Prognostic Model Design for Outcomes after Adult ACL Revision

Eric Sanders, Kevin Chapman, Bradley Hughes

Abstract

A prognostic model for revision ACL outcomes needs a defined population, information boundary, time origin, prediction horizon, and observable endpoint. These choices determine whether a fitted score answers the intended research question. This conceptual methods synthesis develops a model-design specification around an adult revision-management consensus. It distinguishes prognosis from treatment-effect estimation, prevents later information from entering a baseline prediction, and connects validation splits to intended transfer. A revision cohort provides a clinical reference for selection and follow-up, while prediction-study guidance organizes bias assessment and reporting. Discrimination, calibration, and decision consequences remain separate claims. No model is trained and no predictive performance is calculated. The contribution is a testable specification identifying the observations required before a graft-related risk estimate can support a transportable prognostic claim.

Keywords: prognosis; graft outcomes; model validation; calibration; data leakage; transportability.

1. Prediction begins with a clinical state

A prognostic model estimates an outcome conditional on information available at a defined point. Without that point, the same variable can be a legitimate predictor in one study and leaked future information in another. A postoperative scan may be appropriate for updating prognosis after the scan but inappropriate for a model claimed to guide the original graft decision. The model specification should therefore begin with the clinical state and information boundary. Choosing an algorithm before those elements are fixed can produce an accurate answer to an unintended question.

Cao and colleagues' APKASS statement identifies management questions in adult ACL revision (Cao et al., 2026). Those questions can motivate a prognostic study without providing patient-level training data or a validated risk score. The present article asks which outcome would matter, what information would be available at the intended prediction point, and where the resulting estimate should transfer. The consensus is a source of context rather than measured model performance. No baseline score or treatment-selection rule is inferred from an expert agreement percentage.

This article presents a conceptual design. It fits no model, creates no synthetic patient cohort, and reports no new discrimination or calibration estimate. Its central distinction is between a reproducible fitted score and a transportable prediction claim. A score can be computed consistently while relying on information unavailable at the intended time or on relationships that change in another setting. The design must make those possibilities visible. A scientifically useful outcome may be a well-characterized limitation of a model rather than a declaration that it is ready for clinical use.

2. Specify the endpoint and horizon

Structural findings, patient-reported limitation, recurrent instability, and additional surgery are different prediction targets. A model predicting a later treatment decision is partly predicting the care system in which that decision occurs. A model predicting an imaging finding depends on the imaging schedule and interpretation procedure. The endpoint should therefore include its ascertainment, not only a label such as graft failure. A study should preserve how ambiguous or incompletely observed cases were handled and should distinguish confirmed absence from lack of follow-up.

The horizon determines which observations can establish the outcome. A fixed-time risk and an event rate over variable follow-up are not equivalent. Participants enrolled late may not yet have reached the intended horizon, while some participants may undergo another procedure that changes the relevant state. The analysis should describe how observation time and competing events enter the target. A binary label assigned without regard to follow-up duration can make a model appear to distinguish risk when it is actually distinguishing who had enough time to experience or detect the event.

The target population should identify revision history and the care setting in which the prediction is intended. Broad eligibility increases relevance while introducing heterogeneity; narrow eligibility can simplify development while limiting applicability. A generic model name should not conceal either choice. Extending a model to another revision population requires examining predictor availability, outcome ascertainment, and the treatment environment. Shared terminology is not evidence that the learned relationship remains stable. A substantive argument and appropriate validation are needed before the claim is broadened.

3. Predictor timing and measurement

A baseline predictor list should include a source and availability timestamp. Laboratory, imaging, operative, and administrative variables can become available at different stages. A field retrospectively populated in a completed record may not have been available when the prediction would have been made. The development dataset should preserve that distinction. If availability cannot be established, the model's intended timing should be narrowed or the variable excluded from that claim. Otherwise the evaluation can reward information access that the proposed use would not permit.

Biological constructs also need measured proxies. A review can suggest a mechanism, but a broad chart label may represent it poorly. A predictor called graft quality should identify the observation that produced it and the person or procedure responsible. Site-specific coding conventions can create artificial patterns that a model exploits. A data dictionary should preserve units, categories, missingness reasons, and changes over time. A clinically interpretable variable name does not guarantee that measurements from different centres are comparable or that the model uses the intended biological information.

A prognostic association is different from treatment-effect modification. A characteristic predicting an unfavorable outcome does not necessarily identify a patient who benefits more from switching graft strategies. The latter question compares outcomes under alternative interventions and requires additional design assumptions. Management questions in the APKASS statement can motivate either research program, but they do not make them interchangeable. A risk score should not become a graft-selection rule merely because the model contains a treatment variable or a biologically interpretable predictor.

