

Validation Design for a Patient-Centred Spinal Decompression Protocol: Delphi Consensus, Content Validity and Prospective Outcome Evaluation

Badal Biswas^{*1}, Dr. Jagat Bahadur Rawat^{*2}

^{*1}Research Scholar, Department of Clinical Research (Allied Health Sciences), Amaltas University, Dewas, Madhya Pradesh, India

^{*2}Associate Professor, Department of Clinical Research (Allied Health Sciences), Amaltas University, Dewas, Madhya Pradesh, India

Abstract

Clinical protocols for lumbar nerve root compression and radiculopathy are often adopted without formal evidence that their content is valid, that they can be followed in practice or that patients treated under them achieve meaningful improvement. This paper sets out a staged validation design for a patient-centred spinal decompression protocol. Stage one uses a modified Delphi process in which a multidisciplinary panel rates each protocol statement for relevance over two to three rounds, with consensus defined as at least 75% agreement. Stage two quantifies content validity with the item-level and scale-level content validity indices and a modified kappa that adjusts for chance agreement; for a panel of ten, an item is retained when at least eight experts rate it relevant. Stage three is a prospective single-arm pilot that records pain, disability, walking capacity, neurological status, quality of life, satisfaction and complications from baseline to 12 months. Analytical calculations show that detecting a mean improvement of 10 points on the Oswestry Disability Index with a standard deviation of change of 20 requires 34 patients at 80% power and 44 at 90%, whereas estimating the responder proportion to within ± 10 percentage points requires about 97. Stage four embeds the protocol in plan-do-study-act audit cycles. The design gives explicit decision rules for retaining, revising or rejecting protocol content; no patient data are reported.

Index Terms—Delphi method, content validity index, modified kappa, spinal decompression, lumbar radiculopathy, Oswestry Disability Index, sample size, quality improvement

1. Introduction

Decompression surgery for lumbar nerve root compression can relieve leg pain and improve function in selected patients [1], [2], but the benefit depends on how patients are selected, how care is delivered and how outcomes are followed. A structured, patient-centred protocol can make these steps consistent, yet a protocol is only useful if its content is accepted by the clinicians who must use it, if it can be applied in routine practice and if patients managed under it improve by amounts that matter to them.

Validation of a clinical protocol therefore has several layers. Content validity asks whether the protocol includes the right elements and excludes irrelevant ones. Feasibility asks whether the steps can be completed and documented. Outcome evaluation asks whether patients improve and whether harms are acceptable. Each layer needs its own method, its own threshold and its own decision rule. This paper, derived from a doctoral

thesis on a patient-centred decompression protocol, presents an integrated validation design. Its contributions are:

- 1) A four-stage validation framework that links Delphi consensus, content validity indices, a prospective pilot and quality-improvement cycles.
- 2) Quantitative retention thresholds by panel size based on chance-adjusted agreement.
- 3) An outcome set and assessment schedule built on validated patient-reported measures and minimal important change values.
- 4) Sample-size and precision calculations that make the trade-offs in pilot size explicit.

2. Related Work

2.1 Consensus Methods

The Delphi method collects expert judgement anonymously over several rounds with controlled feedback, which reduces the influence of dominant individuals. A systematic review of Delphi studies used to select health-care quality indicators found

wide variation in panel size, number of rounds and definitions of consensus, and recommended clearer reporting [3]. A further review found that consensus was defined in many different ways, most often as a percentage agreement, with 75% as the median threshold, and recommended that the definition be stated in advance [4].

2.2 Content Validity Indices

Content validity is commonly quantified by asking experts to rate each item on a four-point relevance scale. The item-level content validity index (I-CVI) is the proportion of experts rating an item 3 or 4 [5]. Because high agreement can arise by chance in small panels, Polit et al. recommended reporting a modified kappa that adjusts the I-CVI for chance agreement, and proposed that values of 0.74 or above be regarded as excellent [6].

2.3 Outcome Measurement in Lumbar Spine Care

The Oswestry Disability Index (ODI) is a widely used measure of back-related disability [7]. An international consensus proposed a minimal

important change of about 10 points, or 30% improvement from baseline, for the ODI and similar thresholds for pain scales [8]. Studies in surgical populations have reported comparable thresholds, though the value depends on the method used to calculate it [9]. The EQ-5D-5L provides a generic measure of health-related quality of life [10].

