Osteoarthritis and Cartilage Open 7 (2025) 100688

Contents lists available at ScienceDirect

Osteoarthritis and Cartilage Open

journal homepage: www.elsevier.com/journals/osteoarthritis-and-cartilage-open/2665-9131

Development and validation of a generalisable survival model to predict osteoarthritis progression

H.H.T. Li [a,b], L.C. Chan [a], P.K. Chan [c], C. Wen [a,d,*]

[a] Department of Biomedical Engineering, The Hong Kong Polytechnic University, Hong Kong
[b] Department of Prosthetics and Orthotics, Tuen Mun Hospital, Hong Kong
[c] Department of Orthopaedics and Traumatology, The University of Hong Kong, Hong Kong
[d] Research Institute of Smart Ageing, The Hong Kong Polytechnic University, Hong Kong

ARTICLE INFO

Handling Editor: Professor H Madry

Keywords:
Knee osteoarthritis
Domain shift
Transfer learning
Survival analysis

ABSTRACT

Objective: To develop and validate a transfer-learning survival model to predict progression end-stage knee osteoarthritis (esKOA) or receive knee replacement (KR).
Method: A DeepSurv model was trained on the Osteoarthritis Initiative (OAI) dataset with baseline clinical variables, including age, sex, BMI, comorbidities, smoking status, prior knee injury and surgery, pain medication use, use of walking aids, and activity level (9560 knees). A generalisable model was then developed by fine-tuning the OAI-derived model with data from the Multicenter Osteoarthritis Study (MOST) Centre 1 (3002 knees). This model was validated on an independent dataset from MOST Centre 2 (2972 knees). Model performance was evaluated using the concordance index from 1000 bootstrap resamples. SHapley Additive exPlanations (SHAP) were employed to assess changes in feature importance after fine-tuning.
Results: The OAI-derived model performed well within OAI (C-index = 0.75) but fairly on MOST Centre 1 (C-index = 0.61, $p < 0.0001$), indicating domain shift or cross-cohort variation. Similarly, a model trained only on MOST Centre 1 data performed moderately within MOST (C-index = 0.63) but did not generalise to OAI (C-index = 0.60, $p < 0.0001$). After transfer learning, the generalised model maintained performance on OAI (C-index = 0.69) and improved on MOST Centre 1 (C-index = 0.64, $p < 0.0001$) and MOST Centre 2 (C-index = 0.67, $p < 0.0001$). SHAP analysis revealed that heart attack history, diabetes, smoking, and BMI became more influential predictors in the fine-tuned model.
Conclusion: Transfer learning enabled the development of a generalised model for knee OA prognosis that performs consistently across cohorts. By adapting to population-specific risk patterns, this approach enhances model generalisability and reduces bias from ethnic or demographic overrepresentation in training datasets.

1. Introduction

Accurate prognosis of knee osteoarthritis (OA) is crucial for providing timely interventions that can slow disease progression, alleviate symptoms, and improve patients' quality of life [1,2]. Hence, precise prediction models can help clinicians identify individuals at high risk of rapid progression, enabling personalised treatment strategies. While numerous definitions of knee OA progression exist, opting for a clinically meaningful and standardised endpoint is imperative for model applicability [3,4]. In this study, we concentrate on predicting the time to end-stage knee osteoarthritis (esKOA) or knee replacement (KR) as our endpoint measures, as they represent advanced disease states in which patients experience severe symptoms and functional limitations, often necessitating surgical intervention [5]. By targeting esKOA and KR, we aim to provide clinically relevant predictions to inform patient management decisions.

Several survival models have been developed to predict knee OA progression. Nonetheless, their generalisability across different populations remains under investigation [6–8]. Some models are trained and validated within a single population, limiting generalisability across diverse populations. Therefore, these models can perform well on the datasets they were trained on but may struggle to maintain the same level of accuracy when applied to external cohorts. This limitation may stem from differences in data distributions, namely patient

* Corresponding author. ST417, Department of Biomedical Engineering, Faculty of Engineering, The Hong Kong Polytechnic University, Hong Kong.
E-mail addresses: toby.li@connect.polyu.hk (H.H.T. Li), lc-justin.chan@connect.polyu.hk (L.C. Chan), cpk464@yahoo.com.hk (P.K. Chan), chunyi.wen@polyu.edu.hk (C. Wen).

https://doi.org/10.1016/j.ocarto.2025.100688
Received 4 July 2025; Accepted 21 September 2025

demographics and clinical characteristics across various datasets. For instance, the Osteoarthritis Initiative (OAI) and the Multicenter Osteoarthritis Study (MOST), two large longitudinal cohorts commonly used in OA research, exhibit significant variations in these aspects, namely age distribution, BMI ranges, and prevalence of comorbidities. Given that a model fundamentally reflects the data distribution it was trained on, variations in demographics, BMI ranges, and comorbidity prevalence between OAI and MOST may pose a hurdle to developing predictive models that generalise well across diverse populations.

Transfer learning holds a good promise to improve the generalisability of predictive models in knee OA. By leveraging knowledge gained from the source domain, transfer learning adapts the predictive model to enhance the performance on a different target domain with varying data distributions [9]. This approach preserves the high-level robust features learned from the source domain while fine-tuning the model to capture domain-specific characteristics of the target dataset [10]. In particular, transfer learning can adjust for discrepancies in patient demographics, clinical characteristics, and other risk factor distributions, improving the model's applicability to different populations. Prior research in medical imaging has demonstrated the advantages of transfer learning in adapting models to new datasets with different characteristics [11–13]. However, its application in prognostic modelling for knee OA with structured clinical data is still untapped.

Therefore, the objective of this study is to develop and validate a generalisable survival model for knee OA progression using transfer learning. By applying transfer learning, we aim to enhance predictive performance across different cohorts, overcoming the significant limitation of traditional models, and potentially leading to a more reliable prognostic tool for knee OA management across distinct populations.

