A review of observational gait assessment in clinical practice

B. Toro, C. Nester and P. Farren

For physical therapists observational gait assessment is preferable to instrumented gait assessment in the clinical setting. Existing observational gait assessment tools used in clinical practice were reviewed in terms of their validity, repeatability, sensitivity, and specificity. Most gait assessment tools lacked information concerning their validity or repeatability. In addition, most gait assessment tools had not had their sensitivity and specificity adequately evaluated. Current approaches to the development and use of observational gait assessment tools have been characterised by a top down approach, designing what is useful without any understanding of the appropriateness of a tool’s content, construction, and so forth. We would advocate a more structured approach, by assessing gait using the gold standard of instrumented gait analysis and designing an observational gait assessment tool based on the data.

INTRODUCTION

Observational gait assessment is frequently carried out by physical therapists to assess patients’ gait and forms a major aspect of physical therapy practice. For physical therapists observational gait assessment (using the naked eye or video images) is preferable to instrumented gait assessment in the clinical setting (Yack, 1984; Krebs, Edelstein, and Fishman, 1985). Whilst instrumented gait assessment that provides quantitative measures of three-dimensional gait kinematics and kinetics and the electrical activity of muscles remains the gold standard for gait assessment, in the context of routine clinical practice it is still restricted by the fact that it is laboratory based, expensive, and requires a high-level of interpretation skills (Messenger and Bowker, 1987; Davis, 1997; Geurts et al, 1990; Morton, 1999; Coutts, 1999).

The problem with observational gait assessment (OGA), however, is that it is relatively subjective in nature and it has been suggested that this subjectivity may lead to poor validity, reliability, sensitivity, and specificity when compared to the more objective instrumented gait analysis (Krebs, Edelstein and Fishman, 1985; Eastlack et al, 1991; Saleh and Murdoch, 1985; Goodkin and Diller, 1973; Keenan and Bach, 1996.) In addition, these qualities are difficult to test in the context of observation based clinical gait assessment.

In the context of gait analysis, validity refers to the degree to which the gait assessment measures the actual events of gait, but...
there are various levels to this quality. Face validity, that is that the gait assessment tool appears to measure what it is thought to measure, is easily justified, but this is the lowest level of validity and is based entirely on personal opinion. Reasoned arguing that a gait assessment tool adequately reflects and measures selected variables of gait can be used to explore the construct validity of a gait assessment tool. For example, a theoretical construct may be that a gait abnormality will manifest itself in abnormal joint positions of a limb during gait. Having defined the construct (as determined by expert opinion) and that the gait assessment tool has been developed to contain the domains that are required to adequately assess gait, its content is considered valid (content validity). The importance of the expert opinion here is that content validity is entirely context and purpose orientated, and different experts would suggest that the content of gait assessment tools be different for different patient groups and applications. Like construct validity, being reliant on logical arguments in support of the gait assessment tool means that content validity is highly dependent upon clinicians’ genuine expertise in gait assessment and understanding of gait pathologies.

The highest level of gait assessment tool validity (criterion validity) relates to the degree to which the assessment made using the tool reflects actual gait events. This is best achieved by comparing the gait assessment tool to the gold standard measurement of gait, which is instrumented analysis of gait kinematics, kinetics, and muscle activity. Tools that can demonstrate criterion validity have the greatest justification for their use since we can have confidence in their ability to genuinely reflect actual gait events. One important assumption made here, however, is the assumed validity of the gold standard. The validity of three-dimensional gait analysis, for example, is known to be affected by incorrect marker placement and excessive skin movement (della Croce, Cappozzo, and Kerrigan, 1999).

Reliability refers to the repeatability of a gait assessment tool and can be estimated based on the statistical concept of variance between repeated measures of the same gait event by the same rater (intra-rater repeatability) or different raters (inter-rater repeatability). The greater the dispersion of repeated scores of the same event the greater the variance and the lower the repeatability, and the more homogeneous the repeated scores the smaller the variance and the greater the repeatability. Acceptable variation between repeated scores is, of course, a rather subjective threshold and context driven. However, in the context of clinical gait analysis, acceptable variation is perhaps variation between repeated scores that is less than that needed to produce error in subsequent clinical decisions, or conclusions regarding treatment efficacy, based on the gait assessment. Agreement between raters (inter-rater repeatability) can be evaluated using the weighted Kappa coefficient. Intra-rater repeatability can be represented statistically using a variety of methods, such as Pearson product moment for correlation, or Kendall’s coefficient for measures of concordance. It also has been reported by finding the value of the least significant difference (LSD) between the first and the second assessment of the same gait cycle or event (Read et al, 2003). This value corresponds to the smallest difference in the assessment of gait on two different occasions that would be expected to occur due to intra-clinician variation, and not due to an actual change in the gait. As an alternative, Intra-class Correlation Coefficient (ICC) can be used to reflect both the degree of consistency and agreement amongst data. Crucially, and reflecting the complexity of intra and inter rater repeatability, using more than one statistical test will provide a better picture of reliability (Bruton, Conway, and Holgate 2000).

