Development and validation of time-to-event models to predict metastatic recurrence of localized cutaneous melanoma

Guihong Wan, PhD,[a,b,c] Bonnie W. Leung, MD,[a] Mia S. DeSimone, MD, MPH,[d] Nga Nguyen, MD, MPH,[a] Ahmad Rajeh, MS,[a] Michael R. Collier, BS,[a] Hannah Rashdan, BS,[a] Katie Roster, MS,[a] Xu Zhou, MS,[a,e] Cameron B. Moseley, BS,[a] Ajit J. Nirmal, PhD,[b] Roxanne J. Pelletier, MS,[b] Zoltan Maliga, PhD,[b] Gyorgy Marko-Varga, PhD,[f] István Balázs Németh, MD, PhD,[g] Hensin Tsao, MD, PhD,[a] Maryam M. Asgari, MD,[a,h] Alexander Gusev, PhD,[i] Anna M. Stagner, MD,[j] Christine G. Lian, MD,[d] Marc S. Hurlbert, PhD,[k] Feng Liu, PhD,[e] Kun-Hsing Yu, MD, PhD,[c,d] Peter K. Sorger, PhD,[b] and Yevgeniy R. Semenov, MD, MA[a,b]

Background: The recent expansion of immunotherapy for stage IIB/IIC melanoma highlights a growing clinical need to identify patients at high risk of metastatic recurrence and, therefore, most likely to benefit from this therapeutic modality.

Objective: To develop time-to-event risk prediction models for melanoma metastatic recurrence.

Methods: Patients diagnosed with stage I/II primary cutaneous melanoma between 2000 and 2020 at Mass General Brigham and Dana-Farber Cancer Institute were included. Melanoma recurrence date and type were determined by chart review. Thirty clinicopathologic factors were extracted from electronic health records. Three types of time-to-event machine-learning models were evaluated internally and externally in the distant versus locoregional/nonrecurrence prediction.

Results: This study included 954 melanomas (155 distant, 163 locoregional, and 636 1:2 matched nonrecurrences). Distant recurrences were associated with worse survival compared to locoregional/nonrecurrences (HR: 6.21, $P < .001$) and to locoregional recurrences only (HR: 5.79, $P < .001$). The Gradient Boosting Survival model achieved the best performance (concordance index: 0.816; time-dependent AUC: 0.842; Brier score: 0.103) in the external validation.

From the Department of Dermatology, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts[a]; Department of Systems Biology, Harvard Medical School, Boston, Massachusetts[b]; Department of Biomedical Informatics, Harvard Medical School, Boston, Massachusetts[c]; Department of Pathology, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts[d]; School of Systems and Enterprises, Stevens Institute of Technology, Hoboken, New Jersey[e]; Department of Translational Medicine, Lund University, Lund, Sweden[f]; Department of Dermatology and Allergology, University of Szeged, Szeged, Hungary[g]; Department of Population Medicine, Harvard Pilgrim Healthcare, Boston, Massachusetts[h]; Department of Medicine, Dana-Farber Cancer Institute, Boston, Massachusetts[i]; Department of Pathology, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts[j]; and Melanoma Research Alliance, Washington, District of Columbia.[k]

Drs Wan and Leung are co-first authors.

Drs Sorger and Semenov are co-senior authors.

Funding sources: YRS is supported in part by the Department of Defense under Award Number W81XWH2110819, the National Institute of Arthritis and Musculoskeletal and Skin Diseases of the National Institutes of Health under Award Number K23AR080791, and the Melanoma Research Alliance Young Investigator Award.

Patient consent: Not applicable.

IRB approval status: Reviewed and approved by Mass General Brigham Institutional Review Board (Protocol # 2020P002179).

Code availability: The code is available at https://github.com/SemenovLab/Time2Event-MelanomaRecurrence.

Data availability: The data generated for this study can only be shared per specific institutional review board (IRB) requirements. Upon request to the corresponding author, a data sharing agreement can be initiated following institution-specific guidelines. The original figures were uploaded separately.

Accepted for publication August 28, 2023.

Correspondence to: Yevgeniy R. Semenov, MD, MA, Department of Dermatology, Massachusetts General Hospital, Harvard Medical School, 40 Blossom St, Bartlett Hall 6R, Room 626, Boston, MA 02114. E-mail: ysemenov@mgh.harvard.edu, Twitter: @EugeneSemenovMD.

Published online October 4, 2023.

0190-9622/$36.00

https://doi.org/10.1016/j.jaad.2023.08.105

Limitations: Retrospective nature and cohort from one geography.

Conclusions: These results suggest that time-to-event machine-learning models can reliably predict the metastatic recurrence from localized melanoma and help identify high-risk patients who are most likely to benefit from immunotherapy. (J Am Acad Dermatol 2024;90:288-98.)

Key words: clinicopathologic factors; locoregional recurrence; metastatic recurrence; stage I/II melanoma; time-to-event prediction.

INTRODUCTION

Dramatic improvements in cancer cell biology and immunology are constantly pushing forward therapeutic options, and treatment strategies for managing melanoma have continued to evolve over the past decade. Notably, pembrolizumab was approved in December 2021 as adjuvant therapy for patients with high-risk node-negative melanoma (stage IIB/IIC) following complete resection of the primary tumor.[1] However, though there are established surveillance guidelines for advanced (stage III/IV) melanoma, the management of patients with stage IIB/IIC disease is largely up to the physician's discretion, as there is limited evidence supporting specific recommendations. The decision to proceed with adjuvant immunotherapy is based on the recurrence risk, treatment-related toxicities, and patient preferences. This highlights a significant opportunity for developing decision-making tools to stratify patients with localized melanoma (stage I/II) according to their risk for recurrence and mortality.

