# Unsupervised Phenotype-Based Clustering of Clinicopathologic Features in Cutaneous Melanoma

Sarem Rashid[1], Nikolai Klebanov[2], William M. Lin[2] and Hensin Tsao[1,2]

Pathogenic phenotypes in cutaneous melanoma have been vastly cataloged, although these classifications lack concordance and are confined to either morphological or molecular contexts. In this study, we perform unsupervised k-medoids clustering as a machine learning technique of 2,978 primary cutaneous melanomas at Mass General Brigham and apply this information to elucidate computer-defined subsets within the clinicopathologic domain. We identified five optimally separated clusters of melanoma that occupied two distinct clinicopathologic subspaces: a lower-grade partition associated with common or dysplastic nevi (i.e., nevus-associated melanomas) and a higher-grade partition lacking precursor lesions (i.e., de novo melanomas). Our model found de novo melanomas to be more mitogenic, more ulcerative, and thicker than nevus-associated melanomas, in addition to harboring previously unreported differences in radial and vertical growth phase status. The utilization of mixed clinicopathologic variables, reflective of actual clinical data contained in surgical pathology reports, has the potential to increase the biological relevance of existing melanoma classification schemes and facilitate the discovery of new genomic subtypes.

*JID Innovations* (2021);**1:**100047 doi:10.1016/j.xjidi.2021.100047

## INTRODUCTION

Melanoma constitutes approximately 5% of all skin cancers, yet accounts for >75% of skin cancer deaths (Rebecca et al., 2020). Before current molecular and immune therapies, long-term survival was largely correlated with observed clinicopathologic (CP) features refined over decades of epidemiological research (Gupta and Tsao, 2017; Mihm et al., 1971). Breslow (1970) first introduced pathology-based prognostication by demonstrating a positive relationship between tumor thickness and the risk of melanoma metastasis. Subsequent studies identified pathologic features, such as vertical growth phase and lymphovascular invasion, to be significantly associated with metastatic disease and outcome (Elder, 1999). Thus, there had been mixed attempts at feature-based predictions (e.g., thickness) and morphology-based classifications (e.g., superficial spreading melanoma). More recently, there have been efforts to stratify melanoma into outcome-defined classes on the basis of distinct histological features (Amin et al., 2017).

With the advent of genomic technologies, unsupervised machine learning approaches have uncovered latent molecular classes of melanoma that define new relationships between clinical and gene expression changes while underscoring potential etiological pathways (Cancer Genome Atlas Network, 2015; Harbst et al., 2012; Thakur et al., 2019). Through these studies, various molecular types of melanomas have been discovered and designated in accordance with the predominant gene signatures, for example, immune and pigmentation types. One of the recognized assets of unsupervised approaches is the ability to define previously unappreciated relationships between variables that may allow one to formulate novel classifiers. Although this is now standard in transcriptomic studies where gene expression levels (all continuous variables) have been used to identify the classes mentioned earlier, all parameters, including CP features, can be subjected to unsupervised clustering using algorithms that can employ mixed data. With this in mind, we set out to perform class discovery using mixed CP variables (e.g., continuous for thickness, dichotomous for ulceration) from 2,978 melanoma cases seen at Mass General Brigham (MGB), Boston, MA, over a 17-year period to identify the key differentiating factors while elucidating the potential clinical implications.

## RESULTS

### Cluster phenotypes

To determine whether CP features could classify primary cutaneous melanomas into discrete subgroups, we performed unsupervised clustering of our training set data using k-medoids clustering. Clustering variables are specified in Table 1. Of note, the vast majority of melanomas were predicted as early-stage lesions at biopsy (69.3% American Joint Committee on Cancer [AJCC]-predicted stage IA or IB).

[1]*Wellman Center for Photomedicine, Massachusetts General Hospital, Boston, Massachusetts, USA; and* [2]*Department of Dermatology, Massachusetts General Hospital, Boston, Massachusetts, USA*

*Correspondence: Hensin Tsao, Department of Dermatology, Massachusetts General Hospital, Edwards 211, 50 Blossom Street,Boston, Massachusetts 02114, USA. E-mail: htsao@mgh.harvard.edu*

*Abbreviations: AJCC, American Joint Committee on Cancer; CP, clinicopathologic; DNM, de novo melanoma; MGB, Mass General Brigham; NAM, nevus-associated melanoma; PL, precursor lesion; RC, replicate set cluster; RGP, radial growth phase; TC, training set cluster*

*Received 21 April 2021; revised 28 June 2021; accepted 29 June 2021; accepted manuscript published online XXX; corrected proof published online XXX*

*Cite this article as: JID Innovations 2021;1:100047*

© 2021 Published by Elsevier Inc. on behalf of the Society for Investigative Dermatology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

**Table 1. Baseline Summary of Randomized Melanoma Entries for Training and Replicate Data Sets**

| Feature | Training Set | Replicate Set | *P*-value[1] |
|---|---|---|---|
| n | 1,979 | 999 | |
| Thickness (mm), median (IQR) | 1.14 (0.7–2.3) | 1.10 (0.67–2.08) | 0.255 |
| Mitoses (per mm$^2$), median (IQR) | 2.00 (0.8–5.0) | 1.00 (0.0–4.0) | 0.523 |
| Anatomic (Clark) level | | | 0.101 |
| 2 | 92 | 51 | |
| 3 | 535 | 296 | |
| 4 | 1,219 | 604 | |
| 5 | 133 | 48 | |
| Ulceration | | | 0.793 |
| Absent | 1,636 | 822 | |
| Present | 343 | 177 | |
| RGP | | | 0.208 |
| Absent | 489 | 226 | |
| Present | 1,490 | 773 | |
| VGP | | | 0.658 |
| Absent | 19 | 14 | |
| Epithelioid | 1,428 | 713 | |
| Spindled | 114 | 52 | |
| Small | 39 | 21 | |
| Unspecified | | | |
| PL | | | 0.779 |
| Absent | 1,504 | 758 | |
| Common | 234 | 111 | |
| Dysplastic | 220 | 116 | |
| Other[2] | 21 | 14 | |
| TILs | | | 0.944 |
| Nonbrisk/absent | 1,829 | 924 | |
| Brisk | 150 | 75 | |
| VI | | | 0.867 |
| Absent | 1,801 | 911 | |
| Present | 178 | 88 | |
| Regression | | | 0.214 |
| Absent | 1,943 | 974 | |
| Present | 36 | 25 | |
| Sex | | | 0.235 |
| Female | 858 | 456 | |
| Male | 1,121 | 543 | |
| Age (y), median (IQR) | 66.00 (55.00–76.00) | 66.00 (54.00–76.00) | 0.611 |
| Month of diagnosis, median (IQR) | 7.0 (4.0–9.0) | 7.0 (4.0–9.0) | 0.729 |
| Predicted AJCC Stage | | | 0.424 |
| IA | 591 | 315 | |
| IB | 758 | 398 | |
| IIA | 292 | 129 | |
| IIB | 214 | 94 | |
| IIC | 124 | 63 | |

