Reliability of three methods of computer-aided thermal pattern analysis

John Hart, DC, MHS1
Bernard Omolo, PhD2
W.R. Boone, PhD, DC3
Charlene Brown, CT4
Andrew Ashton, BS, DC5

Background: The objective of this study was to assess three methods of computer-aided thermal pattern analysis for a) examiner reliability, b) inter-method differences, and c) determine which method yields the highest percent-similarity between paired test-retest scans.

Methods: Three examiners compared two sets of thermal scans from the same 30 subjects using three different methods of scan alignment. The results were evaluated by the Intraclass Correlation Coefficient and the Wilcoxon signed-rank test, at the 5% level of significance.

Results: Intra and inter-examiner ICC scores for all methods were acceptable (> 0.75). There were no statistically significant differences (at the Bonferroni-corrected level of significance of 0.0004%) in percent similarity of the scans between the three methods.

Conclusions: The results contribute evidence to the reliability of TPC program software. Manually aligning the readings plays an important role in obtaining precise TPC percent-similarities.

(JCCA 2007; 51(3):175–185)

KEY WORDS: thermography, vertebral subluxation, reliability

Contact: John Hart, DC, MHS, Sherman College of Straight Chiropractic, P.O. Box 1452, Spartanburg, South Carolina 29304. Phone: 864-578-8770, ext. 232, Fax: 864-599-4858, jhart@sherman.edu
1 Assistant Director of Research, Sherman College of Straight Chiropractic, Spartanburg, South Carolina.
2 Math and Computer Science, University of South Carolina-Upstate, Spartanburg, South Carolina.
3 Director of Research, Sherman College of Straight Chiropractic, Spartanburg, South Carolina.
4 Administrative Assistant, Research Department, Sherman College of Straight Chiropractic, Spartanburg, South Carolina.
5 Private practice.
© JCCA 2007.
Introduction
Most chiropractic definitions of vertebral subluxation include a misalignment component and a neurological component. One method of neurological assessment proposed involves thermography. One of the two methods of interpretation of thermography consists of assessing the amount of bilateral temperature asymmetry. This method presumes that the greater the asymmetry the more likely the patient’s overall neurology has been compromised. Thermal asymmetry has been noted in a number of health problems, including peripheral nerve involvement. Another method of thermography interpretation involves determining if the asymmetries are persistent; that is, whether the temperature differentials persist as a pattern. Although many factors can influence skin temperature, such as ambient room temperature, sunburn, and blemishes, thermal pattern analysis is based on the following general concepts:

1. Skin temperature is largely regulated by the autonomous nervous system.
2. Dynamic skin temperatures reflect a dynamic healthy nervous system evoking adaptive responses to the environment.

Used by chiropractors since the early 1940s, thermal pattern analysis has been applied in case studies as well as in a study regarding thermal equilibration. Since thermal pattern analysis is one method used by some chiropractors for assessing the neurological component of the vertebral subluxation, the present investigation was designed to assess intra- and inter-examiner reliability in regard to applying three different methods of aligning scans (graphs) for software analysis.

The Scan
The TyTron C-3000 [Titronics Research & Development, Oxford, IA] is a dual-probe infrared instrument that records temperatures on both sides of the spine and also performs a secondary calculation to determine bilateral temperature variations (delta). The TyTron instrument and protocol has been previously described elsewhere. A typical thermographic scan is begun at L5 and continues to the occipital shelf (Figure 1). The scanning procedure has been shown to have high reliability. Each scan is comprised of three vertical lines (readings) displayed on the computer monitor and is comprised of one line for each side of the spine and a third line representing the delta (side-to-side differences) (Figure 2). These lines are also referred to as channels. Figure 3 shows a series of left paraspinal readings. In the past, the simple qualitative observation method of inspection was used to assess the amount of similarity (pattern) between the readings. This is inherently a subjective process, and any analytic techniques which minimize this subjectivity and maximize objectivity are clearly desirable.

