Web Appendix

1. Data Extraction

Each retrieved citation was reviewed by two independently working reviewers. Most of the articles were excluded on the basis of information provided by the title or abstract. Citations that appeared to be appropriate or those that could not be excluded unequivocally from the title and abstract were identified, and the corresponding full text reports were reviewed by the two reviewers. Any disagreement between them was resolved by reviewer consensus. From the included articles, the following data were extracted: patient demographics, preexisting diagnosis, instability, treatment, follow-up, fusion rate, symptoms, and change in symptoms.

2. Study Quality

Articles selected for inclusion were classified by class of evidence. The method used for assessing the quality of evidence of individual studies as well as the overall quality of the body of evidence incorporates aspects of the rating scheme developed by the Oxford Centre for Evidence-based Medicine[1] and used with modification by *The Journal of Bone and Joint Surgery American Volume* (*J Bone Joint Surg Am*),[2] precepts outlined by the Grades of Recommendation Assessment, Development and Evaluation (GRADE) Working Group[3] and recommendations made by the Agency for Healthcare Research and Quality (AHRQ).[4] Each individual study was rated by two different investigators against preset criteria that resulted in an evidence rating (Class of Evidence I, II, III, or IV). Disagreements were resolved through discussion.

Determination of Overall Strength of Evidence

After individual article evaluation, the overall body of evidence with respect to each outcome is determined based on precepts outlined by the Grades of Recommendation Assessment, Development and Evaluation (GRADE) Working Group[1] and recommendations made by the Agency for Healthcare Research and Quality (AHRQ).[5] Qualitative analysis is performed considering the following AHRQ required and

Table 1 Summary of inclusion and exclusion criteria

	Inclusion	Exclusion
Patient	• Adult patients with Chiari malformation with syringomyelia	• Pediatric patients • Trauma • Meningitis • Tumor • Hemorrhage • Arachnoiditis • Chiari w/o syringomyelia
Intervention	• Posterior fossa decompression only	Other decompression: • Foramen magnum • Hindbrain • Cervicomedullary • Suboccipital • Craniovertebral Any other treatment with or without posterior fossa decompression
Outcome	• Recurrent syringomyelia • Residual syringomyelia • Worsening of syringomyelia	
Study design	• Randomized controlled trials • Cohort studies • Case-series with N $\geq$ 10	• Case reports • Historical controls with surgery happening at different point in time • Studies with N $<$ 10 • Non-human • Cadaver • Biomechanical

Table 2 Critical appraisal for studies on residual or recurrent syringomyelia after posterior fossa decompression

Methodological principle	Zhang (2011)	Alfieri (2012)	Batzdorf (2013)	Depreitere (2000)	Ellenbogen (2000)	Fischer (1995)	Garcia-Uria (1981)	Mueller (2005)	Noudel (2011)	Silva (2010)	Vaquero (1990)
Study design											
Prospective cohort study											
Retrospective cohort study											
Case-control study											
Cross-sectional study											
Case-series	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Case-series											
Patients at similar point in the course of their disease or treatment	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Complete follow-up of $\geq$ 80%		✓		✓			✓				
Patients followed long enough for outcomes to occur	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Accounting for other prognostic factors*											
Evidence class	IV	IV	IV	IV	IV	IV	IV	IV	IV	IV	IV

*Authors must consider other factors that might influence patient outcomes

Table 3 Criteria for class of evidence for prognostic studies

		Studies of prognosis	
Class	**Risk of bias**	**Study design**	**Criteria**
I	**Low risk** Study adheres to commonly held tenets of high quality design, execution and avoidance of bias	Good quality cohort[a]	• Prospective design • Patients at similar point in the course of their disease or treatment • F/U rate of $\geq$ 80%[b] • Patients followed long enough for outcomes to occur • Accounting for other prognostic factors[c]
II	**Moderately low risk** Study has potential for some bias; does not meet all criteria for class I but deficiencies not likely to invalidate results or introduce significant bias	Moderate quality cohort	• Prospective design, with violation of one of the other criteria for good quality cohort study • Retrospective design, meeting all the rest of the criteria in class I
III	**Moderately high risk** Study has flaws in design and/or execution that increase potential for bias that may invalidate study results	Poor quality cohort Good quality case-control or cross-sectional study	• Prospective design with violation of 2 or more criteria for good quality cohort, or • Retrospective design with violation of 1 or more criteria for good quality cohort • A good case-control study[d] • A good cross-sectional study[e]
IV	**High risk** Study has significant potential for bias; does not include design features geared toward minimizing bias and/or does not have a comparison group	Poor quality case-control or cross-sectional case-series[d]	• Other than a good case-control study • Other than a good cross-sectional study • Any case-series[f] design

