External validation of the Chicago Chiari Outcome Scale

Clinical article

CHESTER K. YARBROUGH, M.D., JACOB K. GREENBERG, B.A., MATTHEW D. SMYTH, M.D., JEFFREY R. LEONARD, M.D., TAE SUNG PARK, M.D., AND DAVID D. LIMBRICK JR., M.D., PH.D.

Department of Neurosurgery, St. Louis Children’s Hospital, Washington University School of Medicine, St. Louis, Missouri

Object. Historically, assessment of clinical outcomes following surgical management of Chiari malformation Type I (CM-I) has been challenging due to the lack of a validated instrument for widespread use. The Chicago Chiari Outcome Scale (CCOS) is a novel system intended to provide a less subjective evaluation of outcomes for patients with CM-I. The goal of this study was to externally validate the performance of the CCOS.

Methods. Patients undergoing surgery for CM-I between 2001 and 2012 were reviewed (n = 292). Inclusion criteria for this study were as follows: 1) patients receiving primary posterior fossa decompression; 2) at least 5.5 months of postoperative clinical follow-up; and 3) patients ≤ 18 years of age at the time of surgery. Outcomes were evaluated using the CCOS, along with a “gestalt” impression of whether patients experienced significant improvement after surgery. A subgroup of 118 consecutive patients undergoing operations between 2008 and 2010 was selected for analysis of interrater reliability (n = 73 meeting inclusion/exclusion criteria). In this subgroup, gestalt and CCOS scores were independently determined by 2 reviewers, and interrater reliability was assessed using the intraclass correlation coefficient (ICC) and kappa (κ) statistic.

Results. The median CCOS score was 14, and 67% of patients had improved gestalt scores after surgery. Overall, the CCOS was effective at identifying patients with improved outcome after surgery (area under curve = 0.951). The interrater reliability of the CCOS (ICC = 0.71) was high, although the reliability of the component scores ranged from poor to good (ICC 0.23–0.89). The functionality subscore demonstrated a low ICC and did not add to the predictive ability of the logistic regression model (likelihood ratio = 1.8, p = 0.18). When analyzing gestalt outcome, there was moderate agreement between raters (κ = 0.56).

Conclusions. In this external validation study, the CCOS was effective at identifying patients with improved outcomes and proved more reliable than the authors’ gestalt impression of outcome. However, certain component subscores (functionality and nonpain symptoms) were found to be less reliable, and may benefit from further definition in score assignment. In particular, the functionality subscore does not add to the predictive ability of the CCOS, and may be unnecessary. Overall, the authors found the CCOS to be an improvement over the previously used assessment of outcome at their institution.

KeY WORdS • Chiari malformation • posterior fossa decompression • outcome assessment

Abbreviations used in this paper: CCOS = Chicago Chiari Outcome Scale; CM-I = Chiari malformation Type I; ICC = intraclass correlation coefficient; ROC = receiver operating characteristic; SLCH = St. Louis Children’s Hospital.

Chiari malformation Type I (CM-I) is a congenital deformity characterized by herniation of the cerebellar tonsils through the foramen magnum, and it results in a variety of signs and symptoms, ranging from headache to brainstem compromise. Management of CM-I, both operative and nonoperative, comprises a meaningful portion of any pediatric neurosurgery practice. Many treatment strategies have been described, although the optimal treatment remains unclear. A full discussion of the various operative techniques is outside the scope of this study.

Assessment of patient outcomes is complicated by a deficiency of validated outcome measures specific to CM-I. One group has published on patient-reported outcome measures using a patient questionnaire, although the length of the questionnaires involved probably precludes use in the pediatric population. Other groups have published case series with outcomes split into “better, unchanged, or worse” categories. A similar as-
essment algorithm has been in use at our institution, although this classification is highly subjective and is poorly suited for assessing outcomes within or between individuals. More recently, the Chicago Chiari Outcome Scale (CCOS) was developed to address outcomes with a more complete approach to the variety of symptoms patients may experience (Table 1). This scale consists of 4 categories of postoperative outcome (pain, nonpain symptoms, functionality, and complications) that are graded 1 to 4, for a total possible score of 16. The CCOS was retrospectively applied to our patients, and a more detailed analysis of predictors of outcome was performed. This study constitutes the first validation of the CCOS in an external pediatric neurosurgical population.

