Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-01496-3

Self supervised artificial intelligence predicts poor outcome from primary cutaneous squamous cell carcinoma at diagnosis

Check for updates

Nicolas Coudray[1,2,17], Michelle C. Juarez[3,17], Maressa C. Criscito[3,17], Adalberto Claudio Quiros[4,17], Reason Wilken[5], Stephanie R. Jackson Cullison[6], Mary L. Stevenson[3], Nicole A. Doudican[3], Ke Yuan[4,7,8], Jamie D. Aquino[9], Daniel M. Klufas[9], Jeffrey P. North[9], Siegrid S. Yu[9], Fadi Murad[10], Emily Ruiz[10], Chrysalyne D. Schmults[10], Cristian D. Cardona Machado[11,12,13], Javier Cañueto[11,12,13], Anirudh Choudhary[14], Alysia N. Hughes[15], Alyssa Stockard[15], Zachary Leibovit-Reiben[15], Aaron R. Mangold[15], Aristotelis Tsirigos[1,2,16,18] ✉ & John A. Carucci[3,18] ✉

Primary cutaneous squamous cell carcinoma (cSCC) is responsible for ~10,000 deaths annually in the United States. Stratification of risk of poor outcome at initial biopsy would significantly impact clinical decision-making during the initial post operative period where intervention has been shown to be most effective. Using whole-slide images (WSI) from 163 patients from 3 institutions, we developed a self supervised deep-learning model to predict poor outcomes in cSCC patients from histopathological features at initial diagnosis, and validated it using WSI from 563 patients, collected from two other academic institutions. For disease-free survival prediction, the model attained a concordance index of 0.73 in the development cohort and 0.84 in the Mayo cohort. The model's interpretability revealed that features like poor differentiation and deep invasion were strongly associated with poor prognosis. Furthermore, the model is effective in stratifying risk among BWH T2a and AJCC T2, known for outcome heterogeneity.

Cutaneous squamous cell carcinoma (cSCC) is the second most common human cancer, with an estimated incidence of over 1 million cases in the United States and an increasing worldwide incidence over the past 20 years[1–4]. While most cSCCs portend a good prognosis, a subset of tumors are associated with poor outcomes (PO)[5,6] including local recurrence (LR), nodal metastasis (NM), distant metastasis (DM), and disease-specific death (DSD). cSCC causes the majority of keratinocyte carcinoma (KC) deaths in the United States (US); it is estimated that ~10,000 patients die annually from cSCC, which is similar to DSD rates from other cancers such as leukemia, non-Hodgkin lymphoma, and melanoma[7]. Respectively, nodal metastasis and cSCC-specific death occur in approximately 5% and 2% of patients[8,9]. While PO are rare, these values are likely underestimations given

[1]Applied Bioinformatics Laboratories, New York University School of Medicine, New York, NY, USA. [2]Department of Medicine, Division of Precision Medicine, NYU Grossman School of Medicine, New York, NY, USA. [3]The Ronald O. Perelman Department of Dermatology, New York University Grossman School of Medicine, New York, NY, USA. [4]School of Computing Science, University of Glasgow, Glasgow, Scotland, UK. [5]Department of Dermatology, Northwell Health, New York, NY, USA. [6]Department of Dermatology, Thomas Jefferson University, Philadelphia, PA, USA. [7]School of Cancer Sciences, University of Glasgow, Glasgow, Scotland, UK. [8]Cancer Research UK Beatson Institute, Glasgow, Scotland, UK. [9]Department of Dermatology, University of California, San Francisco, San Francisco, CA, USA. [10]Department of Dermatology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. [11]Instituto de Biología Molecular y Celular del Cáncer (Lab 20), Campus Miguel de Unamuno, Salamanca, Spain. [12]Instituto de Investigación Biomédica de Salamanca, CANC-30, Salamanca, Spain. [13]Department of Dermatology, Complejo Asistencial Universitario de Salamanca, Salamanca, Spain. [14]Department of Computer Science, University of Illinois, Urbana-Champain, IL, USA. [15]Mayo Clinic, Scottsdale, AZ, USA. [16]Department of Pathology, New York University School of Medicine, New York, NY, USA. [17]These authors contributed equally: Nicolas Coudray, Michelle C. Juarez, Maressa C. Criscito, Adalberto Claudio Quiros.[18]These authors jointly supervised this work: Aristotelis Tsirigos, John A. Carucci. ✉e-mail: Aristotelis.Tsirigos@nyulangone.org; John.Carucci@nyulangone.org

that there is currently no national cancer registry data for cSCCs that accurately assess incidence and prevalence. In addition, we must recognize that relatively small percentages of poor outcomes from extraordinarily large numbers of patients with cSCC result in significant numbers of patients with metastases and disease-specific death. This in turns results in significant morbidity, mortality and public health cost burden that may be avoidable if highest risk patients are identified early in their course and receive appropriate adjuvant intervention[10]. Furthermore, many epidemiologic studies[11,12] often combine data on keratinocytic carcinomas (basal cell and squamous cell carcinomas) and thus incidence and prevalence rates are widely variable. With an increasing incidence of cSCC, PO may also be rising, making this an underrecognized, although significant, public health entity.

Identifying patients at risk for poor outcome at time of diagnosis can be challenging. Commonly used cSCC staging systems include the American Joint Committee on Cancer Staging Manual 8th edition (AJCC 8)[13,14] and the Brigham and Women's Hospital (BWH) Staging System. The AJCC 8 system utilizes the following factors for cSCC tumor (T) staging: tumor size, deep invasion, perineural invasion (PNI) or evidence of bone invasion. The BWH staging model includes tumor size, poor differentiation, PNI, extension beyond subcutaneous fat and assigns a T stage based on the number of high risk features present. While most POs occur in high stage tumors (BWH T2b and AJCC 8 T3 and above), 25% of POs still occur in low stage tumors, particularly T2a/T2, highlighting the heterogeneity in outcomes, and concomitant difficulties with risk stratification, in this particular subset[15]. Thus, prediction of patients at risk for poor outcome, even in lower stage tumors, is imperative as it may affect management in terms of work up, treatment, postoperative surveillance and early implementation of adjuvant therapy[16,17]. Importantly, identification of which patients may be at risk for POs in this low stage subgroup is ill-defined in the literature and remains under urgent investigation.

While gene expression has predictive power[18,19], such technology may not be accessible to all patients and is associated with high cost. Visual inspection of the distinct histopathological features found in cSCC, on the other hand, remains important in predicting risk of progression and its role in predicting risk of PO in low staged tumors requires further elucidation[20]. Microscopic analysis using hematoxylin and eosin (H&E) stained tissue is the gold standard for diagnosing cSCC and is useful in identifying high risk features such as depth of invasion, PNI, and degree of differentiation. However, light microscopy has a limited role in determining prognosis itself, given potential for sampling error as well as inherent inter-reader variability in histopathologic features. Recently, supervised machine learning algorithms have been used in the cutaneous melanoma realm to identify prognostically important features, response to immunotherapy, one-year disease-free survival and mutation prediction with promising results[21–24]. Current studies looking at machine learning for KC, such as squamous cell carcinoma are, however, limited. Few studies have investigated the use of artificial intelligence (AI) on whole slide images (WSI) of cSCC: Recently, Knuutila et al.[25] trained a supervised ResNet architecture to show that artificial intelligence can predict the risk of metastasis from primary tumor slides of cSCC, but this approach was not amenable to describing which features were used by the classifier.

Herein, our goal was to investigate the histomorphological features associated with cSCC outcomes via a self-supervised deep-learning approach using images from WSI collected from shave, punch, or excisional biopsy specimens of primary cSCC tumors. While supervised approaches are trained to directly learn specific labels (which can often be time-consuming to obtain, require expertise, and may generate bias)[26–28], self-supervised approaches achieve self-discovery of common patterns across unlabeled datasets[29–32]. Supervised approaches are also often described as black boxes whose decision process is difficult to interpret[33,34], which despite an effort to develop interpretation strategies[35], is hindering its acceptance by humans and regulatory approval organizations[34]. Investigating an unbiased and interpretable strategy for prediction of outcome is therefore crucial, and self-supervised learning paradigms currently offer great opportunities for the development of AI in research and medicine[36]. In this study, we used 163 patients from three academic institutions as a development cohort and 563 patients from two other institutions as test cohorts (Supplementary Fig. 1a) and based the core of our study on a Histological Phenotype Pipeline (HPL)[29]. This pipeline (Fig. 1) has recently been developed in a lung cancer study and has shown to cluster significant WSI features in a self-supervised manner, leading to promising results in terms of subtype classification and survival prediction[29]. Furthermore, HPL provides an additional layer of interpretability. Supervised approaches were also considered either as a baseline comparison or as a means to extract further information from the images (Supplementary Fig. 1b).

