# Defining Failure in Revision ACL Research across Structural, Functional, and Patient-Reported Outcomes

Adrian Wells, Gavin Russell, Tyler Sullivan

## Abstract

Failure after revision anterior cruciate ligament reconstruction can refer to structural disruption, recurrent symptoms, functional limitation, or additional treatment. These events are related without being interchangeable. This methodological synthesis develops an outcome framework around an adult revision-management consensus. It attaches an observation window and ascertainment process to structural status, patient reports, performance, and subsequent treatment. It explains why an apparent difference can arise from imaging intensity or follow-up rather than from a different clinical course. Established knee measures, reliability methods, repeated-measure models, and competing-risk frameworks are considered for specified questions. No new outcomes or pooled failure rates are calculated. The contribution is an outcome dictionary and design structure that preserve disagreement among domains instead of hiding it inside an undefined success label. The framework supports future revision studies while providing no patient-specific operative threshold.

Keywords: outcome definition; graft failure; structural integrity; patient-reported outcomes; competing risks; revision surgery.

## 1. Failure names several different observations

A study can describe a graft as intact, a patient as improved, and a clinical course as unsuccessful without logical contradiction. Each statement may refer to a different domain. The difficulty arises when those domains are compressed into a binary success label and their underlying observations disappear. A comparison can then seem to favor a graft because one group underwent fewer scans or had fewer opportunities for additional surgery. Defining failure requires both a clinical concept and a record of how the study attempted to observe it.

Cao and colleagues' APKASS revision ACL statement places failure within a broader management problem (Cao et al., 2026). That context motivates separating an observed state from a decision made in response to it. A participant may have an imaging finding, a functional limitation, or further treatment, and the study should not assume that those observations coincide. The present framework asks which domain an endpoint measures, when it was assessed, and which participants could contribute. It does not reinterpret the consensus as a validated binary definition suitable for every research population.

The proposed framework begins with four domains: structural findings, symptoms and patient-reported function, observed performance, and subsequent treatment. A study may select one as primary while reporting the others separately. The domains can inform one another without becoming substitutes. A patient may avoid repeat surgery despite persistent limitations, and a structural finding may not correspond to the symptom burden captured by a questionnaire. These possibilities should be represented in the data structure before investigators decide whether a composite endpoint is scientifically justified.

## 2. Patient reports require a domain and a population

The KOOS was developed as a self-administered outcome measure for knee injury and osteoarthritis (Roos et al., 1998). Its relevance is that a named instrument contains a specified set of domains rather than an undefined measure of knee success. A study should identify which domain or summary it analyzes and preserve the original scoring rule. Choosing the most favorable domain after inspecting results can change the apparent treatment comparison. Missing items and departures from the intended instrument version should remain visible because they affect what the reported number represents.

The IKDC subjective knee form provides another established approach to patient-reported knee assessment (Irrgang et al., 2001). Agreement or disagreement between instruments can be informative when their content differs. It should not be resolved by assuming that the instrument producing the larger change is more accurate. The study needs a prior account of the construct relevant to its question. Baseline severity also matters: identical numerical changes can occur from different starting points and may have different implications for the final state. A report should therefore retain both the trajectory and the final measurement.

A consistent assessment procedure is needed even when an instrument is well established. Reliability research distinguishes the measurement unit, reader population, and type of agreement being evaluated (Koo and Li, 2016). In a revision study, these choices matter for performance tests and structural judgments as well as for scored questionnaires. Agreement among a trained research panel does not automatically describe an ordinary single assessment. A protocol should retain the original measurements, rater roles, and ambiguous cases so that reproducibility can be examined separately from the biological meaning assigned to the result.

Change scores, final scores, and responder categories are also distinct outcomes. A responder rule needs a threshold justified for the intended population and purpose. A threshold selected because it creates the largest group difference would turn an outcome definition into another fitted parameter. When no defensible threshold is available, a continuous description with its distribution can be more transparent than an apparently decisive success rate. The study should also preserve patients whose experience crosses domains, such as improved pain accompanied by continued restrictions in activity.

## 3. Structural outcomes depend on observation

A structural endpoint should specify the imaging method, anatomical region, classification, reader procedure, and timing. Different schedules can create different opportunities to detect a change. An early examination may observe a different biological stage from a later examination even when both are labeled follow-up. A study should distinguish a scheduled image from an image obtained because symptoms prompted additional investigation. If symptomatic participants are imaged more often, a detected disruption is partly a consequence of the ascertainment process. That process belongs in the analysis rather than being treated as a neutral window onto every participant.

