nature medicine

Article

https://doi.org/10.1038/s41591-023-02643-7

A population-level digital histologic biomarker for enhanced prognosis of invasive breast cancer

Received: 17 May 2023

Accepted: 13 October 2023

Published online: 27 November 2023

Check for updates

Mohamed Amgad[1], James M. Hodge[2], Maha A. T. Elsebaie[3], Clara Bodelon[2], Samantha Puvanesarajah[2], David A. Gutman[4], Kalliopi P. Siziopikou[1], Jeffery A. Goldstein[1], Mia M. Gaudet[5], Lauren R. Teras[2,6] & Lee A. D. Cooper[1,6]

Breast cancer is a heterogeneous disease with variable survival outcomes. Pathologists grade the microscopic appearance of breast tissue using the Nottingham criteria, which are qualitative and do not account for noncancerous elements within the tumor microenvironment. Here we present the Histomic Prognostic Signature (HiPS), a comprehensive, interpretable scoring of the survival risk incurred by breast tumor microenvironment morphology. HiPS uses deep learning to accurately map cellular and tissue structures to measure epithelial, stromal, immune, and spatial interaction features. It was developed using a population-level cohort from the Cancer Prevention Study-II and validated using data from three independent cohorts, including the Prostate, Lung, Colorectal, and Ovarian Cancer trial, Cancer Prevention Study-3, and The Cancer Genome Atlas. HiPS consistently outperformed pathologists in predicting survival outcomes, independent of tumor–node–metastasis stage and pertinent variables. This was largely driven by stromal and immune features. In conclusion, HiPS is a robustly validated biomarker to support pathologists and improve patient prognosis.

- 病理学家将乳腺组织的微观外观划分等级，是定性的，不考虑肿瘤环境中的非癌元素

- 提出组织预后签名HiPS，对乳腺肿瘤微环境引起的生存风险进行全面、可解释的评分

- map细胞和组织结构，来测量上皮、间质、免疫和空间互作特征

- population-level开发，3个队列验证

Breast cancer is the most common malignancy worldwide[1,2]. It is a heterogeneous disease with highly variable survival outcomes that depend on tumor biology, therapeutic regimen, and socioeconomic determinants of health[3,4]. Established prognostic criteria include the American Joint Committee on Cancer (AJCC) tumor–node–metastasis (TNM) staging, Nottingham histologic grading, and intrinsic subtype. Intrinsic subtype can be determined by gene expression profiling or approximated using assessment of the estrogen receptor (ER), progesterone receptor (PR), or the human epidermal growth factor receptor 2 (HER2) expression based on immunohistochemistry (IHC) or in situ hybridization (ISH; Fig. 1a). Starting 2018, the AJCC manual introduced the 'prognostic stage', which combines the traditional TNM stage with ER, PR, and HER2 status and Nottingham grade[3]. This shift in consensus reflects the necessity and complexity of combining multimodal information to place patients along a risk spectrum[5].

In this paper, we present an artificial intelligence system to improve the prognostication of patients with nonmetastatic invasive carcinomas of the breast. We devised a Histomic Prognostic Signature (HiPS) risk score that is consistent with AJCC staging, which combines modalities available to every pathologist: slides stained with

预后标准：

- AJCC 肿瘤-淋巴结-转移；

- 诺丁汉组织学分级；

- 内在subtype通过基因表达谱；免疫组织化学IHC、原位杂交对雌激素受体ER、孕激素受体PR、人表皮生长因子受体HER2来估计

TNM+ER、PR、HER2+Nottingham，结合多模式信息

[1]Department of Pathology, Northwestern University Feinberg School of Medicine, Chicago, IL, USA. [2]Department of Population Science, American Cancer Society, Atlanta, GA, USA. [3]Department of Medicine, John H. Stroger, Jr. Hospital of Cook County, Chicago, IL, USA. [4]Department of Pathology, Emory University School of Medicine, Atlanta, GA, USA. [5]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD, USA. [6]These authors jointly supervised this work: Lauren R. Teras, Lee A.D. Cooper. e-mail: lee.cooper@northwestern.edu

https://doi.org/10.1038/s41591-023-02643-7

- 建立入乳腺癌临床预测指标
- AJCC定义三种以治疗为导向的亚型

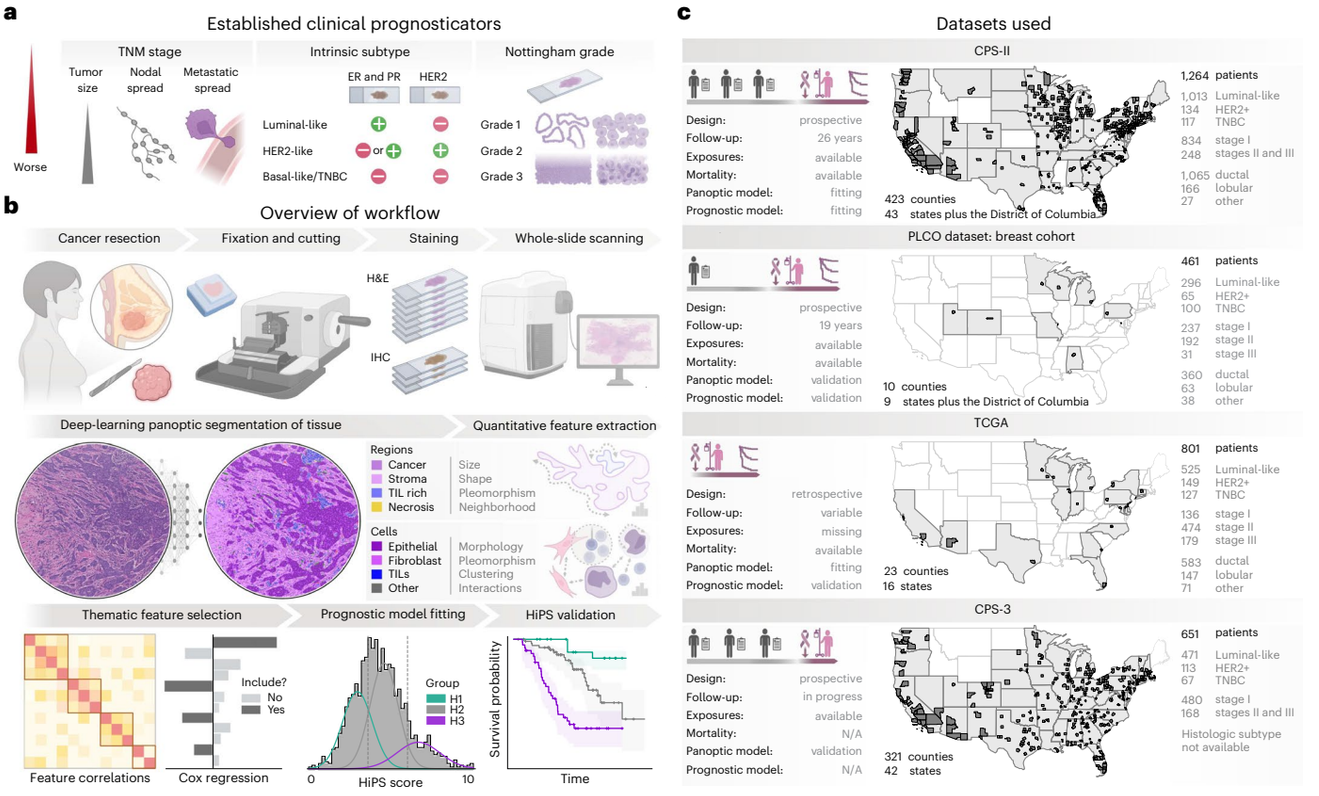


Fig. 1 | Overview of the methodological approach and datasets used. **a**, Established clinical prognosticators in breast cancer. The AJCC staging manual defines three treatment-oriented subtypes: patients with Luminal-like cancer are eligible for hormone therapy, patients with HER2-like cancer are eligible for trastuzumab, and patients with TNBC are not eligible for targeted therapies. We limited our analysis to invasive cancers, not metastatic, at the time of diagnosis. All specimens were routinely assessed using the standard IHC and SH panel (ER, PR, and HER2) and H&E-stained slides. **b**, Our workflow for determining the HiPS. Breast cancer resection specimens were fixed in formalin, embedded in paraffin, cut, stained, and digitally scanned. A panoptic segmentation model identified tissue regions and nuclei in each slide, followed by the computational extraction of interpretable morphologic features. These features include stromal, immune, and spatial interaction features not included in Nottingham grading. The most prognostic features within each biologic theme, combined with ER, PR, and HER2, were used to fit a Cox regression model to cancer-specific survival data. The resultant HiPS score is an interpretable weighted combination of histologic features. In addition, we learned thresholds to identify three distinct prognostic groups. Finally, we validated HiPS using clinical, genomic, and epidemiologic data. **c**, An overview of the datasets used. N/A indicates not applicable. We include patients from almost all geographic regions of the United States, covering 614 counties in 48 states, plus the District of Columbia. CPS-II data were used for prognostic model fitting, and PLCO, TCGA, and CPS-3 were independent validation cohorts. Prediagnostic risk factor exposure data were available for all datasets except TCGA, while survival outcomes were available for all datasets except CPS-3. TCGA and PLCO specimens were exclusively sourced from tertiary medical centers, unlike the CPS datasets, which were mostly sourced from non-tertiary and community hospitals.

首先，使用基于深度学习的计算机视觉对whole-slide images进行定量评估，减轻手动诺丁汉分级中固有的变异性并捕获无法可靠分级的潜在特征

hematoxylin and eosin (H&E) and the ER, PR, and HER2 panel. HiPS addresses three issues with the current standard of care. First, we use advanced deep-learning-based computer vision for quantitative assessment of whole-slide images (WSIs), providing an objective alternative that mitigates variability inherent in manual Nottingham grading and captures latent features that cannot be reliably graded.

其次，全面评估肿瘤微环境TME

- cancer-associated fibroblasts(CAFs) 癌症相关成纤维细胞
- tumor-infiltrating lymphocytes(TILs) 肿瘤浸润淋巴细胞

Second, we comprehensively assess the entire tumor microenvironment (TME) including nonneoplastic elements (Supplementary Figs. 1–5). The Nottingham grading criteria, which assign patients into grades G1–G3, were standardized in the 1930s to the 1990s[6–10]. In the decades since, it has become clear that cancer progression involves multiple pathologic processes, including cancer-permissive inflammation, activation of wound healing and repair cascades, immune modulation, hypoxia-driven metabolic derangements, epithelial-to-mesenchymal transition, and spatio-geometric changes that enable the invasion of cancer cells[11]. Many of these hallmarks are reflected in changes in the density, appearance, and spatial clustering of cancer-associated fibroblasts (CAFs), tumor-infiltrating lymphocytes (TILs), and stromal matrix[12–19]. HiPS measures these morphologic changes using a comprehensive, interpretable set of features not captured by visual grading.

- 癌症允许性炎症
- 激活伤口愈合和修复
- 免疫调节
- 缺氧引起的代谢紊乱
- 上皮间质转化
- 空间几何变化

第三：
- ER、PR、HER2与组织学特征结合，准确评估TME形态的预后重要性，同时考虑内在亚型影响

Third, we combine the standard ER, PR, and HER2 panel with histologic features in a unified framework. This integration enables us to accurately gauge the prognostic importance of the TME morphology while accounting for the effect of intrinsic subtype. Indeed, several recent studies showed that tumor morphology differs with hormone receptor expression[20–23].

HiPS was developed using a large prospective cohort from the Cancer Prevention Study-II (CPS-II) of the American Cancer Society (ACS). Cancer-free participants were enrolled from the general population, unselected for specific characteristics[24]. CPS-II participants who developed breast cancer during cohort follow-up were the patients included in our analysis. We validated HiPS using three independent cohorts, including the Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial, The Cancer Genome Atlas (TCGA), and the Cancer Prevention Study-3 (CPS-3) of the ACS [25–27]. Collectively, the patients were diagnosed in hundreds of US healthcare facilities, ranging from large academic cancer centers to rural facilities (Fig. 1c and

- 这些患者是在数百个美国医疗机构中接受诊断的，从大型学术癌症中心到农村机构

https://doi.org/10.1038/s41591-023-02643-7

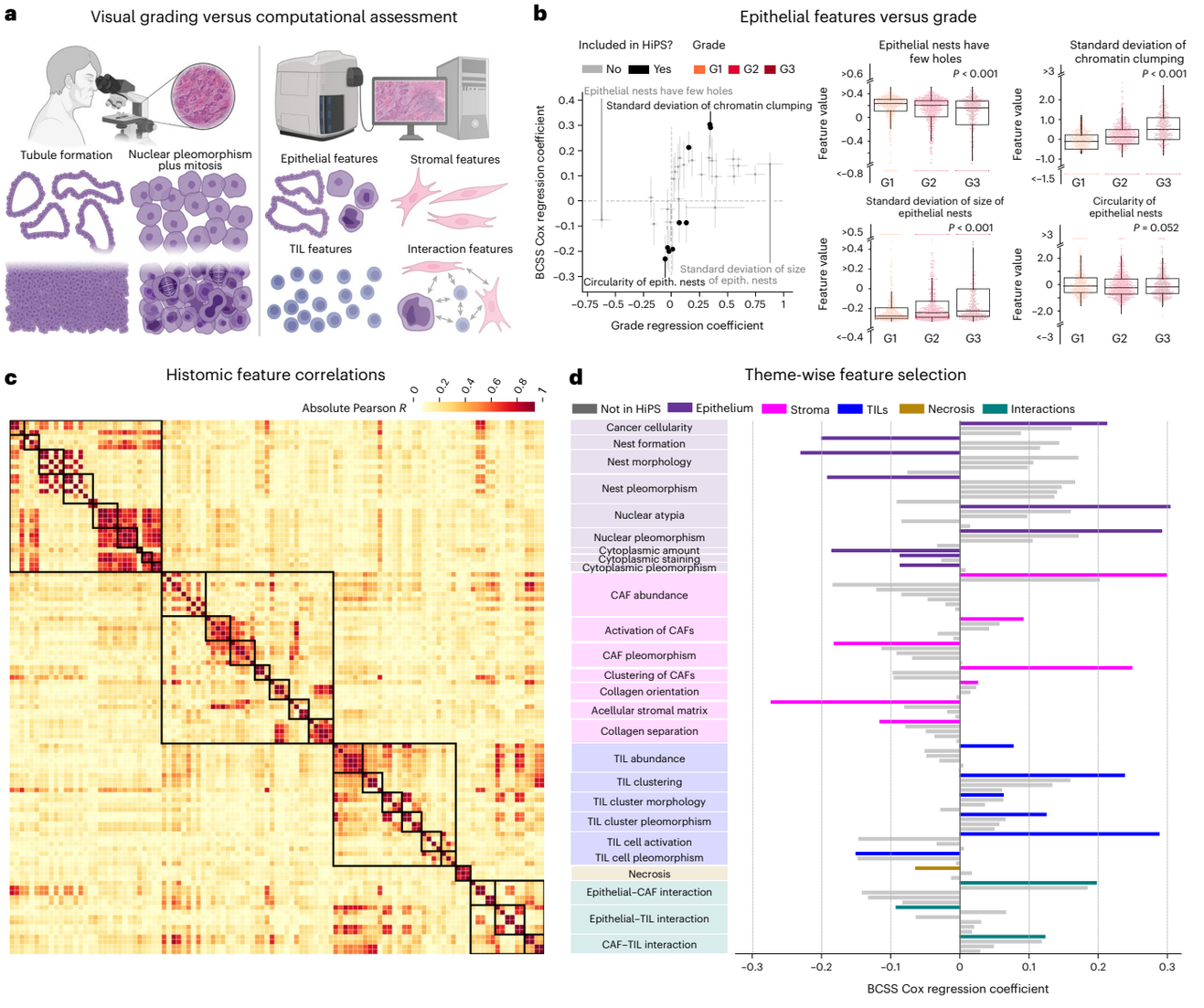


Fig. 2 | Thematic categorization and selection of features using the CPS-II cohort. Supporting results are provided in Supplementary Tables 23–25. **a**, Conceptual differences between visual grading and computational assessment. Pathologists use the Nottingham grading criteria, a visual semiquantitative aggregate score of epithelial tubule formation, nuclear pleomorphism, and mitotic figures. In contrast, our models quantitatively assess the entire TME, including stromal and immune cells, stromal matrix, and spatial interactions. **b**, Left: association of epithelial histomic features with visual grading (ordinal regression) versus association with survival (Cox regression). Error bars represent the standard error. Right: box plots of the feature value distributions of two histomic features highly associated with grade (left) and two highly associated with survival (right). Feature selection for HiPS was guided by the association with survival, not grade. In fact, the standard deviation of epithelial nest size closely captures grades, yet is only modestly prognostic compared with alternative epithelial features. For each box plot, each dot is a single patient, and $n = 467$ (G1), 693 (G2), and 464 (G3) patients. Box plot center line = median; box limits = upper and lower quartiles; whiskers = 1.5× interquartile range. *P* values correspond to the Kruskal–Wallis test, and the exact values are 8.35×10^{-1} (top left), 5.04×10^{-38} (top right), and 2.03×10^{-14} (bottom left). **c**, Inter-feature correlations. The squares represent biological themes and subthemes. Except for interaction features, cross-theme correlations are mostly weak, reflecting the independence of different themes. **d**, Univariable Cox regression coefficients for all 109 histomic features. The most prognostic feature within each of the 26 subthemes was included in the HiPS.

