IMPLEMENTATION AND STATISTICAL EVALUATION OF COMPUTER ASSISTED TW2 METHOD FOR BONE AGE ASSESSMENT

by

Esra Güven

B.S. in Chemistry, Boğaziçi University, 2006

Submitted to the Institute of Biomedical Engineering
in partial fulfillment of the requirements
for the degree of
Master of Science
in
Biomedical Science

Boğaziçi University
March 2009

IMPLEMENTATION AND STATISTICAL EVALUATION OF COMPUTER ASSISTED TW2 METHOD FOR BONE AGE ASSESSMENT

APPROVED BY:

Assoc. Prof. Dr. Albert Güveniş .....
(Thesis Advisor)

Prof. Dr. Bülent Sankur .....

Assistant Prof. Dr. Ata Akın .....

DATE OF APPROVAL: 19.02.2009

ACKNOWLEDGMENTS

I would like to thank my thesis supervisor, Assoc. Prof. Dr. Albert Güveniş for his support and motivation during my thesis study. Without his valuable academic support, experience, and knowledge, this study would not be completed. I would like to thank Dr. Metin Yıldız who shared his hand radiograph archive with us, for his support in Göztepe SSK Hospital.

I would like to thank Mahmut Haktan who has helped me to learn how to use the web-based bone age assessment software in the .Net environment and allowed me to use this software for my thesis. Many thank to my friends in B.U. Biomedical Engineering Institute, to my family and to İlker Peltek. Working and sharing life with you is unforgettable.

And to those who directly and indirectly help me, thank you.

ABSTRACT

IMPLEMENTATION AND STATISTICAL EVALUATION OF COMPUTER ASSISTED TW2 METHOD FOR BONE AGE ASSESSMENT

The most commonly used method for bone age assessment is based on a single x-ray of the hand and wrist. The bones in the x-ray are compared to the bones of a standard atlas, usually "Greulich and Pyle (G&P)". A more complex method also based on hand x-rays is the "Tanner-Whitehouse (TW2)" method, which relies on the systematic evaluation of the maturity of all the bones in the hand and wrist.

In this study, first we implemented the computer assisted TW2 method, then we compared this method with reference to widely used method of G&P using the criteria of accuracy and speed, and lastly we studied how learning and practice affects speed of bone age assessment. We used 50 "bone age" radiographs of the left hand and wrist performed in a large hospital. Data were analyzed using the "method comparison" statistical technique. 20% of the radiographs were then re-analyzed to assess intra-observer variation. The 95% confidence interval for the difference between the two methods was -1.84 to 1.32 years. Intra-observer variation was greater for the G&P method than for the TW2 method (95% confidence limits, -0.77 to 0.97 vs -0.45 to 0.37). The speed of computer based TW2 was close to G&P (1.7 min vs 0.7 min) and increased with practice. Since both methods take reasonable amount of time, computerized TW2 method should be preferred for higher performance in bone age assessment.

Keywords: Bone age; Greulich and Pyle; Computerized Tanner and Whitehouse.

ÖZET

KEMİK YAŞI DEĞERLENDİRMESİ İÇİN BİLGİSAYARLI TW2 METODUNUN UYGULANMASI VE İSTATİSTİKSEL DEĞERLENDİRMESİ

Kemik yaşı tayini için en sık kullanılan metot el ve bileği içeren tek bir x-ışını röntgenine dayanır. Röntgendeki kemikler standart bir atlastaki, genellikle "Greulich ve Pyle (G&P)", kemikler ile karşılaştırılır. El röntgenlerine dayanan diğer bir ve daha karmaşık olan metot ise "Tanner-Whitehouse (TW2)" metodudur. Bu metot, el ve bilekteki tüm kemiklerin olgunlaşma derecelerinin sistematik değerlendirmesine dayanır.

Bu çalışmada, ilk olarak bilgisayarlı TW2 metodu uygulamaya hazır hale getirildi, daha sonra büyük bir hastanede çekilen 50 sol el ve bilek röntgeni kullanılarak yaygın olarak kullanılan G&P metodu referans alınarak doğruluk ve hız kriterleri yönünden karşılaştırıldı ve öğrenme ve pratiğin kemik yaşı değerlendirme hızı üzerindeki etkisi üzerinde çalışıldı. Veriler, "metot karşılaştırma" istatistiksel tekniği kullanılarak analiz edildi. İç gözlemci değişimini değerlendirmek için filmlerin % 20'si tekrar analiz edildi. İki metot arasındaki farka yönelik %95 güvenirlik aralığı -1,84-1,32 yıl olarak bulundu. G&P metodu için iç-gözlemci değişiminin TW2 metoduna göre daha fazla olduğu bulundu (%95 güvenirlik aralığı, -0,77-0,97' ye karşılık -0,45-0,37). Bilgisayarlı TW2 hızının G&P' ye yakın olduğu (1,7 dak.'ya karşılık 0,7 dak) ve pratik yaptıkça hızın arttığı sonucuna varıldı. Her iki metot da kaydadeğer zaman aldığından, kemik yaşı tayininde daha yüksek bir performans elde etmek için bilgisayarlı TW2 metodu tercih edilmelidir.

Anahtar Sözcükler: Kemik yaşı, Greulich ve Pyle, Bilgisayarlı Tanner ve Whitehouse.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

d	Mean difference
n	Number of subject
p	Probability factor
s	Standart deviation of differences
t	t-value

LIST OF ABBREVIATIONS

CL	Confidence Limit
G&P	Greulich and Pyle
TW2	Tanner and Whitehouse
RUS	Radius, Ulna and Short Finger Bones
ROI	Region of Interest
EMROI	Epiphyseal-Metaphyseal Region of Interest
CROI	Carpal Region of Interest

1. INTRODUCTION

1.1 Background and Motivation

Implementation and statistical evaluation of the computerized bone age assessment system will be useful for the pediatric radiology. Bone age assessment is performed about 600-1000 times per year in most of the Education and Research Hospitals. It is not only a time consuming procedure but it also depends on the experience of the physician. The aim of this project is first to implement a new bone age assessment system (computerized TW2), secondly to show the degree of agreement between this new system and the widely used system (manual G&P), and finally to study how learning and practice affects speed of bone age assessment.

1.2 Uses of Skeletal Age Assessment

The bone age or skeletal age assessment is a commonly used procedure for pediatric patients to evaluate their growth disorder, examine their growth disorder, determine their growth potential, and monitor effect of growth therapy. The growth potential of an individual depends largely on the progression of ossification within the epiphysis. Bone age is a measurement of the epiphyseal center development. It is an important procedure in the diagnosis and management of endocrine disorders, diagnostic evaluation of metabolic and growth abnormalities, deceleration of maturation in a variety of syndromes, malformations, and bone dysplasias. A simple method frequently used in bone age assessment is atlas matching by a radiological examination of a left hand and wrist radiograph against a reference set of atlas patterns of normal standards. Although the hand and wrist does not contribute to the height of an individual, the radiograph of this part of the body has been proven valuable and is commonly used in assessment of bone age. After determination of skeletal age, we can compare the results with chronological age. A big difference between these two variables shows an

atypical skeletal development of patient. This procedure is often used in the diagnosis and management of endocrine disorders. Generally it can indicate rate of growth of a patient. Bone age assessment is also used in forensic medicine in such conditions when there is a need to confirm chronological age of a criminal.

1.3 Objectives

The main objectives of this thesis was to implement the computerized "point scoring system" of Tanner and Whitehouse (TW2), to compare bone ages assessed using either the "atlas matching" method of Greulich and Pyle or the computerized TW2 and finally to study how learning and practice affects speed of bone age assessment.