Previous studies have demonstrated that typically up to 30% of patients with localized melanoma experience a disease recurrence.[2,3] Adjuvant immunotherapy, if utilized in the appropriate population, may prevent recurrence and improve patient outcomes. However, this treatment class is associated with significant and potentially permanent toxicities and should therefore be targeted toward patients with the highest risk of recurrence.[4-8] Furthermore, methods to identify patients at the highest risk of recurrence are lacking, underscoring the importance of re-evaluating current management and surveillance patterns to identify those whose risk profile favors systemic therapy. Specifically, prior studies have demonstrated that routine imaging detects only 21% of recurrent cases in patients with stage II disease.[2] Besides, distant metastases following surgical resection of melanoma often portend one of the worst prognoses compared to locoregional recurrences.[9] Also, distant recurrence and locoregional recurrence require different treatment strategies. For example, patients with distant recurrence may need systemic therapies, while patients with locoregional recurrence may benefit from local treatments, such as re-excision and intralesional therapy.[10] Thus, a greater understanding of distant recurrence from localized melanoma is increasingly important.

CAPSULE SUMMARY

- Clinicopathologic factors can be used to identify localized melanomas at high risk of metastatic recurrence and select patients who are most likely to benefit from adjuvant immunotherapy.
- Time-to-event models achieved a concordance index 0.816 and a time-dependent area under the receiver operating characteristic curve 0.842 in the metastatic recurrence prediction.

This study is built on a recent publication from our research group, which successfully developed risk prediction models for any type of recurrence from early-stage melanoma.[11] The goal of the current study is to specifically identify patients at the highest risk of distant recurrence, who are therefore most likely to benefit from adjuvant therapy. Since the timing of the recurrence is an important criterion in evaluating the most appropriate time for systemic intervention, we specifically focus on the methodological approach to time-to-event prediction. Compared to prior studies of melanoma recurrence that relied on conventional statistical models or concentrated on the binary classification of recurrent versus nonrecurrent melanomas,[12,13] our time-to-event machine-learning approach will enable improved representation of the complexity of disease progression and the reality of clinical practice. In this study, we apply time-to-event algorithms from 3 categories including linear, ensemble, and deep learning-based models, and compare their performance in the distant recurrence risk prediction.

METHODS

We developed time-to-event machine-learning models to predict distant recurrence versus

Abbreviations used:

AJCC-8:	American Joint Committee on Cancer eighth edition
AUC:	area under the receiver operating characteristic curve
CCS:	Charlson Comorbidity score
Coxnet:	CoxnetSurvivalAnalysis
CoxPH:	CoxPHSurvivalAnalysis
DFCI:	Dana-Farber Cancer Institute
EHR:	electronic health record
FVM:	first visit due to melanoma
GBS:	GradientBoostingSurvivalAnalysis
HOM:	history of previous melanoma
HR:	hazard ratio
MGH:	Massachusetts General Hospital
RSF:	RandomSurvivalForest
SLNB:	sentinel lymph node biopsy
VGT:	vertical growth phase type

locoregional/nonrecurrence of localized melanomas. We also examined model performances in the prediction of distant versus locoregional recurrence and the prediction of distant recurrence versus nonrecurrence. We additionally predicted distant, locoregional, and nonrecurrence together with competing risk modeling.

Data collection

This study included stage I/II melanomas at the Massachusetts General Hospital (MGH) and Brigham and Women's Hospital/Dana-Farber Cancer Institute (DFCI) between January 2000 and February 2020. Fig 1 shows the flow diagram of how the study population was obtained. All recurrent melanomas were first identified, then nonrecurrent melanomas were 2:1 best matched with recurrent melanomas on the primary melanoma diagnosis year using the "match Controls" function in R 4.1.0.[14] Supplementary Methods 1, available via Mendeley at https://doi.org/10.17632/ft7w5xmbbv.1 describes the inclusion criteria.

A manual review of electronic health records (EHRs) was conducted to ascertain the recurrence status (locoregional, distant, nonrecurrence) and date. Melanomas that were stage IV at the time of recurrence, based on the American Joint Committee on Cancer eighth edition (AJCC-8) staging guidelines, were labeled as "distant recurrence."[15] All recurrent melanomas without distant metastases were labeled as "locoregional recurrence." All melanomas without any recurrence were labeled as "nonrecurrence." The time-to-event was defined as the duration from primary melanoma diagnosis to the first recurrence if the melanoma recurred; otherwise, the duration to the date of death or last follow-up.

We extracted variables of interest (Supplementary Table I, available via Mendeley at https://doi.org/10.17632/ft7w5xmbbv.1) from 2 institutional clinical databases: Research Patient Data Registry[16] and Enterprise Data Warehouse.[17] Supplementary Methods 2, available via Mendeley at https://doi.org/10.17632/ft7w5xmbbv.1 details information extraction. Median household income was extracted from the U.S. Census data using the patient's zip code.[18] ICD codes were used to calculate the Charlson Comorbidity score (CCS).[19] Based on clinical guidelines, sentinel lymph node biopsy (SLNB) is recommended for melanomas with Breslow thickness >1.00 mm and not recommended for primary melanomas that are <0.8 mm thick and non-ulcerated lesions.[20] Primary melanomas that are 0.8 to 1.0 mm thick or are <0.8 mm thick and ulcerated may be offered SLNB after a clinical discussion of the procedure's advantages and disadvantages.[20] We incorporated the SLNB complexity to reflect real-world clinical settings.

Time-to-event machine learning methods

We compared prediction performances of 3 types of supervised time-to-event machine-learning algorithms (Supplementary Methods 3, available via Mendeley at https://doi.org/10.17632/ft7w5xmbbv.1): ensemble models, including GradientBoosting SurvivalAnalysis (GBS)[21] and RandomSurvivalForest (RSF),[22] linear models, including CoxnetSurvival Analysis (Coxnet)[23] and CoxPHSurvivalAnalysis (Cox PH),[24] and deep learning models, including Deep Surv[25] and CoxTime.[26] The DeepHit[27] was applied for competing risk modeling. Models were evaluated internally and externally by the concordance index,[28] time-dependent area under the receiver operating characteristic curve (AUC),[29] and Brier score[30] using: (1) five-fold cross-validation of the entire cohort (internal validation); (2) the MGH cohort for model development and the DFCI cohort for validating independently (external validation). Each experiment was repeated 50 times. Mean and 95% confidence interval (CI) were reported. Experiments were implemented by using scikit-survival 0.18.0[31] and pycox 0.2.3.[32]