Supplementary Tables 1–4 and 29). This diversity supports the broad applicability of HiPS as a generalizable biomarker.

A key advantage of HiPS is interpretability and transparency, being composed of features that correspond to recognizable, well-established biological entities. This enabled us to investigate the biological phenomena underlying HiPS by correlating morphology with pathology reports, messenger RNA (mRNA) expression data, and inferred pathway activations and cell abundance using genomic deconvolution methods. We found that HiPS features are surrogates for high genomic instability, a hypoxic immune microenvironment with activation of wound healing pathways, a myofibroblastic CAF (myCAF) phenotype, and a suboptimal immune response deplete of $CD8^+$ T lymphocytes.

- 可解释性和透明度

Results

Panoptic segmentation for extracting breast TME biomarkers

Before the widespread adoption of deep learning, automated grading workflows relied on detecting tissue regions and nuclei based on shape, texture, and contextual heuristics[28–30]. This approach was largely replaced by deep-learning methods that learn features of relevance

自动分级流程依赖于基于形状、纹理、上下文启发法来检测组织区域和细胞核

深度学习方法可解释性不高

- 显著性热图；LIME；决策树逼近嵌入

to the final prediction[31–33]. By predicting survival directly from image pixels, these end-to-end approaches generally improved prediction accuracy[34–39]. However, this accuracy can come at the cost of explainability. Techniques have been developed to address this shortcoming, including saliency heat maps, Local Interpretable Model-Agnostic Explanations, and Decision Tree Approximation of Learned Embeddings[40–42]. However, these approaches offer post hoc explanations and are vulnerable to confirmation bias[43,44]. Moreover, heat-map-based explanations offer little insight into the directionality or proportionality of influence on the model's prediction.

1. muti-resolution全景分割模型——MuTILs来描绘和分类WSI中的组织区域和细胞核

2. 26个生物主题的109特征，描述大小、形状、纹理、上下文关系——组织学特征

3. 正则化Cox比例风险模型中使用HiPS获得连续风险评分

To address these limitations, we used a 'concept bottleneck modeling' approach to discover visual prognostic biomarkers of the TME (Fig. 1b)[45]. First, we used a multi-resolution panoptic segmentation model called 'MuTILs' to delineate and classify tissue regions and cell nuclei in WSIs (Supplementary Figs. 2–5)[46,47]. Second, we designed a comprehensive set of 109 features from 26 biological themes that describe the size, shape, texture, and contextual relationship of regions and nuclei; these are referred to as 'histomic features' (Fig. 2 and Supplementary Table 5). Using univariable Cox regression of breast-cancer-specific survival (BCSS), we identified the most prognostic feature from each theme (Supplementary Table 23). The 26 most prognostic histomic features, along with the ER, PR, and HER2 expression status, form the HiPS. Finally, we use HiPS features in a regularized Cox proportional hazards model to obtain a continuous risk score, the 'HiPS score', in the range (0, 10). Using Gaussian mixture modeling of the score distribution, we identified thresholds to define HiPS groups H1–H3. Cancers scoring <3.6 are H1, while those scoring ≥6 are H3. For comparison, we fit a control model that combines the Nottingham grade with the ER, PR, and HER2 expression status. Of note, the granularity of staging data we had precluded using the AJCC prognostic stage as our control. We refer to continuous and discrete predictions from the control model as the 'control score' and 'control groups C1–C3' (Supplementary Fig. 7).

- CPS-II是发现队列
- TCGA是半独立验证队列
- PLCO、CPS-3是独立验证队列

Panoptic segmentation was trained using annotations from the Breast Cancer Semantic Segmentation and NuCLS datasets derived from 125 slides from the TCGA cohort, along with 85 annotated slides from the CPS-II cohort[24,25,48–50]. Slide visualization and management utilized the Digital Slide Archive platform[51]. For survival modeling, the CPS-II dataset was used for feature selection and to learn feature weights and score thresholds for HiPS and control scoring and grading. Hence, CPS-II is our discovery cohort, TCGA is a semi-independent validation cohort (used for fitting panoptic segmentation but not the survival model), and PLCO and CPS-3 are independent validation cohorts. Because TCGA and PLCO were selected for particular tumor characteristics and sourced exclusively from tertiary-care centers, they differ from our discovery cohort. Specifically, there are significant differences in the clinical characteristics (Supplementary Tables 1–4) as well as the distributions of HiPS features in these datasets (Supplementary Figs. 25–29). Yet, HiPS performance was robust and consistent across cohorts.

Novel epithelial morphologic features are highly prognostic

Many of the epithelial histomic features correspond to the Nottingham criteria, including epithelial architectural features corresponding to the tubule formation grade and nuclear morphology features corresponding to the nuclear grade (Supplementary Figs. 23 and 24). Although mitotic figure counting is part of the Nottingham criteria, we did not measure this owing to the technical challenge of highly specific mitosis detection. However, we show that the HiPS score is strongly associated with mitotic grade, probably because of the strong correlation between the mitotic grade and the tubule formation and nuclear Nottingham grades (Supplementary Figs. 43 and 44).

We studied this relationship by examining the strength of the features' association with grade versus their association with survival. Histomic features with the strongest Nottingham grade association were not necessarily the most prognostic (Fig. 2b and Supplementary Fig. 19). In the CPS-II and CPS-3 cohorts, features capturing glandular architecture had the highest association with grading, including the mean and variance of the epithelial nest area and perimeter, and the number of holes within epithelial nests and its variance (that is, formation of a glandular lumen). In CPS-II, we found that these features were less prognostic than other architectural features such as the circularity of epithelial nests and epithelial cell clustering within a 64 µm radius. In fact, the most prognostic epithelial features capture nuclear morphology, namely, chromatin clumping of epithelial nuclei (BCSS hazard ratio (HR) = 1.35, $P = 0.001$) and its variance (BCSS HR = 1.34, $P = 0.001$).

Nuclear features had the highest association with Nottingham grading in PLCO and TCGA (Supplementary Fig. 19). These include the size of epithelial nuclei, complexity of the epithelial nuclear boundary, and epithelial nuclear chromatin clumping. In PLCO, these features were also the most prognostic (BCSS HR range = 1.72–3.61), and there was a direct relationship between how prognostic a histomic feature was and how well it was associated with the Nottingham grade. TCGA, on the other hand, had many top prognostic features that were not as strongly associated with grade (Supplementary Figs. 23 and 24), including the number of 'benign-appearing' nuclei per epithelial nest, a feature capturing morphologic similarity to noncancerous tissue (progression-free interval (PFI) HR = 0.48, $P = 0.001$).

The variation between cohorts probably reflects differences in patient populations, variability between pathologists, and limitations of the Nottingham criteria.

Stromal and immune features of the TME are highly prognostic

Our transparent prognostic model reveals the influence of individual features, including (Fig. 3 and Supplementary Table 22):

Global CAF and cancer cell densities. Global density of CAFs was the most important histomic feature and was slightly more influential than cancer cell density. In contrast, global TIL density had little influence.

Nuclear chromatin clumping. The magnitude and cell-to-cell variation in chromatin clumping in cancer cells were highly prognostic, reflecting the nuclear pleomorphism component of Nottingham grading. Surprisingly, chromatin clumping of TILs was also informative.

Stromal matrix and collagen fibrils. The variation in staining of the stromal matrix is the third most influential histomic feature. This reflects interface changes such as stromal desmoplasia. We also find the waviness of collagen fibrils to be weakly influential, reflecting general architectural disruption in advanced cancers. In contrast, the entropy of collagen orientation was not influential, probably because its prognostic value is not independent of other features in HiPS.

Epithelial–stromal cell interactions. These include the clustering of TILs and clustering of CAFs within 64 µm of cancer cells, as well as the density of TILs within 32 µm of cancer cells.

Epithelial nest morphology and architecture. These include cell clustering (that is, formation of nests), nest circularity, and slide-level variation in nest circularity. The amount of cytoplasm per epithelial nest, a robust estimate of the nuclear-to-cytoplasmic ratio, is also influential.

The linearity of our prognostic model also allowed us to calculate subscores that summarize the influence of features belonging to a single theme. Hence, the HiPS score is the sum of six thematic subscores (Fig. 4 and Supplementary Figs. 30–35). In cases in which HiPS altered patients' risk categorization, we found that the combined contribution of epithelial and ER, PR, and HER2 subscores was often less than that of the other subscores. In other words, a patient who harbors high-risk epithelial features may be considered at a lower overall risk because

https://doi.org/10.1038/s41591-023-02643-7

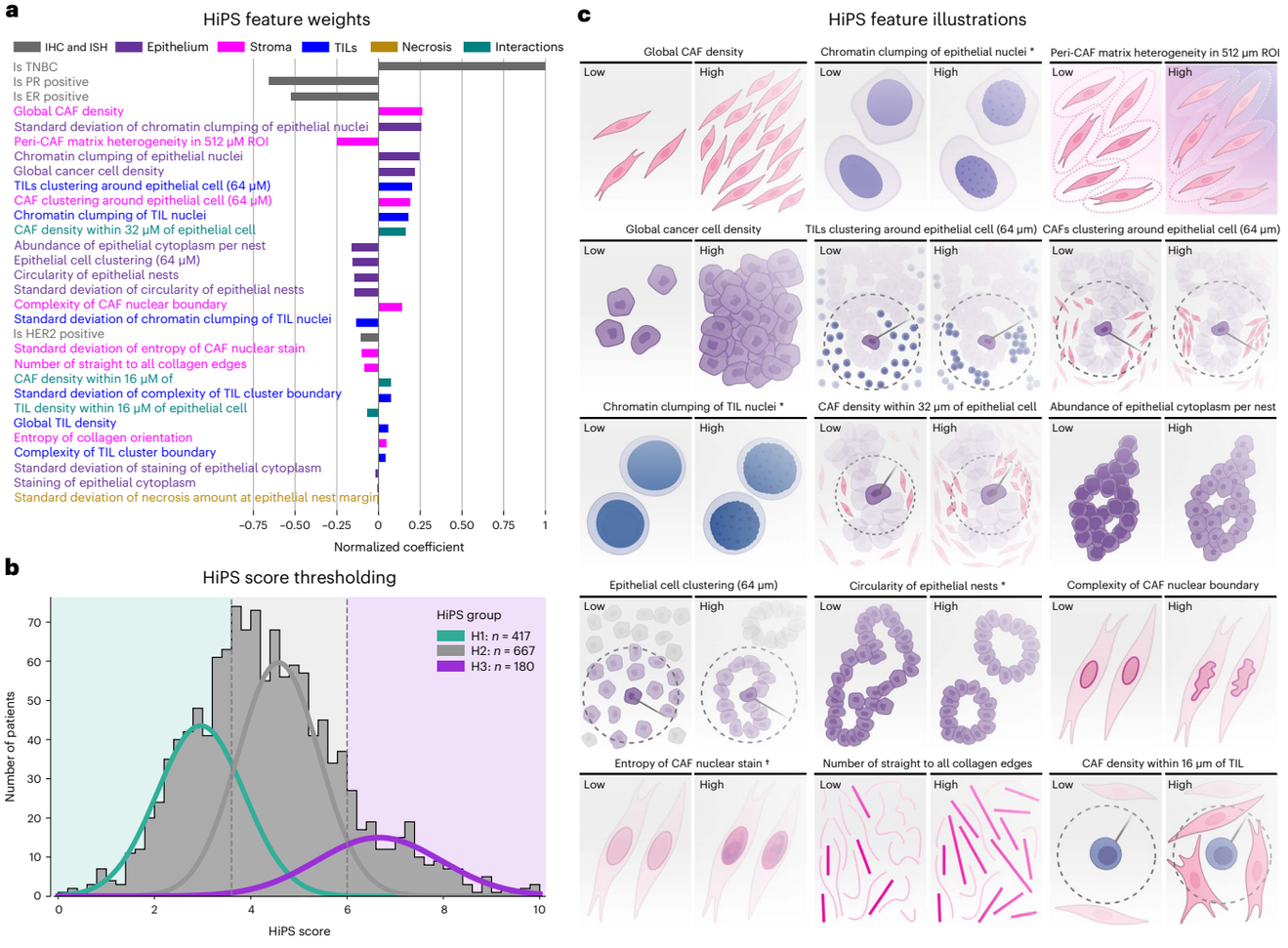


Fig. 3 | The HiPS. a, Relative contribution of histomic features to the HiPS score. HiPS combines 26 computationally derived morphologic descriptors of the TME from H&E WSI scans with the breast IHC and ISH panel. While epithelial features were influential, we found that stromal, immune, and cell–cell interaction features had an equally important prognostic role. **b**, Distribution of the HiPS scores among patients from the CPS-II cohort. The distribution was modeled as a mixture of three Gaussians defining low-risk (H1), intermediate-risk (H2), and high-risk (H3) prognostic groups. **c**, Illustrations of the most influential features on the HiPS score, ordered by importance. The star symbol indicates features whose mean and variance values were both influential, while the cross symbol indicates features whose variance alone was influential. CAF density and acellular stromal matrix heterogeneity were among the top-five features. The morphology and local interactions of CAFs and TILs also played an important role.

of their prognostically favorable stromal features, and vice versa. Note that this principle underlies efforts to use stromal and immune biomarkers to guide immunotherapy de-escalation[15,52].

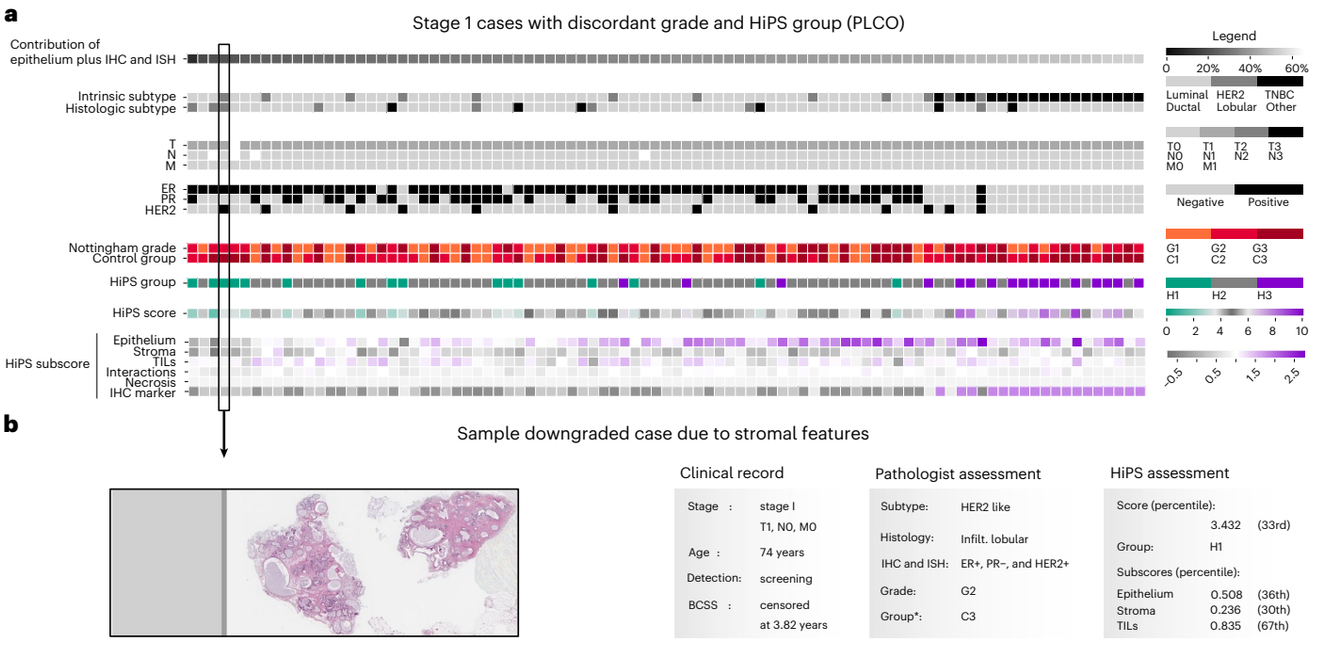
That many stromal and spatial features outweigh epithelial features is consistent with increased recognition of the role of the TME. We hypothesize that the prognostic value of HiPS is partly driven by the diversity of biological phenomena it captures. To test this hypothesis, we performed a sensitivity analysis that fit an alternative score, $HiPS^{epithelial}$, based entirely on epithelial features and the standard IHC and ISH panel (Supplementary Fig. 8). $HiPS^{epithelial}$ was inferior to the full HiPS score, highlighting the value of fully incorporating TME features. Compared with $HiPS^{epithelial}$, HiPS better stratified BCSS in CPS-II (group 3, HR = 6.59 versus HR = 3.97) and PLCO (group 3, HR = 8.3×10^7 versus HR = 7.27; Fig. 5). HiPS and $HiPS^{epithelial}$ had similar prognostic value in TCGA, which comprises more advanced cancers.