2.4 Quality Improvement

Plan–do–study–act (PDSA) cycles test changes on a small scale before adoption. A systematic review found that many published PDSA projects did not follow the method’s key features, such as iterative cycles and prediction of results [11]. The SQUIRE 2.0 guidelines specify how quality-improvement work should be reported [12].

3. Validation Objectives and Questions

Table I lists the validation objectives, the question each one answers and the criterion used to judge success.

TABLE I VALIDATION OBJECTIVES, QUESTIONS AND SUCCESS CRITERIA

Stage	Objective	Question	Success criterion
1	Expert consensus	Do clinicians agree that each statement belongs in the protocol?	$\geq 75\%$ rating 3–4 after ≤ 3 rounds
2	Content validity	Is agreement greater than expected by chance?	$I-CVI \geq 0.78$; $\kappa \geq 0.74$; $S-CVI/Ave \geq 0.90$
3a	Feasibility	Can the protocol be followed and documented?	$\geq 80\%$ of gates documented; follow-up $\geq 80\%$ at 12 months
3b	Outcome evaluation	Do patients improve by a meaningful amount?	Mean ODI change and responder proportion with 95% CI
3c	Safety	Are harms within expected ranges?	Complications recorded and reviewed
4	Continuous improvement	Does audit lead to measured improvement?	Indicators improve across PDSA cycles

4. Stage 1: Modified Delphi Consensus

4.1 Panel and Item Pool

The panel includes spine surgeons, physiotherapists, radiologists, anaesthetists, nurses and, where possible, patient representatives. A panel of 10–15 members balances diversity against the practical burden of several rounds. The item pool is drawn from the protocol domains: eligibility and red flags, clinical and neurological assessment, imaging concordance, conservative care, shared decision-making, decompression planning, peri-procedural safety, rehabilitation and outcome review. Each statement is worded as a single, testable recommendation.

4.2 Rounds, Feedback and Consensus

Fig. 1 summarises the process. In round 1, experts rate each statement on a four-point scale (1 = not relevant, 4 = highly relevant) and may comment or propose new statements. Between rounds, the group distribution and anonymised comments are fed back, and ambiguous statements are reworded. Rounds 2 and 3 repeat the rating for statements without consensus. Consensus is defined in advance, consistent with recommended practice [4], as at least 75% of panel members rating a statement 3 or 4. Statements below 50% are removed, and statements between 50% and 75% after the final round are referred to a moderated panel discussion.

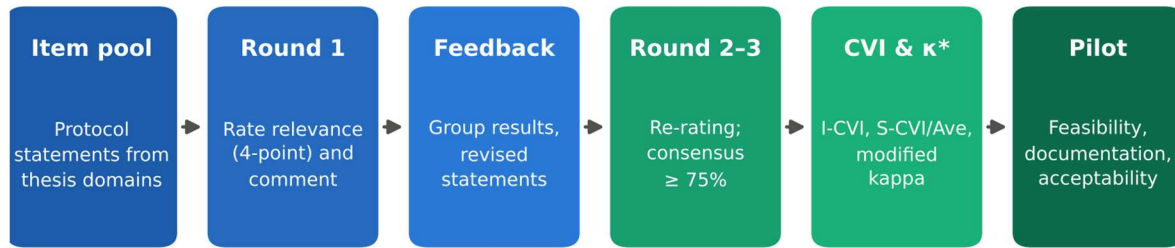


Fig. 1. Modified Delphi and content-validation process, from item pool to pilot implementation.

5. Stage 2: Content Validity

5.1 Indices

For an item rated by N experts, of whom A give a rating of 3 or 4, the item-level index is

$$I-CVI = A / N \quad (1)$$

The scale-level index S-CVI/Ave is the mean I-CVI across items. Following Polit *et al.* [6], the probability of chance agreement on relevance is

$$P_c = [N! / (A! (N - A)!)] \cdot 0.5^{N*} \quad (2)$$

and the modified kappa is

$$\kappa = (I-CVI - P_c) / (1 - P_c) \quad (3)$$

5.2 Retention Thresholds by Panel Size

Fig. 2(a) shows that κ converges towards the I-CVI as the panel grows, because chance agreement becomes less likely. Fig. 2(b) gives the minimum number of experts who must rate an item relevant for it to be retained under the rule $I-CVI \geq 0.78$ and $\kappa \geq 0.74$. Table II lists selected values. The rule becomes slightly less demanding in relative terms for larger panels, but the absolute number of agreeing experts rises steadily.