Supplementary Fig 1, available via Mendeley at https://doi.org/10.17632/ft7w5xmbbv.1 presents the model development and validation pipeline. In Phase 0, we examined model performances using all extracted features. In Phase 1, we conducted feature selection using the MGH cohort. We first investigated the predictive features by conducting permutation importance.[33] Given the multicollinearity of features, we performed hierarchical clustering on features' Spearman rank-order correlations, and a

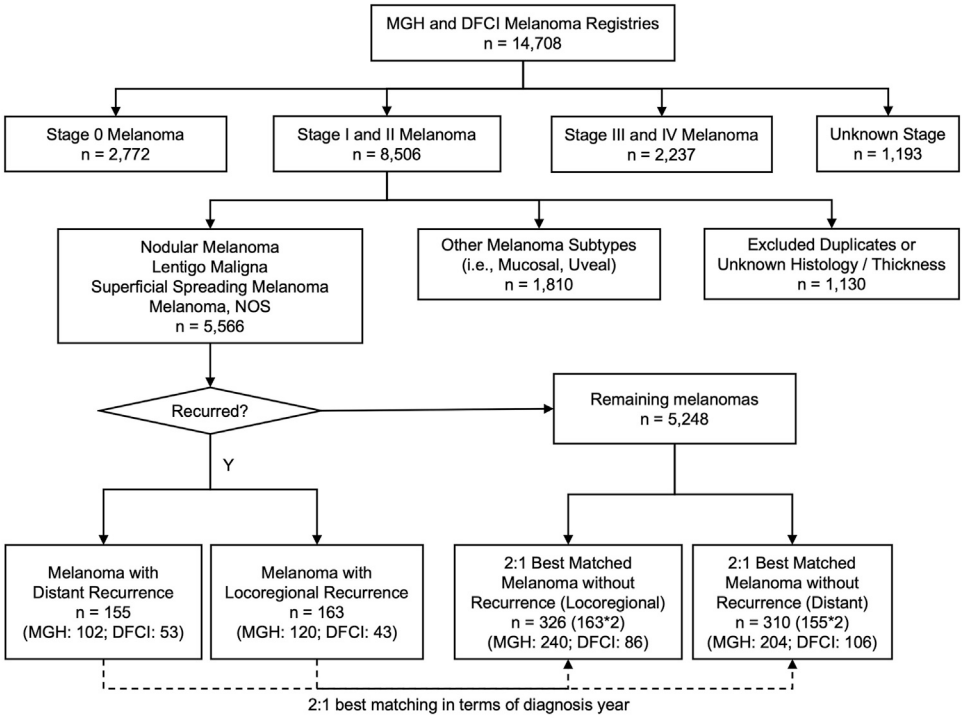

Fig 1. Study Design Flowchart. Stage I and stage II cutaneous melanomas with no evidence of nodal or distant metastasis at the time of primary diagnosis were included. Thus, melanomas with positive sentinel lymph node biopsy and melanomas with microscopic satellites were excluded. Acral, mucosal, uveal, and desmoplastic melanomas were excluded. Melanomas without a pathology report available in electronic health records were excluded. All recurrent melanomas were first identified, then nonrecurrent melanomas were 2:1 best matched with recurrent melanomas on diagnosis year of primary melanoma.

single feature in each highly correlated cluster was retained. In Phase 2, models were evaluated using the selected features. We also ranked the selected features by conducting permutation importance.

Statistical analyses

We conducted Pearson's Chi-squared test for categorical variables and t-test for continuous variables to compare groups. Kaplan-Meier curves for overall survival and distant recurrence free probabilities were utilized. Guarantee-time bias occurs when the outcome of interest competes with mortality (eg, patients must be alive long enough to develop a recurrence). To account for this potential bias, time-varying Cox proportional hazards regression models were used to compute Hazard Ratios (HRs).[34] Statistical analyses were conducted using R 4.1.0.[14]

RESULTS

Participant characteristics

Among the 954 stage I/II primary cutaneous melanomas, 155 melanomas recurred distantly, and 163 melanomas recurred locoregionally. The characteristics of the study population are described in Table I and Supplementary Table III, available via Mendeley at https://doi.org/10.17632/ft7w5xmbbv.1 (all variables). The comparison between distant and locoregional recurrences is presented in Supplementary Table IV, available via Mendeley at https://doi.org/10.17632/ft7w5xmbbv.1.

When comparing the distant recurrence group to the combined locoregional/nonrecurrence group, there were no significant differences in age at diagnosis, race, ethnicity, marital status, CCS, history of previous melanoma (HOM), laterality, presence of regression, and perineural invasion. The percentage of males in the distant recurrence group was higher (70% vs 55%, $P < .001$). The median income in the distant recurrence group was lower (95,000 vs 101,000 dollars, $P = .04$). More patients with stage II melanoma were in the distant recurrence group (59% vs 20%, $P < .001$). Fig 2, *C* presents the Kaplan-Meier curves of distant recurrence free probability stratified by AJCC-8 stage among the study population.

When comparing the distant recurrence group to the locoregional recurrence group (Supplementary