[a]Cohort studies follow individuals with the exposure of interest over time and monitor for occurrence of the outcome of interest.
[b]Applies to cohort studies only.
[c]Authors must consider other factors that might influence patient outcomes and should control for them if appropriate.
[d]A good case-control study must have the all of the following: all incident cases from the defined population over a specified time period, controls that represent the population from which the cases come, exposure that precedes an outcome of interest, and accounting for other prognostic factors.
[e]A good cross-sectional study must have all of the following: a representative sample of the population of interest, an exposure that precedes an outcome of interest (e.g., sex, genetic factor), an accounting for other prognostic factors, and for surveys, at least a 80% return rate.
[f]A case-series design for prognosis is one where all the patients in the study have the exposure of interest. Since all the patients have the exposure, risks of an outcome can be calculated only for those with the exposure, but cannot be compared with those who do not have the exposure. For example, a case-series evaluating the effect of smoking on spine fusion that only recruits patients who smoke can simply provide the risk of patients who smoke that result in pseudarthrosis but cannot compare this risk to those that do not smoke.

additional domains.[4] ►**Table 4** provides an outline of the method used to determine the final SoE.

- *Risk of bias* is evaluated during the individual study evaluation described above. After individual article review, the literature evidence was rated as "HIGH" initially if the majority of the articles are Level I or II. It is rated as "LOW" if the majority were level III or lower. This is the "baseline" strength of evidence, ►**Table 5**. The consistency, directness, precision, and subgroup effects are considered for potential "downgrading" the strength of the body of evidence (one or two levels depending on the degree and number of domain violations).

Criteria Evaluated for "Downgrading"

- *Consistency* refers to the degree of similarity in the effect sizes of different studies within an evidence base. If effect sizes indicate the same direction of effect and if the range of effect sizes is narrow, an evidence base was judged to be consistent. If meta-analyses were conducted, we evaluated the consistency with an "eye ball test." This test consists of a visual appraisal of the forest plots by two independent

Table 4 Methodology outline for determining overall strength of evidence (SoE):

All AHRQ "required" and "additional" domains[a] are assessed. Only those that influence the baseline grade are listed in table. Baseline strength: Risk of bias (including control of confounding) is accounted for in the individual article evaluations. HIGH = majority of articles Level I/II. LOW = majority of articles Level III/IV. DOWNGRADE: Inconsistency[b] of results (1 or 2); Indirectness of evidence (1 or 2); Imprecision of effect estimates (1 or 2); Sub-group analyses not stated *apriori* and no test for interaction (2) UPGRADE: Large magnitude of effect (1 or 2); Dose response gradient (1)					
Outcome	**Strength of evidence**	**Conclusions and comments**	**Baseline**	**DOWGRADE**	**UPGRADE**
Outcome	HIGH	Summary of findings	HIGH Level I/II studies	NO consistent, direct, and precise estimates	NO
Outcome	MODERATE	Summary of findings	LOW Level III studies	NO consistent, direct, and precise estimates	YES Large effect
Outcome	LOW	Summary of findings	HIGH Level I/II studies	YES (2) Inconsistent Indirect	NO

[a]Required domains: risk of bias, consistency, directness, precision. Plausible confounding that would decrease observed effect is accounted for in our baseline risk of bias assessment through individual article evaluation. Additional domains: dose–response, strength of association, publication bias.
[b]Single study = "consistency unknown".

