A comparison of the reproducibility of manual tracing and on-screen digitization for cephalometric profile variables

D. P. Dvortsin, A. Sandham, G. J. Pruim and P. U. Dijkstra
Department of Orthodontics, University Medical Center Groningen, The Netherlands

SUMMARY The aim of this investigation was to analyse and compare the reproducibility of manual cephalometric tracings with on-screen digitization using a soft tissue analysis. A random sample of 20 lateral cephalometric radiographs, in the natural head posture, was selected. On-screen digitization using Viewbox® 3.1.1.9 cephalometric software and manual tracing on a 1:1 printout of the image was carried out twice in different sessions 1 week apart. Differences were analysed using a repeated measurement analysis of variance with method, session, and method–session interaction as explaining variables. The differences were expressed as an absolute percentage of the overall mean.

The findings of the present study indicate that the two measurement methods differ significantly for 11 variables (P=0.001 to P=0.042). The area around stomion was the least reproducible. Except for s−n s−unt, nasal protrusion, with the manual technique, all mean differences between sessions and between methods were less than 1 degree or 1 mm and were, on-screen, smaller for 13 variables compared with those traced manually. Absolute percentage differences of the overall mean were smaller for seven variables with the digital technique and three variables in the manual technique, while four manual variables and one on-screen variable exceeded 2 per cent of the overall mean. Although small significant differences were found, the clinical relevance remains questionable.

Introduction

Standardized lateral cephalometric radiographs (Broadbent, 1931) are widely used as a diagnostic and clinical tool in orthodontics. The radiographic image shows not only the craniofacial and dental structures but also the soft tissue profile. This profile is important in treatment planning and evaluation because of its direct relationship with facial aesthetics. Soft tissue changes, as a result of treatment and growth, may be small in magnitude. It is therefore relevant to understand the errors which exist in the assessment of the soft tissue profile and to identify the most accurate method of analysis.

Determination and interpretation of the various types of errors is relevant in cephalometric studies to ensure correct conclusions are drawn (Houston, 1983; Kamoen and Dermaut, 2001; Schulze et al., 2002). Two types of traits are important in investigations determining errors in measurement: reproducibility which is the ability to produce similar or identical measurement results when measurements are repeated over time, and validity which is the ability to produce measurements which are identical to the actual values of the construct to be measured.

The development of computer software for direct digital imaging and analysis of the soft tissue profile has brought new possibilities which enable manipulation of image quality for greater clarity together with automatic assessment of geometric forms and contrast edges. Few studies have been carried out to compare the measurement errors, using computer software (Eppley and Sadove, 1991; Chen et al., 2000; Kazandjian et al., 2006). Comparison of conventional cephalograms and digital images, recorded simultaneously, shows more reproducible measurements for the digital images due to the better soft tissue visualization (Eppley and Sadove, 1991). In turn, enhancement for on-screen quality has been found to improve the on-screen visibility of the soft tissues (Oka and Trussell, 1978). Without image enhancement, reproducibility of landmark identification was found to be similar with manually traced radiographs, but significantly worse on images of poorer quality (Macrì and Wenzel, 1993; Nimkarn and Miles, 1995). However, these images had not been acquired simultaneously, which might introduce bias because of quality differences between scanning and printing of the original.

Regarding reproducibility of cephalometric variables in conventional (manual) tracings and digitally traced images, some authors emphasize the practical advantages of on-screen image enhancement because options are available to manipulate contrast, grey scale, and accentuate the edges of structures (Oka and Trussell, 1978; Forsyth et al., 1996). However, poorer reproducibility for on-screen images was found compared with cephalometric films using several hard tissue landmarks (Geelen et al., 1998). In contrast, Hagemann et al. (2000) found digitally traced images to have a higher reproducibility compared with manual tracings, which was determined by comparison of the means between two tracing sessions within each technique. Other researchers have concluded that image quality and landmark variation have a greater influence on validity than tracing.
technique (Chen et al., 2000; Ongkosuwito et al., 2002).
However, it is claimed that the reproducibility of landmark
identification and the validity of the digitization method are
equally important in all cephalometric studies (Cooke and
Wei, 1991; Doll et al., 2001).

