

Comparing Reconstructive Grafts under Confounding by Indication

Brandon Fisher, Wesley Harper, Blake Harrison

Abstract

Graft choice in observational surgery studies reflects clinical judgment, prior treatment, anatomy, availability, and patient goals. Those same factors can influence outcomes, complicating comparisons that attribute an observed difference to the graft. This conceptual methods article develops a design for comparative graft research using superior capsular reconstruction and revision ACL reconstruction as separate applications. It specifies eligibility, treatment versions, time zero, follow-up, and outcomes before choosing an adjustment method. Propensity scores and matching are considered as ways to address measured baseline differences, with explicit attention to overlap and remaining bias. The framework distinguishes a material effect from the effect of a complete operative strategy and separates precision from identification. No cohort is newly analyzed and no comparative effect is estimated. The contribution is a practical causal-design specification for studies motivated by graft biology and revision-management evidence.

Keywords: graft selection; confounding by indication; propensity scores; observational surgery; comparative effectiveness.

1. Selection is part of the data-generating process

A graft is not ordinarily chosen by a random mechanism in clinical practice. A surgeon may consider available tissue, previous procedures, anatomical constraints, expected activity, and features that are difficult to encode in a registry. A later outcome comparison therefore begins with groups that may already differ in prognosis. Calling this difference selection bias is only a start. The design needs to identify which selection process matters for the intended treatment contrast and which observations could make that process more transparent. Otherwise increasingly complex adjustment can obscure a poorly defined question.

Cao and colleagues' superior capsular reconstruction review motivates a biological explanation for differences among graft constructs (Cao et al., 2025). Their APKASS statement addresses the management context of adult ACL revision (Cao et al., 2026). Read together as methodological examples, they highlight a central distinction: a reason to choose a graft can also be a reason why its recipients have different outcomes. The shoulder and knee applications remain separate throughout this article. Neither publication is used to supply an unobserved counterfactual outcome, and no patient-specific treatment choice is recommended.

The aim is to define an observational comparison that could test a specified effect. That requires more than adding a list of covariates to a regression. Investigators must decide who could have received either strategy, when the strategies begin, what each strategy contains, and which outcome is followed. The proposed framework makes those decisions explicit before examining whether the observed dataset can support them. A useful result may be that the desired comparison is not identifiable in the available population, rather than a forced numerical estimate with a misleadingly narrow interval.

2. Define the strategies and the population

An intervention should describe the graft construct sufficiently to distinguish relevant versions. Material origin, processing, dimensions, fixation, and accompanying procedures can alter the treatment represented by a broad label. If these versions are pooled, the resulting contrast concerns a mixture determined by the study's practice patterns. That can be a legitimate policy question, but it is not automatically the isolated effect of material biology. A report should state whether the intended intervention is a construct substitution under a standardized procedure or the complete strategy as ordinarily delivered at participating sites.

Eligibility should identify patients for whom both compared strategies were plausible at baseline. A comparison that includes severe anatomical constraints only in one group may have little empirical support for the alternative strategy in those patients. This is a problem of overlap rather than merely a large imbalance to be corrected. Investigators should preserve the reasons a graft was considered unavailable or unsuitable. The resulting restricted comparison may apply to a narrower population, but it can offer a clearer question than a nominally broad comparison supported by extrapolation across nonoverlapping clinical groups.

Time zero needs equal care. Beginning follow-up when one strategy is selected and beginning it only after another strategy is completed can assign different observation opportunities. Staged procedures make this distinction particularly visible. A study should decide whether it compares the initial strategies, their completed reconstructive procedures, or outcomes conditional on reaching a later stage. Each question requires a different population and may exclude different events. The design should not select the most favorable starting point retrospectively. A time diagram can expose these differences before any survival or regression model is fitted.

Clinical research on superior capsular reconstruction supplies one setting in which a full operative context accompanies graft use (Mihata et al., 2013). The MARS cohort supplies another setting for studying graft choice in revision ACL reconstruction (MARS Group, 2014). Their relevance is to the structure of the research problem. This article does not treat the two populations as exchangeable or combine their outcomes. The examples instead motivate recording selection, procedural details, and follow-up at enough resolution to understand what a comparison within each setting would actually estimate.