Table 5 Evidence summary

Baseline quality: HIGH = majority of articles Level I/II. LOW = majority of articles Level III/IV. UPGRADE: Large magnitude of effect (1 or 2 levels); dose response gradient (1 level) DOWNGRADE: Inconsistency of results (1 or 2 levels); indirectness of evidence (1 or 2 levels); imprecision of effect estimates (1 or 2 levels)					
Outcomes	**Strength of evidence**	**Conclusions/comments**	**Baseline**	**UPGRADE (levels)**	**DOWN-GRADE (levels)**
Question 1: What is the average rate of recurrent or residual syringomyelia following posterior fossa decompression as a result of Chiari malformation with associated syringomyelia?					
Recurrent/residual syringomyelia	INSUFFICIENT	Rates of recurrent/residual syringomyelia after posterior fossa decompression in adults range from 0 to 22% with an average across studies of 6.7%. These studies are case series from different populations. Due to the low quality of individual studies and the inconsistency between studies, there is insufficient evidence to establish an expected rate of recurrence; however, we can provide surgeons and patients with a range of estimates to consider.	• LOW	• NO	• Inconsistency (1)
Question 2: What treatment methods have been reported in the literature for managing recurrent or residual syringomyelia after initial posterior fossa decompression?					
Treatments	Not applicable	This was a descriptive key question and therefore an overall strength of evidence is not applicable.	• No studies reported on the management of recurrent/residual syringomyelia after posterior fossa decompression.		

reviewers. Single study evidence bases were judged "consistency unknown (single study)" and downgraded.

- *Directness is* concerned with whether the evidence being assessed reflected a single, direct link between the interventions of interest and the ultimate health outcome; that is, a determination of whether the most clinically relevant outcome was measured or if a surrogate outcome was assessed. Directness also applies to indirect comparisons

of treatment when head to head comparisons of interest could not made within individual studies.

- *Precision* of evidence pertains to the degree of certainty surrounding an estimate of effect for a specific outcome. This is based on whether the estimate of effect reached statistical significance and/or the inspection of confidence intervals around effect estimates. When there are only two subgroups, the overlap of the confidence intervals of the summary estimates of the two groups is considered. No overlap of the confidence intervals indicates statistical significance, but the confidence intervals can overlap to a small degree and the difference still is statistically significant.
- *Subgroup effects*. For evaluating subgroup effects (i.e., heterogeneity of treatment effects), we downgrade if the authors do not state *a priori* their plan to perform subgroup analyses and if there was no test for interaction.

Criteria Used for "Upgrading"

- Finally, if the strength of evidence is less than "HIGH," we "upgrade" the evidence if there is a dose–response association or a strong magnitude of effect.

Strength of Evidence for Existing Systematic Reviews

Level of evidence ratings for Cochrane reviews and other systematic reviews are assigned a baseline score of HIGH if RCTs were used, LOW if observational studies were used. The rating can be upgraded or downgraded based on adherence to the core criteria for methods, qualitative, and quantitative analyses for systematic reviews (there is a reference/evaluation table for this).

The following four possible levels and their definition are reported:

- **High**: High confidence that the evidence reflects the true effect. Further research is very unlikely to change our confidence in the estimate of effect.
- **Moderate**: Moderate confidence that the evidence reflects the true effect. Further research may change our confidence in the estimate of effect and may change the estimate.
- **Low**: Low confidence that the evidence reflects the true effect. Further research is likely to change the confidence in the estimate of effect and likely to change the estimate.
- **Insufficient**: Evidence either is unavailable or does not permit a conclusion.

References

1 Atkins D, Best D, Briss PA, et al; GRADE Working Group. Grading quality of evidence and strength of recommendations. BMJ 2004; 328(7454):1490
2 Wright JG, Swiontkowski MF, Heckman JD. Introducing levels of evidence to the journal. J Bone Joint Surg Am 2003;85-A(1):1–3
3 Balshem H, Helfand M, Schünemann HJ, et al. GRADE guidelines: 3. Rating the quality of evidence. J Clin Epidemiol 2011;64(4):401–406
4 Methods Guide for Effectiveness and Comparative Effectiveness Reviews. AHRQ Publication No. 10(12)-EHC063-EF. Rockville, MD: Agency for Healthcare Research and Quality; April 2012. Available at: www.effectivehealthcare.ahrq.gov
5 West S, King V, Carey TS, et al. Systems to Rate the Strength of Scientific Evidence. Evidence Report/Technology Assessment No. 47 (Prepared by the Research Triangle Institute-University of North Carolina Evidence-based Practice Center, Contract No. 290–97–0011). Rockville, MD: Agency for Healthcare Research and Quality; 2002