Note that computational assessment of epithelial elements alone is also superior to manual grading; $HiPS^{epithelial}$ improves outcome stratification in all three datasets compared with control grouping. This is not entirely explained by increased objectivity from computational versus manual grading, as HiPS includes highly prognostic epithelial features not associated with the Nottingham grade (Fig. 1b).

HiPS is more prognostic than grade in independent datasets

HiPS stratifies patients into three risk groups in CPS-II, PLCO, and TCGA (Fig. 5 and Supplementary Figs. 11–18). Within CPS-II, Nottingham grade identifies patients with distinct BCSS ($P < 0.001$). Control groups, incorporating ER, PR, and HER2, improve this stratification (C3 HR = 4.52 versus G3 HR = 3.99, both $P < 0.001$). HiPS groups are yet more predictive of BCSS (H3 HR = 6.59, log-rank $P < 0.001$). HiPS is also a better predictor of overall survival (OS) than control groups ($P < 0.001$ versus $P = 0.065$). A similar result was found in the PLCO and TCGA validation cohorts. In PLCO, HiPS improved BCSS stratification (log-rank $P < 0.001$ versus $P = 0.037$) and OS stratification ($P = 0.015$ versus $P = 0.308$). We also show that HiPS successfully stratifies OS outcomes in the TCGA dataset, although the result was not statistically significant ($P = 0.055$). We also explored PFI and found that HiPS significantly stratified outcomes ($P = 0.025$)[53]. Control groups could not stratify TCGA patients' OS or PFI outcomes. We hypothesize that

https://doi.org/10.1038/s41591-023-02643-7



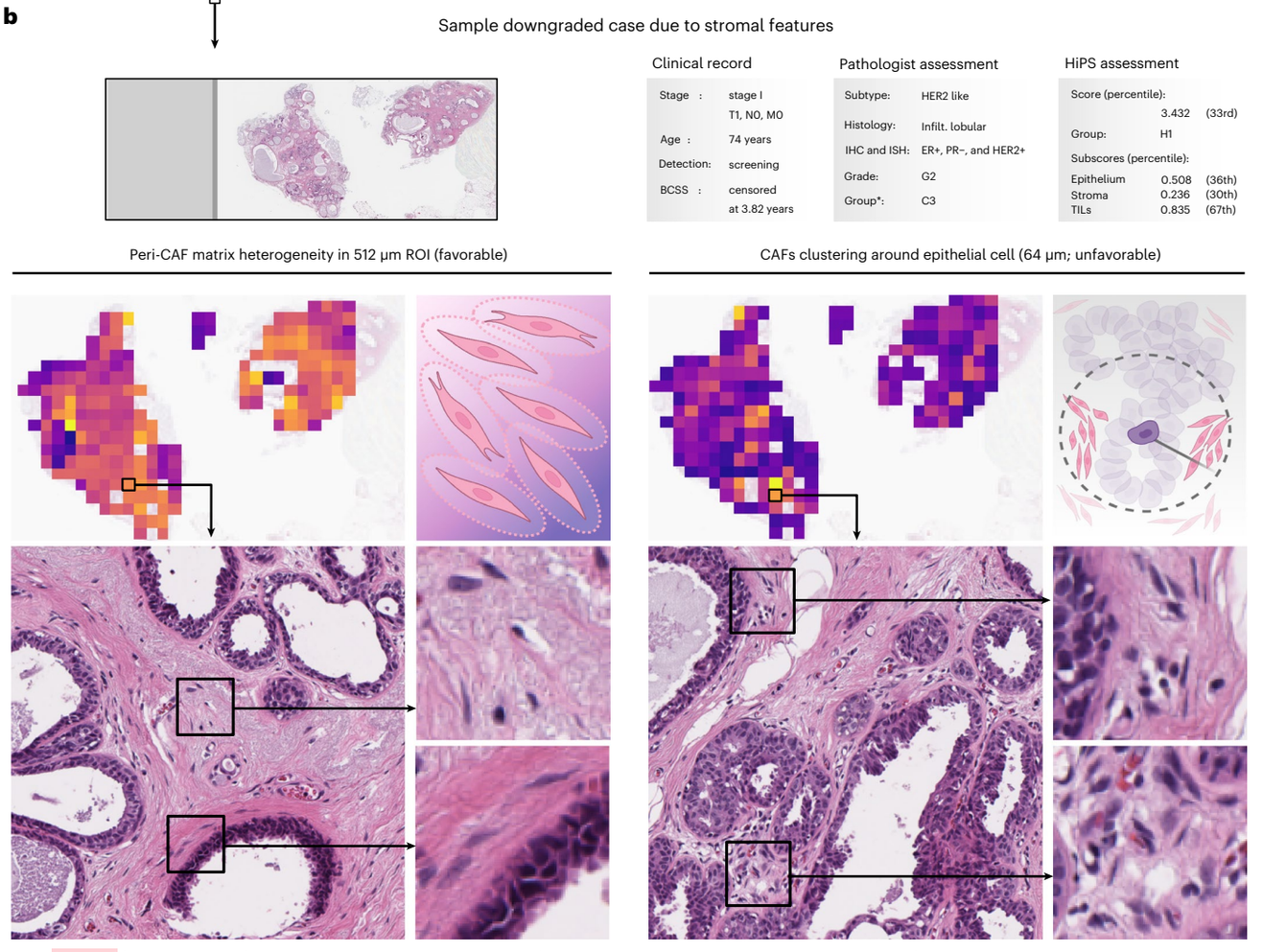


Fig. 4 | Stromal features critically impact the HiPS score and alter risk categorization in stage I cancers. a, Clinicopathologic characteristics of stage I cancers in which HiPS altered the Nottingham risk categorization (91 of 231 patients with stage I cancer). By definition, stage I cancers do not have nodal involvement, so there is a higher importance of histology in guiding clinical decision-making. In the supplement, we show that HiPS improves outcome stratification in this cohort. Patients are sorted by the percent contribution of epithelial and ER, PR, and HER2 features to the HiPS score. We also show the HiPS subscores, each of which summarizes the influence of features within each biological theme. The summation of the six HiPS subscores equals the total HiPS score. Note that non-epithelial features contribute heavily to the total HiPS score in this cohort and were heavily influential in altering the patients' risk categories. **b**, Sample case in which stromal features were heavily influential. Two features are illustrated: (1) The variation in peri-CAF stromal matrix intensity, which reflects stromal interface changes such as desmoplasia and is favorably prognostic, and (2) clustering of CAFs within a 64 µm radius of epithelial cells, which is adversely prognostic.

https://doi.org/10.1038/s41591-023-02643-7

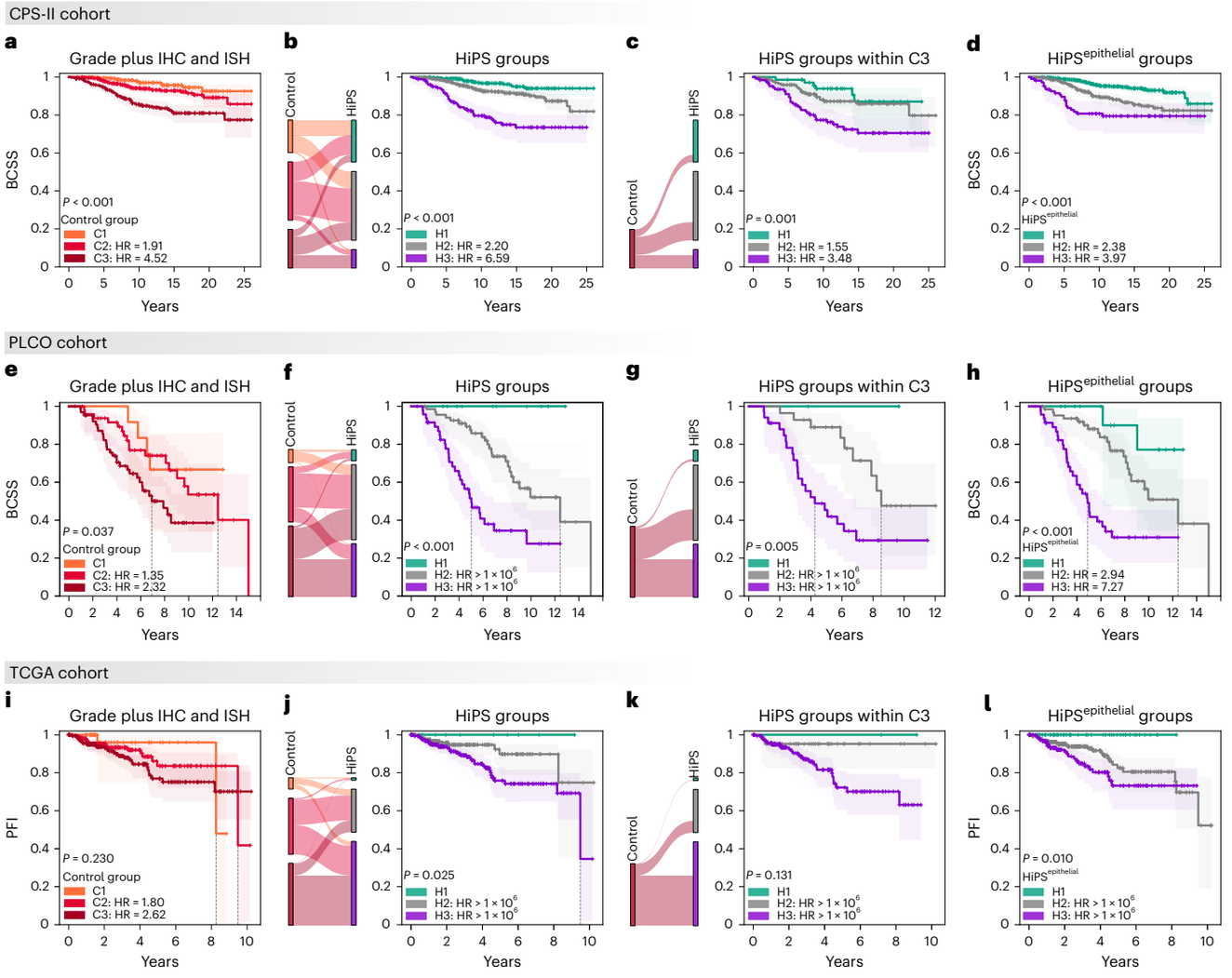


Fig. 5 | Kaplan–Meier analysis of HiPS groups compared with the control groups. Detailed results and at-risk tables are provided in Supplementary Tables 11–16. Error bands represent the exponential Greenwood 95% confidence intervals. CPS-II was our discovery cohort, and PLCO and TCGA were independent validation cohorts. Control prognostic groups were obtained by combining Nottingham grades from pathology reports with the ER, PR, and HER2 panel using the same methodology as the HiPS score. All *P* values represent the log-rank statistic, and all HRs are relative to the lowest-risk category. **a**, BCSS outcomes for patients within the CPS-II cohort using the control groups (Nottingham grade plus ER, PR, and HER2). **b**, Left: Sankey diagram of reclassification from the control group to HiPS in CPS-II. Right: BCSS outcomes for patients in the CPS-II cohort using the HiPS groups. Note the higher HR than that in the control groups in **a**. **c**, Left: Sankey diagram of reclassification from control group C3 to HiPS groups within the CPS-II cohort. Right: BCSS outcomes for control group C3 patients when stratified into HiPS groups. HiPS enables BCSS outcome stratification within control group C3. **d**, BCSS outcome stratification for the CPS-II cohort using the alternative model, HiPS^{epithelial}, which only relies on epithelial histologic features and the ER, PR, and HER2 panel. Group 3 HR values are smaller than those observed with the full HiPS model in **b**. **e**–**h**, BBCS outcomes for the PLCO cohort when using control group assignments (**e**), HiPS group assignments (**f**), HiPS group assignments within control group C3 (**g**), and the HiPS^{epithelial} model (**h**). **i**–**l**, PFI outcomes for the TCGA cohort when using control group assignments (**i**), HiPS group assignments (**j**), HiPS group assignments within control group C3 (**k**), and the HiPS^{epithelial} model (**l**).In TCGA, high censorship rates and missing cause-of-death information make PFI the most suitable outcome measure.

overrepresentation of advanced cases from tertiary-care centers may drive this failure (TNM stages 2–3: 82.8% in TCGA versus 19.6% in CPS-II). TCGA also has fewer observed deaths than CPS-II (9.1% versus 36.6%, $P < 0.001$) owing to the shorter follow-up of TCGA subjects and their younger age relative to the population[54].

Reassigned risks are more consistent with observed outcomes

Novel diagnostics are incrementally useful insofar as they reclassify patients into higher or lower risk (Supplementary Figs. 20–22). To identify clinically relevant risk category reassignments (that is, those impacting survival outcomes), we evaluated HiPS within each control group. In CPS-II, we could identify three prognostically distinct subsets within C2 ($P < 0.001$) and C3 ($P = 0.001$). Likewise, the clinically relevant reassignments in PLCO were those from C3 to H2, identifying a distinct subset of patients with better survival (median BCSS = 8.5 years versus 4.3 years, $P = 0.005$). HiPS also allows some (insignificant) stratification within C3 in TCGA by downgrading a subset of these cases.

