

The LLRS–AIM Limb Deformity Evaluation: A Pilot Study to 10:51 Evaluate Reliability

James J. McCarthy, MD; Sanjeev Sabharwal, MD, Christopher Iobst, MD; Viral Jain, MD; Nate Faulkner; S. Robert Rozbruch, MD; David Klimaski; Emily Eismann
Director, Division of Pediatric Orthopaedic Surgery
Cincinnati Children's Hospital Medical Center
3333 Burnet Avenue–MLC# 2017
Cincinnati, OH 45229
james.mccarthy@cchmc.org; 513–636–4787 telephone; 513–636–3928 facsimile

What was the question?

Can we develop a general limb deformity classification scheme that allows for the pre treatment assessment of a broad range of lower extremity disorders and promotes a uniform method of evaluation?

How did you answer the question?

An outline of a score was developed by first evaluating several other scoring scales and creating a system that can be used for each limb and incorporates standard evaluation techniques. The LLRS AIM classification is a 7 point scale that takes into account deformity, as well as issues related to soft tissue injury, bone quality and underlying health conditions. This was evaluated by 8 surgeons, who rated 10 theoretical patients. They first ranked the patients from simplest to most difficult case, then used the LLRS AIM scale to rate the case. Two or more weeks later, without reference to the first rating, they rated the cases again. Inter-rater reliability was analyzed by a two-way random ANOVA to determine the intraclass correlation between raters ($ICC_{2,k}$). Intra-rater reliability was evaluated by a one-way ANOVA to determine the intraclass correlation over time ($ICC_{1,k}$). Significant differences were analyzed with paired t-tests and a two-way repeated measures ANOVA by rater and trial. Rank scores were assessed for interrater reliability with Kendall's W (coefficient of concordance). The relationship between the LLRS scores and rankings were evaluated through linear regressions.

What are the results?

The LLRS displayed excellent *intrarater* reliability, with highly consistent ratings *over time* (overall $ICC=0.94$, individual rater ICC 's ranged from 0.89 to 1.00). LLRS ratings did not significantly differ over time ($p>0.05$), and on average, raters gave the same exact score for both trials 41% of the time. The LLRS also displayed excellent *interrater* reliability, with highly consistent ratings *between raters* for trial 1 ($ICC=0.97$) and trial 2 ($ICC=0.98$). The LLRS ratings also did not significantly differ between raters ($p=0.86$), over time ($p=0.80$), or between raters over time ($p=0.30$). There was little agreement between raters on their patient rankings ($W=0.023$, $p=0.98$). However, patient rankings were significantly positively correlated with the LLRS–AIMS scores for trial 1 ($R^2 = 0.253$, $p < 0.001$) and trial 2 ($R^2 = 0.229$, $p < 0.001$) both before and after controlling for rater.

What are your conclusions?

This suggests that the LLRS has potential as both a reliable and valid clinical tool.