Sensitivity and specificity are essential qualities of a clinical gait assessment tool. Sensitivity relates to the tool’s ability to identify true and clinically relevant deviations from normal gait or actual changes in gait, for example due to an intervention. Specificity refers to the ability of the tool to correctly identify normal gait and indicate no change in gait from a previous gait assessment. The ideal tool is highly sensitive, so it correctly detects a high percentage of all those
individuals whose gait is abnormal and correctly detects genuine changes in gait. It is also highly specific, so it correctly detects a high percentage of those whose gait is normal as being normal and detects no change in gait when it has not changed. Importantly, however, the change in gait must be considered alongside a genuine measure of the same change in gait, which has been measured using the gold standard method, in this instance quantitative gait analysis.

The preceding statements set the scientific background to this area where clinical practice and scientific measurement and rigor meet. Validity and the different aspects of reliability are benchmarks against which gait assessment tools can be assessed. In this context, the aim of this paper is to explore evidence to support how existing observation-based gait assessment tools meet these benchmarks. The purpose of doing this is two-fold. First, we wish to highlight issues pertinent to the quality of OGA. Second, highlighting where there are gaps in existing knowledge with regard to existing OGA tools, we wish to identify areas where future research is required to either support the use of existing OGA tools or justify the development of new tools.

METHOD

This review considered gait assessment tools that were specifically developed to assess the biomechanics of gait and typically examined ranges of joint motion throughout the gait cycle. We assume that authors chose variables related to joint kinematics (e.g., ankle position at initial contact, knee position at mid-swing) as they describe gait, they relate closely to the variables investigated with quantitative gait analysis, they can inform accepted paradigms of gait pathology, and since they describe specific joints and deformities, they are linked to treatment options and paradigms. Assessment tools that evaluated the overall function or performance of gait or mobility (i.e., Wolfson et al, 1990; Shumway-Cook and Wollacott, 1995, Novacheck, Stout, and Tervo, 2000) were not considered since they may shed little light on the cause of impairment at an individual joint level where many treatment decisions are made. The tools were not specific to a given population.

Between January 2001 and November 2002 we systematically reviewed the Medline, PubMed, CINAHL, Recal, Embase, AMED, PsychINFO, Cochrane Library, National Research Register (NRR), Clinical Trials Database, and the Chartered Society of Physiotherapy databases for literature regarding gait assessment tools used in physical therapy practice. Information also was sought from Internet-based research information sources, the NHS Centre for Reviews and Dissemination, Index to Theses, the Medical Research Council, Physiobase e-mail discussion forum, and conference abstracts not published in journals. The search also included personal communications. The keywords used for all searches, in isolation and in various combinations, were: gait analysis, outcome measures, clinical observation, visual gait assessment, video gait assessment.

RESULTS

“Naked eye” observation of gait

The Benesh Movement Notation (BMN), developed by Rudolph Benesh (Harrison, Atkinson, and De Weerdt, 1992), can be used to record posture and movement sequences by drawing symbols on a five-horizontal line stave, similar to a music transcription stave. Its principal use is within the context of choreography, and although some physical therapists have used it as a basis for posture and movement assessment of patients with walking disabilities, the number to receive formal BMN training was only 21 worldwide in 1992 (Harrison, Atkinson, and De Weerdt, 1992) and since than there has been no further publications using the BMN. It has therefore failed to be widely accepted in clinical practice. Criterion validity of the notation scheme has never been established, however two repeatability studies by Churton (1987), cited by
Harrison, Atkinson, and De Weerdt (1992), demonstrated that less than 3% of 1815 observations made by 11 trained “advanced” notators varied between notators (observations were made of the entire body position during static posture and an undefined movement sequence as performed by one model). The complex notation is designed to detect even small changes in body position and movement to be noted, even facial expression and finger movements, and so sensitivity to change is potentially good. However, sensitivity and specificity studies of the BMN have not been found in the literature. The available literature sheds little light on the appropriateness of BMN for gait assessment and its complex language structure has been and continues to be a significant barrier to its uptake in routine clinical practice.

