Interexaminer Reliability of Supine Leg Checks for Discriminating Leg-Length Inequality

H. Charles Woodfield, RPh, DC,[a,b] B. Burt Gerstman, DVM, MPH, PhD,[c]
Renate Henry Olaisen, DC, MPH,[d] and Dale F. Johnson, PhD[e]

Abstract

Objective: The purpose of this study was to quantify interexaminer reliability of a standardized supine leg check procedure used to screen for leg-length inequality.

Methods: Two doctors of chiropractic used a standardized supine leg check procedure to examine 50 volunteers for leg-length inequality. The order of examination was randomized. The side and magnitude of leg-length inequality were determined to the nearest 1/8 in. Subjects and examiners were blinded. Interexaminer reliability was assessed with a Bland-Altman plot, tolerance table of absolute differences, a quadratic weighted κ statistic for quantitative scores, and a Gwet's first-order agreement coefficient for dichotomous ratings.

Results: The quadratic weighted κ statistic to quantify the reliability of the rating scale was 0.44 (95% confidence interval, 0.21-0.67), indicating moderate reliability. The 2 examiners agreed exactly 32% of the time, within 1/8 in 58% of the time, within 3/16 in 72% of the time, and within 3/8 in 92% of the time. The Bland-Altman plot revealed possible heterogeneity in reliability that requires additional study. The examiners agreed on the presence of a leg-length inequality of at least 1/8 in in 40 (80%) of 50 subjects (first-order agreement coefficient, 0.76), suggesting good agreement for this diagnostic category.

Conclusion: The examiners showed moderate reliability in assessing leg-length inequality at 1/8-in increments and good reliability in determining the presence of a leg-length inequality. (J Manipulative Physiol Ther 2011;34:239-246)

Key Indexing Terms: *Leg-Length Inequality; Chiropractic; Reproducibility of results; Observer variation*

Examination for leg-length inequality (LLI) as a sign of neuromuscular dysfunction and vertebral misalignment is a common screening and diagnostic procedure used in chiropractic and other manual therapies.[1-3] There are several methods for assessing LLI. These include radiographic examination, orthopedic procedures (eg, tape measure), and quick visual and tactile checks. The most common chiropractic methods are visual and tactile checks performed in either the prone or supine position.[1,3] Despite widespread use of such procedures, their reliability and validity are still uncertain.[3-9]

Leg-length inequality is classified as either anatomical or functional. Anatomical LLI refers to measured differences in the bony anatomy of the lower extremity. The criterion standard for identifying and measuring anatomical LLI is computed tomographic scanogram.[10] Knutson[6] estimates that approximately 90% of the population have anatomical differences in leg length, averaging 5.4 mm (approximately 3/16 in), SD of 4.1 mm, whereas noting a difference of more than 20 mm (approximately 3/4 in) is considered clinically significant in contributing to various musculoskeletal pathologies. Anatomical LLI is due to congenital, traumatic, iatrogenic (eg, hip replacement), neoplastic, degenerative, and infectious cause.[6,10]

Functional LLI, as distinct from anatomical LLI, is hypothesized to be the result of asymmetric neurophysiologic responses that arise along the kinetic chain.[3] Although the biomechanical basis of this phenomenon is controversial, reliance on tests for functional LLIs remains at the core of many chiropractic practices and procedures.[1,7,9] There is still limited evidence in the literature to support their clinical relevance.[1,5] To show if these procedures are useful, both reliability and validity are necessary for rational diagnosis and treatment.

[a] Researcher, Upper Cervical Research Foundation, Raleigh, NC.
[b] Life Chiropractic College West, Hayward, CA.
[c] Faculty, Department of Health Science, San Jose State University, San Jose, CA.
[d] Part-time Faculty, Department of Health Science, San Jose State University, San Jose, CA.
[e] Director of Research, Research Department, Life College of Chiropractic West, Hayward, CA.
Submit requests for reprints to: H. Charles Woodfield, RPh, DC, 822 Woodbine Way, Bellingham, WA 98229
(e-mail: *chuckwoodfield@yahoo.com*).
Paper submitted December 6, 2010; in revised form March 23, 2011; accepted April 3, 2011.
0161-4754/$36.00

doi:10.1016/j.jmpt.2011.04.009

Recent reports on prone leg checks suggest good reliability in identifying the side of the short leg in subjects with LLIs and low back pain.[7-9] These studies estimate between 72% and 85% overall agreement and κ statistics between 0.61 and 0.66, indicating good overall agreement. In contrast, little is published in the indexed biomedical literature on supine leg checks (SLCs). One abstract published in the *Journal of Chiropractic Education* suggested that examiners using the SLC could estimate length differences to 1/4 in 87% of the time and to within 1/8 inch 67% of the time.[11] A nonindexed article by Hinson and Brown[12] found intraclass correlation coefficients (ICCs) for interrater reliability that ranged from 0.42 to 0.96 for 9 different examiners.

