J Stoma 2015; 68, 2: 145-158
DOI: 10.5604/00114553.1156237

ORIGINAL ARTICLE

Testing the compatibility of measurement methods in the analysis of cephalometric radiographs

Badanie zgodności metod pomiarowych w analizie zdjęć cefalometrycznych

Zbigniew Paluch[1], Robert Partyka[2], Grażyna Lisowska[3], Maciej Misiołek[3]

[1] **Prywatna Przychodnia Ortodontyczna, Racibórz, Polska**
Private Orthodontic Practice, Racibórz, Poland
Head: dr *Z. Paluch*

[2] **Katedra Anestezjologii, Intensywnej Terapii i Medycyny Ratunkowej, Śląski Uniwersytet Medyczny w Katowicach, Polska**
Department of Anesthesiology, Intensive Treatment and Emergency Medicine, Medical University of Silesia in Katowice, Poland
Head: prof. dr hab. *P. Jałowiecki*

[3] **Katedra i Oddział Kliniczny Otorynolaryngologii i Onkologii Laryngologicznej w Zabrzu, Śląski Uniwersytet Medyczny w Katowicach, Polska**
ENT Department in Zabrze, Medical University of Silesia, Poland
Head: prof. dr hab. *G. Namysłowski*

Abstract

Introduction. *The comparison of the results obtained by different methods has been the subject of many scientific publications over the years. A number of discussions are related to statistical analyses used to test the compatibility of the results obtained by different research techniques. There has been no consensus on the choice of the appropriate statistical method, and so the issue of the compatibility of the results still constitutes the area for further investigation related to the application of methods which would constitute reliable tests indicating a higher precision of their compatibility character.* ***Aim of the study.*** *To test the compatibility of two methods of cephalometric measurements using three comparative methods of diverse sensitivity for the assessment of the compatibility of the tested results.* ***Material and methods.*** *In the first method, cephalometric radiographs were traced and measured manually with trace foil. In the second method, landmarks were identified by a mouse-driven cursor on the screen. Twelve values were measured for randomly selected 14 subjects. Pearson's correlation coefficient, the Mann-Whitney U test and the Wilcoxon test were used to assess the compatibility character of both methods.*

Streszczenie

Wstęp. *Porównywanie wyników otrzymanych różnymi metodami jest od lat częstym tematem publikacji naukowych. Wiele dyskusji dotyczy analiz statystycznych, używanych dla sprawdzenia zgodności wyników uzyskanych za pomocą różnych technik badawczych. Nie osiągnięto konsensusu w kwestii wyboru właściwej metody statystycznej, co skutkuje faktem, iż temat zgodności wyników wciąż stanowi obszar do dalszych poszukiwań stosowania metod, które byłyby wiarygodnymi testami wskazującymi na większą precyzję ich zgodności.* ***Cel pracy.*** *Celem pracy było testowanie zgodności dwóch metod pomiarów cefalometrycznych, przy użyciu trzech metod porównawczych, o zróżnicowanej czułości, dla oceny charakteru zgodności testowanych wyników.* ***Materiał i metody.*** *W pierwszej metodzie zdjęcia cefalometryczne z umieszczoną na nich specjalną kalką zostały wykreślane i mierzone ręcznie. W drugiej metodzie punkty zostały zidentyfikowane przy użyciu kursora myszki na ekranie. Dla losowo wybranych 14 badanych dokonano pomiaru dwunastu wartości. W celu oceny charakteru zgodności obu metod zastosowano współczynnik korelacji Pearsona, test U Manna-Whitney'a i test Wilcoxona.* ***Wyniki.***

KEYWORDS:
cephalometry, compatibility, cephalogram

HASŁA INDEKSOWE:
cefalometria, zgodność, cefalogram

Results. *Both correlations at the significance level of 0.05 and the Mann-Whitney U test did not show statistically significant differences between both methods. Statistically significant differences were obtained for 9 of 12 measured values in the Wilcoxon test at the same level of significance.* ***Conclusions.*** *Correlations and tests based on the comparison of the mean value are not sufficiently sensitive and they do not indicate statistically significant differences between the results.*

Pomiędzy obiema metodami zarówno korelacje na poziomie istotności 0,05, jak i test U Manna-Whitneya nie wykazały statystycznie istotnych różnic. W teście Wilcoxona na wskazanym poziomie istotności, statystycznie istotne różnice otrzymano dla 9 z 12 zmierzonych wartości. ***Wnioski.*** *Korelacje i testy oparte na porównaniu średniej wartości nie są wystarczająco czułe oraz nie wykazują istotnych statystycznie różnic pomiędzy wynikami.*

Introduction

Cephalometric analysis is performed using various tools and different images, i.e. analog or digital pictures. Analog images can be obtained from a laser printer (Laser Imager) as one of the methods of long-term archiving. They can also be available in the case of analog X-ray devices. However, digital image analyses are currently becoming more widely used. The comparison of the results obtained by different methods has been the subject of many scientific publications over the years and it will certainly be continued due to the development of cephalometric techniques in orthodontics.[1-10] There are still discussions related to statistical analyses used to test the compatibility of the results obtained by different research techniques. Some of them – as a correlation method – have been criticized by many researchers[11,12] as an insufficiently sensitive method of comparison. A similar situation was observed in the case of methods based on the comparison of the mean from the obtained results.[13] There is no consensus on the choice of the appropriate statistical method, which results in the fact that the issue of the compatibility of the results still constitutes the area for further investigation related to the application of methods which are reliable tests to indicate a higher precision of their compatibility character. The aim of the comparison is not always the indication of a more precise method, as it is usually obvious for objective reasons.

A number of studies have shown that computerized cephalometric methods are superior to manual methods due to the possibility of using a series of available facilities related to landmark

Wstęp

Analiza cefalometryczna jest wykonywana przy użyciu różnych narzędzi i różnych obrazów, to jest analogowych lub cyfrowych zdjęć. Zdjęcia analogowe mogą być pozyskiwane przez wydruk z drukarek z formy cyfrowej, jako jedna z metod długoczasowego archiwizowania. Mogą być także dostępne w przypadku stosowania jeszcze analogowych aparatów radiologicznych. Jednak obecnie analizy cyfrowego obrazu są stosowane coraz bardziej powszechnie. Porównanie wyników uzyskanych za pomocą różnych metod jest od lat częstym tematem wielu publikacji i z pewnością będzie kontynuowane w związku z rozwojem technik cefalometrycznych w ortodoncji.[1-10] Dyskusje nie milkną odnośnie analiz statystycznych stosowanych do testowania zgodności wyników otrzymanych za pomocą różnych technik badawczych. Część z nich, jako metoda korelacji została skrytykowana przez wielu badaczy,[11,12] jako mało czuły sposób porównania. Podobnie było z metodami, które opierały się na porównaniu średniej z uzyskanych wyników.[13] Brak jest jednoznacznego rozstrzygnięcia, dotyczącego wyboru właściwej metody statystycznej, co powoduje, iż kwestia zgodności wyników wciąż stanowi przestrzeń do dalszych poszukiwań dla zastosowania metod, będących miarodajnymi testami, dla wskazania większej precyzji ich charakteru zgodności. Porównanie nie zawsze ma na celu wskazanie bardziej dokładnej metody, ponieważ jest to zwykle oczywiste ze względów obiektywnych.