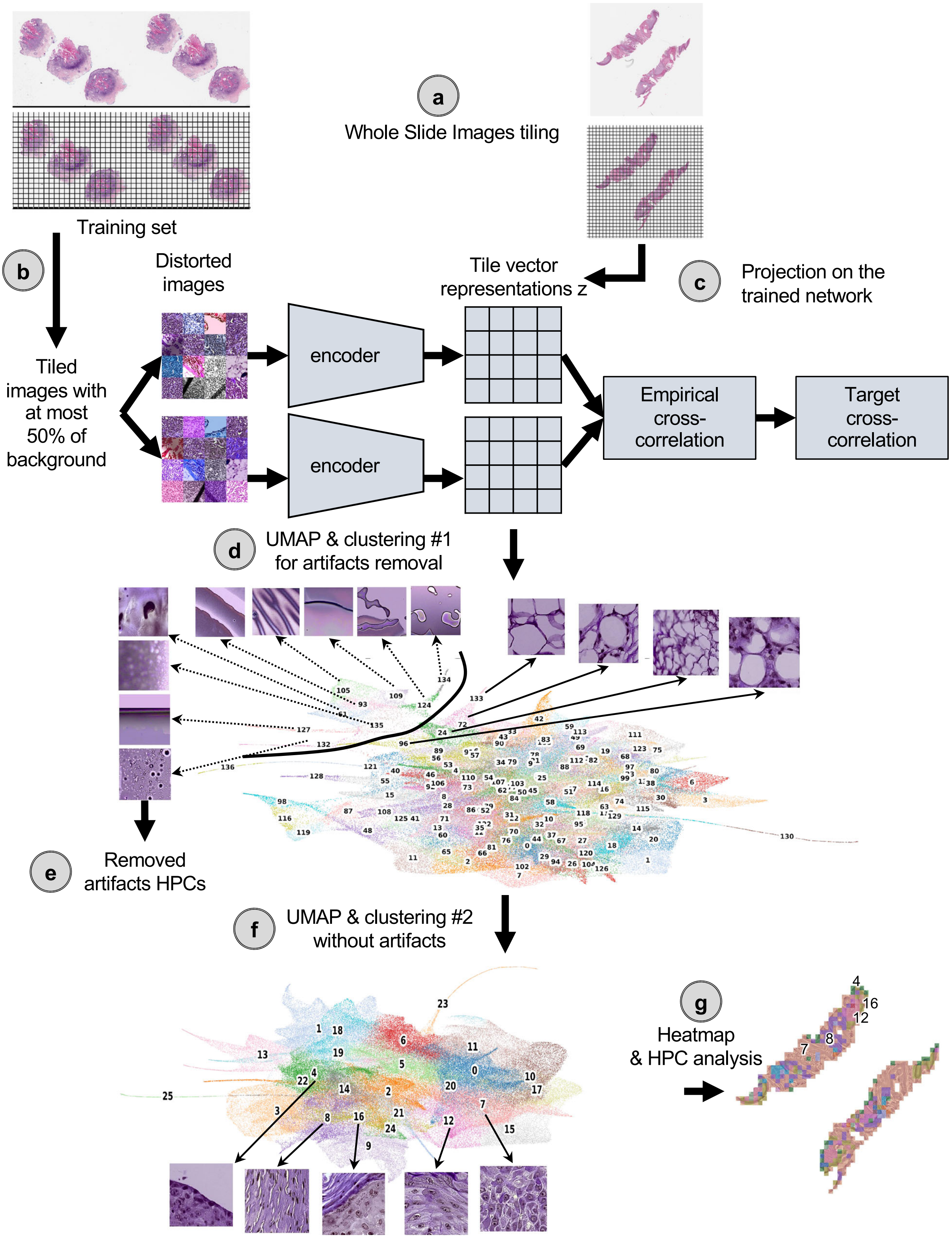
Results

Self-supervised learning highlights histomorphological phenotypes associated with poor and good outcomes

Slides from 167 patients used as the development cohort were collected from three institutions (Supplementary Table 1, Supplementary Fig. 1a): New York University (NYU), University of California San Francisco (UCSF) and Brigham and Women's Hospital (BWH), along with clinical information regarding whether the patient developed a good or poor outcome (see Method section, and Supplementary Fig. 2). As an external cohort, slides from 153 patients were received from the Complejo Asistencial Universitario de Salamanca (CAUSA) and 410 patients from the Mayo Clinic in Arizona (Mayo).

Clinical outcome data was notated as a binary label of good vs. poor outcome, with good outcome indicating no evidence of disease at most recent follow-up and PO representing local recurrence, nodal metastases, distant metastases or disease-specific death. Furthermore, disease-free survival (DFS) data were available for the NYU and UCSF development datasets (median follow-up, 38.0 and 32.7 months respectively), and for the CAUSA and Mayo test datasets (median follow-up, 51.5 and 41.9 months respectively), allowing us to perform Cox regression models with these cohorts. Further patient information regarding age, tumor stage, size, was not available for the UCSF and BWH cohorts. To objectively identify phenotypes which could potentially be predictive of PO, we used a pipeline based on the Barlow-Twins self-supervised approach[37] as described in Fig. 1 (see Method for details). This pipeline has been shown to successfully identify clusters of meaningful phenotypes on lung adenocarcinoma (such as different histological subtypes and types of tissues), and link them to overall and recurrence-free survival[29]. The Barlow-Twins model runs into two identical encoders, batches of samples originating from similar tiles but distorted in different ways. In trying to minimize the empirical cross-correlation between the embeddings on those two networks and the target cross-correlation, images projected into that network will eventually lead to tile vector representations that are more similar to each other if the corresponding images carry similar features. To simplify the analysis of the vector representations, tiles sharing similar phenotypes are clustered into groups called HPC (Histomorphological Phenotype Clusters) using the Leiden algorithm, which has the advantage of being a community detection algorithm with hierarchical clustering and can identify groups of nodes as well as connections between the entities.

Following the strategy explained in the method section to reduce overfitting, we froze a Leiden cluster configuration corresponding to a resolution $r = 0.75$, leading to splitting the dataset into 26 HPCs (Fig. 2a). A PAGA (PArtition-based Graph Abstraction) representation (Fig. 2b) illustrates how the clusters at the top of the graph and of the corresponding UMAP (Fig. 2c) appear more enriched in tiles associated with risk of poor outcome. As expected, at the selected resolution, phenotypes are present in different proportions and the relative size of HPCs vary considerably (Supplementary Figs. 3a, 4a, 5a). However, we can see that patients are well represented, and many of those phenotypes are present to various degrees in most of the patients (Supplementary Figs. 3b, 4b, 5b), with the exception of HPC 25. Similar observation is made regarding the representativity of the different institutions (Supplementary Figs. 3c, 4b, 5b, 6b).


a
Whole Slide Images tiling
Training set
b
Tiled images with at most 50% of background
Distorted images
encoder
encoder
Tile vector representations z
c
Projection on the trained network
Empirical cross-correlation
Target cross-correlation
d
UMAP & clustering #1 for artifacts removal
105
93
109
124
134
133
127
132
135
136
128
130
e
Removed artifacts HPCs
f
UMAP & clustering #2 without artifacts
23
1
18
6
13
19
11
5
0
22
4
14
2
20
10
17
25
3
8
16
21
7
24
12
15
9
g
Heatmap & HPC analysis
4
16
12
8
7

Using the HPL pipeline, each patient can be described by the percentage of tiles contained in each HPC, which in turn can be used as an input into regression models to estimate the HPC outcome prediction power. To assess the variability and impact of Leiden clustering on prediction of outcome, we also ran three-fold cross-validation log regression for binary classification of poor versus good outcome (Supplementary Fig. 6d) and a Cox regressions for survival prediction analysis (Supplementary Fig. 6e) at various resolutions. We observe that the resolution of r = 0.75 selected above also allows for successful predictions in both approaches while preventing overfitting.

Fig. 1 | Adaptation of the self-supervised Histological Phenotype Learning pipeline to study cutaneous squamous cell cancer. a The slides were first tiled into smaller images of 224 ×224 pixels at 0.5 um/pixel (equivalent to a magnification of 20×). **b** A subset of those tiles were used to train the self-supervised Barlow-Twins architecture. **c** Once trained, all the tiles from the three cohorts were then projected onto the trained network to extract their tile vector representations z, a 128 vector coding each image. **d** Those vector representations are then over-clustered using the Leiden approach in order to get homogeneous clusters (called Histomorphological Phenotype Clusters, HPC) and visually identify artifacts from tissue representations. In this UMAP of the tile vector representation z, each dot represents a tile, and each color a different HPC. **e** Tiles belonging to HPCs identified as highly enriched in artifacts are removed from the study. **f** The cleaned dataset is then subject to more detailed analysis and subjected to a new round of Leiden clustering. This UMAP of the cleaned tile vector representations z shows 26 HPCs corresponding to 26 groups of self-identified phenotypes, and representative tile for the top 5 clusters corresponding to the example slides in panel (**c**). **g** The resulting HPCs can then be used to generate heatmaps showing simplified slide representations and analyzed to identify potential correlations between those phenotypes identified by the self-supervised approach and patients' outcome. Here, the heat maps corresponding to the example slide section in panel (**a**) is shown, with the top 5 clusters numbered and corresponding to the ones in panel (**f**). All tiles are shown after Reinhard's color normalization[47].

Self-supervised learning identifies histomorphological phenotype clusters associated with survival

As mentioned, each patient can be described by the proportion of tiles assigned to each HPC, and this simplified phenotype description can in turn be used to study outcome predictability. Computing a univariate c-index analysis on each HPC independently, we observe some HPCs are either correlated with poor or good outcome (Fig. 2d), and for 10 of these HPCs, the trends are confirmed in the two external cohorts (Fig. 2e, Supplementary Fig. 7). Projected on the PAGA representation (Fig. 2f), we notice those related to poor outcomes are located on the upper part of the graph and those related to better outcomes are located on the lower part of the graph.

Using a multivariate log regression approach on the whole HPC vector describing the HPC distribution, we investigated the potential to predict the binary outcome from those HPCs. The good versus poor outcome binary classification run on the 3 development datasets available resulted in an average 3-fold cross-validation AUC of 0.724 (validation set) and 0.689 (test set; Supplementary Fig. 8a). While the performances were lower for the CAUSA cohort (AUC-0.554), they were very good on the Mayo external cohort which contains a mix of Mohs excisions, shave and punch biopsies (AUC ~ 0.844; Supplementary Fig. 8a). This result, as based on a self-supervised approach, allows for an unbiased selection of clusters of tiles which are identified as the most relevant to achieve the classification. It also provides an easy way to identify phenotypes important for such prediction, such as a forest plot analysis which shows the contribution of the main clusters to this prediction (Supplementary Fig. 8b).

As a comparison, we trained the supervised network inception v3, which relies on selected labeled regions. Manual free-text annotations based on consensus agreement were performed by three board certified Mohs micrographic surgeons (M.C., S.R.J, R.W.) and a senior dermatology resident (M.J.) using Aperio's Imagescope interface. Consensus was achieved if all reviewers agreed with the annotation features. We followed the pipeline similarly used by Johannet et al.[22] in their study of melanoma. Here, we first trained the algorithm (Supplementary Fig. 8c, e, Supplementary Tables 2, 3) to identify the regions manually annotated (normal skin, squamous cell carcinoma in situ, invasive squamous cell carcinoma, other, and artifact). Annotations in the 'other' group included dermis, fat, glandular tissue, smooth muscle, cartilage, inflammatory infiltrates, and the presence of ortho and hyperkeratosis. The artifact annotation included negative white space, bubbles or pen markings. Second, we checked whether a selected region (invasive cSCC) or a set of regions of interest (normal, in situ, invasive) could be used to predict a binary outcome (Supplementary Fig. 8d, f). Such a supervised approach achieved performances of AUC = 0.675 on invasive cSCC and 0.671 on the set of regions of interest (normal, in situ and invasive combined) for the development cohort. On the CAUSA and Mayo test cohorts, it achieved performances of AUC = 0.576 and 0.711, respectively, on invasive cSCC, and AUC = 0.598 and 0.726, respectively, on the set of regions of interest. In addition to performing slightly worse than the self-supervised approach (Supplementary Fig. 9 and more metrics in Supplementary Tables 4, 5), this two-step method relies on manual annotations, choices of regions from which the prediction should be made, and little possibility to directly interpret how the model made its decision or what subsets of phenotypes were used. These three bottlenecks are all addressed by the self-supervised approach as described in the next section.