### Methods

Following approval from the Washington University Institutional Review Board, a retrospective review was performed by searching for all consecutive patients treated for CM-I at St. Louis Children’s Hospital (SLCH) over the 12-year period between 2001 and 2012. Inclusion criteria for this study were as follows: 1) patients receiving primary posterior fossa decompression at SLCH between 2001 and 2012; 2) at least 5.5 months of postoperative clinical follow-up; and 3) patients ≤ 18 years of age at the time of surgery. A cutoff of 5.5 months of follow-up was chosen to ensure the availability of longer-term data while also capturing patients whose 6-month appointment occurred shortly before 6 months postoperatively. Exclusion criteria were as follows: 1) previous surgical treatment for CM-I prior to operative treatment at SLCH; and 2) insufficient chart information to evaluate clinical outcome.

A single reviewer analyzed the charts of all patients meeting inclusion criteria. Clinical outcomes were evaluated using the recently published CCOS. As in the initial study describing CCOS, each patient was assigned a gestalt score based on an overall impression of whether he or she benefited significantly from surgery. When being assigned gestalt scores, patients were considered “improved” if all major symptoms improved to a significant degree or resolved entirely, and overall quality of life improved. If minor symptoms persisted, a patient was still considered improved. A patient was considered “unchanged” if the following features were present: some or all major symptoms failed to improve to a significant degree after surgery; new CM-I–related symptoms developed after surgery that caused significant impairment; or a patient required repeat surgery for CM-I–related symptoms or syringomyelia, regardless of final outcome. A patient was considered “worse” if overall quality of life decreased as a result of increased severity of major preoperative symptoms, or if new CM-I–related symptoms developed postoperatively. When analyzing outcomes for headache symptoms, only occipital headaches and headache types present preoperatively were considered. Both CCOS and gestalt outcome scores were assigned based on each patient’s most recent clinical follow-up visit.

To evaluate interrater reliability, a second reviewer independently analyzed the charts of 73 consecutive patients treated between 2008 and 2010 and assigned CCOS and gestalt scores. Both reviewers adhered to the guidelines for the CCOS as originally published. After all score assignments were made, cases with significantly differing CCOS scores were reviewed to identify potential points of ambiguity in the system.

### Statistical Analysis

The ability of the CCOS to distinguish patients with gestalt improved outcome was investigated by examining the area under the receiver operating characteristic (ROC) curve. For this analysis, gestalt improved outcome was used as the gold standard for the disease state. The area under the ROC curve represents the probability that a randomly chosen patient will be correctly assigned to the appropriate outcome category. This area ranges from 0.5 (the test being evaluated has no accuracy) to 1.0 (perfect accuracy). The coordinates of the curve were used to find the CCOS score with optimal sensitivity and specificity for identifying patients with good outcome. To determine which components of the CCOS score were most predictive of gestalt outcome, a logistic regression analysis was done, including the 4 component scores of the CCOS as predictor variables. For both the ROC and logistic regression analyses, “unchanged” and “worse” outcomes were grouped together. Intraclass correlation coefficients (ICCs) and kappa statistics were calculated to test the interrater reliability of the CCOS and gestalt impression scores, respectively. All statistical analyses were done using SPSS version 21 (IBM, Inc.). Probability values < 0.05 were considered significant.

### Table 1: Details of the CCOS

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>pain</td>
<td>worse</td>
</tr>
<tr>
<td>nonpain symptoms</td>
<td>worse</td>
</tr>
<tr>
<td>functionality</td>
<td>unable to attend work or school</td>
</tr>
<tr>
<td></td>
<td>persistent complication, poorly controlled</td>
</tr>
</tbody>
</table>

* As reported in Aliaga et al. meds = medications.
Validation of the Chicago Chiari Outcome Scale

Results

Patient Characteristics

Of the 292 consecutive patients who received primary CM-I decompression between 2001 and 2012 at SLCH (Table 2), 215 were available for analysis. A subgroup of 118 consecutive patients undergoing operations between 2008 and 2010 was selected for analysis of interrater reliability, with 73 patients meeting all inclusion and exclusion criteria. This period was chosen to maximize the availability of relevant medical record data and to provide for adequate follow-up with consecutive patients. Our institution underwent a migration to the current electronic medical records system prior to 2008. In addition to the interrater analysis, all patients were evaluated by one reviewer using both the gestalt outcome and CCOS.

Patient Outcome

Among the 215 included patients, the median CCOS score was 14, and 67% had an improved gestalt outcome. The mean follow-up was 37 ± 29 months. Comparison of the CCOS score and gestalt outcome revealed a correlation between good gestalt outcome and high CCOS score (Fig. 1). All patients with a CCOS score of 15 or greater had improved gestalt outcome, whereas no patient with a CCOS score of 12 or lower had improved outcome.