Abbreviations: AJCC, American Joint Committee on Cancer; IQR, interquartile range; PL, precursor lesion; RGP, radial growth phase; TIL, tumor-infiltrating lymphocyte; VGP, vertical growth phase; VI, vascular invasion.

[1]Statistical tests were performed by the Mann–Whitney test for continuous variables and chi-square testing for categorical variables.

[2]Other category includes the following precursor lesions: melanocytic hyperplasia, nevus of special sites, small dermal melanocytic nest, and lentigo maligna.

Using the silhouette width optimization technique, a five-cluster solution (training set clusters [TCs] 1–5) yielded the highest average silhouette width (silhouette score = 0.215). Visualization of these five clusters in the CP space is shown in the t-distributed stochastic neighbor embedding plot (Figure 1a), and cluster definitions are enumerated in Table 2.

On initial inspection, nevus-associated melanomas (NAMs) appeared to be discretely separated from de novo melanomas (DNMs). TC1 (n = 235) and TC2 (n = 210) tumors tended to be nonulcerated, have higher rates of radial growth phase (RGP), and nearly always harbored benign and dysplastic nevi, respectively. TC1 and TC2 lesions were also less thick than TC3 (n = 644), TC4 (n = 610), and TC5 (n = 280) tumors (respective medians: 0.90 and 0.76 vs. 1.10 mm, 1.10, and 3.60 mm; $P < 0.0001$) with significantly fewer mitoses per mm$^2$ (1.00 and 1.00 per mm$^2$ vs. 1.00, 2.00 and 9.00 per mm$^2$; $P < 0.0001$). In addition, TC2 had the highest rate of brisk tumor-infiltrating lymphocytes (13.3%) among the five clusters ($P = 0.0082$). TC3–5 melanomas were largely de novo lesions. TC5 had the thickest tumors ($P < 0.0001$), the highest densities of mitoses ($P < 0.0001$), and the largest rates of ulceration (78.9% positive, $P < 0.0001$) among the five clusters. TC5 was separated from TC4 by the absence of regression (95.4% vs. 98.0%, respectively; $P < 0.0001$) and sex category (77.9% vs. 4.1% men, respectively; $P < 0.0001$). When compared with TC4 and TC5, TC3 tumors were found almost exclusively in men (95.7% male, $P < 0.001$), had lower mitotic rate ($P < 0.0001$) thresholds, and had increased presence of RGP (84.6% positive, $P < 0.0001$).

Because the algorithm appeared to have identified NAMs and DNMs as the strongest partitions among all the melanomas within the CP space, we examined NAMs and DNMs individually as superclusters. As shown in Figure 1c, DNMs had significantly greater thickness ($P < 0.0001$) and mitotic rate ($P < 0.0001$) than NAMs. DNMs were also associated with increased presence of spindled- or unspecified-type vertical growth phase (6.4% and 22.0% vs. 3.6% and 9.2%, respectively) but decreased presence of RGP (70.9% vs. 90.3%, respectively) ($P < 0.0001$).

We next projected the five TC groups onto clinical features that were not used in the initial clustering (i.e., AJCC staging and pathologic subtype) to determine whether there were any significant correlations. Superficial spreading melanoma was the overall most common subtype for all cluster assignments with the exception of TC5, which demonstrated a predominance of nodular melanomas (26.4% nodular melanomas; $P = 0.0058$ Fisher's exact test; $P < 0.0001$ chi-square test). TC5 contained the highest proportion of AJCC-predicted stage IIB and stage IIC melanomas ($P < 0.001$), with 36.1% and 34.6% for each pathologic stage, respectively, and which is consistent with the increased thickness and ulceration rate. TC1 and TC2 contained the highest proportion of stage IA melanomas (37.4% and 52.9%, respectively) ($P < 0.001$).

Finally, we applied the five-cluster k-medoids algorithm to an independent set of 999 cases randomly selected from our initial cohort of 2,978 cases (replicate set clusters (RCs)