In order to overcome the problem of poor uptake of existing OGA tools, including the BMN, Winter (Winter, 1985) developed the Waterloo Gait Profile Form. It is used either with naked eye examination alone or combined with sagittal plane video recordings. This form was more specific to gait assessment and clinician orientated, and therefore easier to understand than the BMN. It enabled the description of pathological gait characteristics of the trunk, knee, and foot throughout the gait cycle in adults and children. Stick diagram notation and symbols laid out on an examination form allowed the clinician to quickly circle the stick diagram, notation, and symbol that applied to their patient. It assumed that the stick figures, notation, and symbol options covered all possible types of gait pattern, and there is therefore a danger that this tool might not represent all gait patterns, thus having poor content validity. Consequently, there is the possibility that a clinician is forced to choose one of the options on the form, even if it doesn’t adequately represent his or her patient. The meaning of the symbols and angular notation is not always clearly defined, and although there is an extensive appendix to the form, the explanations are not easily interpreted. No data has been presented to demonstrate the criterion validity of the Waterloo gait profile compared to quantitative motion analysis of the trunk, knee, and foot. There have been no reliability or sensitivity and specificity studies reported in the literature.

The form that is probably quoted most frequently in the literature is the Rancho Los Amigos System (Olsson, 1990; Perry 1992). It was developed by the staff at Rancho Los Amigos Hospital in California, USA, to meet the needs of a busy physical therapy department to assess the gait of a wide range of patient groups. It was felt that the staff were able to identify the patient’s gait abnormalities more quickly using the form, hence making the use of instrumented gait assessment less important (Gronley and Perry, 1984). The system is based on a form that is laid out in a tick box format. The user identifies deviations from normal gait by ticking a box for the frontal, sagittal, and transverse plane rotations at the trunk, pelvis, hip, knee, ankle, and toes through the whole gait cycle. Deviations from normal gait are categorised as minor or major, but definitions for these are not available in the general literature. The physical therapist needs to be specifically trained to use this form and have a complete understanding of kinematic terminology and normal gait (Coutts, 1999). Its comprehensive nature leads to the system to be time consuming. We have been unable to find validity, reliability, sensitivity, and specificity studies relating to the Rancho Los Amigos System in the public literature. Working on the basis of absence or presence of deviations from normal, a binary system, it provides little opportunity for noting the subtle differences between patients or assessment intervals, other than the deviation being minor or major. Furthermore, it describes many movements on a binary scale using unspecific wording such as “limited” and “excess.” For instance, the user is asked to indicate if the patient has excess plantar flexion of the foot. How much is excess plantar flexion? If the patient’s plantar flexion reduces after treatment, how much less excess plantar flexion is required before the patient should be considered normal? The lack of clear definition of the boundaries between the different categories such as “inadequate,” “normal,” and “excessive,” would promote
greater reliability. The Rancho Los Amigos System has proved popular and is certainly comprehensive in its coverage of the triplaner movements at the trunk, hip, knee, and ankle. However, lack of definitions in some of the deviations and the lack of differentiation between different gait deviations may compromise its sensitivity and specificity.

The Rivermead Visual Gait Assessment (RVGA) was developed to evaluate the gait of adults with neurological disorders and measures upper limb, trunk, hip, knee, and ankle in stance and swing phases. A four-point scale (0–3) enables the user to grade joint or segment positions as either normal (0), mild (1), moderate (2), or severe (3), and where appropriate indicate the direction of deviation (e.g., forward inclination of trunk or backward inclination of trunk). There is information describing normal gait that is intended to guide the users as to their own individual definitions as to what mild, moderate, and severe deviations from normal might be, but this immediately suggests that a difficulty with inter-rater repeatability might exist. A study by Lord, Halligan, and Wade (1998a) suggested significant criterion-validity of the RVGA when tested with ten patients suffering from Multiple Sclerosis against a 10 metre walking time (\( r = 0.77; \ p < 0.001 \)), the Rivermead Mobility Index (Hsieh, Hsueh, and Mao, 2000), functional balance (Berg et al, 1992), stride length (\( r = -0.61; \ p < 0.005 \)) and step length asymmetry (Robinson and Smidt, 1981). However, these are not appropriate gold standards to which to compare the RVGA and assess whether it can accurately describe the actual gait kinematic and kinetic characteristics of a patient’s gait, a true test of the criterion validity of a gait assessment tool that assesses joint positions and motion. Criterion validity using more appropriate measurements such as quantitative gait kinematics and kinetics has not been established.