Table I. Characteristics of the study population

	Locoregional/nonrecurrence (*N* = 799)	Distant Recurrence (*N* = 155)	*P*-value
Institution			
DFCI	235 (29.4%)	53 (34.2%)	.275
MGH	564 (70.6%)	102 (65.8%)	
Duration of follow-up (year)			
Median [IQR]	6.9 [2.8, 11.2]	5.7 [3.3, 8.9]	<.001
Mortality status			
Alive	599 (75.0%)	60 (38.7%)	<.001
Dead	200 (25.0%)	95 (61.3%)	
Age at diagnosis (year)			
Median [IQR]	62 [51, 72]	63 [54, 72]	.105
Sex			
Female	362 (45.3%)	47 (30.3%)	<.001
Male	437 (54.7%)	108 (69.7%)	
Race			
White	789 (98.7%)	154 (99.4%)	.813
Unavailable/Other	10 (1.3%)	1 (0.6%)	
Ethnicity			
Non-Hispanic	786 (98.4%)	154 (99.4%)	.572
Unavailable/Other	13 (1.6%)	1 (0.6%)	
Histology type			
Lentigo maligna melanoma	55 (6.9%)	14 (9.0%)	<.001
Nodular melanoma	103 (12.9%)	36 (23.2%)	
Superficial spreading melanoma	497 (62.2%)	62 (40.0%)	
Melanoma, not otherwise specified	144 (18.0%)	43 (27.7%)	
Tumor site			
Skin of face	89 (11.1%)	24 (15.5%)	<.001
Skin of lower limb and hip	167 (20.9%)	18 (11.6%)	
Skin of scalp and neck	49 (6.1%)	35 (22.6%)	
Skin of trunk	294 (36.8%)	54 (34.8%)	
Skin of upper limb and shoulder	200 (25.0%)	24 (15.5%)	
AJCC stage			
1A	429 (53.7%)	19 (12.3%)	<.001
1B	213 (26.7%)	45 (29.0%)	
2A	64 (8.0%)	29 (18.7%)	
2B	55 (6.9%)	30 (19.4%)	
2C	38 (4.8%)	32 (20.6%)	
Breslow thickness (mm)			
Median [IQR]	0.7 [0.4, 1.4]	2.1 [1.1, 4.2]	<.001
Anatomic (Clark's) level			
Median [IQR]	4 [2, 4]	4 [4, 4]	<.001
Laterality			
Left	363 (45.4%)	73 (47.1%)	.113
Midline	79 (9.9%)	23 (14.8%)	
Right	357 (44.7%)	59 (38.1%)	
Sentinel lymph node biopsy			
Not indicated	419 (52.4%)	19 (12.3%)	<.001
All nodes negative	312 (39.0%)	102 (65.8%)	
Not performed: unknown reason	17 (2.1%)	5 (3.2%)	
Not performed: due to age/comorbidity	41 (5.1%)	27 (17.4%)	
Not performed: deferred by patient	10 (1.3%)	2 (1.3%)	
Ulceration			
Absent	701 (87.7%)	101 (65.2%)	<.001
Present	94 (11.8%)	50 (32.3%)	
Unavailable	4 (0.5%)	4 (2.6%)	

Continued

Table I. Cont'd

	Locoregional/nonrecurrence (*N* = 799)	Distant Recurrence (*N* = 155)	*P*-value
Mitotic rate (mitoses/mm^2)			
Median [IQR]	1 [0, 3]	4 [1, 12]	<.001
Total surgical margins (cm)			
Median [IQR]	1 [1, 2]	2 [1, 2]	<.001
Tumor infiltrating lymphocytes (TIL)			
Present: brisk	56 (7.0%)	7 (4.5%)	.044
Present: non-brisk	400 (50.0%)	98 (63.2%)	
Absent	160 (20.0%)	25 (16.1%)	
Unavailable	183 (22.9%)	25 (16.1%)	

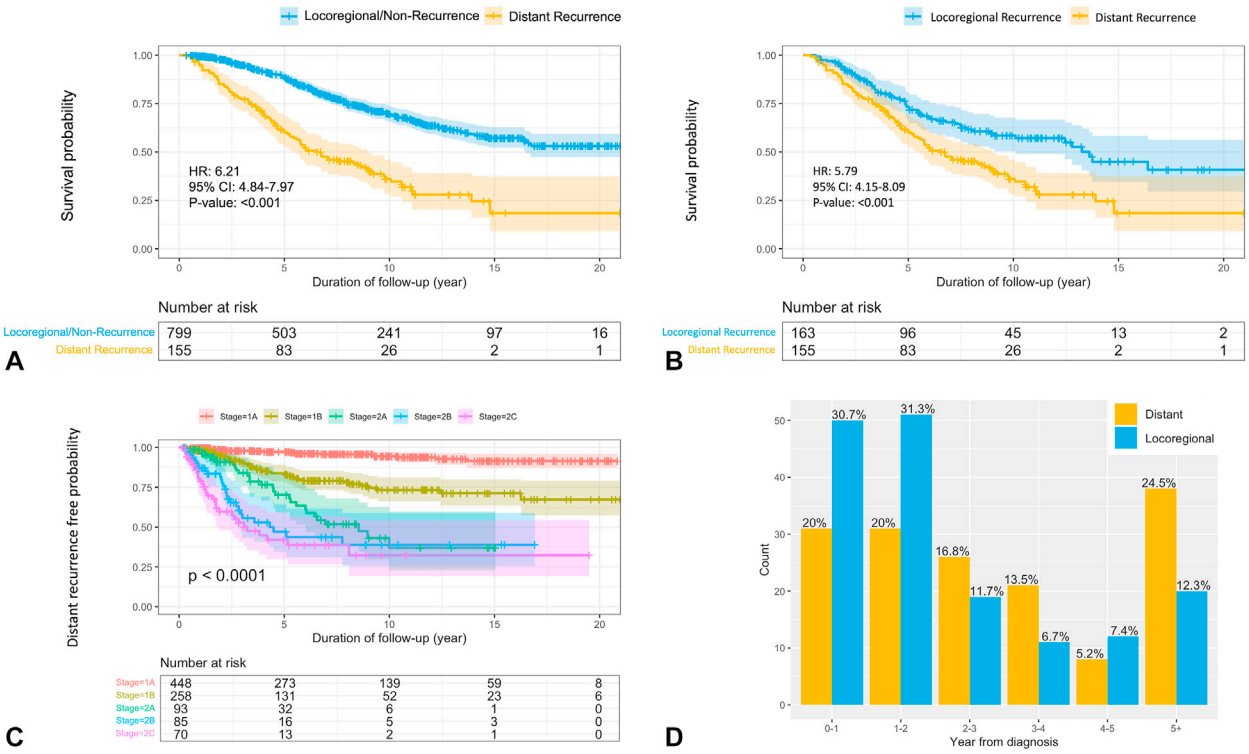

Fig 2. Kaplan-Meier curves and the distribution of recurrent melanomas. **A,** A Kaplan-Meier curve of overall survival comparing the distant recurrence group and the locoregional/nonrecurrence group. The Hazard Ratio for overall survival was computed using a univariate time-varying Cox Proportional Hazards model. **B,** A Kaplan-Meier curve of overall survival comparing the distant recurrence group and the locoregional recurrence group. The Hazard Ratio for overall survival was computed using a univariate time-varying Cox Proportional Hazards model. **C,** A Kaplan-Meier curve for distant recurrence stratified by AJCC stage. **D,** The number of distant recurrences within a period from diagnosis compared to the number of locoregional recurrences.