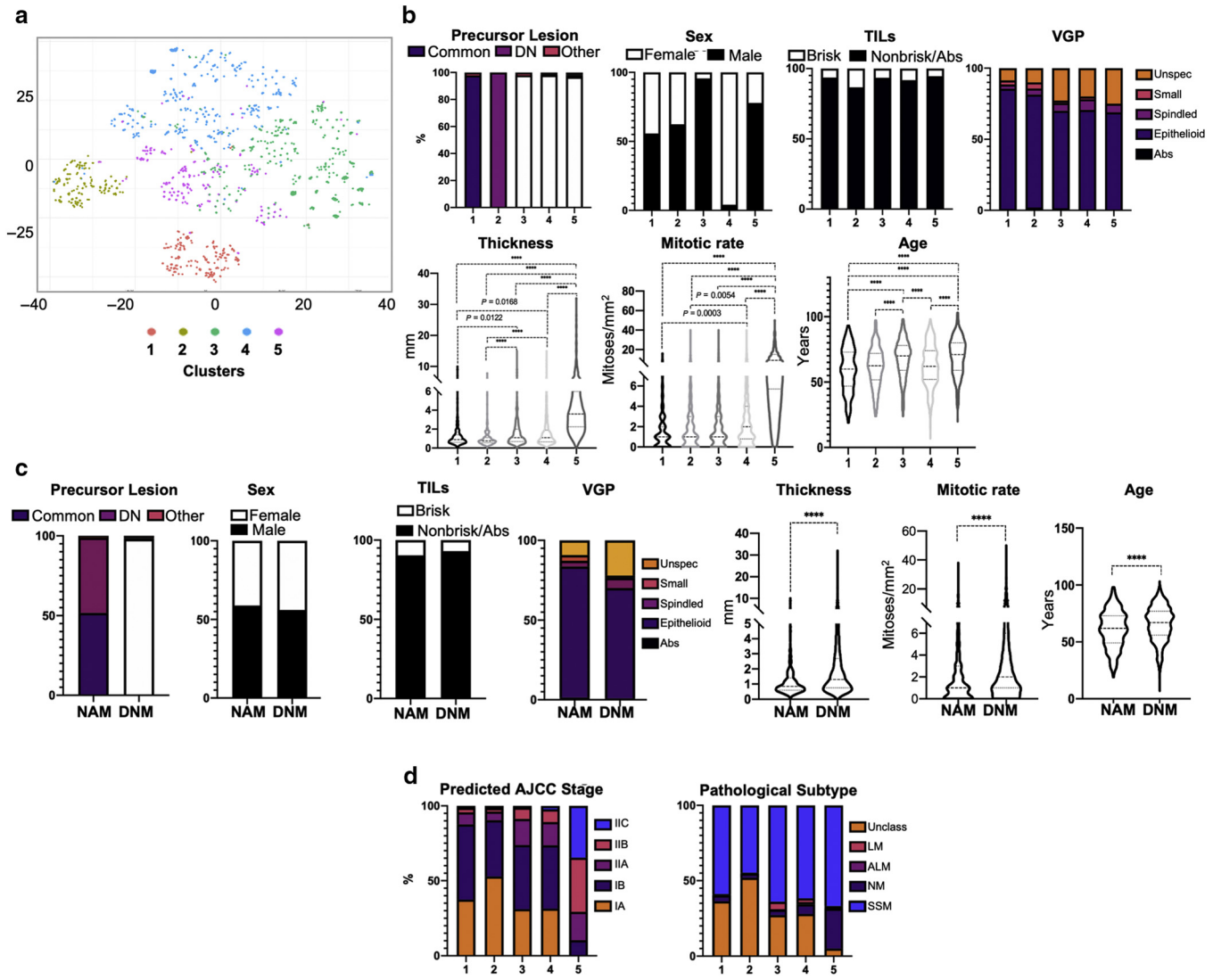


**Figure 1. Training set CP cluster comparisons and projected outcomes.** (**a**) t-SNE, a computational method for visualizing high dimensional data, of k-medoids clustering assignments that demonstrates how the algorithm conceptualizes the data in CP space. (**b**) Individual feature comparisons show distinct organization on the basis of PL status. TC1 and 2 were largely comprised of NAMs and were differentiated by the presence of common or dysplastic lesions, respectively ($P < 0.0001$). Conversely, TC3−5 were made of DNMs. (**c**) Supercluster comparisons demonstrate significantly increased thickness, mitotic rate, and age in DNMs compared with those in NAMs ($P < 0.0001$). (**d**) To the left, we observe a greater proportion of IIA, IIB, and IIC melanomas in TC3−5 ($P < 0.0001$). To the right, the plot shows the highest prevalence of SSM for all TCs. ****Significant associations with $P < 0.0001$. Abs, absent; AJCC, American Joint Committee on Cancer; ALM, acral lentiginous melanoma; CP, clinicopathologic; DN, dysplastic nevi; DNM, de novo melanoma; LM, lentigo maligna; NAM, nevus-associated melanoma; PL, precursor lesion; SSM, superficial spreading melanoma; t-SNE, t-distributed stochastic neighbor embedding; TC, training set cluster; TIL, tumor-infiltrating lymphocyte; VGP, vertical growth phase.

of 1−5) (Figure 2). The clinical features between the training and replication sets were not statistically different (Table 1). Again, the model agnostically separated melanoma entries into two NAM subgroups (RC1 and RC2), distinguished by benign versus dysplastic lesions, and three DNM subgroups (RC3−5) separated by sex and vertical growth phase status (Figure 3). Consistent with the training set, RC3−5 tumors were thicker (respective medians of 1.30, 1.10, and 1.23 mm; $P < 0.0001$) and overall more mitogenic (respective medians of 2.00, 1.00, and 2.00 per mm$^2$; $P < 0.0001$) than RC1 and 2 tumors (Table 3). Altogether, precursor lesion (PL) status remains a strong classifier within the replicate CP space.

## DISCUSSION

Although many histological associations have been identified in melanoma over decades of research, most of these have relied on supervised approaches, such as logistic regression. We employed a fundamentally different approach using unsupervised clustering on multiple CP factors on the basis of a calculated similarity distance (in this case, Gower distance) between each case. With this approach given our data, there appear to be five optimally separated subclasses of melanoma that occupy two distinct CP subspaces: a lower-grade partition (TC1−2) that are associated with common or dysplastic nevi (i.e., NAMs) and a higher-grade partition (TC3−5) that lack PLs (i.e., DNMs). Pathologic reports were

**Table 2. Summary of Clinicopathologic Features for Each Cluster Using the k-Medoids Algorithm (k = 5) on the Training Data Set (n = 1,979)**