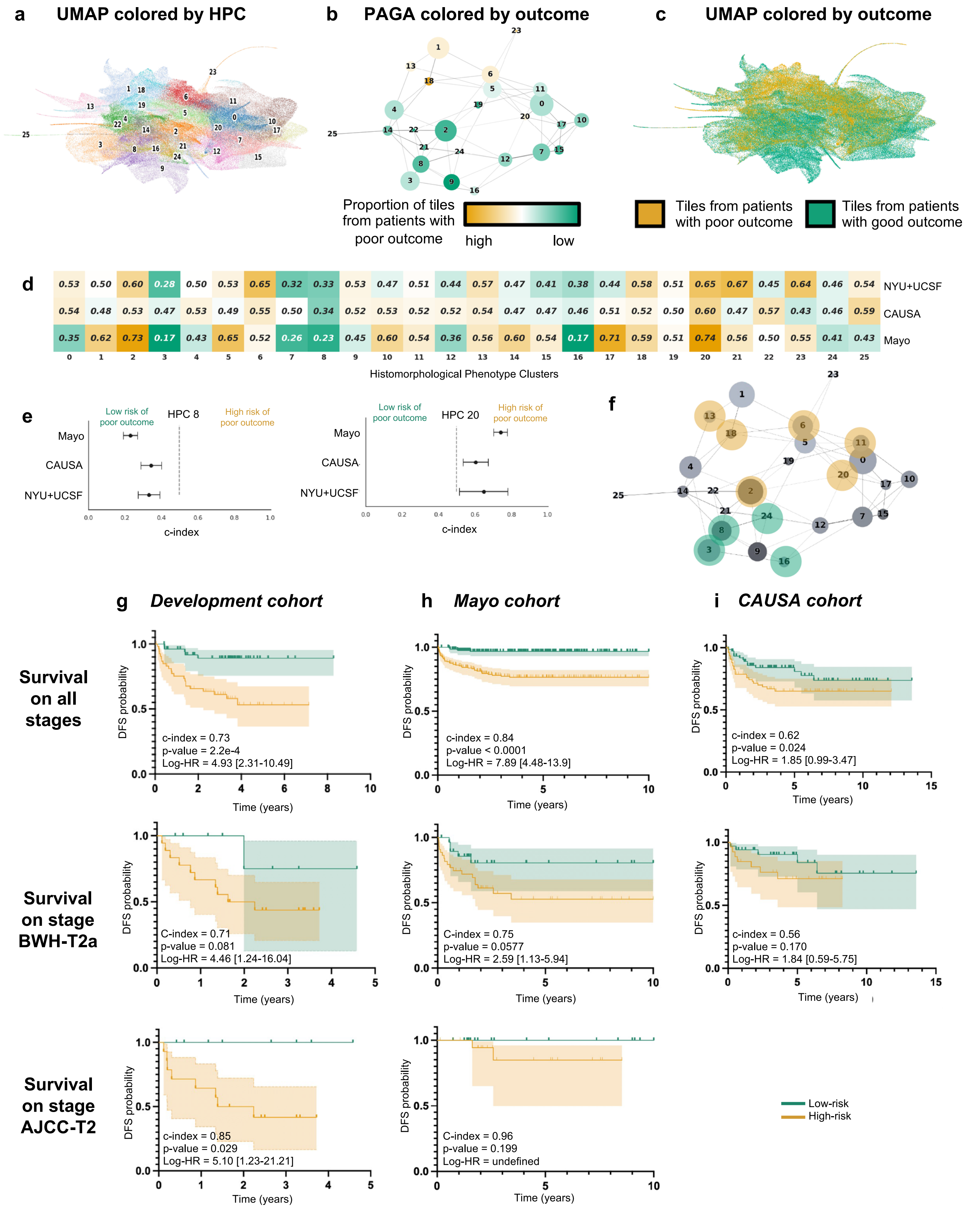
More interestingly, with the two datasets where the DFS data are available, we performed a Cox Regression and obtain a DFS prediction with a Harrell's c-index of 0.73 (0.72 Uno's c-index), and p-value of 2.2e−4 on the cross-validation cohort (Fig. 2g), a Harrell's c-index of 0.84 (0.83 Uno's c-index), and p-value < 0.0001 on the Mayo cohort (Fig. 2h), a Harrell's c-index of 0.62 (0.62 Uno's c-index), and p-value of 2.4e−2 on the CAUSA cohort (Fig. 2i). A forest plot (Supplementary Fig. 10a) and SHAP plot (Supplementary Fig. 10b) were computed to interpret the contributions of the different HPCs to this prediction and understand which phenotypes influence the model's prediction. Expectedly, most of the HPCs identified by the univariate approach appear also relevant in this multivariate approach. This approach is particularly effective at differentiating poor from good outcome in a subset of BWH T2a and AJCC-8 T2 tumors respectively (Fig. 2g–i), which currently face significant outcome heterogeneity thus rendering prognostication challenging. Note that the Kaplan–Meier plots extrapolated from the poor outcome probabilities generated by supervised classifiers show overall much lower performances (Supplementary Figs. 11, 12). This self-supervised approach could potentially address the gap in the current staging systems caused by the relative lack of outcome homogeneity within AJCC T2 and BWH T2a groups.

Cluster interpretation highlights histomorphological phenotypes linked with increased risk of local recurrence, metastasis and increased likelihood of good outcome

The HPL self-supervised approach provides interpretability power that allows us to identify how HPCs are weighted in terms of lower or higher risk of overall PO in our Cox regression model (DFS), or in terms of good/poor outcome in the log regression model. First, 100 random tiles were selected from each HPC and visually analyzed by three board certified Mohs micrographic surgeons (M.C., S.R.J, R.W.) and a senior dermatology resident (M.J.). The details of these observations are shown in Supplementary Table 6 (see Supplementary Fig. 13 for a subset of randomly selected tiles). When we project those observations on the Partition-based Graph Abstraction (PAGA Fig. 3), we observe that HPCs sharing similar phenotypes are linked and located in specific and coherent regions, confirming the coherence of the HPL representation. Further validation was done proceeding similarly on the external cohorts (Supplementary Figs. 14, 15).

In Fig. 4 and in Fig. 5, we show examples of HPCs and tiles associated with higher and lower risk of PO by our Cox model from Fig. 2g.

In Fig. 4a we show randomly selected tiles from the HPCs having the highest influence on the c-index for prediction of higher risk of PO, and in Fig. 4b, c, three heatmaps of patients with PO show the HPCs composition projected on sections of the WSI. For the patient in Fig. 4b, for example, the tumor recurred after 10.5 months and its associated WSI demonstrated high enrichments in HPCs 1, 6 and 20, the latter two with a high SHAP log hazard ratio value synonymous with higher risk of PO. The HPC 6 shows deep invasion of poorly differentiated keratinocytes with mitoses. HPC 20 shows poorly differentiated and pleomorphic keratinocytes with mitoses, and HPC 1 similarly demonstrates poorly differentiated keratinocytes, which have been shown in prior studies to correlate with POs, such as local recurrence[38].



The WSI in Fig. 4c shows a high presence of HPCs 0, 1, 5, 6 and 13, the latter 2 showing a high SHAP log hazard ratio value. HPC 13 shows some pleomorphic features as well as deeply infiltrative tumor cells, of which is a feature associated with POs. Similarly HPC 6 demonstrates deep invasion, poor differentiation, and significant atypia, factors that are recognized to contribute to POs[38]. The SHAP decision plots resulting from our study show how the enrichment or depletion of tiles from certain HPCs weigh into the final decision for a patient whose tumor recurred shortly after surgery (LR, Fig. 4b) and for a patient recurring 46 months after surgery (NM, Fig. 4c).

In Fig. 5a we show randomly selected tiles for the HPCs having the highest influence on the c-index for prediction of a lower-risk of PO. In Fig. 5b, an analysis shows clusters of HPCs which tend to be adjacent on a

Fig. 2 | Unsupervised approach generates clusters enriched in tiles from patients with poor outcome, with good representation of the three cohorts, and predicting disease-free survival while providing tile clusters important for that prediction. **a** UMAP with the 26 Leiden clusters found at resolution 0.75. **b** PAGA representation of the Leiden clusters with node connections. The size of the nodes is proportional to the number of tiles and their color is proportional to the proportion of tiles associated with good/poor outcome patients. **c** UMAP with colors showing tiles associated with good/poor outcome patients (green/orange). Each dot is a tile. **d** Univariate analysis comparing the c-index (average of a 3-fold cross validation) for the prediction of the RFS for the development cohort (NYU + UCSF) and on the external cohorts (CAUSA, Mayo). c-index below 0.5 (green) indicates lower risk of poor outcome, while c-index above 0.5 (orange) indicates higher risk of poor outcome. **e** Details of panel c for two clusters where the development cohort and the external cohort show the same trend (See Supplementary Fig. 6 for all clusters). Error bars show the confidence interval. **f** Projection on the PAGA of the HPCs showing coherent trends for both the cross-validation on the development cohort, and on the external cohorts. **g** Kaplan–Meier curve of predicted high and low risk patients of having a poor outcome from the unsupervised HPL approach using a Cox regression, 3-fold cross-validation on the development cohort (NYU + UCSF). First row is computed using the whole dataset, while second and third show a subset of patient with stage T2a (BWH staging) and T2 (AJCC staging) only. Error bars show 95% confidence interval (CI). 95% CI of hazard ratio (logrank) is shown between brackets. The median value computed on the whole dataset is used to split low from high risk patients. **h** Same as g but using the Mayo as a test cohort. **i** Same as g but using the CAUSA as a test cohort.