Table IV), there were no significant differences in all tumor characteristics except tumor site, laterality, and sentinel lymph node biopsy status. The time from primary diagnosis to distant recurrence was longer than the time to locoregional recurrence (2.6 vs 1.6 years, $P = .003$). The histogram of recurrent melanomas is presented in Fig 2, *D*. Within 2 years, there were fewer distant recurrences than locoregional recurrences (40% vs 64%, $P < .001$).

The median follow-up was 5.7 years (IQR: 3.3-8.9) for distant recurrences, 6.1 years (IQR: 3.2-10.8) for locoregional recurrences, and 7.0 years (IQR: 2.7-11.2) for nonrecurrences (Supplementary Table IV). The distant recurrence group had the highest mortality rate compared to the locoregional recurrence and the nonrecurrence groups (61.3% vs 38.7% vs 21.5%, $P < .001$). The Kaplan-Meier curves for overall survival are presented in Fig 2, *A* and *B*. The

Table II. Time-to-event distant recurrence versus locoregional/nonrecurrence prediction

	Ensemble models		Linear models		Deep learning models	
	RSF	GBS	CoxPH	CoxNet	DeepSurv	CoxTime
All extracted variables						
Concordance index*						
Internal (Mean and 95% CI)	0.836	0.844	0.811	0.805	0.811	0.820
	0.830-0.841	0.839-0.850	0.801-0.822	0.793-0.816	0.799-0.822	0.811-0.829
External (Mean and 95%	0.798	**0.811**	0.711	0.710	0.738	0.749
CI)	0.797-0.799	**0.811-0.812**	0.710-0.712	0.709-0.711	0.728-0.748	0.743-0.755
P-value†	<.001	<.001	<.001	<.001	<.001	<.001
Time-dependent AUC*						
Internal (Mean and 95% CI)	0.867	0.877	0.845	0.846	0.852	0.860
	0.861-0.873	0.869-0.885	0.833-0.857	0.837-0.856	0.839-0.865	0.850-0.869
External (Mean and 95%	0.811	**0.837**	0.738	0.737	0.770	0.779
CI)	0.810-0.812	**0.836-0.838**	0.737-0.739	0.736-0.738	0.760-0.780	0.773-0.785
P-value†	<.001	<.001	<.001	<.001	<.001	<.001
Integrated time-dependent Brier score*						
Internal (Mean and 95% CI)	0.097	0.109	0.108	0.109	0.109	0.108
	0.094-0.101	0.102-0.116	0.101-0.115	0.102-0.116	0.103-0.115	0.103-0.113
External (Mean and 95%	0.107	**0.103**	0.126	0.127	0.127	0.126
CI)	0.106-0.107	**0.103-0.104**	0.126-0.127	0.126-0.127	0.125-0.129	0.123-0.129
P-value†	<.001	.087	<.001	<.001	<.001	<.001
27 selected variables						
Concordance index*						
External (Mean and 95%	0.810	*0.816*	0.774	0.777	0.767	0.772
CI)	0.809-0.811	*0.815-0.816*	0.773-0.774	0.776-0.777	0.756-0.778	0.764-0.78
P-value‡	<.001	<.001	<.001	<.001	<.001	<.001
Time-dependent AUC*						
External (Mean and 95%	0.822	*0.842*	0.792	0.796	0.784	0.786
CI)	0.821-0.823	*0.842-0.843*	0.791-0.793	0.796-0.797	0.772-0.795	0.777-0.795
P-value‡	<.001	<.001	<.001	<.001	.079	.197
Integrated time-dependent Brier score*						
External (Mean and 95%	0.105	*0.103*	0.112	0.110	0.126	0.129
CI)	0.105-0.106	*0.103-0.104*	0.111-0.112	0.108-0.112	0.123-0.128	0.125-0.133
P-value‡	<.001	.823	<.001	<.001	.446	.087

The best performances when all extracted variables were used are highlighted in bold, and the best performances when 27 selected variables were used are highlighted in italics.

*When the concordance index and time-dependent AUC tend to 1, the model is more accurate. When the Brier score tends to 0, the model is more accurate.

†*P*-value: t-test between the internal validation and the external validation.

‡*P*-value: t-test of the external validations when all extracted variables were used and when 27 selected variables were used.

Hazard Ratio of overall mortality for the distant recurrence group was 6.2 (95% CI: 4.8-8.0, $P < .001$) when compared to the combined locoregional/nonrecurrence group and was 5.8 (95% CI: 4.2-8.1) when compared to the locoregional recurrence group.

Performance of prediction models

We built time-to-event models for distant recurrence risk prediction using ensemble, linear, and deep-learning algorithms. We evaluated the model performance first by experimenting on all extracted variables and then on the 27 selected variables (see below for details). The results are presented in Table II.

For experiments on all extracted variables, the ensemble models (GBS and RSF) achieved better performance than the linear and deep-learning models ($P < .001$). In the internal validation, GBS and RSF models achieved similar performance (concordance index: 0.844 vs 0.836, $P = .026$; time-dependent AUC: 0.877 vs 0.867, $P = .057$). In the external validation, GBS outperformed RSF (concordance index: 0.811 vs 0.798, $P < .001$; time-dependent AUC: 0.837 vs 0.811, $P < .001$).