1.4 Outline of the Thesis

Chapter 1 introduces the subject and gives outline of this thesis work. Clinical methods of bone age assessment are discussed in Chapter 2. In Chapter 3, some statistical methods for assessing agreement between two methods of clinical measurement are given. In Chapter 4, previous studies related to this thesis are summarized. In Chapter 5 we will have a look at the computer assisted bone age assessment method. In Chapter 6, details of statistical evaluation of computer assisted TW2 method for bone age assessment are given. Future work and conclusion can be found in the Chapter 7.

2. CLINICAL METHODS OF BONE AGE ASSESSMENT

2.1 Introduction

In this chapter, the type of clinical practice methods for the bone age assessment will be examined and explained.

2.2 Skeletal Age

Skeletal age, a measure of skeletal development of a child, is determined by using standard data that includes measurement of skeletal developments of healthy children. Numerous Roentgen methods have been proposed to assess the bone age according to criteria such as the time of appearance, size and differentiation of the ossification centers. Most commonly used methods are Greulich and Pyle method (G&P), and Tanner and Whitehouse (TW2) method.

2.3 Bone Age Assessment Methods

Greulich and Pyle (G&P) [1] and Tanner and Whitehouse (TW2) [2] methods use atlas matching methods for manual determination of skeletal age. G&P method is easier and faster to use, however TW2 method is more reliable.

Both methods rely on radiographs of the left hand. Prior to detailed examination of the two methods it is useful to mention a little about the anatomy of the hand in order to understand the terminology. Figure 2.1 and Figure 2.2 show the most important bones for skeletal age determination in the hand. In Figure 2.1 we see a schematic representation of the bones of hand skeleton of the hand. Figure 2.2 shows an actual hand radiograph. The most important hand bones are those of the

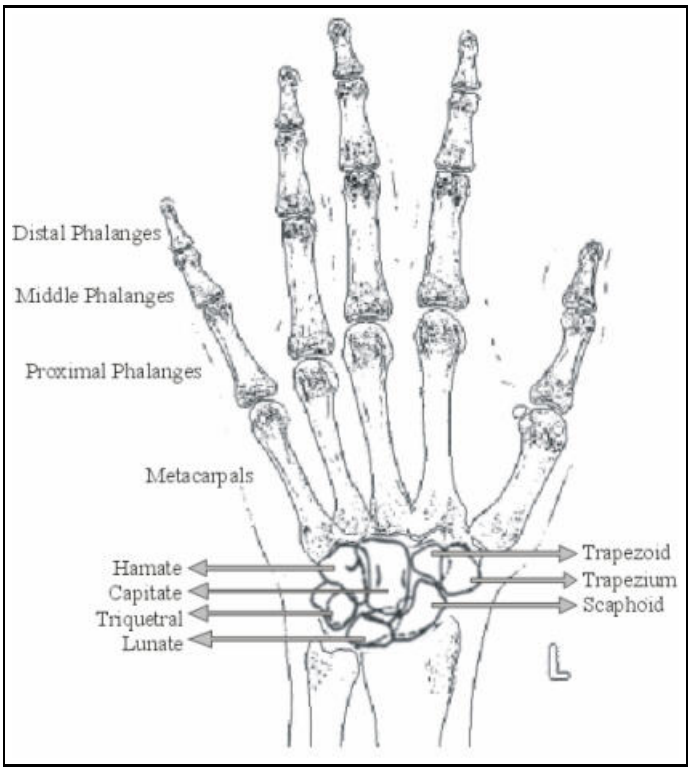


Figure 2.1 Anatomy of left hand

fingers, the proximal, middle and distal phalanges. The thumb has only proximal and distal phalanges. Next there are the carpal bones, Capitate, Hamate, Triquetral, Lunate, Scaphoid, Trapezium and Trapezoid. Between distal phalanges and carpals five metacarpals are observed.

Ossification centers are located between the phalanges of a developing child. Because the radiograph in Figure 2.2 is of an adult subject they can not be shown whereas in Figure 2.3 ossification centers are shown. We can clearly observe three parts, the metaphysis, the epiphysis and the diaphysis. We call this structure as an Epiphyseal-Metaphyseal Region of Interest (EMROI). This structure changes during the maturation of the skeleton. The epiphysis becomes steadily wider and eventually fuses with the metaphysis.

Figure 2.2 Radiograph of left hand

2.3.1 The Greulich and Pyle Method

First studies on the human growth and development started in 1929 at the Western Reverse University School of Medicine in Ohio. A Large number of children of different ages were used in these studies. Left shoulder, elbow, hand, hip and knee radiographs were taken for examination. In the first postnatal year an examination was conducted every three months, from twelve months to five years they were examined in each 6 months and annually thereafter. In total the study ran from 1931 until 1942. In 1937 "Atlas of Skeletal Maturation of Hand" was published by Todd [3].

The Greulich-Pyle has been used extensively by orthopedists and was used by Green and Anderson in compiling their data for growth remaining in children nearing skeletal maturity. Since the Moseley straight line graph was based on the Green-Anderson data, the Greulich-Pyle system is correlated with that graph also. The two methods do not give equivalent bone ages. The Tanner-Whitehouse may be more reproducible, but for the present, the Greulich-Pyle method is still standard for the or-

thopedic use, even though it was derived more than a half century ago on an exclusively white upper-middle class population.

Assessment steps:

1. Assess one bone at a time
2. Locate the Atlas plate that most closely resembles the X-Ray bone
3. Interpolate between Atlas plates.
4. Assess a second time.
5. Average the two twenty-eight assessments.

In the Greulich and Pyle method twenty-eight bones are compared against (normal) the reference atlas. All the bones in the hand and wrist are examined and average of the assessments gives us the bone age of patient.

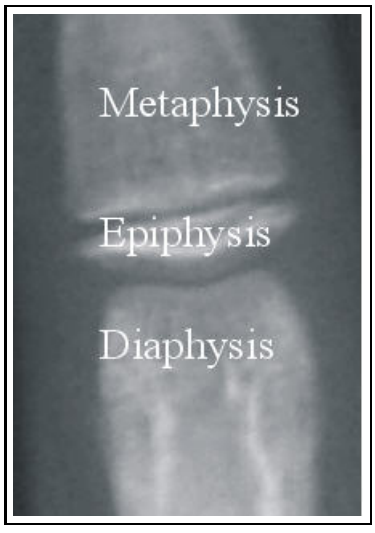


Figure 2.3 The ossification centers on finger

Development stages of each bone are described in the atlas in Figure 2.4. The descriptions in the Figure are more a general guideline to the development of each bone in the hand rather than an instruction on how to rate a bone. Most institutions are

using a more rapid modified version of the original atlas, which is also potentially less accurate. This version is described below

The atlas is divided into two parts, one for the male patients and one for the female patients because females develop quicker than males. Each part contains standard radiographic images of the left hand of children ordered by the chronological age. Skeletal development found in the Greulich and Pyle atlas is given as an example in Figure 2.4.