Since the 1940s, orthogonal-based, upper cervical chiropractors have used a standard SLC procedure as a screening tool in their evaluation of patients.[13-15] In applying this screening procedure to new patients, a threshold of 1/4-in LLI has been used to indicate possible upper cervical vertebral misalignment. Established patients have a lower threshold of 1/8 in according to the SLC procedure. Leg-length inequalities detected at these thresholds indicate the need for further postural and radiographic assessment. Use of these recommendations in patient care is intended to prevent excessive x-ray exposure and imprudent treatment.[13,14,16] Studying the reliability of these thresholds is a necessary prerequisite for future studies that seek to untangle relationships between LLI, upper cervical misalignment, and various types of neurophysiologic pathologies.[17] Thus, the primary purpose of this study was to quantify the interrater reliability of the standardized National Upper Cervical Chiropractic Association (NUCCA) SLC procedure.

Methods

Study Participants and Examination Methods

The protocol of this project (P/N 2010-005) was approved by the Life Chiropractic College West Institutional Review Board (IRB 00007071) on May 11, 2010 using policies derived from Department of Health and Human Services standards and codes of federal regulations Title 45, Part 46 of the Code of Federal Regulations (revised 2005). Volunteers were recruited from students, faculty, staff, and patients at Life Chiropractic College West in Hayward, CA. Recruitment had been encouraged by public posters, e-mails to staff and faculty, and e-mails to interns requesting participation by their patients. Before giving informed consent, volunteers were warned that they may experience some discomfort in their feet and back as they were instructed to recline, lie into, and rise from a supine position on an examination table. They were also advised to avoid participation in the study if unable to lie on their back for up to 15 minutes. Volunteers with lower extremity trauma within the past 30 days were excluded from study. After informed consent, participants completed a basic self-report intake questionnaire that included queries of a history of prior LLI diagnoses, presence of spinal curvature and scoliosis, presence of prior lower extremity trauma, use of a shoe lift or orthotic, and whether they had been a patient of one of the current examiners. Participants ranged in age from 12 to 67 years (mean, 28.8 years; SD, 9.8 years). Of the 50 participants, 29 were male. Table 1 summarizes additional characteristics of the study subjects. Examinations took place on May 22, 2010.

Table 1. *Characteristics of study participants (n = 50)*

Age distribution (y)	n (%)
<20	1 (2)
20-29	36 (72)
30-39	6 (12)
40-49	3 (6)
≥50	4 (8)
Race	
White, non-Hispanic	39 (78)
Hispanic	5 (10)
Asian	6 (12)
Other self-reported information	
Been told 1 leg is shorter	35 (70)
Been told they have spinal curvature or scoliosis	13 (26)
Reported history of injury below pelvis	12 (24)
Was a patient of one of the examiners	5 (10)
Uses shoe lift or orthotic	6 (12)

Participants were assigned to 5 groups of 10 to determine when during the day that they would be examined. As groups were called, study subjects were led by a research assistant approximately 50 yd from the enrollment area to a common staging area. The 2 examination rooms were closed and located in separate wings of the facility and were equidistant from the staging area. Examinations were conducted by experienced, board-certified NUCCA practitioners in these closed rooms. The order of examinations was randomized. Half the subjects were examined first by examiner number 1, and the other half were examined first by examiner number 2. Between examinations, subjects walked the length of a long passageway (approximately 100 yd) to "washout" potential residual effects of their first examination and to reestablish normal posture and gait.

Subjects entered the examination room in their street clothes and normal footwear. Upon entering the examination room, the examiners assessed subjects' footwear. At their discretion, examiners were allowed to put subjects into special examination shoes (Adjusting Shoes; Activator Methods International, Ltd, Phoenix, AZ) if they judged this would better allow them to detect and measure an LLI.

Next, a prepared script was used to instruct the subject to stand squarely at the foot of the adjustment table, sit down, scoot backward on the table using only their hands, lie on their back with head supported by the headpiece, and place

their hands in a folded position atop their abdomen. Examiners then squatted at the foot of the table with their body weight slightly leaning forward on the balls of their feet. The subject's feet were then cupped in the hands of the examiner with the thumbs of the examiner oriented alongside the lateral aspect of the ankle, not allowing the examiner to assume an opposing grip atop the dorsal aspect of the foot. The examiner then exerted a slight headward pressure (no more than 5 lb) to assure the soles of the subject's shoes were in the same plane. Leg lengths where then evaluated with tactile and visual assessments by noting relative heel positions at the shoe-sole interface. The side and extent of LLI were classified as: (*a*) left leg shorter by more than 1/4 in, (*b*) left leg shorter by between 1/8 and 1/4 in, (*c*) left leg shorter by less than 1/8 in, (*d*) legs even, (*e*) right leg shorter by less than 1/8 in, (*f*) right leg shorter by between 1/8 and 1/4 in, or (*g*) right leg shorter by more than 1/4 in. These class intervals are based on diagnostic categories recommended by NUCCA protocol.[13,14,16] Upon completion of the examination, results were recorded on a case report form by the examiner.