Lord, Halligan, and Wade (1998a) found the scores given by an undisclosed number of clinicians in the assessment of 10 patients with Multiple Sclerosis agreed exactly on 63.8% of occasions, suggesting moderate inter-rater repeatability. The raters, who were instructed in the use of the RVGA form, consisted of physical therapists with wide ranging clinical expertise with no specific knowledge of gait analysis. The intra-rater repeatability was evaluated by one rater by assessing the gait of six patients on two separate occasions seven days apart. However, this rater did not assess video recordings of the same gait cycles on these two occasions. This method is flawed because of the likely high variability in patients’ gait between the seven days and the reported intra repeatability data is therefore invalid. The subjective nature of the 0–4 grading system and the incomplete coverage of all possible gait deviations (for example ankle inversion is included but eversion is excluded), reduce the likely sensitivity and specificity of the scores. However, Lord, Halligan, and Wade (1998b) compared two treatment approaches (facilitation and task-orientated) and pre and post treatment scores using the RVGA, walking time, stride length, Berg Balance score (Berg et al, 1992), and Rivermead Mobility Index on 20 patients with Multiple Sclerosis treated three times a week for five to seven weeks. The RVGA, and the other scores, were able to detect the improvements in post-treatment scores.

The Physician Rating Scale (Koman et al, 1994) examines the hip, knee, and the equinus foot in the sagittal plane at undefined phases of the gait cycle; the hind foot at foot strike; and the speed of gait, scoring each on scales ranging from 0–3 and 0–4. It is used to examine gait of children with cerebral palsy, either with naked eye examination alone or together with video recordings. Criterion validity of the Physician Rating Scale (PRS) has not been reported, however, a modified version of the Physician Rating Scale, the Hugh Williamson Gait Laboratory Scale (HWGLS), was shown to have modest criterion validity when compared to quantitative gait assessment of sagittal plane foot and knee joint kinematics (Kappa 0.46–0.61) (Pirpiris et al, 2001). This is the first instance in the available literature that an observational gait assessment tool has been compared to quantitative gait kinematic data.
The fact that this comparison has highlighted some failings of the Physician Rating Scale (modified as HWGLS) demonstrates the importance of establishing the true criterion validity of an observational gait assessment tool. Whilst these results may appear damming, it should be remembered that none of the other tools can be considered superior since they have not been subject to such scrutiny.

Another variant of the PRS, the Observational Gait Scale (OGS), was modified by Boyd and Graham (1999) to improve the assessment of the plantar flexion/knee extension couple in midstance. The OGS was found to have modest validity against quantitative kinematic data (Kappa 0.69, range 0.38–0.94), using two observers and 20 children with spastic diplegia, and limiting the investigation to the first four sections of the OGS (Mackey et al, 2003).

HWGLS failed to achieve acceptable inter-rater repeatability between four experienced raters for 25 children with spastic diplegic gait (Kappa 0.46 for foot-strike) (Pirpiris et al, 2001), though Corry (1995) had previously demonstrated an inter-rater repeatability of Kappa 0.67 for the foot-strike section. The OGS also was found to have modest inter-rater repeatability (Kappa 0.58, range 0.29–0.86) and intra-rater repeatability (Kappa 0.69, range 0.30–0.91) for the first six sections (Mackey et al, 2003) using two experienced observers.

In terms of sensitivity and specificity there are also some indicators. The HWGLS and instrumented kinematic gait analysis were equally able to pick up a treatment effect of surgical intervention and botulinum toxin therapy in the same 25 children, but the treatment effect detected by the HWGLS was larger than that detected with gait analysis, implying a potential overestimation of effect. The original Physician Rating Scale was used to evaluate the effects of Botulinum Toxin-A calf injections on the function of the lower leg with 12 4–11 year old children with cerebral palsy (Koman et al, 1994). An improvement in gait between the botulinum toxin injected children and a placebo group was detected using the PRS (83% and 33% improvement, respectively). The PRS, and its derivative were sensitive to genuine changes in gait, but in these two studies the changes were reasonably large, and therefore its sensitivity to smaller changes remains unclear.