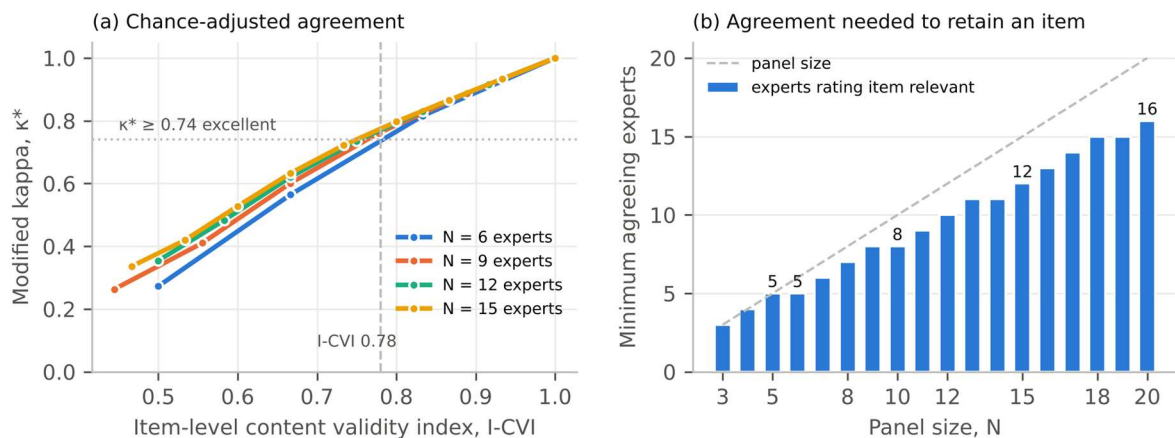


Fig. 2. (a) Modified kappa as a function of the item-level content validity index for panels of 6–15 experts (Eqs. 1–3). (b) Minimum number of experts who must rate an item relevant for retention ($I-CVI \geq 0.78$ and $\kappa \geq 0.74$), by panel size.

TABLE II MINIMUM AGREEMENT REQUIRED TO RETAIN AN ITEM

Panel size N	Minimum agreeing experts	I-CVI at threshold	Interpretation
6	5	0.83	One dissent allowed
8	7	0.88	One dissent allowed
10	8	0.80	Two dissents allowed
12	10	0.83	Two dissents allowed
15	12	0.80	Three dissents allowed
20	16	0.80	Four dissents allowed

5.3 Decision Rules for Items

Items meeting the threshold are retained. Items with I-CVI between 0.70 and 0.78 are revised using expert comments and re-rated. Items below 0.70 are removed unless the panel judges them essential for safety, in which case the reason is recorded. The protocol as a whole is accepted when S-CVI/Ave is at least 0.90. Inter-rater agreement on the concordance matrix and the red-flag gate can additionally be tested with clinical vignettes rated independently by panel members.

6. Stage 3: Prospective Pilot and Outcome Evaluation

6.1 Design and Participants

The pilot is a prospective, single-arm study of adults with radicular leg pain or neurogenic claudication and imaging-confirmed nerve root

compression who enter the protocol. Patients with red-flag presentations are managed on the urgent pathway and analysed separately. All patients who enter the protocol are followed, including those who choose continued conservative care at the shared decision node, so that the outcome of the whole pathway—not only of surgery—is described. The shared-decision step follows an established model [13] and may use a decision aid [14]; its documentation is a feasibility outcome.

6.2 Outcome Measures and Schedule

Table III defines the outcome set. The primary outcome is change in ODI from baseline to 12 months [7]. A responder is defined as a patient who improves by at least 10 points or 30% [8]; sensitivity analyses use alternative thresholds reported for surgical populations [9].

