Interpretable Counterfactual Simulation of Knee Osteoarthritis Progression from qMRI Biomarkers

Gabrielle Hoyer[1,2,3] [0000−0002−1694−1097] and Sharmila Majumdar[1,3] [0000−0002−0201−871X]

[1] Center for Intelligent Imaging, Department of Radiology and Biomedical Imaging, University of California, San Francisco, CA, USA
[2] Department of Bioengineering, University of California, Berkeley, CA, USA
[3] Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, CA, USA
gabbie.hoyer@ucsf.edu

Abstract. Quantitative MRI (qMRI) biomarkers summarize knee joint health, but their value for forecasting progression, and for simulating interventions over time, is not established. Prior work derived a 110-dimensional principal-component (PC) space of knee qMRI biomarkers and reported cross-sectional associations with osteoarthritis (OA) outcomes. We extend that static space into a longitudinal, counterfactual digital twin. It is a recurrent model that rolls the biomarker state forward, predicts event hazard, and simulates the model-implied consequence of altering individual biomarkers or a modifiable clinical factor (body-mass index, BMI). On the publicly available OAI dataset we study two outcomes. For incident OA in initially healthy knees, imaging carries modest prognostic signal (C-index 0.55) that clinical demographics alone do not (0.51), and model-based perturbation responses are directionally consistent with associations from a prior analysis of an overlapping OAI cohort, including a small, correctly signed BMI response in pooled repeated cross-fitting. For total knee replacement (TKR), discrimination appears far higher (0.77), but a stratified analysis shows it is driven by baseline radiographic severity. Within severity strata the imaging signal vanishes. The twin therefore distinguishes genuine biomarker prognosis from a severity-detection confound. This property is essential for trustworthy counterfactual simulation. All results use hardened, leakage-controlled evaluation.

Keywords: Osteoarthritis · Quantitative MRI · Digital twin · Counterfactual simulation · Disease progression.

1 Introduction

Knee osteoarthritis (OA) is a prevalent, progressive joint disease for which quantitative MRI (qMRI) provides rich, noninvasive measures of tissue state, including cartilage T2, cartilage thickness, subchondral bone shape, and meniscal morphology. Two questions follow. Can these biomarkers forecast how a joint will

evolve, and can we simulate the consequence of altering a biomarker or a modifiable risk factor? The first is a prognostic question; the second is counterfactual, and is what would let such a model support reasoning about intervention.

Prior work established a compact representation of knee qMRI by reducing the biomarker panel to a 110-dimensional principal-component (PC) space and reporting cross-sectional associations between these components and OA outcomes [1]. That representation is static, characterizing associations at a single time point without modeling trajectories or simulating interventions over time.

Recurrent survival models and digital twins are established directions in disease progression modeling [4,5,6,12,8]. Existing knee OA progression predictors operate on imaging and clinical data at a single time point [11], and general dynamic survival models such as [12] do not simulate perturbations of the covariate process; perturbation layers on learned simulators are rarely audited for stability or for concordance with prior evidence. The contribution here is not a new architecture. It combines a longitudinal biomarker state space with hardened, leakage-controlled evaluation, and it checks the perturbation layer against prior associations and against a designed severity confound.

We extend this static biomarker space along the temporal and counterfactual axes. Our contributions are threefold. First, we build a hardened recurrent digital twin that predicts the next biomarker state and event hazard from a subject's trajectory. Second, we add a counterfactual layer that simulates biomarker and modifiable-risk-factor interventions, and validate it for concordance against the previously reported cross-sectional associations, including a modifiable body-mass-index (BMI) lever. Third, we use a deliberate contrast between two outcomes, incident OA and total knee replacement (TKR), to show that the framework separates genuine biomarker prognosis from a radiographic-severity confound. The twin is a modest-signal research simulator, so we avoid strong predictions and instead make small, specific claims that outside evidence can corroborate. Formal causal identification of the simulated responses is beyond the present scope and left to future work. Figure 1 shows the framework end to end.