Video-based observational gait assessment (VOGA)

In an attempt to overcome some of the problems associated with naked eye evaluation of gait, such as the speed of movements and only seeing the gait cycle once, video recording of gait has become popular in the clinical setting. What is subsequently derived from the video data varies considerably between the different methods of gait assessment. Some investigators have used goniometers to measure joint angles directly from the television screen. A criterion validity study by Stuberg, Colerick, Blanke, and Bruce (1988) measured hip, knee, and ankle joint ranges of 10 children with cerebral palsy (5 diplegia, 2 quadriplegia, 2 hemiplegia) and nine non-disabled children, which demonstrated intra-class correlation coefficients between 0.84–0.99 when compared to two-dimensional high speed cinematography. Although Embrey, Yates, and Mott (1990) reported very high inter-rater repeatability (ICC 0.934) between two raters who measured knee flexion from the television screen using a small goniometer this was based on assessment of just one subject suffering from cerebral palsy. One of the raters consisted of the principal investigator, possibly introducing some bias into the study. Goniometers could provide a quick and simple semi-instrumented method of gait assessment, but the advantage of quantified data is at the expense of the time consuming process that has proved to be a major barrier to clinicians using instrumented gait assessment to date, and so goniometers are unlikely to be successful in the clinical setting. In addition to this the inter- and intra-rater repeatability of using goniometers on a television screen has not been established and the inter-rater repeatability using goniometers on patients is generally unsatisfactory (Rome and Cowieson, 1996; McDowell et al, 2000).
Hughes and Bell (1994) developed a standardised form to aid structured video-based observational gait assessment of six stroke patients, the hemiplegic gait assessment form. The authors looked at the hip, knee, and ankle in the sagittal and frontal planes during swing and stance phases. Attempts to validate the form using temporal-distance walking data were inconclusive (correlation statistic \( r = 0.60 \); statistically not significant), although temporal-distance parameters are not an appropriate gold standard for a tool designed to indicate joint kinematics. Inter-rater repeatability, tested with three experienced and trained raters, was said to be best for swing phase where Kendall’s coefficient of Concordance was >0.75 and statistically significant at \( p < 0.05 \). Kendall’s Coefficient for the stance phase section was >0.70. The intra-rater repeatability scores for all sections of the tool were found to lie between 0.39 and 0.97 (Kendall’s Coefficient of Concordance) using three raters. A limitation of this element of the study was the failure to randomise the sequence of the gait cycles viewed from the videotapes, introducing the potential for a learned effect, and the large variation of the time between both rating sessions (it varied between 2 and 14 days). Sensitivity and specificity of the hemiplegic gait assessment form has not been demonstrated. The form is not likely to be sensitive to small changes in gait, again because this method uses a three point scale (normal, barely noticeable abnormal, and definitely abnormal) and these categories are too subjective. Other aspects of the rating scale are dependent on the rater’s perception of terms such as “excessive,” “poor,” “too short,” which are not clearly defined.

Hainsworth, Harrison, Sheldon, and Roussounis (1997) used the Rank Scoring System to evaluate the effectiveness of ankle-foot orthoses in 12 children with cerebral palsy (8 diplegic, 4 hemiplegic). This was similar to the scoring system used by Watt et al (1986). The system scores the ankle in terms of its position at initial contact, evidence of toe walking, and whether the foot is in a pronated position. The knee is scored in terms of its position of extension at heel strike and in mid-stance. Evidence of criterion validity of the system has not been found in the literature. The limited categories of assessment for this score question the content validity, though this is highly context dependent and it could be argued that in some instances even these limited assessments might suffice. Inter-rater repeatability, evaluating 12 children with cerebral palsy, was found to be very good with a 90% agreement between three raters, although we have no information with regard to their experience and knowledge of gait assessment. Although the tool was sensitive to changes in sagittal plane ankle motion when ankle foot orthoses were withdrawn, there is little further information on its sensitivity nor its specificity.