W wielu badaniach wskazano, że cefalometryczne metody komputerowe mają przewagę nad manualnymi dzięki możliwości zastosowania sze-

identification and much shorter time. The points of interest can be determined with the mouse cursor, and the lines and angles are measured automatically using a computer program.[3,6,14] However, orthodontists occasionally encounter in one subject two types of images for analysis, i.e. digital and analog taken at different time periods. This situation also occurred when the stability of treatment results was checked after a long period of time and the documentation consisted of different types of registered radiological images. Similar situations occurred in long-term studies which concerned the observation of the growth and aging of the examined anatomical structures, based on different types of radiological images. Manual landmark identification was also necessary in the case of interdisciplinary treatment. Furthermore, many results of scientific research in orthodontics, which are still used, were obtained by manual landmark identification. In a number of situations it is necessary to make sure if the compatibility of the results obtained by both methods is high enough so that they could be used interchangeably.[8-10,15,16]

Aim of the study

The aim of the study was to test the compatibility of two methods of landmark identification used in cephalometric measurements by means of three comparative methods of diverse sensitivity for the assessment of the compatibility of the results tested.

Material and methods

A total of randomly selected 14 lateral cephalometric radiographs (LCRs) were enrolled in the study based on the patient documentation following initial orthodontic diagnosis at the orthodontic outpatient clinic. The inclusion criteria were as follows: the presence of erupted upper and lower primary molars and permanent incisors with closed root apices. Type of malocclusion, sex and age were not considered. Subjects were positioned in the cephalostat with the Frankfort plane parallel to the floor. The cephalograms were obtained using ORTHOPANTOMOGRAPH OP200, TUUSULA FINLAND.

Manual and digital methods of landmark

regu dostępnych udogodnień związanych z identyfikacją punktu oraz wykonanie w znacznie krótszym czasie. Punkty te można wyznaczyć za pomocą kursora myszy, a odległości i kąty mierzone są automatycznie za pomocą programu komputerowego.[3,6,14] Zdarza się czasem, że ortodonci mają do dyspozycji u jednej badanej osoby dwa typy obrazu do analizy: cyfrowy i analogowy, wykonane w różnych okresach czasu. Taka sytuacja powstawała także, gdy sprawdzano stabilność wyników leczenia po długim okresie czasu, a dokumentacja składała sie z różnych rodzajów zarejestrowanych obrazów radiologicznych. Podobne sytuacje występowały w badaniach długoczasowych, dotyczących obserwacji wzrostu i starzenia się badanych struktur anatomicznych na podstawie istniejących różnych typów obrazów radiologicznych. Także w przypadku leczenia interdyscyplinarnego konieczna była ręczna identyfikacja punktów. Wiele wyników badań naukowych w zakresie ortodoncji, które nadal są wykorzystywane, uzyskano na podstawie pomiarów zbieranych metodą ręcznej identyfikacji punktów. W wielu przypadkach jest istotne, aby mieć pewność, że zgodność wyników uzyskanych obiema metodami jest na tyle duża, że można je stosować zamiennie.[8-10,15,16]

Cel pracy

Celem pracy było testowanie zgodności dwóch metod wyznaczania punktów zastosowanych w pomiarach cefalometrycznych, przy użyciu trzech metod porównawczych, o zróżnicowanej czułości, dla oceny charakteru zgodności testowanych wyników.

Materiał i metody

Z dokumentacji pacjentów, którzy przeszli wstępną diagnostykę ortodontyczną w przychodni ortodontycznej wybrano losowo 14 zdjęć bocznych cefalometrycznych (LCR). Kryteriami włączenia były: obecność wyrzniętych z zamkniętymi wierzchołkami korzeni górnych i dolnych pierwszych trzonowców oraz stałych siekaczy. Zarówno typ wady, płeć, jak i wiek kalendarzowy nie były brane pod uwagę przy wyborze. Badanych pozycjonowano w cefalostacie względem płaszczyzny frankfurckiej, równoległej do

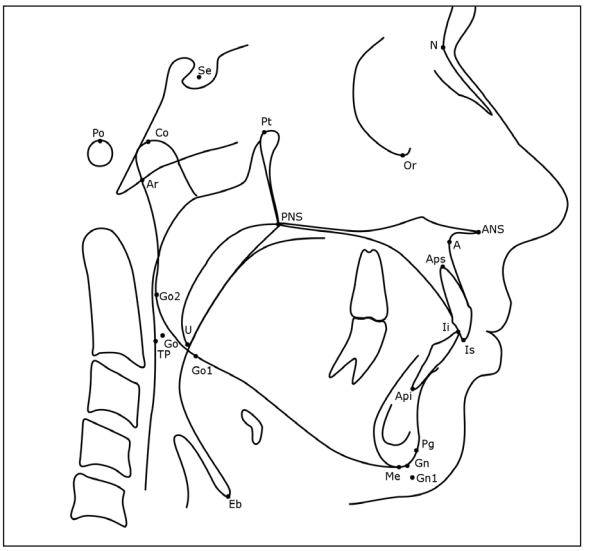

Fig. 1. *Cephalometric landmarks used in this study.*
Punkty cefalometryczne użyte w badaniu.

identification and landmark measurement were used in the present study. Analog images were printed from the digital picture (DryView 6850 Laser Imager). One examiner (ZP) selected 25 landmarks from different anatomic regions (Fig. 1). In the first method, LCRs were traced and measured manually with trace foil Dentaurum, using a 0.35 mm 2H pencil, a cephalometric protractor and a caliper (3M Unitek® Corporation). The LCRs were traced and measured in a darkened, quiet room using a light-viewing box. Each set of landmarks was traced eight times at weekly intervals. Next, five angular and seven linear variables were selected (Table 1). Twelve variables were measured manually eight times – seven linear variables were measured using a caliper and five angular variables were measured with a protractor.

According to the second method, a digital image format was copied from the compact disc to the hard drive. Next, the digital picture was displayed on the 20-inch LCD monitor (Dell Ultra Sharp 2007FP TCO99 Monitor) with a resolution of 1600 x 1200px/60Hz. The monitor was placed in a darkened room. DesignCAD software combined with the author's (ZP) original overlays package (Orto-TestPor-ZPaluch) written in Microsoft Visual

podłogi. Cefalogramy były wykonane z użyciem ORTHOPANTOMOGRAPH OP200, TUUSULA FINLAND.