3. Put candidate covariates on a timeline

A baseline variable can be a common cause of treatment and outcome, a predictor of outcome alone, or a consequence of earlier treatment history. Its analytical role depends on the target question. Causal diagrams provide a way to state those assumptions (Greenland et al., 1999). The diagram should distinguish information available before graft selection from observations made during or after the procedure. A postoperative imaging finding cannot serve as an ordinary baseline adjustment variable when it may lie on the pathway through which the graft affects the later outcome.

Prior surgery is not a single uniform characteristic. Its details can affect current anatomy, available tissue, expectations, and the surgeon's strategy. A registry field recording only the number of prior operations may leave important selection differences unresolved. Conversely, collecting every possible detail without an analysis plan can produce unstable models and obscure the key variables. The proposed design begins with a small causal specification, then asks whether the measurements adequately represent the relevant histories. Unmeasured clinical judgment should remain an explicit limitation rather than being assumed to disappear when many administrative fields are included.

Site and calendar period can represent both treatment availability and broader changes in care. If a graft is used only at a particular site or becomes available late in recruitment, a material comparison may largely compare teams or eras. The data should show these distributions before adjustment. A model can include site indicators, but it cannot create within-site treatment variation where none exists. The interpretation should follow the actual support. A study estimating a strategy difference across sites needs a stronger argument about comparability than a study with meaningful variation within the same care environments.

4. Propensity scores organize measured differences

The propensity score describes the probability of receiving a treatment conditional on observed baseline covariates under a specified assignment model (ROSENBAUM and RUBIN, 1983). It can support matching, weighting, or stratification, but its value depends on the design and variables used. A high-performing treatment classifier is not the objective. Near-perfect prediction can signal poor overlap, leaving few comparable patients. Investigators should examine the treatment distributions and covariate balance relevant to the target population rather than interpreting a strong assignment prediction as proof that confounding has been controlled.

Matching methods provide a framework for constructing more comparable observed groups (Stuart, 2010). A graft study should report which patients were matched, which were excluded, and how the resulting population differs from the original cohort. The estimate after matching applies to that constructed population under the stated assumptions. Matching does not remove the need for an outcome model or uncertainty calculation appropriate to the matched design. Nor does it address an unmeasured factor simply because the measured factors appear balanced. The balance table should be accompanied by an account of the clinical information that remains unavailable.

Weighting similarly changes the population represented by the analysis. Large weights can indicate that a few observations stand in for many patients with limited treatment alternatives. Truncating those weights changes the procedure and can introduce another bias-variance tradeoff. A report should preserve the weight distribution, any truncation rule, and the resulting effective representation of the groups. It should also state the estimand, such as an average effect in an eligible population or an effect among recipients of one strategy. These targets should not be exchanged after the results are inspected.

5. Precision does not establish identification

Heteroskedasticity-consistent covariance estimation addresses a variance assumption in a specified model (White, 1980). It does not fix treatment misclassification, lack of overlap, or unmeasured confounding. A narrow robust interval can therefore surround a biased treatment contrast. This distinction is especially important when observational datasets become large. More records can improve numerical precision without making the groups more comparable. A complete interpretation should discuss the direction and plausible importance of remaining design problems rather than equating statistical significance with a credible causal effect.

Patients treated by the same surgeon or centre may share unobserved practices and outcome-ascertainment patterns. Cluster-robust inference provides tools for considering dependence within such groups (Cameron and Miller, 2015). Its reliability still depends on the number and structure of clusters. It also addresses a different question from whether site-level treatment assignment is confounded. A standard error correction cannot repair a comparison in which all recipients of one graft come from a single distinctive centre. The design should separate the identification problem from the dependence problem and identify what evidence would improve each.

ROBINS-I offers a structured approach to evaluating bias in nonrandomized intervention studies (Sterne et al., 2016). In a graft comparison, the assessment should retain the connection between a specific bias domain and a possible distortion of the intended effect. Treating the tool as a checklist completed after the analysis would miss much of its value. It can instead guide data collection and the decision about whether a comparison should be attempted. A clearly described limitation is more useful than an unqualified effect estimate whose assumptions cannot be assessed from the report.

6. Follow-up and deviations from the strategy

Subsequent treatment and rehabilitation can both influence outcomes and respond to early recovery. A study must decide whether it evaluates an initial graft strategy under ordinary subsequent care or a sustained protocol. Adjusting indiscriminately for postoperative variables can remove part of the strategy's effect or condition on consequences of early failure. A detailed time diagram helps identify this problem. More elaborate longitudinal methods may be appropriate for some questions, but they require additional measurements and assumptions. This article proposes no unobserved rehabilitation history and performs no longitudinal effect estimation.