Ideally, a thermal scan stops at exactly the same point for every scan; however, in actuality this does not happen...
due to examiner variations when attempting to stop at a
given point on the spine for the purpose of comparing
successive scans. Consequently, the readings being com-
pared in the TPC do not line-up precisely (Figure 4). The
TPC reads data points between any two scans from side-
to-side, for 350 to 400 rows of temperatures on each side
of the spine; therefore, it is important that the readings be
vertically aligned as evenly as possible to achieve the
most precise calculation (Figure 5). The process of verti-
cal alignment can itself be a source of subjectivity. An-
other source of subjectivity of the analysis pertains to the
extensions of the upper and lower data points in the
graphic display (Figure 6). Extending the data points
could include up to a maximum of 20 rows of tempera-
tures, as part of 350 to 400 rows overall (Figure 5). Thus,
the examiner has the option of leaving the data in situ or
to slide the graphs to enhance vertical alignment as well
as extending the scans from the top or bottom if the data
are deemed to be useable. The sliding of the scans and
extending of the data points are herein referred to as mod-
ifications.

Due to these options, the question arises as to whether
the different approaches affect the outcome of the TPC
analysis. Further to the findings of Owens et al., this
study seeks to:
(a) Assess the intra-examiner reliability for each of three thermal pattern calculation (TPC) methods;
(b) Assess the inter-examiner reliability for each of three TPC methods;
(c) Assess the by-examiner inter-method differences of the three TPC methods; and
(d) Determine which of the three TPC methods yields the highest percent-similarity between paired test-retest scans.

Methods
The study was approved by the Sherman College IRB and participants signed a consent form. Using a convenience sampling method, 30 students (16 females, 14 males) were recruited from one of the authors’ classes for thermographic scanning purposes. For feasibility reasons, we did not perform random sampling; nevertheless, we have no reason to believe that our sample should yield data so biased to the extent of diametrically changing our
conclusions. Participants’ ages ranged between from 22–55 years (mean = 28.9 years, median = 26 yrs).

Each volunteer was scanned twice with the TyTron C–3000, with 10 minutes between the scans. The scans were performed by the principal investigator who had approximately six years of thermal scanning experience with this thermal instrument. From the TyTron program, the readings were then exported into a notepad (text) file consisting of several hundred rows of numerical temperatures for each reading. The longer the length of the participant’s spine, the more rows of temperatures there were. The notepad files were then imported into the TPC software where the two readings were graphically displayed and compared for similarity of slope as previously described.\textsuperscript{12}

A convenience sample of three independent examiners was selected from available research department employees. One of the examiners had approximately four years experience working with the TPC program, while the other two examiners each had approximately one month of experience using the TPC. A brief training session was given by the principal investigator for the other two examiners for the purpose of establishing consistency of procedures. Each examiner compared 60 thermal scans from 30 different participants (two scans per participant).

**TPC Methods 1–3**
Method 1 (Figure 6) involved vertically aligning the two scans from each of the 30 participants. When a scan is vertically aligned, the entire scan is moved for a best fit according to the examiner’s visual assessment. Once the examiner is satisfied, the “calculate” button is clicked and the TPC percent similarity (of slope) is returned. A second procedure of deleting or including extra data points

---

*The same data points as depicted in Figure 3 are also presented in this Figure. The graphs have been vertically aligned, but no modifications of data points (above the upper bar or below the lower bar) have been included (arrows).*
Thermal pattern analysis

(Figure 5), if deemed usable, was also a step in the protocol. Method 2 involved using the same amount of vertical alignment as in Method 1 but did not include the option of modifying extra data points (Figure 5). Method 3 (Figure 4) was entirely computer generated; the only requirement of the examiner was to press the “calculate” key for the TPC software to calculate the percent similarity.

The examiners performed Method 1 in the first session, Method 2 in a second session and Method 3 in a third session. The sessions ranged between from one to three weeks apart. It was considered highly unlikely that the examiners would recall how they had previously aligned or modified data on 30 sets of graphs from the previous session. However, if an examiner did recall any aspects of performing a prior method, that recall could possibly have influenced performance of the next method. Consequently, the precaution was taken to have each examiner attest, in writing, to their level of recall regarding the degree of aligning or modifying data that had transpired during the previous session. All reported they had no recall of how they had manipulated the scanned data while performing Method 1 or Method 2. As an additional measure to protect against any recognition of data, each thermograph scan was given a different identification number for each of the two trials and the sequence of the numbers changed as well. Recall was not

*Data points have been vertically aligned by the examiner to visually achieve the best match for the left and right side graphs. Modification of extra data points above the upper bar and below the lower bar, to include usable data, has been incorporated as well. If the examiner chooses to include data points above the upper bar or below the lower bar, as was done with Method 1, the amount of extra data points always defaults to the shorter graph. The darker reading is longer than the lighter reading (arrow) due to different lengths of scans (arrows).
an issue in Method 3, as that method was entirely computer driven, without any input from the examiners in regard to aligning or otherwise modifying data. Thus, each examiner was blinded to results of his/her previous session(s) as well as to any of the results of the other examiners. The examiners made hand-written percent notations for their respective calculated percent agreements of slope between the scans being compared in the TPC software. The examiners were assessed as to whether they each agreed on the resultant percent similarity returned by the TPC software after necessary alignment and optional data points procedures were addressed. These results were then manually entered into a spreadsheet for statistical analyses.