Do identyfikacji punktów i ich pomiarów użyto metod: ręcznej i cyfrowej. Analogowe zdjęcia były drukowane z obrazu cyfrowego przez Laser Imager 6850 DRYVIEW. Jeden specjalista (ZP) wyznaczał 25 punktów z różnych okolic anatomicznych (Fig. 1). W pierwszej metodzie LCR były wykreślane i mierzone ręcznie na specjalnej kalce firmy Dentaurum, przy użyciu ołówka 0,35 mm 2H, cefalometrycznego kątomierza i suwmiarki firmy 3M Unitek® Corporation. LCR były umieszczone na negatoskopie, wykreślane i mierzone w ciemnym, cichym pokoju. Każdy zestaw punktów był wyznaczany ośmiokrotnie w odstępach tygodniowych. Następnie, wybrano 5 kątowych i 7 liniowych zmiennych (Tab. 1). Ośmiokrotnie mierzono ręcznie dwanaście zmiennych – 7 liniowych zmiennych mierzono cyrklem, a 5 zmiennych kątowych mierzono kątomierzem.

Według drugiej metody, cyfrowy obraz skopiowano z płyty CD na dysk twardy. Następnie obraz cyfrowy został wyświetlony na 20 calowym monitorze LCD (Dell Ultra Sharp 2007FP TCO99 Monitor) o rozdzielczości 1600 x 1200 pikseli/60Hz. Monitor był umieszczony w ciemnym pokoju. Do analizy zdjęć radiologicznych użyto oprogramowania DesignCAD w połączeniu z autorskim (ZP) programem (Orto-TestPor-ZPaluch) napisanym w Microsoft Visual C++. NET. Uzyskane wyniki były eksportowane do plików tekstowych w formacie CVS. W celu wyeliminowania błędów systematycznych, w metodzie ręcznej pomiary cefalometryczne skorygowano z uwzględnieniem współczynnika powiększenia, a obrazy cyfrowe skalibrowano. Każdy z obrazów miał podziałkę, na której wyznaczono dwa punkty. Wszystkie wymiary liniowe były izotropowo skalowane, aby uzyskać wartości rzeczywiste. Za pomocą oprogramowania Orto-TestPor-ZPaluch autor wyznaczył ośmiokrotnie punkty kursorem myszki na monitorze w przerwach tygodniowych. Następnie punkty były łączone w linie i kąty, które mierzono także programem Orto-TestPor-ZPaluch.

Celem analizy statystycznej było porówna-

Table 1. Variables used for this study

ANS-Me	Anterior lower facial height
Eb-PNS	Vertical airway length (where Eb was the deepest point of the epiglottis)
Se-N	Anterior cranial base (where Se was the midpoint of the entrance to the sella turcica)
ANS-PNS	Palatal plane
U-TP	Retropalatal airway space (where U was the most inferior posterior point of the uvula; whereas TP landmark was redefined as the construction point on the posterior pharyngeal wall at the intersection of the line from U parallel to the Po-Or line)
Co-A	Midfacial length
Co-Gn	Mandibular length (where Gn was the point of the intersection of Pt-Gn1 on the contour of the chin; whereas Gn1 indicated the construction point at the intersection of Go-Me and N-Pg; and Go was the point of the intersection of two lines of the region of the gonial angle of Ar-Go2 and Me-Go1)
NSL-NL	Angle formed between SN (NSL) and the palatal (NL) planes
NSL-ML	Angle between SN and the mandibular (ML) planes
ML-NL	Vertical jaw relationship
NL-U1	Angle formed between the palatal plane and upper incisor axis (U1)
ML-L1	Angle formed between the mandibular plane and lower incisor axis (L1)

C++.NET was used to analyze the radiographs. The obtained results were exported to text files in the CVS format. In order to eliminate systematic errors, in the manual method cephalometric measurements were corrected with consideration given to the zoom factor and the digital images were calibrated. Each image had a scale provided, on which two landmarks were identified. All linear measurements were isotropically scaled to obtain real values. Using the Orto-TestPor-ZPaluch software, the author designated landmarks eight times at weekly intervals, with a cursor on the monitor. Next, the landmarks were connected into lines and angles, which were also measured using Orto-TestPor-ZPaluch.

The aim of the statistical analysis was to compare the mean values from the measurements obtained eight times using analog and digital cephalometry with the use of three statistical methods.

The correlation analysis was used to determine whether the values of two measurement methods were related and what type of dependency existed. To compare the results, two non-parametric

nie średnich wartości z pomiarów wykonanych ośmiokrotnie za pomocą analogowej i cyfrowej cefalometrii przy użyciu trzech metod statystycznych.

Analiza korelacji była użyta do ustalenia, czy wartości dwóch metod pomiarowych były powiązane oraz jaki istniał rodzaj zależności. W celu porównania wyników zastosowano dwie metody nieparametryczne test U Manna-Whitney'a (M-W) oraz test par Wilcoxona (W). Każda z zastosowanych metod wymagała innych założeń. Dlatego też otrzymane wyniki były odpowiednio zinterpretowane. W pierwszym z testów potraktowano zmierzone wartości jako dwie niezależne grupy. W drugim wzięto pod uwagę różnicę wartości uzyskanych w dwóch metodach dla poszczególnych osób. Obie metody nieparametryczne zostały wybrane, ponieważ były najbardziej odpowiednie dla prób o małych liczebnościach, z jakimi mieliśmy do czynienia w badaniu. Przeprowadzone testy nieparametryczne były przydatne w celu stwierdzenia, czy wyniki porównania były uzależnione od rodzaju metody, jak również ze względu

methods were employed i.e. the Mann–Whitney U test (M-W) and The Wilcoxon signed-rank test (W). Since each method required different assumptions, the obtained results were appropriately interpreted. In the first test, the measured values were treated as two independent groups. In the second test, a difference of the values obtained in two methods for particular subjects was considered. Both non-parametric methods were chosen because they were the most suitable for small size samples, as is the case in the current study. The conducted non-parametric tests were useful to determine whether the comparison results depended on the type of the method. They were also useful due to different test sensitivity and the range of each test. The W test examined the hypothesis on the zero value of the median of the differences between values obtained with two methods for particular cephalograms. The M–W test checked the equality of the mean value for both samples tested. In both cases it was assumed that the tested variable was measured on an ordinal scale, which allowed ranking the differences. R-Project program was used to conduct both tests.

Results

Based on the obtained measurements, which were repeated eight times, mean values and the standard deviation (SD) were calculated for fourteen cephalograms. This procedure was performed for all twelve examined values for both methods. Tables 2 and 3 show the comparison results of the values obtained by manual tracing and by a computer method, respectively. All the performed tests considered the mean values in the tables.

The correlation analysis was used to determine whether the values of two measurement methods were related, and what kind of dependency existed. The Pearson's correlation coefficient "r" for a sample size of fourteen data pairs was calculated (second column in Table 4). Only the average value measurements were considered in the analysis, due to a low value of SD for each subject compared to the differences between them.