HiPS is prognostic independent of stage and other variables

HiPS score and groups provide significant prognostic value for BCSS in the CPS-II cohort, independent of cancer stage and tumor size (Supplementary Tables 6 and 7, and Supplementary Fig. 17; HiPS score

https://doi.org/10.1038/s41591-023-02643-7

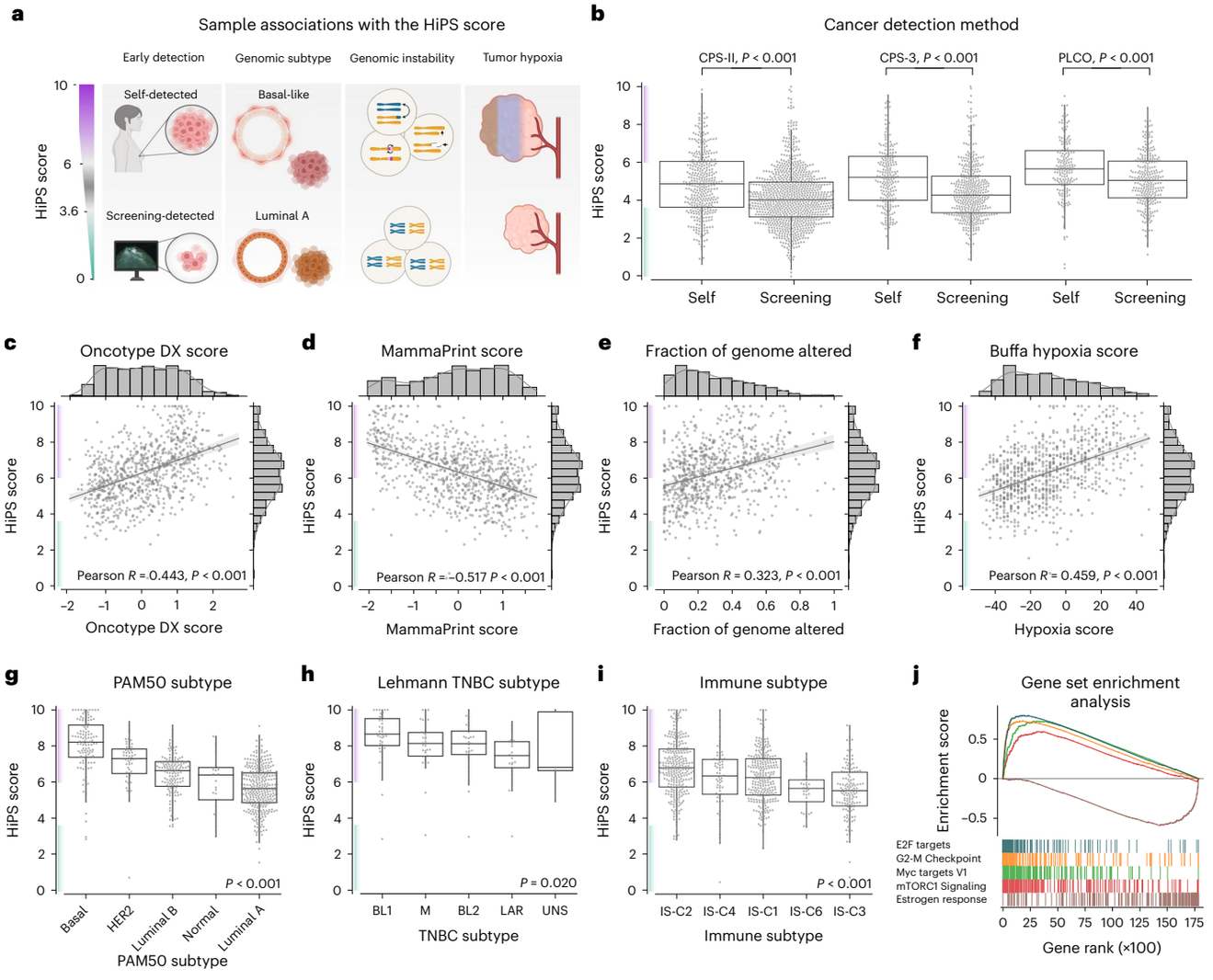


Fig. 6 | The HiPS score is consistent with established risk profiles. In each of the plots shown, each dot is a single patient, and the green and purple *y*-axis ranges indicate H1 (scores <3.6) and H3 (scores ≥6), respectively. All box plot center lines = median, box limits = upper and lower quartiles, and whiskers = 1.5× interquartile range. Supporting results and sample sizes are provided in Supplementary Tables 1, 2, and 27. All *P* values are two sided. *P* values in **b** represent the independent two-sample *t*-test, and those in **g**–**i** represent one-way ANOVA. **a**, Some of the epidemiological and genomic associations discussed. **b**, The distribution of HiPS scores by cancer detection method. Cancers detected using screening programs had lower HiPS scores than self-detected ones, probably reflecting early detection before developing high-risk features. From left to right, $n = 370, 860, 225, 402, 179$, and 279 patients. **c**,**d**, Scatterplots of HiPS scores versus the Oncotype DX (**c**) and MammaPrint RNA (**d**) assays of breast cancer recurrence risk and (inverse) metastasis risk, respectively. Exact *P* values are 3.341×10^{-39} (**c**) and 5.11×10^{-55} (**d**). **e**,**f**, Scatterplots of HiPS scores versus the fraction of genome altered (**e**) and the composite Buffa hypoxia score (**f**). Higher genome alteration and tumor hypoxia correlate directly with the HiPS score. **g**, HiPS score distributions within the PAM50 genomic subtypes. Basal and Luminal A cancers have the highest and lowest HiPS scores, consistent with subtype risk profiles. From left to right, $n = 122, 60, 151, 23$, and 354 patients. **h**, HiPS score distributions within genomic TNBC subtypes. BL1 cancers have the highest scores, while LAR cancers have the lowest scores. UNS, unspecified TNBC subtype. From left to right, $n = 44, 35, 32, 23$, and 5 patients. **i**, Distribution of HiPS scores within cancer IS. IS-C3 is associated with better clinical outcomes, whereas IS-C1 and IS-C2 are associated with worse outcomes. From left to right, $n = 287, 69, 259, 32$, and 143 patients. **j**, GSEA of HiPS scores using the MSigDB Hallmarks collection. High-HiPS tumors have upregulation of proliferation and cell cycle dysregulation gene sets and downregulation of early estrogen response genes.

HR = 1.28, $P < 0.001$; H3 HR = 3.62, $P < 0.001$). The control score and groups were also independently prognostic (Supplementary Tables 8 and 9; control score HR = 1.24, $P < 0.001$; C3 HR = 3.56, $P < 0.001$). When used in an expanded multivariable Cox regression model, HiPS scoring and groups retained their prognostic value (HiPS score HR = 1.41, $P < 0.001$; H3 HR = 3.45, $P = 0.006$). In addition to tumor stage and size, this model also included demographic factors (age, menopausal status, race, smoking, body mass index), expression of basal markers cytokeratin 5/6 (CK5/6) or epidermal growth factor receptor (EGFR), detection method (screening versus self-detected), and treatment (chemotherapy, radiation therapy, and targeted therapy; Supplementary Tables 10 and 11). The control models had similar prognostic values (Supplementary Tables 12 and 13; control score HR = 1.33, $P < 0.001$; C3 HR = 3.56, $P = 0.003$).

We also found that HiPS scoring, but not HiPS grouping, was independently predictive of OS within PLCO after adjusting for cancer stage, tumor size, and patient age (Supplementary Tables 14 and 15, and Supplementary Fig. 18; HiPS score HR = 1.15, $P = 0.029$). Here we

assessed OS rather than BCSS owing to missing data and sample size constraints. In contrast, the control models were not prognostic when adjusted for the same covariates (Supplementary Tables 16 and 17).

Likewise, HiPS scoring, but not HiPS grouping, is predictive of OS in the TCGA cohort independent of cancer stage and tumor size (Supplementary Tables 18 and 19; HiPS score HR = 1.22, $P = 0.009$). The control models had no independent prognostic value (Supplementary Tables 20 and 21). Neither HiPS nor control scoring was predictive of PFI.

HiPS is prognostic in ER+, Luminal-like, and HER2+ cancers

HiPS improves outcome stratification compared with grading in ER+, Luminal-like, and HER2+ cancers. As expected, the improvement was more dramatic in the CPS-II discovery cohort than in the independent cohorts (Supplementary Figs. 36–41). Although HiPS improves BCSS stratification of ER− cancers in CPS-II, the effect did not generalize to other cohorts. Neither grading nor HiPS could stratify patients with triple-negative breast cancer (TNBC) in any cohort.

Within the CPS-II ER+ subcohort, HiPS grouping performs better than Nottingham grading, both in terms of OS ($P < 0.001$ versus $P = 0.411$) and BCSS (H3 HR = 6.76 versus G3 HR = 3.83, all $P < 0.001$). We note that Nottingham grades were almost indistinguishable in terms of median OS (20.2–20.5 years), whereas H3 patients had a shorter median OS of 14.9 years compared with 20.6 years and 20.2 years of H1 and H2, respectively. Similarly, HiPS stratifies BCSS outcomes in the PLCO ER+ subcohort ($P = 0.004$), unlike Nottingham grading ($P = 0.163$). Finally, HiPS results in improved, albeit not statistically significant, stratification of OS and PFI outcomes of the TCGA ER+ subcohort.

Because most Luminal-like cancers are ER+, the prognostic effect of HiPS is similar in these subgroups. In CPS-II Luminals, HiPS results in insignificant stratification of OS and BCSS outcomes compared with Nottingham grading (OS: HiPS $P = 0.001$ versus Nottingham $P = 0.565$; BCSS: HiPS $P < 0.001$, G3 HR = 6.73, versus Nottingham $P < 0.001$, G3 HR = 4.33). In TCGA Luminals, HiPS grouping results in a visible improvement in OS and PFI stratification but is not statistically significant. In contrast, neither Nottingham nor HiPS could significantly stratify outcomes in the PLCO Luminal-like cohort.

Finally, HiPS grouping of CPS-II HER2+ cancers results in a dramatic improvement in the stratification of outcomes compared with grading (OS: $P = 0.035$ versus $P = 0.561$; BCSS: $P = 0.005$ versus $P = 0.541$). Likewise, HiPS improves OS stratification in HER2+ cancer in PLCO ($P = 0.051$ versus $P = 0.690$). Missing cause-of-death data precluded BCSS assessment in this subcohort. Neither Nottingham nor HiPS grouping could significantly stratify the HER2+ TCGA subcohort.

HiPS is consistent with established prognostic surrogates

The HiPS score is higher in cancers with high-risk clinical and genomic features (Fig. 6). Cancers that were detected using mammographic screening had lower HiPS scores than those self-detected in the CPS-II, CPS-3, and PLCO cohorts (all $P < 0.001$). This finding is consistent with recommendations for cancer screening by the ACS. This association persisted within TNM stage I (all $P \leq 0.001$) and stages II–III ($P = 0.002$, $P = 0.014$, and $P < 0.001$ in CPS-II, CPS-3, and PLCO, respectively). Screening-detected cancers derived their low-risk scores not only from their ER, PR, and HER2 status but because they also had favorable histology. This was evidenced by the lower HiPS scores in screening-detected cancers within Luminal-like cancers (all $P \leq 0.001$), HER2-like cancers ($P < 0.001$ for CPS-II and CPS-3; PLCO not significant), and TNBC ($P = 0.037$ and $P = 0.006$ for CPS-II and PLCO, respectively; CPS-3 not significant).

The 2018 AJCC update included the addition of Oncotype DX, a 21-gene RNA expression assay predicting recurrence risk, as a stage modifier[55,56]. MammaPrint is another commonly used RNA expression assay that uses a 70-gene panel to predict risk of metastasis (Supplementary Fig. 45)[57,58]. Using a research-based estimate of these assays, we show that HiPS is significantly correlated with both scores in TCGA (Oncotype DX: $r = 0.443$, $P < 0.001$; MammaPrint: $r = -0.517$, $P < 0.001$)[59].

PAM50 gene expression subtypes had significantly different HiPS scores in TCGA ($P < 0.001$), with Basal and Luminal A cancers having the highest and lowest scores, respectively (8.05 ± 1.46 versus 5.64 ± 1.23). Higher HiPS scores were also observed in cancers expressing the Basal IHC and ISH markers EGFR (in CPS-II and CPS-3, $P < 0.001$) and CK5/6 (in CPS-II, $P < 0.001$)[60–65].

Within TNBCs, HiPS score distributions differed by genomic subtype ($P = 0.02$), with Basal-Like 1 (BL1) and Luminal Androgen Receptor (LAR) having the highest and lowest scores, respectively (8.53 ± 1.36 versus 7.30 ± 1.37). This is consistent with the mutational burden in BL1 cancers compared with LAR (2.1 versus 1.8 mutations Mb^{-1})[60]. HiPS scores were inversely correlated with immune cell infiltration within the TNBC microenvironment, determined using the xCell genomic deconvolution assay ($r = -0.179$, $P = 0.039$) or the immune score from estimation of stromal and immune cells in malignant tumor tissues using expression data ($r = -0.168$, $P = 0.05$)[61,62].

The HiPS score was significantly correlated with composite scores measuring the accumulation of genetic alterations, including the aneuploidy score ($r = 0.228$, $P < 0.001$) and the fraction of genome altered ($r = 0.323$, $P < 0.001$)[63,64]. It was also correlated with various composite measurements of tumor hypoxia, including the Buffa, Ragnum, and Winter hypoxia scores ($r = 0.355–0.459$, all $P < 0.001$)[65–68].

Finally, using gene set enrichment analysis (GSEA), we show that tumors with high HiPS scores had upregulation of gene sets related to cell proliferation, such as *E2F Targets*, *G2-M Checkpoint*, *Myc Targets V1*, *mTORC1 Signaling*, and *Mitotic Spindle*, and downregulation of estrogen response genes[69,70].

HiPS reflects effectiveness of the cancer immune response

Using gene expression data for TCGA enabled us to obtain mechanistic insights into the biological phenomena that underlie HiPS. In GSEA with the Azimuth single-cell gene set collection, we found that chromatin clumping of TIL nuclei was associated with an immune response predominantly mediated by $CD4^+$ T-helper (Th) cells and inversely associated with $CD8^+$ T cell infiltration (Supplementary Fig. 50)[71–73]. In addition, GSEA using the Hallmarks set showed that chromatin clumping of TILs was positively associated with protein secretion, probably because of B cells and plasma cells, and inversely associated with TNF-α signaling via NF-kB, a critical pathway in T cell activation[69,70,74,75]. We did not find an association with regulatory T cell expression. Together, these findings indicate that tumors with a high HiPS score lack effective $CD8^+$ response and instead have an immune reaction predominated by ineffectual B cells, plasma cells, and $CD4^+$ T cells[72,73,76]. This may explain why average chromatin clumping of TILs was an adverse prognostic feature, while its variance was protective, signifying some $CD8^+$ involvement in the anticancer immune response.

HiPS scores differed significantly by cancer immune subtypes (IS; $P < 0.001$), being higher in IS-C1 and IS-C2, known to have poor survival, than in the favorable subtype IS-C3 (ref. 77). In addition, there was a moderate, significant correlation between the HiPS score and the *Wound Healing* gene set that is definitional of immune subtypes IS-C1 and IS-C2 (Fig. 6 and Supplementary Fig. 48; $r = 0.454$, $P < 0.001$)[77,78]. Both subtypes are known to have a low Th1:Th2 ratio which indicates, albeit indirectly, that Th2 polarization is another high-risk immune feature that increases the HiPS score[76,79].

HiPS captures CAF subtypes and mesenchymal transition

CAF subtypes associated with different survival risks have been reported[80,81]. We found that high complexity of CAF nuclear boundaries was significantly associated with the abundance of myCAFs, as determined by gene expression profiling (Supplementary Fig. 47)[81]. This is consistent with the known higher risk associated with the myCAF phenotype, which is in turn a subset of the CAF-S1 subtype known to have an immunosuppressive function[81,82]. This finding was also supported by the enrichment of myogenesis and smooth-muscle-defining gene

sets in slides with complex CAF nuclear boundaries (Supplementary Figs. 49 and 50)[69–71].

Many of the top features in HiPS are probably surrogates for CAF hypoxia, activation, regulation of extracellular matrix stiffness, and their crosstalk with cancer cells with consequent epithelial-to-mesenchymal transition (EMT)[83,84]. The *EMT* gene set was significantly enriched in cancer with a high global CAF density, complex CAF boundaries, and homogeneous stromal matrix surrounding CAFs (Supplementary Fig. 49). In addition, the interleukin-6 (IL-6) pathway was enriched in tumors with high CAF density and complex CAF nuclear boundaries, consistent with previous reports on the role of IL-6 in CAF-mediated induction of *EMT*[83,85]. Finally, the inverse association between oxidative phosphorylation gene sets and top HiPS CAF features, along with the inverse association between transforming growth factor-β (TGF-β) signaling and peri-CAF stromal matrix homogeneity, points to a TGF-β-mediated hypoxic response that is associated with CAF activation, consistent with other studies[82,84,86].

Discussion

We described the development, validation, and utilization of a comprehensive histologic signature for prognosticating invasive nonmetastatic breast cancer using population-based datasets. HiPS is a multimodal score that combines scanned H&E slides with the ER, PR, and HER2 panel to place patients along a risk spectrum based on the prognostic favorability of their TME. Our method combines deep learning for robust panoptic segmentation of tissue regions and cell nuclei, morphological processing to extract hypothesis-driven features, and Cox regression modeling to maximize interpretability. This approach enabled us to leverage recent advances in deep learning for pattern recognition while addressing interpretability issues that limit widespread adoption of deep learning in clinical settings. We developed and validated our score using two population-level cohorts from the CPS-II and CPS-3 studies and two diverse datasets from the TCGA and PLCO trial. In total, we used data from 3,177 patients, with tissue samples from 614 counties in 48 states. These data have a high variability in patient demographics, slide preparation and staining protocols, WSI scanners, and other preanalytical factors. This diversity supports the generalizability of our findings.