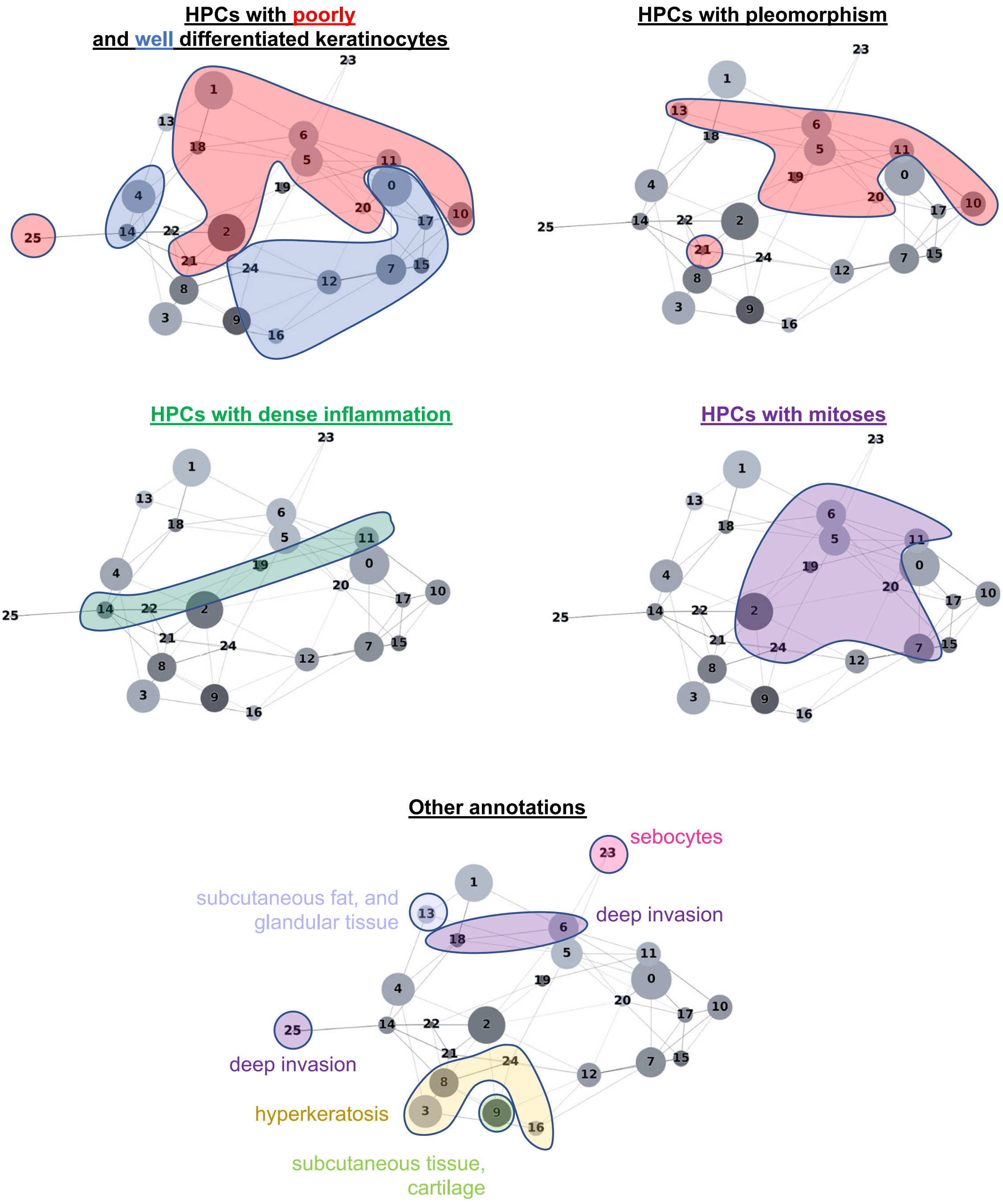

HPCs with poorly and well differentiated keratinocytes
HPCs with pleomorphism
HPCs with dense inflammation
HPCs with mitoses
Other annotations
sebocytes
subcutaneous fat, and glandular tissue
deep invasion
deep invasion
hyperkeratosis
subcutaneous tissue, cartilage

Fig. 3 | PAGA graph shows a coherent organization of features found on cutaneous squamous cell carcinoma whole slide images. Annotations provided by a group of Mohs surgeons, of which included tiles randomly selected from the development cohort (NYU + UCSF + BWH) for each HPC (annotation taken from Supplementary Table 6) and are projected on the PAGA graph from Fig. 2b.

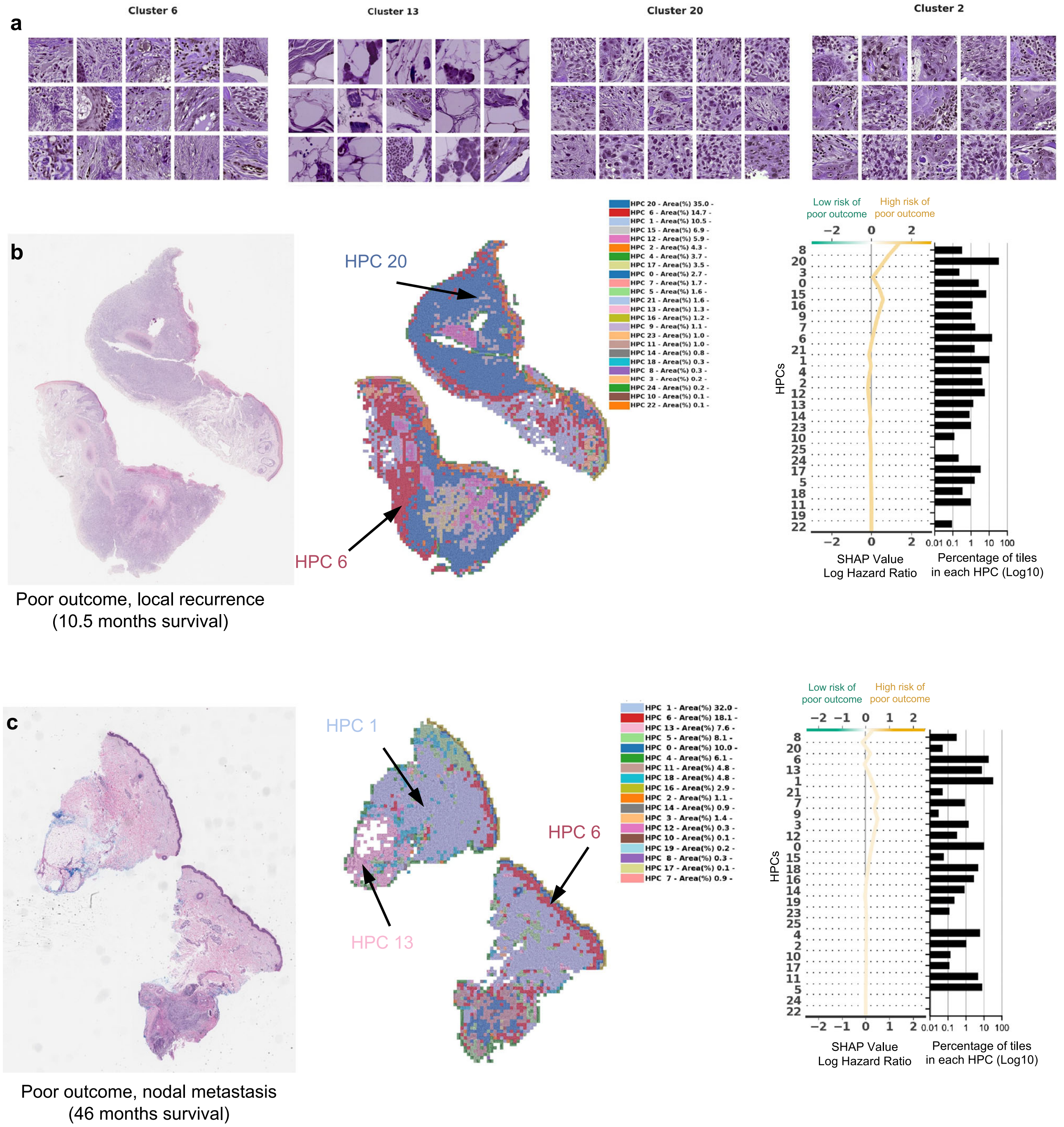


Fig. 4 | Example of tiles from HPCs associated with higher risk of poor outcome. **a** Example of tiles randomly selected from certain HPCs leading to risk prediction of poor outcome. **b**, **c** Examples of data from patients with poor outcome shortly after surgery (10.5 months, local recurrence) and with poor outcome a few years after surgery (46 months, nodal metastasis). For each case, a small portion of the original slide is shown as well as the corresponding heatmap and the associated SHAP decision plot. The color of the heatmap shows the HPC associated with each tile, with the proportion of tile belonging to each HPC shown in the legend (percentages computed over the whole slide(s) available for each patient). The top of the SHAP decision plot shows the predicted value which determines the color of the curve. Reading from bottom to top, the SHAP values for each HPC are cumulatively summed, and the HPCs are ordered according to the absolute SHAP weight. On the right, the proportion of tiles associated with each cluster is shown on a Log10 scale. All tiles are shown after Reinhard's color normalization[47].

slide, by checking, for each tile belonging to a given HPCs, what HPCs were associated with its adjacent tiles. Two groups of HPCs identified as lower risk of PO were seen as having relatively high interactions: HPCs 7, 12 and 16 as one group and HPCs 3, 8 and 24 as the other. A commonality of all these HPCS was the presence of well-differentiated keratinocytes and the lack of atypia or pleomorphism, which would be expected for a tumor with good outcome. In Fig. 5c, d, we show the slides associated with two patients who were followed up for more than three years with no evidence of disease. Those slides show the presence of HPCs 3, 8 and 24 seen in Fig. 5b, as well as HPCs 7, 12 and 16. Similarly, it is likely that the good differentiation, lack of pleomorphism and relatively non-specific features of hyperkeratosis can be attributed to these findings.

Because the HPCs that are associated with good outcome are those that have normal appearing keratinocytes, intuitively, the slide selection could impact the resulting percentage of tiles and therefore final classification. To explore this, we first generated a dendrogram with the heatmap color