Figure 2 presents the results of twelve measured values. The first parameter (horizontal axes) was

na różną czułość testów i zakres tego, co dany test bada. Test W testował hipotezę o zerowej wartości mediany różnic wartości otrzymanych dwoma metodami dla poszczególnych zdjęć cefalometrycznych, test M-W testował równość wartości średniej dla obu prób badawczych. W obu przypadkach zakładano, że testowana zmienna mierzona była na skali porządkowej, co umożliwiło rangowanie różnic. Do przeprowadzenia obu testów posłużono się programem R-Project.

Wyniki

Na podstawie uzyskanych pomiarów, które powtórzono ośmiokrotnie, dla 14 badanych cefalogramów obliczono wartości średnie i odchylenie standardowe (SD). Niniejsza procedura była przeprowadzona dla wszystkich 12 badanych wartości dla obu metod. Tabela 2 przedstawia wyniki porównania wartości otrzymanych w ręcznej metodzie, a tabela 3 wartości otrzymane metodą cyfrową. Wszystkie wykonane testy uwzględniały średnie wartości w tabelach.

Analizę korelacji użyto do ustalenia, czy wartości dwóch metod pomiarowych były powiązane i jaki był rodzaj zależności. Współczynnik Pearsona korelacji r obliczono dla próby wielkości 14 par danych (druga kolumna w tabeli 4). W analizie uwzględniono jedynie średnie wartości pomiarów ze względu na niską wartość SD dla każdego badanego w porównaniu z różnicami pomiędzy nimi.

Na rycinie 2 przedstawiono wyniki dwunastu zmierzonych wartości. Pierwszy parametr (poziome osie) otrzymano z LCR, które były zmierzone ręcznie, a drugą zmienną (pionowa oś) mierzono za pomocą programu komputerowego. Punkty były rozmieszczone wokół lub na prostej (o nachyleniu równym współczynnikowi Pearsona), który przedstawiał typ korelacji.

W celu uzyskania statystycznie istotnego współczynnika korelacji, obliczyliśmy 95% przedział ufności, to jest zakres wartości, które zawierają 95% współczynnik przedziału ufności. Tabela 4 przedstawia wartości współczynnika Pearsona dla dwóch metod pomiarów dla 12 wartości (kolumna A w tabeli 4) i 95% przedział ufności (kolumna B w tabeli 4).

Table 2. Manual measurements

		1	2	3	4	5	6	7	8	9	10	11	12	13	14
Linear measurements [mm]															
ANS-Me	$\bar{x}$	60.5	53.8	68.7	50.8	61.1	66.9	70.6	64	57.8	63.2	66	70.5	55.1	60.3
	σ	0.6	0.5	0.4	0.7	0.4	0.6	0.5	0.5	0.5	0.6	0.5	0.4	1.1	0.6
Eb-PNS	$\bar{x}$	58.1	46.4	69.8	49.3	55.9	66.5	64.1	65.1	43.9	61.2	64.2	58.2	49.5	53.5
	σ	3.1	0.4	0.5	0.4	0.5	0.3	1.3	0.7	0.5	0.8	0.7	0.4	0.6	0.6
Se-N	$\bar{x}$	65.3	60.8	64.3	62.5	61.8	65.8	63.3	65.7	65.4	71.7	62.6	72.0	60.8	63.9
	σ	0.2	0.4	0.9	0.5	0.3	0.5	0.7	0.6	0.5	0.6	0.4	0.4	0.8	0.6
ANS-PNS	$\bar{x}$	53.2	47.6	52.2	51.9	47.5	52.9	55.3	49.3	48.6	51.8	53.3	54.1	46.2	52.2
	σ	1.8	1.3	0.5	0.9	1.2	0.7	0.7	0.6	0.8	1.5	0.5	0.7	1.0	1.1
U-TP	$\bar{x}$	7.4	12.4	10.0	7.3	6.1	13.8	10.9	8.3	13.5	11.6	11.6	12.3	13.6	4.3
	σ	0.8	0.7	0.3	0.3	1.1	1.0	0.2	0.2	0.9	0.6	0.7	0.5	0.7	0.2
Co-A	$\bar{x}$	85.1	76.0	84.3	79.6	82.8	85.7	85.9	83.7	77.7	84.6	88.9	90.3	77.7	84.2
	σ	0.8	1.0	1.1	1.2	0.9	0.5	1.5	1.2	0.6	0.8	1.0	0.5	0.4	0.7
Co-Gn	$\bar{x}$	105.8	94.0	113.4	99.5	105.3	115.8	116.3	102.8	100.0	111.6	110.3	113.2	98.9	101.9
	σ	0.6	0.8	1.4	1.3	1.1	0.6	1.4	1.3	0.7	1.2	0.3	0.3	0.3	0.5
Angular measurements [°]															
NSL-NL	$\bar{x}$	7.4	9.7	15.3	10.1	11.3	4.8	8.0	6.8	6.1	8.0	7.9	2.5	5.5	6.4
	σ	1.1	0.7	1.0	0.6	0.4	0.9	0.9	0.6	0.8	1.2	0.8	0.9	1.6	0.9
NSL-ML	$\bar{x}$	29.8	30.8	42.5	28.0	36.7	28.8	32.5	40.6	29.4	35.4	37.1	33.9	32.5	30.2
	σ	1.3	0.8	0.6	0.6	0.4	0.8	0.5	0.6	0.7	0.7	0.6	1.2	2.0	0.3
ML-NL	$\bar{x}$	22.8	21.4	27.3	18.0	25.6	24.1	25.4	34.1	23.7	28.9	29.3	31.8	27.1	23.9
	σ	0.8	0.7	0.6	0.6	1.2	0.4	1.9	0.7	0.9	3.9	0.6	0.5	1.2	0.8
NL-U1	$\bar{x}$	117.3	108.8	108.2	115.3	108.8	112.4	110.4	114.1	100.1	107.5	112.1	112.1	105.7	105.9
	σ	1.6	1.0	1.5	0.5	1.1	1.7	0.8	0.9	1.2	1.7	1.8	1.2	1.8	1.6
ML-L1	$\bar{x}$	96.9	99.2	92.7	103.6	86.9	83.0	88.3	93.5	92.2	94.1	99.0	102	86.4	99.2
	σ	1.3	7.3	1.3	1.7	1.4	1.8	1.3	1.8	1.5	0.6	1.2	1.1	2.1	1.6

obtained from the LCRs which were measured manually, and the second variable (vertical axis) was measured by a computer program. The points were distributed around or on the straight line (with a slope equal to the Pearson's coefficient), which illustrated the correlation type.

