European Journal of Orthodontics, 2017, 1–7
doi:10.1093/ejo/cjx002

Original article

ROC surface assessment of the ANB angle and Wits appraisal's diagnostic performance with a statistically derived 'gold standard': does normalizing measurements have any merit?

Hans L.L. Wellens[1], Ellen A. BeGole[2] and Annemarie M. Kuijpers-Jagtman[1]

[1]Department of Orthodontics and Craniofacial Biology, Radboud University Nijmegen Medical Centre Nijmegen, The Netherlands, [2]Department of Orthodontics, College of Dentistry, University of Illinois at Chicago, USA

Correspondence to: Hans L.L. Wellens, Researcher, Department of Orthodontics and Craniofacial Biology, Radboud University Nijmegen Medical Centre, 309 Tandheelkunde, P.O. Box 9101, 6500 HB Nijmegen, The Netherlands; E-mail: wellens.hans@telenet.be

Summary

Objective: To assess the ANB angle's and Wits appraisal's diagnostic performance using an extended version of Receiver Operating Curve (ROC) analysis, which renders ROC surfaces. These were calculated for both the conventional and normalized cephalometric tests (calculated by exchanging the patient's reference landmarks with those of the Procrustes superimposed sample mean shape). The required 'gold standard' was derived statistically, by applying generalized Procrustes superimposition (GPS) and principal component analysis (PCA) to the digitized landmarks, and ordering patients based upon their PC2 scores.

Methods: Digitized landmarks of 200 lateral cephalograms (107 males, mean age: 12.8 years, SD: 2.2, 93 females, mean age: 13.2 years, SD: 1.7) were subjected to GPS and PCA. Upon calculating the conventional and normalized ANB and Wits values, ROC surfaces were constructed by varying not just the cephalometric test's cut-off value within each ROC curve, but also the gold standard cut-off value over different ROC curves in 220 steps between -2 and 2 standard deviations along PC2. The volume under the resulting ROC surfaces (VUS) served as a measure of overall diagnostic performance. The statistical significance of the volume differences was determined using permutation tests (1000 rounds, with replacement).

Results: The diagnostic performance of the conventional ANB and Wits was remarkably similar for both Class I/II (81.1 and 80.75% VUS, respectively, $P > 0.05$). Normalizing the measurements improved all VUS highly significantly (91 and 87.2 per cent, respectively, $P < 0.001$).

Conclusion: The conventional ANB and Wits do not differ in their diagnostic performance. Normalizing the measurements does seem to have some merit.

Introduction

The orthodontic diagnostic toolset conventionally comprises both a clinical and radiological investigation; the latter usually consisting of a panoramic radiograph and/or peri-apical series, as well as a lateral (and anteroposterior) cephalogram. Although most contemporary orthodontic textbooks recommend the routine use of lateral cephalometry for diagnostic purposes (1, 2), its impact on clinical practice seems to be somewhat limited (3–7). A recent meta-analysis concluded that 'cephalograms are not routinely needed for treatment planning in Class II malocclusions' (8), while another suggested 'lateral cephalometric radiographs have been used without adequate scientific evidence', and that 'there is an urgent need to improve

lateral cephalometry's diagnostic efficiency and therapeutic efficacy' (9). The efficacy of diagnostic imaging was defined by Fryback and Thornbury (10) using a six level hierarchical model, the first three of which pertain to the images' technical quality (level one), diagnostic accuracy (level two), and influence on the practitioner's diagnostic thinking (level three).

Since the addition of a lateral cephalogram has been found to cause few changes to treatment plans formulated without it (3–7), lateral cephalometry seems to score low in level three. This is usually attributed to the technical problems it is fraught with, such as image enlargement and structural blurring, doubling and shrouding (11), which would seem to impact mainly level one. Level two might be influenced by difficulties associated with choosing, precisely defining and pinpointing landmarks (11), while geometrical distortion might play a role as well. The latter seems to be linked to the highly variable nature of the cephalometric reference landmarks and planes (12), thereby allowing individuals with the same cephalometric values to exhibit markedly differing intermaxillary relationships, while the opposite may hold true as well (13–17). One solution to this predicament might be to exchange the patient's highly variable reference framework with a fixed one, by superimposing a template on the digitized patient landmarks using Procrustes superimposition, and performing the measurements from the superimposed template's reference landmarks, instead of the patient's (12) (Fig. 1). Albeit unconventional, this approach significantly improved the correlation between the 'normalized' measurements, as compared to the conventional ones (12). Correlations however do not represent a measure of diagnostic performance.