- 用Cox回归建模来最大化可解释性

We show that the HiPS score is a strong, independent predictor of survival outcomes in nonmetastatic ER+ and HER2+ cancers, and that it is concordant with known epidemiologic and genomic risk profiles. In addition, HiPS has a high correlation with the Oncotype DX and MammaPrint gene expression assays, which are widely used to predict recurrence and metastasis risk[55–58]. While these assays have been a tremendous leap forward in personalized care, they present limitations in terms of cost, access, and processing time. With a cost of US$3,000, Oncotype DX costs Medicare hundreds of millions of dollars per year[87]. Furthermore, not all facilities provide molecular testing, limiting accessibility and further exacerbating cancer disparities[88]. Finally, Oncotype DX testing has been associated with treatment delays owing to the lengthy processing time needed[89].

- 与现有的Oncotype DX和MammaPrint有高度相关性

- 原有办法，成本、访问性、处理时间有局限性

The body of work on computational analysis of breast pathology is large and includes Ki67 and mitosis quantification, detection of lymph node metastases, classification of histologic subtypes, characterization of immune infiltrates, prediction of hormone receptor status, and prognostication[90]. Approaches that design features rationally have typically focused on a theme, such as immune infiltrates or epithelium, or have investigated specific features such as collagen fiber organization or stromal cellular density[91,92]. A recent study describes an epithelial signature that improves prognostication when combined with Oncotype DX in ER+ node-negative invasive breast cancer, but this signature focused on a limited feature set that excluded stromal and TIL morphology[93]. Comprehensive feature sets have been evaluated in predicting pathway activation and clinically relevant phenotypes such as PD-L1 expression in multiple cancers[94].

End-to-end learning has yielded impressive results for diagnostic applications including a study of breast cancer diagnosis from adjacent stroma, detection of cancer in prostate biopsies and metastases in lymph nodes, and prediction of primary origin for metastases[34,95–97]. Methods that learn from patient-level labels typically offer explanations in the form of saliency heat maps that localize relevant regions. However, localization alone does not entirely explain the features being interpreted by a model, leaving human intuition to fill the gap[98]. For a problem such as cancer detection, localization may provide an adequate explanation. Prognosis, however, typically requires holistic evaluation of features at multiple scales, with some features being protective and some being associated with adverse outcomes. Relevant features may be colocalized, which further challenges explanation by saliency heat maps[95].

Quantitative HiPS features were developed in a hypothesis-driven manner to quantify distinct biological phenomena[94,99]. As a result, we exceeded expert performance using established grading criteria. This success is partly driven by capturing stromal, immune, and spatial clustering features not typically assessed. However, we exceeded human performance even when we limited our analysis to epithelial morphology. These gains may be due to the quantitative nature of HiPS compared with visual estimates[100]. Nonetheless, we also showed that epithelial features modestly correlated with the Nottingham criteria are highly prognostic, consistent with previous indirect evidence[23].

Compared with previous works, our segmentation models enable us to untangle the influence of CAFs, acellular stroma, and TILs[91,101]. In particular, we found that stromal interface changes reflecting desmoplasia or collagen disorder were favorably prognostic, consistent with previous works[92,101]. We also found an adverse prognostic value of CAF density, both globally and within 64 µm of epithelial cells, which may be proxies for wound healing and epithelial–CAF interaction[13,102]. This was supported by the increased activation of the Wound Healing pathway in cancers with a high HiPS score. Also, consistent with a previous study, we found CAF clustering around epithelial cells to be adversely prognostic[103]. Some of our measurements also focused on phenotypic differences in the appearance of CAFs. We found that increased average complexity of the CAF nuclear boundary is adversely prognostic, possibly reflecting myofibroblastic differentiation and EMT of leading cancer cells as they acquire a CAF-like morphology[13,86]. Supporting this interpretation, we observed higher HiPS scores in cancers with a high proportion of myCAFs, as well as enrichment of smooth muscle gene sets in high-HiPS cancer. EMT pathways were also enriched in high-HiPS tumors.

We found TIL clustering and morphology to be more relevant than their abundance. This may reflect the inconsistent prognostic nature of TIL abundance in ER+ (unfavorable) versus TNBC and HER2+ cancers (favorable)[104,105]. In particular, the spatial clustering of TILs within 64 µm of epithelial cells is highly prognostic, probably reflecting TIL–TIL interactions that fuel the inflammatory response sustaining or modulating cancer progression[79,106]. Moreover, we found that the morphology of TIL nuclei is prognostic. High average nuclear chromatin clumping is an adverse prognostic biomarker, while the TIL-to-TIL variation in chromatin clumping is favorable. The biological significance of these findings is unclear but could represent different lymphocyte subsets and degrees of differentiation[107]. Using gene expression data, we found the HiPS score to be directly associated with increased $CD8^+$ T cells and inversely associated with $CD4^+$ T cell abundance. In combination, our results are most consistent with the mechanistic model put forth in a previous study[80] and extended by other studies whereby myCAFs maintain an immunosuppressive TME by a TGF-β-mediated reduction of $CD8^+$ infiltration and retention of $CD4^+$ T cells[82]. Future work using IHC, ISH, and molecular studies could further characterize the biological phenomena underlying HiPS. For example, the correlation between chromatin clumping and $CD8^+$ T cells, and its anti-correlation with $CD8^+$ T cells, should be confirmed using IHC markers such as CD3, CD4, CD8,

and FoxP3 (to identify regulatory T cells). We focused on H&E to ensure applicability in routine settings[108].

We want to highlight some of the limitations of our approach. The prognostic value of histomic features is dependent on multiple factors, not just fundamental biology. For example, the robustness of algorithms in consistently capturing the same phenomena is an important consideration. We described each tissue region and cell nucleus by a set of morphological and spatial features, which were aggregated using weighted mean and variance to obtain per-patient results. This results in the loss of potentially useful information, such as the outsized influence of small foci of angioinvasion. However, there is an inevitable tradeoff between modeling complexity and interpretability[43,45]. Our results show that HiPS is not prognostic within ER− and TNBC cancers, which may reflect a differential effect of histomic features depending on intrinsic subtype. We intend to explore this in future works. It is also worth noting that genomic analyses were performed using the TCGA dataset, which does not reflect the population. We also note that the genomic analysis presented rely on post hoc correlations; while useful, these do not necessarily provide a mechanistic model. Specifically, we acknowledge that the role of the immune system in progression of luminal breast cancer remains unclear. We intend to explore these and other limitations in future works. Finally, we acknowledge the observational nature of this analysis. Most of the histomic features have correlates in tumor biology, but the causative relationship is unclear.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information, details of author contributions and competing interests, and statements of data and code availability are available at https://doi.org/10.1038/s41591-023-02643-7.

References

1. *Global Cancer Facts & Figures 4th Edition* (American Cancer Society, 2018).
2. Siegel, R. L., Miller, K. D., Wagle, N. S. & Jemal, A. Cancer statistics, 2023. *CA Cancer J. Clin.* **73**, 17–48 (2023).
3. American Joint Commission on Cancer *AJCC Cancer Staging Manual 2017* (Springer International Publishing, 2017).
4. Coughlin, S. S. Social determinants of breast cancer risk, stage, and survival. *Breast Cancer Res. Treat.* **177**, 537–548 (2019).
5. Li, X. et al. Validation of the newly proposed American Joint Committee on Cancer (AJCC) breast cancer prognostic staging group and proposing a new staging system using the National Cancer Database. *Breast Cancer Res. Treat.* **171**, 303–313 (2018).
6. Scarff, R. W. & Handley, R. S. Prognosis in carcinoma of the breast. *Lancet* **232**, 582–583 (1938).
7. BLACK, M. M., OPLER, S. R. & SPEER, F. D. Survival in breast cancer cases in relation to the structure of the primary tumor and regional lymph nodes. *Surg. Gynecol. Obstet.* **100**, 543–551 (1955).
8. Bloom, H. J. & Richardson, W. W. Histological grading and prognosis in breast cancer; a study of 1409 cases of which 359 have been followed for 15 years. *Br. J. Cancer* **11**, 359–377 (1957).
9. Elston, E. W. & Ellis, I. O. Method for grading breast cancer. *J. Clin. Pathol.* **46**, 189–190 (1993).
10. Elston, C. W. & Ellis, I. O. Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. *Histopathology* **19**, 403–410 (1991).
11. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674 (2011).
12. Cardenas, M. A., Prokhnevska, N. & Kissick, H. T. Organized immune cell interactions within tumors sustain a productive T-cell response. *Int. Immunol.* **33**, 27–37 (2021).
13. Sahai, E. et al. A framework for advancing our understanding of cancer-associated fibroblasts. *Nat. Rev. Cancer* **20**, 174–186 (2020).
14. Liu, T., Zhou, L., Li, D., Andl, T. & Zhang, Y. Cancer-associated fibroblasts build and secure the tumor microenvironment. *Front. Cell Dev. Biol.* **7**, 60 (2019).
15. Savas, P. et al. Clinical relevance of host immunity in breast cancer: from TILs to the clinic. *Nat. Rev. Clin. Oncol.* **13**, 228–241 (2016).
16. Ha, S. Y., Yeo, S.-Y., Xuan, Y. & Kim, S.-H. The prognostic significance of cancer-associated fibroblasts in esophageal squamous cell carcinoma. *PLoS ONE* **9**, e99955 (2014).
17. Conklin, M. W. et al. Aligned collagen is a prognostic signature for survival in human breast carcinoma. *Am. J. Pathol.* **178**, 1221–1232 (2011).
18. Provenzano, P. P. et al. Collagen reorganization at the tumor–stromal interface facilitates local invasion. *BMC Med.* **4**, 38 (2006).
19. Shekhar, M. P., Werdell, J., Santner, S. J., Pauley, R. J. & Tait, L. Breast stroma plays a dominant regulatory role in breast epithelial growth and differentiation: implications for tumor development and progression. *Cancer Res.* **61**, 1320–1326 (2001).
20. Couture, H. D. et al. Image analysis with deep learning to predict breast cancer grade, ER status, histologic subtype, and intrinsic subtype. *NPJ Breast Cancer* **4**, 30 (2018).
21. Rawat, R. R. et al. Deep learned tissue "fingerprints" classify breast cancers by ER/PR/Her2 status from H&E images. *Sci. Rep.* **10**, 7275 (2020).
22. Gamble, P. et al. Determining breast cancer biomarker status and associated morphological features using deep learning. *Commun. Med.* **1**, 14 (2021).
23. Bychkov, D. et al. Outcome and biomarker supervised deep learning for survival prediction in two multicenter breast cancer series. *J. Pathol. Inform.* **13**, 9 (2022).
24. Calle, E. E. et al. The American Cancer Society Cancer Prevention Study II Nutrition Cohort: rationale, study design, and baseline characteristics. *Cancer* **94**, 2490–2501 (2002).
25. Cancer Genome Atlas NetworkComprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
26. Zhu, C. S. et al. The Prostate, Lung, Colorectal and Ovarian Cancer (PLCO) screening trial pathology tissue resource. *Cancer Epidemiol. Biomark. Prev.* **25**, 1635–1642 (2016).
27. Patel, A. V. et al. The American Cancer Society's Cancer Prevention Study 3 (CPS-3): recruitment, study design, and baseline characteristics. *Cancer* **123**, 2014–2024 (2017).
28. Haralick, R. M., Shanmugam, K. & Dinstein, I. Textural features for image classification. *IEEE Trans. Syst. Man Cybern.* **SMC-3**, 610–621 (1973).
29. Doyle, S., Agner, S., Madabhushi, A., Feldman, M. & Tomaszewski, J. Automated grading of breast cancer histopathology using spectral clustering with textural and architectural image features. In *Proc. 2008 5th IEEE Int. Symposium on Biomedical Imaging: From Nano to Macro* 496–499 (IEEE, 2008).
30. Gurcan, M. N. et al. Histopathological image analysis: a review. *IEEE Rev. Biomed. Eng.* **2**, 147–171 (2009).
31. Litjens, G. et al. A survey on deep learning in medical image analysis. *Med. Image Anal.* **42**, 60–88 (2017).
32. Liu, Y., Han, D., Parwani, A. V. & Li, Z. Applications of artificial intelligence in breast pathology. *Arch. Pathol. Lab. Med.* https://doi.org/10.5858/arpa.2022-0457-RA (2023).
33. Abels, E. et al. Computational pathology definitions, best practices, and recommendations for regulatory guidance: a white paper from the Digital Pathology Association. *J. Pathol.* **249**, 286–294 (2019).

34. Campanella, G. et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat. Med.* **25**, 1301–1309 (2019).
35. Lu, M. Y. et al. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nat. Biomed. Eng.* **5**, 555–570 (2021).
36. Mobadersany, P. et al. Predicting cancer outcomes from histology and genomics using convolutional networks. *Proc. Natl Acad. Sci. USA* **115**, E2970–E2979 (2018).
37. Bychkov, D. et al. Deep learning based tissue analysis predicts outcome in colorectal cancer. *Sci. Rep.* **8**, 3395 (2018).
38. Chen, R. J. et al. Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. *IEEE Trans. Med. Imaging* **41**, 757–770 (2022).
39. Duanmu, H. et al. A spatial attention guided deep learning system for prediction of pathological complete response using breast cancer histopathology images. *Bioinformatics* **38**, 4605–4612 (2022).
40. Selvaraju, R. R. et al. Grad-CAM: visual explanations from deep networks via gradient-based localization. *Int. J. Comput. Vis.* **128**, 336–359 (2020).
41. Ribeiro, M. T. et al. "Why should i trust you?": explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1135–1144 (ACM, 2016).
42. Amgad, M. et al. Explainable nucleus classification using Decision Tree Approximation of Learned Embeddings. *Bioinformatics* **38**, 513–519 (2022).
43. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**, 206–215 (2019).
44. Leavitt, M. L. & Morcos, A. Towards falsifiable interpretability research. Preprint at arxiv.org/abs/2010.12016 (2020).
45. Koh, P. W. et al. Concept bottleneck models. in *Proc. 37th Int. Conf. on Machine Learning* (eds III, H. D. & Singh, A.) Vol. 119, 5338–5348 (PMLR, 2020).
46. Kirillov, A., He, K., Girshick, R., Rother, C. & Dollar, P. Panoptic segmentation. in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)* (2019).
47. Amgad, M., Salgado, R. & Cooper, L. L. A panoptic segmentation approach for tumor-infiltrating lymphocyte assessment: development of the MuTILs model and PanopTILs dataset. Preprint at *medRxiv* https://doi.org/10.1101/2022.01.08.22268814 (2023).
48. Amgad, M., Salgado, R. & Cooper, L. A. D. MuTILs: a multiresolution deep-learning model for interpretable scoring of tumor-infiltrating lymphocytes in breast carcinomas using clinical guidelines. Preprint at *medRxiv* https://doi.org/10.1101/2022.01.08.22268814 (2022).
49. Amgad, M. et al. Structured crowdsourcing enables convolutional segmentation of histology images. *Bioinformatics* **35**, 3461–3467 (2019).
50. Amgad, M. et al. NuCLS: a scalable crowdsourcing approach and dataset for nucleus classification and segmentation in breast cancer. *Gigascience* **11**, giac037 (2022).
51. Gutman, D. A. et al. The Digital Slide Archive: a software platform for management, integration, and analysis of histology for cancer research. *Cancer Res.* **77**, e75–e78 (2017).
52. Schmid, P. et al. Pembrolizumab plus chemotherapy as neoadjuvant treatment of high-risk, early-stage triple-negative breast cancer: results from the phase 1b open-label, multicohort KEYNOTE-173 study. *Ann. Oncol.* **31**, 569–581 (2020).
53. Liu, J. et al. An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell* **173**, 400–416.e11 (2018).
54. Wang, X. et al. Characteristics of The Cancer Genome Atlas cases relative to U.S. general population cancer cases. *Br. J. Cancer* **119**, 885–892 (2018).
55. Kalinsky, K. et al. 21-gene assay to inform chemotherapy benefit in node-positive breast cancer. *N. Engl. J. Med.* **385**, 2336–2347 (2021).
56. Paik, S. et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N. Engl. J. Med.* **351**, 2817–2826 (2004).
57. van't Veer, L. J. et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530–536 (2002).
58. van de Vijver, M. J. et al. A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.* **347**, 1999–2009 (2002).
59. Howard, F. M. et al. Integration of clinical features and deep learning on pathology for the prediction of breast cancer recurrence assays and risk of recurrence. *NPJ Breast Cancer* **9**, 25 (2023).
60. Lehmann, B. D. et al. Multi-omics analysis identifies therapeutic vulnerabilities in triple-negative breast cancer subtypes. *Nat. Commun.* **12**, 6276 (2021).
61. Aran, D., Hu, Z. & Butte, A. J. xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol.* **18**, 220 (2017).
62. Yoshihara, K. et al. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.* **4**, 2612 (2013).
63. Hoadley, K. A. et al. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell* **173**, 291–304.e6 (2018).
64. Berger, A. C. et al. A comprehensive pan-cancer molecular study of gynecologic and breast cancers. *Cancer Cell* **33**, 690–705.e9 (2018).
65. Bhandari, V. et al. Molecular landmarks of tumor hypoxia across cancer types. *Nat. Genet.* **51**, 308–318 (2019).
66. Buffa, F. M., Harris, A. L., West, C. M. & Miller, C. J. Large meta-analysis of multiple cancers reveals a common, compact and highly prognostic hypoxia metagene. *Br. J. Cancer* **102**, 428–435 (2010).
67. Winter, S. C. et al. Relation of a hypoxia metagene derived from head and neck cancer to prognosis of multiple cancers. *Cancer Res.* **67**, 3441–3449 (2007).
68. Ragnum, H. B. et al. The tumour hypoxia marker pimonidazole reflects a transcriptional programme associated with aggressive prostate cancer. *Br. J. Cancer* **112**, 382–390 (2015).
69. Liberzon, A. et al. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.* **1**, 417–425 (2015).
70. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
71. Hao, Y. et al. Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587.e29 (2021).
72. DeNardo, D. G. et al. Leukocyte complexity predicts breast cancer survival and functionally regulates response to chemotherapy. *Cancer Discov.* **1**, 54–67 (2011).
73. Mahmoud, S. M. A. et al. Tumor-infiltrating CD8+ lymphocytes predict clinical outcome in breast cancer. *J. Clin. Oncol.* **29**, 1949–1955 (2011).
74. Oh, H. & Ghosh, S. NF-κB: roles and regulation in different CD4(+) T-cell subsets. *Immunol. Rev.* **252**, 41–51 (2013).
75. Olkhanud, P. B. et al. Tumor-evoked regulatory B cells promote breast cancer metastasis by converting resting CD4+ T cells to T-regulatory cells. *Cancer Res.* **71**, 3505–3515 (2011).
76. Varn, F. S., Mullins, D. W., Arias-Pulido, H., Fiering, S. & Cheng, C. Adaptive immunity programmes in breast cancer. *Immunology* **150**, 25–34 (2017).