To obtain a statistically significant correlation coefficient, 95% confidence interval was calculated, that is the range of values that contain

Wyniki z 11 danych wskazywały na bardzo silną ($0,9 < r < 1$) oraz dodatnią korelację liniową pomiędzy dwiema zmiennymi. Ponadto, przedziały ufności były bardzo wąskie, co potwierdziło silną istotność współczynnika korelacji. Dla jednej zmiennej ANS-PNS (płaszczyzna podniebienna NL) uzyskano stosunkowo szeroki przedział ufności. Sytuacja ta oznaczała, że istniała możliwość wystąpienia dużego odchylenia, a wyniki te miały

Table 3. Software – computerized measurements

		1	2	3	4	5	6	7	8	9	10	11	12	13	14
Linear measurements [mm]															
ANS-Me	$\bar{x}$	59.5	52.3	67.9	50.3	61.2	67.4	70.6	62.9	57	62.6	64.9	69.1	58.9	59.9
	σ	0.56	0.23	0.24	0.46	0.3	0.21	0.61	0.56	0.2	0.3	0.16	0.34	0.94	0.33
Eb-PNS	$\bar{x}$	50.2	45	70	48.9	56.2	66.8	62.5	64.8	44.7	60.7	63.1	57.6	50.3	54.6
	σ	0.42	0.66	0.71	0.59	0.46	0.37	0.65	0.72	1.07	0.87	0.43	0.83	0.46	0.37
Se-N	$\bar{x}$	65.7	60.1	63.6	62.1	62	66.5	64.3	66	65.5	71.9	62.8	72.3	62.7	64.2
	σ	0.1	0.28	0.4	0.15	0.22	0.29	0.31	0.22	0.54	0.31	0.25	0.3	0.28	0.5
ANS-PNS	$\bar{x}$	52.7	47.5	52.5	52.7	49.1	53.6	55	50.3	48.0	50.3	50.9	55.3	48.1	50.3
	σ	2.09	1.37	0.51	1.44	1.42	1.41	2.13	0.69	0.68	1.47	1.04	0.58	1.65	0.95
U-TP	$\bar{x}$	7.5	13.2	10.4	7.7	10.4	14.4	11.5	9.4	14.0	12.2	11.9	13.0	13.2	6.4
	σ	0.42	0.83	0.34	0.28	3.45	0.53	0.21	0.47	0.82	0.32	0.76	0.48	0.93	0.6
Co-A	$\bar{x}$	84.7	73.9	83.6	77.7	82.3	84.2	86.7	83.6	76.5	82.2	86.3	89.0	79.5	83.6
	σ	0.61	0.86	0.33	0.84	0.66	0.74	1.04	1.41	0.8	0.68	2.02	0.65	0.81	0.73
Co-Gn	$\bar{x}$	106	92	112.3	98.1	105.3	115.3	116.3	101.9	98.4	110.6	109.5	112.7	101.8	102.2
	σ	0.48	0.74	0.43	0.73	0.67	0.73	0.91	0.84	0.46	0.63	0.99	0.48	0.56	0.43
Angular measurements [°]															
NSL-NL	$\bar{x}$	7.6	8.9	14.7	9.4	11.3	4.4	7.5	6.1	6.2	7.5	6.6	2.9	1.4	6.0
	σ	0.96	0.3	0.49	0.38	0.61	0.52	0.44	0.4	0.37	0.49	0.57	0.68	0.95	0.33
NSL-ML	$\bar{x}$	29.3	30.4	41.5	27.5	35.8	28.8	33	40.8	29.6	35.3	36.8	33.5	31.8	30.0
	σ	0.52	0.68	0.31	0.43	0.95	0.55	1.11	1.05	0.36	0.54	0.23	0.42	0.68	0.54
ML-NL	$\bar{x}$	21.74	21.56	26.75	18.10	24.53	24.48	25.49	34.73	23.46	27.79	30.27	30.60	30.44	24.00
	σ	1.26	0.65	0.53	0.59	0.94	0.45	1.15	0.82	0.48	0.66	0.59	0.76	0.61	0.23
NL-U1	$\bar{x}$	117.6	108.0	107.1	115.5	110.5	111.6	110.4	114	100.7	109.3	108.9	112.6	104.4	106.2
	σ	1.11	0.53	1.98	0.78	0.89	1.97	0.76	0.57	1.87	1.57	0.79	1.33	1.82	1.73
ML-L1	$\bar{x}$	98.4	101.3	90.2	106.7	86.4	81.9	88.6	95.6	93.3	95.2	97.8	102.9	89.3	99.8
	σ	0.64	2.03	0.78	2.66	2.13	0.6	1.45	2.08	1.27	1.05	0.94	2.39	1.49	0.87

95% confidence correlation coefficient. Table 4 presents the values of the Pearson's coefficient for two measurement methods for twelve values (column A in Table 4) and 95% confidence interval (column B in Table 4).

The results from eleven data indicated a very strong ($0.9 < r < 1$) and positive linear correlation between the two variables. Moreover, the confidence intervals were very narrow, which

małą wiarygodność. Dla 12 wartości pokazano na wykresie zależności pomiędzy dwiema zmiennymi. Większość punktów była umieszczona centralnie wokół prostej określającej zależność między poszczególnymi zmiennymi otrzymanymi ze współczynnika Pearsona.

Dla wszystkich 12 mierzonych wielkości wyliczono wartość p dla testu M-W, z próby 14 zmiennych dla poszczególnych wartości obu me-