The ability of the CCOS to distinguish patients with good gestalt outcome was plotted using an ROC curve (Fig. 2). The ROC plots sensitivity against 1-specificity, and shows the probability that a randomly chosen patient would be assigned the correct gestalt outcome based on CCOS score. The area under the ROC curve was 0.951, suggesting excellent prediction of gestalt outcome. Sensitivities and specificities for the CCOS were also calculated to determine the optimal CCOS value to use as a cutoff score (Table 3). These results showed that a CCOS value of 14 has excellent sensitivity (0.95) and good specificity (0.746) for identifying patients with good gestalt outcome.

To test whether certain components of the CCOS have a stronger relationship than others with gestalt outcome, logistic regression for each subscore was performed (Table 4). The results of this analysis showed increased likelihood of improved gestalt outcome with higher CCOS subscores for all domains except functionality. Although the confidence intervals for the odds ratios were wide, the pain and nonpain component scores appeared to have a particularly large impact on the probability of a positive surgical outcome compared with the complications score. Given that the functionality subscore was not a significant predictor of gestalt outcome, we also tested a regression model that excluded this component of the CCOS. The likelihood ratio statistic was 1.8 (p = 0.18), indicating that the inclusion of the functionality component did not significantly improve the overall predictive ability of the model.

Reliability Assessment of CCOS

All patients who underwent Chiari decompression

![Fig. 1. Bar graph showing a comparison of CCOS scores and gestalt outcomes. A clear alignment of higher CCOS score with better gestalt outcome is shown, with all patients scoring 15 or 16 on the CCOS having improved gestalt outcome. Conversely, all except 1 patient with a CCOS score of 10 or lower had worse than preoperative status on gestalt assessment.](image1)

![Fig. 2. Graph showing the ROC for CCOS. The ROC had an area under the curve of 0.951, showing excellent fit of the CCOS to predict outcome on the gestalt assessment.](image2)
were calculated, showing moderate to excellent agreement between 2008 and 2010 and met inclusion criteria were analyzed by independent reviewers (Table 5). The ICCs for complications (0.89), pain (0.63), and nonpain signs and symptoms (0.58). The ICC for the functionality score (0.23) was poor. A review of cases in disagreement showed different interpretations of the functionality score guidelines, as well as inconsistent results due to limited record documentation. Overall, the 2 reviewers obtained identical total CCOS scores for 43% of patients. In the patients with differing CCOS scores, the mean difference between reviewers was 1.4 (range 1–3). The kappa statistic for gestalt outcome assessment was 0.56, showing moderate agreement. However, after identifying the functionality score as having lower interrater reliability, the average composite CCOS difference was 1.4, with the most common disagreement being 1 point. This difference could result from retrospective bias, because the clinical charts were not designed to capture fully the CCOS. However, despite some inconsistencies, the composite CCOS score showed good interrater agreement relative to gestalt outcome. Furthermore, our logistic regression analysis of the CCOS showed that each subcomponent of the scale except for functionality had a strong impact on the likelihood for an improved outcome (Table 4).

In our statistical assessment of the CCOS, there was a clear correlation between higher CCOS score and gestalt outcome (Fig. 1). Additionally, 2 independent raters showed moderate to good agreement in composite score and all subscores of the CCOS except the functionality score (Table 5). In cases in which we observed disagreement, the average composite CCOS difference was 1.4, with the most common disagreement being 1 point. This difference could result from retrospective bias, because the clinical charts were not designed to capture fully the CCOS. However, despite some inconsistencies, the composite CCOS score showed good interrater agreement relative to gestalt outcome. Furthermore, our logistic regression analysis of the CCOS showed that each subcomponent of the scale except for functionality had a strong impact on the likelihood for an improved outcome (Table 4).

Discussion

Chiari malformation Type I is a common referring diagnosis to pediatric neurosurgical practices. Assessment of patient outcomes is complicated by lack of validated, quantifiable outcome measures. A gestalt assessment algorithm of “better, unchanged, or worse” has been in use at our institution and others, although this assessment has obvious weaknesses in its failure to provide quantifiable results. Aliaga et al. developed the CCOS to address these weaknesses. Although the CCOS was applied to patients by its developers at their home institution, it has not been externally verified. This study constitutes the first validation of the CCOS by an external pediatric neurosurgical practice.

In our statistical assessment of the CCOS, there was a clear correlation between higher CCOS score and gestalt outcome (Fig. 1). Additionally, 2 independent raters showed moderate to good agreement in composite score and all subscores of the CCOS except the functionality score (Table 5). In cases in which we observed disagreement, the average composite CCOS difference was 1.4, with the most common disagreement being 1 point. This difference could result from retrospective bias, because the clinical charts were not designed to capture fully the CCOS. However, despite some inconsistencies, the composite CCOS score showed good interrater agreement relative to gestalt outcome. Furthermore, our logistic regression analysis of the CCOS showed that each subcomponent of the scale except for functionality had a strong impact on the likelihood for an improved outcome (Table 4).