Diagnostic performance is usually determined using Receiver Operating Characteristic curve analysis (ROC); a procedure which originated in signal analysis (18), but has since found widespread application in medicine (19, 20). ROC curve analysis plots the sensitivity (or true positive ratio) versus 1-specificity (or false-positive ratio) for a full range of possible values of the diagnostic test's cut-off value. The area under the resulting curve serves as a measure of diagnostic performance: the larger the surface area under the curve (the closer the curve approaches the upper left corner of the graph), the more powerful the test. A test characterized by a diagonal ROC curve (from lower left to upper right) has no discriminatory power whatsoever. Anything below 60 per cent area under the curve is usually designated 'very poor', whereas between 60 and 70 per cent

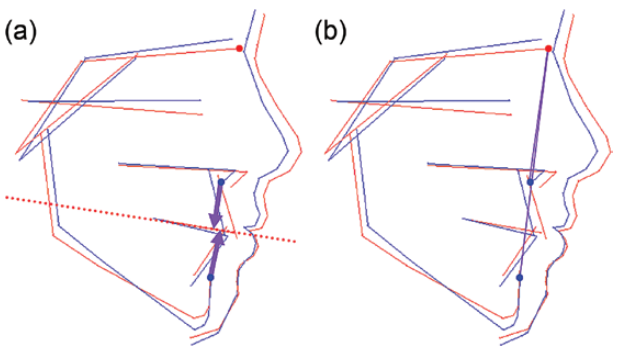

Figure 1. (a) The patient's configurations is shown in blue, the Procrustes superimposed sample mean shape in red. The normalized Wits appraisal is obtained by dropping perpendiculars from the patient's points A and B (in blue) onto the superimposed sample mean shape's occlusal plane (the dotted red line). (b) Similarly, the normalized ANB angle is obtained by measuring the angle between the patient's points A and B (in blue) and the superimposed sample mean shape's point N (in red).

UAC, tests are usually classified as 'poor', between 70 and 80 as 'fair', between 80 and 90 as 'good', and anything above as 'excellent'. ROC curve analysis is however dichotomous in nature, requiring clearly discernible health states in order to provide the black-or-white diagnostic result required to determine the test's diagnostic power (21). This would seem to align poorly with the continuous spectrum of facial variation present in the orthodontic patient population (22). Also, ROC curve analysis requires a gold standard (an ideally infallible 'reference test' which provides the correct answer to the diagnostic question) (21), which until recently did not seem to be available.

McIntyre and Mossey (23), Halazonetis (24) and later Akli *et al.* (25) proposed adopting a geometric morphometric approach to cephalometry, based upon the combined application of Procrustes superimposition and principal component analysis to previously digitized landmark coordinates; a methodology which is used ubiquitously in biology and anthropology for the analysis of shape (26, 27). Procrustes superimposition centres, scales and rotates landmark configurations to minimize the distance between the corresponding points using the least squares criterion (26, 27) (Fig. 2a, Supplementary Animation 1), while principal component analysis finds the directions in multivariate space along which the superimposed configurations vary most, in decreasing order (26, 27) (Fig. 2b, Supplementary Animation 2). In earlier studies, the first and second principal components (PCs; i.e. the major directions of variance), were found to predominantly describe variation in the vertical (dolichofacial versus brachyfacial morphology) and anteroposterior dimensions (Class II versus Class III), respectively (22, 24, 25, 28) (Fig. 2b, Supplementary Animation 2). A plot of PC1 versus PC2 may be therefore be used as a map, characterizing a patient's horizontal and vertical skeletal makeup, while also allowing for inter-patient comparison in terms of the same variables (Fig. 2b). Additionally, the distribution of the PC scores may be used to categorize patients: a logical approach would be to designate those patients belonging to the central portion of the PC1 score distribution (e.g. PC1 mean ± 1 SD) as being normodivergent, and those in between PC2 mean ± 1 SD as being skeletal Class I. Two recent publications provided some tools for delineating those regions of the PC1-PC2(-PC3) shape space containing patients which would be regarded as normo-, hypo- and hyper-divergent and skeletal Class I, II and III (25, 28).

The aim of this investigation was to compare the diagnostic performance of the ANB angle and Wits appraisal using ROC analysis, whereby the 'gold standard' is derived statistically, by classifying patients based upon the distribution of the PC2 scores resulting from the combined application of Procrustes superimposition and principal component analysis (25, 28). Furthermore, we introduce in an extension of ROC analysis, whereby the gold standard cut-off is varied as well, resulting in ROC-surfaces instead of curves (29–32). Finally, we aimed to compare the diagnostic performance of the ANB angle and Wits appraisal measurements to their normalized counterparts (obtained by superimposing the sample mean shape on the patient's landmarks and measuring from the sample mean shape's reference structures).