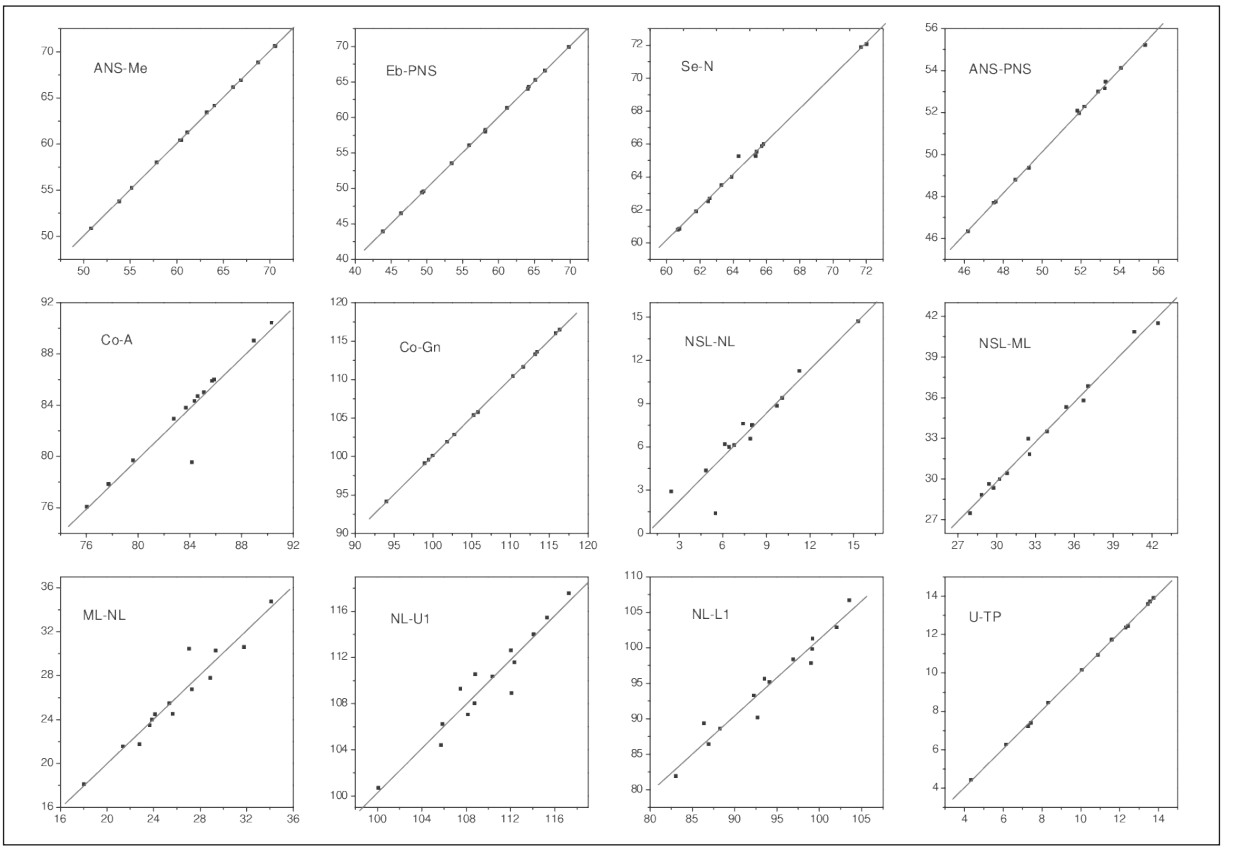

Fig. 2. *The relationship between the Pearson's coefficient and the distribution of points.*
Zależność pomiędzy współczynnikiem Pearsona korelacji r a rozkładem punktów.

confirmed a strong significance of the correlation coefficient. For one variable ANS-PNS (palatal plane-NL) a relatively wide confidence interval was obtained. This situation meant that there existed the possibility of a large deviation and these results had small reliability. The diagram illustrates the relationships between two variables for twelve values. Most points were placed centrally around a straight line defining the relationship between individual variables obtained from the Pearson's coefficient.

For all twelve measured values the p-value was calculated for the M-W test, of fourteen variables for particular values for both methods. The obtained values are presented in Table 4 (column C). The obtained "p" values were high and in all twelve cases there were no premises to reject the null hypothesis on equality of the mean value for both samples.

tod. Otrzymane wartości przedstawia tabela 4 (kolumna C). Otrzymane wartości p były duże i we wszystkich 12 przypadkach nie było przesłanek do odrzucenia hipotezy zerowej mówiącej o równości wartości średniej dla obu prób.

Podobnie dla 12 zmierzonych wielkości wyliczono wartość p dla testu W. Każda para stanowiła pomiar dla jednej osoby dokonany dwiema metodami. Otrzymane wyniki zestawiono w tabeli 4 (kolumna D). Gdy wartość p była mniejsza niż 0,01, wskazywało to na wysoki poziom istotności. Gdy wartość p znajdowała się między 0,01 a 0,05, to wynik był istotny. W obu przypadkach hipoteza zerowa mówiąca o równości median była odrzucana. Dla trzech zmiennych ML-NL, NL-U1 oraz ML-L1 wyliczona wartość p była większa niż standardowy próg istotności 0,05. Jednak dla dwóch pierwszych wartość p była wysoka. Nie było więc podstaw, by założyć, że mediany były

Table 4. The values of the Pearson's coefficient, 95% confidence interval, p-value Test Mann-Whitney U and p-value Test Wilcoxon

	A	B	C	D
Variables	Pearson's correlation coefficient	95% confidence interval	p-value Test Mann-Whitney U	p-value Test Wilcoxon
		Linear measurements [mm]		
ANS-Me	0.9769	(0.9269, 0.9929)	0.8036	0.00134
Eb-PNS	0.9619	(0.8812, 0.9881)	0.8388	0.02148
Se-N	0.9789	(0.9330, 0.9935)	0.7345	0.00061
ANS-PNS	0.8845	(0.6675, 0.9630)	0.7304	0.00513
U-TP	0.9564	(0.8646, 0.9864)	0.7688	0.00232
Co-A	0.9569	(0.8661, 0.9865)	0.9268	0.03381
Co-Gn	0.9862	(0.9558, 0.9957)	0.8036	0.00061
		Angular measurements [°]		
NSL-NL	0.9454	(0.8326, 0.9829)	0.5409	0.00647
NSL-ML	0.9954	(0.9851, 0.9986)	0.8743	0.04944
NL-U1	0.9589	(0.8724, 0.9872)	0.9459	0.74770
ML-L1	0.9744	(0.9190, 0.9920)	0.9999	0.90320
ML-NL	0.9635	(0.8860, 0.9887)	0.7688	0.15310

Similarly, for twelve measured values the p-value was calculated for the W test. Each pair constituted the measurement for one subject by two methods. The obtained results are presented in Table 4 (column D). The p-value < 0.01 indicated a high level of significance. The result was significant if the p-value was between 0.01 and 0.05. In both cases the null hypothesis on the equality of medians was rejected. For three variables ML-NL, NL-U1 and ML-L1, the calculated p-value was higher than the standard significance threshold of 0.05. However, the p-value was high for the first two values. Therefore, there was no basis to assume that medians were different. For the third value, the issue could not be solved on the basis of the test.

Discussion

Three comparative methods of different sensitivity were used for precise comparison and

różne. Dla trzeciej wielkości kwestia nie była do rozstrzygnięcia na podstawie tego testu.

Dyskusja

Trzy metody porównawcze o zróżnicowanej czułości były użyte do precyzyjnego porównania i do testowania zgodności dwóch metod pomiarowych, co skutkowało wiarygodną oceną charakteru zgodności testowanych wyników.

Dla określenia, czy wartości dwóch metod pomiarowych były powiązane i jaki istniał rodzaj zależności, użyto analizy korelacji i dwóch testów statystycznych. Pacjenci, których cefalogramy poddano analizie mieli różne wartości wieku szkieletowego oraz różne typy wad zgryzu. Z tych przyczyn porównania średnich wartości uzyskanych z pomiarów dla każdego badanego ze średnimi dla całej populacji nie mogły być testem ważności. Ponadto porównanie średniej i zmien-

to test the compatibility of two measurement methods, which resulted in a reliable assessment of the compatibility of the tested results.

