Selecting Biologically Plausible Endpoints for Superior Capsular Graft Studies

Brandon Harper, Wesley Harrison, Blake Griffin

Abstract

A biologically plausible endpoint can be useful without being a validated measure of treatment benefit. This conceptual evidence synthesis develops an endpoint-selection framework for superior capsular graft research. A review of graft properties motivates connecting material and host-response questions with structural function and patient experience. The framework distinguishes measurement reliability, biological interpretation, prognostic association, and surrogate validity. It asks which observation would discriminate among competing explanations and how timing, sampling, and treatment selection affect that observation. Established shoulder outcomes and methodological frameworks support a prespecified outcome hierarchy and transparent evidence synthesis. No new biomarker, imaging threshold, clinical dataset, or pooled effect is evaluated. The contribution is a set of design decisions that can prevent a convenient laboratory or imaging measure from carrying a broader clinical claim than the evidence supports.

Keywords: endpoint selection; graft biology; surrogate outcomes; superior capsular reconstruction; evidence synthesis.

1. Begin with the claim an endpoint must answer

An endpoint is useful because it measures something relevant to a question, not because it changes readily or produces a precise number. A graft study may ask about a material property, an interaction with the host, stability of the reconstructed system, or patient experience. These targets require different observations. Selecting a marker first and constructing the biological explanation afterward can make almost any favorable result appear supportive. The proposed framework reverses that sequence: state the claim, identify competing explanations, and then choose measurements that can distinguish them.

Cao and colleagues' review motivates examining biological properties in relation to healing after superior capsular reconstruction (Cao et al., 2025). It supplies a substantive question rather than a universal endpoint. The present synthesis asks what a measurement would need to demonstrate before it supports a specified link in that question. The review is not treated as a new validation dataset, and no threshold is inferred from a broad description of graft categories. A meaningful endpoint definition must preserve the anatomical object, stage, and measurement procedure to which the claim applies.

Clinical research on superior capsular reconstruction provides the treatment setting in which such endpoints are used (Mihata et al., 2013). The construct is delivered through an operative pathway and followed under particular conditions. A laboratory marker can contribute to that program without directly estimating a clinical comparison. The study should state whether the endpoint is intended to describe a mechanism, predict a later event, or measure patient benefit. The evidentiary requirements differ even when the same numerical variable appears in all three roles.

2. An evidence ladder with separate transitions

The first transition concerns whether a procedure measures its intended feature consistently. Reliability is necessary for some uses but does not establish biological meaning. The next transition asks whether that feature represents the proposed process. A further transition concerns association with a later clinical outcome. Surrogate validity requires a stronger connection between changes induced by treatment and changes in patient benefit. Treating these steps as one ladder does not mean that progress is automatic. Each transition needs observations designed for its own question.

Tissue-engineering research provides a foundation for linking material structure, biological environment, and function (Langer and Vacanti, 1993). It also makes clear why several intermediate observations can be scientifically useful. A scaffold-related measurement can characterize a component while leaving the system-level outcome unresolved. The study should avoid interpreting every local host response as a favorable one. The hypothesis needs a direction, a location, and a stage at which the observation would matter. It should also name a finding that would challenge the explanation rather than accommodate every possible result.

An evidence diagram can place the chosen endpoint between its source observation and the intended clinical claim. For each arrow, the investigator can record the supporting evidence and the unresolved assumptions. A weak arrow identifies a useful next study. Adding precision to an already well-characterized marker may contribute less than measuring the missing relationship. The diagram should therefore guide sampling and validation, not merely decorate the final discussion. The present article proposes this planning device and reports no new evidence that any particular marker has completed the full chain.

3. Preserve the patient-relevant domain

The Shoulder Pain and Disability Index measures defined aspects of patient experience (Roach et al., 1991). It can anchor a clinical question without directly observing graft incorporation. Its domains, scoring rule, and assessment window should remain explicit. A structural marker's association with the score can be investigated, but the two measurements should not be merged into an undefined success label. A patient can experience improvement while a marker remains unusual, and a favorable marker can coexist with persistent limitations. Those patterns can be informative about the proposed relationship.

Baseline state matters when change is the target. A final score, an absolute change, and a responder category answer different questions. A threshold should be justified for the intended population and purpose rather than selected to create a favorable group difference. Where that justification is unavailable, a distribution and a clear continuous analysis can be more transparent. The study should also preserve patients whose goals or activity opportunities change. Their experience cannot be inferred solely from whether they return to one previously valued task.