Methodology

The methodology has been published in detail previously (28). Briefly, two hundred consecutive lateral cephalometric radiographs (107 males, mean age: 12.8 years, SD: 2.2, range: 7.4–19.1; 93 females, mean age: 13.2 years, SD: 1.7, range 8.3–19.6) were collected, using the following inclusion criteria: only pre-treatment radiographs, absence of craniofacial syndromes, only Caucasian

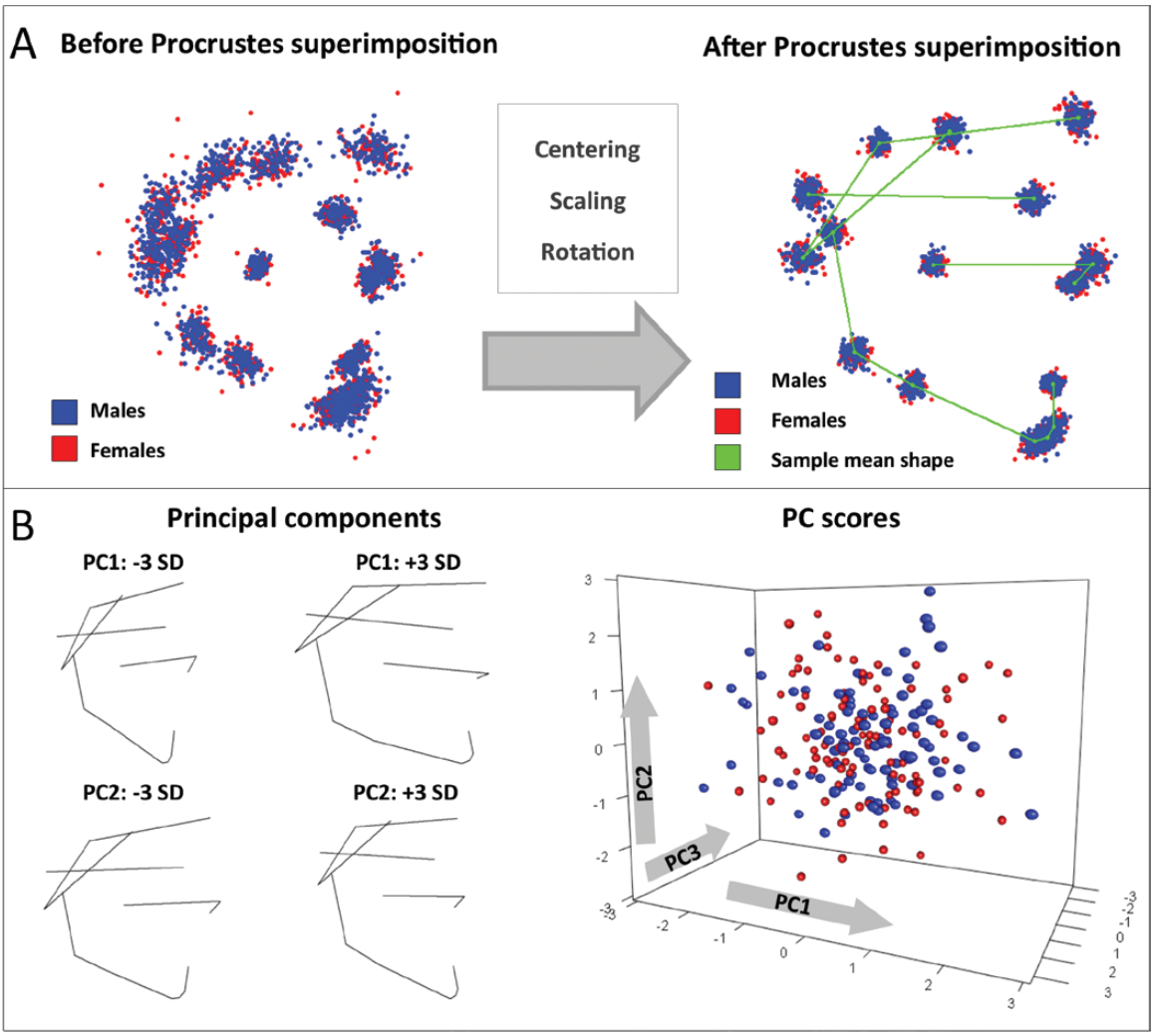

Figure 2. (a) On the left, the original coordinates of sixteen skeletal landmarks before Procrustes superimposition are shown. The right image shows the same landmarks after centring, scaling and rotating the configurations in order to minimize the squared distance between the corresponding landmarks. The sample mean shape is depicted in green. (b) The first two principal components are shown on the left, by deforming the sample mean shape three standard deviations along the respective PCs. The right pane depicts the PC scores associated with the first three principal components. Each dot represents the value of a particular patient on principal components one, two and three. Since 16 landmarks were digitized, each patient thus has a 28-score long 'address' in multivariate space (4 degrees of freedom were lost to centring, scaling and rotation of the landmark configurations). Not all principal components do however represent biologically meaningful information: in this scenario, only the first five PCs are biologically 'interpretable'.

patients, only radiographs taken in occlusion, and absence of gross movement artifacts. Patients had to be at least seven and no older than 20 years to be included in the sample. The required sample size was calculated beforehand based on an estimation of the number of subjects present in the tails of a normally distributed sample.

