from Leung TN, Pang MW et al, the observed Z-scores for the 5th percentile and 95th percentile were -0.8983 and 2.6627. A total of 710 measurements that were actually less than the 5th percentile were missed diagnoses, and 3,479 measurements were wrongly classified as larger than the 95th percentile. The sensitivity of screening fetuses with measurements < 5th percentile was only 23.1%, although the specificity was 100%, and the Youden index was 0.231. Thus, the value was too low to be used as a

### Table 4. Mean, standard deviation (SD) and the results of the Kolmogorov-Smirnov test of Z-score.

<table>
<thead>
<tr>
<th>Author</th>
<th>Number</th>
<th>Range of gestational weeks</th>
<th>Population</th>
<th>Mean</th>
<th>SD</th>
<th>Z of K-S test</th>
<th>P</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leung TN, et al</td>
<td>18536</td>
<td>12-40wks</td>
<td>China (Hong Kong)</td>
<td>0.8651</td>
<td>1.09136</td>
<td>1.532</td>
<td>0.018</td>
<td>-0.053</td>
<td>-0.545</td>
</tr>
<tr>
<td>Chitty LS, et al</td>
<td>19588</td>
<td>12-42wks</td>
<td>UK</td>
<td>-0.1322</td>
<td>0.8719</td>
<td>1.989</td>
<td>0.001</td>
<td>0.01</td>
<td>0.911</td>
</tr>
<tr>
<td>Papageorghiou AT, et al</td>
<td>19515</td>
<td>14-42wks</td>
<td>Pelotas, Brazil; Turin, Italy; Muscat, Oman; Oxford, UK; Seattle WA, USA; Shunyi County in Beijing, China; the central area of Nagpur, India; and the Parklands suburb of Nairobi, Kenya</td>
<td>0.5845</td>
<td>1.3923</td>
<td>1.755</td>
<td>0.004</td>
<td>0.043</td>
<td>0.029</td>
</tr>
</tbody>
</table>

doi:10.1371/journal.pone.0159733.004
diagnostic test. The sensitivity and specificity of screening for fetuses with measurements >95th percentile were 100% and 80.2%, respectively, and the Youden index was 0.802.

The observed Z-scores for the 5th percentile and 95th percentile when using the reference equations from Chitty LS, Altman DG et al were -1.530 and 1.2746. A total of 280 measurements that were actually less than the 5th percentile and 586 measurements that were actually larger than the 95th percentile were missed diagnoses. The sensitivity of screening fetuses with measurements <5th percentile and >95th percentile was 72.1% and 39.8%, respectively, and the specificity was 100% for both percentiles. Thus, the Youden index was 0.721 for identifying measurements that were <5th percentile and 0.398 for identifying measurements that were >95th percentile.

Finally, when using the reference equations from Papageorghiou AT et al (2014), the observed Z-score for the 5th and 95th percentile were -1.6622 and 2.9296. Only 25 measurements were wrongly classified as less than the 5th percentile, and 3,436 measurements were wrongly classified as greater than the 95th percentile. The sensitivity of screening fetuses with measurements <5th percentile and >95th percentile were both 100%, and the specificity was 99.9% and 81.5%, respectively. Thus, the Youden index was 0.999 for identifying measurements that were <5th percentile and 0.815 for identifying measurements that were >95th percentile.

**Discussion**

Because of ethnic heterogeneity, we were initially inclined to believe that reference curves calculated from data from our own nation and ethnic population would be more applicable. Surprisingly, our findings were very different. The Z-score distribution curves of the measurements
<table>
<thead>
<tr>
<th></th>
<th>Number of observation</th>
<th>Range of gestational weeks</th>
<th>Expected 5&lt;sup&gt;th&lt;/sup&gt; perc (Zscore)</th>
<th>Observed 5&lt;sup&gt;th&lt;/sup&gt; perc (Zscore)</th>
<th>N&lt;5&lt;sup&gt;th&lt;/sup&gt; perc expected</th>
<th>N&lt;5&lt;sup&gt;th&lt;/sup&gt; perc calculated</th>
<th>TP</th>
<th>FP</th>
<th>TN</th>
<th>FN</th>
<th>Number wrongly classified</th>
<th>Se</th>
<th>Spe</th>
<th>Youden</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leung TN, et al,</td>
<td>18536</td>
<td>12-40wks</td>
<td>-1.645</td>
<td>-0.8983</td>
<td>923</td>
<td>213</td>
<td>213</td>
<td>0</td>
<td>17613</td>
<td>710</td>
<td>710</td>
<td>23.1%</td>
<td>100%</td>
<td>0.231</td>
</tr>
<tr>
<td>Chitty LS, et al</td>
<td>19588</td>
<td>12-42wks</td>
<td>-1.645</td>
<td>-1.530</td>
<td>1002</td>
<td>722</td>
<td>722</td>
<td>0</td>
<td>18586</td>
<td>280</td>
<td>280</td>
<td>72.1%</td>
<td>100%</td>
<td>0.721</td>
</tr>
<tr>
<td>Papageorghiou AT, et al</td>
<td>19515</td>
<td>14-42wks</td>
<td>-1.645</td>
<td>-1.6622</td>
<td>976</td>
<td>1001</td>
<td>976</td>
<td>25</td>
<td>18514</td>
<td>0</td>
<td>25</td>
<td>100%</td>
<td>99.9%</td>
<td>0.999</td>
</tr>
</tbody>
</table>

TP: True positive; FP: False positive; FN: False negative; TN: True negative; Se: sensitive; Spe: Specificity.

doi:10.1371/journal.pone.0159733.t005
Table 6. Results of screening for fetuses with measurements > 95th percentile.