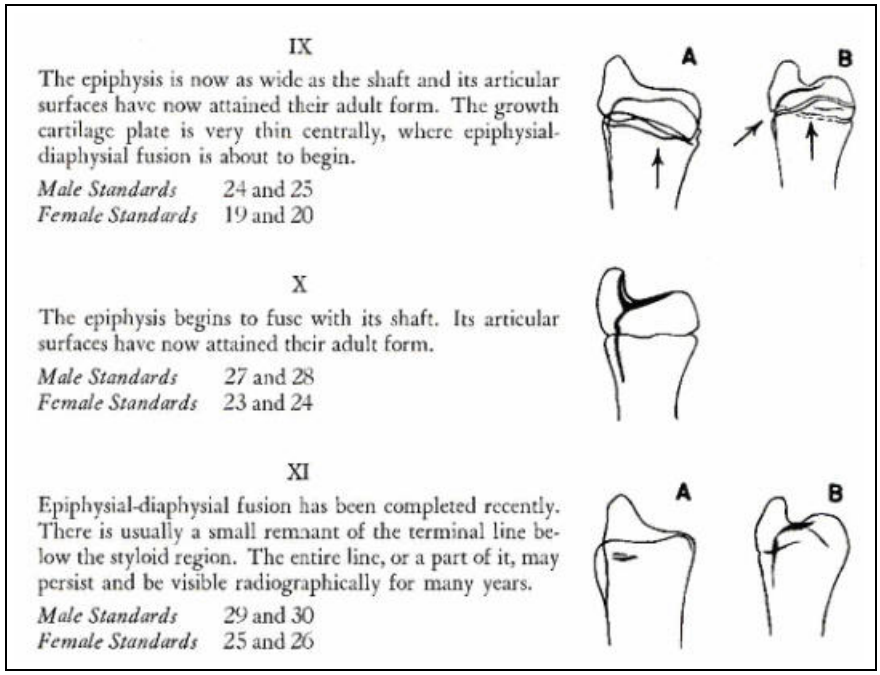


Figure 2.4 Example of the descriptions of the skeletal development

To measure skeletal age of a patient, first of all the radiograph of the patient is compared with the image in the atlas that matches with the chronological age of the patient. Next, one should compare it with the adjacent images representing both younger and older children. There are some maturity indicators when comparing the radiograph of the patient against an image in the atlas. These features can vary with the race, age and sex of the child. In the younger children the presence or absence of the certain carpal or epiphyseal ossification centers are often pointers for the physician about the skeletal age of a child. In older children the shape of the epiphyses and the

amount of fusion with the metaphysis is a good indicator for the skeletal age, carpal ossification centers did not differ at that time. Once the atlas image that most resembles the radiograph is found, the physician should conduct a more detailed examination of the individual bones and epiphyses. So we can find the skeletal age when the matching radiograph has been found.

2.3.2 The Tanner and Whitehouse (TW2) Method

Tanner and Whitehouse knew about the Greulich and Pyle atlas and the way in which it was used. They noticed several aspects of the method that they felt needed to be improved. According to them the subjectivity of the matching process was an obvious weakness of the G&P method. Physicians generally look at the whole radiograph at once and then compare it with an image in the atlas. Often a specific radiograph does not match any of the images in the atlas exactly, and little guidance is given on how to balance out the discrepancies that arise from one bone being more or less advanced than its match in the atlas.

The scale used for expressing the maturity was another aspect of the G&P method TW2 didn't like. Each standard radiograph in the atlas has the age of the child from whom it was taken associated with it. Therefore maturity is measured on an age scale. It could even be seen as the predicted age, because the matching process gives the most likely chronological age of the child being matched, as judged from the radiograph. This most likely age is then said to be the skeletal age. Tanner and Whitehouse deemed that a new, more sophisticated system was needed which would not make use of an age scale for maturity measurements. In their view a maturity scale should be defined in a manner which does not directly relate to age. This would allow them to produce a set of "maturity standards" of any given population by studying the relationship between maturity and age.

To build standard atlas radiographs of up to 12 years old, healthy children were used in the period of six months. Changes in the shape and density markings of each

bone were recorded in standard atlas. After this the different stages were identified, these stages had to be universally present in all individuals.

Features that were only present in the bones of particular subjects were excluded; also absolute size was always ignored. The exact number of stages for each bone was chosen so that differences between consecutive stages were neither so small as to cause confusion and rating error, nor so large that significant information was lost.

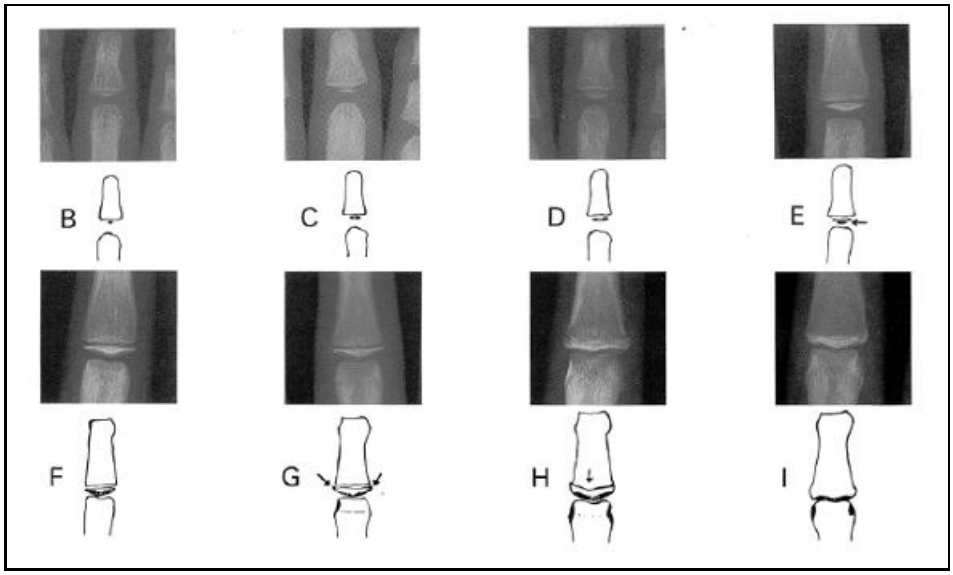


Figure 2.5 Example images of a given ROI in the hand

Figure 2.5 shows an example of the images used to find the stage (B through I) of a given ROI (in this case the middle phalanx of the third finger) in the hand. When the stage is known the ROI can be scored. The development of the epiphysis is clearly showed.

A score had to be assigned to each bone in the hand for each stage. Many of the bones in the hand and wrist are giving approximately the same information about maturity. This is especially true for the bones in the fingers. Therefore it is not desirable to just take an average of all the scores in a hand since this would give the nineteen finger bones together a much greater importance than the seven carpals or the radius and the ulna. A system of weighted scores was developed in order to correct

for this. Also, the scores of boys and girls are different because girls mature faster than boys. Especially in older children the carpal bones do not provide any useful information about the skeletal maturity. To allow the exclusion of the carpals or the specific examination of them, three separate scoring systems were developed:

1. RUS analysis (13 ROI are examined, no carpal bones are involved)
2. Carpal analysis (only the 7 carpal bones are examined)
3. Full TW2 analysis (20 ROI are examined)

The first system named RUS (Radius, Ulna and Short finger bones) contains the wrist, the bones of the thumb and the third and fifth fingers. The second system contains only the carpal bones. The third system is the most extensive one featuring 20 bones of the hand. These bones include bones of the wrist, the carpal bones, and bones of the thumb and the third and fifth fingers, which is also the sum of the first two systems. Our software is based on the full TW2 analysis system.

Using the TW2 method is relatively straightforward; first one chooses a suitable scoring system (TW2, RUS or carpal). A suitable scoring system is mostly determined by the calendar age of the patient.