2 Data and Cohorts

We use the publicly available Osteoarthritis Initiative (OAI) dataset and the qMRI-derived PCA biomarkers of [1]. The state is 110-dimensional and spans eleven tissue families (bone shape, T2, and cartilage thickness for femur, patella, and tibia; lateral and medial meniscus), each represented by ten PCs, across up to seven visits spanning nearly a decade.

We study two outcomes with deliberately different cohort definitions. For incident OA (a prognostic task), we select baseline-healthy knees (Kellgren–Lawrence grade KL$\leq$1), $n = 746$ with 357 progression events. By construction this cohort excludes the radiographic-severity shortcut, isolating genuine prognostic signal. For TKR (a severity-mixed task), we admit all baseline grades (KL$\geq$0), $n = 1113$ with 253 events; the baseline-healthy TKR cohort is event-

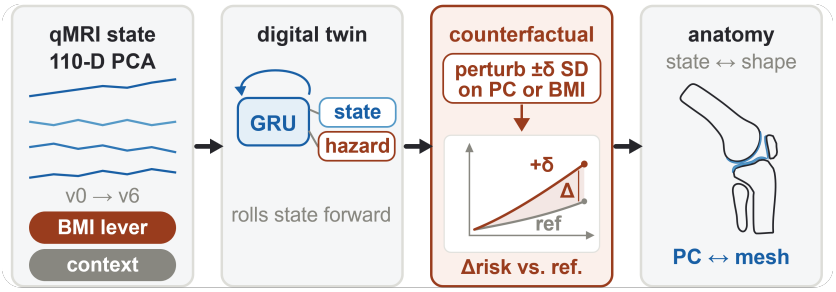

Fig. 1. Overview of the counterfactual qMRI twin, read left to right. The qMRI state is the 110-dimensional PCA representation of a subject's biomarker trajectory across visits v0 to v6, together with the modifiable BMI lever and held context features. A recurrent GRU rolls this state forward and emits two outputs, a next-state prediction and an event hazard. The counterfactual stage perturbs one PC or the BMI lever by $\pm\delta$ SD, re-rolls the twin, and reads off the change in cumulative risk against the unperturbed reference. The inset is schematic and shows no results. Biomarker components map back to knee bone and cartilage shape, linking the twin to the static PCA space of [1].

starved (22 events) and unusable. A prior-surgery exclusion was declared in the cohort configuration but was inert, since the assembled table carries no surgical-history field; the cohorts are therefore defined by the KL criteria alone. We additionally consider clinical features. These are BMI (time-varying, modifiable) and age, sex, and baseline quadriceps lean mass (static context). The muscle measure is quadriceps lean mass, not whole-thigh muscle. Throughout, the KL grade, a labor-intensive expert radiographic read, is used only for cohort selection and as a diagnostic. It is never a model input or a deployable predictor.

3 Method

3.1 Recurrent qMRI twin.

The twin is a gated recurrent unit (GRU) that maps a subject's biomarker-state sequence to (i) a next-state prediction and (ii) a per-interval event hazard. The input state comprises the 110 PCs together with eleven binary indicators encoding informative missingness (whole-family acquisition gaps reach 23.5% for meniscus), giving a 121-dimensional input; the clinical variant additionally rolls BMI forward as a state feature, holds age, sex, and quadriceps lean mass as context, and extends the indicator set with BMI and muscle missingness flags (127-dimensional, comprising the 110 PCs, BMI, three context features, and thirteen indicators). We selected the architecture (hidden size 64, single layer, $\sim$41k parameters) by a capacity sweep under hardened training. Both discrimination and next-state fidelity stayed flat across capacity (a 17$\times$ increase in parameters changed RMSE by $<$2%), so we retained the smallest model.