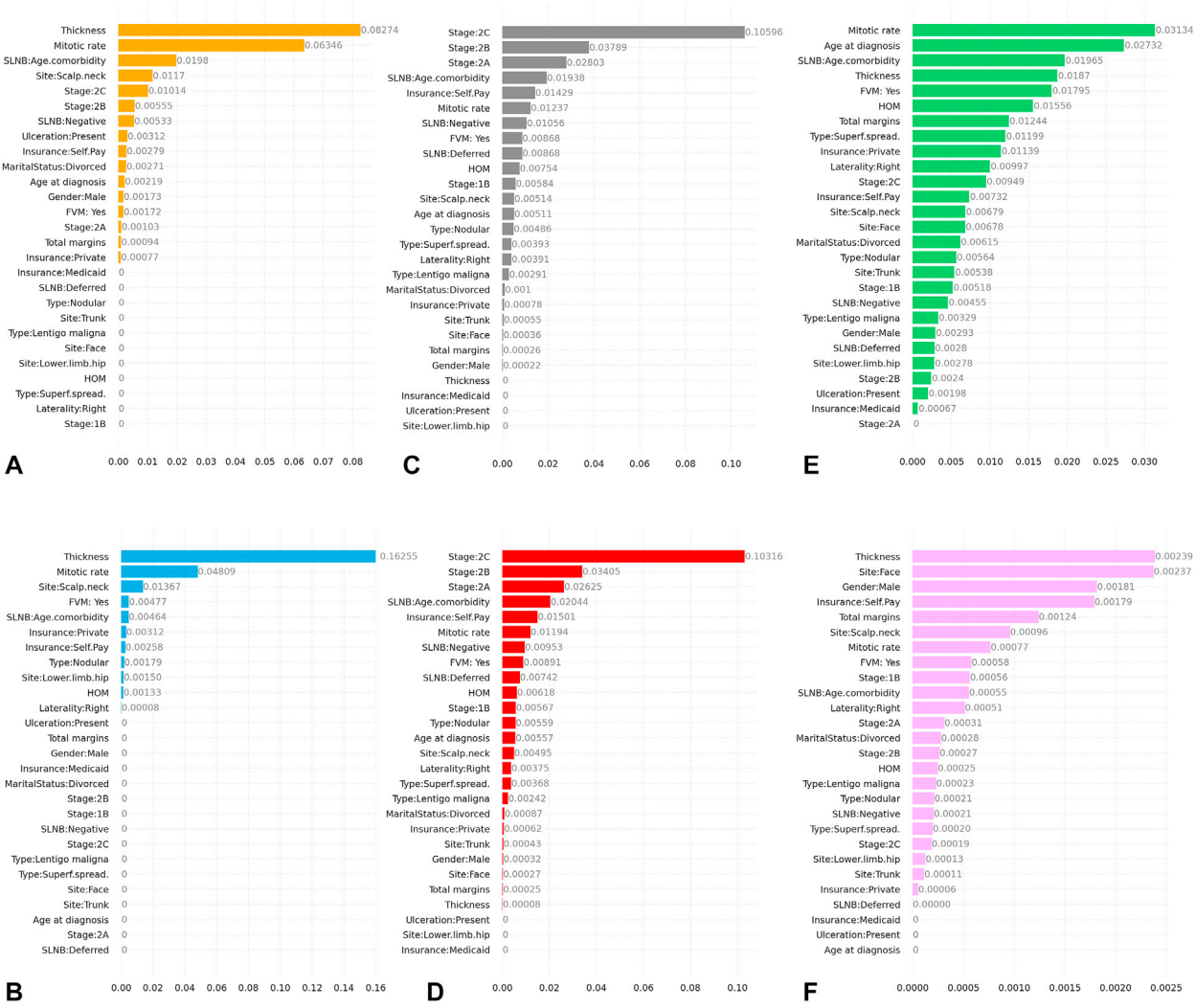

Fig 3. Mean feature importance in the time-to-event distant recurrence versus locoregional/nonrecurrence prediction on the MGH cohort. First column (**A** and **B**): Ranking of the 27 selected features in the MGH cohort by the ensemble models (RSF and GBS, respectively). Second column (**C** and **D**): Ranking of the 27 selected features in the MGH cohort by the linear models (CoxPH and Coxnet, respectively). Third column (**E** and **F**): Ranking of the 27 selected features in the MGH cohort by the deep learning models (DeepSurv and CoxTime, respectively).

We selected features by ranking all extracted features based on permutation importance (Supplementary Figs 2-4, available via Mendeley at https://doi.org/10.17632/ft7w5xmbbv.1) and conducting correlation analysis (Supplementary Fig 5, available via Mendeley at https://doi.org/10.17632/ft7w5xmbbv.1). Breslow thickness and mitotic rate were the 2 most important features in the GBS and RSF models (Supplementary Fig 2). AJCC stage, mitotic rate, and vertical growth phase type (VGT) appeared to be important in the linear models (Supplementary Fig 3). Breslow thickness, health insurance type, and total surgical margins were the most important features in the deep-learning models (Supplementary Fig 4). Other important features included SLNB, ulceration, melanoma histology type, age at diagnosis, tumor site, first visit due to melanoma (FVM), HOM, patient sex, and laterality. We initially selected 30 one-hot encoded features and conducted a correlation analysis. "Stage:1A", "SLNB: Not indicated", and "Laterality: Left" were removed since they were highly correlated with the other variables (Supplementary Fig 5). The final 27 selected features are specified in Supplementary Table II, available via Mendeley at https://doi.org/10.17632/ft7w5xmbbv.1.

When experimenting on the 27 selected features, all models achieved better performance in terms of concordance index in the external validation

($P < .001$). Ensemble and linear models achieved better time-dependent AUC ($P < .001$) compared to the results when all available features were included. The GBS models outperformed other models in the external validation (concordance index: 0.816; time-dependent AUC: 0.842, integrated time-dependent Brier score: 0.103). The results of the internal validation are presented in Supplementary Table V, available via Mendeley at https://doi.org/10.17632/ft7w5xmbbv.1.

Fig 3 displays the importance ranking of the selected features in each model. Breslow thickness, mitotic rate, SLNB:Age.comorbidity (inability to perform SLNB due to age or comorbidity of the patient), and Site:Scalp.neck were the common top 5 features in the ensemble models. Stage:2C, Stage:2B, Stage:2A, SLNB:Age.comorbidity, and Insurance:Self. Pay were the top 5 features in the linear models. Mitotic rate, Breslow thickness, Age at diagnosis, SLNB: Age.comorbidity, Site:Face, and Gender:Male played important roles in the deep-learning models.