Statistics
Since the examiners were not randomly selected, the Fixed Effects (Model 3) Intraclass Correlation Coefficient (ICC) was used\(^1\) to determine the extent of intra- and inter-examiner agreement for all three methods. The minimum acceptable ICC score was set at \(0.750\)\(^1\) with an alpha level of 0.05 (5% level of significance). The ICC test assessed the TPC percents (slope similarity) obtained by the examiners.

The Kolmogorov-Smirnov and Shapiro-Wilk statistics demonstrated that the data were not normally distributed. Consequently, the non-parametric Wilcoxon signed-rank test, which does not assume normal distribution, was used to assess intra- and inter-examiner differences as well as differences between methods. The alpha level for the Wilcoxon test was set at 0.05 (5% level of significance) The ICC, Kolmogorov-Smirnov, Shapiro-Wilk, and Wilcoxon signed-rank tests were performed with the Statistical Package for the Social Sciences [SPSS Version 14.0, Chicago, IL].

Results

TPC methods
Visual depictions of the three methods are shown in Figures 4–6. The same thermographic temperatures (data points) are represented in these three figures, with explanations provided for each of the methods.

Examiner agreements
Method 1 intra-examiner reliability (repeatability) ICC scores ranged from 0.860 to 0.987 (Table 1). Inter-examiner ICC scores for this method ranged from 0.799 to 0.926 (Table 2). Method 2 intra-examiner reliability ICC scores ranged from 0.791 to 0.982 (Table 1). Inter-examiner reliability ICC scores for this method ranged from 0.805 to 0.941 (Table 2). Because Method 3 required no modifications to the graphs, intra and inter-examiner comparisons naturally did not reveal any significant differences and ICC scores were a perfect 1.00, reflecting the reproducibility of the TPC software (Tables 1 & 2).

Examiner differences
There were no significant differences for intra and inter-examiner agreement for any of the methods (Tables 1 and 2). There were 14 significant differences found (out of 36 comparisons) when comparing Method 3 to Methods 1 and 2 (Table 5). Methods 1 and 2 produced higher percentages of similarity when comparing paired test-retest graphs than Method 3 (Table 5). As shown in Table 5, there was a small, clinically trivial difference in percent-similarity between Methods 1 and 2, varying from 0.3 to 0.4, while a substantially larger difference in percent-similarity occurred between Methods 1 and 3 (4.0–7.1) and Methods 2 and 3 (4.3 to 7.5).

Discussion
Initially, in the present study, there was interest to investigate if a difference in experience using the TPC software would affect the outcome. Only one of the three examiners had considerable experience with the TPC software. The second examiner was a chiropractic student while the third was a chiropractic technician, both of whom had approximately one month of experience with the TPC. Thus, variation of technical training with the TPC software was considerable; however the high degree of repeatability of the three examiners suggests that the TPC procedures are readily learned, thereby potentially providing a valuable technology for a variety of practice settings.

The results show that different examiners can demonstrate a high level of within-method agreement, regardless of which of the three tested methods they apply; nevertheless, although Methods 1 and 2 maximized percent-similarity by affording the examiner the option of visually aligning the scanned data before importing to the TPC, and indicated good inter-method agreement with
each other, one-third of their results differed significantly from Method 3 (p < 0.05), the latter of which required no input from the examiner (Table 4).

Limitations of the Study

Because of the many multiple comparisons in the study, the original 5% level of significance (alpha) was subjected to a Bonferroni correction that considerably lowered the alpha (to 0.0004). This in turn, resulted in none of the inter-method comparisons being able to demonstrate any statistically significant differences (where with the traditional alpha of 0.05, there were 14), which suggests that some of our analyses may be subject to a Type 2 error (accepting an erroneous hypothesis). There is controversy as to whether Bonferroni adjustment should be used and consensus is lacking as to when it is appropriate.

The study participants do not represent a random sample from the general population (they were generally healthy, young adult students), so that the generalizability of our findings may be diminished; however, because there was a wide variation of thermal patterns (an important criterion in reliability testing), we judge our sample profile to be adequate for the purposes of our reliability analyses.