77. Thorsson, V. et al. The immune landscape of cancer. *Immunity* **48**, 812–830.e14 (2018).
78. Chang, H. Y. et al. Gene expression signature of fibroblast serum response predicts human cancer progression: similarities between tumors and wounds. *PLoS Biol.* **2**, e7 (2004).
79. Saltz, J. et al. Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. *Cell Rep.* **23**, 181–193.e7 (2018).
80. Costa, A. et al. Fibroblast heterogeneity and immunosuppressive environment in human breast cancer. *Cancer Cell* **33**, 463–479. e10 (2018).
81. Li, B. et al. Cell-type deconvolution analysis identifies cancer-associated myofibroblast component as a poor prognostic factor in multiple cancer types. *Oncogene* **40**, 4686–4694 (2021).
82. Mhaidly, R. & Mechta-Grigoriou, F. Fibroblast heterogeneity in tumor micro-environment: role in immunosuppression and new therapies. *Semin. Immunol.* **48**, 101417 (2020).
83. Asif, P. J., Longobardi, C., Hahne, M. & Medema, J. P. The role of cancer-associated fibroblasts in cancer invasion and metastasis. *Cancers* **13**, 4720 (2021).
84. Kim, I., Choi, S., Yoo, S., Lee, M. & Kim, I.-S. Cancer-associated fibroblasts in the hypoxic tumor microenvironment. *Cancers* **14**, 3321 (2022).
85. Ebbing, E. A. et al. Stromal-derived interleukin 6 drives epithelial-to-mesenchymal transition and therapy resistance in esophageal adenocarcinoma. *Proc. Natl Acad. Sci. USA* **116**, 2237–2242 (2019).
86. Yu, Y. et al. Cancer-associated fibroblasts induce epithelial-mesenchymal transition of breast cancer cells through paracrine TGF-β signalling. *Br. J. Cancer* **110**, 724–732 (2014).
87. Mariotto, A. et al. Expected monetary impact of Oncotype DX score-concordant systemic breast cancer therapy based on the TAILORx trial. *J. Natl Cancer Inst.* **112**, 154–160 (2020).
88. Davis, B. A. et al. Racial and ethnic disparities in Oncotype DX test receipt in a statewide population-based study. *J. Natl Compr. Canc. Netw.* **15**, 346–354 (2017).
89. Losk, K. et al. Factors associated with delays in chemotherapy initiation among patients with breast cancer at a comprehensive cancer center. *J. Natl Compr. Canc. Netw.* **14**, 1519–1526 (2016).
90. Yousif, M. et al. Artificial intelligence applied to breast pathology. *Virchows Arch.* **480**, 191–209 (2022).
91. Abubakar, M. et al. Tumor-associated stromal cellular density as a predictor of recurrence and mortality in breast cancer: results from ethnically diverse study populations. *Cancer Epidemiol. Biomark. Prev.* **30**, 1397–1407 (2021).
92. Li, H. et al. Collagen fiber orientation disorder from H&E images is prognostic for early stage breast cancer: clinical trial validation. *NPJ Breast Cancer* **7**, 104 (2021).
93. Chen, Y. et al. Computational pathology improves risk stratification of a multi-gene assay for early stage ER+ breast cancer. *NPJ Breast Cancer* **9**, 40 (2023).
94. Diao, J. A. et al. Human-interpretable image features derived from densely mapped cancer pathology slides predict diverse molecular phenotypes. *Nat. Commun.* **12**, 1613 (2021).
95. Bejnordi, B. E. et al. Deep learning-based assessment of tumor-associated stroma for diagnosing breast cancer in histopathology images. In *Proc. IEEE Int. Symp. Biomed. Imaging* 929–932 (2017).
96. Lu, M. Y. et al. AI-based pathology predicts origins for cancers of unknown primary. *Nature* **594**, 106–110 (2021).
97. Ehteshami Bejnordi, B. et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* **318**, 2199–2210 (2017).
98. Ghassemi, M., Oakden-Rayner, L. & Beam, A. L. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit. Health* **3**, e745–e750 (2021).
99. Bilal, M. et al. Development and validation of a weakly supervised deep learning framework to predict the status of molecular pathways and key mutations in colorectal cancer from routine histology images: a retrospective study. *Lancet Digit. Health* **3**, e763–e772 (2021).
100. Mercan, C. et al. Deep learning for fully-automated nuclear pleomorphism scoring in breast cancer. *NPJ Breast Cancer* **8**, 120 (2022).
101. Beck, A. H. et al. Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. *Sci. Transl. Med.* **3**, 108ra113 (2011).
102. Karagiannis, G. S. et al. Cancer-associated fibroblasts drive the progression of metastasis through both paracrine and mechanical pressure on cancer tissue. *Mol. Cancer Res.* **10**, 1403–1418 (2012).
103. Yuan, Y. et al. Quantitative image analysis of cellular heterogeneity in breast tumors complements genomic profiling. *Sci. Transl. Med.* **4**, 157ra143 (2012).
104. He, L. et al. Association between levels of tumor-infiltrating lymphocytes in different subtypes of primary breast tumors and prognostic outcomes: a meta-analysis. *BMC Womens Health* **20**, 194 (2020).
105. Denkert, C. et al. Tumour-infiltrating lymphocytes and prognosis in different subtypes of breast cancer: a pooled analysis of 3771 patients treated with neoadjuvant therapy. *Lancet Oncol.* **19**, 40–50 (2018).
106. AbdulJabbar, K. et al. Geospatial immune variability illuminates differential evolution of lung adenocarcinoma. *Nat. Med.* **26**, 1054–1062 (2020).
107. Huang, Z. et al. Artificial intelligence reveals features associated with breast cancer neoadjuvant chemotherapy responses from multi-stain histopathologic images. *NPJ Precis. Oncol.* **7**, 14 (2023).
108. Amgad, M. et al. Report on computational assessment of tumor infiltrating lymphocytes from the International Immuno-Oncology Biomarker Working Group. *NPJ Breast Cancer* **6**, 16 (2020).

Methods

Clinical datasets

TCGA clinical and WSI data were obtained from the National Cancer Institute Genomic Data Commons portal: gdc.cancer.gov. Updated survival outcomes data were obtained from the supplemental files from previous studies[53,60,61]. Breakdown of the Nottingham grade components in TCGA into pleomorphism grade, tubule formation grade, and mitotic grade was obtained from previously parsed pathology reports[109,110]. The degree of lymphovascular invasion was also obtained from pathology reports[110].

The CPS studies are prospective observational cohort studies of cancer risk factors, morbidity, and mortality that the ACS conducted. Deidentified clinical and imaging data were shared with Emory University and Northwestern University through data-sharing agreements. Adult men and women with no personal history of cancer were enrolled in 1982 in CPS-II and 2006–2013 in CPS-3 (refs. 27,111,112). CPS-II participants enrolled in a subcohort followed for cancer incidence and mortality starting in 1992 and 1993[24]. CPS-II cancer incidence follow-up is complete through 2017 (mortality through 31 December 2018). In CPS-3, follow-up is ongoing but currently complete through 2018 (mortality through 2017). We included female participants who developed breast cancer and for whom tissue slides were available for scanning. The included participants were diagnosed between the start of follow-up (CPS-II: 1992–1993; CPS-3: 2006–2013) and the last administrative date (CPS-II: 2017; CPS-3: 2018).

The CPS-II study recruited cancer-free individuals (not selected for any clinical factors, such as family history of cancer) and then followed them forward in time via surveys and linkages to state cancer registries and the National Death Index. The subset of CPS-II participants used to develop HiPS were diagnosed with breast cancer during CPS-II Nutrition Survey cohort follow-up. Breast cancer case ascertainment was conducted by self-reporting on biennial surveys (>90% survey response rate) followed by medical record verification. Tissue blocks from the diagnosis of CPS-II participants were collected from hospitals ranging from community to tertiary-care centers. Generalizability of any study to the entire US population is difficult; however, the unselected strategy of participant enrollment and case ascertainment provides a more generalizable set of breast cancer tissues and outcome data than other available resources today (for example, TCGA).

The PLCO Cancer Screening Trial was designed to evaluate the efficacy of specific screening procedures in reducing mortality rates associated with prostate, lung, colorectal, and ovarian cancers[26]. This PLCO trial adopted a randomized, controlled design and recruited participants from 1993 to 2001. Cancer incidence data were amassed until 31 December 2009, while mortality information was collected through 2015. PLCO female participants who developed breast cancer after enrollment in the trial were included in our study. Deidentified clinical and WSI data were obtained with permission from the National Cancer Institute Cancer Data Access System.

TCGA genomic data

Breast cancer genomic subtype classification; precalculated Buffa, Winter, and Ragnum hypoxia scores; fraction of genome altered; and aneuploidy scores were all obtained from the PanCancer Atlas on Genomic Data Commons: gdc.cancer.gov/about-data/publications/pancanatlas ref. [63,64]. mRNA expression data were also obtained from the PanCancer Atlas[64]. Genomically determined immune subtypes and related pathway activation data were obtained from the PanImmune dataset[77]. Angiogenesis and lymphangiogenesis scores were also obtained from the PanImmune dataset, in which they were scored using known gene expression signatures using bulk RNA sequencing profiles[77]. The proportion of myCAFs, determined using gene expression deconvolution, was obtained from a previous study[81]. Finally, the correlation between HiPS features and cell abundance by genomic deconvolution was determined using published xCell deconvolution scores[61].

Geographic data processing

US state and county data presented in Fig. 1 correspond to facilities where the tissue samples were sourced. Cartographic boundary data were obtained from census.gov for 2021: census.gov/geographies/mapping-files/time-series/geo/cartographic-boundary.html. Facilities in cities that intersect multiple counties were assigned the county that maximally intersects the city, as determined using simplemaps.com/data/us-cities. Facilities in small towns were assigned to the nearest city.

WSI data acquisition and management

CPS-II and CPS-3 slides were scanned at Emory University using the NanoZoomer 2.0-HT slide scanner by Hamamatsu Photonics at a 40× magnification (0.23 µm per pixel). They were saved as .ndpi files. PLCO WSIs were obtained utilizing the Aperio AT2 (Leica) scanner at a 40× magnification using standard settings. They were saved as .svs files at the Cancer Genomics Research Laboratory at the National Cancer Institute. TCGA slides were scanned at different institutions that contributed samples to the data repository; details can be found at the Genomic Data Commons portal (portal.gdc.cancer.gov), and a detailed analysis of site-specific characteristics was provided in a previous study[113].

All WSI management was done using the Digital Slide Archive software platform: digitalslidearchive.github.io/digital_slide_archive ref. [51]. This platform includes a Mongo database that is accessible through a user interface as well as programmatically through RestfulAPI. In addition, we utilized the associated image processing library (HistomicsTK) and WSI visualization and annotation interface (HistomicsUI).

Programming and statistical analysis

All computer programming for clinical and WSI data analysis was done using the Python 3.8 and Bash programming languages. Deep-learning models were developed using the Pytorch library (v. 1.7.1). Statistical tests for specific experiments were discussed with relevant figure and table captions.

Unless stated otherwise, all statistical tests were two sided, and all measurements were taken from distinct patient samples. Pearson and Spearman correlations, independent-sample *t*-test, and the one-way ANOVA statistics were calculated using the 'pearsonr', 'spearmanr', 'ttest_ind', and 'f_oneway' methods of the 'scipy.stats' library (v.1.5.4) using default settings. The 'lifelines' Python package (v.0.27.8) was used for survival analysis, including the 'KaplanMeierFitter' and 'CoxPHFitter', and 'multivariate_logrank_test' using default parameters. For all box plots shown, graphical elements represent the standard representation: center line, median; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range; and points, outliers.

Illustrations

Multiple figure panels were created with BioRender.com. Most plots were created using the 'seaborn' and 'matplotlib Python' libraries and compiled using 'Inkscape'.

Panoptic segmentation model training

We used our panoptic segmentation convolutional neural network (CNN) model, 'MuTILs', to delineate tissue regions and cell nuclei in the slide, a task we hereafter refer to as 'panoptic segmentation' given that it combines semantic segmentation (regions) and object detection (nuclei; Fig. 1b)[46,47]. Supplementary Figs. 2–5 illustrate MuTILs segmentation results on five representative WSIs from the PLCO dataset. We had previously published details of the training and validation procedure, but we summarize it below for convenience[47].

The MuTILs architecture is composed of two U-Net CNN models that work in parallel to segment regions and nuclei at resolutions of 1 µm and 0.5 µm per pixel (MPP), respectively[114]. It is a multi-resolution, multi-task, biologically inspired CNN architecture. To transfer information from the low-resolution branch to the high-resolution branch,

we utilized concatenation, inspired by the HookNet architecture[115]. Region predictions were used to impose constraints on the nucleus class inference to ensure compatibility of the region-level and cell-level predictions. This was achieved using class-specific attention maps, which were derived by modeling the nucleus class previous probability as a linear combination of the corresponding region probability vector. We trained MuTILs on WSI data from two publicly available datasets: the Breast Cancer Semantic Segmentation dataset (region segmentation) and the NuCLS dataset (nucleus detection)[49,50]. Region-level and nucleus-level data were reconciled to produce a single panoptic segmentation dataset[48]. In addition, we expanded the 'training' data by annotating 85 slides from the CPS-II cohort to improve representation of less-advanced cases.

To evaluate our model's performance, we partitioned the slides using fivefold internal–external cross-validation; that is, different institutions were assigned different training and testing folds. All testing set annotations were produced or reviewed by pathologists, as described in previous work[49,50]. MuTILs accuracy statistics include:

- Cancer region segmentation: Sørensen–Dice (DICE) = 82.7 ± 0.4
- Stromal region segmentation: DICE = 80.8 ± 0.4
- Cancer nucleus detection: area under the receiver operating characteristic (AUROC) = 95.9 ± 3.2
- CAF nucleus detection: AUROC = 91.0 ± 3.6
- Lymphocyte nucleus detection: 93.0 ± 1.1
- Nucleus detection micro-average: 92.6 ± 2.8
- Nucleus detection macro-average: 82.7 ± 5.0

Other forms of validation were used as well, including qualitative examination of inference results and correlation between computational and manual TIL scores (Spearman $R = 0.58$, $P < 0.001$). Please refer to our previous work for further details[48].