For some analyses, we considered class intervals c, d, and e to be near even. This is consistent with NUCCA practice, which considers any LLI of less than 1/8 in inconsequential for diagnostic and treatment choices.

Statistical Methods

Data were double entered and validated using version 3.1 of the EpiData Reliable Data System (EpiData Association, www.epidata.dk, Odense, Denmark). The analysis module of EpiData was used to merge data and cross-tabulate rating pairs.

A quadratic weighted κ statistic was used to quantify reliability. This approach was used because it accounts for closeness of rating pair discrepancies while simultaneously adjusting for random agreement. Quadratic weights were selected because our class intervals were not uniformly spaced, quadratic weights are routinely used for ordinal rating scales, and the quadratic weighting derives a κ that is equivalent to a ICC.[18] Benchmarks of Landis and Koch[19] are adopted as a way to help readers interpret κ results. We note that these benchmarks do not address imprecision estimates and other κ nuances (including possible heteroscedasticity).

We used WinPEPI (PEPI-for-Windows, Jerusalem, Israel) to calculate κ point estimates and 95% confidence intervals (CIs).[20] Confidence intervals were calculated without assuming a κ of 0 (formula 18.36).[21] See Abramson[20] for technical details about WinPEPI. WinPEPI is a public domain software that can be downloaded from http://www.brixtonhealth.com/.

Gwet's first-order agreement coefficient (AC1) was used to quantify the extent of agreement for binary ("nominal") ratings (eg, presence/absence of LLI). This is especially important when marginal totals in two-by-two concordance tables are asymmetric. Tables with asymmetric marginal totals tend to yield overly low κ values in the face of high agreement—a phenomenon known as "the κ paradox."[22,23] The AC1 statistic corrects for the κ paradox while providing a chance-corrected index of agreement.[24] The AC1 is preferable to other prevalence- and bias-adjusted κ statistics, such as Prevalence-Adjusted Bias-Adjusted Kappa (PABAK).[25]

Analysis for quantitative differences in assessments was achieved by using class interval midpoints, with class interval midpoints as follows: (*a*) subjects classified as "even" were assigned an LLI of 0 in; (*b*) subjects with LLI classified as "less than 1/8 in" were assigned a value of 1/16 or 0.0625 in; (*c*) subjects with LLI classified as "between 1/8 and 1/4 in" were assigned a value of 3/16 or 0.1875 in; (*d*) subjects with LLI classified as "at least 1/4 in" were assigned a value of 5/16 or 0.3125 in. Assigning an expectation of 5/16 in for observations in this last category could underestimate the midpoint of this interval if a large number of subjects in the subgroup had LLIs that exceeded 3/8 in. However, only 6% of the observations fell into this high LLI class interval (4 of 50 for examiner number 1 and 2 of 50 for examiner number 2). In addition, the examiners reported no LLIs in excess of 1/2 in in post–study debriefings. Therefore, we believe that 5/16 in is a reasonable estimate for the average LLI for these observations.

Calculations of mean values and SDs were based on class interval midpoints. Let m_i represent the midpoint of interval i and f_i represent its frequency. Then,

$$\bar{x} = \frac{\sum f_i m_i}{\sum f_i}$$

and

$$s = \sqrt{\frac{\Sigma f_i (m_i - \bar{x})^2}{(\Sigma f_i) - 1}}.$$

These are, respectively, the weighted average of class interval midpoints and the root mean square error of class interval midpoints using observed frequency as interval weights.

In addition, a Bland-Altman plot was used to explore patterns of discrepancies for all 50 rating pairs. The Bland-Altman plot (1986) is a popular method for exploring the reliability for quantitative biomedical measurements.[26] It is superior to any omnibus index of reliability (whether a weighted κ or second-order agreement coefficient (AC2) statistic) because it displays all rating pair differences and reveals patterns that may otherwise be hidden by a summary index. The Bland-Altman plot was created with PASW (SPSS) release 18 (IBM Corporation, Somers, NY).