Training minimizes an equally weighted sum of a mean-squared next-state term over the full input vector and a per-interval binary cross-entropy on the hazard logits, with padded steps masked out; the next-state head is a regression and the hazard head is a discrete-time survival classifier. Folds are split at the subject level (one knee per participant, so no participant spans train and test) and stratified on the event indicator, and each repeat draws a distinct partition seed.

Three hardening choices proved essential; we report them here for reproducibility. First, validation loss bottomed out early and then climbed, leaving final-epoch weights miscalibrated, so we checkpoint on best validation loss. Second, we impute per fold and encode missingness. Imputation and standardization statistics are fit on training subjects only within each fold, which closes a global-imputation leak; the indicators are context-only and never intervened on. Third, we re-confirm the operating point on independent seeds. Models are implemented in PyTorch 2.4.0.

3.2 Counterfactual simulation.

Each simulation shifts one standardized state feature by $\pm\delta$ standard deviations (SD) at baseline and then rolls the twin forward. The reported effect is the change in cumulative risk against the unperturbed reference trajectory. These are model-based perturbation responses of the trained twin; absent a causal identification argument they are not estimates of intervention effects in patients. Context features and missingness indicators are held fixed and guarded against intervention; BMI is the one modifiable clinical lever. To assess specificity we define a null group of features that prior work found unassociated with the outcome. Perturbing standardized inputs out of distribution induces a nonspecific drift, so we do not assume the null group is flat. We plot it explicitly and require genuine anchors to separate from it, holding our claims to ± 1 SD, a clinically plausible magnitude at which the null drift is small. The anchor directions come from prior cross-sectional analyses of the same parent OAI resource with overlapping participants (the exact overlap is not recoverable from the released artifacts), so concordance is an internal consistency check rather than external validation.

3.3 Evaluation.

Table 1 reports seed-confirmed means $\pm$ standard deviation from 5-fold $\times$ 3-repeat cross-validation with pooled out-of-fold predictions, with out-of-fold coverage and the absence of train/test leakage verified on every fold. Figure 2 instead evaluates the persisted stratified holdout of the imaging-only twin, Figure 3 shows a single stratified-holdout instance of the imaging-plus-clinical twin, and the BMI result uses the pooled cross-fitting analysis of Section 4.4. The hazard head outputs a discrete conditional hazard h_t per observed inter-visit interval, with no adjustment for interval length (intervals span 12 or 24 months), and cumulative risk through interval T is $1 - \prod_{t=1}^{T}(1 - h_t)$. The interval bracketing the

Table 1. Discrimination (seed-confirmed C-index ± std) across feature sets and outcomes. KL grade is deliberately not a column: it is an expert radiographic label, not a deployable predictor (see Table 2).

Outcome	imaging-only	clinical-only	imaging + clinical
OA (n=746)	0.549 ± 0.002	0.509 ± 0.004	0.558 ± 0.001
TKR (n=1113)	0.772 ± 0.005	0.583 ± 0.005	0.762 ± 0.002

event time is labeled one; all other observed intervals, including those after an event, are labeled zero, and censoring for non-events is administrative at the end of follow-up. Discrimination and calibration at a horizon include subjects whose status at that horizon is known, meaning an event by the horizon or follow-up beyond it; calibration uses ten equal-count bins of predicted risk with observed rates taken as unadjusted event fractions within bins. Competing events are not modeled; in the OA cohort 22 subjects underwent TKR during follow-up and continue to contribute intervals, and mortality is not recorded in the assembled table.

4 Results

4.1 Twin validity.

On incident OA, the twin discriminates modestly (seed-confirmed cross-validated C-index 0.549, Table 1; the held-out-subject evaluation in Fig. 2 is consistent, with a wide interval), with hazard calibration that tracks the diagonal once best-validation checkpointing is applied, and a next-state RMSE near 0.84 SD averaged over the three imaging feature groups on held-out subjects. This beats a predict-the-mean baseline but still points to limited dynamics. We report this modest performance plainly, as it is the basis for the conservative, specific claims that follow.