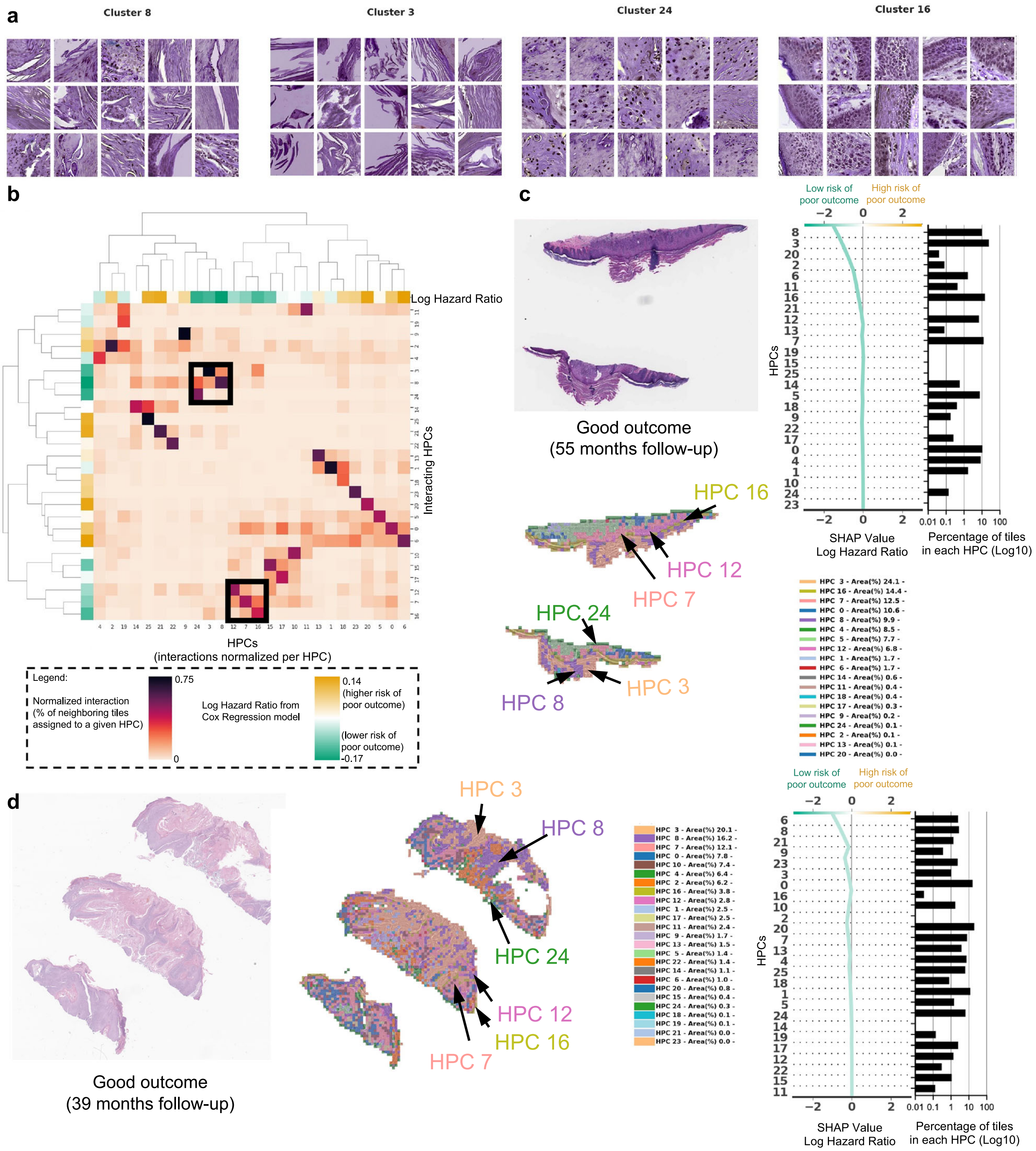


Fig. 5 | Example of tiles from HPCs associated with lower risk of poor outcome. **a** Example of tiles randomly selected from certain HPCs leading to prediction of good outcome. **b** The interaction analysis between HPCs shows two groups of HPCs which tend to be adjacent on slides; each column shows the normalized proportion of interactions each tile associated with a given HPC has with HPCs associated with its adjacent tiles. The dendrograms correspond to bi-hierarchical clustering of HPCs. **c**, **d** Examples of data from patients who have not recurred and have been followed for more than three years. For each case, a small portion of the original slide is shown as well as the corresponding heatmap and the associated SHAP decision plot. The color of the heatmap shows the HPC associated with each tile, with the proportion of tile belonging to each HPC shown in the legend (percentages computed over the whole slide(s) available for each patient). The top of the SHAP decision plot shows the predicted value which determines the color of the curve. Reading from bottom to top, the SHAP values for each HPC are cumulatively summed, and the HPCs are ordered according to the absolute SHAP weight. On the right, the proportion of tiles associated with each cluster is shown on a Log10 scale. All tiles are shown after Reinhard's color normalization[47].

showing the HPC composition per slide for all patients with multiple slides, and performed hierarchical clustering (Supplementary Fig. 16a), demonstratingthat for most cases, slides belonging to the same patient are clustered nearby showing similar overall composition. Then for each patient, we computed the variance, among their own slides, on the percentage of tiles associated with each cluster (intra-patient variance), and normalized it by the variance across patients (inter-patient variance of tile percentages associated with each cluster). Supplementary Fig. 16b, c shows that the ratio intra-patient variance versus inter-patient variance is above 1 for most patients and most HPCs, meaning more homogeneity among slides belonging to the same patient. Finally, we computed per slide instead of per patient SHAP decision plots (Supplementary Fig. 16d), showing that the conclusions differ only slightly among slides from a given patient.

Other differences and potential confounding factors between the development and the test set are the type of biopsy and the source of the anatomic site. While the NYU cohort had nearly as many slides from excision as from shave biopsies, the CAUSA test cohort is exclusively composed of wide local excision, and the Mayo cohort, on the other hand, is mostly dominated by shave biopsy (Supplementary Table 1). We notice that most HPCs contain tiles from all types of preparations (Supplementary Figs. 3e, 4d). In some rare HPCs, the relative proportion of tiles between excision and biopsies varies more (Supplementary Figs. 3d, 4c): for example, HPCs 1 and 13, located at the top left of the PAGA graph and contain subcutaneous fat/tissues, are relatively more rare in shave biopsies than excisions, which is consistent considering that the depth of the shave samples is thinner than excisions which are often full-thickness. While the survival prediction pipeline seems to perform well on both the shave and excision biopsies specimens from the Mayo cohort (Supplementary Fig. 17a–c), we cannot rule out that with larger training set, a biopsy-type specific Cox Regression approach would lead to even better performances, specially for wide local excisions specimens like those used in the CAUSA cohort. Finally, because the anatomic sites are so diverse compared to the number of patients available, conclusions regarding potential biases are more difficult to draw but we notice as well that most HPCs contain at least a few tiles from most sites (Supplementary Figs. 3f, g, 4e, f, 5c, d). At this stage, we did not identify either impact from the tissue source on the survival prediction, though larger per source cohorts would be needed (Supplementary Fig. 17d, e).

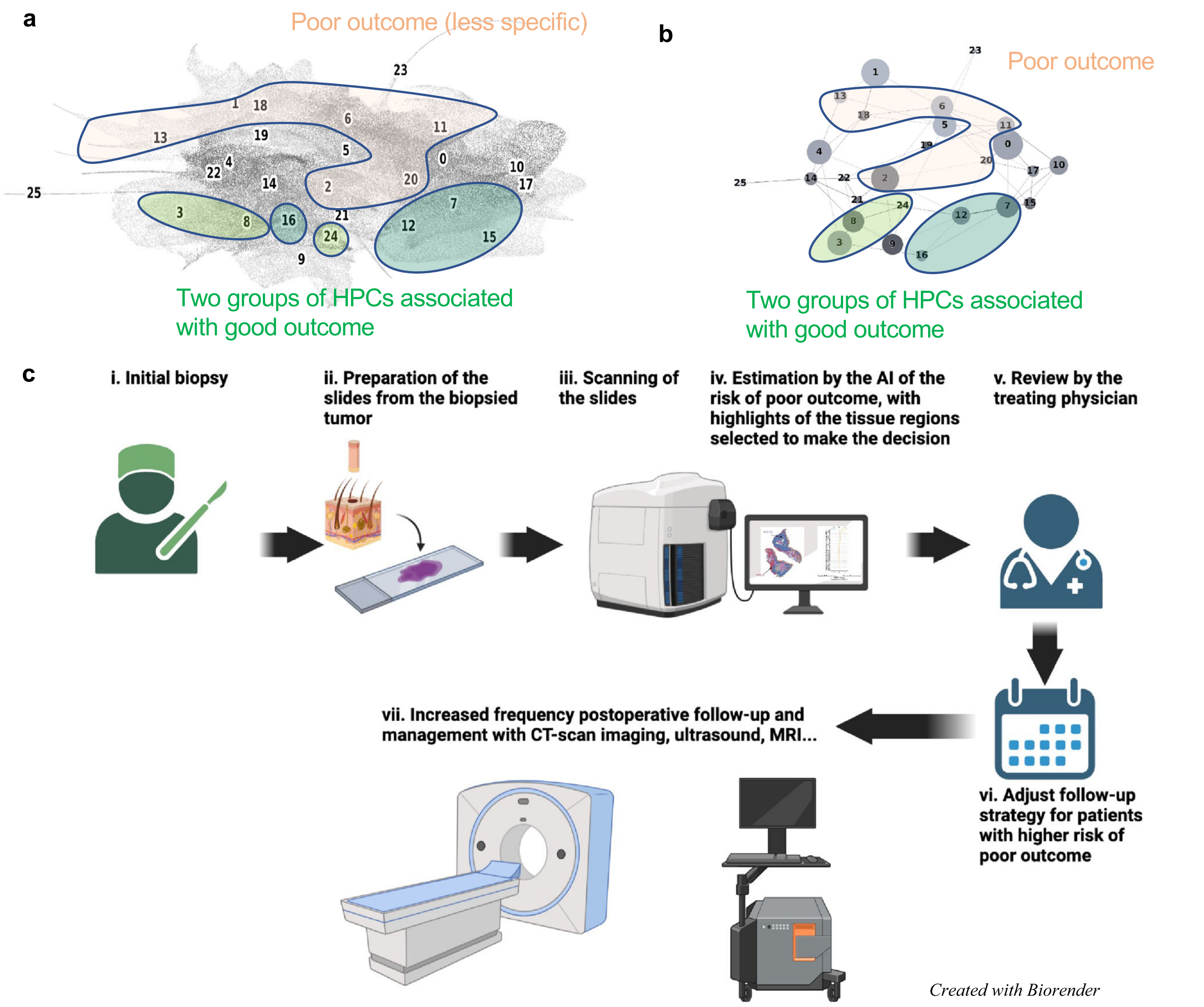
Discussion

In this study, we demonstrate that the self-supervised approach of HPL can be successfully applied to analyze sets of cSCC WSIs from initial biopsy samples from different institutions, grouping in a coherent manner a variety of histopathological features linked to good and POs. The ability to obtain prognostic information from biopsy slides alone is significant as it may guide clinical decision making regarding treatment and surveillance of patients with potential PO. The HPCs identified were used to predict the good versus poor outcome with an area under the curve (AUC) of ~0.7 and the disease-free survival (DFS) with a c-index of 0.73 (p-value = 2.2e−4) on the cross-validation of the development cohort, and AUC ~ 0.58–0.73 and c-index~0.62–0.84 on the test cohorts. The performance remains compelling in a subset of AJCC-8 T2 and BWH T2a tumors (c-indexes of 0.85 and 0.71, respectively, on the cross-validation cohort; 0.96 and 0.75, respectively, on the Mayo cohort; and 0.56 for the BWH T2a on the CAUSA cohort). The performance achieved here can also be placed in the context of other methods to which it could be potentially combined to further refine the precision of the outcome. For example, Zhao et al.[19] showed that protein expression of AXIN2 and SNAIL have a c-index of 0.69 in predicting recurrence-free survival, and that, although their available clinicopathological data alone had little prediction power (c-index of 0.40), the c-index was increased to 0.75 when combining clinicopathological with the protein expression. Using supervised deep-learning architectures, few studies have explored the predictability of cSCC from WSIs, all of which were unable to pinpoint which features were used by the algorithm to make the decision. Focusing on prediction of metastasis in a cohort of 104 patients harboring cSCC, Knuutila et al.[25] achieved AUCs within 0.629–0.689, performing better (AUC = 0.747) when restricting the study to those that recurred rapidly (within 180 days). This study was done using either the whole slide or tumor regions manually annotated by pathologists. On the other hand, using two cohorts of 54 melanoma patients, Comes et al.[24] achieved AUCs = 0.667–0.695 in predicting the one-year disease free survival using regions of interest manually pre-selected by pathologists. In addition to its performance, the advantages of the HPL pipeline, initially developed on lung cancer[29], is that its training is self-supervised and does not require any manual pre-annotations. Furthermore, it provides an additional layer of interpretation highlighting which phenotypes weighed in favor of higher or lower risk prediction.