Reliability can also be explored by creating a "tolerance table" of the absolute differences of rating pairs. The tolerance table lists counts and cumulative percentages falling below set class intervals. The

Table 2. *Interrater agreement for LLI, side, and estimated magnitude of shorter leg, κ = 0.44 (95% CI, 0.21-0.67)*

	Examiner 1							
Examiner 2	Left >1/4 in	Left 1/8-1/4 in	Left <1/8 in	"Even"	Right <1/8 in	Right 1/8-1/4 in	Right >1/4 in	Total
Left >1/4 in	2	1	0	0	0	0	0	3
Left 1/8-1/4 in	4	4	0	5	0	6	2	21
Left <1/8 in	0	1	0	1	0	0	1	3
Even	0	0	0	0	0	0	0	0
Right <1/8 in	0	0	0	0	0	0	0	0
Right 1/8-1/4 in	0	0	0	2	0	6	3	11
Right >1/4 in	0	2	0	3	0	3	4	12
Total	6	8	0	11	0	15	10	50

percentage of ratings that fall below specific cut points can be viewed as potentially acceptable levels of tolerance for that cut point.

Sample size requirements for this study were based on achieving a margin of error of 0.2 κ units with 90% confidence while assuming a 50/50 split of right-leg/left-leg LLIs and an expected κ of 0.5. To achieve these conditions, 49 paired observations were needed. Note that this sample size calculation is based on statistical precision, not statistical power, because the goal was to estimate κ, not test it for statistical significance against a null hypothesis. The sample size of 50 pairs is also justified in other recently published studies on the reliability of LLI examinations, which used 45 and 46 observations, respectively.[8,9] Finally, this study was limited to 50 pairs so that the study could be completed in a timely manner. A sample size of 50 pairs, thus, provided a balance of reasonable precision and timely results.

Results

Table 2 cross-tabulates results for all 50 rating pairs. Counts on the diagonal from the upper left to lower right of the table represent perfectly concordant ratings. Cells near this diagonal represent "close" ratings. The opposite diagonal, from the upper right to the lower left, represents perfectly discordant ratings. The quadratic weighted κ, which takes into account the closeness of rating pairs, is 0.44, suggesting moderate agreement overall.[19] The 95% CI for κ is 0.21 to 0.67, suggesting a range of κs consistent with fair to substantial agreement.

Table 3 lists frequencies for absolute differences in rating pairs. Thirty-two percent of the ratings were in perfect agreement, 58% of the rating pairs were within 1/8 in of each other, 72% were within 3/16 in, 92% were within 3/8 in, and 100% were within 1/2 in. The mean absolute difference was 0.169 in or approximately 1/6 of an inch (SD, 0.162 in).

Figure 1 displays a Bland-Altman plot. This plot displays differences in rating pairs (y-axis) against averages of rating pairs (x-axis). For example, study subject number 11 on this plot shows an LLI rating of −0.3125 (left leg shorter by approximately 5/16 in) by both examiners. Thus, the difference is 0; and the mean rating is −0.3125. In contrast, study subject number 5 received an LLI rating of 0.3125 by examiner 1 (right leg shorter by approximately 5/16 in) and a rating of −0.1875 by examiner 2 (left leg shorter by 3/16 in). Thus, the difference is 0.3125 − (−0.1875) = 0.5000; and the mean is [0.3125 + (−0.1875)]/2 = 0.0625. The Bland-Altman plot suggests excellent agreement in the left-most and right-most wings of the graph and poor agreement in the middle of the graph, near mean ratings of zero.

Table 3. *Tolerance table: absolute difference in raters' assessments*

Absolute difference LLI (in)	Decimal equivalent	Approximate mm equivalent	Count	%	Cumulative percent
0	.0000	–	16	32	32
1/16	.0625	2	1	2	34
1/8	.1250	3	12	24	58
3/16	.1875	5	7	14	72
5/16	.3125	8	3	6	78
3/8	.3750	10	7	14	92
1/2	.5000	13	4	8	100
Total	–	–	50	100	–

The examiners agreed on the presence of an LLI in 39 (78%) of the 50 rating pairs. The κ for this result is 0.00, maximum attainable κ* of 0 (note presence of "the κ paradox"); and the AC1 statistic is 0.73. If we consider an LLI of less than 1/8 in as evidence of "evenness" (a standard NUCCA definition), then there were 40 (80%) agreements in the 50 rating pairs for a κ of 0.13 (maximum attainable κ of 0.13; κ paradox) and AC1 statistic of 0.75. Thus, the overall agreement for detecting whether an LLI is present is good according to the "new model" benchmarks of Gwet.[24]

Table 4 exhibits results for the 39 paired observations in which both examiners provided data about the side of the shorter leg. The examiners agreed on 28 (72%) of these pair ratings (κ = 0.45; 95% CI, 0.19-0.71). The AC1 coefficient

* Ceiling value given the observed marginal totals.