4.2 Imaging versus clinical signal.

We write imaging signal for discrimination attributable to the 110 PC features and clinical signal for discrimination attributable to BMI, age, sex, and quadriceps lean mass. Table 1 contrasts feature sets across both outcomes. For incident OA, clinical demographics alone are near chance (0.509), imaging adds +0.050 over them, and clinical adds essentially nothing over imaging (+0.009). In the cohort that structurally excludes the severity shortcut, imaging is what carries the modest prognostic signal. For TKR, imaging alone scores much higher (0.772), and Section 4.3 examines why.

4.3 TKR discrimination is severity detection.

Stratifying TKR by baseline KL (Table 2) reveals that the high number is a confound. Within KL≥2, which contains 91% of events, the twin collapses to

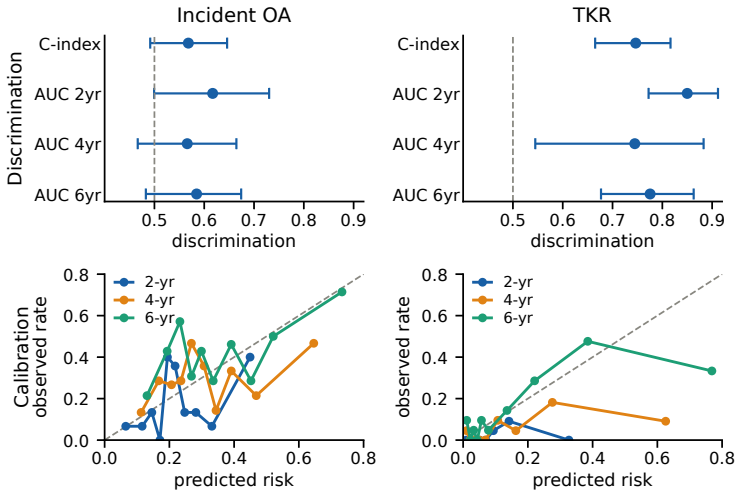

Fig. 2. Twin validity (Incident OA, left; TKR, right), evaluated on the persisted stratified holdout of the imaging-only twin (n=150 OA, n=223 TKR). Top: discrimination (C-index and time-dependent AUC at 2/4/6 years) versus the chance line. Bottom: hazard calibration (predicted versus observed, ten equal-count bins, among subjects with known status at each horizon) at 2/4/6 years. TKR discriminates strongly and calibrates cleanly; OA shows honest, modest discrimination with calibration tracking the diagonal.

0.544, near chance. Baseline radiographic grade alone is highly prognostic for surgery, but it is an expensive expert label rather than a deployable predictor, and we report it solely as a diagnostic. The twin's apparent TKR performance is therefore re-derived radiographic severity, not biomarker prognosis. This contrast is what lends credibility to the OA prognostic claim, since the same framework, applied to a severity-mixed cohort, correctly exposes the absence of fine-grained biomarker signal.

4.4 Counterfactual concordance and the BMI lever.

For incident OA (Fig. 3), protective and risk anchors move risk in the direction of the prior cross-sectional associations, and all five anchors are concordant. At ± 1 SD the protective group is more negative than the null group (-0.045 versus -0.015; paired difference -0.031, subject-level bootstrap interval -0.034 to -0.028 within the holdout instance). As cautioned, the null group drifts at larger magnitudes, so we anchor claims at ± 1 SD and draw the null explicitly. The modifiable BMI lever carries the smallest plotted effect and is the one response we examined for stability across retraining. The single trained instance in Fig. 3 yields a weak, non-monotone BMI perturbation response, with a model-implied cumulative-risk difference of -0.005 at $+1$ SD. Three-repeat five-fold cross-fitting (15 fitted twins), pooled out of fold over all 746 subjects, instead

Table 2. TKR discrimination stratified by baseline KL grade (confound diagnostic). Twin values are pooled out-of-fold cross-validated C-indexes averaged over three repeated partitions of the model in Table 1; the seed-confirmed 0.772 ± 0.005 there averages three further independent partition seeds of the same model. Stratified rows subset the same pooled predictions, and the KL$\leq$1 value rests on 22 events, spanning 0.62 to 0.70 across repeats. KL-alone scores the baseline grade in-sample on the full cohort, with no fitted model and no folds; it is a contextual diagnostic only, not a like-for-like comparator, and its overall value largely reflects the between-stratum contrast that stratification removes.

