12-1-2012

Intra- and inter-examiner Reliability of Direct Facial Soft Tissue Measurements Using Digital Calipers

Nikolay Mollov
Marquette University, nikolay.mollov@marquette.edu

Jose A. Bosio
Marquette University, jose.bosio@marquette.edu

Jessica E. Pruszynski
Medical College of Wisconsin

Thomas S. Wirtz
Marquette University

Accepted version. *Journal of the World Federation of Orthodontists*, Vol. 1, No. 4 (December 2012): e157-e161. DOI.

Intra- and inter-examiner reliability of facial soft tissue measurements

Nikolay Mollov
School of Dentistry, Marquette University
Medical College of Wisconsin
Milwaukee, WI
José Antônio Bosio
School of Dentistry, Marquette University
Medical College of Wisconsin
Milwaukee, WI
Jessica Pruszynski
School of Dentistry, Marquette University
Medical College of Wisconsin
Milwaukee, WI
Thomas Wirtz
School of Dentistry, Marquette University
Medical College of Wisconsin
Milwaukee, WI

Abstract

Background: The objective of this study is to determine if facial soft tissue measurements using digital calipers can be reliably taken by the same examiner and by a large group of examiners.

Materials and Methods: Ten examiners performed a set of 18 in-clinic measurements on ten female and ten male dental students using a digital

Journal of the World Federation of Orthodontists, Vol. 1, No. 4 (December 2012): pg. e157-e161. DOI.

caliper twice over a three week period. The intra-class correlation coefficient and the Shrout-Fleiss method were used for the statistical analysis.
Results: Anthropometric intra-examiner reliability was high for all measurements (none fell below R=0.934). However, inter-examiner reliability exhibited a wide range of values, some reliable [alR-alL (R=0.922) and sn-ul (R=0.926)], and others unreliable [base of nose (R=0.590), mouth height (R=0.585) and B' – gn (R=0.623)].
Conclusion: Soft tissue measurements of clearly identifiable points measured by the same examiner produced highly consistent, accurate and reliable measurements. Soft tissue points with poor definition resulted in average-to-poor reliabilities measurements.

Keywords: soft tissue, facial measurements, calipers, reliability.

Introduction

Plastic surgery, otolaryngology and dentistry share a common field of interest called facial esthetics. In order to establish this commonality and evaluate the modification of anatomical structures, anthropometric measurements can be used. In the early 1980s, Farkas,[1] a plastic surgeon, established the proper dimensional norms for the head, face and ear by measuring a sample of 1312 subjects. Later on, Farkas and Posnick[2] determined the proportions of the developing head by performing a series of anthropometric studies on 140 soft tissue parameters from approximately 1600 patients.

As technology evolved, more sophisticated methods were developed to analyze and quantify what precisely makes the human face attractive. Peck and Peck[3] compared the lateral and frontal photographs of a number of individuals who were previously "acclaimed" to be "possessing the most pleasing facial esthetic qualities" to cephalometric measurements. Similarly, studies were conducted[4,5] to compare certain cephalometric and anthropometric measurements taken directly off the face. Some authors[6,7,8] looked solely at cephalograms, and attempted to develop soft tissue standards for orthognathic surgery treatment planning. Cephalograms also showed us that conventional orthodontics can alter certain soft tissue structures.[9] Ferring and Pancherz[10] examined the "divine proportions of the growing face" by means of photographic evaluation[11,12] and used twenty-first century technology to recreate a

three-dimensional digital image of the study participants, or face modeling, and performed the desired measurements on that model. Newer technology has made possible to combine cone-beam computer tomography (CBCT), digital photography and increasingly more powerful computer systems in order to study facial soft tissues. Studies show that CBCT can accurately reproduce the identification of soft-tissue facial landmarks and facial tissue depth measurement.[13,14] However, an anthropometric measurement on a CBCT reproduced image is still hindered by the software's rendering of the patient's skin texture, color, facial line angles, light reflection, etc. Latest techniques such as image fusion allow us to superimpose a 3D photograph on a CBCT image.[15] While there are some errors with these methods, it is a promising development of anthropometric measurements in the digital world adding more accuracy to the facial soft tissue measurements.