4. Development is a complete pipeline

Preprocessing, imputation, variable selection, transformation, and hyperparameter tuning are parts of the fitted procedure. Estimating them using the full dataset before splitting can transfer information from the evaluation observations. The training boundary should apply to the complete pipeline rather than only the final algorithm. This is particularly relevant when repeated observations from one patient share stable features. Randomly splitting visits can allow a model to recognize a familiar patient while being described as evaluated on new patients. The split manifest should make the grouping explicit.

Model-selection bias can arise when performance estimates are reused to choose the final procedure (Varma and Simon, 2006). A nested evaluation separates that choice from assessment of the selected pipeline. The outer split should represent the intended transfer challenge, while the inner process uses only the permitted development observations. Limited sample size can make the procedure variable,

especially for uncommon outcomes. That variability is a result to report. It should not be resolved by repeatedly revising the model against the same final evaluation set until an attractive score appears.

Reporting guidance such as TRIPOD emphasizes transparent specification of prediction-model development and validation (Collins et al., 2015). A graft study should state whether the reported result is apparent performance, internal validation, temporal validation, or evaluation in another centre. These labels describe different information environments. The report should also preserve the number of patients and events supporting each stage. A long feature list and a sophisticated algorithm do not compensate for a sparse outcome or an evaluation sample that contributes little independent variation relevant to the claim.

5. Match validation to intended transfer

Structured cross-validation offers a framework for withholding related observations together (Roberts et al., 2017). A model intended for future patients in the same care system faces a different challenge from one intended for a new centre or a changed practice era. A centre-held-out split tests some aspects of geographical or organizational transfer, while a temporal split can expose changes in treatment and measurement. Neither automatically reproduces every deployment condition. The study should state what novelty the validation introduces and what information remains shared with development.

Risk of bias and applicability need separate assessment. PROBAST provides a framework for evaluating prediction-model studies along those dimensions (Wolff et al., 2019). A technically careful internal validation can still concern a population or outcome unlike the intended application. Conversely, a relevant population can be analyzed with a biased evaluation procedure. The assessment should connect each concern to the model's intended use rather than collapse all domains into a vague quality label. This helps identify whether the next useful step is better methods, new data, a narrower claim, or some combination.

Site variation can affect both calibration and the predictor-outcome relationship. A model can learn a site's treatment availability, imaging schedule, or threshold for additional surgery rather than a stable biological relation. Such information may improve prediction within that system while weakening transfer. Validation should therefore inspect performance by the relevant environments and document differences in measurement. A study should avoid treating a site identifier as a universal solution. It can encode context in development without providing an appropriate value or relationship for a genuinely new site.

6. Discrimination, calibration, and consequence

Discrimination concerns ordering patients by outcome risk, while calibration concerns agreement between predicted probabilities and observed frequencies. Prediction-performance frameworks emphasize evaluating several aspects rather than relying on one summary statistic (Steyerberg et al., 2010). A model can rank patients reasonably while systematically overestimating risk. A graft study should show the outcome prevalence and the range over which predictions are supported. An impressive aggregate metric can conceal poor performance in a clinically important range or a small subgroup with very uncertain estimates.

Calibration methods developed for neural networks offer useful examples of separating confidence from accuracy (Guo et al., 2017). Their applicability to a particular graft-risk model must be evaluated rather than assumed. A calibration plot is itself estimated from finite data and can be unstable when outcomes are rare. The report should preserve uncertainty and avoid interpreting a smooth fitted curve as exact. Recalibration in a new setting also uses new information; its performance should be distinguished from

the performance of the unchanged original model. Both can be informative when their data requirements are explicit.

Statistical uncertainty does not resolve an incorrectly defined target or a biased sample. Heteroskedasticity-consistent inference addresses a variance issue in a specified model (White, 1980). It cannot repair future-information leakage, differential endpoint ascertainment, or a mismatch between development and intended use. The same principle applies to a narrow interval around a performance statistic. More precise evaluation of the wrong task does not establish clinical relevance. A complete interpretation should describe the measurement and design limits alongside numerical uncertainty rather than allowing the latter to stand in for all forms of uncertainty.

7. Deferral and evaluation boundaries

A model may encounter a patient whose measurements or clinical context are outside its development support. Selective prediction offers a framework for examining error and coverage when a system can decline to decide (Geifman and El-Yaniv, 2017). Applying that idea requires a defined action and a meaningful evaluation. Deferral might indicate that an input is missing or that the model has not been validated in the relevant population. It should not be counted automatically as a successful prediction. The resulting workload and the fraction of patients left without a supported estimate belong in the report.