An ICC for inter-method analysis was not performed because of the differing protocols required in the different methods; thus it was believed that there would not be an apple-to-apples comparison being made which was thought to be a violation of ICC calculations.

These findings are preliminary and should be verified with further research. Furthermore, in light of the growing body of evidence demonstrating the reliability of the scanning method as well as to the TPC software, other validation aspects of thermal pattern analysis now require evaluation. While this study contributes information in regard to examiner reliability in applying more than one method of importing thermal scan data into the TPC software, as well as how those methods affect the TPC re-

---

Table 1  Three Methods Intra-examiner ICC Scores and P-values

<table>
<thead>
<tr>
<th>Channels</th>
<th>Examiners</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Left</td>
<td>Delta</td>
<td>Right</td>
<td>Left</td>
</tr>
<tr>
<td>Method 1</td>
<td>ICC</td>
<td>0.987</td>
<td>0.959</td>
<td>0.976</td>
</tr>
<tr>
<td></td>
<td>p-value</td>
<td>0.7</td>
<td>0.8</td>
<td>0.9</td>
</tr>
<tr>
<td>Method 2</td>
<td>ICC</td>
<td>0.980</td>
<td>0.969</td>
<td>0.982</td>
</tr>
<tr>
<td></td>
<td>p-value</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
</tr>
<tr>
<td>Method 3</td>
<td>ICC</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>p-value</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

1. Thirty subjects each had two thermographic scans taken 10 minutes apart. Three examiners twice compared the 30 pairs of scans by three different methods. Each method was applied in separate sessions (see Methods).
2. Intraclass Correlation Coefficients. All ICC scores exhibited the minimum acceptable level of 0.750 or greater. Method 3 was perfectly correlated as it involved no input from the examiner other than to press “calculate” on the computer’s keyboard. There were no significant differences between trials 1 and 2 for any of the three methods.
3. P values represent difference according to the Wilcoxon test (alpha= 0.05) and was used to assess differences between Trials 1 and 2.
Table 2  Three Methods\(^1\) Inter-examiner Agreement (ICC)\(^2\)

<table>
<thead>
<tr>
<th>Method</th>
<th>Trial 1 / Trial 2</th>
<th>Trial 1 / Trial 2</th>
<th>Trial 1 / Trial 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trial 1</td>
<td>0.829</td>
<td>0.799</td>
<td>0.918</td>
</tr>
<tr>
<td>Trial 2</td>
<td>0.910</td>
<td>0.881</td>
<td>0.926</td>
</tr>
<tr>
<td>Method 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trial 1</td>
<td>0.815</td>
<td>0.806</td>
<td>0.941</td>
</tr>
<tr>
<td>Trial 2</td>
<td>0.825</td>
<td>0.805</td>
<td>0.936</td>
</tr>
<tr>
<td>Method 3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trial 1</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Trial 2</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

1. Thirty subjects each had two thermographic scans taken 10 minutes apart. Three examiners twice compared the 30 scans by three different methods. Each method was applied in separate trials (see Methods). All ICC scores exhibited the minimum acceptable level of 0.750 or greater. Method 3 was perfectly correlated as it involved no input from the examiner other than to press “calculate” on the computer’s keyboard.

2. Intraclass Correlation Coefficients. All ICC scores exhibited the minimum acceptable level of 0.750 or greater. Method 3 was perfectly correlated as it involved no input from the examiner other than to press “calculate” on the computer’s keyboard. There were no significant differences between trials 1 and 2 for any of the three methods.

Table 3  Inter-examiner Differences for Three Methods\(^1\)