The aim of the present investigation was to compare the
reproducibility of cephalometric soft tissue measurements,
using on-screen digitization with image enhancement, and
manual tracing.

Materials and methods

Lateral cephalometric images of patients recorded in a
standardized head posture (Solow and Tallgren, 1971) \((n =
20, 12\) males, \(8\) females; mean age 12 years, standard
deviation = 0.33) were selected at random from the digital
archive of the University Medical Center Groningen. An
alphabetical list of patients was scrutinized with inclusion
criteria based on a skeletal Class I dental base relationship
and availability of a complete radiographic image of the
soft tissue profile. Every 10th patient who satisfied these
guidelines was included. Standardized head posture
cephalometric radiographs had been obtained by instructing
the standing subject to relax and look into a distant mirror
during exposure of the radiograph. No rigid
fixation of the
head took place, but light support from loose ear rods of the
X-ray machine ensured there was no lateral rotation of the
head. The radiographic images had been acquired digitally
(ProMax, DiMax2 Digital Cephalometric Unit, Planmeca,
Helsinki, Finland) with a resolution quality of 2272 pixels
width and 2045 pixels height at a 24 bit depth.

The 20 images were then printed out on high-quality
paper (to avoid absorption spreading), at a scale of 1:1.
Using acetate tracing paper and a 2H pencil, the radiographs
were traced in daylight using the soft tissue analysis based
on that proposed by Sarnäs and Solow (1980). This analysis
was selected because it provides a detailed linear and
angular dimensional analysis of the facial profile, which
includes the lips, nose, and overall profile angle. A second
tracing of the 20 radiographs was carried out 1 week later
by the same clinician (DPD).

The on-screen images were digitized and analysed
according to the profile analysis (Figure 1), using Viewbox
3.1.1.9 software® (dHAL, Kifissia, Greece), developed for
handling digital cephalometric data. To assist point
placement, where necessary, the images were zoomed,
enhanced for contrast, and adapted for auto-grey levels.
Point placement was then carried out using an on-screen
cursor. Several points were located by novel computer-
aided techniques, e.g. geometric calculation of sella after
digital contouring of sella turcica, visualization of the
deepest points of curves, tangent lines, and contrast edge
enhancement.

Coordinate measurements were recorded for the defined
soft tissue profile points (Figure 1), and linear and angular
values were calculated from a variable definition file. The
second digitizing session was carried out 1 week later.

Statistical analysis

Data were analysed using the Statistical Package for Social
Sciences version 12.0 for Windows (SPSS Inc., Chicago,
Illinois, USA). Data were checked visually for normal
distribution using normal probability plots. A repeated measure
analysis of variance (ANOVA) was performed to analyse the
influence of measurement conditions on measurement results:
it also included tests for sphericity. Measurement conditions
in this study were method (on-screen digitization and paper
tracings) and sessions (first and second) for each method. The
overall mean value for each variable was also calculated,
which was the average of the four measurements. The
difference between the sessions for both techniques and the
overall mean was then expressed as an absolute percentage of
the overall mean value per variable.

Results

The data were normally distributed and no violations in
sphericity were found. For both the linear and angular
variables, mean values with their standard deviations for the
two sessions per method and the \(P\) values of the ANOVA
analysis are presented in together with the differences
between the sessions per method and their absolute
percentage of the overall mean. When traced on-screen, 13
variables showed smaller mean differences between the
sessions for each method, and for four variables, this value
was smaller when traced by hand.