In our study we found that enrichment in HPCs 3 and 8 correlated with a prediction of good outcome from cSCC. Both HPCs demonstrate well-differentiated keratinocytes, lack of atypia or pleomorphism, and the relatively non-specific presence of hyperkeratosis. Alternatively, enrichment in HPCs 7 and 16, which feature well-differentiated cells and lacked pleomorphism, deep invasion, or significant atypia, also correlated, to a lesser extent towards a lower risk of PO. Additionally, among the HPCs identified that correlated with higher risk of PO, the two major phenotypes identified were severe pleomorphism with poor differentiation and deep invasion. This result is consistent not only with previous studies but also with the current BWH and AJCC-8 staging systems[38]. In a recent meta analysis, Zakhem et al. revealed that tumors with invasion beyond the subcutaneous fat were associated with a statistically significant risk of LR and DSD[39–41]. Moreover, ulcerated tumors, poorly differentiated tumors, PNI, lymphovascular invasion, desmoplastic stroma and immunosuppression are all significantly associated with POs[38,42]. These attributes have been defined as key defining features of aggressive behavior by cSCC and some are considered in staging. It is for this reason that the ability of our machine learning algorithm to detect these features, at time of biopsy, is significant. Our system may offer a standardized method for feature identification given the potential for inherent inter-reader variability in identification of high risk histopathologic features by dermatopathologists. Poor differentiation and invasion beyond the subcutaneous fat have been associated with an increased risk of metastasis[43]. More interestingly, the different clusters identified in this study and their association with certain types of outcomes are located in well-defined and coherently connected regions of the UMAP and of the PAGA graphs (Fig. 6a, b). The two sets of HPCs associated with good outcome (HPCs 3, 8 and 24 as one group, and 7, 12 and 16 on the other) and identified as having high numbers of interactions on the slides are each connected in the PAGA and located on the lower right side of the UMAP. On the other hand, the top side of the UMAP is dominated by HPCs associated with POs.

In this study, we uniquely used a self-supervised learning followed by community-based clustering to predict cSCC-free survival, while describing the phenotypes of the clusters weighing the most in these findings. Predictive clusters for PO included those with poor differentiation whereas lack thereof, enrichment in non-specific hyperkeratosis without atypia tended to favor prediction of good prognosis. Importantly, we demonstrate significant potential to optimize clinical decision-making in that this approach is particularly efficient at differentiating PO risk in low stage tumors (BWH T2a and AJCC-8 T2 tumors). This addresses a large gap in the literature relating to outcome homogeneity between low stage tumors (BWH T2a and AJCC T2) given that ~25% of POs occur in low stage tumors[15]. Gupta et al. previously evaluated risk factors for poor outcomes in stage T2a cSCC and identified a predictive model for those at risk of poor outcomes using major and minor criteria with high specificity (97.4%) but low sensitivity (7.7%) (15). Thus, while this model is highly accurate at identifying tumors without poor outcomes, it does not perform as well at identifying tumors that will develop poor outcomes. Additionally, this study was not externally validated. While direct comparison of these models is outside the scope of this study, we hope that our findings may add to the literature demonstrating an adjunctive method for accurately identifying patients at risk for poor outcome on an externally validated cohort. Overall, we believe incorporation of these data along with existing clinical information may augment identification of highest risk patients and would allow for rationally based, focused clinical follow up that may lead to the development of algorithms for further imaging and work up.

Our results suggest that prognostication of cSSC can benefit from self-supervised learning to not only assist clinicians in predicting outcomes but also highlight histomorphological patterns associated with these outcomes. These findings play a significant role in patient care, as prognostic information from initial biopsy slides alone may guide clinical decision making with regard to diagnostic workup, treatment and surveillance of patients with high risk for PO. Clinicians may use the information gained from self-supervised learning as an adjunct in clinical decision making and assigning


a
Poor outcome (less specific)
Two groups of HPCs associated with good outcome
b
Poor outcome
Two groups of HPCs associated with good outcome
c
i. Initial biopsy
ii. Preparation of the slides from the biopsied tumor
iii. Scanning of the slides
iv. Estimation by the AI of the risk of poor outcome, with highlights of the tissue regions selected to make the decision
v. Review by the treating physician
vi. Adjust follow-up strategy for patients with higher risk of poor outcome
vii. Increased frequency postoperative follow-up and management with CT-scan imaging, ultrasound, MRI...

Created with Biorender

Fig. 6 | Specific HPCs are correlated with poor outcome. a, **b** Projection on the UMAP and PAGA graph of the HPCs associated with high and low risk of poor outcome. **c** Ultimately, we anticipate such a deep-learning tool, which identifies patients at higher risk with poor outcome and provides histomorphological interpretability, could assist treating physicians in making decisions on an increased post-operative follow-up and management strategy. Panel (**c**) created with biorender.com.

pre-test probabilities for patients that might be at higher risk for PO and benefit from further workup and management. (Fig. 6c). For example, patients identified at high-risk may be deemed appropriate for pre-operative imaging, more frequent follow up or removal of a primary tumor with complete margin control and enhanced pathologic staging provided by Mohs micrographic surgery[44,45].

Development of, and access to, large datasets will be crucial to further validate and expand the current study. Ultimately, the ability to assess the risk of PO at time of initial diagnosis could provide the basis to establish and test diagnostic and therapeutic protocols that could ultimately optimize clinical outcome.

Methods

Ethics approval

The NYU study number is 20-01740 and is classified as non-human research, therefore was not subject to IRB review at NYU. UCSF received approval for expedited IRB review (protocol #21-34087) and BWH (protocol# 2021P000701). The cohort from the Complejo Asistencial Universitario de Salamanca was approved by the local IRB. The cohort from the Mayo Clinic was subject to IRB review (ID #21-012833, from Clinicopathologic and Multi-Omic Stratification of Cutaneous Squamous Cell Carcinoma).

Dataset

Datasets of shave or punch biopsy specimens were collected from three institutions each with separate IRB approval processes (Supplementary Table 1, Supplementary Fig. 1a): 119 slides from 42 patients were collected at New York University (NYU), 95 slides from 95 patients were collected at the University of California San Francisco (UCSF) and 40 slides from 40 patients were collected at the Brigham and Women's Hospital (BWH). For cases at NYU, multiple slides from a single lesion were analyzed per patient. For slides from BWH and UCSF, single slides of the highest yield resolution were obtained for ease of logistical coordination between sites. Due to lack of information or poor slide quality (e.g., out of focus), fourteen slides were removed from the study (Supplementary Fig. 2). We therefore ended up including slides from 163 patients diagnosed with cSCC on initial biopsy who developed good or poor outcomes (NYU: 38 patients, $n = 31$ and 7, respectively; UCSF: 85 patients, $n = 58$ and 27, respectively; BWH: 40 patients, $n = 20$ and 20, respectively). The size of the training cohort was determined by the samples available in the institutions selected to compose the training cohort, and due to the limited amount available, it was determined by sample size calculation approaches. To benefit from the best resolution available, whole slide images were scanned on an Aperio AT2 scanner and captured at 0.25 um/pixel at 40 X using JPEG2000 compression and stored as a svs pyramidal file.

De-identified slides from other collaborating institutions were sent to NYU Langone Health's Dermatologic Surgery & Cosmetic Associates Office. The scanned images did not contain any patient information. De-identified samples were simply classified as good versus POs. All de-identified physical slides were stored at NYU Langone Health's Dermatologic Surgery & Cosmetic Associates Office until the analysis was complete. Upon completion of analysis, the slides were returned to the original institution.

Adult patients 18–89 years old with existing slides of biopsy proven cSCC obtained prior to January 1, 2021 were included. Patients were excluded if they were outside the specified age range and had no histological confirmation of cSCC. Patients treated at NYU Langone Health's Dermatologic Surgery & Cosmetic Associates Office were identified for inclusion by a NYULH study team member based on the above inclusion and exclusion criteria above. Patients at NYU were identified using retrospective chart review of those with poor outcome, and thereon manual review of slides were performed to select those with the best slide quality. Patients treated at collaborating institutions were identified by individuals at those institutions based on the inclusion and exclusion criteria.

Patients were classified as having a PO if the tumor was successfully treated but the tumor came back at any time in the future, either in the form of local recurrence (LR), nodal metastasis (NM), distant metastasis (DM) or if the patient had a disease-specific death (DSD). Otherwise, if no tumor was detected at subsequent visits, the patients were classified as having a good outcome. In total, 119 patients were associated with good outcomes and 44 with POs. For UCSF and NYU, the times to LR, NM, DM or DSD were also available, giving further granularity into the disease-free survival (DFS) analysis. However, this data was not available for the BWH cohort.

In addition, two external cohorts were obtained (Supplementary Table 1): 156 slides from 153 patients (112 good outcome, 41 poor outcome) from the Complejo Asistencial Universitario de Salamanca (CAUSA) scanned with MoticEasyScan One (Motic, Hong Kong) and 411 slides from 410 patients (455 good outcome, 55 poor outcome) from the Mayo Clinic (Mayo) scanned with a Leica's GT450 scanner (Leica Biosystems).