The correlation analysis and two statistical tests were used to determine whether the values of two measurement methods were related, and what type of dependency existed. Subjects whose radiograms were analyzed had a different skeletal age and presented different forms of malocclusion. For this reason, comparing the average values obtained from measurements for each subject with the average for the whole population could not be a test of validity. Moreover, comparing the mean and variance for the sample obtained for both methods would not be a sensitive test to compare methods. Only the average values of the measurements were considered for the correlation analysis due to the low SD value for each subject compared to the differences between them.

The present study showed that measurements of analog cephalometric radiographs and the measurements with a computer program could provide high compatibility results, which was also found in other studies.[9,18] Our analysis did not verify which method was better, but it indicated that both methods were consistent. The computer cephalometric analysis did not introduce significant differences in landmark identification compared to the manual method. However, this method showed that statistically significant differences existed as regards landmark identification between analog and digitized LCRs. Most authors concluded that the differences between landmark measurements on analog cephalometric radiographs and those identified on their digitized counterparts were statistically significant, yet clinically acceptable.[8,9,17]

Considering the frequent criticism of the correlation method as the way to compare measurement methods, the comparison was not based on one method only. By subjecting data to several tests, different results were obtained, which allowed drawing conclusions that provided a broader view on the issue. Therefore, the results obtained from the correlation analysis would indicate a high compatibility of both methods. Similarly, in the M-W test, the obtained p-values

ności dla próby badawczej dla obydwu metod nie byłoby czułym testem porównania metod. Do analizy korelacji brano pod uwagę jedynie średnie wartości pomiarów, ze względu na niską wartość SD dla każdego badanego w porównaniu z różnicami pomiędzy nimi.

To badanie pokazało, że pomiary analogowych zdjęć cefalometrycznych i za pomocą programu komputerowego, mogły zapewnić wysoką zgodność wyników, co zostało również potwierdzone w innych badaniach. [9,18]

Nasza analiza nie sprawdzała, który sposób był lepszy, lecz pokazała, że obie metody były zgodne. Komputerowa analiza cefalometryczna nie wprowadzała istotnych różnic w wyznaczaniu punktów w porównaniu do metody ręcznej. Jednak metoda ta wykazała, że istniały statystycznie istotne różnice w identyfikacji punktów między analogowymi a cyfrowymi LCR. Większość autorów uważała, że różnice pomiędzy pomiarami punktów na analogowych zdjęciach cefalometrycznych a pomiarami zidentyfikowanymi na ich cyfrowych odpowiednikach były istotne statystycznie, ale klinicznie akceptowalne. [8,9,17]

Biorąc pod uwagę częstą krytykę metody korelacji, jako sposobu na porównanie metod pomiarowych, nie oparto porównania tylko na jednej metodzie. Poddając dane kilku testom, otrzymano zróżnicowane wyniki, które pozwoliły na wyciągnięcie wniosków, dających szersze spojrzenie na postawiony problem. Dlatego też wyniki uzyskane z analizy korelacji wskazywałyby na wysoką zgodność obu metod. Podobnie w przypadku testu M-W otrzymane wartości p nie skłoniły badacza do odrzucenia zerowej hipotezy o równości wartości średniej i rozkładów. Jednakże wynik uzyskany z testu W wskazał na zgoła odmienne wnioski, co do zgodności obu metod. Jako najczulszy z przeprowadzonych testów uwzględniający zależność wyników otrzymanych dwiema metodami u konkretnego badanego i biorący pod uwagę porównanie median test dał wynik negatywny, co do hipotezy o zgodności wartości median obu pomiarów dla większości badanych wielkości (9 na 12). Otrzymane wyniki był zbieżne z krytyką stosowania korelacji w celu uzyskania potwierdzenia zgodności

did not incline the researcher to reject the zero hypothesis on the equality of the mean value and distributions. However, the result obtained from the W test showed completely different conclusions regarding the compatibility of both methods. As the most sensitive of the conducted tests which considered the relationship of the results obtained by two methods in a particular subject and the comparison of medians, the test gave a negative result regarding the hypothesis on the compatibility of median values of both measurements for the majority of the measured values (9/12). The obtained results were consistent with the criticism of the application of correlation to obtain confirmation of the compatibility of methods. In their studies, *Altman* and *Bland*[11] provided a series of arguments explicitly demonstrating insensitivity of the method to test the compatibility.

Another test that was employed was the M-W test, which gave a positive result. It was based on the comparison of the mean value, and was unreliable considering a wide range of dispersion of the measured values. Furthermore, it was possible to venture an opinion that neither of the tests that considered the comparison of the mean could be a good test to show the compatibility of the two applied methods. Only tests that examine the result of the differences of particular measurements on the whole measured scale (the whole range) provided a reliable view of the compatibility (or incompatibility) character when two methods were compared.

One of the methods recommended by *Bland* and *Altman*[12] was the image analysis of mutual relationships of the values measured using two methods and a visual assessment of the obtained results. However, the analysis conducted in this manner required considerable experience and skill to formulate theses. It was a subjective assessment. Considering the fact that medical conclusions were based on such premises, the analysis could be considered promising. To rely on the objective result, the W test was performed which considered both differences in particular pairs of measurements and the scale of these differences.

On the basis of the obtained p-values, the metod. *Altman* i *Bland*[11] w swoich pracach przytoczyli szereg argumentów jednoznacznie wskazujących na nieczułość tej metody do testowania zgodności.

Drugim z zastosowanych testów był test M-W, który dał pozytywny wynik. Ten test opierał się na porównaniu wartości średniej i był testem niemiarodajnym, ze względu na szeroki zakres rozrzutu mierzonych wielkości. Co więcej, można było pokusić się o stwierdzenie, że żaden z testów biorących porównanie średniej nie mógł być dobrym testem na wykazanie zgodności dwóch stosowanych metod. Jedynie testy, badające wynik różnic poszczególnych pomiarów na całej mierzonej skali (całym zakresie), dały miarodajny obraz charakteru zgodności (bądź też niezgodności) przy porównywaniu dwóch metod.

Jedną z metod proponowanych przez *Blanda* i *Altmana*[12] była analiza rysunku wzajemnych zależności wielkości zmierzonych obiema metodami i naoczna ocena uzyskanych rezultatów. Jednak tak przeprowadzona analiza wymagała sporego doświadczenia i umiejętności w formułowaniu stawianych tez. To była ocena subiektywna. Biorąc pod uwagę, że jednak wnioski medyczne w dużej mierze opierały się na takich przesłankach, można było uznać ją za obiecującą. Aby oprzeć się na obiektywnym wyniku przeprowadzono test W, który uwzględnił zarówno różnice w poszczególnych parach pomiarów, jak i skalę tych różnic.

