Imaging Evidence in Adult ACL Revision: Reliability, Timing, and Interpretation

Cameron Griffin, Jordan Palmer, Austin Wallace

Abstract

Imaging contributes to revision anterior cruciate ligament research through observations made at particular times under specified acquisition and interpretation procedures. An image does not directly reveal the entire process of incorporation or establish which management response improves outcomes. This methodological synthesis develops an imaging-evidence framework around an adult revision ACL consensus. It distinguishes acquisition, reliability, construct validity, and change over time. Scheduled and symptom-triggered assessments are considered separately, together with repeated observations and informative missingness. Established knee outcomes provide a patient-relevant link without being treated as direct tissue measurements. No image set is newly analyzed and no diagnostic accuracy or healing rate is calculated. The contribution is a reproducible measurement record and a sequence of validation questions that can clarify what an imaging result supports in a prospective revision study.

Keywords: graft incorporation; imaging reliability; repeated measurements; reference standards; missing data; reconstructive surgery.

1. An image observes a stage of a process

An imaging assessment observes selected features under an acquisition protocol. In revision ACL research, a finding can describe anatomy or structural status without identifying the full process that produced it. The research question should specify the feature, the region, and the time at which it is meaningful. A signal change should not be renamed a biological property unless an appropriate validation supports that interpretation. The proposed framework keeps the measured feature available even when its clinical meaning remains uncertain.

Cao and colleagues' APKASS statement places assessments inside the management context of adult ACL revision (Cao et al., 2026). This motivates a distinction between observing a feature and deciding how to respond. The imaging marker and the management decision require different evidence. A marker can be repeatable without proving that acting on it improves a later outcome. The present article develops those validation questions and does not claim that a particular image category establishes the need for a procedure or identifies the best graft strategy.

The analysis is conceptual. No image set, pathology specimen, or patient record is newly assessed. The proposed framework separates five steps: defining the construct, acquiring an image, interpreting a measurement, evaluating its reliability, and assessing whether it predicts or explains a meaningful outcome. A weakness at one step cannot always be repaired at another. For example, highly consistent readers can agree on a feature that is weakly related to incorporation. A strong association with an outcome can also coexist with poor reproducibility that limits individual interpretation.

2. Specify what the marker is intended to represent

A marker may describe continuity, thickness, signal intensity, interface appearance, or another prespecified feature. These are different observed objects. A study should identify the anatomical region and the measurement support rather than use a broad healing label. Regional heterogeneity can matter: one interface can appear different from another part of the construct. Averaging across regions may be appropriate for some questions, but it should not conceal the possibility that the limiting mechanism is local. The protocol should preserve regional information when the hypothesis depends on it.

A reference standard also needs a stated role. Direct tissue observations, operative findings, and later clinical outcomes can provide complementary evidence, but none is an unquestionable universal truth for every imaging question. A tissue sample can be spatially limited, and additional surgery can select patients with particular symptoms. A comparison against either reference should describe that selection. The analysis must distinguish agreement with an observed reference from validation of the entire biological concept. Otherwise the word gold standard can hide uncertainty rather than clarify it.

The timing of the marker is part of its definition. An early signal can reflect a different stage from a later one, so a common follow-up label is insufficient when visit windows are broad. Investigators should retain the date, permitted window, and reason for acquisition. A change between two scans can then be interpreted in relation to the interval actually observed. A cross-sectional comparison of patients scanned at different stages should not be described automatically as the within-person trajectory of incorporation.

3. Reliability before biological interpretation

Reader reliability concerns the consistency of an interpretation procedure. Intraclass correlation coefficients require a specified model, measurement unit, and form of agreement (Koo and Li, 2016). Reporting one unnamed coefficient can therefore leave the evaluation ambiguous. A study should identify whether it concerns one reader or an average of readers, absolute agreement or consistency, and the population of readers to which it intends to generalize. The choice should follow the intended use. Reliability of an averaged research panel does not automatically describe an ordinary single-reader assessment.

For continuous measurements, Bland and Altman offer a complementary approach based on differences between methods or readers (Bland and Altman, 1986). Examining the direction and spread of differences can reveal systematic offsets and variability related to measurement magnitude. A high correlation can coexist with important disagreement. Repeated images or several readers require an analysis that preserves that structure. The present article uses the principle to define a research task; it supplies no invented limits of agreement for graft imaging and does not assume that a particular error magnitude is clinically acceptable.