<table>
<thead>
<tr>
<th>Examiners</th>
<th>Left Channel</th>
<th>Delta Channel</th>
<th>Right Channel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A-B</td>
<td>0.9 / 0.5</td>
<td>0.8 / 0.6</td>
<td>0.6 / 0.8</td>
</tr>
<tr>
<td>A-C</td>
<td>0.4 / 0.4</td>
<td>0.2 / 0.4</td>
<td>0.4 / 0.6</td>
</tr>
<tr>
<td>B-C</td>
<td>0.4 / 0.9</td>
<td>0.3 / 0.6</td>
<td>0.6 / 0.8</td>
</tr>
<tr>
<td>Method 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A-B</td>
<td>0.9 / 0.3</td>
<td>0.9 / 0.6</td>
<td>0.8 / 0.8</td>
</tr>
<tr>
<td>A-C</td>
<td>0.2 / 0.2</td>
<td>0.2 / 0.3</td>
<td>0.4 / 0.6</td>
</tr>
<tr>
<td>B-C</td>
<td>0.2 / 0.8</td>
<td>0.3 / 0.6</td>
<td>0.6 / 0.7</td>
</tr>
<tr>
<td>Method 3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A-B</td>
<td>1.0 / 1.0</td>
<td>1.0 / 1.0</td>
<td>1.0 / 1.0</td>
</tr>
<tr>
<td>A-C</td>
<td>1.0 / 1.0</td>
<td>1.0 / 1.0</td>
<td>1.0 / 1.0</td>
</tr>
<tr>
<td>B-C</td>
<td>1.0 / 1.0</td>
<td>1.0 / 1.0</td>
<td>1.0 / 1.0</td>
</tr>
</tbody>
</table>

1. The numbers provided represent the p value of the Wilcoxon signed-rank test (alpha = 0.05 [Bonferroni corrected = .0004]) for inter-examiner differences across each trial separately. Thirty subjects each had two thermographic scans taken 10 minutes apart. Three examiners twice compared the 30 scans by three different methods. Each method was applied in separate sessions (see Methods).

Table 4  Inter-method Differences by Examiner and Trial\(^1\)

<table>
<thead>
<tr>
<th>Examiner</th>
<th>Trial</th>
<th>1 vs 2</th>
<th>1 vs 3</th>
<th>2 vs 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>0.7</td>
<td>0.8</td>
<td>0.1</td>
</tr>
<tr>
<td>A</td>
<td>2</td>
<td>0.8</td>
<td>0.9</td>
<td>0.07</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>0.6</td>
<td>0.8</td>
<td>0.1</td>
</tr>
<tr>
<td>B</td>
<td>2</td>
<td>0.9</td>
<td>0.8</td>
<td>0.2</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>0.8</td>
<td>0.8</td>
<td>0.3</td>
</tr>
<tr>
<td>C</td>
<td>2</td>
<td>0.9</td>
<td>0.9</td>
<td>0.2</td>
</tr>
</tbody>
</table>

1. Each examiner twice compared two graphs from 30 subjects for analysis of percent similarity. There were no significant differences between the methods.
Thermal pattern analysis

Conclusions

The TPC software program itself was shown in our study to be acceptably reliable. Aligning the readings, as used in Methods 1 and 2, appears to play an important role in maximizing TPC percent-similarity, and both of these methods descriptively demonstrated acceptable within- and between-method reliability. (Recall that Method 2, incorporated the same amount of vertical aligning as Method 1, but omitted the extra data inclusions.) Our study also provides evidence that the key TPC procedures we used are easily learned and reliable. The findings of this study prompted the initiation of improvements to the TPC program which enables the software to automatically align thermographs, thereby eliminating another area of examiner subjectivity in the use of the TPC. Future studies should involve: a) a random selection process for participants (from the general outpatient population) and examiners in order to better generalize the findings, (b) comparing Methods 1 and 2 with software that automatically aligns graphs for analysis, and c) investigating the relationship between TPC findings and other indicators of vertebral subluxation (construct validity).

Conflict of Interests

None declared

References


Table 5  TPC Means and Standard Deviations of Percent Similarity for Thirty Pairs of Thermographs

<table>
<thead>
<tr>
<th>Method</th>
<th>Left</th>
<th>Delta</th>
<th>Right</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>72.4</td>
<td>70.0</td>
<td>72.3</td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>7.8</td>
<td>8.1</td>
<td>9.5</td>
</tr>
</tbody>
</table>

| Method 2 | Mean | 72.7 | 70.4 | 72.7 |
| Std. Dev. | 7.7  | 8.6  | 9.8  |

| Method 3 | Mean | 68.4 | 62.9 | 67.3 |
| Std. Dev. | 10.1 | 12.7 | 11.1 |

Differences in Percent Similarity Between the Three Methods

<table>
<thead>
<tr>
<th>Methods</th>
<th>1:2</th>
<th>1:3</th>
<th>2:3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.3</td>
<td>4.0</td>
<td>4.3</td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>0.4</td>
<td>7.1</td>
<td>7.5</td>
</tr>
</tbody>
</table>

1. Thirty subjects each had two thermographic scans taken 10 minutes apart. Three examiners twice compared the 30 pairs of scans by three different methods. Each method was applied in separate trials (see Methods).