2. Method

2.1. Data acquisition, inclusion and exclusion criteria

In this study, both the OAI and MOST datasets were leveraged to develop and validate survival models using DeepSurv to predict the time until progression to esKOA or KR (Fig. 1).

The OAI is a prospective longitudinal cohort study designed to identify risk factors associated with the onset and progression of knee OA. A total of 4796 participants aged 45–79 years were enrolled on the study at baseline, with annual follow-up visits extending up to 96 months. The OAI dataset encompasses comprehensive clinical information such as baseline demographic characteristics, anthropometric measurements, comorbidities, imaging data, and disease outcomes. The OAI is registered at ClinicalTrials.gov (ID: NCT00080171).

The MOST is the other prospective, multicenter cohort study examining risk factors for knee OA development and progression in 3026 older adults aged 50–79 years. Baseline data were collected, with yearly follow-up extending up to 84 months. The MOST dataset comprises data from two clinical sites, referred to as Centre 1 and Centre 2. Centre 1 corresponds to the Alabama site, which shares a higher proportion of Black participants, while Centre 2 refers to the Iowa site, which is predominantly White. Similar to OAI, it contains essential clinical information, including demographic data, anthropometric measurements, comorbidities, imaging findings, and disease outcomes. The MOST is registered at ClinicalTrials.gov (ID: NCT00294581).

Knees that had undergone KR or had reached esKOA at or before the baseline examination were excluded to prevent left-censoring, as including them would invalidate time-to-event estimation. It is important to note that both knees from each individual were included in the analysis, even if one knee had already progressed to esKOA or KR. This inclusion was made to simulate real-world clinical scenarios, where patients could experience asymmetric disease progression and may still benefit from prognostic insights for the contralateral knee. Thus, the model is better positioned to capture the full spectrum of disease risk and improve its practical relevance.

On the other hand, right-censored data were incorporated in the analysis, represented by cases where esKOA and KR had not occurred by the conclusion of the study but may transpire in the future. These right-censored cases provide valuable information regarding the minimum duration of follow-up for individuals who have not yet progressed to esKOA or receive KR.

2.2. Covariates and outcome measures

In this study, both the OAI and MOST datasets were studied to develop and validate survival models using DeepSurv to predict the time until progression to esKOA or KR [14]. DeepSurv is a deep learning-based variant of the Cox Proportional Hazards model. It preserves the multiplicative feature of the hazard function while substituting the linear component with a neural network to model non-linear relationships between covariates, further improving the predictive performance, especially in high-dimensional settings [14]. Hence, in this study, the DeepSurv model was selected to learn a hazard function using baseline clinical variables, which were selected based on a comprehensive review of the knee OA pathophysiological mechanisms from recent literature [1,2,15,16]. It encompassed demographic variables (age, sex, education level), anthropometric measurements (BMI), comorbidities (diabetes, heart attack history, stroke history, and smoking status), and knee-related factors (knee injury history, prior knee surgery, use of pain medication, use of walking aids, and activity level in the last seven days) (Table 1). These covariates were chosen based on their clinical relevance for practical and scalable implementation. Thus, features that are objective, accessible, and require minimal specialised

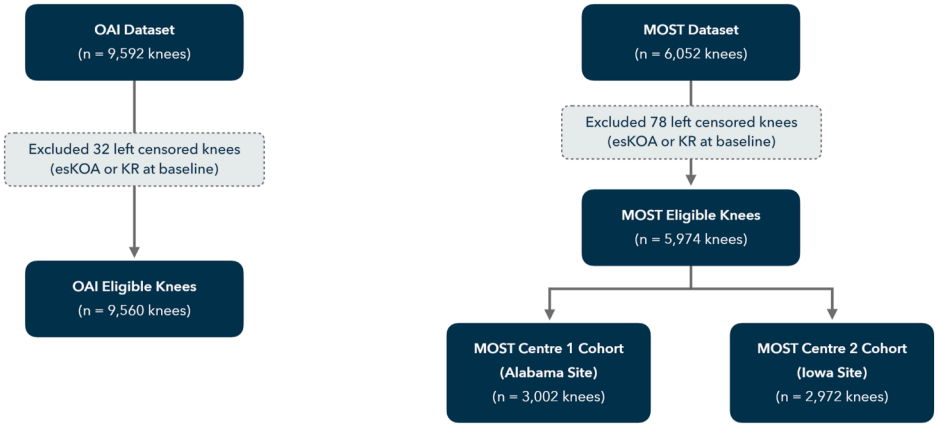

Fig. 1. The schematic on the OAI and MOST datasets utilisation for OAI base model and generalised model development.

Table 1
An overview of the features used in this study.