| Feature | TC1 | TC2 | TC3 | TC4 | TC5 | *P*-value[1] |
|---|---|---|---|---|---|---|
| n | 235 | 210 | 644 | 610 | 280 | |
| Thickness (mm), median (IQR) | 0.90 (0.60–1.39) | 0.76 (0.56–1.10) | 1.10 (0.70–2.00) | 1.10 (0.66–1.90) | 3.60 (2.25–5.89) | <0.0001 |
| Mitoses (per $mm^2$), median (IQR) | 1.00 (0.00–2.00) | 1.00 (0.00–2.80) | 1.00 (0.00–3.00) | 2.00 (0.80–4.00) | 9.00 (5.80–14.90) | <0.0001 |
| Anatomic (Clark) level | | | | | | <0.0001 |
| 2 | 21 (9) | 13 (6.3) | 28 (4.6) | 30 (5.2) | 0 (0) | |
| 3 | 82 (34.9) | 87 (41.4) | 177 (27.5) | 162 (26.6) | 27 (9.6) | |
| 4 | 131 (55.7) | 108 (51.4) | 404 (62.7) | 383 (62.8) | 193 (68.9) | |
| 5 | 1 (0.4) | 2 (1) | 35 (5.4) | 35 (5.7) | 60 (21.4) | |
| Ulceration | | | | | | <0.0001 |
| Absent | 225 (95.7) | 195 (92.9) | 608 (94.4) | 549 (90) | 59 (21.1) | |
| Present | 10 (4.3) | 15 (7.1) | 36 (5.6) | 61 (10) | 221 (78.9) | |
| RGP | | | | | | <0.0001 |
| Absent | 33 (14) | 10 (4.8) | 99 (15.4) | 128 (21) | 219 (78.2) | |
| Present | 202 (86) | 200 (95.2) | 545 (84.6) | 482 (79) | 61 (21.8) | |
| VGP | | | | | | <0.0001 |
| Absent | 2 (0.9) | 4 (1.9) | 9 (1.4) | 4 (0.7) | 0 (0) | |
| Epithelioid | 199 (84.7) | 167 (79.5) | 442 (68.6) | 427 (70) | 193 (68.9) | |
| Spindled | 7 (3) | 9 (4.3) | 35 (5.4) | 46 (7.5) | 17 (6.1) | |
| Small | 7 (3) | 9 (4.3) | 11 (1.7) | 12 (2) | 0 (0) | |
| Unspecified | 20 (8.5) | 21 (10) | 147 (22.8) | 121 (19.8) | 70 (25) | |
| PL | | | | | | <0.0001 |
| Absent | 0 (0) | 0 (0) | 633 (98.3) | 599 (98.2) | 272 (97.1) | |
| Common | 230 (97.9) | 0 (0) | 0 (0) | 0 (0) | 4 (1.4) | |
| Dysplastic | 0 (0) | 210 (100) | 0 (0) | 6 (1) | 4 (1.4) | |
| Other[2] | 5 (2.1) | 0 (0) | 11 (1.7) | 5 (0.8) | 0 (0) | |
| TILs | | | | | | 0.0082 |
| Nonbrisk/absent | 220 (93.6) | 182 (86.7) | 602 (93.5) | 560 (91.8) | 265 (94.6) | |
| Brisk | 15 (6.4) | 28 (13.3) | 42 (6.5) | 50 (8.2) | 15 (5.4) | |
| VI | | | | | | <0.0001 |
| Absent | 201 (85.5) | 186 (88.6) | 571 (88.7) | 569 (93.3) | 274 (97.9) | |
| Present | 34 (14.5) | 24 (11.4) | 73 (11.3) | 41 (6.7) | 6 (2.1) | |
| Regression | | | | | | 0.0023 |
| Absent | 233 (99.1) | 208 (99) | 637 (98.9) | 598 (98) | 267 (95.4) | |
| Present | 2 (0.9) | 2 (1) | 7 (1.1) | 12 (2) | 13 (4.6) | |
| Sex | | | | | | <0.0001 |
| Female | 104 (44.3) | 79 (37.6) | 28 (4.3) | 585 (95.9) | 62 (22.1) | |
| Male | 131 (55.7) | 131 (62.4) | 616 (95.7) | 25 (4.1) | 218 (77.9) | |
| Age (y), median (IQR) | 60.00 (47.00–73.00) | 62.50 (52.00–72.00) | 70.00 (59.00–78.00) | 62.00 (52.00–74.00) | 71.00 (59.00–80.00) | <0.0001 |
| Month of diagnosis, median (IQR) | 7.0 (4.0–9.0) | 7.0 (4.0–10.0) | 6.0 (4.0–9.0) | 7.0 (4.0–9.0) | 6.0 (4.0–9.0) | 0.233 |

Abbreviations: IQR, interquartile range; PL, precursor lesion; RGP, radial growth phase; TC, training set cluster; TIL, tumor-infiltrating lymphocyte; VGP, vertical growth phase; VI, vascular invasion.

[1]Statistical tests were performed by Mood's median test for continuous variables and by chi-square testing for categorical variables.

[2]Other category includes the following precursor lesions: melanocytic hyperplasia, nevus of special sites, small dermal melanocytic nest, and lentigo maligna.

randomized into training and testing (replicate) sets to evaluate model consistency, and the testing results showed a concordant five-cluster model. Our model was able to reliably segregate melanomas on the basis of PL status. However, it is important to highlight that models trained on such data may not be applicable to other data formats.

Although the relationship between NAMs and DNMs has been carefully studied in the past (Cymerman et al., 2016; Dessinioti et al., 2021; Lin et al., 2015; Martin-Gorgojo et al., 2018; Pampena et al., 2017; Reiter et al., 2021; Sheen et al., 2017), the unsupervised clustering suggests that NAMs are more similar to each other across all CP features than DNMs. In other words, NAMs (e.g., TC1–2) tend to be less mitogenic, less ulcerative, and less thick than DNMs (TC3–5). Although some of these associations have been previously described (Reiter et al., 2021; Sheen et al., 2017), our unsupervised algorithm agnostically identified a group of melanomas that is associated with less aggressive features despite the lack of melanoma-specific mortality data to supervise the clustering. Furthermore, the model has reported supercluster differences in both vertical growth phase and RGP statuses, which have not, to our knowledge, been previously described in the literature. Between TC1 and TC2, tumors with precursor dysplastic nevi have a slightly higher rate of