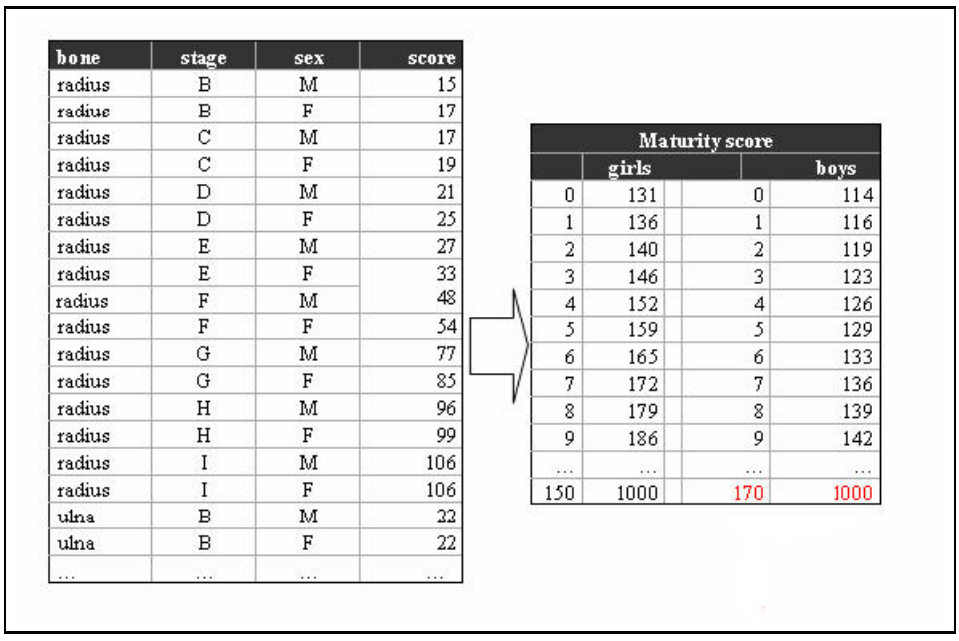
Next one looks up the bones in the radiograph that are associated with that particular system. Those particular bones are then compared to a series (see Figure 2.5) of reference images in the TW2 book. These reference images each represent one developmental stage (B-I) and are backed up by textual descriptions. In these descriptions of a stage, important changes in the bones are described. Also, the descriptions often contain requirements that a bone must meet in order to be classified a certain stage.

The reference image that most resembles the bone from the radiograph is chosen. Of course the bone must also meet the requirements in the description. The stage

corresponding to a reference image is assigned to the bone. When the stage of a certain bone is known the score associated with that particular bone and stage can be looked up. After all the required bones have been analyzed and all the needed scores have been gathered, the total score is determined by adding all the separate scores. The total maturity score is linked to a certain skeletal age via a conversion table (see Figure 2.6). Bone age of the patient can be calculated with the maturity score: Bone age = 1 + Maturity sore / 10.

For example, if our total maturity score value is 1000, then the bone age value can be calculated as 18 from the formula. So, with TW2 method we can calculate the bone ages up to 18 for males, and up to 16 for females.

User chooses one stage for each one of the bones that corresponds one individual maturity score for each one of them. With the sum of the individual scores user can define the maturity score from two different scales one for male and one for female by using the conversion table in Figure 2.6.



bone	stage	sex	score
radius	B	M	15
radius	B	F	17
radius	C	M	17
radius	C	F	19
radius	D	M	21
radius	D	F	25
radius	E	M	27
radius	E	F	33
radius	F	M	48
radius	F	F	54
radius	G	M	77
radius	G	F	85
radius	H	M	96
radius	H	F	99
radius	I	M	106
radius	I	F	106
ulna	B	M	22
ulna	B	F	22
...	...	...	...

Maturity score			
	girls		boys
0	131	0	114
1	136	1	116
2	140	2	119
3	146	3	123
4	152	4	126
5	159	5	129
6	165	6	133
7	172	7	136
8	179	8	139
9	186	9	142
...	...	...	...
150	1000	170	1000

Figure 2.6 Conversion chart for skeletal age

3. STATISTICAL METHODS FOR ASSESSING AGREEMENT BETWEEN TWO METHODS OF CLINICAL MEASUREMENT

3.1 Introduction

In clinical measurement, a new measurement method is often compared with an established measurement method to measure the agreement between them. If the agreement is sufficient, the old method may be replaced by the new one. This should not be confused with calibration. In calibration, known quantities are measured by a new method and the result is then compared with the true value. When two methods are compared neither provides a true value, only the degree of agreement can be assessed. But such investigations are often analyzed inappropriately usually by using correlation coefficients as an indicator of agreement [4]. The use of correlation is misleading. An alternative and better technique will be described in this chapter.

3.2 Sample Data

A set of data presented in Table 3.1 are given as an example for expression of statistical approach which was used in this thesis.

The first and second measurements are performed by two different methods on the same subjects. The evaluation of these values are given in this chapter.

Table 3.1
Sample data

Subject	First measurement(l/min)	Second measurement(l/min)
1	494	512
2	395	430
3	516	520
4	434	428
5	476	500
6	557	600
7	413	364
8	442	380
9	650	658
10	433	445
11	417	432
12	656	626
13	267	260
14	478	477
15	178	259
16	423	350
17	427	451

3.3 Plotting Data

The first step is to plot the data and draw the line of equality on which all points would lie if the two meters gave exactly the same reading every time (Figure 3.1).

The degree of agreement can be easily seen from the below graph. A more informative type of plot will be showed in this chapter.

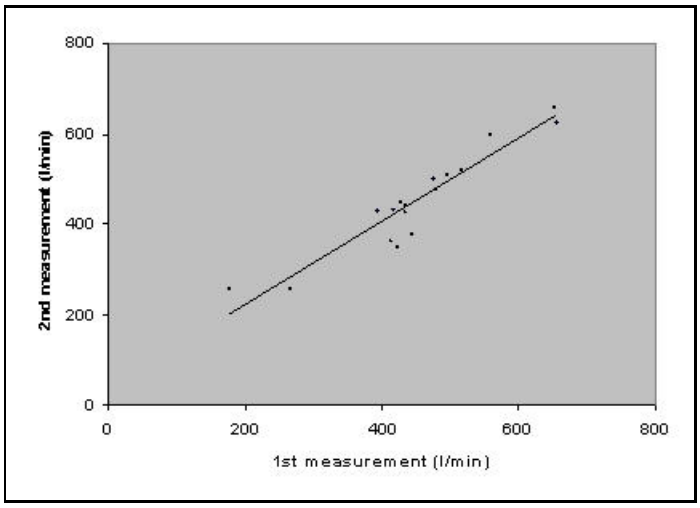


Figure 3.1 First measurement and second measurement, with line of equality

3.4 Inappropriate Use of Correlation Coefficient

The second step is usually to calculate the correlation coefficient (r) between the two methods. The quantity r measures the strength and the direction of a linear relationship between two variables. The mathematical formula for computing r is given in Eq. 3.1.

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2}\sqrt{n(\sum y^2) - (\sum y)^2}} \tag{3.1}$$

where n is the number of pairs of data, x is the first measurement and y is the second.

The value of r is such that -1 $\leq$ r $\leq$ +1.

The plus and minus signs are used for positive linear correlations and negative linear correlations, respectively.

For the data in Figure 3.1, r = 0.94 (p $<$ 0.001). The null hypothesis here is that the measurements by the two methods are not linearly related. The probability is very small and we can safely conclude that the measurements by the two methods are related. However, this high correlation does not mean that the two methods agree:

1. r measures the strength of a relation between two variables, not the agreement between them. We have perfect agreement only if the points in Figure 3.1 lie along the line of equality, but we will have perfect correlation if the points lie along any straight line.

2. A change in scale of measurement does not affect the correlation, but it certainly affects the agreement. For example, if one plots the weights of a number of children against the half-weights of those children, in the style of Figure 3.1, we should get a perfect straight line with slope 2.0. The correlation would be 1.0, but the two measurements would not agree.

3. Correlation depends on the range of the true quantity in the sample. If this is wide, the correlation will be greater than if it is narrow. For those subjects whose first measurement is less than 500 l/min, r is 0.88 while for those with second measurement r is 0.90. Both are less than the overall correlation of 0.94.

4. The test of significance may show that the two methods are related, but the two methods designed to measure the same quantity may not be related. For example, the high correlation of 0.94 for our own data conceals considerable lack of agreement between the two instruments, which is shown below.

3.5 Measuring Agreement

A plot of the difference between the methods against their mean may be more informative. Figure 3.2 displays considerable lack of agreement between the first and second measurements, with discrepancies of up to 80 l/min, these differences are not obvious from Figure 3.1.