All images were collected using a Planmeca Proline XC (Planmeca Oy, Helsinki, Finland) by the first author, using appropriate settings and a standardized technique. The radiographs were then loaded in Viewbox (dHal software version 4.0.1.7, Kifissia, Greece), in order to identify the position of sixteen skeletal landmarks (Fig. 3). Cephalometric enlargement was compensated for during the digitizing process. The obtained coordinates were then exported to R (http://www.r-project.org) for further processing. The digitized skeletal coordinates of the pooled sample were superimposed using generalized Procrustes superimposition (GPS, Fig. 2a, Supplementary Animation 1) (26, 27, 33, 34), and stereometrically projected onto tangent space (26, 27, 34), after which the male and female mean shapes were calculated. The significance of the morphological difference between them as well as their mean age difference, was tested using a 10 000 round permutation test.

The GPS superimposed and projected landmark coordinates were then subjected to principal component analysis (26, 27, 34, 35), after which the principal component scores and their standard deviations were calculated (Fig. 2b, Supplementary Animation 2). This allowed us to objectively classify patients in terms of their intermaxillary relationships based upon each patient's PC2 score. We then calculated the corresponding (conventional) ANB angle and Wits appraisal values, as well as their normalized counterparts. The latter were determined by Procrustes-superimposing the (pooled) sample mean shape on the patient's landmarks, and measuring the ANB angle using the superimposed mean shape's point N as reference structure. Similarly, the Wits value was determined using the superimposed sample mean shape's occlusal plane, after rescaling to true size.

Next, ROC curves were constructed for the conventional and normalized ANB angle and Wits appraisal (ANBc/WitsC and ANBN/

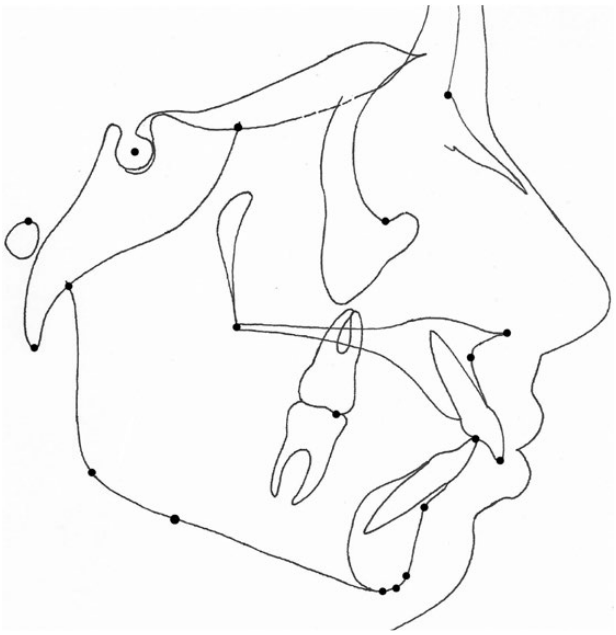

Figure 3. The sixteen digitized skeletal and dental landmarks: sella, nasion, porion, orbitale, anterior and posterior nasal spine, basion, articulare, points a and b, pogonion, gnathion, menton, antegonial notch, gonion, spheno-ethmoidale, mesiobuccal cusp tip of the upper first molar, upper and lower incisal edge. For their definitions, please refer to (12).

WitsN, respectively), by plotting sensitivity versus 1-specificity for the full range of possible cut-off values. This process was repeated 220 times while incrementally increasing the 'gold standard' cut-off value between minus two and two standard deviations on the PC2 axis. Every cycle's gold standard diagnosis was determined by comparing each patient's PC2 score to that cycle's gold standard cut-off value: patients with PC2 scores smaller than the gold standard PC2 score cut-off were designated 'Class II', and those with equal or larger PC2 scores 'Class I'. When placing the resulting 220 ROC curves side-by-side, a ROC surface was created (i.e. a three-dimensional mesh; Fig. 3, Supplementary Webpage 3), the volume under which would add up to one for a perfect test (i.e. dimensions 1 × 1 × 1), and 50 per cent for an indiscriminate test. After calculating the ROC surface volumes of the classic and normalized measures, the statistical significance between them was calculated by randomly permuting the 220 corresponding ROC curves between the two measures under investigation with replacement (1000 rounds), and calculating the volume of the resulting randomly permuted ROC surfaces. Since 1000 randomized ROC surfaces were generated, 499 500 volume differences were thus calculated, the number of which exceeding the original one determined the significance of the difference. The significance level was set at 5 per cent. The digitizing error was determined in previous investigations involving the same material, and was found to be non-significant (22, 28).