Features/Risk Factors	OAI Dataset			MOST Dataset			Standardised Categories/Unit
	Variable	Description	Categories/Unit	Variable	Description	Categories/Unit	
Sex	P02SEX	Gender	0: Male 1: Female	SEX	Sex/Gender	0: Female 1: Male	0: Male 1: Female
Education Level	V00EDCV	Highest grade or year of school completed	0: Less than high school graduate 1: High school graduate 2: Some college 3: College graduate 4: Some graduate school 5: Graduate degree	V0EDUC	Highest grade or year of school completed	1: Some elementary school 2: Elementary school (completed grade 8) 3: Some high school 4: High school graduate (completed grade 12) 5: Some college 6: College graduate 7: Some graduate school 8: Graduate degree	1: Less than high school graduate 2: High school graduate 3: Some college 4: College graduate 5: Some graduate school 6: Graduate degree
Activity Level in the Last Seven Days	V00WORKAMT	Occupational activity level, past 7 days	1: Sitting 2: Sitting/standing/ walking 3: Walking/handling <50 lbs 4: Walking/handling >50 lbs	V0WKPA	Type phys activity required at work, PAST 7 D	1: Sitting 2: Sitting/standing/ walking 3: Walking/handling <50 lbs 4: Walking/handling >50 lbs	1: Sitting 2: Sitting/standing/walking 3: Walking/handling <50 lbs 4: Walking/handling >50 lbs
Smoking Habit	V00SMKNOW	Smoke cigarettes now	0: No 1: Yes	V0SKNOW	Smoke now, CURRENT	0: No 1: Yes	0: No 1: Yes
Stroke History	V00STROKE	Had stroke, cerebrovascular accident, blood clot or bleeding in brain, or transient ischemic attack	0: No 1: Yes	V0STROKE	Stroke, cerebrovascular accident, TIA, EVER	0: No 1: Yes 8: Don't know	0: No 1: Yes
Diabetes History	V00DIAB	Have diabetes	0: No 1: Yes	V0DIABT	Diabetes, CURRENT	0: No 1: Yes 8: Don't know	0: No 1: Yes
Heart Attack History	V00HRTAT	Ever had heart attack	0: No 1: Yes	V0HRTAT	Heart attack, EVER	0: No 1: Yes 8: Don't know	0: No 1: Yes
Use of Pain Medication	P01KPMED	Either knee, used medication for pain, aching or stiffness, past 12 months	0: No 1: Yes	V0ARTHRX	Taking selected arthritis meds every or almost every day, CURRENT	0: No 1: Yes 8: Don't know/ Refused	0: No 1: Yes
Knee Injury History	P01INJR (right); P01INJL (left)	Ever injured badly enough to limit ability to walk for at least two days	0: No 1: Yes	V0LAR (right); V0LAL (left)	Injury limited ability to walk at least 2 days	0: No 1: Yes 8: Don't know/ Refused	0: No 1: Yes
Knee Surgery History	P01KSURGR (right); P01KSURGL (left)	Ever have surgery or arthroscopy	0: No 1: Yes	V0SURGR (right); V0SURGL (left)	Knee surgery, EVER	0: No 1: Yes 8: Don't know/ Refused	0: No 1: Yes
Use of Walking Aids	V00WLKAID	Using walking aid such as cane	0: No 1: Yes	V0AID	Completed walk using a walking aid	0: No 1: Yes 8: Don't know/ Refused	0: No 1: Yes
Body Mass Index	P01BMI	Body mass index	kg/m^2	V0BMI	Body Mass Index (BMI) CURRENT, kg/m**2	kg/m^2	kg/m^2
Age	V00AGE	Age	Year	AGE	Age at baseline	Year	Year

assessment were prioritised, supporting the potential for community-wide, low-cost OA risk screening.

The prediction outcome was defined as the time to progression to esKOA or undergoing KR. EsKOA was defined as reaching a Kellgren-Lawrence (KL) grade 4 with moderate-to-intense symptoms, indicated by Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) pain and function scores greater than 11 [17]. This composite endpoint captures patients who do not opt for KR despite severe symptoms, as well as those who receive KR earlier due to debilitating knee pain [18,19].

2.3. Selection of source domain, adaptation, and independent validation sets

The larger OAI cohort served as the source domain for learning knee OA-related patterns to develop the OAI base model. Recognising potential domain-specific variations due to demographic and lifestyle differences between cohorts, transfer learning was employed to adapt a pre-trained OAI base model to MOST data. Since the MOST data originated from two centres, i.e. Centre 1 and Centre 2, statistical comparisons were conducted to compute the domain mismatch relative to the OAI (Tables 2 and 3). MOST Centre 1 demonstrated greater shifts in data distribution from the OAI and was therefore chosen as the adaptation cohort to address substantial distributional differences. MOST Centre 2 was held out as an independent test set for validation across datasets and isolated from any training or fine-tuning processes to provide an unbiased evaluation of predictive performance. The two MOST centres were intentionally treated separately rather than combining them to retain distinct population-level characteristics, given that previous studies have shown that factors such as race, income, and familiarity with surgery significantly alter the likelihood of undergoing KR [20,21]. Combining the centres for the analysis could underestimate these differences, hence limiting our ability to assess the model generalisability between clinically meaningful population subgroups.

2.4. Data harmonisation and missing data handling

To enable transfer learning between OAI and MOST, predictors were harmonised. Each candidate variable's question text and response options were reviewed, and then the original categories were mapped to a shared encoding. Most categorical variables had equivalent response options and were directly mapped. However, when options differed, they were collapsed to the coarsest common categorisation. For instance, education was aligned by collapsing MOST's elementary and some high school categories into “less than high school graduate,” resulting in a six-level scale consistent with OAI. Missing, unknown, or refused responses were standardised to a common code. Continuous variables such as age and BMI were aligned in units. Regarding variable levels, both person-level (e.g., age, sex, BMI) and knee-level variables (e.g., pain, injury, surgery) were included in the modelling, with the unit of analysis determined at the knee level. Thus, person-level variables were replicated across both knees for each individual to facilitate knee-level prediction while maintaining consistency. Missing data in both OAI and MOST were addressed using multiple imputation by chained equations to handle incomplete cases [22].

2.5. Initial model establishment and testing

To systematically evaluate baseline performance and quantify the impact of domain differences before introducing transfer learning, both the OAI dataset and the identified MOST Centre 1 adaptation cohort were randomly partitioned into training (80 %) and testing (20 %) subsets. This configuration produced four distinct training and evaluation pathways.

First, the OAI base model was trained exclusively on the 80 % OAI training subset and was assessed on its corresponding 20 % OAI test split to obtain the in-domain baseline concordance index (C-index) and average Area under the Receiver Operating Characteristic Curve (AUC). Second, the same OAI base model was tested on the MOST Centre 1

Table 2
Baseline characteristics of the OAI, MOST Centres 1 and 2 datasets.