The plot of difference against mean shows also any possible relationship between the measurement error and the true value. The true value is not known, and the mean of the two measurements is the best estimate.

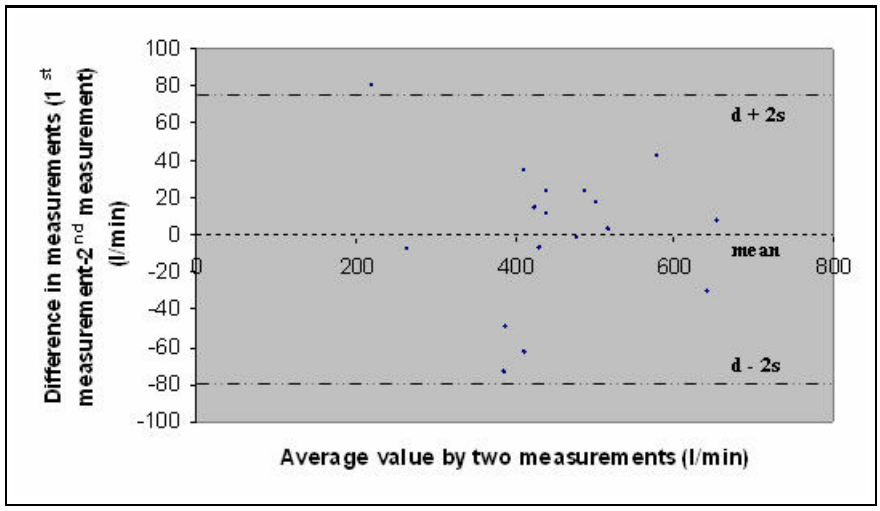


Figure 3.2 Difference against mean for sample data

For our sample data, there is no obvious relation between the difference and the mean. Therefore, we can summarize the lack of agreement by calculating the bias, estimated by the mean difference d, and the standard deviation of the differences (s). For the sample data the mean difference (first measurement minus second measurement) is -2.1 l/min and s is 38.8 l/min. Most of the differences would be expected to lie between d - 2s and d + 2s (Figure 3.2). If the differences are Normally distributed (Gaussian), 95% of differences will lie between these limits (or, more precisely, between d - 1.96s and d + 1.96s).

Provided differences within d +/- 2s would not be clinically important, we could use the two measurement methods interchangeably. We shall refer to these as the "limits of agreement". For our sample data we get:

d - 2s = -2.1 - (2x38.8) = -79.7 l/min

d + 2s = -2.1 + (2x38.8) = 75.5 l/min

Thus, the first measurement may be 80 l/min below or 76 l/min above the second measurement, which would be unacceptable for clinical purposes. This lack of agreement is by no means obvious in Figure 3.1.

3.6 Repeatability

The repeatability of the two methods of measurement limits the amount of agreement which is possible. Poor repeatability means that there is considerable variation in repeated measurements on the same subject. Then the agreement between the two methods is poor too. When the old method is the more variable one, even a new method which is perfect will not agree with it.

Repeatability is measured by taking repeated measurements on a series of subjects. Then a plot similar to Figure 3.2, showing differences against mean for each subject can be drawn. Then the mean and standard deviation of the differences are calculated as before. The mean difference should here be zero since the same method was used. If the mean difference is significantly different from zero, the data can not be used to assess repeatability because either knowledge of the first measurement is affecting the second or the process of measurement is altering the quantity.

3.7 Discussion

In the analysis of measurement method comparison data, the correlation coefficient as shown here is not appropriate. These misleading analyses should be replaced by a method that is simple both to do and to interpret. Further, the same method may be used to analyze the repeatability of a single measurement method.

4. PREVIOUS WORK

Previous studies will be discussed in this chapter:

Several small studies have compared the two methods [5] [6] and have suggested that there is close agreement between them. However, the data from these studies were analyzed by regression analysis, which is inappropriate for this type of comparison.

In Milner's study [6], comparison of two methods of bone-age estimation has been made using the hand and wrist radiographs of 66 boys and 58 girls. An estimate made using the specialist auxological technique of Tanner and Whitehouse was compared with three separate estimates obtained by the atlas matching method of Greulich and Pyle (1971). Two of the latter estimates were carried out by two dedicated radiologists whose results showed close agreement. The third estimate made using the Greulich and Pyle method was derived from the routine reports of a number of radiologists at initial presentation. Only in one narrow age-range for girls was there a significant inter-observer difference, and the reasons for this are discussed. Estimates made using the method of Greulich and Pyle were younger than those made using that of Tanner and Whitehouse. There was a linear relationship between the two methods for the boys but not for the girls. It is suggested that atlas matching methods still have a valuable place in non-specialist hospitals concerned with initial diagnosis rather than with the long-term care of growth problems.

King [5] analyzed the Tanner and Whitehouse II twenty bone (TW2) method of bone age assessment, and compared it with the Greulich and Pyle (G&P) method. 50 previous bone ages were independently re-calculated by each of three registrars using both techniques, with the time taken to perform each assessment being recorded. For each method the interobserver variation was analyzed in terms of the spread of results. The intraobserver variation in TW2 was determined by comparing the bone age originally reported with that subsequently calculated on the same film by the same

registrar. The average spread of results was 0.74 years for TW2 method, and 0.96 years for the GP method and this difference is not statistically significant at the 5% level. The average intraobserver variation to TW2 was 0.33 years, but with 95% confidence limits of -0.87 to +1.53 years. The average time taken was 7.9 min for TW2 and 1.4 min for G&P assessments. It was concluded that the G&P method gave similar reproducibility and was faster than the TW2 method. Following clinical discussion the routine departmental bone age assessment method was changed from the TW2 to the G&P method.

In addition to these studies, what I want to mention here in more details is Edward's Study [7] : They compared the rapid Greulich and Pyle method, as used commonly in clinical practice, with the TW2 method in a large group of subjects. Data were analyzed using the more appropriate "method comparison" technique as we did in our study:

A number of bone age radiographs of the left hand, including the wrist and distal radius were analyzed. The children were aged between 2 and 18 years. The radiographs were assessed according to the method of Greulich and Pyle and then the same radiographs were also assessed by the TW2 method.

Thirty nine of the radiographs ($\tilde{1}0\%$) were then reassessed by both methods by the same observers to assess intra-observer variation for each method.

Their mean age disparity was 0.38 years ($p < 0.01$). Their study was the first of this type to use method comparison scatter plots instead of regression analysis. The 95% confidence interval for the difference between the two methods was 2.28 to -1.52 years. In clinical practice an error of

Their measured intra-observer variation was greater for the Greulich and Pyle method than for the TW2 method (95% confidence limits, -2.46 to 2.18 vs -1.41 to 1.43).

They concluded that the Greulich and Pyle and TW2 methods produce different values for bone age, which are significant in clinical practice. This disagrees with previous smaller studies, all of which were performed by the use of regression analysis, which is an inappropriate statistical technique for this type of study. In addition, they have shown that the TW2 method is more reproducible than the Greulich and Pyle method. They hypothesized also that the rapid Greulich and Pyle method, as used in common clinical practice, is potentially less accurate than the more rigorous time consuming approach.

5. THE COMPUTER ASSISTED BONE AGE ASSESSMENT

5.1 Introduction

In the previous chapter we reviewed the previous studies related to my study. We have had three suggestions in order to improve and advance Edward' s study. We have proposed to use of computer assisted TW2 instead of time consuming version of TW2, to compare rapid G&P and computerized TW2 in terms of accuracy and also speed, and to obtain the more accurate results in shorter time. In this chapter, we will have a look at our computerized bone age assessment method [8].