The majority of the methods used to perform facial soft tissue measurements are extremely resource consuming and very impractical in terms of study set-up. Although a variety of measuring methods is available, direct clinical measurement is a simpler method for investigating soft tissue facial landmarks. The cost is lowered as is the simplicity of the study design. The limitations are related to the landmark identification and the measurement acquisition consistency of the different investigators. If changes to the facial soft tissue are to be measured before and after orthodontic using direct clinical measurements, the reliability of the operators needed to be reported.

Shaner[16] compared the reliability of thirteen caliper-taken facial measurements done by one examiner on two participants with the landmarks either being marked or not-marked on the subject's face. To the best of our knowledge no studies have been conducted using a large number of evaluators to collect facial measurements from volunteer subjects in two different times.

The aim of this study is to determine if consistent facial soft tissue measurements using digital calipers can be obtained by the same examiner and to determine the reliability of these same measurements among a large group of examiners.

Journal of the World Federation of Orthodontists, Vol. 1, No. 4 (December 2012): pg. e157-e161. DOI.

Materials and Methods

This study was approved by the Institutional Review Board. (Protocol # HR-2083)

Twenty dental students were recruited and randomly selected to participate as subject of measurement in the project. The group was comprised of 10 male and 10 female students. Participants were to be excluded should they have exhibited congenital facial abnormalities, as well as those undergoing medical/pharmacological treatment that could produce distortion of normal facial landmarks.

Ten examiners were selected from the postgraduate orthodontic program and the undergraduate dental students. One examiner was a full time faculty member. The examiner population was comprised of 5 females and 5 males, however this distribution of males/females was purely coincidental as it was not considered prior to the examiners selection. Due to the number of examiners standard calibration was not feasible. Instead, the examiners were provided with a detailed write-up and a Power-Point presentation on how to identify the facial landmarks (Figure 1). The examiners practiced in clinic to identify the facial landmark points and took measurements on each other. (Figure 2)

Facial measurements were taken in the school orthodontic clinic using an 8 mm sliding digital Mitutoyo calipers (Aurora, IL). The measurement error for all Mitutoyo calipers was identical as per the company's description (0.01mm). The examiners were paired in teams – one examiner took the measurements while the other recorded the data. The participant subjects were seated in the dental chair with their head relaxed and in an upright position. In order, to establish a repeatable position of the mandible, the study participants were guided into mandibular rest position and asked to remain with their lips relaxed. Measurements were recorded in the standard form for all participant subjects (Figure 2). The study participants were recalled approximately a month later and the whole procedure was repeated.

The intra-class correlation coefficient was used to determine both the intra- and inter-investigator reliabilities. This correlation

Journal of the World Federation of Orthodontists, Vol. 1, No. 4 (December 2012): pg. e157-e161. DOI.

coefficient is a general measure of agreement between two or more raters. The Shrout-Fleiss method was used to compute both, the inter- and the intra- investigator reliabilities.

Results

The reliability coefficients for the 18 facial soft tissue measurements and the intra-examiner and inter-examiner differences with a 95% confidence interval are shown in Table 1. The first five measurements were considered horizontal, whereas the last 13 were considered vertical.

Intra-examiner differences

All 10 examiners showed consistently high intra-examiner reliability between T1 and T2. None of the calculated reliabilities fell below R=0.934. The least reliable measurements were nasal width at base of the nose, soft tissue B point to gnathion and mouth height. Even for those 3 measurements, the average reliabilities varied between R=0.934 to R=0.943. The 18 measurements exhibited very high reliabilities with nasal width (al-al, R=0.992), middle third of the face (N'- Sn, R= 0.989), and upper lip length (Sn – UL, R=0.992) showing the highest reliabilities.

Inter-examiner differences

When comparing the measurements among the 10 examiners, a larger reliabilities distribution was found. The reliabilities for the 18 measurements can be placed in three distinct groups. Group one is made up of a few measurements showing consistently high reliabilities. Those include alR-alL (R=0.922) and sn-ul (R=0.926). As noted before, those same two measurements also showed very high intra-examiner reliability.

Significant reliability measurements differences are seen in the second group with a larger number of measurements showing poor reliability. Most notable are nasal width at base of nose (R=0.590), mouth height (R=0.585) and B' – gn (R=0.623). The first two measurements also showed the lowest intra-examiner reliability.

Journal of the World Federation of Orthodontists, Vol. 1, No. 4 (December 2012): pg. e157-e161. DOI.