Features/Risk Factors	Categories	OAI (n = 9560 knees)	MOST Centre 1 (n = 3002 knees)	MOST Centre 2 (n = 2972 knees)
Sex	0: Male	3975 (41.6 %)	1223 (40.7 %)	1167 (39.3 %)
	1: Female	5585 (58.4 %)	1779 (59.3 %)	1805 (60.7 %)
Age, years	Mean (Standard deviation)	61.1 (9.19)	62.3 (8.24)	62.6 (7.97)
Education Level	1: Less than high school graduate	334 (3.52 %)	193 (6.4 %)	68 (2.3 %)
	2: High school graduate	1210 (12.8 %)	694 (23.1 %)	781 (26.3 %)
	3: Some college	2281 (24.1 %)	884 (29.4 %)	714 (24.0 %)
	4: College graduate	1998 (21.1 %)	544 (18.1 %)	602 (20.3 %)
	5: Some graduate school	791 (8.3 %)	221 (7.4 %)	256 (8.6 %)
	6: Graduate degree	2864 (30.2 %)	466 (15.5 %)	551 (18.5 %)
Body Mass Index (BMI), kg/m^2	Mean (Standard deviation)	28.6 (4.83)	31.0 (6.17)	30.4 (5.72)
Activity Level in the Last Seven Days	1: Sitting	1904 (28.6 %)	463 (23.8 %)	450 (20.2 %)
	2: Sitting/standing/walking	2755 (41.4 %)	857 (44.1 %)	927 (41.6 %)
	3: Walking/handling <50 lbs	1735 (26.1 %)	534 (27.5 %)	699 (31.4 %)
	4: Walking/handling >50 lbs	258 (3.9 %)	90 (4.6 %)	152 (6.8 %)
Smoking Habit	0: No	3806 (85.5 %)	1889 (62.9 %)	1163 (39.1 %)
	1: Yes	646 (14.5 %)	1113 (37.1 %)	1809 (60.9 %)
Stroke History	0: No	9100 (97.1 %)	2816 (93.8 %)	2809 (94.5 %)
	1: Yes	276 (2.9 %)	186 (6.2 %)	163 (5.5 %)
Diabetes History	0: No	8596 (92.3 %)	2488 (82.9 %)	2710 (91.2 %)
	1: Yes	719 (7.7 %)	514 (17.1 %)	262 (8.8 %)
Heart Attack History	0: No	9146 (98.0 %)	2785 (92.8 %)	2857 (96.1 %)
	1: Yes	188 (2.0 %)	217 (7.2 %)	115 (3.9 %)
Use of Pain Medication	0: No	4280 (44.8 %)	695 (23.2 %)	728 (24.5 %)
	1: Yes	5272 (55.2 %)	2307 (76.8 %)	2244 (75.5 %)
Knee Injury History	0: No	6901 (72.9 %)	2227 (74.2 %)	2203 (74.1 %)
	1: Yes	2568 (27.1 %)	775 (25.8 %)	769 (25.9 %)
Knee Surgery History	0: No	8314 (87.1 %)	2608 (86.9 %)	2647 (89.1 %)
	1: Yes	1229 (12.9 %)	394 (13.1 %)	325 (10.9 %)
Use of Walking Aids	0: No	9427 (98.6 %)	2923 (97.4 %)	2927 (98.5 %)
	1: Yes	133 (1.4 %)	79 (2.6 %)	45 (1.5 %)

Table 3
Statistical comparisons between OAI and MOST Centres with effect sizes.

Features/Risk Factors	OAI vs. MOST Centre 1 *p*-value	OAI vs. MOST Centre 1 Effect Size	OAI vs. MOST Centre 2 *p*-value	OAI vs. MOST Centre 2 Effect Size
Sex	0.427	V = 0.01	<0.05	V = 0.02
Age, years	<0.001	Cohen's d = −0.15	<0.001	Cohen's d = −0.19
Education Level	<0.001	V = 0.19	<0.001	V = 0.18
Body Mass Index (BMI), kg/m2	<0.001	Cohen's d = −0.47	<0.001	Cohen's d = −0.32
Activity Level in the Last Seven Days	<0.001	V = 0.05	<0.001	V = 0.10
Smoking Habit	<0.001	V = 0.50	<0.001	V = 0.48
Stroke History	<0.001	V = 0.07	<0.001	V = 0.06
Diabetes History	<0.001	V = 0.13	0.060	V = 0.02
Heart Attack History	<0.001	V = 0.13	<0.001	V = 0.05
Use of Pain Medication	<0.001	V = 0.19	<0.001	V = 0.18
Knee Injury History	0.167	V = 0.01	0.189	V = 0.01
Knee Surgery History	0.750	V = 0.00	<0.01	V = 0.02
Use of Walking Aids	<0.001	V = 0.04	0.685	V = 0.00

Note: Continuous variables were compared using the *t*-test (effect size: Cohen's d); Categorical variables were compared using the Chi-square test (effect size: Cramér's V).

cohort's 20 % test set, in order to illustrate any performance decline owing to domain mismatch. Third, a model was trained from scratch with the 80 % MOST Centre 1 train split and then evaluated on the OAI test set to assess how an entirely MOST-trained model generalises back to the OAI domain. Finally, the aforementioned MOST Centre 1-based model was also evaluated on its own 20 % in-domain test partition, offering a benchmark for the adaptation cohort itself. In short, these four evaluations resulted in a preliminary yet comprehensive view of cross-domain performance disparity (Fig. 2).

The DeepSurv [23] model comprises a multi-layer perceptron module as the feature extractor and a linear survival head for final prediction. The multi-layer perceptron module is composed of 3 layers with 170 hidden units and ReLU [24] as the non-linear activation function. To facilitate regularisation and reduce overfitting risk, a dropout rate of 0.2 was employed in each layer. Subsequently, the linear survival head projects the 170-dimensional feature vectors into the survival risk score. The models were trained using a partial log likelihood function as the loss function with Adam optimiser [25], 1e-4 learning rate, 1e-5 weight decay, and maximum training epochs of 100 at a batch size of 64. All model and training hyperparameters were optimised using a random search approach with 5-fold cross-validation in the train set, the selected hyperparameters and parameter search space are recorded in Supplementary Table 1. The deep learning models were developed using Pytorch v1.10.1 and scorch v0.13.0 libraries in Python 3.7.6.