Panoptic segmentation of WSIs

After we had trained the MuTILs model, we had five sets of weights, each trained on a different fold of training data obtained from a unique set of hospitals (internal–external cross-validation)[116]. We relied on this to reduce the overall bias in our predictions by employing a form of model ensembling that does not increase overall run time and hence more closely reflects realistic clinical deployment. We obtained the region-adjacency graph for each WSI at a very low resolution (20 MPP). Using the 'rag_threshold' function in the 'histolab' library, the low-resolution image was segmented into superpixels using the simple linear iterative clustering algorithm ('slic', from 'scikit-image' (v.0.18.0), scikit-image.org), which were then clustered using *K*-means based on the intensity values[117,118]. We set the 'slic' parameters to 128 segments, a compactness of 10, and a threshold of 9. Each contiguous segment was assigned a single set of MuTILs model weights. In effect, every low-resolution superpixel was predicted by one set of weights, and each set of model weights could be assigned multiple superpixels (Supplementary Fig. 6). Using this approach minimized edge artifacts related to systemic biases in predictions from different model weights.

First, we excluded white space and red, green, and blue markings (felt pen or inking) using the relevant 'histolab' functions (forked and modified from v.0.6.0, available as a submodule of github.com/PathologyDataScience/HiPS). These functions rely on color thresholding. Next, we tiled each WSI into 512 × 512 µm square regions of interest (ROIs), keeping only tiles with at least 50% tissue composition. Second, we performed deconvolution-based color normalization using the Macenko method, restricting the color normalization to the tissue mask using the 'HistomicsTK' library (v.1.2.10)[119]. Third, each ROI was assigned a single set of MuTILs model weights for inference depending on which superpixel it maximally overlaps with. Once we had obtained the CNN inference results for regions and cells, we saved a WSI segmentation mask for later use in feature extraction.

Histomic feature extraction

Using our previously trained and validated panoptic segmentation model (MuTILs), we were able to delineate all tissue regions and nuclei within WSIs automatically. These regions and nuclei were then used to extract several morphological and contextual features. Because each slide contains thousands of regions and up to a million nuclei, these features were aggregated to obtain a single mean and variance value per patient, as we discuss in the next section. After feature aggregation, each patient sample was thus described by a set of 109 numbers, that is, the patient-level histomic features. Supplementary Table 5 describes each of these features, and we provide a general overview below.

Global features. These include overall cancer cell density, overall TIL density, the global amount of necrosis within the WSI, and other aggregate measures. These values were normalized to the amount of tissue analyzed.

Region morphology. Standard size and shape measurements were extracted for each tissue region, focusing on epithelial and TIL-dense regions. We segment tissue regions at a 1 MPP resolution. The output is saved as a WSI mask, which is then loaded for region feature extraction at a 2 MPP resolution. Saving the WSI mask enabled us to extract features from contiguous tissue regions that may span multiple ROIs; hence, this approach mitigates the effects of artifacts related to the ROI edge. To obtain individual regions from the semantic segmentation mask, we first did binary dilation of the mask using a 5 × 5 pixel selem (that is, 10 × 10 µm) to remove small artifacts. Then, we removed holes smaller than 48^2 µm^2 in area. Finally, we performed a connected component analysis using a connectivity of 2. Only tissue regions with an area larger than 128^2 µm^2 were considered.

Region morphology features were extracted using the 'scikit-image' library's 'regionprops' method (v.0.18.0). Boundary complexity was measured using fractal (box counting) dimension, as implemented by N. P. Rougier in the GitHub repository gist.github.com/rougier/e5eafc276a4e54f516ed5559df4242c0. We note that the MuTILs model detects contiguous TIL aggregates at one MPP resolution. Hence, we did not need to use global clustering or graph-based methods to identify TIL cluster boundaries as they were determined in a data-driven manner using semantic segmentation.

Standard nuclear morphology. Nuclear features were extracted for the three nuclear superclasses: epithelial, non-TIL stromal (mostly CAFs), and TILs. Nuclear size, shape, staining intensity, boundary complexity, edges (chromatin clumping), and texture were extracted using the 'HistomicsTK' library's 'compute_nuclei_features' method: github.com/DigitalSlideArchive/HistomicsTK. The HistomicsTK implementation is primarily based on the 'scikit-image' library's 'regionprops' method for the size and shape features, Canny edge detectors and Haralick features for texture, and fractal dimension for boundary complexity[28].

Deep nuclear morphology. We observed that nuclei do not always have typical morphology and that they are often ambiguous and difficult to classify in H&E images without IHC markers. Deep-learning models produce a classification probability vector, and we used those probabilities to capture the degree of conformity of nuclei to various classifications of interest.

TIL activation. This refers to the probability that a particular nucleus is classified as a plasma cell, divided by the overall probability that the nucleus belongs to the TIL superclass. In the ground truth, the plasma cell class was not determined using IHC, nor was it limited to the most typical morphology. Hence, it refers to large TILs, including plasma cells and others.

Epithelial nuclear atypia. This refers to the probability that a certain nucleus is classified as a cancer cell divided by the total probability that it belongs to the epithelial class.

CAF epithelialization. This refers to the probability that a nucleus is classified as a fibroblast, divided by the sum of its fibroblast and cancer cell classification probabilities.

Cytoplasmic texture and staining. The delineation of cytoplasmic boundaries in H&E is unreliable, so we calculated texture statistics within 4 µm of nuclear boundaries. This search area is determined by dilating nuclear boundaries. The 'HistomicsTK' library was used for extracting these features.

Local cell density. Local cell density was defined as the average number of cells of a certain class within a predefined radius from the 'central' cell. The central nucleus can have the same class as the surrounding nuclei; for example, the 'LocalTILsDensity32uM' metric measures how many TILs are within 32 µm of the typical TIL. Alternatively, the central and surrounding can have different classifications; for example, 'TILsDensityWithin32uMOfEpithCell' measures how many TILs are within 32 µm of the typical epithelial cell. This statistic was calculated using a fast K-D tree implementation, loosely based on the implementation by S. P. Ingram in the following GitHub repository: github.com/SamPIngram/RipleyK.

Local cell clustering. Local cell clustering was based on Ripley's K-function at a single distance, which is a measure of clustering beyond that expected from random chance[120,121]. We obtained this metric by normalizing local cell density estimates to 'complete spatial randomness'. For example, there is a higher chance that more lymphocytes will surround another lymphocyte by random chance, just because there are so many of them. Hence, high density does not necessarily indicate clustering beyond random chance. On the other hand, just a few fibroblasts surrounding each other may result in a high clustering value since they are (globally) less dense, so there is a lower chance of this dense local aggregation occurring by random chance. The radii used for the calculation of local cell density and clustering were 16, 32, and 64 µm.

Region composition. Region composition refers to the cellular composition of various histopathological tissue compartments. For example, 'NoOfLowGradeNucleiPerEpithNest' measures the number of epithelial nuclei that were considered low grade by the MuTILs model, per epithelial nest. Region composition metrics also enabled us to estimate the nuclear-to-cytoplasmic ratio. Again, cytoplasmic boundaries cannot be precisely determined in H&E slides, so we relied on the following heuristic to calculate the nuclear-to-cytoplasmic ratio: divide the total nuclear area within an epithelial nest by the overall area of the epithelial nest.

Region neighborhood composition. Region masks were morphologically dilated to identify the tissue and cellular composition within 128 µm and 256 µm of the edge. For example, 'CAFDensityAtEpithNestMargin' measures the density of CAFs within 128 µm of epithelial nests.

Acellular stromal matrix and collagen. Supplementary Fig. 10 illustrates features that capture stromal matrix, including abstract texture and intensity measurements, as well as a more sophisticated analysis of the separation, length, and disorder of collagen fiber orientations. We captured collagen disorder by three separate approaches.

First, we hypothesized that collagen separation and stromal matrix discoloration (for example, desmoplasia) would reflect on abstract intensity and texture measurements from the collagen stroma. 'PeriCAFMatrixHeteroIn512uMROI' is a feature that captures the variation in stromal matrix at the interface between desmoplastic and quiescent stroma. The metric is calculated by measuring the average intensity within a very thin rim around each fibroblast and calculating the variance in that intensity across a 512 × 512 µm squared ROI. This metric is related to one of the prognostic stromal features described in a previous study that relied on the absolute difference in intensity between neighboring contiguous stromal regions[101]. In this study, the approach may have been liable to some confounding by segmentation errors or non-stromal matrix elements such as small vessels and vacuoles. To address this issue, we relied on the peri-fibroblast stromal matrix within 4 µm. All images were color normalized using the Macenko method to maximize robustness to staining and scanner differences[119].

Second, we took a direct approach whereby we detected the collagen fibers themselves, largely following the methodology described in a previous study[92]. This analysis used 256 × 256 µm ROIs with at least 30% stroma and 20% tumor at a 0.5 MMP resolution. We used a Canny edge detection algorithm to detect the interface where collagen fibers separate. Then, we used connected component analysis to isolate individual edges. Fibers with a minor-to-major axis ratio <0.2 were considered straight fibers and were further admitted to calculate the collagen fiber orientation disorder metric described in the previous study[92]. This measures the degree of disorder in collagen orientation, calculated from a length-weighted orientation co-occurrence matrix. One difference between our implementations is that we masked out nuclei before applying the edge detection in order to minimize confounding by nuclear material.

Finally, we also measured collagen entropy indirectly by calculating the entropy of orientations of fibroblast nuclei within a certain radius of each other. We hypothesized that in some settings, fibroblast nuclei might be more reliably detected than collagen fibers.

Histomic feature aggregation and processing

Histomic features had to be aggregated to obtain patient-level data from tissue-region-level, cell-level, ROI-level, and slide-level data. For each patient, we had 1,000+ tissue regions, 100,000+ nuclei, multiple ROI-level histomic features (including descriptors of the acellular stroma), and 1–3 slides. Supplementary Fig. 9 illustrates the ROI-level spatial variability in two influential features before WSI-level aggregation: 'ChromatinClumpingOfEpithNuclei' and 'PeriCAFMatrixHeteroIn512uMROI'. Supplementary Table 5 provides details of the various levels of aggregation used for different features, and we provide an overview below:

- Individual tissue region morphology and neighborhood information were aggregated to obtain per-WSI mean and standard deviation.
- Nuclear clustering, interaction features, and global densities were obtained for the entire WSI to avoid artifacts related to artificial ROI boundaries. No aggregation was needed.
- Individual nuclear morphology was first aggregated per 512 × 512 µm ROI using mean and standard deviation. Each ROI was assigned a 'saliency score', which is maximized for ROIs with a high composition of adjacent epithelial and stromal tissue regions, hence prioritizing the cancer–stroma interface[48]. The saliency score is calculated as follows:

$$\text{Saliency} = (\text{Area of stroma within 32 µm of an epithelial nest} \times \text{Area of epithelial tissue}) / (\text{Area of non} - \text{necrotic tissue}).$$

Note that solely relying on stroma within 32 µm from the tumor is inadequate, as ROIs with scattered, spaced-out tumor nests would get a high score even though there is little tumor. The per-ROI data were aggregated using saliency-weighted mean and standard deviation, restricted to the top 128 most-salient ROIs.

- Features describing the acellular stromal matrix and collagen were aggregated per WSI using saliency-weighted mean and standard deviation of the per-ROI data, restricted to the top 256 most-salient 256 × 256 µm ROIs. Saliency was calculated in the same way as previously described.
- Finally, per-WSI data from multiple slides were averaged to obtain per-patient information.

After aggregation, all histomic features were *z*-scored relative to the CPS-II mean, omitting missing values. When features had missing values, this was because of unavailable tissue, such as when there were no TIL clusters to calculate cluster morphology. These values were filled using 10-nearest neighbors imputation using the CPS-II cohort.

Thematic classification of histomic features

Histomic features fall under five major themes, which are further divided into 26 subthemes. Themes and subthemes were engineered in a hypothesis-driven manner and are meant to capture distinct biological phenomena and processes (Fig. 2). The themes are epithelial features, stromal features, TIL features, necrosis abundance, and features capturing various interactions between different cells. Subthemes encompass specific phenomena. For example, the TIL theme is subdivided into TIL abundance, TIL clustering, TIL cluster morphology, TIL cluster pleomorphism, and so on. We expect histomic features within the same themes and subthemes to be correlated with each other and less correlated with other themes. The exception, of course, is the interactions theme, which may or may not be correlated with others. Examining the absolute correlation matrix, we can see that the correlation is stronger towards the diagonal, within the squares representing themes and subthemes (Fig. 2). One of the themes we focused on is characterizing cancer-associated stroma as standard Nottingham grading does not capture it. Stromal features include morphological descriptors of fibroblast nuclei, characterization of TILs, and detailed analysis of the stromal matrix, which is primarily composed of type I collagen fibers (Supplementary Fig. 10). All of these measurements were assessed as potential prognostic indicators, and the most prognostic ones were admitted into the final prognostic score.

HiPS prognostic model fitting

The most prognostic feature from each biological subtheme at the univariable level was admitted into this model, along with the standard IHC marker panel: ER expression, PR expression, and HER2+ overexpression (Fig. 2). In addition, we created a composite metric for triple-negative status (TNBC), defined as the absence of all ER, PR, and HER2 markers. The rationale for incorporating IHC markers is to find histomic features that provide excess prognostic values beyond that already defined by the expression of hormone and HER2 receptors. We did not want to learn histological features that were highly correlated with, say, TNBC status as they would not be clinically helpful. In a clinical setting, the practicing pathologist and oncologist always have access to at least the histological slide and the ER, PR, and HER2 IHC and ISH markers[3,122].

A total of 30 features (26 subthemes and 4 IHC panel features) were entered into an elastic-net regularized Cox proportional hazards survival model, predicting BCSS with the CPS-II patient cohort[123]. The optimal hyperparameters (alpha and L1 ratio) for model regularization were obtained by cross-validation (Supplementary Figs. 7 and 42). The trained model was then used to predict the log partial hazard for the entire training population, and the predictions were scaled such that the resultant continuous score ranges from 0 (lowest risk) to 10 (highest risk). In addition, we modeled the predicted risk scores as a mixture of three Gaussian curves, representing the low-, intermediate-, and high-risk populations. The points where curves cross were then considered a data-driven cutoff for dividing patients into three survival risk groups.

To ensure a fair head-to-head comparison of our HiPS, we used the same methodology to develop a baseline model fit only to the Nottingham grade and the standard ER, PR, and HER2 panel (Supplementary Fig. 7). This baseline model also yields a risk score in the range 0–10 and three risk groups. As grading is discrete, unlike our histomic features, the resulting histogram of predicted risks contained discrete bins, so a mixture of Gaussian could not faithfully represent it. Instead, we divided the score range into three equal intervals.

Multivariable Cox proportional hazards regression

After we had learned the optimum combination of histomic features comprising the HiPS and control models, and the optimum thresholds to learn discrete risk groups, we produced the following features for each patient:

- HiPS score: This number ranges from 0 to 10.
- Histomic prognostic group: one of three risk groups: H1, H2, and H3, corresponding to low, intermediate, and high risk.
- Control score: ranges from 0 to 10, using the baseline model (Nottingham grade and ER, PR, and HER2).
- Control group: one of three risk groups, C1, C2, and C3, corresponding to low, intermediate, and high risk.

Using the CPS-II cohort, we fit multivariable models to predict BCSS using each of the risk scores and groups. There were missing clinical data, so we explored two multivariable models. The first controls only for pathologic stage and tumor size and is a robust model with maximal sample size. We also fit another model using a smaller set of patients with complete clinical information on pathologic TNM stage, tumor size, whether the cancer was detected using proactive screening, menopausal status at diagnosis, race, smoking history, age at diagnosis, body mass index, and expression of basal markers CK5/6 or EGFR.

We explored the independent prognostic value of the risk scores and groups in the PLCO and TCGA datasets as well. However, we were limited by the constraints of missing data limiting the power of prognostic modeling. The clinico-epidemiologic variables for which data were missing differed for different patients; for example, we may have the cause-of-death information but not the race. Hence, we used OS in PLCO and TCGA and always controlled for TNM stage and tumor size. Guided by model fit statistics (primarily the concordance index), we also controlled for patient age in PLCO and gene expression intrinsic subtype in TCGA.

We checked hazard proportionality using the 'check_assumption' function of the 'lifelines' package using a *P* value threshold of 0.01: lifelines.readthedocs.io/en/latest/jupyter_notebooks/Proportional%20hazard%20assumption.html. We also visually inspected the scaled Schoenfeld residuals as needed. Depending on context, we prioritized answering the preset hypotheses over post hoc changes to model design[124,125].