Results

Neither the male/female shape difference (Fig. 4), nor the age difference between them was statistically significant (P = 0.1926 and 0.1818, respectively, 10 000 permutation rounds). Both groups were therefore pooled for further analysis. Table 1 lists the diagnostic performance of the conventional and normalized ANB and Wits, expressed as the volume under the ROC surface (expressed in percentages). The conventional ANB angle and Wits appraisal performed remarkably similar, at about 80 per cent volume under the ROC surface (Table 1) (Fig. 5). The difference between them (0.34 per cent, Table 2) was not significant (P = 0.402, Table 2). Normalizing the measurements increased the volumes for both Wits and ANB, although the latter improved about 10 per cent (Fig. 6), as opposed to 7 per cent for the former (WitsN: 87.15 per cent, ANBN: 91.04 per cent, Table 1). The difference between them was highly significant (P < 0.001, Table 2). All pairwise comparisons were highly significant, except that between the conventional ANB and Wits (Table 2).

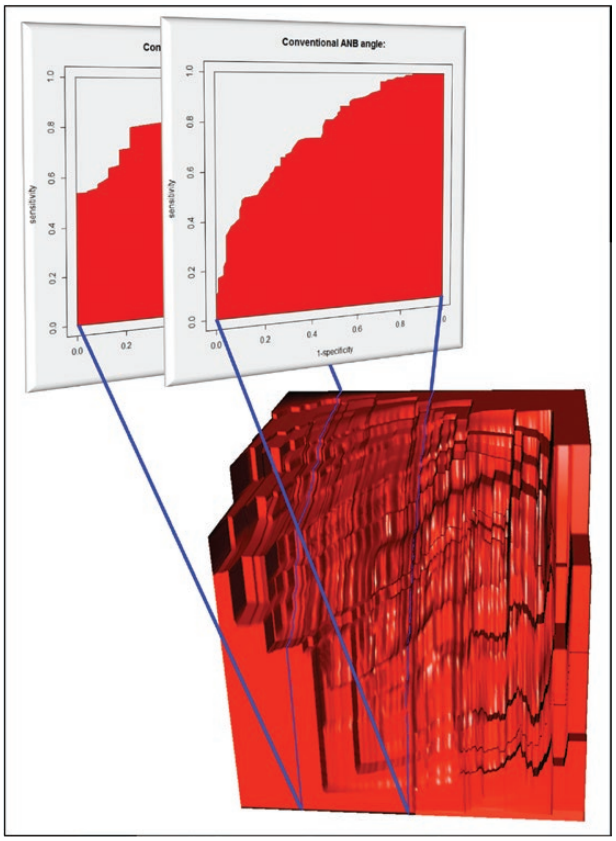

Figure 4. ROC surface for the conventional ANB angle, together with two of the 220 ROC curves that make up the ROC surface. Their position within the ROC surface is depicted in purple.

Table 1. Volume under the surface values for the conventional and normalized ANB and Wits measurements.

	VUS
ANBC	81.092
ANBN	91.037
WitsC	80.746
WitsN	87.148

VUS: Volume under surface (percentage).

Discussion

To date, surprisingly few studies are available about the diagnostic performance of currently available lateral cephalometric tests (8, 9). Whereas conventionally the results of newly introduced tests were

often correlated to those of existing ones to assess or compare performance, more recent studies increasingly rely on ROC analysis for this purpose. In the absence of a true gold standard for cephalometric diagnosis, some authors have chosen to classify their study subjects based upon occlusion (36, 37) or existing cephalometric analyses, applied either singly (38) or combined (39). Other studies have included profile assessments (40, 41), while one study applied a Delphi approach to establish their gold standard (41). Akli *et al.* (25) proposed combining Procrustes superposition and principal component analysis, and using the underlying distribution of PC1, 2 and 3 scores to provide a more objective, statistically based classification methodology based upon craniofacial morphology (i.e. craniofacial shape; Fig. 2). As such, this methodology might serve as a 'gold standard' in ROC analysis for the assessment of vertical growth pattern and mandibulomaxillary discrepancy.

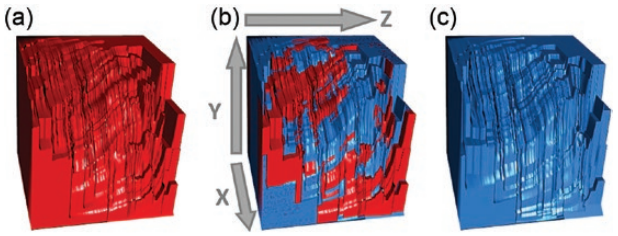

Figure 5. (a and c) Resulting ROC surfaces for the conventional ANB angle and Wits appraisal, respectively. (b) Superimposition of the ANBc and WitsC ROC surfaces. The x-axis represents 1-specificity, the y-axis the sensitivity, and the z-axis the gold-standard cut-off in between –2 and 2 SD along PC2 (220 steps). The difference between both ROC surfaces was non-significant ($P > 0.05$, Table 2).