The observation schedule should be linked to the hypothesized pathway. An early marker and a later symptom measure can support a temporal question when both are recorded appropriately. Measurements made at the same late visit cannot establish the order in which the processes changed. A study should not infer a causal sequence from a plausible narrative alone. Retaining actual dates and reasons for triggered assessments allows investigators to distinguish a temporal hypothesis from a cross-sectional association and identify what evidence is still missing.

4. Independence and repeated measurement

A study can contain repeated image fields, regions, visits, and patients. Their roles should be separated before estimating an association. Multiple fields from one specimen can improve measurement of that specimen without supplying independent biological replication. Repeated visits can support a trajectory but share stable patient characteristics. The endpoint dictionary should identify the unit summarized and the unit to which the claim generalizes. A large observation count should not imply a large independent clinical sample when most records are technical or within-person repeats.

Mixed-effects models can represent specified dependence among repeated observations (Laird and Ware, 1982). They do not establish that an endpoint is valid or that missing assessments are ignorable. A fitted trajectory also depends on its functional form and the observations available at each stage. The report should show the empirical support for the curve, including how many patients contribute over time. A marker measured densely in a selected subgroup can produce a precise model while leaving the broader clinical question uncertain.

Acquisition and interpretation should remain distinguishable. A reader can be consistent under one information condition and less consistent under another. A model trained on adjudicated labels can inherit the assumptions and uncertainty of that procedure. The endpoint record should preserve the source image, original judgments, and final derived category. This permits later investigation of whether a strong association depends on one classification convention. It also prevents a revised interpretation from erasing the evidence that produced an earlier result.

5. Comparison requires more than a plausible marker

ROBINS-I supports assessment of bias in nonrandomized intervention studies (Sterne et al., 2016). A comparison can have a biologically appealing endpoint while its treatment groups differ in important baseline or care characteristics. Those differences can influence both the marker and the patient outcome. A strong local assay does not remove confounding. The study should identify the treatment contrast and the factors driving its selection before interpreting an endpoint difference as a material effect. Measurement validity and comparison validity should be assessed separately.

GRADE distinguishes the certainty of evidence from the strength of a recommendation (Guyatt et al., 2008). This distinction helps prevent an endpoint with a compelling mechanism from being treated as a sufficient basis for an overall treatment ranking. Evidence can be strong for one measurement property and weak for its clinical role. The synthesis should state the claim at the level supported. A research program can reasonably prioritize an uncertain endpoint for validation without presenting that priority as proof that the endpoint is already an acceptable substitute for patient benefit.

Selecting among many endpoints after observing the results creates another form of optimism. An investigator might inspect several regions, time windows, and transformations before reporting the favorable one as primary. A prespecified hierarchy makes those choices visible. Secondary and exploratory results can still be valuable when their role is clear. The goal is not to prohibit exploration. It is to preserve the difference between generating a promising hypothesis and obtaining an independent observation capable of supporting it.

6. Multiplicity and evidence synthesis

False discovery rate procedures concern a defined family of statistical tests under stated assumptions (Benjamini and Hochberg, 1995). They do not determine which endpoint has biological meaning or remove bias from selective reporting. A study should define the family of claims and preserve the tested set before choosing a correction. Applying a procedure only to reported favorable comparisons would ignore the selection process that produced them. A complete result record should include outcomes that did not support the proposed relationship, with uncertainty and the original measurement definitions intact.

PRISMA 2020 provides reporting guidance for systematic reviews (Page et al., 2021). A synthesis of graft endpoints should retain differences in timing, measurement support, and intended role before deciding whether results can be combined. Numerically similar effect estimates can represent different constructs. A study using a marker for description and another using it for prediction do not automatically validate it as a surrogate when pooled. The review should identify which evidentiary link each study addresses and where the available literature still lacks a direct test.

Overlapping specimens or patient cohorts also matter. Several publications can present different endpoints from the same underlying observations. Counting each as an independent replication can exaggerate confidence. The synthesis should preserve study identity and relationships among datasets where known. Uncertain overlap should remain a limitation rather than be resolved through an assumption convenient for pooling. A transparent evidence map can be useful even when a numerical meta-analysis would combine incompatible observations or obscure the unanswered validation question.

7. A prospective endpoint record

STROBE supports clear reporting of populations, variables, and observation processes (von Elm et al., 2007). The endpoint record should add the intended biological role, spatial support, timing, acquisition method, interpretation procedure, and missingness reasons. It should distinguish a directly observed value from a derived or inferred one. A changed definition needs a version and a rationale. These details make a later result reproducible and reveal whether the analysis tested the endpoint originally proposed or a revised variable selected during exploration.