TABLE III OUTCOME MEASURES, INSTRUMENTS AND TIME POINTS

Domain	Instrument	Time points	Role
Disability	Oswestry Disability Index (0–100)	Baseline, 6 wk, 3, 6, 12 mo	Primary outcome
Pain	Numeric rating scale, back and leg (0–10)	All time points	Secondary
Walking capacity	Self-reported walking distance or timed walk	Baseline, 6 wk, 3, 6, 12 mo	Secondary
Neurological status	Motor, sensory and reflex examination	All time points	Safety and secondary
Quality of life	EQ-5D-5L [10]	Baseline, 3, 6, 12 mo	Secondary
Satisfaction	Satisfaction and goal attainment	6 wk, 3, 6, 12 mo	Patient-centred
Complications	Adverse events, reoperation, readmission	2 wk to 12 mo	Safety
Process	Gate documentation; follow-up completion	Continuous	Feasibility

6.3 Sample Size and Precision

For the primary outcome, the number of patients needed to detect a mean within-patient change Δ with a paired analysis is approximately

$$n = [(z_{1-\alpha/2} + z_{1-\beta}) \sigma_d / \Delta]^2 + z_{1-\alpha/2}^2 / 2 \quad (4)$$

where σ_d is the standard deviation of change. Fig. 3(a) evaluates Eq. (4) at $\alpha = 0.05$. With $\Delta = 10$ ODI points and $\sigma_d = 20$, 34 patients are needed for 80% power and 44 for 90%; if σ_d is 25, the figures rise to 51 and 68. Detecting a smaller change of 5 points with $\sigma_d = 20$ requires 128 patients. Because the protocol includes non-operative pathways, the change in the whole cohort may be smaller than in surgical series, so the design assumes $\Delta = 10$ and $\sigma_d = 20$ as a minimum.

A pilot is often judged by the precision of its estimates rather than by power. The half-width of a 95% confidence interval for the responder proportion p is

$$h = 1.96 \sqrt{[p(1-p)/n]} \quad (5)$$

Fig. 3(b) shows that a half-width of ± 10 percentage points requires 97 patients when $p = 0.5$ and 81 when $p = 0.7$. Allowing for 15% loss to follow-up, the pilot targets about 115 enrolled patients, which provides adequate power for the primary outcome and acceptable precision for the responder proportion.

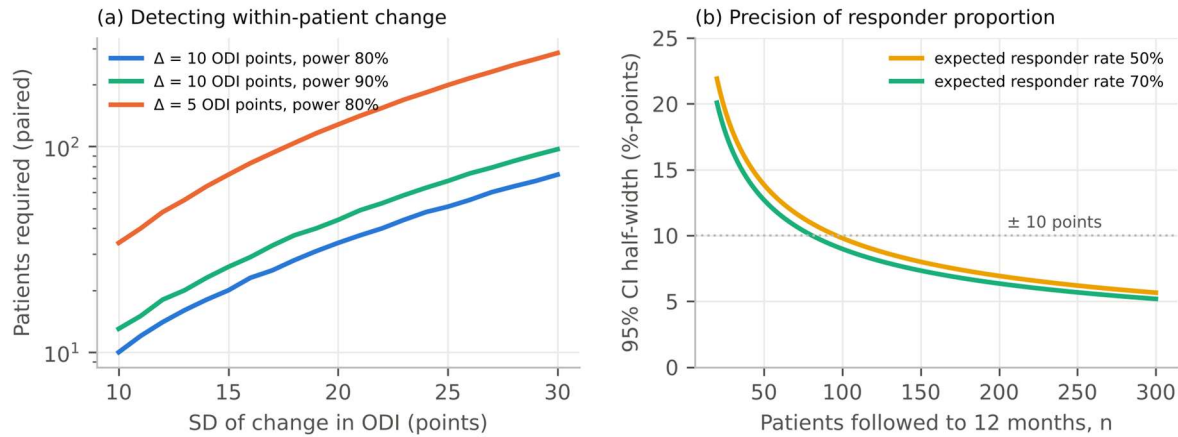


Fig. 3. (a) Patients required to detect a mean within-patient ODI change, by standard deviation of change, effect size and power (Eq. 4). (b) Half-width of the 95% confidence interval for the responder proportion by number of patients followed to 12 months (Eq. 5).

6.4 Analysis Plan and Missing Data

Continuous outcomes are summarised as means with 95% confidence intervals and analysed with linear mixed models that include all available time points, which remain valid when data are missing at random. The responder proportion is reported with a Wilson confidence interval. Pre-specified subgroups include surgical versus non-surgical pathway, stenosis versus disc herniation and baseline severity. Missing data are described by time point and reason; multiple imputation is used in a sensitivity analysis. Complications are

reported as counts and proportions with exact confidence intervals.