Traditional ROC analysis is somewhat limited by its dichotomous nature: by calculating the probability that a test will correctly identify the diseased and healthy patient in a pair, it provides the diagnostic power of a rather blunt, 'yes-or-no' test to whether disease is present, therefore applying strictly to two class problems (19–21). Although extensions to the ROC methodology to accommodate three-class and multi-class situations have recently appeared in the literature (30–32), these do not seem to apply to orthodontic cephalometry either: the absence of a 'disease state' in the true sense generally precludes the formulation of straightforward, clear-cut, and universally applicable cut-off points for the different cephalometric classes. Instead, the continuous spectrum of craniofacial variability present in the orthodontic population begs the diagnostically more sophisticated question of whether patients are relatively more or less prognathic/retrognathic instead of just Class I, II or III, not just at all possible cut-off points of the pertaining cephalometric measure, 'but also of the gold standard' (i.e. it basically represents an 'infinite class problem'). This study therefore proposed a modification of the recently published extensions to the ROC methodology (29–32), by varying the cut-off points of the cephalometric measure under investigation 'within in each ROC curve', as well as the cut-off points of the gold standard 'over different ROC curves'. In doing so, the traditional three-class diagnostic problem (i.e. Class I, II or III) is reduced to a 2-class one (since the 'less Class II/more Class II' diagnostic question is equivalent to the 'more Class III/less Class III' question), albeit applied over a broad range of gold-standard cut-off values. When placed next to one another, the combined ROC curves generate a ROC surface, the volume under which serves as a more sophisticated measure of overall diagnostic performance.

When applying this approach to the conventional and normalized ANB angle and Wits appraisal, the conventional measurements

Table 2. Results of the permutation test in order to compare the VUS measurements for the conventional and normalized ANB and Wits.

Permutation test:	Orig. vol. diff.	Randomly permuted volume difference:				
(1000 rounds)		Mean (%)	SD (%)	Max. (%)	>Orig.	*P*
ANBc/ANBn	9.95	0.48	0.36	2.60	0	<0.001
ANBc/WitsC	0.34	0.33	0.25	1.84	201007	0.402
ANBn/WitsN	3.89	0.34	0.26	2.07	0	<0.001
ANBc/WitsN	6.06	0.42	0.31	2.26	0	<0.001
ANBn/WitsC	10.29	0.48	0.36	2.54	0	<0.001
WitsC/WitsN	6.40	0.37	0.28	2.08	0	<0.001

Orig. vol. diff.: Original volume difference (%).
>Orig.: The number of iterations in which the volume difference exceeded the original.

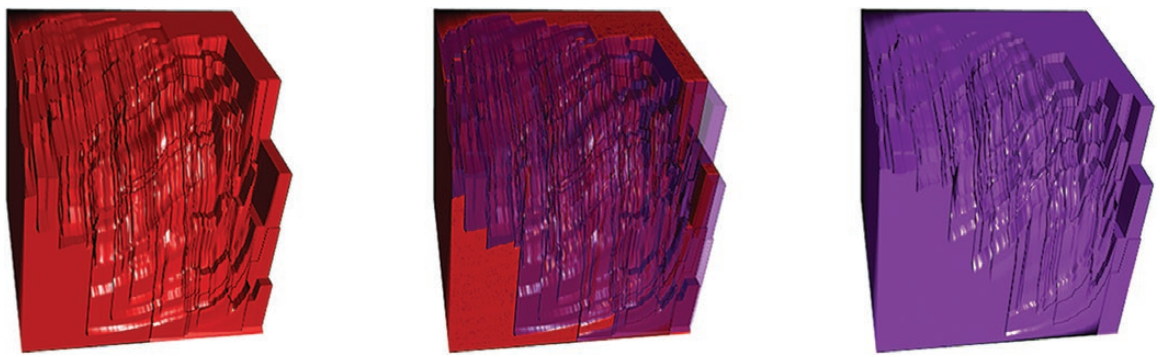

Figure 6. ROC surfaces for the conventional (on the left, in red) and normalized (on the right, in purple) ANB angle. In the middle the superimposition of both surfaces is visualized. The improvement in VUS resulting from the normalization was almost 10 percent (Table 2).

were found to perform strikingly similar, at about 80 per cent volume under the surface (P = 0.40, Table 2). In an earlier publication (28), those regions of the PC1-PC2 plot which would be regarded as skeletal Class I according to both measurements were identified, and found to be distinctly different in shape, whereby the Wits appraisal's region shape was almost identical to that defined by the underlying distribution of the PC2 scores. We might therefore expect the Wits appraisal to slightly outperform the ANB angle, which was found to not be the case. Tables 1 and 2 do seem to suggest that normalizing the measurements has some merit: the volumes under the ROC surfaces of the normalized measurements were found to be significantly higher compared to those of the conventional measurements. The observed improvements (Table 2) lend credence to the hypothesis that cephalometric diagnostic confusion may be explained, in part, by the high inter-individual variability of the reference landmarks and planes used in orthodontic cephalometric tests (12, 28). It is interesting to note that normalizing the ANB angle led to a larger improvement compared to the Wits appraisal (9.9 versus 6.4 %, respectively, Table 2). This might be explained by a slightly more pronounced susceptibility of the Wits appraisal to changes in the cant of the occlusal plane (17), for which the normalizing procedure is not able to compensate fully.