Categorical interpretations require similar discipline. A binary intact-versus-disrupted label may conceal uncertain or partially observed cases. An adjudication procedure can improve consistency, but it should retain the original reader decisions and reasons for disagreement. If a model is later trained on the adjudicated labels, the record can reveal whether it is learning a stable visual distinction or a convention imposed on ambiguous observations. Excluding ambiguous cases can make accuracy appear stronger while narrowing the population. Their frequency and characteristics should remain part of coverage reporting.

4. Acquisition variation can imitate change

Scanner settings, patient positioning, sequence selection, and processing can affect an imaging feature. If these conditions change between visits, an apparent biological trajectory can partly reflect a measurement change. A prospective protocol should standardize the conditions most relevant to its marker and record departures. Standardization does not require pretending that every clinical site has identical equipment. It requires enough metadata to distinguish plausible biological variation from acquisition variation and to identify the conditions under which a marker has actually been evaluated.

Calibration objects or repeated acquisition can help characterize some sources of technical variation, but their relevance depends on the marker. A phantom can test a scanner response without reproducing the full tissue environment. A repeat scan can estimate short-term measurement variability without

identifying long-term biological change. The validation report should state which uncertainty each comparison addresses. Accumulating several technical checks is useful when their roles are explicit; it does not automatically create a complete validation of graft incorporation or of a treatment effect.

Imaging schedules can also differ through care decisions. A participant with symptoms may receive additional imaging, while an asymptomatic participant contributes only a scheduled scan. Comparing all available images without representing that process can overrepresent concerning periods. The data should distinguish protocol-driven and clinically triggered acquisitions. Their observations may both be valuable, but they need not estimate the same distribution. A research trajectory should not be reconstructed by treating every scan as if it occurred under a common independent schedule.

5. Repeated measurements and missing observations

Mixed-effects models provide a framework for representing dependence among repeated observations (Laird and Ware, 1982). A prospective graft study might estimate average marker change while allowing person-level variation. Its model should state the time scale, functional form, and covariance assumptions. The method does not determine whether the marker is biologically valid or whether the missing scans are ignorable. A fitted smooth trajectory can appear persuasive even when few participants are observed at its later end. Reporting the number and distribution of measurements over time helps reveal that limitation.

Missing-data methods depend on assumptions about the process that produced the missingness (Little and Rubin, 2002). An absent scan can arise from recovery, dissatisfaction, access barriers, additional treatment, or technical failure. Those reasons can relate differently to the unobserved marker. A study should record them where possible and distinguish a failed acquisition from a missed visit. Imputation can be part of an analysis, but it should not convert an unknown structural outcome into an apparently observed one. Sensitivity analyses should examine plausible departures from the primary assumption rather than merely repeat the same model under different software settings.

Revision or other additional surgery can change the construct being observed. A later scan may then describe a different intervention rather than continued follow-up of the original graft. The study needs a rule for that transition. It might analyze the original construct until the additional procedure, treat the procedure as an outcome, or define a broader strategy-level trajectory. These are different questions. The record should preserve the timing and reason for the procedure so that the chosen analysis does not silently combine observations of different constructs into one biological history.

6. From a marker to a comparative outcome

A marker can be associated with a later clinical result because both respond to underlying tissue condition, treatment history, or activity. That association does not establish that changing the marker would improve the clinical result. Evidence-grading frameworks help distinguish the certainty of different claims (Guyatt et al., 2008). A reliable imaging measure can therefore support description or prognosis before it supports use as a surrogate for treatment benefit. The study should name the intended role and identify what additional evidence would be needed to move from one role to another.

Clinical cohorts such as MARS illustrate the need to place graft-related observations inside a complete follow-up process (MARS Group, 2014). Their relevance to this framework is the relationship among treatment selection, patient characteristics, and subsequent outcomes. A proposed imaging substudy should report how participants were selected from the broader cohort and whether their measurement schedule differs. Detailed images in a selected group do not automatically strengthen generalization to everyone enrolled. The additional resolution can improve one evidentiary link while leaving the sampling limitation intact.