Most of the remaining measurements can be placed in group three which showed reliabilities that fell somewhere in between the extremes with mouth width (chR-chL, R=0.863), the third of the face (Ha – Na', R=0.827; Na' – sn, R=0.899; sn – gn', R=0.867), measurements around the mouth (stL – LVB, R = 0.865; stU-stL=0.882) being the most consistent. Measurements between the left and right commissures differed greatly (sn – ch L R = 0.758; sn – ch R = 0.837).

No significant differences were found between horizontal and vertical measurements. Both categories feature some reliable and some unreliable measurements.

Discussion

In order to evaluate future changes in the soft tissue contour before and after orthodontic treatment within a large sample, a strong reliability test is necessary. This study was designed to evaluate the reliability of soft tissue measurements performed on a dental student volunteer sample.

The time it took to acquire the measurements was not recorded. However, we made a general observation where most of the 18 measurements were collected in less than 4 minutes. If we were to only acquire the reliable measurements in future studies, this time can be greatly reduced, and clinical measurements can be performed without disrupting office flow.

Some particular measurements were different from those performed in previous studies.[2,17,18,19] However, the majority of facial landmarks used in the study (Figure 2) were developed, similarly to the points used by Farkas [2]. The one exception was stomion upper (stU) and stomion lower (stL). In Farkas' description[1], stomion was a point described by the intersection of the facial midline with the "horizontal labial fissure of the gently closed lips." In our study, the participants were requested to relax their mandible and, consequently, their lips were also relaxed, which often resulted in an interlabial gap. Thus, the lower-most point of the upper lip and the upper-most point

of the lower lip (both crossing the imaginary facial midline) were defined as stomion upper and stomion lower.

Burstone[20] used cephalometric headplates in lieu of measurements taken directly from the living. He believed that those measurements would diminish accuracy associated with soft-tissue flexibility. He also stated that the time factors was relevant, since the operator could not be as leisurely with that method and the patient could not be expected to hold a given pose for a long period of time. However, transverse measurements, as it was acquired in this study, were not noted on those cephalometric headplates, and tracing errors would also have to be investigated.

Farkas[1] identified three particular sources of error – improper measuring technique, problems with the measuring instruments and improper identification of the facial landmarks. The authors attempted to eliminate the first two by training all ten examiners prior to the study and by having the ten examiners use the exact same caliper model, as well as measuring the sample on the same day and in the same clinical setting. Thus, the only variable that could produce error among the different examiners was the facial landmark identification.

Intra-examiner Reliability

The examiners in this study exhibited very high intra-investigator reliability for essentially all measurements. Shaner used two examiners to measure similar anthropometric facial measurements and found the majority of the measurements were in good agreement, similar to the findings of this study.[16] Farkas[1] also found minimal differences in measurements when looking at one examiner over different time points.

The present findings showed that the examiners consistently pick the same points. However, without a gold standard to identify some of these points, i.e. those that are not easily identifiable due to them overlying a bony structure (zygion, gnathion) or those that require several different angles for precise identification (pronasale), the precise determination of the points becomes difficult. Thus, while we can say that the examiners consistently picked the same point we

cannot state with certainty if those points were the correct ones or if they were what the examiner believed was the correct point.

Inter-examiner Reliability

Inter-examiner reliability showed a much larger variation. This was confirmed by previous studies. Mommaerts et al.[21] investigated several distances similar to those measured in this project and found the majority of those to be unreliable. The measurements that showed the highest reliability involved points that were very easy to identify - in their study, the pupils in the interpupillary distance measurement, supraorbitale, gnathion. The distance between the two zygomatic points (right and left) was found to not be reliable similar to the results of this study.

Geerts et al.[22] attempted to evaluate the reliability of measuring the vertical dimension of rest by essentially measuring the distance between pronasale and gnathion with a caliper. They used an examiner sample of N=20 (1 patient, 1 measurement, 10 times) and found good inter-examiner reliability for those two points. This was confirmed in another study[23] that attempted to evaluate the measurement of the vertical dimension of rest using pronasale and an additional point on the chin. In our particular study, the measurements involving pronasale fell in the second group – while the reliability was acceptable it was not ideal. This again is dependent on the points that comprise the particular measurement – those that involve clearly identifiable points produce, as expected, a much more reliable measurement. The least reliable measurement was the mouth height (ls-li). This result was possibly generated due to subject difficulties to maintain their lips relaxed during measurement acquiring.