Reader agreement is a separate question from biological validity. Bland and Altman's method-comparison work emphasizes that correlation alone does not establish agreement (Bland and Altman, 1986). For continuous structural measurements, signed differences and their relation to magnitude can expose systematic disagreement. Categorical integrity judgments require a compatible agreement analysis and a record of ambiguous cases. A highly reproducible measurement can still represent an imperfect marker of functional incorporation. Conversely, disagreement can reflect a poorly specified boundary between categories rather than a failure of the imaging modality itself.

The event time can be uncertain when structural status is assessed intermittently. A disruption first observed at a scheduled examination occurred within an interval rather than necessarily on the examination date. Assigning all events to that date can distort an apparent temporal pattern. A study should retain the last known intact assessment and the first assessment indicating disruption. Whether an interval-censored method is needed depends on the research question and observation schedule. The essential first step is preserving the available timing information so that the analysis does not imply precision that was never observed.

## 4. Performance and repeated trajectories

Observed performance can reveal capacities that a symptom questionnaire does not directly measure. Its interpretation depends on task instructions, equipment, effort, familiarization, and the conditions under which the test was performed. A participant may decline or be unable to complete a task for reasons related to the outcome of interest. Excluding that observation without recording the reason can make the analyzed group appear more capable. A useful protocol distinguishes a missing test, an invalid attempt, and an observed inability to complete the task. These states should not be silently collapsed into one blank field.

Repeated-measure models can represent changes across follow-up and some dependence among observations from the same participant (Laird and Ware, 1982). Their value depends on a specified mean trajectory and covariance structure. They do not automatically repair selective follow-up or make an unscheduled measurement comparable with a scheduled one. A graft study should state whether its primary question concerns average trajectory, time to a defined event, or status at a prespecified visit. Using every available measurement can improve precision for one question while changing the sample contributing to another.

Heteroskedasticity-consistent covariance estimation addresses variance assumptions in an otherwise specified model (White, 1980). It does not validate the definition of a functional outcome or account for every repeated-person dependency. This distinction matters when a large number of visits creates an impression of a large number of independent patients. The reporting record should list participants, observations, and the distribution of observations per participant. A study comparing trajectories should also show when the groups were actually observed, rather than relying on an average follow-up duration that hides different schedules.

## 5. Additional treatment is an informative decision

Repeat surgery is important but combines clinical need with a decision process. Symptoms, imaging, expectations, access, surgeon assessment, and patient preference can influence whether an additional procedure occurs. Two groups can therefore differ in repeat-surgery rates even when their underlying structural outcomes are similar. The endpoint remains useful when interpreted as subsequent treatment under the observed care system. It should not be renamed structural failure without evidence supporting that equivalence. The study should record the indication and timing of the additional procedure where feasible.

Some subsequent events also alter the opportunity to observe the event of interest. A different major procedure, death, or withdrawal from a particular care pathway can change whether later graft-related treatment is observed. Competing-risk methods such as the Fine-Gray framework address specified cumulative-incidence questions (Fine and Gray, 1999). They do not decide which events belong in the competing category or what causal interpretation is justified. That choice should follow the clinical and

observation process. A method selected only because it is available in software can produce a technically correct estimate for an unintended question.

An event diagram can help. It would begin with the index procedure and distinguish structural detection, persistent symptoms, additional treatment, and loss of observation. Arrows would identify possible sequences rather than assume that every patient moves through one fixed order. A structural event can be detected before a treatment decision, after it, or not at all. Preserving the sequence allows investigators to ask targeted questions about the clinical course. A single terminal success label discards much of this information and can conceal which component explains a group difference.

## 6. Composite endpoints need an explicit rationale

Baseline status should remain attached to each outcome domain. A limitation documented before reconstruction is different from a new limitation during follow-up, even when the final score is identical. The opposite limb is not automatically an unaffected reference, and a patient's recalled earlier function can have a different measurement basis from a contemporaneous assessment. A prospective outcome dictionary should identify these reference states and preserve their source. This allows a later analysis to distinguish recovery, persistent burden, and newly developed problems without redefining the endpoint after the trajectory is known. It also makes an absent baseline visible as missing information rather than an assumed normal starting point.