The proposed approach differs from previously published methodology (29–32): in medical settings it is hardly ever possible to vary all the different classes' gold standard cut-off to the extent possible in cephalometrics, since there usually are few if any degrees to 'being ill'. This allowed us to rephrase the three-class diagnostic question into a 2-class one, resulting in a somewhat differently shaped ROC surface. There naturally is a practical limit on the number of cut-off points at which the test can be evaluated: it is of little use to evaluate the test at many more levels as there are patients in the sample. Also, we decided to assess the diagnostic performance in between 2 standard deviations above and below the mean PC2 score, due to the dwindling number of patients above/below this limit.

From the clinical point of view, the most sophisticated methodology currently available for assessing intermaxillary relationships would seem to be the use of PC2 scores (22, 24, 25), which unfortunately requires the availability of a relatively large database of ethnicity-specific patient coordinates in order to scrutinize craniofacial variability (i.e. GPS and PCA). Another potential problem is the abstract nature of the PC2 scores (due to the lack familiarity to the clinical orthodontist). This might be circumvented simply by assigning patients the accompanying cephalometric value of the sample mean shape, deformed to each patient's position in the PC1–PC2 map, as proposed previously (measurement by proxy) (28). Since this procedure applies the same 'ruler' to all patients, it also prevents geometric distortion, although it again requires the availability of a coordinate database. Notwithstanding the trivial nature of providing such databases in anonymized form, the normalization procedure offers improved diagnostic performance in the absence thereof: only the sixteen coordinates of the sample mean shape are required to calculate the normalized values. The latter also represents values that are more familiar to most orthodontists.

Conclusion

The ANB angle and Wits appraisal were found to perform very similarly at about 80 percent area under the surface. Normalizing the ANB and Wits improved all VUS highly significantly, with almost 10 and 6.6 percent, respectively.

Supplementary material

Supplementary data are available at *European Journal of Orthodontics* online.

Conflict of interest

None to declare.

References

1. Graber, T., Vanarsdall, R. and Vig, K. (2005) *Orthodontics: Current Principles and Techniques*. Mosby, Philadelphia, Pasadena, 4th edn.
2. Proffit, W.R., Fields, H.F., Jr and Sarver, D.M. (2006) *Contemporary Orthodontics*. *Elsevier Health Sciences*, St. Louis, Missouri.
3. Atchison, K.A., Luke, L.S. and White, S.C. (1991) Contribution of pretreatment radiographs to orthodontists' decision making. *Oral surgery, oral medicine, and oral pathology*, 71, 238–245.
4. Han, U.K., Vig, K.W., Weintraub, J.A., Vig, P.S. and Kowalski, C.J. (1991) Consistency of orthodontic treatment decisions relative to diagnostic records. *American Journal of Orthodontics and Dentofacial Orthopedics*, 100, 212–219.
5. Hansen, K. and Bondemark, L. (2001) The influence of lateral head radiographs in orthodontic diagnosis and treatment planning. *European Journal of Orthodontics*, 23, 452–453.
6. Nijkamp, P.G., Habets, L.L., Aartman, I.H. and Zentner, A. (2008) The influence of cephalometrics on orthodontic treatment planning. *European Journal of Orthodontics*, 30, 630–635.
7. Devereux, L., Moles, D., Cunningham, S.J. and McKnight, M. (2011) How important are lateral cephalometric radiographs in orthodontic treatment planning? *American Journal of Orthodontics and Dentofacial Orthopedics*, 139, e175–e181.
8. Durão, A.R., Alqerban, A., Ferreira, A.P. and Jacobs, R. (2015) Influence of lateral cephalometric radiography in orthodontic diagnosis and treatment planning., *The Angle orthodontist*, 85, 206–210.
9. Rischen, R.J., Breuning, K.H., Bronkhorst, E.M. and Kuijpers-Jagtman, A.M. (2013) Records needed for orthodontic diagnosis and treatment planning: a systematic review., *PLoS ONE*, 8, e74186.
10. Fryback, D.G. and Thornbury, J.R. (1991) The efficacy of diagnostic imaging. *Medical Decision Making*, 11, 88–94.
11. Baumrind, S. and Frantz, R.C. (1971) The reliability of head film measurements. *American Journal of Orthodontics*, 60, 111–127.
12. Wellens, H. (2009) Improving the concordance between various anteroposterior cephalometric measurements using Procrustes analysis. *European Journal of Orthodontics*, 31, 503–515.
13. Taylor, C.M. (1969) Changes in the relationship of nasion, point A, and point B and the effect upon ANB. *American Journal of Orthodontics*, 56, 143–163.
14. Freeman, R.S. (1981) Adjusting A-N-B angles to reflect the effect of maxillary position. *The Angle orthodontist*, 51, 162–171.
15. Binder, R.E. (1979) The geometry of cephalometrics. *Journal of Clinical Orthodontics: JCO*, 13, 258–263.
16. Hussels, W. and Nanda, R.S. (1984) Analysis of factors affecting angle ANB. *American Journal of Orthodontics*, 85, 411–423.
17. Roth, R. (1982) The 'Wits' appraisal - its skeletal and dento-alveolar background. *European Journal of Orthodontics*, 4, 21–28.
18. Swets, J.A., Green, D.M., Getty, D.J. and Swets, J.B. (1978) Signal detection and identification at successive stages of observation. *Perception & Psychophysics*, 23, 275–289.
19. Metz, C.E. (1978) Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, 8, 283–298.
20. Hanley, J.A. and McNeil, B.J. (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143, 29–36.
21. Swets, J.A. (1979) ROC analysis applied to the evaluation of medical imaging techniques. *Investigative Radiology*, 14, 109–121.
22. Wellens, H.L., Kuijpers-Jagtman, A.M. and Halazonetis, D.J. (2013) Geometric morphometric analysis of craniofacial variation, ontogeny and