Secondary analyses

We additionally evaluated model performances in the distant versus locoregional recurrence prediction (Supplementary Table VI, available via Mendeley at https://doi.org/10.17632/ft7w5xmbbv.1) and the distant versus nonrecurrence prediction (Supplementary Table VII, available via Mendeley at https://doi.org/10.17632/ft7w5xmbbv.1). All models did not achieve satisfactory performance in the distant versus locoregional recurrence prediction. The best concordance index and time-dependent AUC in the external validation were less than 0.7 (Supplementary Table VI). The models achieved better concordance index and time-dependent AUC in the prediction of distant versus nonrecurrence (Supplementary Table VII) compared to the results in the prediction of distant versus locoregional/nonrecurrence (Supplementary Table IV), especially in the internal validation ($P < .01$ for all models). For competing risk modeling, the model achieved concordance index of 0.767 (95% CI: 0.760-0.773) and Brier score of 1.49 (95% CI: 0.147-0.151) in the external validation.

DISCUSSION

Overall, we found that models trained to predict distant recurrence versus locoregional/nonrecurrence had better discriminative prediction than those trained to predict distant versus locoregional recurrence. While all models showed satisfactory performance, ensemble models had greater discriminating power (external: time-dependent AUC 0.822-0.842) than linear and deep-learning models (external: time-dependent AUC 0.784-0.796). Since patients with distant recurrence have significantly worse survival, these models can help identify patients who would benefit most from adjuvant immunotherapy. The enhanced performance of ensemble models is likely due to the combination of multiple diverse models used in the methodology and their ability to excel with a relatively modest volume of data.

In our models, Breslow thickness, mitotic rate, tumor site, and the reason for SLNB deferral were ranked as the most important features, which is largely consistent with the literature.[35,36] Similarly, our prior study found Breslow thickness and mitotic rate to be the most predictive features for overall primary melanoma recurrence.[11] Furthermore, another recent study demonstrated that the binary classification power of stage-related features alone has limitations and may lead to missed recurrent cases.[13] Our findings, in conjunction with prior studies, highlight that there are features beyond those included in AJCC staging criteria that can help predict recurrence with improved accuracy to inform clinical recommendations and adjuvant therapy consideration in patients with early-stage melanoma.

Compared with our prior work,[11] this study focuses on distant metastases and delineates between distant and locoregional recurrences, given the increased risk of mortality and potential benefits from adjuvant immunotherapy for those at risk of distant recurrence. In addition, we included a comprehensive set of time-to-event algorithms. This study underscores opportunities for more sophisticated risk-stratifying tools to assess a patient's risk for recurrence to inform patient selection for adjuvant therapy. By identifying those at the highest risk of distant recurrence, prediction models can prevent overtreatment and optimize adjuvant therapy benefits. A recent trial showed that adjuvant treatment with pembrolizumab for up to approximately 1 year reduced the risk of distant metastases by two-fold in patients with stage IIB/IIC melanoma.[37] Patients often describe the possibility of recurrence as a psychological and emotional burden, as rapid disease progression can occur after a melanoma recurrence.[38] Management considerations, including prognosis, potential adverse events, treatment burden, and costs to the healthcare system and the individual, are complex.[39,40] Although the occurrence of grade 3 or worse immune-related adverse events is relatively uncommon, some patients experience life-long toxicities or toxicities that require long-term therapy, such as systemic corticosteroid or hormone replacement.[37] As our study illustrates, time-to-event machine-learning algorithms demonstrate the

improved ability to capture the complexities of melanoma recurrence that are underrepresented in the current literature and likely hold the promise of future advancements in precision medicine.[38]

Limitations of this study include the retrospective nature of clinical data extraction, where some variables were unavailable. Nonetheless, our machine-learning models performed satisfactorily in the external validation. Also, our models did not include immunosuppression information and were developed using a relatively small cohort of patients seen in an overall similar geography. Future studies should leverage time-to-event algorithms with immunosuppression information and larger generalized cohorts to guide clinically deployable prediction models. Furthermore, the incorporation of prognostic molecular markers will strengthen the discriminatory power and advance personalized treatment.

Conflicts of interest

YRS is an advisory board member/consultant and has received honoraria from Incyte Corporation, Castle Biosciences, Galderma, and Sanofi outside of the submitted work.