7. Stage 4: Audit and Continuous Improvement

After the pilot, the protocol is maintained through repeated PDSA cycles (Fig. 4). Each cycle selects one indicator—for example, the proportion of patients with a documented concordance judgement—sets a target, predicts the effect of a change, tests it on a small scale and compares the results with the prediction. Explicit prediction and iteration address the weaknesses observed in many published PDSA projects [11], and reporting follows SQUIRE 2.0 [12].

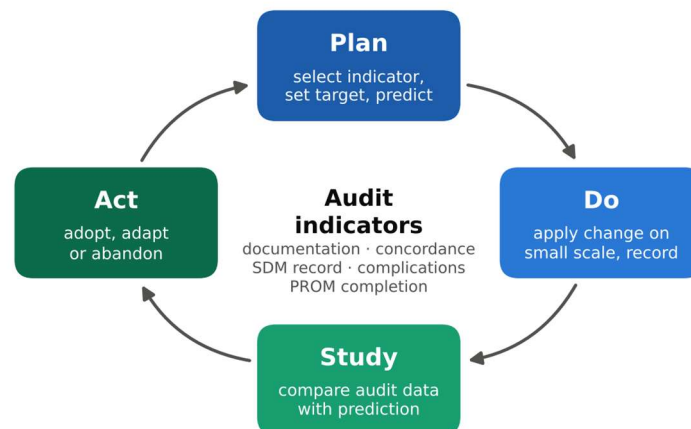


Fig. 4. Plan–do–study–act cycle for continuous improvement of the protocol, with the audit indicators used to select and evaluate changes.

8. Ethical Considerations

Delphi participation requires informed consent from panel members and confidentiality of individual ratings. The pilot requires institutional ethics approval, written informed consent and registration before enrolment. Because the protocol does not change the indications for surgery beyond accepted practice, the main ethical risks concern data protection and the burden of follow-up. Data are pseudonymised, stored securely and analysed only in aggregate. The shared decision node supports the patient's right to decline surgery without loss of follow-up or care.

9. Discussion

The validation design separates questions that are often merged: whether clinicians agree with a protocol, whether that agreement exceeds chance, whether the protocol can be followed and whether patients improve. Each stage has a predefined threshold and decision rule, which reduces the risk that thresholds are adjusted after the results are known—a weakness identified in reviews of Delphi studies [3], [4]. The content validity analysis shows that small panels need near-unanimous agreement to exceed chance, which argues for a panel of at least ten members.

The sample-size analysis highlights a practical tension. A pilot sized to detect a 10-point mean ODI change can be modest, but precise estimation of the responder proportion needs roughly three times as many patients. Because the responder proportion is what patients and clinicians usually want to know, the design is sized for precision. The main limitation is the single-arm design: improvement over time cannot be attributed to the protocol itself, because natural history and regression to the mean also contribute. A controlled or stepped-wedge comparison would be needed to estimate the effect of implementing the protocol. The thresholds in this paper are illustrative calculations, not results; actual values will depend on the observed variability of outcomes.

10. Conclusion

This paper presented a four-stage design for validating a patient-centred spinal decompression protocol: modified Delphi consensus, chance-adjusted content validity, a prospective pilot with

validated outcome measures and continuing PDSA audit. Explicit thresholds were derived for item retention by panel size, and calculations showed that about 34–44 patients are needed to detect a 10-point mean ODI change, while about 97 are needed to estimate the responder proportion within ± 10 percentage points. Future work will convene the expert panel, complete the Delphi rounds and conduct the pilot, after which a controlled evaluation can test whether the protocol improves outcomes compared with usual care.