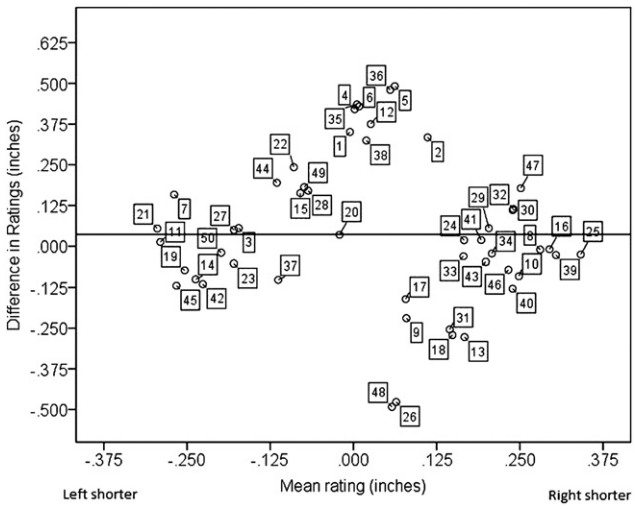

Fig 1. *Bland-Altman plot. Numbers in the plot represent subject identification numbers. Data points have been jittered to show overlapping points.*

for this observation was 0.44. (Note the absence of the κ paradox.) Thus, agreement for identifying the side of the shorter leg is moderate.

Discussion

As far as the authors are aware, this article represents the first publication on the reliability of the SLC procedure to appear in the indexed biomedical literature. This study adds to the body of evidence on the reliability of chiropractic checks for leg length inequality. It uses a double-blinded approach to investigate the interexaminer reliability of the procedure. The order of examinations was randomized, and a washout was used between examinations to mitigate potential biases that may have otherwise resulted from the first examination. The diagnostic procedure and diagnostic thresholds mirror practices currently used to make clinical decisions.

The point estimate for the κ reliability statistic is 0.44, indicating moderate agreement. This κ statistic accounts for the closeness of discrepant rating pairs, providing a statistic that is equivalent to an ICC. The κ statistic of 0.44 is the "maximally likely" estimate based on data in this specific sample. The CI for the κ parameter is 0.21 to 0.67, thus accounting for sampling imprecision. The CI is consistent with reliabilities that can be characterized as fair to substantial.[19]

Cicchetti[27] recommends a minimal sample size of at least 98 rating pairs when using the weighted κ statistic for a 7-point ordinal scale. Given our limited sample size of 50, we reanalyzed our data using 5 class intervals by combining the "near-even categories" (>1/4 in short left leg, 1/4 to 1/8 in short left leg, "near even," 1/8 to 1/4 in short right leg, >1/4 in short right leg). The quadratic weighted κ statistic for the 5-level rating scale was essentially unchanged (κ = 0.45; 95% CI, 0.23-0.68). The minimum required sample size for a 5-ordinal scale rating system according to Cicchetti[27] is 50.

Table 4. *Interrater agreement for side of the shorter leg in which both examiners offered a rating on the side of the short leg, κ = 0.45 (95% CI, 0.19-0.71)*

	Examiner 1		
Examiner 2	Left	Right	Total
Left	12	9	21
Right	2	16	18
Total	14	25	39

Using a threshold of between 0 and less than 1/8 in, examiner number 1 rated 11 subjects as even; examiner number 2 rated those same 11 subjects as 6 with a short left leg and 5 with a short right leg. We cannot assume how examiner number 1 would have rated the side of the short leg had he been encouraged to choose, so we have excluded these 11 rating pairs from consideration.

The SLC is used as a screening procedure in NUCCA practice. Two screening thresholds are applied. In new patients, a 1/4-in threshold determines the need for further diagnostic testing. In established patients, a 1/8-in threshold is used. Unlike other LLI diagnostic procedures, the side of the LLI detected in a SLC is relatively unimportant because the detection of a leg-length asymmetry is a "go/no-go" type of clinical decision.[3] The examiners in our study agreed on the presence of a functional LLI of at least a 1/8 in in 40 (80%) of the 50 rating pairs (AC1, 0.75). According to the new model benchmarks of Gwet,[24] this indicates good agreement.

The Bland-Altman plot suggests excellent agreement in the wings of the graph and poor agreement in the middle of the graph. The reason for this heteroscedasticity (unequal variance) may include several factors including: (*a*) a mix-up of "right" and "left" (errare humanum est); (*b*) positioning difficulties, examination inconsistencies, and other types of examination errors; and (*c*) heterogeneity—a 2-population hypothesis: the ability to derive reliable results in some patient populations but not in others. Investigating the outliers on the Bland-Altman plot provides clues about these potential sources of error.