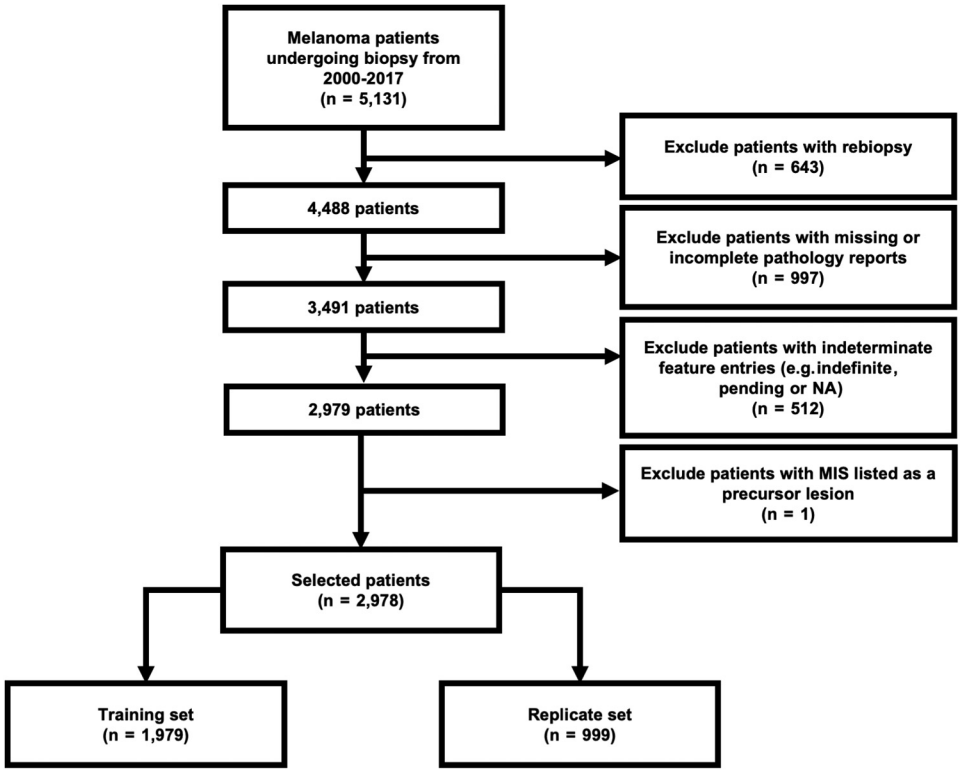


**Figure 2. Enrollment and attrition for patient entries in the clustering analysis.** MIS, melanoma in situ; NA, not applicable.

brisk tumor-infiltrating lymphocytes, suggesting that melanomas associated with dysplastic nevi may be more immunogenic than those contiguous with a common nevus. Numerous previous studies have reported increased thickness, ulceration, nodular subtype, pathologic stage, and/or mortality in DNMs compared with those in NAMs (Cymerman et al., 2016; Martin-Gorgojo et al., 2018; Pampena et al., 2017). One year before, Lin et al. (2015) reported significant associations in ulceration status, age, pathologic subtype, and truncal site in NAMs; however, their study exclusively selected patients undergoing sentinel lymph node biopsy. Furthermore, investigators have shown a central anatomic site (e.g., head, neck, and truncal areas) predisposition for DNMs and increased likelihood of nodular subtype DNMs that are truncally located (Dessinioti et al., 2021). Pandeya et al. (2018) have also shown a significantly increased proportion of *BRAF(V600E)* mutations in NAMs. Taken together, the literature suggests significant biological differences supported by this model that seem to link phenotype with a molecular senescence model of melanogenesis. In this model, the intrinsic or acquired presence of certain pathogenic mutations in banal nevi such as *BRAF(V600E)* triggers an oncogene-induced senescence (Shain and Bastian, 2016). Various interactions from downstream effectors and the tumor microenvironment drive the acquisition of somatic passenger mutations (e.g., *TERT*, *CDK2NA*); therefore, facilitating a small minority of banal nevi to eventually progress to melanoma.

Interestingly, we isolated a group of thick DNMs in men with minimal RGP and the highest rates of mitoses and ulceration (TC5). TC4 is a female-enriched DNM class that appears to harbor less aggressive features than TC3, however more aggressive ones than TC5. Again, without access to survival information, the algorithm has identified sex as a critical discriminant within the DNM subspace. To date, there is no clear etiology for the male survival disadvantage, although the neglect of sun-protective behaviors, sex steroid interactions, oxidative stress exposure, and vitamin D metabolism have all been suspected to play a role in compromised cancer immune surveillance (Gamba et al., 2013).

The unsupervised clustering also suggests a notable correlation between higher-partition melanomas (DNMs) and stage, predicted according to the eighth edition of AJCC guidelines using tumor thickness and ulceration (Figure 1d). TC3 and TC4 demonstrated nearly identical stage compositions. In particular, this may suggest that male de novo tumors are costaged with female de novo tumors with slightly increased rates of mitoses and angioinvasion. Both CP features have been reported as independent adverse prognosticators of tumor size, depth of invasion, and/or mortality in previous studies (Crowson et al., 2006; Scolyer et al., 2020). Of note, mitotic rate was removed as a T-stage parameter in the eighth edition of the AJCC melanoma staging system, although a collection is still recommended for inclusion in biopsies (Amin et al., 2017). Although this model should not be used to replace the current staging guidelines, the phenotypic associations presented certainly encourage the investigation of a broader set of parameters to inform personalized melanoma prognostication.

Altogether, the oncogene-induced senescence hypothesis suggests a distinct pathogenesis for growth-arrested NAMs that involves either dysregulation of the MAPK signaling pathway, intrinsic changes to the cellular microenvironment, or a combination of both (Damsky and Bosenberg, 2017). Lesions in a male host are expected to be less resilient against