A composite may be useful when its components reflect a coherent clinical objective, but convenience alone is insufficient. Combining an imaging finding with repeat surgery can make the result depend strongly on which component is more common or more intensively observed. The study should report component frequencies and the rule for counting multiple events in the same patient. It should also explain whether all components have similar importance for the intended interpretation. An apparent reduction in the composite can coexist with no improvement in its most consequential component.

An alternative is a structured outcome profile. The profile reports the primary endpoint together with prespecified secondary domains, without pretending that every domain can be reduced to one ranking. This approach can expose tradeoffs and support later evidence synthesis. It also requires restraint in interpretation: investigators should not declare a treatment favorable by choosing whichever profile component happens to improve. A hierarchy of questions and a complete result table make that judgment reviewable. The proposed profile is a design suggestion, not an outcome set validated by new patient data in this article.

## 7. Reporting a result that can be compared

STROBE emphasizes the information needed to interpret an observational study's design and findings (von Elm et al., 2007). For the present problem, the essential additions are an outcome dictionary, ascertainment schedule, handling of ambiguous observations, and the relationship among component endpoints. Each reported denominator should identify which participants were eligible to contribute. A failure proportion among those returning for imaging and a repeat-procedure proportion among all enrolled participants answer different questions. Presenting both is useful when the distinction is explicit and the missingness is visible.

The revision-management question can be investigated more clearly when the observed state and the treatment response remain distinguishable. A future study could use the proposed framework to identify where structural, symptomatic, and functional outcomes agree or diverge. It should also preserve domains that remain unobserved. This is a demand for clearer measurement, not a rule for turning several imperfect

observations into one clinical decision. A finding can contribute to the research program even when it does not support an overall ranking of reconstructive strategies.

Missingness should be represented within each outcome domain rather than with one generic follow-up indicator. Missing-data methods depend on assumptions about how unobserved values relate to the available information (Little and Rubin, 2002). A participant can supply a questionnaire while missing imaging, or undergo another procedure before a scheduled test. Treating those patterns as equivalent can change the endpoint the analysis estimates. A prospective dictionary should preserve the reason and timing of each absent observation, allowing sensitivity analyses to address plausible observation processes instead of filling every blank under an unexplained default.

Failure becomes scientifically useful when it names a reproducible observation at a stated time under a known ascertainment process. Structural disruption, patient-reported limitation, performance, and additional treatment can all matter, but they should remain identifiable. The proposed framework preserves that detail and clarifies the role of repeated-measure and competing-risk methods. Its contribution is a more interpretable outcome architecture for graft research. It supplies no new failure rate and does not decide which reconstructive strategy is best for a particular patient.

## References


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.

Roos, Ewa M., et al. "Knee Injury and Osteoarthritis Outcome Score (KOOS): Development of a Self-Administered Outcome Measure." *Journal of Orthopaedic & Sports Physical Therapy*, vol. 28, no. 2, 1998, pp. 88-96. https://doi.org/10.2519/jospt.1998.28.2.88.

Irrgang, James J., et al. "Development and Validation of the International Knee Documentation Committee Subjective Knee Form." *The American Journal of Sports Medicine*, vol. 29, no. 5, 2001, pp. 600-613. https://doi.org/10.1177/03635465010290051301.

Koo, Terry K., and Mae Y. Li. "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research." *Journal of Chiropractic Medicine*, vol. 15, no. 2, 2016, pp. 155-163. https://doi.org/10.1016/j.jcm.2016.02.012.

Bland, J. Martin, and Douglas G. Altman. "Statistical Methods for Assessing Agreement between Two Methods of Clinical Measurement." *The Lancet*, vol. 327, no. 8476, 1986, pp. 307-310. https://doi.org/10.1016/S0140-6736(86)90837-8.

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.

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.

Fine, Jason P., and Robert J. Gray. "A Proportional Hazards Model for the Subdistribution of a Competing Risk." *Journal of the American Statistical Association*, vol. 94, no. 446, 1999, pp. 496-509. https://doi.org/10.1080/01621459.1999.10474144.

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.

Little, Roderick J. A., and Donald B. Rubin. *Statistical Analysis with Missing Data*. Wiley, 2002. https://doi.org/10.1002/9781119013563.