Read, Hillman, Hazlewood, and Robb (2003) described the development of the Edinburgh Visual Gait Score, which evaluates the position of the trunk, pelvis, hip, knee, ankle, and foot in the sagittal and coronal planes, plus the transverse plane position of the pelvis. Observations are made on a 3-point scale of normal and moderate and marked deviations from normal. The validity of the scale has been measured by comparing videotaped gait sequences taken from five children using a Vicon system. Agreement between the gait score and the quantitative kinematics for the ten numerical gait items that measured ranges of movement of the ankle, knee, hip, and pelvis ranged from 47% for maximum knee extension in stance to 83% for maximum ankle dorsiflexion in swing (mean 64% agreement). Inter-rater repeatability for all 17 items on the gait score, across five experienced raters ranged from 96% for initial contact to 55% for knee extension in terminal swing (mean 70% inter-rater repeatability). Intra-rater repeatability was reported to be good for all five raters, demonstrated by a mean Least Significant Difference (LSD) of 3.20 (range 2.63—4.01).

Sensitivity and specificity have been explored by demonstrating changes in the total gait scores due to surgical intervention, which we would expect to have gross effects of joint positions and gait. The mean score reduction on the scale was 4.2 points (ranging from +0.3
to 8.5) where the two patients undergoing multilevel surgery showed the largest score changes. The ability of the gait score to detect smaller changes in gait due to less invasive treatment methods has not been shown. The gait score is accompanied by extensive guidelines giving normal gait data and defining terminology used within the form. Whilst the fact that the scoring system has a score to indicate overall change in gait, this might not inform the clinician on what to treat and how to treat.

In order to reduce the problem of subjectivity, attempts have been made to enhance the means by which we make visual observations. Hillman et al., (1988), for example, demonstrated that their Segment Rotation Indicators improved the ability of raters to correctly identify transverse plane hip position of one normal subject during walking (mean difference of measurements was 0.3° with a 95% confidence interval from -1.7 to 2.2°). This was achieved by having a visual aid to facilitate the evaluation of joint position, when compared to simply looking at the limbs themselves. The raters consisted of four senior physical therapists and four orthopaedic senior registrars who were experienced in estimating transverse plane rotations visually. Perhaps, using these types of attachments might prove important in establishing good inter/intra-rater repeatability in the future. Owing to their size the thigh rotation indicators are limited to children over the age of around six years, whose leg rotation is not too severe, as otherwise the indicators will interfere with the opposing limb. Validity of these indicators has been explored by comparing the results of the observations to 3-D gait analysis. The maximum disagreement between estimated rotation angle and that measured using instrumented kinematic analysis was 3.8°. Sensitivity and specificity of the Segment Rotation Indicators have not been demonstrated.

Table 1 presents the available information about observation based gait assessment tools that we have reviewed.

### DISCUSSION AND CONCLUSION

Whilst it is not appropriate to compare directly different gait assessment tools, as each is designed for different purposes, there are some problems that are common to several tools, particularly with regard to the methods used to create evidence supporting the use of the tools. When criterion validity was evaluated, it was sometimes evaluated by comparing the assessment from the OGA tool to an existing tool that did not assess the same dimension. For example, a tool assessing joint position would be compared to stride length, walking distance, instead of the gold standard instrumented kinematic analysis. How can comparing stride length to joint position tell us that the assessment of joint position is valid? The concept of comparing one OGA tool to a gold standard appears to have been misinterpreted by some as meaning the comparison of an OGA tool to any other gait measures, even if they assess different gait parameters to the OGA tool being evaluated. This is not appropriate, the gold standard against which any OGA tool should be evaluated should itself measure the same gait parameters as the OGA.

The issues of face, content, and construct validity of gait assessment tools have not been explicitly addressed in the literature. We assume that in the context of the authors own work that these were taken for granted. If face and content validity had been low then the authors would most likely have chosen a different gait assessment tool to evaluate. The authors must at least have believed that the tool they used and its contents could adequately represent gait.