<table>
<thead>
<tr>
<th>Number of observation</th>
<th>Range of gestational weeks</th>
<th>Expected 95th perc (Zscore)</th>
<th>Observed 95th perc (Zscore)</th>
<th>N &gt; 95th perc expected</th>
<th>N &gt; 95th perc observed</th>
<th>TP</th>
<th>FP</th>
<th>TN</th>
<th>FN</th>
<th>Number wrongly classified</th>
<th>Se</th>
<th>Spe</th>
<th>Youden</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leung TN, et al</td>
<td>18536</td>
<td>12-40wks</td>
<td>1.645</td>
<td>2.6627</td>
<td>923</td>
<td>4402</td>
<td>923</td>
<td>14134</td>
<td>0</td>
<td>3479</td>
<td>100%</td>
<td>80.2%</td>
<td>0.802</td>
</tr>
<tr>
<td>Chitty LS, et al</td>
<td>19588</td>
<td>12-42wks</td>
<td>1.645</td>
<td>1.2746</td>
<td>974</td>
<td>388</td>
<td>388</td>
<td>18614</td>
<td>586</td>
<td>586</td>
<td>39.8%</td>
<td>100%</td>
<td>0.398</td>
</tr>
<tr>
<td>Papageorghiou AT, et al</td>
<td>19515</td>
<td>14-42wks</td>
<td>1.645</td>
<td>2.9296</td>
<td>960</td>
<td>4396</td>
<td>960</td>
<td>15119</td>
<td>0</td>
<td>3436</td>
<td>100%</td>
<td>81.5%</td>
<td>0.815</td>
</tr>
</tbody>
</table>

TP: True positive; FP: False positive; FN: False negative; TN: True negative; Se: sensitive; Spe: Specificity.

doi:10.1371/journal.pone.0159733.t006
appeared to be normal, but none of them matched the expected standard normal distribution. The only published reference curve from a Chinese population (Leung TN, Pang MW et al) had very limited diagnostic value in correctly identifying fetuses with short femur length, with a Youden index of only 0.231. In the clinical arena, however, the significance of correctly diagnosing a fetus with a short femur length is much greater than that of identifying a longer FL. Thus, this curve is not the appropriate reference chart for our Chinese population. This may be due to the fact that this curve was calculated from data from one Hong Kong hospital and that hereditary differences may exist between women in that region and women in northern China.

The classic reference curve published by Chitty LS, Altman DG et al in 1994 has been widely used. The Youden index for recognizing a FL $< 5^{th}$ percentile is 0.721. A Youden index of more than 0.7 is generally considered to be a better diagnostic test. Although its diagnostic value for recognizing a FL $> 5^{th}$ percentile is lower (Youden index 0.398), we still considered this reference curve to be valuable in the clinical arena.

Papageorghiou AT et al published the study of the Fetal Growth Longitudinal Study of the INTERGROWTH-21st Project. This was a multi-center, population-based longitudinal study that aroused widespread concern in the obstetrics academic community globally because of its scientific design. Doubts still remain. The data were collected from 8 countries, and we were uncertain about the applicability of the reference curve in our population. We were surprised to discover that this chart had very high sensitivity and specificity for correctly identifying measurements below the 5th percentile and measurements greater than the 95th percentile, with Youden indices of 0.999 and 0.815, respectively. Thus, this reference chart was found to have the best diagnostic value in our study.

**Limitations**

Our study evaluated different reference standards for measuring femoral length (FL) in a Chinese population. However, it has some limitations.

Recent studies about customized percentiles that are adjusted or customized based on sex and maternal characteristics, such as height, weight, parity, and ethnic origin, have conveyed the idea that one size does not fit all [10]. Instead of customized percentiles, our study used population percentiles for the following reasons: first, it was not permitted to report fetal sex on ultrasound because of Chinese government policies, and fetal sex and weight are two of the most important factors in customized percentiles. Therefore, despite its advantages, the clinical use of customized growth charts is limited in terms of retrospective value and clinical decision making in China. Second, our data were collected in 2013, when China was enforcing the one-child policy, which meant that most families had only one child (6,284 of 7,194 women were primiparous in our study). Taking these unique national policies into consideration, we still used the traditional population percentiles in our study. We attempted to address this limitation by setting exclusion criteria designed to reduce individual differences, for example, by excluding women who were not ethnically Chinese, had twin or multiple pregnancies, had disorders that may affect fetal growth and had severe pregnancy complications. This policy may change some day when people no longer prefer boys to girls. We thus expect that the customized growth charts will present a real benefit to mothers.

**Conclusions**

It is important to choose an appropriate chart when defining what is normal. The Papageorghiou reference curve for FL seems to best fit our population. It may be time to change our reference curve for femur length.
Supporting Information

S1 File. Original data of the manuscript. This is the original data of the manuscript. (XLS)

S2 File. AJE certificate for language editing of the manuscript. (PDF)

S3 File. Editing for proper English language by AJE with tracked changes of the original manuscript after the initial revision. The data of this version was different from the publication one because the exclusion criteria had changed after the second revision. (DOC)

Author Contributions

Conceived and designed the experiments: BL HY Y. Wei. Performed the experiments: BL. Analyzed the data: BL. Contributed reagents/materials/analysis tools: HY CW RS WM Yongqing Wang LS ZC LJ Yunfeng Wang Ying Sun JL LW Yufeng Sun Ying Sun JL LW Yufeng Sun XZ TL HC LY. Wrote the paper: BL HY.

References