The threshold for acting on a prediction is a separate question from the quality of the probability estimate. It depends on the available decisions, consequences, and patient priorities. A model that predicts an unfavorable outcome under observed care does not automatically identify the action that would reduce that risk. A prospective impact study would need to evaluate the model within a defined decision process. This article proposes no such threshold and no individualized treatment policy. It identifies the evidence required before a risk score can be interpreted beyond prognosis.

8. A reproducible study specification

The proposed record contains the population, time origin, horizon, endpoint ascertainment, predictor availability, preprocessing, model selection, validation split, and complete performance results. It preserves exclusions and missingness without treating them as inconvenient details. Each version of the model should be linked to the data and code used to develop it. This makes a later update distinguishable from a reproduction. An external assessment should also state what local adaptation or recalibration occurred before the reported result, since those steps change the empirical claim.

The management context in the APKASS statement can guide a substantive prediction question while leaving the model design open to empirical evaluation. Its role remains distinct from measured performance. A well-designed study can investigate which predictors retain information under a realistic transfer challenge and where the proposed use exceeds the observations collected. The strongest result may be a clear applicability limit rather than a broadly framed accuracy claim. The present framework supplies that research specification and no fitted prognostic model.

The MARS cohort supplies a clinical reference for investigating graft choice and outcomes within revision ACL reconstruction (MARS Group, 2014). It does not provide a validated risk model merely by containing potential predictors and follow-up. A proposed development study would still need a defined endpoint, information timing, selection process, and evaluation design. This distinction makes it possible to use a clinically relevant cohort responsibly without treating its size or familiarity as a substitute for validation. The model claim should follow the observations actually used and the novelty introduced by the evaluation.

9. Distillation inside a prognostic pipeline

Fractal Generative Distillation motivates examining economical transfer of learned features (J. Wang et al., 2026). In a proposed revision-outcome model, the relevant object could be a compact image representation feeding a separately defined predictor. This is not equivalent to demonstrating prognosis with the cited method. A representation optimized for a teacher task can preserve information unrelated to the intended clinical endpoint. The study would therefore need to test the downstream prediction as well as the fidelity of the intermediate feature transfer, using the same information point and horizon specified for the original model.

Distillation and its tuning belong inside the development boundary. Patient images from the final evaluation setting should not enter teacher adaptation or feature selection unless the task explicitly permits that information. A locally adapted student and an unchanged student would represent different transfer claims. The comparison should preserve discrimination, calibration, coverage, event counts, and resource costs for both. A smaller encoder can maintain average accuracy while changing behavior in a rare or uncertain subgroup, which matters even when the aggregate teacher-student agreement is high.

No teacher, student, or predictor is trained here. The proposed comparison also does not identify who benefits from a different graft strategy. It concerns prediction under a stated observation and care process. Its value would lie in establishing whether a computational reduction preserves that limited predictive claim, and where it fails. The method provides a research option rather than a shortcut around endpoint definition, bias assessment, or external validation.

10. Conclusion

Graft-risk prediction becomes credible through a clear target, a complete information boundary, and validation that resembles the intended setting. Algorithm choice is only one part of that evidence. The framework developed here connects reporting, bias assessment, calibration, and deferral to the underlying clinical measurement problem. It offers a design for future research without claiming a trained model or a new clinical benefit. The result is a more testable account of what a prediction score would need to demonstrate before it supports a transportable prognostic claim.

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.

Collins, Gary S., et al. "Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement." *Annals of Internal Medicine*, vol. 162, no. 1, 2015, pp. 55-63. https://doi.org/10.7326/m14-0697.

Wolff, Robert F., et al. "PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies." *Annals of Internal Medicine*, vol. 170, no. 1, 2019, pp. 51-58. https://doi.org/10.7326/m18-1376.

Steyerberg, Ewout W., et al. "Assessing the Performance of Prediction Models." *Epidemiology*, vol. 21, no. 1, 2010, pp. 128-138. https://doi.org/10.1097/ede.0b013e3181c30fb2.

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.

Varma, Sudhir and Richard Simon. “Bias in error estimation when using cross-validation for model selection.” *BMC Bioinformatics*, vol. 7, no. 1, 2006. https://doi.org/10.1186/1471-2105-7-91.

Roberts, David R., et al. “Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure.” *Ecography*, vol. 40, no. 8, 2017, pp. 913-929. https://doi.org/10.1111/ecog.02881.

Guo, Chuan, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. “On Calibration of Modern Neural Networks.” *Proceedings of the 34th International Conference on Machine Learning*, PMLR, vol. 70, 2017, pp. 1321-1330. https://proceedings.mlr.press/v70/guo17a.html.

Geifman, Yonatan, and Ran El-Yaniv. “Selective Classification for Deep Neural Networks.” *Advances in Neural Information Processing Systems*, vol. 30, 2017. https://arxiv.org/abs/1705.08500.

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.