Unlike Shaner's[16], this study did not attempt to mark the landmarks on the participant subject faces for two reasons. First, we wanted to allow the examiners to identify the points themselves, and second, we did not want to spend an excessive amount of time marking the points and acquiring the measurements. Landmark identification relationship between different examiners needed to be proven strong, as well as how successfully could these examiners

Journal of the World Federation of Orthodontists, Vol. 1, No. 4 (December 2012): pg. e157-e161. DOI.

reproduce that landmark identification from T1 to T2. Marking the points on the face would have defied the purpose of the study.

Lastly, while observing the reliability between the set of horizontal versus the set of vertical measurements no differences were found. Both groups had some reliable measurements and some poor ones. This is probably due to the reliabilities being dependent on how easy it is to define the facial landmarks as opposed to how the measuring device is being held.

Conclusion

Clearly identifiable points measured with digital calipers by the same examiner produced highly consistent, accurate and reliable results. Soft tissue points with poor definition yielded average-to-poor reliabilities. Significant differences were found when different examiners identified the same point. No differences were noted between vertical and horizontal measurements.

Aknowledgments

The authors want to thank Dr. Arthur Hefti, for the valuable statistical advices.

This manuscript has been supported in part by AAO (FFA) and AAOF (OFDFA) funding awards.

Bibliography

1. Farkas LG. Anthropometry of the Head and Face in Medicine. New York, NY, USA: Elesvier; 1981.
2. LG F, JC P. Growth and development of regional units in the head and face based on anthropometric measurements. Cleft Palate-Craniofacial Journal 1992;29:301-329.
3. Peck H, Peck S. A concept of facial esthetics. Angle Orthod 1970;40:284-318.
4. Farkas LG, Tompson B, Phillips JH, Katic MJ, Cornfoot ML. Comparison of anthropometric and cephalometric measurements of the adult face. J Craniofac Surg 1999;10:18-25; discussion 26.

5. Budai M, Farkas LG, Tompson B, Katic M, Forrest CR. Relation between anthropometric and cephalometric measurements and proportions of the face of healthy young white adult men and women. J Craniofac Surg 2003;14:154-161; discussion 162-153.
6. Arnett GW, Bergman RT. Facial keys to orthodontic diagnosis and treatment planning. Part I. Am J Orthod Dentofacial Orthop 1993;103:299-312.
7. Arnett GW, Bergman RT. Facial keys to orthodontic diagnosis and treatment planning. Part II. Am J Orthod Dentofacial Orthop 1993;103:395-411.
8. Arnett GW, Jelic JS, Kim J, Cummings DR, Beress A, Worley Jr CM et al. Soft tissue cephalometric analysis: diagnosis and treatment planning of dentofacial deformity. Am J Orthod Dentofacial Orthop 1999;116:239-253.
9. Park Y-C, Burstone C. Soft-tissue profile - Fallacies of hard-tissue standards in treatment planning. AM J ORTHOD DENTOFAC ORTHOP 1986;90:52-62.
10. Ferring V, Pancherz H. Divine proportions in the growing face. Am J Orthod Dentofacial Orthop 2008;134:472-479.
11. Ferrario VF, Sforza C, Schmitz JH, Miani Jr A, Serrao G. A three-dimensional computerized mesh diagram analysis and its application in soft tissue facial morphometry. Am J Orthod Dentofacial Orthop 1998;114:404-413.
12. Ferrario VF, Sforza C, Serrao G, Colombo A, Ciusa V. Soft tissue facial growth and development as assessed by the three-dimensional computerized mesh diagram analysis. Am J Orthod Dentofacial Orthop 1999;116:215-228.
13. Medelnik J, Hertrich K, Steinhäuser-Andresen S, Hirschfelder U, Hofmann E. Accuracy of anatomical landmark identification using different CBCT- and MSCT-based 3D images: an in vitro study. J Orofac Orthop 2011;72:261-278.
14. Fourie Z, Damstra J, Gerrits PO, Ren Y. Accuracy and reliability of facial soft tissue depth measurements using cone beam computer tomography. Forensic Sci Int. 2010;199:9-14.
15. Maal TJ, Plooij JM, Rangel FA, Mollemans W, Schutyser FA, Bergé SJ. The accuracy of matching three-dimensional photographs with skin surfaces derived from cone-beam computed tomography. Int J Oral Maxillofac Surg 2008;37:641-646.
16. Shaner DJ, Bamforth JS, Peterson AE, Beattie OB. Technical Note: Different techniques, different results – a comparison of photogrammetric and caliper-derived measurements. Am J Phys Antropol 1998;106:547-552.