Our first aim was to update database of left hand radiographs of patients in the software. The user selects the appropriate stages by matching the bone regions of the film to the presented images on the web. The bone age is then calculated automatically. The system was developed using the .Net environment and the C+ programming language and is expected to be used both for clinical and educational purposes. Our next aim was concentrating on comparing these two methods in terms of reliability and speed.

5.2 Computerized Bone Age Assessment Method

Being the software is web-based, it is easily reachable, user-friendly and has no difficulties of installing. The user chooses manually the corresponding stages of the 20 hand and carpal bones utilizing the web-based interface of the TW2 software.

Explanations of stages can be read when the mouse is scrolled over the images (Figure 5.1, Figure 5.2).

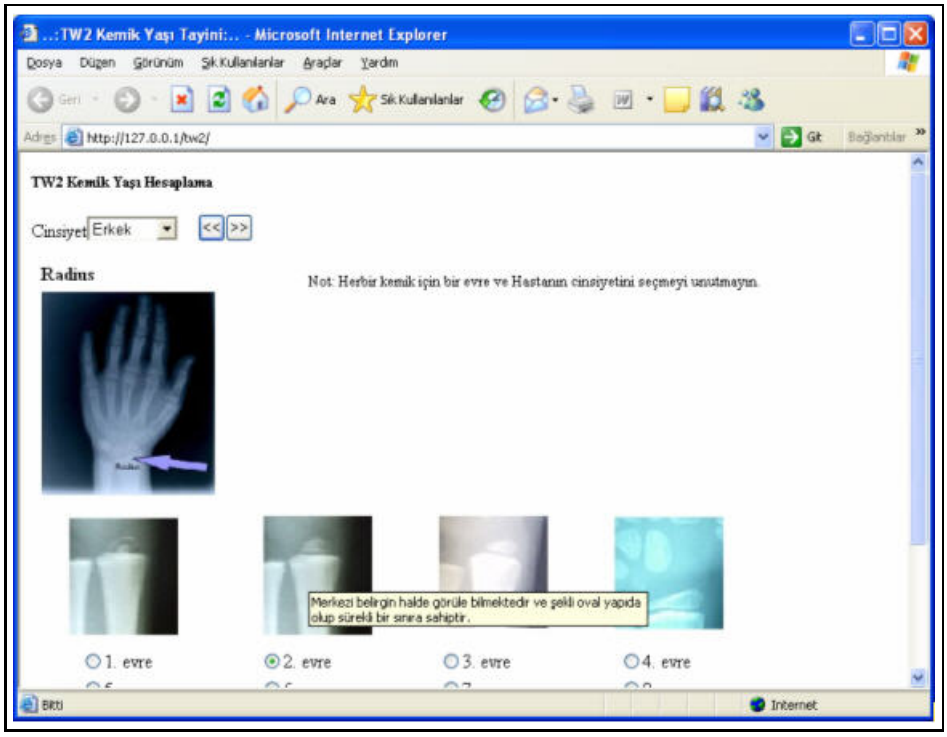


Figure 5.1 The web based bone age assessment system

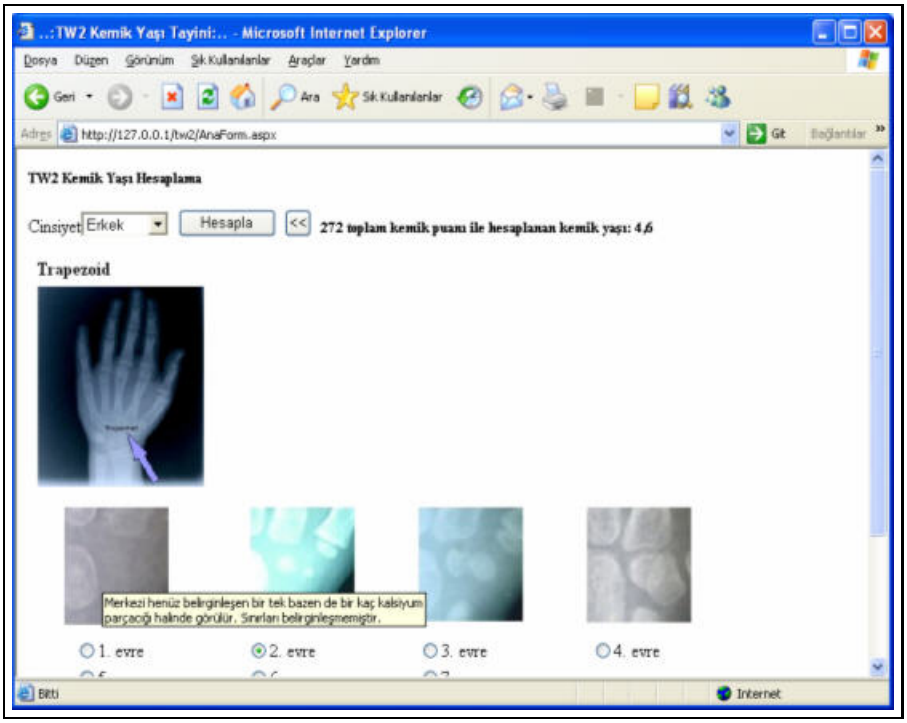
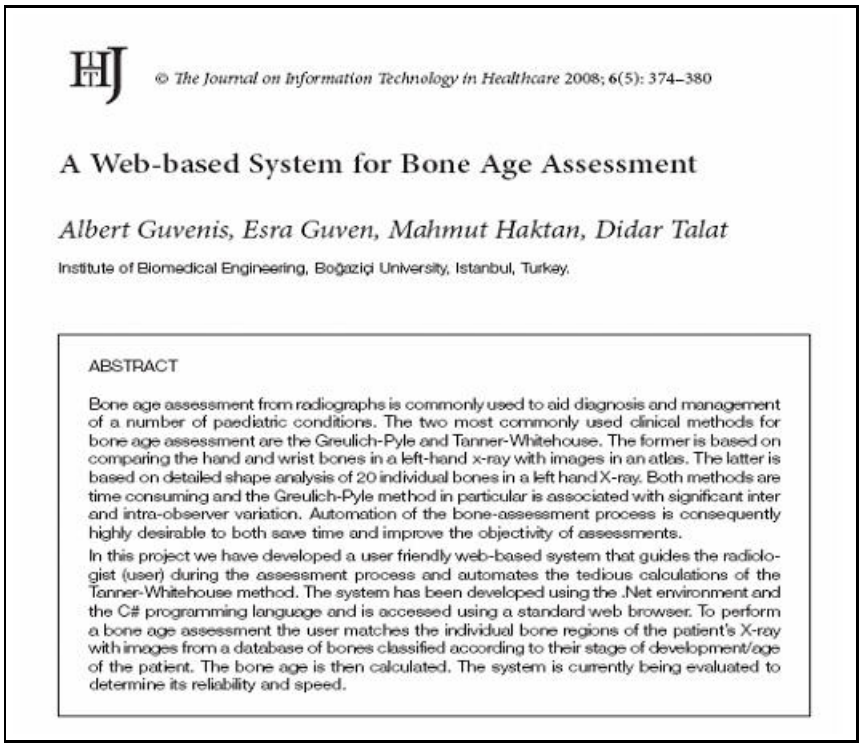


Figure 5.2 Web-Based manual TW2 skeletal age calculation



HTJ © The Journal on Information Technology in Healthcare 2008; 6(5): 374–380

A Web-based System for Bone Age Assessment

Albert Guvenis, Esra Guven, Mahmut Haktan, Didar Talat

Institute of Biomedical Engineering, Boğaziçi University, Istanbul, Turkey.