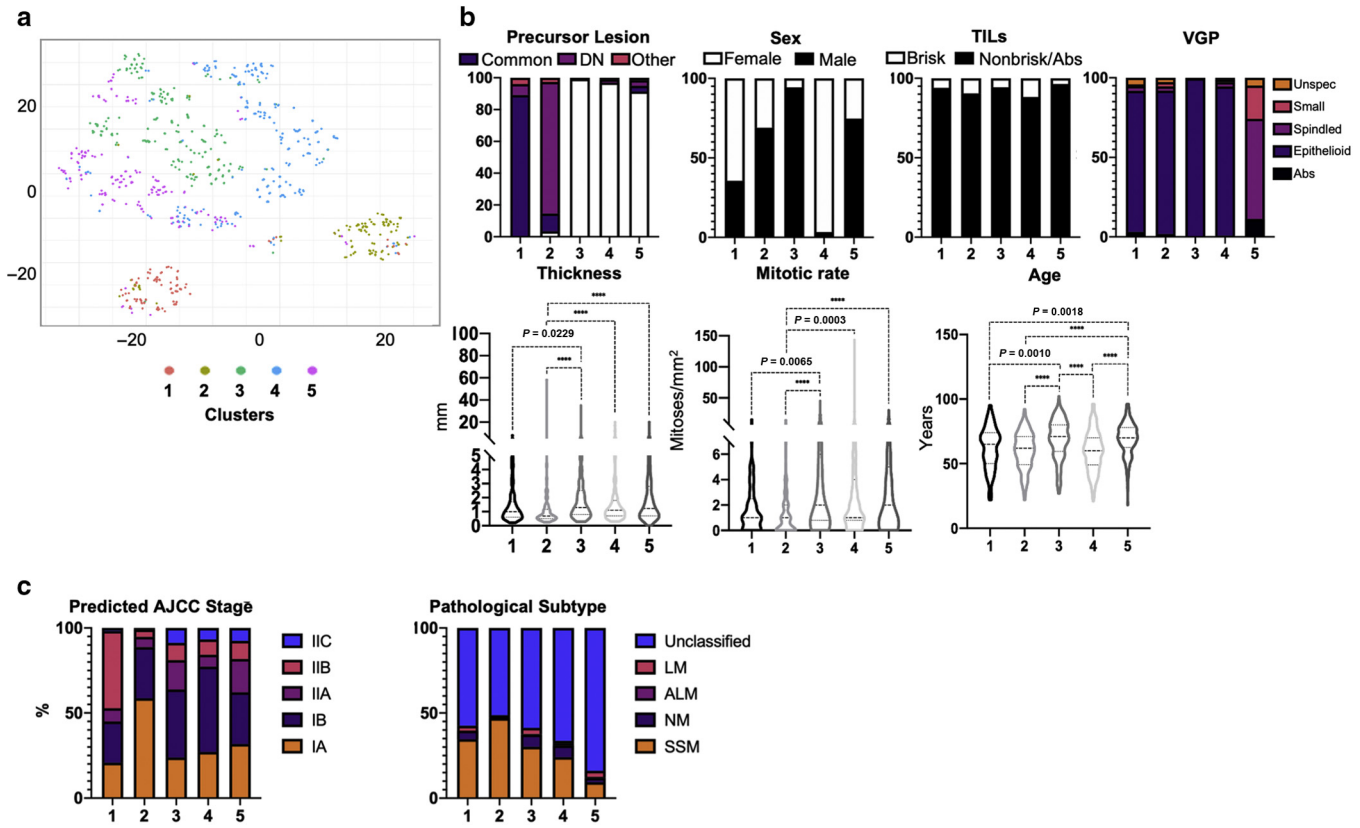


**Figure 3. CP comparisons and projections for RCs.** (**a**) Visualization of RC assignments using t-SNE. (**b**) Again, individual feature distributions showed precursor lesion status to be the strongest partition within the CP space. RC1−2 tumors tended to be NAMs, whereas RC3−5 tumors tended to be DNMs with increased thickness and mitotic rate. (**c**) RC3−5 demonstrated the highest proportion of stage IIC melanomas ($P < 0.0001$). SSM was the most common pathologic subtype across all replicate set clusters. ****Significant associations with $P < 0.0001$. Abs, absent; AJCC, American Joint Committee on Cancer; ALM, acral lentiginous melanoma; CP, clinicopathologic; DN, dysplastic nevi; DNM, de novo melanoma; LM, lentigo maligna; NAM, nevus-associated melanoma; RC, replicate set cluster; SSM, superficial spreading melanoma; t-SNE, t-distributed stochastic neighbor embedding.

innate defenses of the host immune response, which may modulate melanoma development through an independent mechanism. The inclusion of molecular correlates, for example, *BRAF*-mutant melanocytic nevi, may further unravel hidden classifiers within our data and would therefore be valuable to include in a clustering approach in future studies.

## MATERIALS AND METHODS

### Study subjects

Clinical and pathologic data were obtained using the MGB (formerly Partners) Healthcare Research Patient Data Registry and were approved by the MGB institutional review board. We queried patients of all ages with a medical record–level diagnosis of personal history of malignant melanoma of the skin at all available MGB sites. Data included demographics variables, and free-text surgical pathology report texts were requested for dates starting from 1 January 2000 to 11 October 2017. Tumor biopsies had been obtained from patients as part of routine care with appropriate informed consent.

### Data preprocessing and variable selection

From the 24,016 entries retrieved in the original query, we excluded nonmelanoma pathology reports, reports with missing accompanying clinical data, and reports that were pertaining to sentinel lymph node biopsies. For patients with rebiopsies due to microsatellites or metastatic disease, the thickest entry was included. Next, we randomized the remaining sample in an approximately 2:1 ratio to compose the training and replicate sets (Figure 2).

Pathologic features were extracted from free-text pathology synoptic reports using regular expression text matching (grep base R function, version 3.6.2). Features with ≥25% missing data were excluded from the analysis. Observations with incomplete data (or that were not available) in variables with <25% missing data were also excluded. Anatomic location and laterality were not extracted and therefore not included in our final clustering analysis.