A Mix-up of "Right" and "Left" (Errare Humanum Est)

There were 7 observations with 3/8-in discrepancies between the rating pairs (subject nos. 1, 2, 4, 6, 12, 35, and 38) and 4 rating pairs with 1/2-in discrepancies between the rating pairs (nos. 5, 26, 36, and 48). In 9 (82%) of these 11 rating pairs, the magnitude of the LLIs was identical; but the short leg was discordant. We might assume that the examiners just did not agree on the side of the short leg. However, they may have also observed left and recorded right (and vice versa). It is easy to see how this type of error could occur in some

of these 9 discordances because in the supine position, a presenting short left leg is on the examiner's right side (and vice versa). Wrong-side confusions in surgery, radiology, and other medical procedures occur more frequently than most people realize.[28-30] The current study did not put any markings on the patient or table to indicate right and left, but future studies on this topic will incorporate such markings to reduce the potential for this type of error.

Positioning Difficulties, Examination Inconsistencies, and Other Types of Examination Errors

Our study was intended to emulate the clinical experience using the established SLC procedure. Reliable LLI measurements, however, require consistent procedural and examination practices. We are able to identify 2 sources of clinical judgment that may have influenced the reproducibility of results. These are (1) inconsistent use of special examination shoes and (2) inconsistent methods in positioning of subjects into supine position. The examiners varied in their decisions to use special examination shoes. Examiner number 1 used the special examination shoes in 25 subjects, whereas examiner number 2 used the special examination shoes in 7 subjects. In addition, examiner number 1 put the shoes on the subjects when they were in the supine position, whereas examiner number 2 required subjects to place the shoes on themselves. Nguyen et al[7] reported a high level of examiner disagreement when special examination shoes were used. The inconsistent use of special examination shoes in our study may have confounded the results. Future studies may benefit from controlling for shoe wear in an attempt to improve consistency and reduce possible confounding.

We observed inconsistencies in the manner and extent to which the examiners repositioned subjects. Inconsistency in positioning may influence the degree of neck extension and alter sensory and neuromuscular responses in subjects.[13,31-34] In our study, examiner number 1 had 13 positioning difficulties compared with 3 such reported difficulties for examiner number 2. In addition, the angle of the headpieces of the tables was not identical. During and after the study was complete, inconsistencies were noticed in the height and inclination of the headpiece of the examination tables. These types of inconsistencies may have contributed to discordant results.

In contrast to our study, several other LLI reliability studies did not reposition subjects between examinations. Instead, in past studies, examiners moved from subject to subject.[7-9,12] Stationary subjects removed 1 source of variability but does not account for interexaminer comparisons of the entire LLI procedure used in practice. We addressed the interrater reliability of the entire LLI check procedure, including setup, positioning, and evaluation. For reliability study results to be relevant, subjects should be studied in near clinical conditions.[35]

Heterogeneity—a 2-Patient Population Hypothesis

This hypothesis suggests that lack of agreement in some pair ratings might derive from use of the procedure in a heterogeneous patient population: one population in which LLI ratings are stable, and another population, unstable, producing temporary unstable results in the SLC procedure. An asymptomatic population with relatively stable neuromusculoskeletal fixations may express consistent leg-length asymmetries, even when ambulating between examinations. In contrast, the absence of neuromusculoskeletal fixations may express itself as unstable LLI ratings. Inherent patient instability may limit the usefulness of LLI testing, particularly when applied indiscriminately to a mix of symptomatic and symptom-free populations.[7]

Limitations

Study participants were derived predominantly from staff, students, and patients of a chiropractic college. In addition, the study did not screen for the presence or absence of pain. The sample, therefore, may not be representative of a typical clinical population. Future study must recruit subjects within this demographic.

Inclusion and exclusion criteria for LLI investigation vary greatly. Prior LLI studies have restricted the study population by prescreening for established LLI and presence of back pain. Interexaminer reliability may increase in a symptomatic study population.[8] One study excluded subjects that received a recent chiropractic adjustment.[9] These facts are relevant because high levels of interexaminer agreement can be expected with leg-length discrepancies greater than 4 mm (approximately 3/16 in).[36] During post–study debriefing, examiners in our study described subjects as atypical, with observed LLIs less than what was normally observed in their practices. In the absence of prescreening, examiners were challenged to discriminate small leg-length inequalities, consistent with nonsymptomatic patient population. Our population limits the generalizability of the findings.