GSEA

GSEA was performed using the Python package 'GSEAPy (v.1.0.6)': github.com/zqfang/GSEApy/, which is a Python implementation of GSEA and a wrapper for Enrichr[70,126,127]. We obtained mRNA expression data from the TCGA PanCancer Atlas supplementary file data_mrna_seq_v2_rsem_zscores_ref_all_samples.txt[64]. The data file contained log-transformed mRNA expression data that have been *z*-score normalized relative to the expression distribution of all samples from all cancers ('RNA Seq V2 RSEM'). After restricting the data to breast cancer patients, genes with >80% missing data were removed from the analysis. We then calculated the Pearson correlation coefficients between HiPS scores and histomic features with the mRNA expression data. Using the 'prerank' module, we ran the analysis with 4 threads, between 5 and 1,000 genes per gene set, and using 1,000 phenotype permutations. Depending on the hypothesis, we ran the analyses using either the 'MSigDB_Hallmark_2020' or 'Azimuth_Cell_Types_2021' collections, imported directly from gseapy[11,71].

Ethics statement

All study participants participated voluntarily and provided informed consent. CPS-II and CPS-3 data sharing was approved through the Emory University Institutional Review Board, approval numbers IRB00045780 and IRB00059007. Deidentified CPS-II and CPS-3 data were obtained through a data-sharing agreement between the ACS, Emory University, and Northwestern University. Deidentified PLCO data were obtained with permission from the National Cancer Institute Cancer Data Access System.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Supplementary Table 27 contains our calculated histomic feature values, HiPS scores and subscores, and related data for the TCGA cohort. We provide this to facilitate reproducibility and to act as a resource for the scientific community. TCGA clinical data and WSIs are publicly available at gdc.cancer.gov. The Breast Cancer Semantic Segmentation dataset is available at github.com/PathologyDataScience/BCSS, and the NuCLS dataset is available at sites.google.com/view/nucls. These datasets were combined to produce the PanopTILs dataset, available at sites.google.com/view/panoptils. Requests for ACS data from the CPS-II or CPS-3 studies should be submitted to maddison.hall@cancer.org. Requests for PLCO data should be submitted at cdas.cancer.gov/learn/plco. Breast cancer genomic subtypes, hypoxia scores, fraction genome altered, aneuploidy scores, and mRNA expression profiles were obtained from the Genomic Data Commons Pancancer Atlas: gdc.cancer.gov/about-data/publications/pancanatlas. Immune subtypes and related pathway activations, as well as angiogenesis and lymphangiogenesis scores, were obtained from the PanImmune dataset: gdc.cancer.gov/about-data/publications/panimmune. CAF subtype abundance data for TCGA were obtained from a previous study[81]. xCell cell type abundance data for TCGA were obtained from a previous study[61].

Code availability

The code is publicly available at github.com/PathologyDataScience/HiPS. Processing of histology images was performed using HistomicsTK (v.1.2.10, github.com/DigitalSlideArchive/HistomicsTK), histolab (v.0.6.0, github.com/histolab/histolab), and scikit-image (v.0.18.0, scikit-image.org). Analysis of clinical outcomes data was performed using Lifelines (v.0.27.8, github.com/CamDavidsonPilon/lifelines). Enrichment analysis with RNA profiles was performed using GSEAPy (v.1.0.6, gseapy.rtfd.io). Additional Python libraries used for database management, graphical plotting, scientific calculations, and other tasks include numpy v.1.19.4, pandas v.1.1.5, SQLAlchemy v.1.3.21, scipy v.1.5.4, scikit-learn v.0.23.2, imageio v.2.9.0, pillow v.8.0.1, matplotlib v.3.3.3, seaborn v.0.11.0, torch v.1.7.1, torchvision v.0.8.2, and pyvips v.2.1.15.

References

109. Ping, Z. et al. A microscopic landscape of the invasive breast cancer genome. *Sci. Rep.* **6**, 27545 (2016).
110. Thennavan, A. et al. Molecular analysis of TCGA breast cancer histologic types. *Cell Genom.* **1**, 100067 (2021).
111. Garfinkel, L. Selection, follow-up, and analysis in the American Cancer Society prospective studies. *Natl Cancer Inst. Monogr.* **67**, 49–52 (1985).
112. Stellman, S. D. & Garfinkel, L. Smoking habits and tar levels in a new American Cancer Society prospective study of 1.2 million men and women. *J. Natl Cancer Inst.* **76**, 1057–1063 (1986).
113. Howard, F. M. et al. The impact of site-specific digital histology signatures on deep learning model accuracy and bias. *Nat. Commun.* **12**, 4423 (2021).
114. Ronneberger, O., Fischer, P. & Brox, T. U-Net: convolutional networks for biomedical image segmentation. In *Proc. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (eds Navab, N. et al.) 234–241 (Springer, 2015).
115. van Rijthoven, M., Balkenhol, M., Siliņa, K., van der Laak, J. & Ciompi, F. HookNet: multi-resolution convolutional neural networks for semantic segmentation in histopathology whole-slide images. *Med. Image Anal.* **68**, 101890 (2021).
116. Steyerberg, E. W. & Harrell, F. E. Prediction models need appropriate internal, internal-external, and external validation. *J. Clin. Epidemiol.* **69**, 245–247 (2016).
117. Marcolini, A. et al. histolab: a Python library for reproducible digital pathology preprocessing with automated testing. *SoftwareX* **20**, 101237 (2022).
118. Achanta, R. et al. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Trans. Pattern Anal. Mach. Intell.* **34**, 2274–2282 (2012).
119. Macenko, M. et al. A method for normalizing histology slides for quantitative analysis. in *2009 IEEE Int. Symposium on Biomedical Imaging: from Nano to Macro* 1107–1110 (IEEE, 2009); https://doi.org/10.1109/ISBI.2009.5193250
120. Ripley, B. D. The second-order analysis of stationary point processes. *J. Appl. Probab.* **13**, 255–266 (1976).
121. Amgad, M., Itoh, A. & Tsui, M. M. K. Extending Ripley's K-function to quantify aggregation in 2-D grayscale images. *PLoS ONE* **10**, e0144404 (2015).
122. Lester, S. C. et al. Protocol for the examination of specimens from patients with invasive carcinoma of the breast. *Arch. Pathol. Lab. Med.* **133**, 1515–1538 (2009).
123. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–22 (2010).
124. Campbell, H. & Dean, C. B. The consequences of proportional hazards based model selection. *Stat. Med.* **33**, 1042–1056 (2014).
125. Stensrud, M. J. & Hernán, M. A. Why test for proportional hazards? *JAMA* **323**, 1401–1402 (2020).
126. Fang, Z., Liu, X. & Peltz, G. GSEApy: a comprehensive package for performing gene set enrichment analysis in Python. *Bioinformatics* **39**, btac757 (2023).
127. Chen, E. Y. et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinform.* **14**, 128 (2013).

Acknowledgements

We express sincere appreciation to all CPS-II and CPS-3 participants and to each member of the study and biospecimen management group. We would like to acknowledge the contributions to this study from central cancer registries supported through the Centers for Disease Control and Prevention's National Program of Cancer Registries and cancer registries supported by the National Cancer Institute's Surveillance Epidemiology and End Results Program. We thank the National Cancer Institute for access to NCI's data collected by the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial. We are grateful to the annotation team for the Breast Cancer Semantic Segmentation and NuCLS datasets. We would also like to acknowledge F.M. Howard and A.T. Pearson (University of Chicago) for providing us with the research-use Oncotype DX and MammaPrint scores for TCGA. Figures 1–4 and 6, and multiple supplementary figures, were created in part using BioRender.com. This work was supported by the US National Institutes of Health grants U01CA220401 and U24CA19436201. The ACS funds the creation, maintenance, and updating of the CPS-II and CPS-3 cohorts.

Author contributions

M.A. and L.A.D.C. conceived of the research idea. M.A. carried out data analysis, model development, and model validation. M.A., M.A.T.E.,

and L.A.D.C. wrote the paper, and J.M.H., C.B., K.P.S., J.A.G., M.M.G., and L.R.T. edited the paper. J.M.H. and S.P. performed data curation for the Cancer Prevention Studies cohorts. K.P.S. provided expertise on breast cancer pathology. C.B., M.M.G., and L.R.T. provided expertise on breast cancer epidemiology and population science and assisted with the interpretation of results. D.A.G. provided assistance with computing and data visualization. L.R.T. and L.A.D.C. jointly supervised the work.

Competing interests

L.A.D.C. has invention disclosures registered at the Northwestern Office of Innovation and New Ventures, consults for Tempus, and advises Veracyte and Targeted Bioscience. D.A.G. holds stock options in Histowiz LLC and is a cofounder and stockholder of Switchboard, MD. The other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41591-023-02643-7.

Correspondence and requests for materials should be addressed to Lee A. D. Cooper.

Peer review information *Nature Medicine* thanks Po-Hsuan Cameron Chen and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Lorenzo Righetto, in collaboration with the *Nature Medicine* team.

nature portfolio

Corresponding author(s): Lee A.D. Cooper

Last updated by author(s): Oct 12, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used for data collection.
Data analysis	Code is publicly available at: github.com/PathologyDataScience/HiPS. Processing of histology images was performed using HistomicsTK (v.1.2.10, github.com/DigitalSlideArchive/HistomicsTK), histolab (v.0.6.0, github.com/histolab/histolab), and scikit-image (v.0.18.0, scikit-image.org). Analysis of clinical outcomes data was performed using Lifelines (v.0.27.8, github.com/CamDavidsonPilon/lifelines). Enrichment analysis with RNA profiles was performed using GSEAPy (v.1.0.6, gseapy.rtfd.io). Additional python libraries used for database management, graphical plotting, scientific calculations, and other tasks includes: numpy v.1.19.4, pandas v.1.1.5, SQLAlchemy v.1.3.21, scipy v.1.5.4, scikit-learn v.0.23.2, imageio v.2.9.0, pillow v.8.0.1, matplotlib v.3.3.3, seaborn v.0.11.0, torch v.1.7.1, torchvision v.0.8.2, and pyvips v.2.1.15.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Table S27 contains our calculated histomic feature values, HiPS scores and subscores, and related data for the TCGA cohort. We provide this to facilitate reproducibility and to act as a resource for the scientific community. TCGA clinical data and WSIs are publicly available at: gdc.cancer.gov. The Breast Cancer Semantic Segmentation dataset is available at: github.com/PathologyDataScience/BCSS, while the NuCLS dataset is available at: sites.google.com/view/nucls. These datasets were combined to produce the PanopTILs dataset, available at: sites.google.com/view/panoptils. Requests for American Cancer Society data from the CPS-II or CPS-3 studies should be submitted to maddison.hall@cancer.org. Requests PLCO data should be submitted at: cdas.cancer.gov/learn/plco. Breast cancer genomic subtypes, hypoxia scores, fraction genome altered, aneuploidy scores, and mRNA expression profiles were obtained from the Genomic Data Commons Pancancer Atlas: gdc.cancer.gov/about-data/publications/pancanatlas. Immune subtypes and related pathway activations, as well as angiogenesis and lymphangiogenesis scores were obtained from the PanImmune dataset: gdc.cancer.gov/about-data/publications/panimmune. CAF subtype abundance data for TCGA was obtained from: Li, B. et al. Cell-type deconvolution analysis identifies cancer-associated myofibroblast component as a poor prognostic factor in multiple cancer types. Oncogene 40, 4686–4694 (2021). xCell cell type abundance data for TCGA was obtained from: Aran, D., Hu, Z. & Butte, A. J. xCell: digitally portraying the tissue cellular heterogeneity landscape. Genome Biol 18, 220 (2017).

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Not applicable.
Reporting on race, ethnicity, or other socially relevant groupings	Provided in Table S3.
Population characteristics	Provided in Tables S1-4.
Recruitment	- TCGA dataset: Participant recruitment details can be accessed through the National Cancer Institute Genomic Data Commons (GDC) portal: https://gdc.cancer.gov/. - The CPS studies are prospective observational cohort studies of cancer risk factors, morbidity, and mortality that the American Cancer Society conducted. Deidentified clinical and imaging data was shared with Emory University and Northwestern University through data sharing agreements. Adult men and women with no personal history of cancer were enrolled in 1982 in CPS-II and 2006-2013 in CPS-3. CPS-II participants enrolled in a subcohort followed for cancer incidence and mortality starting in 1992/1993. CPS-II cancer incidence follow-up is complete through 2017 (mortality through 12/31/2018). In CPS-3, follow-up is ongoing but currently complete through 2018 (mortality through 2017). We included female participants who developed breast cancer and for whom tissue slides were available for scanning. The included participants were diagnosed between the start of follow-up (CPS-II: 1992-1993, CPS-3: 2006-2013) and the last administrative date (CPS-II: 2017, CPS-3: 2018). Details are provided in the following publications: |_ Calle, E. E. et al. The American Cancer Society Cancer Prevention Study II Nutrition Cohort: rationale, study design, and baseline characteristics. Cancer 94, 2490–501 (2002). |_ Patel, A. V et al. The American Cancer Society's Cancer Prevention Study 3 (CPS-3): Recruitment, study design, and baseline characteristics. |_ Garfinkel, L. Selection, follow-up, and analysis in the American Cancer Society prospective studies. Natl Cancer Inst Monogr 67, 49–52 (1985). |_ Stellman, S. D. & Garfinkel, L. Smoking habits and tar levels in a new American Cancer Society prospective study of 1.2 million men and women. J Natl Cancer Inst 76, 1057–63 (1986). Cancer 123, 2014–2024 (2017). - The Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial was designed to evaluate the efficacy of specific screening procedures in reducing mortality rates associated with prostate, lung, colorectal, and ovarian cancers. This PLCO trial adopted a randomized, controlled design and recruited participants from 1993 to 2001. Cancer incidence data were amassed until December 31, 2009, while mortality information was collected through 2015. PLCO female participants who developed breast cancer after enrollment in the trial were included in our study. Deidentified clinical and WSI data were obtained with permission from the National Cancer Institute Cancer Data Access System. Details are provided in the following publication: Zhu, C. S. et al. The Prostate, Lung, Colorectal and Ovarian Cancer (PLCO) Screening Trial Pathology Tissue Resource. Cancer Epidemiol Biomarkers Prev 25, 1635–1642 (2016).
Ethics oversight	All study participants participated voluntarily and provided informed consent. CPS-II and CPS-3 data sharing was approved through the Emory University Institutional Review Board, approval number IRB00045780 and IRB00059007. Deidentified CPS-II and CPS-3 data were obtained through a data sharing agreement between the American Cancer Society, Emory University, and Northwestern University. Deidentified PLCO data was obtained with permission from the National Cancer Institute Cancer Data Access System.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We used deidentified data from preexisting studies that we did not design. Details regarding sample size choice are published with each of the publicly available datasets used, including The Cancer Genome Atlas (https://www.cancer.gov/ccg/research/genome-sequencing/tcga), Cancer Prevention Studies (https://www.cancer.org/research/population-science/cancer-prevention-and-survivorship-research-team/acs-cancer-prevention-studies.html) and the PLCO trial (https://cdas.cancer.gov/plco/).
Data exclusions	All non-metastatic breast cancer cases were included in the analysis.
Replication	- We provide all code needed to replicate this study at: github.com/PathologyDataScience/HiPS - To replicate TCGA cohort findings, we provide Table S27 which contains our calculated histomic feature values, HiPS scores and subscores, and related data for the TCGA cohort. - Requests for CPS-II or CPS-3 data should be submitted to the ACS. - Requests PLCO data should be submitted at: https://cdas.cancer.gov/learn/plco/.
Randomization	Not applicable because we used retrospective data from preexisting studies that we did not design. Our study is an exploratory study, not a prospective clinical trial.
Blinding	Not applicable, because our study is an exploratory study based on prior published retrospective data, not a prospective clinical trial.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	Not applicable. We used deidentified clinical data from the preexisting TCGA, CPS-II, CPS-3, and PLCO studies.
Study protocol	Not applicable. We used deidentified clinical data from the preexisting TCGA, CPS-II, CPS-3, and PLCO studies.
Data collection	Not applicable. We used deidentified clinical data from the preexisting TCGA, CPS-II, CPS-3, and PLCO studies.
Outcomes	Not applicable. We used deidentified clinical data from the preexisting TCGA, CPS-II, CPS-3, and PLCO studies.