Self-supervised-based analysis

The self-supervised-based study was based on the Histomorphological Phenotype Learning (HPL) through self-supervised learning and community detection pipeline developed by Quiros et al.[29] and summarized as follows and in Fig. 1 (and Supplementary Fig. 1). Using the DeepPATH tools[46], images from the 3 datasets were first tiled (removing those where the background covers more than 75% of the tile, and applying color normalizing using the Reinhard's method[47]), and converted to a h5 file such as each tile fed to the self-supervised pipeline has a field of view of 224 ×224 pixels at a pixel size of about 0.5 um (corresponding to a magnification of 20×, Fig. 1a). The 2,069,052 resulting tiles were split such that 40% of the tiles from each dataset were combined and used to train the self-supervised Barlow-Twins algorithm based network[37] (Fig. 1b). After training (Supplementary Fig. 6a), all the tiles were projected into the 128 dimension z tile representation vector of the trained network (Fig. 1c) and are represented by UMAPs[48] and PAGA[49] in this manuscript (Fig. 1d). As a filtering step, a first Leiden clustering[50] was achieved using a resolution of 7 in order to obtain a large ($n = 136$) number of Histomorphological Phenotype Clusters (HPCs) and over-cluster and increase the chance of having homogeneous HPCs. Those HPCs were visually inspected to identify those containing artifacts (air bubbles, blurring, dust, etc. Fig. 1d), with the goal to remove from the rest of study the tiles from HPCs representing artifacts. We identified 9 clusters containing artifacts which were removed from the dataset. Those artifacts are all contained within regions protruding from the rest of the UMAP (Fig. 1e), and were exclusively artifacts without any underlying tissue. The resulting 1,998,932 tiles were then used for the rest of the study. Next, another round of Leiden clustering was applied to the remaining tiles (Fig. 1f), and each HPC was mapped back to the slide of each patient (Fig. 1g). Each patient is therefore described by a patient vector representation which is embedded in the percentage of tiles associated with each HPC.

Considering the small development dataset, the analyses were done using a 3-fold cross-validation approach to study the variability of the approach while allowing each set to have enough samples. In each fold, a different third was used as a test set, while the remaining tiles are split between training (80%) and validation (20%). To ensure representativity and proper split, folds were generated randomly with a single constraint on the RFS to ensure each train and test sets have similar Kaplan–Meier profiles. After Leiden clustering, each whole slide image (WSI) can be represented by a codebar called a WSI vector representation which describes the distribution of tiles in each HPC. When a patient has more than one slide available, those can also be aggregated into a "patient vector representation". Logistic and/or Cox proportional hazards regressions have been run using the patient vector representations from the training sets, and evaluated using the validation and test sets left. Similar to the previous study on lung cancer[29], the performance was analyzed on a set of cluster configurations via n-fold cross validation to estimate variability at a given Leiden resolution, the Wald test being used to measure the significance on each regression and using Fischer's method to combine the p-values. Once done, we locked down a fold for further analysis.

As the resolution parameter r of the Leiden clustering algorithm is increased, the UMAP appears as split into more HPCs (Supplementary Fig. 6c, green curve). However, as the number of HPC increases and gets smaller, in terms of average number of tiles, the risk of obtaining institution or patient-specific clusters increases (Supplementary Fig. 6c, purple and cyan curves), which would be a sign of over-fitting. Indeed, increasing the number of clusters too much increases the risk of detecting features which are patient or institution specific (which may be caused by the fact that the 3 cohorts were stained in 3 different institutions and scanned on 2 different scanners, or may be related to some other phenotypes specific to natural variations between individuals). However, we are interested in finding common patterns across the three institutions, with enough meaningful (or compact) clusters to describe the diversity of common patterns found in this disease. Therefore, to select the best Leiden cluster resolution for the subsequent analysis of the HPCs, we checked, for each resolution: 1- the average patient and institution presence in HPCs (see details below); 2- the performance of the binary classifier (good versus poor outcome) via the AUC (Area Under the receiver operating Curve) of the logistic regression approach; 3- the performance of the Cox Regression for survival prediction. The average patient presence (Supplementary Fig. 6c) is defined as the average percentage of patients present in the HPCs at a given resolution, either counting all patients even if only 1 of their tiles belong to a certain HPC, or using a 1% threshold. Similarly, the institution presence (Supplementary Fig. 6c) is defined as the average percentage of institutions present in the HPCs at a given resolution, either counting all institutions, or only counting those with at least 1% of their tiles associated with a given HPC. Despite the small size of our cohort and limited number of institutions involved, it allows us to get a sense of the potential generalization of the study, and these averages will tend to get smaller as the HPCs become more patient or institution specific. We notice that at resolutions higher than $r = 0.75$, these averages decrease, showing that more HPCs become specific and less generalizable across patients and institutions.

Binary classification between good and poor outcome was done using all three development samples, while survival analysis data was only available for the NYU and UCSF datasets. Those analyses were done using a three-fold cross validation approach (Supplementary Fig. 6d, e), using folds consistent with those used for Leiden cluster determination to study report influence on variations between different Leiden clustering runs (Supplementary Fig. 2). The logistic and cox regressions were done following the approach detailed in Quiros et al.[29]. Briefly, WSI vector representations were built for each patient to describe the percentage of tiles associated with each

HPC, and center log-ratio transformation was applied to use those in linear models. A three fold cross-validation analysis was performed such as, for each fold, one third of the patients is used as a test set, and the rest is used for the training/validation process. For each fold, the regressions were fit using the training set and assessed with the validation and test sets.

Elastic-net penalty models were used for regression where we optimized the alpha parameter (final value of 0.25) for the logistic regression analysis, and the alpha and l1 ratio parameters (to final values of 0.35 and 0.01 respectively) for the Cox regression analysis. After having locked a cluster configuration, the medium of the hazard predictions on the training set was used to define the threshold between the low and high risk groups used on the test set and shown in Fig. 2g. The statistical significance between the two groups is measured using the log rank test and a *p*-value threshold of 0.05.

In addition to the cross-validation results on the development cohort, the generalizability of those trained networks was tested by inferences on the two external cohorts, CAUSA and Mayo.

Cluster analysis

UMAPs[48] (Uniform Manifold Approximation and Projection) and PAGA[49] (partition-based graph abstraction) were used to visualize the tile vector representations and resulting Leiden clusters. PAGA provides an additional layer of interpretability by preserving the topology where edges between the nodes denote statistically relevant connectivity between HPCs.

For each HPC, 100 tiles were randomly selected and visually interpreted (blinded from the positions of the HPCs on the PAGA) by three board certified Mohs surgeons and a senior dermatology resident (M.C., S.R.J., R.W.) who freely annotated and labeled each of them (no features to choose from were supplied). Consensus annotations, which arrived if all reviewers agreed with the annotation features, are shown in Supplementary Table 6. Annotations were then mapped into the PAGA (Fig. 3 for the development cohort, and, for the external test cohorts, in Supplementary Figs. 14, 15), where interesting connections between nodes can be seen despite the pathologists' annotations having been done without knowledge of the PAGA.

Analyses of the correlation between the clusters and external annotations were done following the HPL pipeline[29]: SHAP (SHapley Additive exPlanations) and Forest plots were used to evaluate how each HPC affects the log odds ratio of patients. The SHAP values were calculated across each test set of 3-fold cross-validation analyses. The Forest plots are based on the log hazard ratio of Cox proportional hazards model over the train sets of a 3-fold cross-validation. The coefficients were averaged across fold and combined *p*-values with Fisher's combined probability test. Correlations with pathologic diagnostic and type of recurrence (LR or overall metastases) was achieved using Spearman's rank correlation with a significance threshold of 0.01 on the *p*-values (adjusted with the Benjamini/Hochberg[51] method for false discovery rate). Overall metastases include both nodal and distant ones.

Furthermore, Cox proportional hazards regressions univariate analysis was performed on each cohort using a 3-fold cross validation approach.

Supervised analysis

To explore whether the performance of the outcome prediction in a supervised manner, we used DeepPATH[46] and followed an approach comparable to the one used to predict response to Melanoma treatment in Johannet et al.[22], training inception v3[52] twice at a magnification of 20x: first to automatically segment the slides, second to predict the outcome from selected segmented regions. For the segmentation, a 3 and 5-class network were trained. In the 3-class approach explored, the network was trained to identify the following classes: regions of interest, artifacts and other features (muscle, bone, cartilage, hair follicles, nerve…). The goal was therefore to simply be able to sort out the artifacts and other features regions judged irrelevant by the team to later predict outcome. In the 5-class approach, the network was designed to split more precisely the "regions of interest", and it was therefore trained to identify the following classes: invasive SCC, in-situ SCC, normal epidermis, artifacts and other features. Next, we trained a network to study the predictability of the good versus poor outcome using the regions of interest only, or using the invasive SCC only.

Data availability

Reasonable requests for cohort data may be addressed to the corresponding authors.

Code availability

The codes are written in Python and available on github. The supervised approach relies on the DeepPATH (https://github.com/ncoudray/DeepPATH). The un-supervised approach relies on the Histomorphological Phenotype Learning pipeline (https://github.com/AdalbertoCq/Histomorphological-Phenotype-Learning) and more details on the code and options used in this study are reported in https://github.com/ncoudray/AI-analysis-of-cutaneous-squamous-cell-carcinoma/.

Received: 13 November 2023; Accepted: 4 February 2025;
Published online: 15 February 2025

References

1. Lomas, A., Leonardi-Bee, J. & Bath-Hextall, F. A systematic review of worldwide incidence of nonmelanoma skin cancer. *Br. J. Dermatol.* **166**, 1069–1080 (2012).
2. Waldman, A. & Schmults, C. Cutaneous Squamous Cell Carcinoma. *Hematol. Oncol. Clin. North Am.* **33**, 1–12 (2019).
3. Stang, A. et al. Incidence and mortality for cutaneous squamous cell carcinoma: comparison across three continents. *J. Eur. Acad. Dermatol. Venereol.* **33**, 6–10 (2019).
4. Leiter, U. et al. Incidence, Mortality, and Trends of Nonmelanoma Skin Cancer in Germany. *J. Invest. Dermatol.* **137**, 1860–1867 (2017).
5. Van Lee, C. B. et al. Recurrence rates of cutaneous squamous cell carcinoma of the head and neck after Mohs micrographic surgery vs. standard excision: a retrospective cohort study. *Br. J. Dermatol.* **181**, 338–343 (2019).
6. Lansbury, L., Bath-Hextall, F., Perkins, W., Stanton, W. & Leonardi-Bee, J. Interventions for non-metastatic squamous cell carcinoma of the skin: systematic review and pooled analysis of observational studies. *BMJ* **347**, f6153–f6153 (2013).
7. Rogers, H. W., Weinstock, M. A., Feldman, S. R. & Coldiron, B. M. Incidence Estimate of Nonmelanoma Skin Cancer (Keratinocyte Carcinomas) in the U.S. Population, 2012. *JAMA Dermatol.* **151**, 1081–1086 (2015).
8. Karia, P. S., Han, J. & Schmults, C. D. Cutaneous squamous cell carcinoma: Estimated incidence of disease, nodal metastasis, and deaths from disease in the United States, 2012. *J. Am. Acad. Dermatol.* **68**, 957–966 (2013).
9. Eigentler, T. K. et al. Survival of Patients with Cutaneous Squamous Cell Carcinoma: Results of a Prospective Cohort Study. *J. Invest. Dermatol.* **137**, 2309–2315 (2017).
10. Stevenson, M. L. et al. Use of Adjuvant Radiotherapy in the Treatment of High-risk Cutaneous Squamous Cell Carcinoma With Perineural Invasion. *JAMA Dermatol.* **156**, 918–921 (2020).
11. Lewis, K. G. & Weinstock, M. A. Trends in nonmelanoma skin cancer mortality rates in the United States, 1969 through 2000. *J. Invest. Dermatol.* **127**, 2323–2327 (2007).
12. American Cancer Society. Facts & Figures. Paperpile, https://paperpile.com/app/p/47ca1717-2267-0081-9d88-750bb97346e9 (2023).
13. Karia, P. S., Morgan, F. C., Califano, J. A. & Schmults, C. D. Comparison of Tumor Classifications for Cutaneous Squamous Cell Carcinoma of the Head and Neck in the 7th vs 8th Edition of the AJCC Cancer Staging Manual. *JAMA Dermatol.* **154**, 175–181 (2018).
14. Amin, M. B. et al. The Eighth Edition AJCC Cancer Staging Manual: Continuing to build a bridge from a population-based to a more

'personalized' approach to cancer staging. *CA Cancer J. Clin.* **67**, 93–99 (2017).