The factor method was significant for 11 of the 17
variables and approached significance for one variable, \(s_{0(i)}\)
to ML \((P = 0.059)\). The factor session was significant for
seven of the 17 variables and approached significance for
\(s_{s}−n_{s}=sm_{s}\) and NFL/NCL \((P = 0.085\) and \(P = 0.078,\)
respectively). The interaction term method \(×\) session was
significant for two variables, \(s_{0(i)}\) to ML and \(s_{0(s)}\) to OLs,
and it approached significance for three variables \(prn\) to \(n_{s}−
ss_{s}\) \((P = 0.070)\), \(s−n_{s}=unt\) \((P = 0.082)\), and NFL/NCL
\((P = 0.067)\).

Absolute percentage errors of the overall mean are
summarized in Figure 1. These errors ranged from 0.01
to 2.16 per cent except for two obvious outliers, Is to
NCL and li to NCL, which were not included due to the
scale of the \(y\)-axis. For the remaining 15 variables, 12
showed smaller percentual errors for on-screen tracings
and three errors were smaller for the manual tracing

Discussion

In this study, the errors in on-screen digitization were
compared with those in manual tracings at various levels.
Reproducibility still remains controversial as Geelen et al. (1998) found no significant differences for direct digital versus scanned or printed manual tracings of 20 radiographs. The reproducibility in that study for the digital modalities was slightly lower. It must be, however, borne in mind that the images for manual tracing had been digitally enhanced before printing, confirming the need for digital enhancement.

The area around stomion was the least reproducible because some subjects had their lips in contact while in others the lips were slightly apart when relaxed. Stomion is difficult to define, particularly in subjects with a lips apart posture, where the lowest dependent point on the upper lip (sto) and the highest point on the lower lip (sto) have to be estimated. The cephalometric software auto-assists in defining a single stomion, but the positions of two profile curves had, in some of the cases, to be adapted manually. The two extreme outliers for percentual deviations (the prominence of the upper and the lower lips) included the most prominent points of the lips. This is in agreement with Cooke and Wei (1991), who found lip prominence points to be poor landmarks. As the profile form, NCL/NFL, shows a relatively low percentual deviation (Table 1 and Figure 2), it may be assumed that lip distance to NCL is difficult to measure accurately. An interesting finding was the significant interaction between the session and the measurement method for the height of the lower lip (sto to ML) and the upper lip contact position (sto to OL). Although both variables contain the above landmarks, this significant difference cannot be satisfactorily explained. If the observer had a subjective preference for one measurement method, this may result in a systematic difference between the two methods. If a learning curve is present or if the observer is tired between the sessions, systematic difference may occur. There is no reason to assume that the behaviour of the observer changed per session for the different measuring techniques for sto to ML and sto to OL. Although both variables contain the above landmarks, this significant difference cannot be satisfactorily explained. If the observer had a subjective preference for one measurement method, this may result in a systematic difference between the two methods. If a learning curve is present or if the observer is tired between the sessions, systematic difference may occur. There is no reason to assume that the behaviour of the observer changed per session for the different measuring techniques for sto to ML and sto to OL. Although both variables contain the above landmarks, this significant difference cannot be satisfactorily explained. If the observer had a subjective preference for one measurement method, this may result in a systematic difference between the two methods. If a learning curve is present or if the observer is tired between the sessions, systematic difference may occur. There is no reason to assume that the behaviour of the observer changed per session for the different measuring techniques for sto to ML and sto to OL. Although both variables contain the above landmarks, this significant difference cannot be satisfactorily explained. If the observer had a subjective preference for one measurement method, this may result in a systematic difference between the two methods. If a learning curve is present or if the observer is tired between the sessions, systematic difference may occur. There is no reason to assume that the behaviour of the observer changed per session for the different measuring techniques for sto to ML and sto to OL. Although both variables contain the above landmarks, this significant difference cannot be satisfactorily explained. If the observer had a subjective preference for one measurement method, this may result in a systematic difference between the two methods. If a learning curve is present or if the observer is tired between the sessions, systematic difference may occur. There is no reason to assume that the behaviour of the observer changed per session for the different measuring techniques for sto to ML and sto to OL. Although both variables contain the above landmarks, this significant difference cannot be satisfactorily explained. If the observer had a subjective preference for one measurement method, this may result in a systematic difference between the two methods. If a learning curve is present or if the observer is tired between the sessions, systematic difference may occur. There is no reason to assume that the behaviour of the observer changed per session for the different measuring techniques for sto to ML and sto to OL. Although both variables contain the above landmarks, this significant difference cannot be satisfactorily explained. If the observer had a subjective preference for one measurement method, this may result in a systematic difference between the two methods. If a learning curve is present or if the observer is tired between the sessions, systematic difference may occur. There is no reason to assume that the behaviour of the observer changed per session for the different measuring techniques for sto to ML and sto to OL. Although both variables contain the above landmarks, this significant difference cannot be satisfactorily explained. If the observer had a subjective preference for one measurement method, this may result in a systematic difference between the two methods.