2.6. *Generalised model establishment and testing*

A "generalised model" was developed by fine-tuning the OAI base model with the complete 80 % training subset of the MOST Centre 1 cohort, a process involving the preservation of the higher-level features learned from OAI but allowing the adjustment of final-layer parameters to the MOST Centre 1 cohort. The purpose of this fine-tuning process was to produce a model that could retain strong performance on the source domain (OAI) while improving prediction accuracy for the target domain (MOST Centre 1).

Afterwards, the generalised model underwent a holistic evaluation on three distinct test sets to assess its robustness and generalisability across various domains. Firstly, the generalised model was tested on the 20 % OAI test split to verify whether the model retained its predictive accuracy within the original OAI domain after fine-tuning. This evaluation step was essential to detect any potential degradation in performance, commonly referred to as catastrophic forgetting, that might have resulted from the fine-tuning adjustments made with MOST Centre 1 data [26].

Secondly, the generalised model was assessed on the 20 % testing subset of the MOST Centre 1 cohort. This aimed to measure the effectiveness of the generalised model in adapting to the domain-specific characteristics of the MOST Centre 1 data.

Thirdly, to investigate the model's ability to generalise to entirely new data, the fine-tuned model was evaluated on the MOST Centre 2 data. This dataset had been completely excluded from any training or fine-tuning stages, ensuring that the evaluation represented the true out-of-sample performance.

The generalised model employs the same network architecture as the initial model. However, to perform supervised finetuning on the other dataset, a smaller learning rate of 5e-6 was chosen. During the fine-tuning procedure, the model was first trained for 10 epochs with the weights in the feature extraction module frozen, adjusting only the survival prediction head. Later, all learnable weights were unfrozen, and

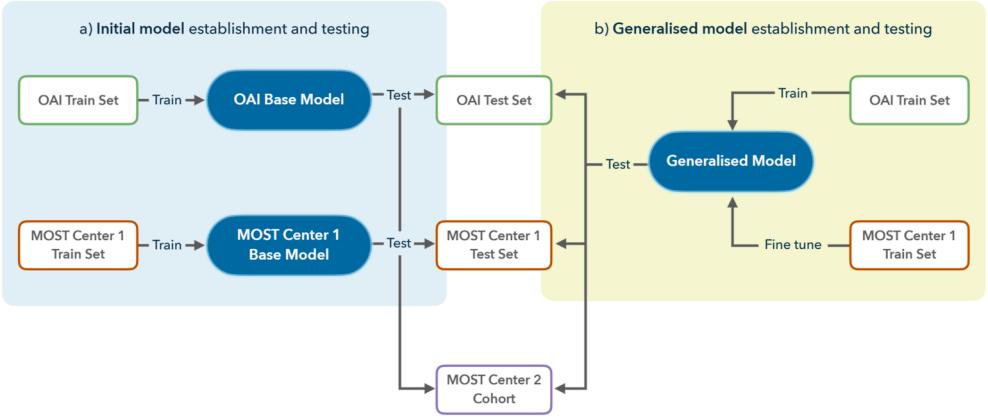

Fig. 2. Workflow of transfer learning and model evaluation.

the model was trained for an additional 10 epochs for final enhancement. Similar to the previous section, the model training hyperparameters were optimised using random search with 5-fold cross-validation in the train set, the selected hyperparameters and parameter search space are reported in Supplementary Table 1.

2.7. Feature importance analysis

To understand how transfer learning can modify the importance of various predictors, SHapley Additive exPlanations (SHAP) analysis was conducted. For both the baseline OAI model and the generalised model, SHAP values were computed for each covariate when applied to the independent MOST Centre 2 cohort to indicate each feature's contribution to the survival prediction. The normalised absolute SHapley Additive exPlanations values were then visualised in a bubble plot, indicating shifts in the magnitude of feature importance resulting from the adaptation procedures. This enabled the identification of variables that gain or lose emphasis under fine-tuning, demonstrating how the OAI model's perspective adjusted to MOST's local context.

2.8. Statistical analysis

Baseline characteristics were summarised as mean and standard deviation for continuous variables and as counts with percentages for categorical variables from the OAI and MOST cohorts. Differences between each cohort were assessed using appropriate statistical test. For continuous variables such as age and BMI, t-tests were performed to compare OAI against each MOST centre. The results were then followed by calculations of Cohen's d to compute the magnitude of the mean differences. Similarly, chi-square tests were conducted to identify significant differences in distributions compared to OAI for categorical variables, namely sex, smoking status, and various comorbidities. The size of these differences was then quantified with Cramér's V.

To ensure the data structure was preserved after the multiple imputation by chained equations implementation for missing values and the random train-test partitioning, T-tests and Chi-square tests were used to verify no significant differences before and after imputation ($p > 0.05$) for the integrity of the dataset.

Hyperparameters of all models were optimised by grid-search through 5-fold cross-validation, maximising the C-index. Model performance was evaluated using the C-index and average AUC to gauge the predictive accuracy of the survival models for time-to-event outcomes. Each model was then subjected to 1000 iterations of bootstrap sampling with replacement from the test split to estimate the distributions of the C-index and average AUC for each model. The resulting distributions were used to assess the statistical significance of performance differences with t-tests. Differences in model performance were considered statistically significant for $p < 0.05$ between models.

3. Results

3.1. Baseline model performance on OAI and MOST cohorts before adaptation

After applying the exclusion criteria, 32 knees were removed from the OAI dataset and 78 knees from the MOST dataset due to missing outcome or covariate data. This resulted in 9560 knees in the final OAI cohort and 6002 knees in the MOST cohort. It is noted that domain shifts between datasets affected model performance across cohorts. The OAI base model yielded a C-index of 0.75 (±0.022) on the OAI test set (Table 4). However, when this initial model was evaluated directly to the MOST Centre 1 dataset, its performance dropped to a C-index of 0.61 (±0.029) ($p < 0.0001$). A further decrease was observed on MOST Centre 2, held out for independent validation, where the C-index declined to 0.60 (±0.027) ($p < 0.0001$). These results highlighted the challenges of cross-domain generalisability when applying a model trained on one population directly to another without any adjustment.