15. Gupta, N. et al. Identifying Brigham and Women's Hospital stage T2a cutaneous squamous cell carcinomas at risk of poor outcomes. *J. Am. Acad. Dermatol.* **86**, 1301–1308 (2022).
16. Work Group et al. Guidelines of care for the management of cutaneous squamous cell carcinoma. J. Am. Acad. Dermatol. **78**, 560–578 (2018).
17. Jennings, L. & Schmults, C. D. Management of high-risk cutaneous squamous cell carcinoma. *J. Clin. Aesthet. Dermatol.* **3**, 39–48 (2010).
18. Wysong, A. et al. Validation of a 40-gene expression profile test to predict metastatic risk in localized high-risk cutaneous squamous cell carcinoma. *J. Am. Acad. Dermatol.* **84**, 361–369 (2021).
19. Zhao, G. et al. AXIN2 and SNAIL expression predict the risk of recurrence in cutaneous squamous cell carcinoma after Mohs micrographic surgery. *Oncol. Lett.* **19**, 2133–2140 (2020).
20. Yanofsky, V. R., Mercer, S. E. & Phelps, R. G. Histopathological variants of cutaneous squamous cell carcinoma: a review. *J. Skin Cancer* **2011**, 210813 (2011).
21. Yacob, F. et al. Weakly supervised detection and classification of basal cell carcinoma using graph-transformers on whole slide images. *Sci Rep* **13**, 7555, https://doi.org/10.1038/s41598-023-33863-z (2023).
22. Johannet, P. et al. Using Machine Learning Algorithms to Predict Immunotherapy Response in Patients with Advanced Melanoma. *Clin. Cancer Res.* **27**, 131–140 (2021).
23. Kim, R. H. et al. Deep Learning and Pathomics Analyses Reveal Cell Nuclei as Important Features for Mutation Prediction of BRAF-Mutated Melanomas. *J. Invest. Dermatol.* **142**, 1650–1658.e6 (2022).
24. Comes, M. C. et al. A deep learning model based on whole slide images to predict disease-free survival in cutaneous melanoma patients. *Sci. Rep.* **12**, 20366 (2022).
25. Knuutila, J. S. et al. Identification of metastatic primary cutaneous squamous cell carcinoma utilizing artificial intelligence analysis of whole slide images. *Sci. Rep.* **12**, 9876 (2022).
26. Sali, R. et al. Deep Learning for Whole-Slide Tissue Histopathology Classification: A Comparative Study in the Identification of Dysplastic and Non-Dysplastic Barrett's Esophagus. *J. Personal. Med.* **10**, 141 (2020).
27. Dimitriou, N., Arandjelović, O. & Caie, P. D. Corrigendum: Deep Learning for Whole Slide Image Analysis: An Overview. *Front. Med.* **7**, 419 (2020).
28. Shmatko, A., Ghaffari Laleh, N., Gerstung, M. & Kather, J. N. Artificial intelligence in histopathology: enhancing cancer research and clinical oncology. *Nat Cancer* **3**, 1026–1038 (2022).
29. Quiros, A. C. et al. Self-supervised learning in non-small cell lung cancer discovers novel morphological clusters linked to patient outcome and molecular phenotypes. https://arxiv.org/pdf/2205.01931.pdf (2022).
30. Chen, Z., Li, X., Yang, M., Zhang, H. & Xu, X. S. Optimization of deep learning models for the prediction of gene mutations using unsupervised clustering. *Hip Int.* **9**, 3–17 (2023).
31. Kim, J. et al. Author Correction: Unsupervised discovery of tissue architecture in multiplexed imaging. *Nat. Methods* **19**, 1662 (2022).
32. Chen, R. J. et al. Scaling Vision Transformers to Gigapixel Images via Hierarchical Self-Supervised Learning. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2022).
33. Poon, A. I. F. & Sung, J. J. Y. Opening the black box of AI-Medicine. *J. Gastroenterol. Hepatol.* **36**, 581–584 (2021).
34. Van der Laak, J., van der Laak, J., Litjens, G. & Ciompi, F. Deep learning in histopathology: the path to the clinic. *Nat. Med.* **27**, 775–784 (2021).
35. Guidotti, R. et al. A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv.* **51**, 1–42 (2018).
36. Rajpurkar, P., Chen, E., Banerjee, O. & Topol, E. J. AI in health and medicine. *Nat. Med.* **28**, 31–38 (2022).
37. Zbontar, J., et al. Twins: Self-Supervised Learning via Redundancy Reduction. In: *Proceedings of the 38th International Conference on Machine Learning* (eds. Meila, M. & Zhang, T.) vol. 139 12310–12320 (PMLR, 2021).
38. Zakhem, G. A., Pulavarty, A. N., Carucci, J. & Stevenson, M. L. Association of Patient Risk Factors, Tumor Characteristics, and Treatment Modality With Poor Outcomes in Primary Cutaneous Squamous Cell Carcinoma. *JAMA Dermatol.* **159**, 160 (2023).
39. Brancaccio, G. et al. Risk Factors and Diagnosis of Advanced Cutaneous Squamous Cell Carcinoma. *Dermatol. Pract. Concept* **11**, e2021166S (2021).
40. Que, S. K. T., Zwald, F. O. & Schmults, C. D. Cutaneous squamous cell carcinoma: Incidence, risk factors, diagnosis, and staging. *J. Am. Acad. Dermatol.* **78**, 237–247 (2018).
41. Rowe, D. E., Carroll, R. J. & Day, C. L. Jr. Prognostic factors for local recurrence, metastasis, and survival rates in squamous cell carcinoma of the skin, ear, and lip. Implications for treatment modality selection. *J. Am. Acad. Dermatol.* **26**, 976–990 (1992).
42. Campoli, M., Brodland, D. G. & Zitelli, J. A prospective evaluation of the clinical, histologic, and therapeutic variables associated with incidental perineural invasion in cutaneous squamous cell carcinoma. *J. Am. Acad. Dermatol.* **70**, 630–636 (2014).
43. Schmults, C. D., Karia, P. S., Carter, J. B., Han, J. & Qureshi, A. A. Factors Predictive of Recurrence and Death From Cutaneous Squamous Cell Carcinoma. *JAMA Dermatol.* **149**, 541 (2013).
44. Gibson, F. T., Murad, F., Granger, E., Schmults, C. D. & Ruiz, E. S. Perioperative imaging for high-stage cutaneous squamous cell carcinoma helps guide management in nearly a third of cases: A single-institution retrospective cohort. *J. Am. Acad. Dermatol*. **88**, 1209–1211 (2023).
45. Canavan, T. N. et al. A cohort study to determine factors associated with upstaging cutaneous squamous cell carcinoma during Mohs surgery. *J. Am. Acad. Dermatol.* **88**, 191–194 (2023).
46. Coudray, N. et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat. Med.* **24**, 1559–1567 (2018).
47. Reinhard, E., Adhikhmin, M., Gooch, B., Shirley, P. & July-Aug. Color transfer between images. *IEEE Comput. Graph. Appl.* **21**, 34–41 (2001).
48. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.* **3**, 861 (2018).
49. Wolf, F. A. et al. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol.* **20**, 59 (2019).
50. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* **9**, 5233 (2019).
51. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Methodological* **57**, 289–300 (1995).
52. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2016).

Acknowledgements

The authors thank the team of the Center of Biospecimen Research and Development from the New York University for their help in digital pathology and histology for this project. The Center of Biospecimen Research and Development, RRID:SCR_018304, is partially supported by the Cancer Center Support Grant P30CA016087 at the Laura and Isaac Perlmutter Cancer Center. Parts of the computational analysis for this work were supported by the NYU Langone High Performance Computing (HPC) Core's resources, and we thank the HPC team for their support. We would like to thank the Genome Technology Center (GTC) for expert library preparation and sequencing, and the Applied Bioinformatics Laboratories (ABL) for providing bioinformatics support and helping with the analysis and interpretation of the data. GTC and ABL are shared resources partially supported by the Cancer Center Support Grant P30CA016087 at the Laura and Isaac Perlmutter Cancer Center. This work has used computing resources at the NYU School of Medicine High Performance Computing (HPC) Facility. Other funding Sources include NCI/NIH Cancer Center Support Grant P30CA016087 (A.T.). A.C.Q. is supported by a scholarship from School of Computing Science, University of Glasgow.

Author contributions

N.C. and A.C.Q. designed and executed the experiments, and wrote the python codes. N.A.D helped obtain IRB approval and provided logistic coordination between all institutions. M.C.C., S.R.J., R.W. and M.C.J. provided slide annotations and histological assessment of HPCs. N.C., M.C.J. and M.C.C. prepared the manuscript which was later edited or approved by all co-authors. N.C., A.C.Q, K.Y. and A.T provided deep-learning expertise. M.C.C., S.R.J., R.W., M.C.J., and J.A.C. provided squamous cell cancer expertise. R.W, M.L.S., D.M.K, S.Y., J.C., F.M., C.D.S., C.D.C.M., J.C., A.C., A.N.H., A.S., Z.L.-R. and A.R.M. provided the slides and corresponding data. J.A.C. and A.T. jointly supervised the study.

Competing interests

The authors declare the following competing interests: A.T. is a co-founder of Imagenomix; N.C. is a scientific advisor for Imagenomix. The other authors declare that they have no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-025-01496-3.

Correspondence and requests for materials should be addressed to Aristotelis Tsirigos or John A. Carucci.

Reprints and permissions information is available at http://www.nature.com/reprints