Na podstawie uzyskanych wartości p, hipoteza o równości median była odrzucona dla 9 zmierzonych wartości. Tak więc istniały miarodajne przesłanki wskazujące na różnice obu metod. Jednak badanie nie pokazało, jak istotne było to w zakresie diagnozowania lub jednoczesnego stosowania obu tych metod w indywidualnym przypadku. Postępowanie dotyczące wykorzystania tych metod podlegać będzie ocenie prowadzącego badanie i często będzie zależeć od czynników, których nie da się uwzględnić w analizach statystycznych. Przy takich wnioskach ważną staje się ocena precyzji dokonywanych pomiarów. Pomiar uznany za bardziej precyzyjny, wydaje się być lepszy, ponieważ jest obarczony mniejszym ryzykiem błędu. Ocena zgodności wyników zawsze

hypothesis on the equality of medians was rejected for nine measured values. Therefore, there were reliable premises indicating the differences of both methods. However, the study failed to demonstrate how significant it was in terms of diagnosis or simultaneous application of these two methods in an individual case. The procedure regarding the application of these methods will be subject to assessment by the researcher and it will frequently depend on the factors that cannot be considered in statistical analyses. With such conclusions, the assessment of the measurement precision is crucial. The measurement regarded as a more precise one appears to be better as the risk of error is lower. The assessment of the compatibility of the results will always depend on the precision of particular measurements.

Conclusions

The correlation analysis demonstrated how the measurement technique affects the measured variable. Most of the results proved to be equivalent for these two methods. The current study did not test whether the precision and the results were superior in one method, but it only indicated that one of these two methods could be used interchangeably in some cases. The assessment of the compatibility of the results will depend on the precision of particular measurements. On the basis of the obtained results from the Mann-Whitney U test and the comparison of these results with the Wilcoxon test results, none of the statistical tests which considers the comparison of the mean will be a proper test to indicate the compatibility of the two applied methods. Only tests that consider the differences between particular measurements for the whole examined range provide a reliable possibility for compatibility assessment.

będzie zależeć od dokładności poszczególnych pomiarów.

Wnioski

Analiza korelacji pokazała, jak technika pomiaru wpływa na zmierzoną zmienną. Większość wyników okazała się równoważna do tych metod. Niniejsze badanie nie sprawdzało, czy precyzja i wyniki były lepsze w jednej metodzie, lecz pokazało jedynie, że jeden z tych dwóch sposobów mógłby być stosowany zamiennie w niektórych przypadkach. Ocena zgodności wyników będzie zależeć od dokładności poszczególnych pomiarów. Na podstawie uzyskanych wyników z testu U Manna-Whitney'a i porównania ich z wynikami z testu Wilcoxona stwierdzono, że żaden z testów biorących pod uwagę porównanie średniej nie będzie dobrym testem na wykazanie zgodności dwóch stosowanych metod. Jedynie testy biorące pod uwagę różnice pomiędzy poszczególnymi pomiarami dla całego badanego zakresu dają miarodajną możliwość badania zgodności.

References

1. *Bruntz LQ, Palomo JM, Baden S, Hans MG:* A comparison of scanned lateral cephalograms with corresponding original radiographs. Am J Orthod Dentofacial Orthop 2006; 130: 340-348.
2. *Roden-Johnson D, English J, Gallerano R:* Comparison of hand-traced and computerized cephalograms: Landmark identification, measurement, and superimposition accuracy. Am J Orthod

Dentofacial Orthop 2008; 133: 556-564.

3. *Tsorovas G, Linder-Aronson Karsten A:* A comparison of hand-tracing and cephalometric analysis computer programs with and without advanced features-accuracy and time demands. Eur J Orthod 2010; 32: 721-728.
4. *Paluch Z, Wojtyna J, Misiołek M:* The influence of nasopharyngeal patency on the morphology of nasomaxillary complex. Acta Odont Scand 2013; 71: 1599-1605.
5. *Gribel BF, Gribel MN, Frazaão DC, McNamara JA Jr, Manzi FR:* Accuracy and reliability of craniometric measurements on lateral cephalometry and 3D measurements on CBCT scans. Angle Orthod 2011; 81: 26-35.
6. *Olmez H, Gorgulu S, Akina E, Bengi AO, Tekdemir İ, Ors F:* Measurement accuracy of a computer-assisted three-dimensional analysis and a conventional two-dimensional method. Angle Orthod 2011; 81: 375-382.
7. *Tan SSW, Ahmad S, Moles DR, Cunningham SJ:* Picture archiving and communications systems: a study of reliability of orthodontic cephalometric analysis. Eur J Orhod 2011; 33: 537-543.
8. *Huja SS, Grubaugh EL, Rummel AM, Fields HW, Beck FM:* Comparison of hand-traced and computer-based cephalometric superimpositions. Angle Orthod 2009; 79: 428-435.
9. *Celik E, Polat-Ozsoy O, Memikoglu TUT:* Comparison of cephalometric measurements with digital versus conventional cephalometric analysis. Eur J Orthod 2009; 31: 241-246.
10. *Santoro M, Jarjoura K, Cangialosic TJ:* Accuracy of digital and analogue cephalometric measurements assessed with the sandwich technique. Am J Orthod Dentofacial Orhop 2006; 129: 345-351.
11. *Altman DG, Bland JM:* Measurement in medicine: the analysis of method comparison studies. Statistician 1983; 32: 307-317.
12. *Bland JM, Altman DG:* Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986; i: 307-310.
13. *Sayinsu K, Isik F, Trakyali G, Arun T:* An evaluation of the errors in cephalometric measurements on scanned cephalometric images and conventional tracings. Eur J Orthod 2007; 29: 105-108.
14. *Chen S, Chen Y, Yao CJ, Chang H:* Enhanced speed and precision of measurement in a computer-assisted digital cephalometric analysis system. Angle Orthod 2004; 74: 501-507.
15. *Pancherz H:* The mechanism of Class II correction in Herbst appliance treatment. A cephalometric investigation. Am J Orthod 1982; 82: 104-113.
16. *Ricketts RM:* Cephalometric analysis and synthesis. Angle Orthod. 1961; 31: 141-156.
17. *Bondevik O:* Dentofacial changes in adults: a longitudinal cephalometric study in 22-33 and 33-43 year olds. J Orofac Orthop 2012; 73: 277-288.
18. *Fontes AM, Joondeph DR, Bloomquist DS, Greenlee GM:* Long-term stability of anterior open-bite closure with bilateral sagittal split osteotomy. Am J Orthod Dentofacial Orthop 2012; 142: 792-800.
19. *Borrie F, Thomson D, McIntyre GT:* Precision of measurements on conventional negative 'bones white' and inverted greyscale 'bones black' digital lateral cephalograms. Eur J Orthod 2012; 34: 57-61.
20. *Naoumova J, Lindman R:* A comparison of manual traced images and corresponding scanned radiographs digitally traced. Eur J Orthod 2009; 31: 247-253.

Address: 47-400 Racibórz, ul. Podwale 1/1
Tel.: +4832 4150555
e-mail: zbemalikp@op.pl

Received: 13th January 2015
Accepted: 14th February 2015