17. Gosman SD. Anthropometric method of facial analysis in orthodontics. Am J Orthod 1950;36:749-762.

18. Naini FB, Moss JP, Gill DS. The enigma of facial beauty: esthetics, proportions,deformity and controversy. Am J Orthod Dentofacial Orthop 2006;130:277–282.
19. Fourie Z, Damstra J, Gerrits PO, Ren Y. Evaluation of anthropometric accuracy and reliability using different three-dimensional scanning systems. Forensic Sci Int. 2011;207:127-134.
20. Burstone C. The integumental profile. Am J Orthod 1958;44:1-25.
21. Mommaerts MY, Moerenhout BA. Reliability of clinical measurements used in the determination of facial indices. J Craniomaxillofac Surg 2008;36.:279-284.
22. Geerts GA SM, Nel DG. A comparison of the accuracy of two methods used by pre-doctoral students to measure vertical dimension. J Prosthet Dent 2004;91:59-66.
23. Sakar O, Sülün T, Kurt H, Gençel B. Reliability and comparison of two facial measurements to detect changes of occlusal vertical dimension in complete denture wearers. Gerodontology 2011;28:205-208.

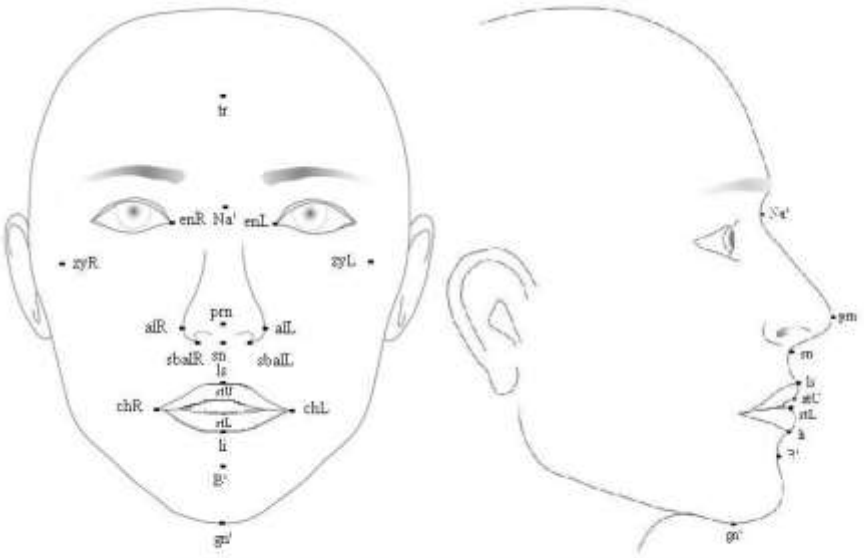
Figure 1. Description of the points utilized for the soft tissue measurements.

trichion (tr)	The point on the hairline in the midline of the forehead. Note: for this project, not participants with visible hair loss or abnormally high hairline were selected for participation
soft tissue nasion (Na')	The soft tissue covering the point located in the midline of both the nasal root and the nasofrontal suture
endocanthion (en)(Left or Right)	The point at the inner commissure of the eye fissure
zygion (zy)(L or R)	The most lateral point of each zygomatic arch; identified by trial measurements. **Note**: in this project left and right are identified, when applicable, in order to help the investigators in communicating with the study examiners
pronasale(prn)	The most protruded point of the apex nasi
alare (al)(L or R)	The most lateral point on each alar contour
subnasale (sn) (L or R)	The midpoint of the columnella base at the apex of the angle where the lower border of nasal septum and the surface of the upper lip meet
subalare (sbal)(L or R)	The point at the lower limit of each alar base, where the alar base disappears into the skin of the upper lip
labiale superius (ls)	The midpoint of the upper vermillion line
labiale inferius (li)	The midpoint of the lower vermillion line
cheilion (ch)(L or R)	The point located at each labial commissure
stomion (st) (Upper and Lower)	The imaginary point at the crossing of the vertical facial midline and the horizontal labial fissure between the upper/lower lip and the oral cavity as seen from a frontal view. Note: in this project the study participants were asked to relax their lips, hence the visible border of each lip was used as the horizontal landmark
soft tissue B point (B')	The deepest curvature of the soft tissue between the lower lip and the chin point
gnathion (gn')	The lowest median landmark of the lower border of the mandible