References

- [1] J. N. Weinstein et al., "Surgical versus nonsurgical therapy for lumbar spinal stenosis," *N. Engl. J. Med.*, vol. 358, no. 8, pp. 794–810, 2008, doi: [10.1056/NEJMoa0707136](https://doi.org/10.1056/NEJMoa0707136).
- [2] W. C. Peul et al., "Surgery versus prolonged conservative treatment for sciatica," *N. Engl. J. Med.*, vol. 356, no. 22, pp. 2245–2256, 2007, doi: [10.1056/NEJMoa064039](https://doi.org/10.1056/NEJMoa064039).
- [3] R. Boukdedid, H. Abdoul, M. Loustau, O. Sibony, and C. Alberti, "Using and reporting the Delphi method for selecting healthcare quality indicators: A systematic review," *PLoS ONE*, vol. 6, no. 6, Art. no. e20476, 2011, doi: [10.1371/journal.pone.0020476](https://doi.org/10.1371/journal.pone.0020476).
- [4] I. R. Diamond et al., "Defining consensus: A systematic review recommends methodologic criteria for reporting of Delphi studies," *J. Clin. Epidemiol.*, vol. 67, no. 4, pp. 401–409, 2014, doi: [10.1016/j.jclinepi.2013.12.002](https://doi.org/10.1016/j.jclinepi.2013.12.002).
- [5] M. R. Lynn, "Determination and quantification of content validity," *Nurs. Res.*, vol. 35, no. 6, pp. 382–386, 1986, doi: [10.1097/00006199-198611000-00017](https://doi.org/10.1097/00006199-198611000-00017).
- [6] D. F. Polit, C. T. Beck, and S. V. Owen, "Is the CVI an acceptable indicator of content validity? Appraisal and recommendations," *Res. Nurs. Health*, vol. 30, no. 4, pp. 459–467, 2007, doi: [10.1002/nur.20199](https://doi.org/10.1002/nur.20199).
- [7] J. C. T. Fairbank and P. B. Pynsent, "The Oswestry Disability Index," *Spine*, vol. 25, no. 22, pp. 2940–2953, 2000, doi: [10.1097/00007632-200011150-00017](https://doi.org/10.1097/00007632-200011150-00017).
- [8] R. W. J. G. Ostelo et al., "Interpreting change scores for pain and functional status in low back pain: Towards international consensus regarding minimal important change," *Spine*, vol. 33, no. 1, pp. 90–94, 2008, doi: [10.1097/BRS.0b013e31815e3a10](https://doi.org/10.1097/BRS.0b013e31815e3a10).

- [9] A. G. Copay, S. D. Glassman, B. R. Subach, S. Berven, T. C. Schuler, and L. Y. Carreon, "Minimum clinically important difference in lumbar spine surgery patients: A choice of methods using the Oswestry Disability Index, Medical Outcomes Study questionnaire Short Form 36, and pain scales," *Spine J.*, vol. 8, no. 6, pp. 968–974, 2008, doi: [10.1016/j.spinee.2007.11.006](https://doi.org/10.1016/j.spinee.2007.11.006).
- [10] M. Herdman et al., "Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L)," *Qual. Life Res.*, vol. 20, no. 10, pp. 1727–1736, 2011, doi: [10.1007/s11136-011-9903-x](https://doi.org/10.1007/s11136-011-9903-x).
- [11] M. J. Taylor, C. McNicholas, C. Nicolay, A. Darzi, D. Bell, and J. E. Reed, "Systematic review of the application of the plan–do–study–act method to improve quality in healthcare," *BMJ Qual. Saf.*, vol. 23, no. 4, pp. 290–298, 2014, doi: [10.1136/bmjqs-2013-001862](https://doi.org/10.1136/bmjqs-2013-001862).
- [12] G. Ogrinc et al., "SQUIRE 2.0 (Standards for QQuality Improvement Reporting Excellence): Revised publication guidelines from a detailed consensus process," *BMJ Qual. Saf.*, vol. 25, no. 12, pp. 986–992, 2016, doi: [10.1136/bmjqs-2015-004411](https://doi.org/10.1136/bmjqs-2015-004411).
- [13] G. Elwyn et al., "Shared decision making: A model for clinical practice," *J. Gen. Intern. Med.*, vol. 27, no. 10, pp. 1361–1367, 2012, doi: [10.1007/s11606-012-2077-6](https://doi.org/10.1007/s11606-012-2077-6).
- [14] D. Stacey et al., "Decision aids for people facing health treatment or screening decisions," *Cochrane Database Syst. Rev.*, no. 4, Art. no. CD001431, 2017, doi: [10.1002/14651858.CD001431.pub5](https://doi.org/10.1002/14651858.CD001431.pub5).