On the other hand, the model trained exclusively on MOST Centre 1 data achieved a moderate C-index of 0.63 (±0.023) on its own validation set. Similarly, this MOST-trained model performed poorly on the OAI test set and the independent MOST Centre 2 subset, with the C-index dropping to 0.60 (±0.025) and 0.60 (±0.023) ($p < 0.0001$), respectively, indicating its limited ability to generalise back to the OAI and MOST Centre 2 domains.

3.2. Improved predictive accuracy across cohorts with the generalised model

Transfer learning effectively improved model performance across different datasets. By fine-tuning the OAI base model using MOST Centre 1 data, the generalised model maintained a reasonable level of predictive accuracy on the OAI test set, achieving a C-index of 0.69 (±0.026). Despite the decrease compared to the original OAI model's in-domain performance, it indicated that the majority of the predictive power was retained post fine-tuning.

But most importantly, the generalised model exhibited improved performance on the MOST datasets. On MOST Centre 1, the model achieved a C-index of 0.64 (±0.026), surpassing both the OAI base model and the MOST Centre 1-trained model within this domain ($p < 0.0001$). The most notable improvement was observed on the independent MOST Centre 2 subset, where the C-index rose to 0.67 (±0.022) from the generalised model. This represented the highest performance among all models evaluated on MOST Centre 2 ($p < 0.0001$), implying that the fine-tuning process effectively adapted the model to capture domain-specific characteristics without overfitting.

3.3. Comorbidities emerge as stronger predictors after transfer learning

The shifts in feature importance emphasised comorbidities as key predictors in the fine-tuned model when evaluated on the independent MOST Centre 2 cohort (Fig. 3). After fine-tuning, SHAP analysis revealed increased importance of features such as heart attack history, diabetes history, smoking habit, and BMI. The emphasis on broad demographic characteristics or generalised treatment patterns, such as age, sex, prior knee surgery, and use of pain medication generally reduced. Meanwhile, the feature importance regarding the use of walking aids and activity levels remained largely unchanged. These changes demonstrated a shift in the feature contributions following fine-tuning.

4. Discussion

In this study, a generalisable survival model was developed and validated to predict the time to esKOA or KR with transfer learning. By

Table 4
Summary of model performance (C-index, average AUC, and standard deviation) under different training and testing scenarios.

Training Set	Testing Set					
	OAI		MOST Centre 1		MOST Centre 2	
	C-index	Average AUC	C-index	Average AUC	C-index	Average AUC
OAI	0.75 (0.022)	0.73 (0.027)	0.61 (0.029)	0.62 (0.040)	0.60 (0.027)	0.61 (0.043)
MOST Centre 1	0.60 (0.025)	0.62 (0.031)	0.63 (0.023)	0.61 (0.040)	0.60 (0.023)	0.61 (0.036)
Generalised Model (OAI model fine-tuned on MOST Centre 1 data)	**0.69 (0.026)**	**0.73 (0.027)**	**0.64 (0.026)**	**0.68 (0.044)**	**0.67 (0.022)**	**0.69 (0.044)**

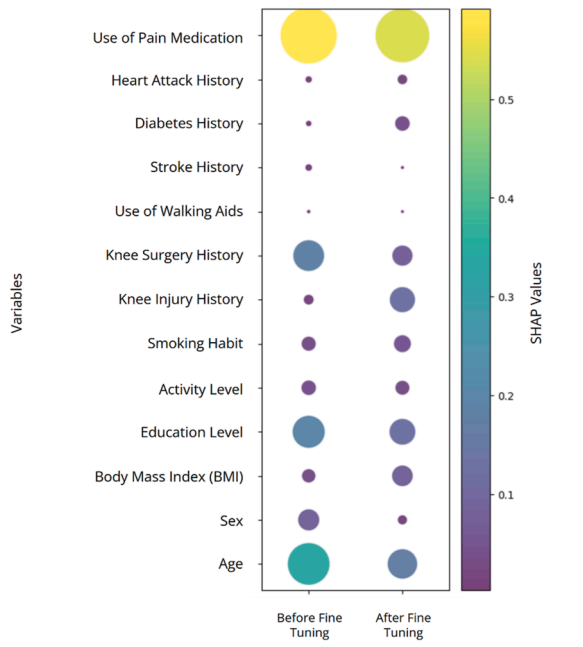

Fig. 3. The SHapley Additive exPlanations bubble plot illustrating feature importance before (left) and after (right) fine-tuning.

fine-tuning a DeepSurv model initially trained on the OAI dataset with data from MOST Centre 1, a balanced predictive performance was attained across different cohorts. The generalised model demonstrated improved C-index on the MOST Centre 1 and MOST Centre 2 datasets, from 0.61 to 0.64 and from 0.60 to 0.67, respectively, while retaining a modest level of predictive accuracy on the OAI dataset (decreasing from 0.75 to 0.69). SHapley Additive exPlanations analysis revealed shifts in feature importance after fine-tuning, with increased emphasis on factors such as heart attack history, diabetes, smoking habit, and body mass index (BMI). These changes aligned with the higher prevalence of these risk factors among subjects who progressed to esKOA or underwent KR in the MOST cohort.

Our findings suggest that transfer learning can effectively mitigate domain shifts between cohorts, enhancing model applicability in diverse clinical settings [9,27]. The fine-tuning process allowed the model to adjust to subtle demographic and clinical differences without significant degradation of performance on the original dataset, minimising the risk of catastrophic forgetting [28,29]. This approach holds promise for broader clinical applications, where models trained on one population can be adapted for use in another with differing characteristics.

Notably, the increased importance of comorbidities such as heart attack history and diabetes after fine-tuning reflects their stronger association with negative outcomes in the MOST population. For instance, among progressing subjects, the prevalence of heart attack history and diabetes history increased in MOST Centre 1 compared to OAI. The mean BMI was also higher in the MOST Centre 1 cohort (Table 5). These shifts highlight the model's ability to adapt to population-specific patterns in survival prediction with a transfer learning strategy.