The management context in the APKASS statement can help identify a consequential imaging question, but the marker still requires validation for that question. A useful next study would specify the acquisition, reference observation, schedule, and downstream outcome in a defined revision population. It would assess whether the proposed relationship is observed rather than select a favorable marker after examining alternatives. Negative and ambiguous results would remain informative when they show which evidentiary link has not been established.

7. A reusable imaging-evidence record

Verification can be selective even when image acquisition is scheduled. A more invasive reference assessment may be available only for patients undergoing another procedure. Those patients can differ from the people whose images are never verified in that way. Comparing imaging with the available operative findings would then estimate agreement in a selected group. The report should preserve the reason the reference became available and avoid generalizing the resulting accuracy to every imaged patient. A prospective validation design could combine consistent imaging with several complementary reference observations, each with its own selection record. This would not create an infallible reference, but it would make the limits of each comparison visible. It would also prevent an apparently strong validation result from concealing the very clinical process that determined which cases could be checked.

STROBE provides a reporting foundation for the observational components of such a study (von Elm et al., 2007). The imaging record should add acquisition parameters, reader identity, blinded interpretation procedures, adjudication, region definitions, and reasons for missing or additional examinations. The report should state which readers had access to symptoms, previous scans, or treatment allocation. Such information can alter interpretation and should not be left to assumptions about routine practice. The design may legitimately permit some contextual information, but the validity claim must match the information actually available.

The FAIR principles support linking these objects through identifiers and reusable metadata (Wilkinson et al., 2016). An image version, a reader's original judgment, an adjudicated label, and a later derived summary should remain distinguishable. Updating an interpretation should not erase the observation history. Access and consent constraints can be preserved while making the permissible analysis reproducible. The aim is a record that lets another investigator reconstruct the measurement and understand its uncertainty, rather than a folder of images accompanied by an untraceable final healing label.

A patient-reported outcome can anchor the clinical question without being a reference measurement of tissue incorporation. The IKDC subjective knee form supplies a defined knee assessment (Irrgang et al., 2001). A prospective study could investigate its relationship with a prespecified imaging feature while keeping both original measurements. Agreement between the two is not required for either to be useful, and disagreement can identify a limitation in a proposed surrogate interpretation. The study should state whether it asks about description, prognosis, or treatment benefit before using that relationship as evidence for a broader claim.

8. A compact imaging encoder as a separately validated measurement component

Fractal Generative Distillation offers a teacher-student feature-transfer approach (J. Wang et al., 2026). A proposed use in revision imaging would concern a compact encoder, not the biological process itself. The research question would be whether its outputs preserve the distinctions in a predefined measurement task under matched image and reference access. A category reproduced from a teacher remains a model-derived judgment. It should not be relabeled an observed tissue state or used to bypass the reliability and reference-selection questions already required by the imaging framework.

A meaningful test would hold patient groups, acquisition definitions, labels, and downstream scoring fixed while comparing the teacher and student. The evaluation should include uncertain reader cases and incomplete visualization, not only clear examples that are easy to classify. Scanner or site features can be predictive shortcuts, and efficient transfer can preserve those shortcuts as well as useful image information. Performance should therefore be assessed on the acquisition changes relevant to the proposed use. A difference in computing cost would be reported alongside the corresponding change in measurement error and coverage.

The experiment is a proposal rather than a result of this article. It would need a declared training boundary, stored model versions, and an independently assessed evaluation set. Teacher agreement alone would not establish that the imaging marker represents incorporation or that acting on it improves outcomes. The computational comparison can strengthen one measurement link only if its own evidence is collected. It cannot supply the clinical or biological links that remain outside the image task.

9. Conclusion

Imaging contributes most when the observed feature and its evidentiary role are explicit. Acquisition reliability, reader agreement, biological interpretation, and clinical relevance are complementary questions. The framework developed here connects them while preserving their differences. It offers a design for testing whether a proposed marker observes incorporation consistently and meaningfully. It reports no new healing rate or diagnostic accuracy and supplies no individual treatment recommendation. Its conclusion is that a transparent measurement chain can make both positive and negative imaging findings more scientifically useful.

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.

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.

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.

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.

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.

Wilkinson, Mark D., et al. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data*, vol. 3, 2016, article 160018. https://doi.org/10.1038/sdata.2016.18.

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

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.

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.

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.