In an attempt to permit the examiners to use their clinical judgment and emulate clinical conditions, our examiners relied on subjective visual and tactile assessment. Measurements were not confirmed using objective devices such as rulers, the Anatometer (Benesh Corporation, Monroe, MI); x-ray examinations; or computed tomographic scanograms. In contrast, other studies investigating the reliability of LLI assessments have relied on measuring devices as part of their assessment.[4,5,8,37]

Our study had the relatively narrow research focus of examining interrater reliability for 2 particular examiners. The extent to which our results apply to less experienced

practitioners is uncertain. In addition, our study did not verify if LLI existed in the absence of a criterion standard (if indeed there is one) for determining the presence of a functional LLI. Inconsistencies remain in how to screen for functional LLIs.[5]

Our examiners experienced difficulties positioning participants for examination and were allowed to decide whether to reposition subjects. Complexities arising from subject movement, positioning, and repositioning were not controlled for in an attempt to emulate clinical conditions.

One of the examiners had prior clinical experience with 20% of the subject population. Prior experience with these subjects may have influenced these observations through knowledge of previous examinations, thus introducing an unintended source of bias.

Our limited sample size prevented us from studying reliability within subgroups. In addition, the limited sample size produced a 95% CI for κ of 0.21 to 0.67, which is consistent with fair to superior reliability. A larger study would remedy these limitations.[19] Several sample size requirement scenarios were considered for future study. Note that sample size requirements depend on several underlying assumptions, including the underlying κ value, desired precision, and nonresponse rates. Using an assumed expected κ value of 0.45, a margin of error of 0.10 and a nonresponse rate of 15% will require 358 paired observations.[20]

Reliability is distinct from validity, both of which are required for rational and effective clinical decision making. Wright and Feinstein[35] consider unreliable clinical measurements to be the result of 3 sources of variability: (*a*) the patient, (*b*) the procedure, and (*c*) the clinician. Investigation to identify patients with possible unstable leg lengths may be necessary to diminish this source of variability. Inconsistencies in implementing the procedure and clinician performance can be addressed through additional standardization of the procedure and training of clinicians.[9,36,38,39] Future studies are necessary to establish reliable standards in each of these 3 areas. We, therefore, intend to pursue a follow-up study in which we use a population-based sample, more refined examination technique, and a sample that is at least 5 times the current sample size.

Conclusion

Two experienced examiners independently estimated the side and magnitude of the LLI to within approximately 1/8 of an inch in 50 volunteer subjects. Overall results indicate moderate agreement between examiners. The Bland-Altman plot, however, suggested strong reliability in the wings of the graph and weak reliability in the center of the graph, indicating possible heterogeneity in reliability by subgroup. The significance and impact of this phenomenon require further investigation.

Practical Applications

- Two trained examiners demonstrated moderate agreement in determining the side and magnitude of length inequalities using a standardized supine leg check procedure.

Acknowledgment

The authors thank Dr. Lee Yardley, Dr. Kim Khauv and Dr. Michael Zabelin for their contributions to this study.

Funding Sources and Potential Conflicts of Interest

This study was funded in part by the Upper Cervical Research Foundation, the Tao Foundation, and Life Chiropractic College West. No additional conflicts of interest were reported for this study.

References

1. Mannello DM. Leg length inequality. J Manip Physiol Ther 1992;15:576-90.
2. Walker BF, Buchbinder R. Most commonly used methods of detecting spinal subluxation and the preferred term for its description: a survey of chiropractors in Victoria, Australia. J Manip Physiol Ther 1997;20:583-9.
3. Knutson GA. Anatomic and functional leg-length inequality: a review and recommendation for clinical decision-making. Part II. The functional or unloaded leg-length asymmetry. Chiropr Osteopat 2005;13:12.
4. Cooperstein R, Morschhauser E, Lisi A, Nick TG. Validity of compressive leg checking in measuring artificial leg-length inequality. J Manip Physiol Ther 2003;26:557-66.
5. Cooperstein R, Morschhauser E, Lisi AJ. Cross-sectional validity study of compressive leg checking in measuring artificially created leg length inequality. J Chiropr Med 2004; 3:91-5.
6. Knutson GA. Anatomic and functional leg-length inequality: a review and recommendation for clinical decision-making. Part I, anatomic leg-length inequality: prevalence, magnitude, effects and clinical significance. Chiropr Osteopat 2005;13:11.
7. Nguyen HT, Resnick DN, Caldwell SG, Elston EW, Bishop BB, Steinhouser JB, et al. Interexaminer reliability of activator methods' relative leg-length evaluation in the prone extended position. J Manip Physiol Ther 1999;22:565-9.
8. Schneider M, Homonai R, Moreland B, Delitto A. Interexaminer reliability of the prone leg length analysis procedure. J Manip Physiol Ther 2007;30:514-21.
9. Holt KR, Russell DG, Hoffmann NJ, Bruce BI, Bushell PM, Taylor HH. Interexaminer reliability of a leg length analysis procedure among novice and experienced practitioners. J Manip Physiol Ther 2009;32:216-22.
10. Sabharwal S, Kumar A. Methods for assessing leg length discrepancy. Clin Orthop Relat Res 2008;466:2910-22.
11. Hinson R, Pfleger B. Pre and post-adjustment supine leg-length estimation. J Chiropr Educ 2000;14:37-8.
12. Hinson R, Brown S. Supine leg length differential estimation: an inter- and intra-examiner reliability study. Chiropr Res J 1998;5:17-22.