Stratum	n	events	twin C-index	KL-alone (diagnostic)
Overall	1113	253	0.767	0.863
KL$\leq$1	746	22	0.656 (unmeasurable)	0.546
KL$\geq$2	367	231	0.544 ($\approx$chance)	0.606

yields a monotone response of +0.028 at +1 SD (subject-level bootstrap CI +0.027 to +0.031, conditional on the fitted folds) and +0.096 at +3 SD. This interval does not quantify variability from retraining or alternative fold assignments. For this weak BMI lever the single-instance curve is not a stable summary, so the pooled cross-fitted response is the reported BMI result; the substantive claim is anchored at +1 SD, while +3 SD is a model extrapolation. The pooled response is small and correctly signed. For TKR the anchors are poorly concordant and BMI shows no lever, consistent with a hazard driven by severity rather than by per-feature biomarker structure.

5 Discussion

The central message is a dissociation; discrimination magnitude and counterfactual interpretability are not the same, and they can trade off. A twin can score highly by re-deriving radiographic severity while carrying no modifiable biomarker signal, as TKR shows; conversely, a modest discriminator can carry genuine, interpretable signal when the cohort is designed to exclude the dominant confound, as incident OA shows. Cohort design, here restricting OA analysis to baseline-healthy knees, is what isolates true prognosis, and the counterfactual layer with its modifiable BMI lever is what makes the resulting twin usable for hypothesis-generating reasoning about intervention.

Several limitations are intrinsic to the approach, and we state them plainly. Absolute signal is modest, and the $\sim$0.84 SD next-state fidelity bounds how far autoregressive rollouts can be trusted; out-of-distribution drift limits high-magnitude counterfactual claims, which is why we anchor at $\pm$1 SD. The OA and TKR cohorts differ by construction and are not the same patients; the muscle proxy is a baseline static quadriceps measure, with the time-varying whole-thigh measure excluded for missingness; and we study a single dataset. Uncertainty is reported at the subject level for the BMI response only. Planned extensions add bootstrap intervals across the anchor set, time-dependent Brier scores and calibration slope with numbers at risk, and a quantitative measure of distance from

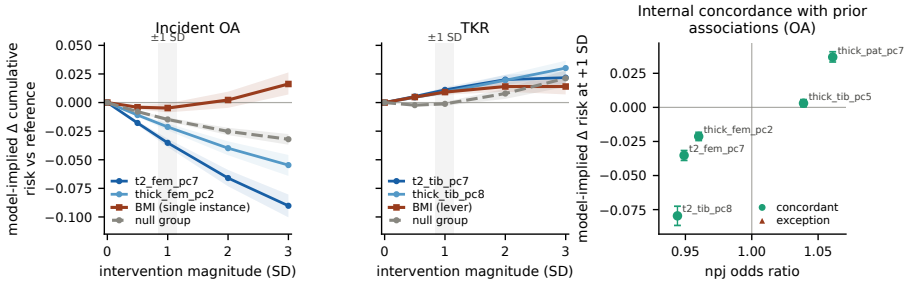

Fig. 3. Model-based perturbation responses and internal concordance with prior associations. (A) Incident OA: imaging-anchor responses separate from the plotted null group in one trained instance; the BMI trace from that instance is weak and non-monotone. (B) TKR: anchors show poor concordance and BMI follows the null in the same single-instance analysis; panels A–B share axes with the ±1 SD region shaded. (C) Model-implied risk differences at +1 SD versus odds ratios from an earlier analysis of the same parent OAI resource with overlapping cohorts, an internal consistency check rather than external validation. Responses in all three panels come from one trained instance; the reported BMI result is the pooled out-of-fold response of Section 4.4.