ANOVA showed that significant differences existed between the two methods for 11 outcome variables and between sessions for seven outcome variables. Clinically, these findings indicate that the two measurement methods differ significantly for 11 variables. It is, however, not possible to distinguish which technique results in the most reproducible outcomes because of lack of a reference value or gold standard. This problem, earlier described by Houston (1983), still remains the main barrier for any statements on validity in cephalometrics. Additionally, ANOVA ‘only’ analyses whether significant differences exist in the tracing results. The magnitude of the differences must be analysed in post hoc procedures. Nevertheless, Buschang et al. (1987) claimed that full factorial ANOVA was adequate to analyse reproducibility of a cephalometric analysis compared with the method error.

Figure 1 Soft tissue points—ct: chin tangent point. Lowest point on the NCL line; ft: frontal tangent point. Upper tangent point of the NFL line; li: labrale inferius. Most prominent point on the protrabium of the upper lip; ls: labrale superius. The most prominent point on the prolabilum of the upper lip; Int: lower nasal tangent point; ns: soft tissue nasion; pg: soft tissue pogonion. The soft tissue point overlying pg; prn: pronasale. The most prominent point on the apex of the nose; sn: subnasale. The deepest point of the nasolabial curvature, sm: soft tissue supramentale. The deepest point of the mental fold (sm to li); upper lip protrusion (sto to NL), depth of nasolabial curvature (sn to lnt); lower lip prominence (li to NCL), depth of mental fold (sm to li–pg), lower lip prominence (li to NCL), upper lip contact position (sto to OL), lower lip contact position (sto to OL). Angular variables—nasal protrusion (s–ns–unt), upper lip protrusion (s–ns–ss), lower lip protrusion (s–ns–sm), sagittal soft tissue relationship (ss–ns–sm), soft tissue chin protrusion (s–ns–pg), profile form (NFL/NCL).

The mean differences between sessions and methods were, except for s–ns–unt in the manual technique, all less than 1 degree or 1 mm, which is in broad agreement with previous studies. Santoro et al. (2006) found the same rates using a sandwich technique for detailing an intracranial cephalometric analysis with both techniques. A mean difference of 1 mm for linear variables between sessions using the digital technique and 1.09 mm in the conventional technique were also found by Hagemann et al. (2000). This is also in the line with the results of Chen et al. (2000), who compared 10 scanned hardcopies with their originals using ANOVA and found differences of up to 1 mm between methods.
Table 1 Overview of the mean values per session per measurement method, P values of the repeated measurements ANOVA with method, session, and method and session interactions. Significant values are shown in bold. The mean differences between the sessions per measurement technique are expressed as Δ Man (Manual) and Δ Dig (Digital), which are also expressed as the absolute percentage of the overall mean (|percent| Δ).