Where repeatability of existing gait assessment tools was tested, experts in gait assessment were often used and their ability to correctly and repeatably identify events in complex gait pathologies is unlikely to reflect the ability of a broader and less specialised user population. Clinician performance in using an observation based gait assessment tool will depend upon experience, training, and underlying knowledge of gait, which will clearly
<table>
<thead>
<tr>
<th>Name of Tool</th>
<th>Criterion Validity</th>
<th>Intra(^1)/Inter(^2) Rater Repeatability</th>
<th>Sensitivity/Specificity</th>
<th>Number of Subjects</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Naked Eye Tools</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physician Rating Scale (PRS) (Koman et al, 1994)</td>
<td>not published</td>
<td>modest (Kappa 0.67(^2), rater n = 2 (Corry, 1995))</td>
<td>not published</td>
<td>12 children with cerebral palsy</td>
</tr>
<tr>
<td>HWGLS (modified version of PRS) (Pirpiris et al, 2001)</td>
<td>modest (Kappa 0.46–0.61) compared to instrumented kinematic gait analysis</td>
<td>modest (Kappa 0.46(^2), rater n = 4)</td>
<td>sensitive to change* found over-estimation of treatment effect compared to instrumented gait analysis</td>
<td>25 children with cerebral palsy</td>
</tr>
<tr>
<td><strong>Observational Gait Scale</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scale (OGS) (modified version of PRS) (Mackey et al, 2003)</td>
<td>modest (Kappa 0.69) compared to instrumented kinematic gait analysis</td>
<td>modest (Kappa 0.69(^1), rater n = 2)</td>
<td>not published</td>
<td>20 children with cerebral palsy</td>
</tr>
<tr>
<td>Rivermead Visual Gait Assessment (Lord et al, 1998a and b)</td>
<td>significant* ((r = 0.77)), compared to temporal-distance measures, balance, Rivermead Mobility Index (RMI)</td>
<td>modest (Kappa 0.58(^2), rater n = 2)</td>
<td>significant improvement* ((p &lt; 0.01)) compared to temporal-distance measures, balance, Rivermead Mobility Index (RMI)</td>
<td>10 adults with Multiple Sclerosis</td>
</tr>
<tr>
<td>Benesh Movement Notation (BMN) (Harrison et al, 1992)</td>
<td>not published</td>
<td>very good (&gt;98%)(^1), rater n = 11</td>
<td>not published</td>
<td>one able-bodied adult</td>
</tr>
<tr>
<td>Waterloo Gait Profile (Winter 1985)</td>
<td>valid*, compared to instrumented kinematic gait analysis. No data given</td>
<td>not published</td>
<td>not published</td>
<td>not applicable</td>
</tr>
</tbody>
</table>
| Rancho Los Amigos System (1970–80) (Rancho Los Amigos Medical Center) | not published | not published | not published | not published | (continued)
<table>
<thead>
<tr>
<th>Name of Tool</th>
<th>Criterion Validity</th>
<th>Intra(^1)/Inter(^2) Rater Repeatability</th>
<th>Sensitivity/Specificity</th>
<th>Number of Subjects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Video-Based Tools</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Edinburgh Visual Gait Score</td>
<td>modest (64% agreement) compared to instrumented kinematic gait analysis</td>
<td>good(^*) (LSD 3.20 points)(^1), rater (n = 5), good (70%)(^2), rater (n = 5)</td>
<td>sensitivity to change(^*), but not compared to other measures, no statistical data given.</td>
<td>4 children with cerebral palsy 1 able-bodied child</td>
</tr>
<tr>
<td>(Read et al, 2003)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Segment Rotation Indicators</td>
<td>very good (3.8(^\circ) max disagreement compared to Instrumented gait analysis)</td>
<td>very good (mean 0.3(^\circ) variation between raters)(^2), rater (n = 8)</td>
<td>not published</td>
<td>1 able-bodied subject</td>
</tr>
<tr>
<td>(Hillman et al., 1998)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Goniometers</td>
<td>Valid(^*) (ICC 0.84−0.99) compared to 2-D kinematic gait analysis</td>
<td>not published</td>
<td>not published</td>
<td>10 children with cerebral 9 able-bodied children</td>
</tr>
<tr>
<td>(Stuberg et al, 1988)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hemiplegic Gait Analysis Form</td>
<td>modest ((r = 0.60)), compared to temporal-distance measures</td>
<td>low−very high (Kendall’s 0.39−0.97)(^1) rater (n = 3), high (Kendall’s 0.71−0.90)(^2), rater (n = 3)</td>
<td>not published</td>
<td>6 adults with stroke</td>
</tr>
<tr>
<td>(Hughes and Bell, 1994)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rank Scoring System</td>
<td>not published</td>
<td>good(^*) (90%)(^2), rater (n = 3) (Hainsworth et al, 1997)</td>
<td>not published</td>
<td>12 children with cerebral palsy</td>
</tr>
<tr>
<td>(Watt et al, 1896)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\(^*\)Direct quotation from publication.
differ in different clinical settings. Basing repeatability ratings for a gait assessment tool on assessment of expert clinicians is contrary to the aims of developing clinically focussed gait assessment tools, namely increasing the use of gait assessment in the wider population of clinicians managing gait disorders. Good inter-rater repeatability is essential for the sharing of clinical information between clinicians and for multi-center research, although even the gold standard of quantitative gait analysis has been shown to have questionable inter-rater repeatability (Skaggs et al., 2000). Intra-rater repeatability of the tools has been assessed less frequently than inter-rater repeatability. An inherent problem in assessing intra-rater repeatability is the fact that the second or third assessment of the same gait pattern may be influenced by prior knowledge and experience gained in the preceding assessments. Adequate time lapse between assessments and randomising the order of assessments should reduce the effect of this.