REFERENCES

1. Killock D. Pembrolizumab reduces recurrence risk in stage II melanoma. *Nat Rev Clin Oncol*. 2022;19(6):359. https://doi.org/10.1038/s41571-022-00638-w
2. Bleicher J, Swords DS, Mali ME, et al. Recurrence patterns in patients with stage II melanoma: the evolving role of routine imaging for surveillance. *J Surg Oncol*. 2020;122(8):1770-1777. https://doi.org/10.1002/jso.26214
3. von Schuckmann LA, Hughes MCB, Ghiasvand R, et al. Risk of melanoma recurrence after diagnosis of a high-risk primary tumor. *JAMA Dermatol*. 2019;155(6):688-693. https://doi.org/10.1001/jamadermatol.2019.0440
4. Palmieri DJ, Carlino MS. Immune checkpoint inhibitor toxicity. *Curr Oncol Rep*. 2018;20(9):72. https://doi.org/10.1007/s11912-018-0718-6
5. Nguyen N, Wan G, Ugwu-Dike P, et al. Influence of melanoma type on incidence and downstream implications of cutaneous immune-related adverse events in the setting of immune checkpoint inhibitor therapy. *J Am Acad Dermatol*. 2023;88(6):1308-1316. https://doi.org/10.1016/j.jaad.2023.02.014
6. Leung BW, Wan G, Nguyen N, et al. Increased risk of cutaneous immune-related adverse events in patients treated with talimogene laherparepvec and immune checkpoint inhibitors: a multi-hospital cohort study. *J Am Acad Dermatol*. 2023;88(6):1265-1270. https://doi.org/10.1016/j.jaad.2023.02.017
7. Leung B, Zhang S, Nguyen N, et al. 259 tissue-specific homing in cutaneous immune-related adverse events. *J Invest Dermatol*. 2022;142(8):S44. https://doi.org/10.1016/j.jid.2022.05.266
8. Wan G, Leung B, DeSimone M, et al. 217 time-to-event machine learning prediction of metastatic recurrence of localized melanoma. *J Investig Dermatol*. 2023;143(5):S37.
9. Soong SJ, Harrison RA, McCarthy WH, Urist MM, Balch CM. Factors affecting survival following local, regional, or distant recurrence from localized melanoma. *J Surg Oncol*. 1998;67(4):228-233. https://doi.org/10.1002/(sici)1096-9098(199804)67:4<228::aid-jso4>3.0.co;2-a
10. Pasquali S, Hadjinicolaou AV, Chiarion Sileni V, Rossi CR, Mocellin S. Systemic treatments for metastatic cutaneous melanoma. *Cochrane Database Syst Rev*. 2018;2(2):CD011123. https://doi.org/10.1002/14651858.CD011123.pub2
11. Wan G, Nguyen N, Liu F, et al. Prediction of early-stage melanoma recurrence using clinical and histopathologic features. *NPJ Precis Onc*. 2022;6:79. https://doi.org/10.1038/s41698-022-00321-4
12. Matheson JA, Te Marvelde L, Mailer S, et al. Prospective evaluation of prognostic indicators for early recurrence of cutaneous melanoma. *Melanoma Res*. 2017;27(1):43-49.
13. Wan G, Leung B, Nguyen N, et al. The impact of stage-related features in melanoma recurrence prediction: a machine learning approach. *JAAD Int*. 2022;10:28-30. https://doi.org/10.1016/j.jdin.2022.08.014
14. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing; 2018.
15. Gershenwald JE, Scolyer RA. Melanoma staging: American Joint Committee on Cancer (AJCC) 8th edition and beyond. *Ann Surg Oncol*. 2018;25:2105-2110.
16. Nalichowski R, Keogh D, Chueh HC, Murphy SN. Calculating the benefits of a research patient data Repository. *AMIA Annu Symp Proc*. 2006;2006:1044.
17. *The Enterprise Data Warehouse (EDW): Creating the Foundation for Effective Healthcare Improvement Analytics*. Health Catalyst; 2015.
18. United States Census Bureau. Selected Income Characteristics, 2006—2020 American Community Survey 5-year Estimates. United States Census Bureau; 2020.
19. Roffman CE, Buchanan J, Allison GT. Charlson comorbidities index. *J Physiother*. 2016;62:171.
20. Wong SL, Faries MB, Kennedy EB, et al. Sentinel lymph node biopsy and management of regional lymph nodes in melanoma: American Society of Clinical Oncology and Society of Surgical Oncology clinical practice guideline update. *J Clin Oncol*. 2018;36:399-413.
21. Hothorn T, Buhlmann P, Dudoit S, Molinaro A, van der Laan MJ. Survival ensembles. *Biostatistics*. 2006;7(3):355-373. https://doi.org/10.1093/biostatistics/kxj011
22. Hemant I, Udaya BK, Eugene HB, Michael SL. Random survival forests. *Ann Appl Stat*. 2008;2(3):841-860. https://doi.org/10.1214/08-AOAS169
23. Simon N, Friedman J, Hastie T, Tibshirani R. Regularization paths for Cox's proportional hazards model via coordinate descent. *J Stat Software*. 2011;39(5):1.
24. Kalbfleisch JD. *The Efficiency of Cox's Likelihood Function for Censored Data. Springer Series in Statistics:119-129*. 1977.
25. Katzman JL, 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*. 2018;18(1):24. https://doi.org/10.1186/s12874-018-0482-1
26. Kvamme H, Borgan Ø, Scheel I. Time-to-event prediction with neural networks and Cox regression. *J Mach Learn Res*. 2019;20(129):1-30.
27. Lee C, Zame W, Yoon J, van der Schaar M. DeepHit: A Deep Learning Approach to Survival Analysis With Competing Risks. *Proceedings of the AAAI Conference on Artificial Intelligence*. 2018;32(1). https://doi.org/10.1609/aaai.v32i1.11842
28. Antolini L, Boracchi P, Biganzoli E. A time-dependent discrimination index for survival data. *Stat Med*. 2005;24(24):3927-3944.
29. Lambert J, Chevret S. Summary measure of discrimination in survival models based on cumulative/dynamic time-dependent ROC curves. *Stat Methods Med Res*. 2016;25(5):2088-2102. https://doi.org/10.1177/0962280213515571

30. Brier GW. Verification of forecasts expressed in terms of probability. *Mon Weather Rev*. 1950;78(1):1-3.
31. Pölsterl S. Scikit-survival: a Library for time-to-event analysis built on top of scikit-learn. *J Mach Learn Res*. 2020; 21(212):1-6.
32. Havakv. Havakv/pycox: survival analysis with pytorch. GitHub. Accessed December 9, 2022. https://github.com/havakv/pycox
33. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 2011;12:2825-2830.
34. Zhang Z, Reinikainen J, Adeleke KA, Pieterse ME, Groothuis-Oudshoorn CGM. Time-varying covariates and coefficients in Cox regression models. *Ann Transl Med*. 2018;6(7):121. https://doi.org/10.21037/atm.2018.02.12
35. Ahmed I. Malignant melanoma: prognostic indicators. *Mayo Clin Proc*. 1997;72(4):356-361. https://doi.org/10.4065/72.4.356
36. Azzola MF, Shaw HM, Thompson JF, et al. Tumor mitotic rate is a more powerful prognostic indicator than ulceration in patients with primary cutaneous melanoma: an analysis of 3661 patients from a single center. *Cancer*. 2003;97(6):1488-1498. https://doi.org/10.1002/cncr.11196
37. Luke JJ, Rutkowski P, Queirolo P, et al. Pembrolizumab versus placebo as adjuvant therapy in completely resected stage IIB or IIC melanoma (KEYNOTE-716): a randomised, double-blind, phase 3 trial. *Lancet*. 2022;399(10336):1718-1729. https://doi.org/10.1016/S0140-6736(22)00562-1
38. Livingstone A, Milne D, Dempsey K, et al. Should i have adjuvant immunotherapy? An interview study among adults with resected stage 3 melanoma and their partners. *Patient*. 2021;14(5):635-647. https://doi.org/10.1007/s40271-021-00507-1
39. Leung B, Nguyen N, Choi MS, et al. Clinical and histopathologic risk factors for early-stage melanoma recurrence. *J Investig Dermatol*. 2022;142:S113.
40. Leung BW, Collier MR, Tiu BC, et al. Patterns in utilization of health care services and medications among patients with cutaneous immune-related adverse events: a population-level cohort study. *J Am Acad Dermatol*. 2023;88(5):1215-1218. https://doi.org/10.1016/j.jaad.2022.12.042