Figure 2 - Soft Tissue Measurements

Dimension	Facial Landmark	Measurement (mm)
Horizontal Measurements	1. Zygomatic Width (zyR – zyL)	
	2. Mouth Width (chR - chL)	
	3. Nasal Width at widest nostrils (alR – alL)	
	4. Nasal Width at Base of Nose (sbalR – sbalL)	
	5. Intraorbital Width (enR – enL)	
Vertical Mesurements	6. Hairline-Nasion (tr – Na')	
	7. Nasion – SubNasale (Na' – sn)	
	8. SubNasale – Gnathion (sn – gn')	
	9. Nasion – Tip of Nose (Na' – prn)	
	10. Stomion Lower – Soft Tissue B point (li – B')	
	11. Soft Tissue B point – Gnathion (B' – gn')	
	12. SubNasale – Right commissure (sn – chR)	
	13. SubNasale – Left commissure (sn – chL)	
	14. Tip of Nose - upper lip (prn - ls)	
	15. Mouth height (ls - li)	
	16. SubNasale to Upper Lip (sn - ls)	
	17. Lower Lip Thickness (stL - li)	
	18. Interlabial Gap (stU – stL) - if lips are incompetent	

Journal of the World Federation of Orthodontists, Vol. 1, No. 4 (December 2012): pg. e157-e161. DOI.



"These figures were published in the Am J Orthod Dentofac Orthoped, Vol 103(5):395-411, G. William Arnett and Robert T. Bergman: Facial keys to orthodontic diagnosis and treatment planning. Part II, Copyright Elsevier, 1993." They were slightly modified for the purpose of this study.

Table 1. Inter- and Intra-examiner reliability estimates (mean and standard deviation) and 95% confidence intervals.

Measurement	Caliper Inter-examiner Reliability	Caliper Intra-examiner Reliability
Zygomatic Width - **(zyR – zyL)**	0.696 (0.55-0.837)	0.958 (0.924-0.981)
Mouth Width - **(chR – chL)**	0.863 (0.774-0.932)	0.984 (0.972-0.993)
Nasal Width at Widest Nostrils - **(alR – alL)**	0.922 (0.866-0.963)	0.992 (0.985-0.996)
Nasal Width at Base of Nose - **(sbalR – sbalL)**	0.590 (0.428-0.765)	0.935 (0.882-0.970)
Intraorbital Width - **(enR – enL)**	0.775 (0.65-0.884)	0.972 (0.949-0.987)
Hairline – Nasion - **(tr – Na')**	0.827 (0.723-0.914)	0.980 (0.963-0.991)
Nasion – SubNasale - **(Na' – sn)**	0.899 (0.83-0.951)	0.989 (0.98-0.995)
SubNasale – Gnathion - **(sn – gn')**	0.867 (0.78-0.935)	0.985 (0.973-0.993)
Nasion – Tip of Nose - **(Na' – prn)**	0.763 (0.635-0.877)	0.97 (0.946-0.986)
Stomion Lower – Soft Tissue B point - **(li – B')**	0.706 (0.562-0.843)	0.96 (0.928-0.982)
Soft Tissue B point – Gnathion - **(B' – gn')**	0.623 (0.465-0.788)	0.943 (0.897-0.974)
SubNasale – Right commissure - **(sn – chR)**	0.837 (0.736-0.919)	0.981 (0.965-0.991)
SubNasale – Left commissure - **(sn – chL)**	0.758 (0.628-0.874)	0.969 (0.944-0.986)
Tip of Nose - upper lip - **(prn - ls)**	0.850 (0.755-0.926)	0.983 (0.969-0.992)
Mouth height - **(ls - li)**	0.585 (0.423-0.762)	0.934 (0.88-0.97)
SubNasale to Upper Lip - **(sn - ls)**	0.926 (0.872-0.965)	0.992 (0.986-0.996)
Lower Lip Thickness - **(stL - li)**	0.865 (0.778-0.934)	0.985 (0.972-0.993)
Interlabial Gap - **(stU – stL)**	0.882 (0.803-0.942)	0.987 (0.976-0.994)

Journal of the World Federation of Orthodontists, Vol. 1, No. 4 (December 2012): pg. e157-e161. DOI.