| Variable | Manual session | Manual session | Digital session | Digital session | ANOVA, method (P) | ANOVA, session (P) | ANOVA, method × session (P) | Overall x | Δ Man | Δ Dig | |percent| Δ Man | |percent| Δ Dig |
|----------|----------------|----------------|----------------|----------------|-------------------|-------------------|-----------------------------|-----------|-------|-------|--------------|------------|--------------|------------|
| n_s–sn   | 49.00 (3.43)   | 48.23 (3.69)   | 48.49 (3.63)   | 47.92 (3.44)   | 0.458             | 0.019            | 0.624                        | 48.41     | -0.78 | -0.57 | 1.60          | 1.18       |
| n_s–prn  | 41.38 (2.57)   | 40.58 (2.82)   | 41.26 (3.36)   | 40.88 (3.26)   | 0.855             | 0.008            | 0.110                        | 41.02     | -0.8  | -0.39 | 1.95          | 0.94       |
| prn to n_s–ss_s | 15.10 (1.58)   | 15.08 (1.62)   | 15.68 (1.71)   | 15.52 (1.80)   | 0.004             | 0.209            | 0.070                        | 15.34     | -0.03 | -0.17 | 0.16          | 1.08       |
| sto to NL | 24.20 (3.37)   | 24.05 (3.37)   | 23.56 (3.05)   | 23.83 (3.04)   | 0.008             | 0.596            | 0.147                        | 23.91     | -0.15 | 0.27  | 0.63          | 1.13       |
| sn to lat–ls | 7.73 (0.88)    | 7.70 (0.92)    | 6.90 (0.95)    | 7.04 (0.93)    | <0.001            | 0.478            | 0.257                        | 7.34      | -0.03 | 0.14  | 0.34          | 1.84       |
| ls to NCL | -1.33 (1.84)   | -1.30 (1.96)   | 1.39 (1.80)    | 1.29 (1.77)    | 0.004             | 0.575            | 0.236                        | 0.01      | 0.03  | 0.1   | 222.22        | 888.89     |
| sto to ML | 39.25 (3.67)   | 40.01 (3.59)   | 40.11 (3.68)   | 39.91 (3.52)   | 0.059             | 0.029            | 0.002                        | 39.83     | 0.8   | -0.2  | 2.01          | 0.50       |
| sm to li–pg   | 4.63 (1.41)    | 4.73 (1.41)    | 3.79 (1.79)    | 3.78 (1.70)    | <0.001            | 0.476            | 0.390                        | 4.23      | 0.1   | -0.01 | 2.37          | 0.24       |
| li to NCL   | -0.03 (2.53)   | 0.10 (2.54)    | 0.08 (2.48)    | 0.16 (2.46)    | 0.928             | 0.123            | 0.656                        | 0.08      | 0.1   | 0.06  | 131.15        | 72.13      |
| sto to OL–s | 5.13 (1.71)    | 5.25 (1.61)    | 5.37 (1.91)    | 5.25 (1.96)    | 0.576             | 0.931            | 0.046                        | 5.25      | 0.13  | -0.12 | 2.38          | 2.19       |
| sto to OL–s | 4.93 (1.87)    | 4.98 (1.81)    | 3.85 (3.08)    | 3.81 (3.04)    | 0.042             | 0.945            | 0.560                        | 4.39      | 0.05  | -0.04 | 1.14          | 0.91       |
| s–n_s–ss_s | 117.65 (5.31)  | 119.13 (4.91)  | 115.53 (4.97)  | 115.94 (4.14)  | 0.002             | 0.015            | 0.082                        | 117.06    | 1.48  | 0.42  | 1.26          | 0.35       |
| s–ns–ss_s | 94.35 (4.10)   | 94.33 (3.77)   | 92.01 (4.26)   | 92.33 (3.69)   | 0.001             | 0.020            | 0.213                        | 93.03     | 0.88  | 0.32  | 0.94          | 0.34       |
| s–ns–ss_s | 84.10 (4.58)   | 84.70 (4.35)   | 83.02 (4.78)   | 83.24 (4.34)   | 0.002             | 0.023            | 0.190                        | 83.76     | 0.6   | 0.22  | 0.72          | 0.26       |
| ss_s–ns–ss_s | 9.38 (1.56)    | 9.58 (3.16)    | 8.99 (3.17)    | 9.11 (1.54)    | <0.001            | 0.085            | 0.575                        | 9.26      | 0.2   | 0.12  | 2.16          | 1.24       |
| s–ns–pg   | 85.10 (4.81)   | 85.75 (4.62)   | 84.00 (5.01)   | 84.21 (4.63)   | 0.001             | 0.020            | 0.131                        | 84.77     | 0.65  | 0.21  | 0.77          | 0.25       |
| NFL/NCL   | 145.53 (4.63)  | 145.30 (4.67)  | 145.43 (4.70)  | 145.42 (4.75)  | 0.922             | 0.078            | 0.067                        | 145.42    | -0.23 | -0.01 | 0.15          | 0.01       |
The overall mean per variable might be considered as the gold standard as this value averages the most number of measurements. In fact, Chen et al. (2000) used a similar approach by averaging the measurements of the different operators. If this assumption is correct, the difference between sessions per method expressed as a percentage of the overall mean would be an assumption for the validity of that variable. In that case, the validity was better for the digital technique for 12 variables and for the manual technique for three variables (excluding the two outliers, ls to NCL and li to NCL) as these variables also showed a significant interaction between the methods.