In spite of these promising results, several limitations in this study warrant consideration. First, both the OAI and MOST datasets predominantly include older adults from the United States, which may limit the generalisability of our findings to younger populations or other ethnic and geographic groups, given that genetic predispositions and lifestyle differences can influence the prevalence and progression of knee OA. Interestingly, even within these US-based cohorts largely consisting of white participants, meaningful differences in demographic and clinical characteristics between MOST Centre 1 and Centre 2 were observed. This suggests that significant heterogeneity can exist even within seemingly similar populations, reinforcing the need for an adaptable model. It is therefore reasonable to expect even greater potential challenges when implementing predictive models in ethnically diverse settings. For example, studies have shown that older Chinese women have a higher prevalence of both radiographic and symptomatic knee OA

Table 5
Prevalence of risk factors among esKOA/KR + subjects in OAI and MOST Centre 1.

Features/Risk Factors	Categories	OAI (n = 440 knees)	MOST Centre 1 (n = 477 knees)
Sex	0: Male	171 (38.9 %)	137 (28.7 %)
	1: Female	269 (61.1 %)	340 (71.3 %)
Age, years	Mean (Standard deviation)	63.8 (8.10)	64.4 (7.42)
Education Level	1: Less than high school graduate	7 (1.6 %)	39 (8.2 %)
	2: High school graduate	67 (15.2 %)	121 (25.4 %)
	3: Some college	117 (26.6 %)	138 (28.9 %)
	4: College graduate	89 (20.2 %)	79 (16.6 %)
	5: Some graduate school	41 (9.3 %)	25 (5.2 %)
	6: Graduate degree	119 (27.0 %)	75 (15.7 %)
Body Mass Index (BMI), kg/m^2	Mean (Standard deviation)	29.8 (4.69)	33.2 (7.23)
Activity Level in the Last Seven Days	1: Sitting	69 (15.7 %)	77 (16.1 %)
	2: Sitting/standing/walking	281 (63.9 %)	327 (68.6 %)
	3: Walking/handling <50 lbs	80 (18.2 %)	66 (13.8 %)
	4: Walking/handling >50 lbs	10 (2.3 %)	7 (1.5 %)
Smoking Habit	0: No	420 (95.5 %)	445 (93.3 %)
	1: Yes	20 (4.5 %)	32 (6.7 %)
Stroke History	0: No	422 (95.9 %)	450 (94.3 %)
	1: Yes	18 (4.1 %)	27 (5.7 %)
Diabetes History	0: No	398 (90.5 %)	414 (86.8 %)
	1: Yes	42 (9.5 %)	63 (13.2 %)
Heart Attack History	0: No	428 (97.3 %)	456 (95.6 %)
	1: Yes	12 (2.7 %)	21 (4.4 %)
Use of Pain Medication	0: No	367 (83.4 %)	362 (75.9 %)
	1: Yes	73 (16.6 %)	115 (24.1 %)
Knee Injury History	0: No	269 (61.1 %)	315 (66.0 %)
	1: Yes	171 (38.9 %)	162 (34.0 %)
Knee Surgery History	0: No	300 (68.2 %)	324 (67.9 %)
	1: Yes	140 (31.8 %)	153 (32.1 %)
Use of Walking Aids	0: No	435 (98.9 %)	467 (97.9 %)
	1: Yes	5 (1.1 %)	10 (2.1 %)

compared to their White counterparts in the US, implying that genetic factors may contribute to these disparities [30]. Therefore, our US-trained model may not perform as well in populations with dissimilar ethnic backgrounds without additional adaptation and validation.

Second, the challenge of data harmonisation can still persist, particularly when integrating unstructured clinical data. Thus, establishing consistent data standards is vital for robust model adaptation in the future. Overcoming these barriers is paramount for the clinical implementation of such models on a global scale. Particularly, as models transition from curated datasets to real-world workflows, we should expect greater heterogeneity and messy inputs, such as missing fields, artifacts and out-of-distribution cases. To handle this, future work shall incorporate strategies such as test-time training [31,32], which allows the model to adapt to the data distribution shift on-the-fly in an unsupervised manner.

Third, our model utilised only routine clinical variables. While practical for routine operation and implementation, it may not capture all relevant factors influencing knee OA progression. Recruiting additional data types, such as biomarkers, medical images, or biomechanical data, could enhance predictive performance [33–35]. Future studies should explore the integration of these factors to develop more comprehensive and robust models.

While the predictive performance of the generalised model remains modest, the value of this study lies in its methodological contribution toward improving model generalisability with transfer learning. Our results show that transfer learning can mitigate performance degradation when a model encounters new populations with distinct demographic and clinical characteristics, which is relevant given the real-world challenge of deploying machine learning models across diverse health systems and populations. The ability to adapt existing models without retraining from scratch could contribute to the scalability and sustainability of predictive modelling frameworks, particularly in resource-limited settings.

In summary, this study highlights the potential of transfer learning to improve the generalisability of survival models for forecasting knee OA progression across distinct cohorts. By fine-tuning a DeepSurv model trained on the OAI dataset using data from MOST Centre 1, the generalised model yielded improved predictive performance on the independent cohort from MOST Centre 2. In addition, the study underscores how transfer learning can adapt to domain-specific risk patterns without the need for complete retraining. While the model's predictive accuracy remains modest, the methodological approach provides a scalable and adaptable framework for improving model robustness across distinct populations. Future efforts should focus on integrating additional data modalities and validating this approach in more diverse settings to advance toward more robust OA prognostic tools.

Author contributions

HL, LC, PC and CW conceived the study. HL and LC collected data. All authors contributed to the writing of the manuscript and approved the final version.

Role of the funding source

This work is supported by the Hong Kong Polytechnic University Shen Zhen Research Institute “百城百園” scheme, Research Institute of Smart Ageing, and Hong Kong Innovation Technology Fund for Better Living (ITB/FBL/B046/22/S).

Declaration of competing interest

The authors have no relevant competing interests to disclose.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ocarto.2025.100688.