the training distribution for perturbed states, alongside conditional or anatomically constrained perturbations that respect correlations among components. Future work includes architecture exploration, formal identification analysis asking which perturbation responses are compatible with the data under stated assumptions, and a 3D anatomical-mesh visualization of the counterfactual evolution.

To make the perturbation simulator easier to inspect, we built a browser-based research interface. A user selects a biomarker component or the BMI lever, sets a perturbation magnitude in standard deviations, and the twin re-rolls to show the model-implied change in cumulative risk alongside the affected region on the bone surface; the surface difference is read against the trained twin rather than amplified for display. Supplementary material illustrates the interface with a worked femur cartilage T2 (PC 7) perturbation at three doses, and the results reported here do not depend on it.

6 Conclusion

A hardened longitudinal counterfactual qMRI twin recovers specific, modest biomarker effects for incident knee OA that concord with previously reported associations from the same parent cohort, including a small, correctly signed BMI response in the pooled repeated cross-fitting analysis. Through a deliberate OA-versus-TKR contrast, it distinguishes genuine prognostic signal from a radiographic-severity confound, an essential property for trustworthy counterfactual simulation.

References

1. Hoyer, G., Gao, K.T., Gassert, F.G., et al.: Foundations of a knee joint digital twin from qMRI biomarkers for osteoarthritis and knee replacement. npj Digit. Med. **8**, 118 (2025). https://doi.org/10.1038/s41746-025-01507-3
2. National Institute of Arthritis and Musculoskeletal and Skin Diseases (NIAMS): Osteoarthritis Initiative (OAI): a multi-center, longitudinal, prospective observational study of knee osteoarthritis. U.S. Department of Health and Human Services, National Institutes of Health. https://nda.nih.gov/oai (2024)
3. Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: EMNLP, pp. 1724–1734 (2014)
4. Faraggi, D., Simon, R.: A neural network model for survival data. Stat. Med. **14**(1), 73–82 (1995)
5. Katzman, J.L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., Kluger, Y.: DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Med. Res. Methodol. **18**(1), 24 (2018)
6. Lee, C., Zame, W.R., Yoon, J., van der Schaar, M.: DeepHit: a deep learning approach to survival analysis with competing risks. In: AAAI, pp. 2314–2321 (2018)
7. Harrell, F.E., Califf, R.M., Pryor, D.B., Lee, K.L., Rosati, R.A.: Evaluating the yield of medical tests. JAMA **247**(18), 2543–2546 (1982)
8. Björnsson, B., et al.: Digital twins to personalize medicine. Genome Med. **12**(1), 4 (2020)
9. Pearl, J.: Causality: Models, Reasoning, and Inference. 2nd edn. Cambridge University Press (2009)
10. Kellgren, J.H., Lawrence, J.S.: Radiological assessment of osteo-arthrosis. Ann. Rheum. Dis. **16**(4), 494–502 (1957)
11. Tiulpin, A., Klein, S., Bierma-Zeinstra, S.M.A., Thevenot, J., Rahtu, E., van Meurs, J., Oei, E.H.G., Saarakkala, S.: Multimodal machine learning-based knee osteoarthritis progression prediction from plain radiographs and clinical data. Sci. Rep. **9**, 20038 (2019). https://doi.org/10.1038/s41598-019-56527-3
12. Lee, C., Yoon, J., van der Schaar, M.: Dynamic-DeepHit: a deep learning approach for dynamic survival analysis with competing risks based on longitudinal data. IEEE Trans. Biomed. Eng. **67**(1), 122–133 (2020). https://doi.org/10.1109/TBME.2019.2909027