Although the differences were statistically significant for several variables, the clinical relevance of the differences is limited. Excluding the two outliers, only for a minority of measurements (four variables for the manual measuring technique and one variable for the digital measuring technique) did the difference exceed 2 per cent of the overall mean (Table 1). This limit of a 2 per cent error was arbitrarily chosen as a clinically acceptable measurement error as previous assessments using this statistical technique are lacking.

Theoretically incorrect landmark identification, performed twice during hand tracing and during on screen digitization, does not result in different measurement results between the methods. Thus, an excellent reproducibility exists combined with poor validity. This possible source of bias may be present in the present data but cannot be determined because of the lack of a gold standard. A quality-related bias might be hidden in the printing process of the original images. This problem is hard to quantify, but is recognized by the authors to be of possible presence and importance. Further, the size sample might be a limitation in this study, but a sample of 20 radiographs was chosen because this is comparable with other cephalometric research (Battagel, 1993; Macri and Wenzel, 1993; Lim and Foong, 1997; Ongkosuwito et al., 2002; Kazandjian et al., 2006).

Thus, it appears that the main difficulties in the field of cephalometry remain the clinical relevance of the error and the lack of a gold standard for the cephalometric variables. The clinical relevance of the error can be dependent on the purpose of the analysis, whether this is for research or daily orthodontics, but may be limited. Moreover, the expertise of the operator has been claimed to impact on validity of measurements and the human errors introduced in the manual measurement procedure are reduced with the digital technique (Chen et al., 2004). There is a decrease in the number of errors due to the reduction in the number of procedures with the digital technique (Chen et al., 2004). The absence of a reference value, in turn, undermines many statistical statements concerning validity because of the individual variability of the human face. Finally, comparison of the manual tracing technique with on-screen digitization also has a number of practical aspects. Software with an automatic edge definition feature is a promising tool for more accurate and reproducible cephalometrics (Kazandjian et al., 2006). Further advantages such as the progress in digital acquisition of cephalometric images and data handling and storage stimulate further development of the on-screen tracing technique.

Conclusions

The findings of the present study indicate that the two measurement methods were statistically significantly different for 11 variables ($P = 0.001$ to $P = 0.042$). The area around stomion was the least reproducible.

Except for $s$–$n$–$unt$ in the manual technique, all mean differences between sessions and between methods were less than 1 degree or 1 mm. More variables were reproducible on-screen than with the manual technique. The clinical relevance of the registered differences remains questionable.

Address for correspondence

Dimitri P. Dvortsin
Department of Orthodontics
University Medical Center Groningen
Hanzaplein 1
P.O. Box 30.001
9700 RB Groningen
The Netherlands
E-mail: d.dvortsin@umo.umcg.nl

References

Broadbent B H 1931 A new X-ray technique and its application to orthodontia. Angle Orthodontist 1 : 45–66