References

[1] A. Mahmoudian, et al., Early-stage symptomatic osteoarthritis of the knee—time for action, Nat. Rev. Rheumatol. (2021) 1–12.
[2] E.M. Roos, N.K. Arden, Strategies for the prevention of knee osteoarthritis, Nat. Rev. Rheumatol. 12 (2) (2016) 92–101.
[3] C. Kokkotis, et al., Machine learning in knee osteoarthritis: a review, Osteoarthr. Cartil. Open (2020) 100069.
[4] A. Jamshidi, J.-P. Pelletier, J. Martel-Pelletier, Machine-learning-based patient-specific prediction models for knee osteoarthritis, Nat. Rev. Rheumatol. 15 (1) (2019) 49–60.
[5] G.A. Hawker, Who, when, and why total joint replacement surgery? The patient's perspective, Curr. Opin. Rheumatol. 18 (5) (2006) 526–530.
[6] A. Jamshidi, et al., Machine learning–based individualized survival prediction model for total knee replacement in osteoarthritis: data from the osteoarthritis initiative, Arthr. Care Res. 73 (10) (2021) 1518–1527.
[7] Q. Liu, et al., Prediction models for the risk of total knee replacement: development and validation using data from multicentre cohort studies, Lancet Rheumatol. 4 (2) (2022) e125–e134.
[8] H.H.T. Li, et al., An interpretable knee replacement risk assessment system for osteoarthritis patients, Osteoarthritis and Cartilage Open 6 (2) (2024) 100440.
[9] S.J. Pan, Q. Yang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng. 22 (10) (2009) 1345–1359.
[10] G. Vrbančič, V. Podgorelec, Transfer learning with adaptive fine-tuning, IEEE Access 8 (2020) 196197–196211.
[11] A.W. Salehi, et al., A study of CNN and transfer learning in medical imaging: advantages, challenges, future scope, Sustainability 15 (7) (2023) 5930.
[12] P. Kora, et al., Transfer learning techniques for medical image analysis: a review, Biocybern. Biomed. Eng. 42 (1) (2022) 79–107.
[13] P.S.Q. Yeoh, et al., Transfer learning-assisted 3D deep learning models for knee osteoarthritis detection: data from the osteoarthritis initiative, Front. Bioeng. Biotechnol. 11 (2023) 1164655.
[14] J.L. Katzman, et al., DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network, BMC Med. Res. Methodol. 18 (2018) 1–12.
[15] J.L. Bowden, et al., Core and adjunctive interventions for osteoarthritis: efficacy and models for implementation, Nat. Rev. Rheumatol. (2020) 1–14.
[16] H.A. Wieland, et al., Osteoarthritis—An untreatable disease? Nat. Rev. Drug Discov. 4 (4) (2005) 331–344.
[17] J.B. Driban, et al., Defining and evaluating a novel outcome measure representing end-stage knee osteoarthritis: data from the osteoarthritis initiative, Clin. Rheumatol. 35 (2016) 2523–2530.
[18] G.A. Hawker, et al., Determining the need for hip and knee arthroplasty: the role of clinical severity and patients' preferences, Med. Care 39 (3) (2001) 206–216.
[19] R.F. Loeser, J.A. Collins, B.O. Diekman, Ageing and the pathogenesis of osteoarthritis, Nat. Rev. Rheumatol. 12 (7) (2016) 412–420.
[20] S.A. Ibrahim, et al., Differences in expectations of outcome mediate African American/white patient differences in “willingness” to consider joint replacement, Arthritis Rheum. 46 (9) (2002) 2429–2435.
[21] P.W. Groeneveld, et al., Racial differences in expectations of joint replacement surgery outcomes, Arthritis Care Res.: Off. J. Am. Coll. Rheumatol. 59 (5) (2008) 730–737.
[22] M.J. Azur, et al., Multiple imputation by chained equations: what is it and how does it work? Int. J. Methods Psychiatr. Res. 20 (1) (2011) 40–49.
[23] J.L. Katzman, et al., DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network, BMC Med. Res. Methodol. 18 (1) (2018) 24.
[24] A.F. Agarap, Deep learning using rectified linear units (relu), arXiv preprint arXiv: 1803.08375, 2018.
[25] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, arXiv preprint arXiv:1412.6980, 2014.
[26] I.J. Goodfellow, et al., An empirical investigation of catastrophic forgetting in gradient-based neural networks, arXiv preprint arXiv:1312.6211, 2013.
[27] K. Weiss, T.M. Khoshgoftaar, D. Wang, A survey of transfer learning, J. Big Data 3 (2016) 1–40.
[28] S. Ruder, et al., Transfer learning in natural language processing, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials, 2019.
[29] S. Niu, et al., A decade survey of transfer learning (2010–2020), IEEE Transact. Artif. Intell. 1 (2) (2021) 151–166.
[30] Y. Zhang, et al., Comparison of the prevalence of knee osteoarthritis between the elderly Chinese population in Beijing and whites in the United States: the Beijing osteoarthritis study, Arthritis Rheum.: Official Journal of the Am. Coll. Rheumatol. 44 (9) (2001) 2065–2071.
[31] Y. Sun, et al., Test-time training with self-supervision for generalization under distribution shifts, in: International Conference on Machine Learning, PMLR, 2020.
[32] Y. Sun, et al., Learning to (learn at test time): rnns with expressive hidden states, arXiv preprint arXiv:2407.04620, 2024.
[33] V. Pedoia, et al., MRI and biomechanics multidimensional data analysis reveals R2-R1ρ as an early predictor of cartilage lesion progression in knee osteoarthritis, J. Magn. Reson. Imag. 47 (1) (2018) 78–90.
[34] Nelson, et al., A machine learning approach to knee osteoarthritis phenotyping: data from the FNIH biomarkers consortium, Osteoarthr. Cartil. 27 (7) (2019) 994–1001.
[35] P. Ornetti, et al., Gait analysis as a quantifiable outcome measure in hip or knee osteoarthritis: a systematic review, Jt. Bone Spine 77 (5) (2010) 421–425.