Sensitivity and specificity of the observational gait assessment tools were not explicitly assessed in any of the tools we reviewed, with the exception of the Edinburgh gait score (Read et al., 2003), but attempts were made to assess whether a tool could detect changes in gait after treatment, and the magnitude of that change. As with validity testing, properly assessing sensitivity and specificity requires quantitative measures of gait, so that false positives detection or false negative detection of gait abnormalities or changes in gait can be identified.

The methods by which observation based gait assessment tools are assessed must be considered in the context of the patient population and the nature of the gait pathologies studied. A tool that is reliable for one patient group may not be reliable for another group, due to differences in the magnitude of gait pathology between patient groups and the fact that some pathologies and changes in gait are easier to distinguish than others. For example, in cerebral palsy gait, we might expect an assessment of whether initial contact is by toe strike or heel strike to be very reliable since it is relatively easy to distinguish these two gross characteristics, whereas detection of more subtle differences in gait, such as a change in the magnitude of various movement at the knee, are likely to be less reliable. Information regarding the precise clinical population and the gait pathologies of the patients was rarely sufficiently detailed in the literature we reviewed and thus the context for future use of the tools we reviewed remains unclear. The credibility of the testing of the validity, repeatability, sensitivity, and specificity qualities of a gait assessment tool also is dependent on the sample size used, especially when the patients suffer from complex pathologies, as is the case with cerebral palsy and adult neurological disorders. Such patient groups tend to demonstrate great variability in gait patterns and therefore a wide range of patients within specific pathologies need to be analysed. The sample sizes used in the studies we have reviewed were frequently very small, as low as 1–5 subjects, and our confidence in the results of such studies should also be low.

Importantly, some fundamental work is missing in the available literature. The basic relationship between what we can observe and the actual events of gait has not been evaluated. Most gait assessment tools lack evidence for any relationship to the gold standard quantitative gait assessment. If we investigate this relationship, we could develop valid categories of joint positions and joint motion based on categories of patients that have been identified through analysis of quantitative gait kinematics data. This gold standard data should be the starting point for the bottom up development of an observation based gait assessment tool, not something to which we compare an assessment tool based on theoretical categories. If we had valid categories based on prior quantitative data, then testing inter- and intra-rater repeatability would have greater value, and the quantitative data are essential for accurate assessment of sensitivity and specificity.

In summary two problems exist in the field of observation based gait assessment tools. First, there has been no systematic approach to developing and testing the current tools, and
existing tools appear to have grown out of clinical need and practical use, rather than a scientific scrutiny of pathological gait. Second, the little data that does exist regarding the quality of observation based gait assessment tools is of little value without details of the specific patient and clinician populations involved.

References


Corry IS 1995 The ‘Koman scale.’ Queen’s University of Belfast. Thesis/Dissertation


Hillman SJ, Hazlewood ME, Loudon IR, Robb JE 1998 Can transverse plane rotations be estimated from video gait analysis? Gait and Posture 8: 87–90


Rome K, Cowieson F 1996 A reliability study of the universal goniometer, fluid goniometer, and electrogoniometer...
for the measurement of ankle dorsiflexion. Foot and Ankle International 17: 28–32
Winter DA 1985 Concerning the scientific basis for the diagnosis of pathological gait and for rehabilitation protocols. Physiotherapy Canada 37: 245–252