PL entries were classified as 0 (Absent), 1 (Common), 2 (Dysplastic), and 3 (Other). Lesions included in the Other category are listed in Tables 1 and 2. From our data, six lesions had melanoma in situ designated as a PL, although this was not uniformly categorized across the data set. As such, we did not consider in situ melanomas as a PL but rather the earliest phases of the invasive melanoma for five of the six entries. The remaining entry was not a precursor to invasive melanoma and was therefore excluded from the analysis. Furthermore, tumor-infiltrating lymphocytes were stratified into two categories: nonbrisk/absent versus brisk. For older entries reporting mitoses per high power field, the following conversion was used: 1 mitosis per 10 high power fields = 0.4 mitosis per $mm^2$. Predicted AJCC stage for cluster assignments was

**Table 3. Summary of Clinicopathologic Features for Each Cluster Using the k-Medoids Algorithm (k = 5) on the Replicate Data Set (n = 999)**

| Feature | TC1 | TC2 | TC3 | TC4 | TC5 | *P*-value[1] |
|---|---|---|---|---|---|---|
| n | 101 | 116 | 285 | 299 | 198 | |
| Thickness (mm), median (IQR) | 1.00 (0.61–1.69) | 0.69 (0.51–1.13) | 1.30 (0.80–2.50) | 1.10 (0.70–1.80) | 1.23 (0.70–2.77) | <0.0001 |
| Mitoses (per $mm^2$), median (IQR) | 1.00 (0.00–2.00) | 1.00 (0.00–2.00) | 2.00 (0.80–6.00) | 1.00 (0.80–4.00) | 2.00 (0.00–5.00) | <0.0001 |
| Anatomic (Clark) level | | | | | | <0.0001 |
| 2 | 8 (8.2) | 14 (12.3) | 11 (4) | 8 (2.8) | 10 (5.5) | |
| 3 | 23 (22.8) | 64 (55.2) | 70 (24.6) | 94 (31.4) | 45 (22.7) | |
| 4 | 67 (66.3) | 36 (31) | 194 (68.1) | 180 (60.2) | 127 (64.1) | |
| 5 | 3 (3) | 2 (1.7) | 10 (3.5) | 17 (5.7) | 16 (8.1) | |
| Ulceration | | | | | | <0.0001 |
| Absent | 97 (96) | 111 (95.7) | 210 (73.7) | 245 (81.9) | 159 (80.3) | |
| Present | 4 (4) | 5 (4.3) | 75 (26.3) | 54 (18.1) | 39 (19.7) | |
| RGP | | | | | | <0.0001 |
| Absent | 19 (18.8) | 6 (5.2) | 74 (26) | 70 (23.4) | 57 (28.8) | |
| Present | 82 (81.2) | 110 (94.8) | 211 (74) | 229 (76.6) | 141 (71.2) | |
| VGP | | | | | | <0.0001 |
| Absent | 3 (3) | 3 (2.6) | 0 (0) | 2 (0.7) | 7 (3.5) | |
| Epithelioid | 86 (85.1) | 100 (86.2) | 285 (100) | 242 (80.9) | 0 (0) | |
| Spindled | 3 (3) | 3 (2.6) | 0 (0) | 7 (2.3) | 39 (19.7) | |
| Small | 1 (1) | 3 (2.6) | 0 (0) | 4 (1.3) | 13 (6.6) | |
| Unspecified | 8 (7.9) | 7 (6) | 0 (0) | 44 (14.7) | 139 (70.2) | |
| PL | | | | | | <0.0001 |
| Absent | 0 (0) | 4 (3.4) | 283 (99.3) | 290 (97) | 181 (91.4) | |
| Common | 90 (89.1) | 13 (11.2) | 1 (0.4) | 0 (0) | 7 (3.5) | |
| Dysplastic | 7 (6.9) | 96 (82.8) | 0 (0) | 6 (2) | 7 (3.5) | |
| Other[2] | 4 (4) | 3 (2.6) | 1 (0.4) | 3 (1) | 3 (1.5) | |
| TILs | | | | | | 0.0057 |
| Nonbrisk/Absent | 95 (94.1) | 105 (90.5) | 269 (94.4) | 264 (88.3) | 191 (96.5) | |
| Brisk | 6 (5.9) | 11 (9.5) | 16 (5.6) | 35 (11.7) | 7 (3.5) | |
| VI | | | | | | 0.2833 |
| Absent | 90 (89.1) | 103 (88.8) | 255 (89.5) | 280 (93.6) | 183 (92.4) | |
| Present | 11 (10.9) | 13 (11.2) | 30 (10.5) | 19 (6.4) | 15 (7.6) | |
| Regression | | | | | | 0.9428 |
| Absent | 99 (98) | 114 (98.3) | 277 (97.2) | 292 (97.7) | 192 (97) | |
| Present | 2 (2) | 2 (1.7) | 8 (2.8) | 7 (2.3) | 6 (3) | |
| Sex | | | | | | <0.0001 |
| Female | 36 (35.6) | 80 (69) | 269 (94.4) | 10 (3.3) | 148 (74.7) | |
| Male | 65 (64.4) | 36 (31) | 16 (5.6) | 289 (96.7) | 50 (25.3) | |
| Age (y), median (IQR) | 65.00 (50.00–73.00) | 62.00 (49.50–70.75) | 71.00 (59.50–80.00) | 60 (49.00–70.00) | 70.50 (63.00–78.00) | <0.0001 |
| Month of diagnosis, median (IQR) | 8.0 (4.0–10.0) | 5.0 (3.0–9.0) | 6.0 (4.0–9.0) | 7.0 (4.0–9.0) | 6.0 (4.0–9.0) | 0.178 |

Abbreviations: IQR, interquartile range; PL, precursor lesion; RGP, radial growth phase; TC, training set cluster; TIL, tumor-infiltrating lymphocyte; VGP, vertical growth phase; VI, vascular invasion.

[1]Statistical tests were performed by Mood's median test for continuous variables and by chi-square testing for categorical variables.

[2]Other category includes the following precursor lesions: melanocytic hyperplasia, nevus of special sites, small dermal melanocytic nest, and lentigo maligna.

calculated on the basis of the T-stage criterion only because information on regional and distant metastasis is not contained within MGB pathology reports. A summary of patient characteristics for each data set is described in Table 1. Notably, the pathologic features included in MGB pathology reports include additional prognostic parameters beyond those utilized by the eighth edition of the AJCC staging system. Thus, replication of this model may be confined to a limited number of datasets.

### Computational and statistical methods

Clustering analysis was performed using statistic software R (version 4.0.2; R Development Core Team). Data randomization was performed using the default rand operation in Microsoft Excel (Excel, version 16.40, Microsoft, Redmond, WA). Cluster member features were compared using chi-square testing for nominal variables and the Mann–Whitney test for continuous variables.