modularity in a cross-sectional sample of modern humans. *Journal of Anatomy*, 222, 397–409.
23. McIntyre, G.T. and Mossey, P.A. (2003) Size and shape measurement in contemporary cephalometrics. *European Journal of Orthodontics*, 25, 231–242.
24. Halazonetis, D.J. (2004) Morphometrics for cephalometric diagnosis. *American Journal of Orthodontics and Dentofacial Orthopedics*, 125, 571–581.
25. Akli, E., Marinaki, L. and Halazonetis, D.J. (2015) Selecting subjects with high craniofacial shape homogeneity for clinical trials. *American Journal of Orthodontics and Dentofacial Orthopedics*, 148, 1026–1035.
26. Dryden, I.L. and Mardia, K.V. (1998) *Statistical Shape Analysis*. Wiley, Chichester, New York.
27. Zelditch, M.L., Swiderski, D.L. and Sheets, H.D. (2012). *Geometric Morphometrics for Biologists: A Primer.* Academic Press, Amsterdam, The Netherlands, 2nd edn.
28. Wellens, H.L.L. and Kuijpers-Jagtman, A.M. (2016) Connecting the new with the old: modifying the combined application of Procrustes superimposition and principal component analysis, to allow for comparison with traditional lateral cephalometric variables. *European Journal of Orthodontics*, 38, 569–576.
29. Mossman, D. (1999) Three-way ROCs. *Medical Decision Making*, 19, 78–89.
30. Dreiseitl, S., Ohno-Machado, L. and Binder, M. (2000) Comparing three-class diagnostic tests by three-way ROC analysis. *Medical Decision Making*, 20, 323–331.
31. Nakas, C.T. and Yiannoutsos, C.T. (2004) Ordered multiple-class ROC analysis with continuous measurements. *Statistics in Medicine*, 23, 3437–3449.
32. Li, J. and Fine, J.P. (2008) ROC analysis with multiple classes and multiple tests: methodology and its application in microarray studies. *Biostatistics*, 9, 566–576.
33. Gower, J.C. and Dijksterhuis, G.B. (2004) *Procrustes Problems*. Oxford University Press, New York.
34. Claude, J. (2008) *Morphometrics with R*. Springer, New York, 1st edn.
35. Hotelling, H. (1933) Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24, 417–441.
36. Han, U.K. and Kim, Y.H. (1998) Determination of Class II and Class III skeletal patterns: receiver operating characteristic (ROC) analysis on various cephalometric measurements. *American Journal of Orthodontics and Dentofacial Orthopedics*, 113, 538–545.
37. Freudenthaler, J.W., Celar, A.G. and Schneider, B. (2000) Overbite depth and anteroposterior dysplasia indicators: the relationship between occlusal and skeletal patterns using the receiver operating characteristic (ROC) analysis. *European Journal of Orthodontics*, 22, 75–83.
38. Kumar, S., Valiathan, A., Gautam, P., Chakravarthy, K. and Jayaswal, P. (2012) An evaluation of the Pi analysis in the assessment of anteroposterior jaw relationship. *Journal of Orthodontics*, 39, 262–269.
39. Neela, P.K., Mascarenhas, R. and Husain, A. (2009) A new sagittal dysplasia indicator: the YEN angle. *World Journal of Orthodontics*, 10, 147–151.
40. Baik, C.Y. and Ververidou, M. (2004) A new approach of assessing sagittal discrepancies: the Beta angle. *American Journal of Orthodontics and Dentofacial Orthopedics*, 126, 100–105.
41. Anderson, G., Fields, H.W., Beck, F.M., Chacon, G. and Vig K.W. (2006) Development of cephalometric norms using a unified facial and dental approach. *Angle Orthodontist*, 76, 612–618.