Loss to follow-up can also differ between strategies for reasons related to recovery, dissatisfaction, access, or additional treatment elsewhere. A comparison among observed returners can then represent a selected subgroup. Recording contact attempts, reasons for missing visits, and outcomes from linked permitted records can clarify the observation process. Sensitivity analyses should target plausible missingness mechanisms rather than assume that one imputation procedure resolves all uncertainty. The study should distinguish the outcome it failed to observe from the patient's actual outcome, which remains unknown rather than conveniently favorable.

7. A transparent comparative study record

Availability can be measured separately from preference. A construct unavailable at a site creates a different selection constraint from a construct considered but declined for a documented clinical reason. Combining those states into one untreated group can obscure why overlap is absent. The prospective record should preserve the alternatives considered at the decision point where feasible. A sensitivity analysis could then distinguish a comparison among patients with documented alternatives from a broader comparison requiring stronger assumptions. This does not make clinical documentation a randomization mechanism; it improves the description of the assignment process.

Negative or inconclusive findings also need a design interpretation. An imprecise comparison after a defensible restriction can be more informative than a precise comparison created through unsupported extrapolation. The report should identify which population was lost by the restriction and what new observations would restore support. That explanation turns a limitation into a concrete plan for the next study.

The proposed record begins with eligibility, treatment versions, time zero, and outcomes. It then lists baseline confounders with their timing, observed overlap, the adjustment design, balance diagnostics, and the population represented after adjustment. It preserves deviations, subsequent treatment, loss of observation, and dependence by site or surgeon. This is a research specification, not a list of clinical instructions. Its purpose is to make the intended effect and the empirical support visible so that another investigator can challenge the design before accepting its numerical result.

Biological hypotheses from the shoulder review can help identify candidate effect modifiers, but those modifiers should be measured before the relevant treatment contrast where the question requires it. Revision-management considerations from the knee statement can help explain why treatment selection is

not random. Their joint contribution is to improve the causal question, not to supply an answer across procedures. A future study could test a narrowly defined strategy contrast within each setting and report how its population differs from the clinical populations that motivated it.

Comparative graft research becomes more credible when selection is treated as a scientific object rather than an inconvenience left to regression. Propensity methods, matching, robust covariance, and bias assessment have useful roles after the comparison is defined. They cannot substitute for a population with plausible alternatives or for measurements of the factors driving selection. The framework developed here identifies those requirements and preserves the limits of the available evidence. It produces a design for a testable comparison, while leaving the actual comparative effect to a study with the observations needed to estimate it.

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.
- Cao, Mingde, et al. "APKASS 2024 consensus statement on anterior cruciate ligament reconstruction, Part II: Management of anterior cruciate ligament revision in adults." *Asia-Pacific Journal of Sports Medicine, Arthroscopy, Rehabilitation and Technology* 45 (2026): 74-81.
- The MARS Group. "Effect of Graft Choice on the Outcome of Revision Anterior Cruciate Ligament Reconstruction in the Multicenter ACL Revision Study (MARS) Cohort." *The American Journal of Sports Medicine*, vol. 42, no. 10, 2014, pp. 2301-2310. <https://doi.org/10.1177/0363546514549005>.
- 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>.
- Greenland, Sander, Judea Pearl, and James M. Robins. "Causal Diagrams for Epidemiologic Research." *Epidemiology*, vol. 10, no. 1, 1999, pp. 37-48. <https://doi.org/10.1097/00001648-199901000-00008>.
- 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>.
- ROSENBAUM, PAUL R., and DONALD B. RUBIN. "The central role of the propensity score in observational studies for causal effects." *Biometrika*, vol. 70, no. 1, 1983, pp. 41-55. <https://doi.org/10.1093/biomet/70.1.41>.
- Stuart, Elizabeth A. "Matching Methods for Causal Inference: A Review and a Look Forward." *Statistical Science*, vol. 25, no. 1, 2010, pp. 1-21. <https://doi.org/10.1214/09-sts313>.
- White, Halbert. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica*, vol. 48, no. 4, 1980, pp. 817-838. <https://doi.org/10.2307/1912934>.
- Cameron, A. Colin, and Douglas L. Miller. "A Practitioner's Guide to Cluster-Robust Inference." *Journal of Human Resources*, vol. 50, no. 2, 2015, pp. 317-372. <https://doi.org/10.3368/jhr.50.2.317>.