Based on the observed inconsistency in the assignment of the functionality subscore and its uncertain impact on the relationship between CCOS score and gestalt outcome, we examined the scoring methods and contribution of this subscore to the composite CCOS score in detail. Indeed, we found some ambiguity in scoring functionality; specifically it can be difficult to distinguish between subscores of 2 (able to work or go to school < 50%) and 3 (able to work or go to school > 50%), and between subscores of 3 (able to work or go to school > 50%) and 4 (fully functional). We found that many patients have some minor symptoms that may or may not result from CM-I, but cannot be ruled out based on clinical examination or from the documentation available in the clinical chart. The subscore of 3, in particular, comprises a large group of potential patients who are not fully symptom free, but...
Validation of the Chicago Chiari Outcome Scale

are not completely debilitated either. Delving deeper, we found that the functionality subscore did not contribute significantly to the predictive ability of the logistic regression model; thus, our findings indicate that the CCOS may be improved by clarifying the definition and scoring of the functionality subscore or perhaps by removing it altogether.

An additional area of ambiguity within the CCOS is the headache subscore. Headaches are common in the general population, with yearly prevalence greater than 50% in adolescents and children.16 Thus, a large portion of these patients will ultimately have clinical courses complicated by a headache syndrome unrelated to CM-I. To control for this fact in our patient population, we limited recurrent headache syndromes to occipital headaches, posttussive or exertional headaches, or headaches similar to that which the patient suffered at initial presentation. Although we think that this is a rational limitation, it invites some ambiguity by introducing the examiner’s interpretation of what is truly related to CM-I. This factor may be inescapable, because headaches are subjective experiences. Nonetheless, we believe that some attempt to focus the evaluation of headache symptoms is essential.

The CCOS is a strong step forward in providing an outcomes assessment tool for treatment of patients with CM-I. However, it does not fully remove subjectivity from the clinical assessment, because different interpretations are possible in at least the headache and functionality subscores. Additionally, we believe that there is some difficulty in adequately separating patients with severe symptoms who have minor improvements and patients who suffer complications but also show symptomatic improvement. There are relatively few patients with subscores of “1” in any category, and scores for “improved” (median CCOS Score 15) and “unchanged or worsened” (median CCOS Score 13) gestalt outcomes show significant clustering at the higher end of the CCOS. Thus, although we are able to validate that the CCOS adequately predicts gestalt outcome, additional refinements may be necessary to overcome the challenges described above.

Conclusions

In this paper we have evaluated the CCOS at our institution and found it to be more consistently applied between examiners than the previously used gestalt assessment. This score is clearly an improvement over the status quo, and should allow for more subtle analysis of a large and challenging patient population. As currently defined, the functionality subscore allows for more ambiguity in interpretation than the other subscores without augmenting the CCOS’s ability to estimate outcomes. Overall, the CCOS represents a major advance in assessing clinical outcomes in patients with CM-I, but this tool may benefit from further refinement, including modifying or eliminating the functionality subscore.

Acknowledgment

We acknowledge Michael Wallendorf, Ph.D., for his suggestions regarding statistical analysis.

Disclosure

This study was supported in part by the Park-Reeves Syringomyelia Research Consortium, and by the Washington University Institute of Clinical and Translational Sciences grants UL1 TR000448 and TL1 TR000449 from the National Center for Advancing Translational Sciences. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. The authors report no conflict of interest concerning the materials or methods used in this study or the findings specified in this paper.

Author contributions to the study and manuscript preparation include the following. Conception and design: Yarborough, Greenberg, Limbrick. Acquisition of data: Yarborough, Greenberg. Analysis and interpretation of data: Yarborough, Greenberg. Drafting the article: Yarborough, Greenberg. Critically revising the article: all authors. Reviewed submitted version of manuscript: all authors. Approved the final version of the manuscript on behalf of all authors: Yarborough. Study supervision: Limbrick.

References


Manuscript submitted September 30, 2013. Accepted March 10, 2014. Please include this information when citing this paper: published online April 11, 2014; DOI: 10.3171/2014.3.PEDS13503. Address correspondence to: Chester K. Yarbrough, M.D., Department of Neurosurgery, Washington University School of Medicine, 660 S. Euclid Ave., Campus Box 8057, St. Louis, MO 63110. email: yarbroughc@wudosis.wustl.edu.