ABSTRACT

Bone age assessment from radiographs is commonly used to aid diagnosis and management of a number of paediatric conditions. The two most commonly used clinical methods for bone age assessment are the Greulich-Pyle and Tanner-Whitehouse. The former is based on comparing the hand and wrist bones in a left-hand x-ray with images in an atlas. The latter is based on detailed shape analysis of 20 individual bones in a left hand X-ray. Both methods are time consuming and the Greulich-Pyle method in particular is associated with significant inter and intra-observer variation. Automation of the bone-assessment process is consequently highly desirable to both save time and improve the objectivity of assessments.

In this project we have developed a user friendly web-based system that guides the radiologist (user) during the assessment process and automates the tedious calculations of the Tanner-Whitehouse method. The system has been developed using the .Net environment and the C# programming language and is accessed using a standard web browser. To perform a bone age assessment the user matches the individual bone regions of the patient's X-ray with images from a database of bones classified according to their stage of development/age of the patient. The bone age is then calculated. The system is currently being evaluated to determine its reliability and speed.

Figure 5.3 The Journal based on web-based TW2 method

To each bone is assigned an individual score that depends on both stage and gender. The sum of these scores determines the maturity score (See Figure 2.6).

A journal based on this software was published by me, my thesis supervisor Albert Guvenis, Mahmut Haktan and Didar Talat in "The Journal on Information Technology in Healthcare 2008; 6(5):374-380" (Figure 5.3). According to the abstract this software system is currently being evaluated to determine its reliability and speed. My thesis is based on this evaluation, so it was completed already.

6. STATISTICAL EVALUATION OF COMPUTER ASSISTED TW2 METHOD FOR BONE AGE ASSESSMENT

We tested the performance of web-based TW2 scoring system. Two criteria should be tested in our study, the accuracy in comparison to the manual G&P method and the difference between speeds when using this web-based system.

6.1 Materials and Methods

A number (50) of bone age radiographs of the left hand, including the wrist and distal radius, performed in Göztepe SSK hospital for assessment of bone age were analyzed. The children were aged between 4 and 16 years (children aged $<$ 4 years were excluded because bone age assessment from radiographs of the wrist in this age group is unreliable).

Experiments were done by rating each radiograph according to the computerized TW2 method and manual G&P method. Each time, the time needed per method was recorded. For each radiograph twenty hand and carpal bones were examined. The radiographs were assessed by a succession of a Pediatric Endocrinology Specialist. Then, twelve ($\tilde{2}0\%$) of the radiographs were reassessed by the same observer to assess the intra-observer variation for each method.

6.2 Results and Discussion

Statistical analysis involved comparison of bone age assessed by these two methods. Results are shown on a scatter graph (Figure 6.1) plotting mean age as calculated by the two methods against the age disparity between the two methods.

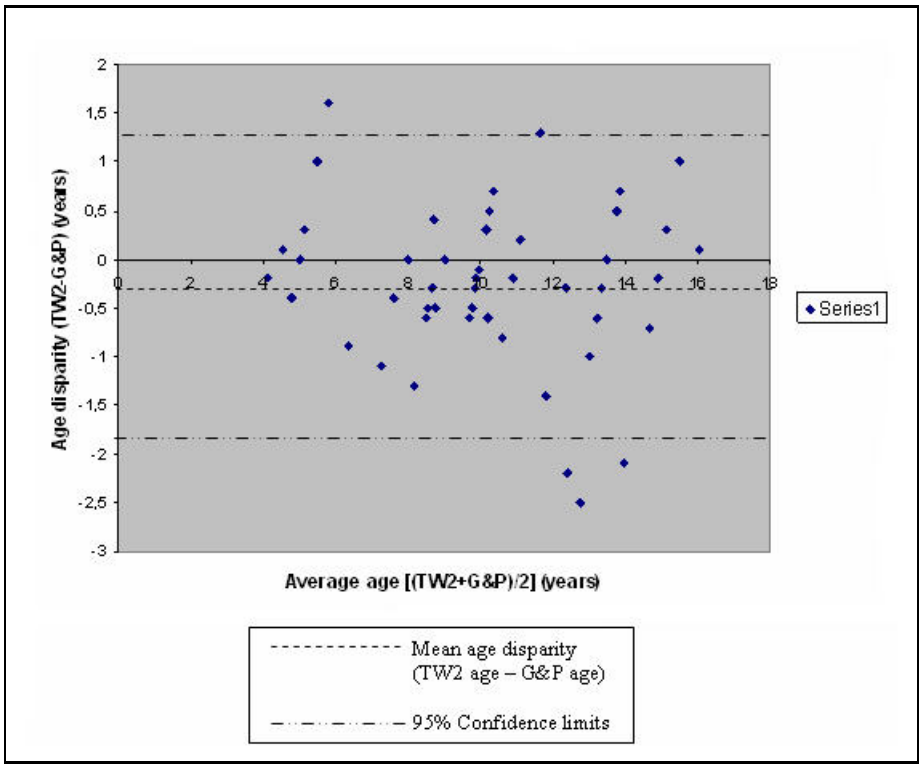


Figure 6.1 Age disparity versus average age of G&P and TW2

The same statistical technique was then used to assess the repeated studies ($\tilde{2}0\%$ of the radiographs) to measure the intra-observer variation. This variation can be compared with the variation between the two methods (Table 6.1).

Time for TW2 method varies between 180 to 90 seconds, whereas time for G&P method varies between 93 to 25 seconds.

Table 6.1
Intra-observer variation and variation between methods

	Mean age disparity(years)	SD of disparity(years)	95%CL
Intra-observer variation of G&P method	0.1	0.44	-0.77 to 0.97
Intra-observer variation of TW2 method	-0.04	0.21	-0.45 to 0.37
Variation between methods	-0.26	0.81	-1.84 to 1.32

For intra-observer variation, calculation of mean age disparity is subtracting first reading from second reading.

For variation methods, calculation of mean age disparity is subtracting G&P age from TW2 age.

Our mean age disparity was 0.26 years (Figure 6.1) and this difference is significantly different from zero (p = 0.005).

We used the dependent means t-test, because in our case the same group was measured twice. The t-value is computed by using Eq. 6.1.

$$t = \frac{d}{s \div \sqrt{n}} \tag{6.1}$$

Where, D is the mean difference between the two observations. s is the standard deviation of the differences, and n is the number of subjects.

The differences against mean bone age are readily interpretable. The range of differences is easily discernible, which is important in clinical usage. The 95% confidence interval for the difference between the two methods is -1.84 to 1.32. In clinical practice an error of this size is unacceptably large.

Our t-value was 2.92, and therefore our p-value was calculated as 0.005 by using a simple p-value calculator or even excel.

If our data were re-analyzed using the inappropriate method of regression analysis (as used in previous studies) the r value obtained was 0.97, which initially appears highly impressive. However, it does not convey the relevant information about absolute and maximum differences between the results of the two techniques. The correlation coefficient measures the strength of an association between two variables, not the agreement between them; furthermore, the wider the range of values being compared (in our case from 4 to 16 years), the greater the correlation. You can also see in Figure

6.1 that the use of correlation is misleading.

Our measured intra-observer variation (Table 6.1) is greater for the Greulich and Pyle method than for the TW2 method (95% confidence limits, -0.77 to 0.97 vs -0.45 to 0.37). This magnitude of intra-observer variation seen for the Greulich and Pyle method probably accounts for much of the discrepancy between the two methods. The subjects on whom the two bone age methods were originally based came from very different social backgrounds. Greulich and Pyle studied American children of high socioeconomic status in the 1940s, whereas Tanner and Whitehouse studied Scottish children of low socioeconomic status in the 1950s. All of the above factors probably contributed to the higher intra-observer variation seen with the Greulich and Pyle method. However, the greatest potential source of error probably comes from the fact that we compared the overall appearance of the radiographs with the standard reference radiographs to obtain the best match. Although this is the approach commonly used, this is not the method originally suggested by Greulich and Pyle. If this more time consuming approach had been used in our study, it is possible that both intra-observer variation and variation between methods would have been reduced.