Pathologic reports were compiled into a mixed continuous–categorical data frame and were used to calculate a Gower dissimilarity matrix, implemented using the daisy function from the cluster package (version 2.1.1; https://cran.r-project.org/package=cluster). For clustering assignments, the k-medoids algorithm (using the pam function from cluster) iteratively determined the minimum overall cost within a cluster to select partition medoids, defined as representative objects for each cluster. Optimal cluster number (k) was

determined computationally; the average silhouette width is the method of choice for the k-medoids algorithm. We optimized a priori cluster number for k = 2−10.

### Data availability statement

The deidentified datasets analyzed during this study are available on reasonable request.

### ORCIDs

Sarem Rashid: http://orcid.org/0000-0002-4001-2335
Nikolai Klebanov: http://orcid.org/0000-0001-8310-3029
Hensin Tsao: http://orcid.org/0000-0002-2204-2071

### AUTHOR CONTRIBUTIONS

Conceptualization: SR, HT; Data Curation: NK; Formal Analysis: SR; Methodology: SR, NK, HT; Software: SR; Supervision: WML, HT; Validation: SR; Writing - Original Draft Preparation: SR, HT; Writing - Review and Editing: NK, WML

### ACKNOWLEDGMENTS

This work was funded through the generous donors of the Massachusetts General Hospital (MA, Boston). We thank the pathologists at Mass General Brigham for providing pathologic biopsy reports and would like to dedicate these results to the numerous patients with melanoma. All work for this study was performed in Boston, MA.

### CONFLICT OF INTEREST

The authors state no conflict of interest.

### REFERENCES

Amin MB, Edge S, Greene F, Byrd DR, Brookland RK, Washington MK, et al., editors. AJCC cancer staging manual. 8th ed. New York, NY: Springer International Publishing; 2017.

Breslow A. Thickness, cross-sectional areas and depth of invasion in the prognosis of cutaneous melanoma. Ann Surg 1970;172:902−8.

Cancer Genome Atlas Network. Genomic classification of cutaneous melanoma. Cell 2015;161:1681−96.

Crowson AN, Magro CM, Mihm MC. Prognosticators of melanoma, the melanoma report, and the sentinel lymph node. Mod Pathol 2006;19(Suppl. 2):S71−87.

Cymerman RM, Shao Y, Wang K, Zhang Y, Murzaku EC, Penn LA, et al. De novo vs nevus-associated melanomas: differences in associations with prognostic indicators and survival. J Natl Cancer Inst 2016;108:djw121.

Damsky WE, Bosenberg M. Melanocytic nevi and melanoma: unraveling a complex relationship. Oncogene 2017;36:5771−92.

Dessinioti C, Geller AC, Stergiopoulou A, Dimou N, Lo S, Keim U, et al. A multicentre study of naevus-associated melanoma vs. de novo melanoma, tumour thickness and body site differences. Br J Dermatol 2021;185:101−9.

Elder D. Tumor progression, early diagnosis and prognosis of melanoma. Acta Oncol 1999;38:535−47.

Gamba CS, Clarke CA, Keegan TH, Tao L, Swetter SM. Melanoma survival disadvantage in young, non-Hispanic white males compared with females. JAMA Dermatol 2013;149:912−20.

Gupta S, Tsao H. Epidemiology of melanoma. In: Loda M, Mucci L, Mittelstadt ML, Van Hemelrijck M, Cotter MB, editors. Pathology and epidemiology of cancer. New York, NY: Springer International Publishing; 2017. p. 591−611.

Harbst K, Staaf J, Lauss M, Karlsson A, Måsbäck A, Johansson I, et al. Molecular profiling reveals low- and high-grade forms of primary melanoma [published correction appears in Clin Cancer Res 2012;18:5829]. Clin Cancer Res 2012;18:4026−36.

Lin WM, Luo S, Muzikansky A, Lobo AZ, Tanabe KK, Sober AJ, et al. Outcome of patients with de novo versus nevus-associated melanoma. J Am Acad Dermatol 2015;72:54−8.

Martin-Gorgojo A, Requena C, Garcia-Casado Z, Traves V, Kumar R, Nagore E. Dysplastic vs. common naevus-associated vs. de novo melanomas: an observational retrospective study of 1,021 patients. Acta Derm Venereol 2018;98:556−62.

Mihm MC Jr, Clark WH Jr. From L. The clinical diagnosis, classification and histogenetic concepts of the early stages of cutaneous malignant melanomas. N Engl J Med 1971;284:1078−82.

Pampena R, Kyrgidis A, Lallas A, Moscarella E, Argenziano G, Longo C. A meta-analysis of nevus-associated melanoma: prevalence and practical implications. J Am Acad Dermatol 2017;77:938−45.e4.

Pandeya N, Kvaskoff M, Olsen CM, Green AC, Perry S, Baxter C, et al. Factors related to nevus-associated cutaneous melanoma: a case-case study. J Invest Dermatol 2018;138:1816−24.

Rebecca VW, Somasundaram R, Herlyn M. Pre-clinical modeling of cutaneous melanoma. Nat Commun 2020;11:2858.

Reiter O, Kurtansky N, Nanda JK, Busam KJ, Scope A, Musthaq S, et al. The differences in clinical and dermoscopic features between in situ and invasive nevus-associated melanomas and de novo melanomas. J Eur Acad Dermatol Venereol 2021;35:1111−8.

Scolyer RA, Rawson RV, Gershenwald JE, Ferguson PM, Prieto VG. Melanoma pathology reporting and staging. Mod Pathol 2020;33(Suppl. 1):15−24.

Shain AH, Bastian BC. From melanocytes to melanomas [published correction appears in. Nat Rev Cancer 2020;20:355. Nat Rev Cancer 2016;16:345−58.

Sheen YS, Liao YH, Lin MH, Chen JS, Liau JY, Liang CW, et al. Clinicopathological features and prognosis of patients with de novo versus nevus-associated melanoma in Taiwan. PLoS One 2017;12:e0177126.

Thakur R, Laye JP, Lauss M, Diaz JMS, O'Shea SJ, Poźniak J, et al. Transcriptomic analysis reveals prognostic molecular signatures of stage I melanoma. Clin Cancer Res 2019;25:7424−35.

**This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-nd/4.0/**