13. Gregory RR. Model for the supine leg check. Upper Cervical Monograph 1979;2:1-5.
14. Wiedemann RI. The supine leg check. In: Thomas M, editor. NUCCA: protocols and perspectives. 1st ed. Monroe: National Upper Cervical Chiropractic Association; 2002. p. 3-1-3-8.
15. Eriksen K. Leg length inequality. In: Eriksen K, editor. Upper cervical subluxation complex-A review of the chiropractic and medical literature. 1st ed. Philadelphia: Lippincott, Williams, and Wilkins; 2004. p. 131-62.
16. NUCCA standards of practice and patient care. 1st ed. Monroe: National Upper Cervical Chiropractic Association; 1994. p. IV-3.
17. Bakris G, Dickholtz M, Meyer PM, Kravitz G, Avery E, Miller M, et al. Atlas vertebra realignment and achievement of arterial pressure goal in hypertensive patients: a pilot study. J Hum Hypertens 2007;21:347-52.
18. Fleiss JL, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educ Psychol Meas 1973;33:613-9.
19. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-74.
20. Abramson JH. WINPEPI updated: computer programs for epidemiologists, and their teaching potential. Epidemiol Perspect Innovations 2011;8:1. Available from: http://www.epi-perspectives.com/content/8/1/1.
21. Fleiss JL, Levin B, Cho Paik M. Statistical methods for rates and proportions. 3rd ed. Wiley: Hoboken; 2003.
22. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. J Clin Epidemiol 1990;43:543-9.
23. Cicchetti DV, Feinstein AR. High agreement but low kappa: II. Resolving the paradoxes. J Clin Epidemiol 1990;43:551-8.
24. Gwet KL, editor. Handbook of inter-rater reliability. 2nd ed. Gaithersburg: Advanced Analytics, LLC; 2010.
25. Gwet LI. Computing inter-rater reliability and its variance in the presence of high agreement. Br J Math Stat Psychol 2008;61:29-48.
26. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;1:307-10.
27. Cicchetti DV. Assessing inter-rater reliability for rating scales: resolving some basic issues. Br J Psychiatry 1976;129:452-6.
28. Simon JW, Ngo Y, Khan S, Strogatz D. Surgical confusions in ophthalmology. Arch Ophthalmol 2007;125:1515-22.
29. Elghrably I, Fraser SG. An observational study of laterality errors in a sample of clinical records. Eye (Lond) 2008;22:340-3.
30. Sangwaiya MJ, Saini S, Blake MA, Dreyer KJ, Kalra MK. Errare humanum est: frequency of laterality errors in radiology reports. AJR Am J Roentgenol 2009;192:W239-44.
31. Magoun HW. Caudal and cephalic influences of the brain stem reticular formation. Physiol Rev 1950;30:459-74.
32. Rhines R, Magoun HW. Brain stem facilitation of cortical motor response. J Neurophysiol 1946;9:219-29.
33. Grostic J. Dentate ligament—cord distortion hypothesis. Chiropr Res J 1988;1:47-55.
34. Knutson GA. SLC as a determinant of physiologic/postural LLI-a case study. Chiropr Res J 2000;7:8-13.
35. Wright JG, Feinstein AR. Improving the reliability of orthopaedic measurements. J Bone Joint Surg Br 1992;74:287-91.
36. Fryer G. Factors affecting the intra-examiner and inter-examiner reliability of palpation for supine medial malleoli asymmetry. Int J Osteopathic Med 2006;9:58-65.
37. DeBoer KF, Harmon RO, Savoie S, Tuttle CD. Inter- and intra-examiner reliability of leg length differential measurement: a preliminary study. J Manip Physiol Ther 1983;6:61-6.
38. Degenhardt BF, Snider KT, Snider EJ, Johnson JC. Interobserver reliability of osteopathic palpatory diagnostic tests of the lumbar spine: improvements from consensus training. J Am Osteopath Assoc 2005;105:465-73.
39. List T, John MT, Dworkin SF, Svensson P. Recalibration improves inter-examiner reliability of TMD examination. Acta OdontolScand 2006;64:146-52.