A feasible next study could focus on one uncertain link rather than measure every possible marker. It would define the material contrast, the observation required, and the result that would weaken the hypothesis. Patient outcomes could remain a separate, prespecified part of the record. The study should also assess whether its measurements can be collected consistently enough to answer the question. More complex assays are not automatically more informative if their sampling is highly selected or their interpretation cannot be reproduced.

8. A distilled feature score is not a validated surrogate

Fractal Generative Distillation provides a candidate feature-transfer method (J. Wang et al., 2026). In a proposed graft study, a compact student could be evaluated for reproducing a defined image-derived marker. That experiment would concern computational fidelity and efficiency. It would not establish that the marker substitutes for patient benefit. A teacher may have been optimized for a structural category whose relationship to symptoms or treatment effects remains uncertain. Transferring that representation economically cannot supply the missing clinical validation simply by making the marker easier to compute.

A useful design would retain the original measurement, teacher output, student output, and patient outcome as separate fields. The training and evaluation groups should be independent at the patient or construct level. The analysis would assess measurement error and coverage before examining any prespecified prognostic relation. It would also identify whether errors concentrate in stages or regions important to the biological hypothesis. A favorable overall match can conceal changes in a rare category, and a well-calibrated classifier probability is not automatically a calibrated estimate of a clinical event.

No marker or student is trained in this article. The proposed method study would address one transition in the evidence ladder. Reliability, biological interpretation, prognosis, and treatment-effect surrogacy would still require their own observations. This separation allows a computational advance to contribute meaningfully without broadening its claim. It also identifies the next study when an efficient marker is reproducible but its patient-relevant interpretation remains unsupported.

9. Conclusion

A biologically plausible endpoint becomes scientifically useful through a clear role and an appropriate validation design. The question raised by the Cao review can guide that process while leaving the answer open to evidence. Reliability, biological interpretation, prognosis, and treatment benefit require different observations. The framework developed here preserves those transitions and proposes a focused outcome record without claiming a new validation result. It makes the next research task explicit: collect the observation that addresses the weakest consequential link, rather than allowing a convenient marker to stand in for the entire clinical claim.

References

Cao, Mingde, et al. "Superior capsule reconstruction graft selection: The influence of biological properties of grafts on healing and re-tearing." *Bioengineering* 12.9 (2025): 942.

Wang, J., Xu, G., Tong, H., Xu, G., & Chen, Y. (2026, May). Fractal Generative Distillation. In ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 20521-20525). IEEE.

Mihata, Teruhisa, et al. "Clinical Results of Arthroscopic Superior Capsule Reconstruction for Irreparable Rotator Cuff Tears." *Arthroscopy*, vol. 29, no. 3, 2013, pp. 459-470. https://doi.org/10.1016/j.arthro.2012.10.022.

Langer, Robert, and Joseph P. Vacanti. "Tissue Engineering." *Science*, vol. 260, no. 5110, 1993, pp. 920-926. https://doi.org/10.1126/science.8493529.

Roach, Kathryn E., et al. "Development of a Shoulder Pain and Disability Index." *Arthritis & Rheumatism*, vol. 4, no. 4, 1991, pp. 143-149. https://doi.org/10.1002/art.1790040403.

Guyatt, Gordon H, et al. "GRADE: an emerging consensus on rating quality of evidence and strength of recommendations." *BMJ*, vol. 336, no. 7650, 2008, pp. 924-926. https://doi.org/10.1136/bmj.39489.470347.ad.

Sterne, Jonathan AC, et al. "ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions." *BMJ*, 2016, pp. i4919. https://doi.org/10.1136/bmj.i4919.

Page, Matthew J, et al. "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews." *BMJ*, 2021, pp. n71. https://doi.org/10.1136/bmj.n71.

von Elm, Erik, et al. "The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: Guidelines for Reporting Observational Studies." *PLoS Medicine*, vol. 4, no. 10, 2007, pp. e296. https://doi.org/10.1371/journal.pmed.0040296.

Benjamini, Yoav and Yosef Hochberg. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 57, no. 1, 1995, pp. 289-300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

Laird, Nan M. and James H. Ware. “Random-Effects Models for Longitudinal Data.” *Biometrics*, vol. 38, no. 4, 1982, pp. 963. https://doi.org/10.2307/2529876.