In addition to these, if we have a look at age dependent agreement, we can say that the agreement between two methods is worse for ages more than 10. 95% confidence limits are -0,22 $\pm$ 1,17 for ages less than 10, whereas the limits are -0,31 $\pm$1,95 for ages more than 10. Our measured intra-observer variation is greater for ages more than 10 (95% confidence limits, -0.22 to 0.17 vs -0.65 to 0.55). The reason for these results comes probably form the fact that in older children the carpal bones do not provide any useful information about the skeletal maturity.

Our study is the second of this type to use method comparison scatter plots instead of regression analysis. But we took the speed of both methods also into consideration. Average time of the last 5 readings for computerized TW2 method was 1,7 min, whereas average time of the last 5 readings for G&P method was 0,7 min. Normally, TW2 assessment without the aid of computer takes about 15 min., which is too long and not practical for clinical purposes. By using this software, time for TW2

assessment was reduced from 15 min to 2 min.

Furthermore, as one more practices with computer assisted TW2, one becomes faster (see Figure 6.2). The average time for the first 5 readings was 3,3 min., whereas the average time for the last 5 readings was 1,7 min.

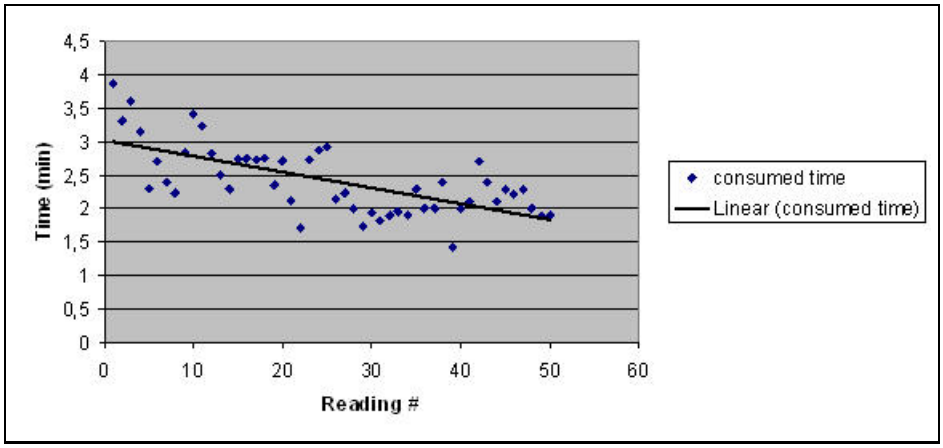


Figure 6.2 Decreased time values consumed during measurement with computer assisted TW2 method

7. CONCLUSIONS AND FUTURE WORK

We conclude that the Greulich and Pyle and computerized TW2 methods produce different values for bone age, which are significant in clinical practice. This disagrees with previous smaller studies, which were analyzed by using correlation coefficients, which is an inappropriate statistical technique for this type of study. Our results were very similar to Edward' s study. We have shown that the TW2 method has higher intra-observer repeatability than the Greulich and Pyle method. Therefore we suggest that; since both methods take relatively little time, computerized TW2 method should be preferred for higher performance in bone age assessment.

To sum up, computer assisted TW2 method is more reproducible, reasonably fast, widely accessible and also in Turkish language. Therefore we recommend the computerized TW2 method for bone age assessment.

In the future, further development of this software can be done by further testing with larger sample sizes and also by implementing the system in more hospitals. Also, the system can be put to work for educational purposes

APPENDIX A. RESULTS OF THE SAMPLES

The results of the samples are given in Table A.1 and Table A.2. The results continue in Table A.2

Table A.1
Results of the samples

Subject ID	Web-based TW2 bone age calculation (years)	Time for TW2 method (years)	Bone age with manual G&P method (sec)	Time for G&P method (sec)
Ö.S.	8.9	232	8.5	52
M.D.	13.5	199	13.5	93
N.G.	9.9	210	10.5	48
N.Ç.	7.5	189	8.8	44
G.E.B.	8.5	138	8.8	50
D.T.	5.9	161	6.8	35
Ö.F.B.	14	146	13.5	49
N.T.	14.2	133	13.5	52
Y.A.	11.1	171	12.5	46
F.A.	8.5	205	9	70
G.D.	10.3	194	10	64
N.Ş.	8.3	170	8.8	40
P.S.	11.5	138	14	35
A.H.	8	165	8	38
G.K.	4	166	4.2	45
A.G.S.	6.7	164	7.8	38
G.A.	9.9	165	10	55
M.P.	12.3	142	11	46
O.C.	4.6	162	4.5	55
K.Y.	10.8	127	11	42
Ş.D.	14.3	102	15	35
M.D.	12.9	164	13.5	36
İ.Y.	9.7	172	10	40
M.N.A.	10.5	175	10	70
D.Y.	9.5	129	10	75

Table A.2
Results of the samples

Subject ID	Web-based TW2 bone age calculation (years)	Time for TW2 method (years)	Bone age with manual G&P method (sec)	Time for G&P method (sec)
M.K.T.	12.5	133	13.5	34
M.Ç.	6	121	5.5	46
E.Ö.	14.8	104	15	52
M.K.T.	13.2	116	13.5	55
U.K.	5.3	169	5.5	65
E.Y.	7.4	114	7.8	42
H.B.	12.9	117	15	36
C.Ç.	15.3	114	15	30
G.G.	8.2	142	8.8	44
B.C.Y.	6.6	126	5	40
F.Y.	4.6	124	5	48
O.B.	9	144	9	55
Ş.E.	16	85	15.5	25
Ö.G.	16.1	119	16	35
İ.T	11.2	136	11.5	65
D.A.	9.4	167	10.5	64
N.A.	8.3	149	8.8	57
E.D.	9.8	137	10.5	48
B.Y.	12.2	153	12.5	70
O.D.	10.7	150	10	40
P.Ş.	11.3	144	13.5	28
M.Ş.	10.2	145	11	35
E.N.	14.3	113	15	40
M.V.Ö.	9.9	163	10	55
M.K.	5	150	5	42

REFERENCES

1. Greulich W.W, P. S., and W. A.M., *A Radiographic Standard of Reference for the Growing Hand and Wrist*, Chicago: Case Western Reserve University, 1971.

2. Tanner J.M., W. R., and C. N., *Assessment of Skeletal Maturity and Prediction of Adult Height*, London: Academic Press, 2nd ed., 1983.

3. W.W., G., and P. S.I., *Radiographic Atlas of Skeletal Development of Hand and Wrist*, Standford University Press, 2nd ed., 1971.

4. J.M., B., and A. D.G, "Statistical method for assessing agreement between two methods of clinical measurement," *Lancet*, Vol. 8, pp. 307–310, Feb 1986.

5. King D.G., S. D., and O. M.P., "Reproducibility of bone ages when performed by radiology registrars: an audit of tanner and whitehouse 2 versus greulich and pyle methods," *British Journal of Radiology*, Vol. 67, pp. 848–851, 1994.

6. Milner G.R., L. R., and K. R., "Assessment of bone age: a comparison of the greulich and pyle and the tanner and whitehouse methods," *Clinical Radiology*, Vol. 37, pp. 320–327, 1986.

7. Bull R.K., E. P., and K. P. F. S., "Bone age assessment: a large scale comparison of the greulich and pyle, and tanner and whitehouse (tw2) methods," *Arch Dis Child*, Vol. 81, pp. 172–173, Aug 1999.

8. Güven, E., *Computer assisted TW2 method*, Boğaziçi University, 1st ed., 2008. Available: http:/www.bme.boun.edu.tr/esraguven.