Multimodal Models to Stratify Ovarian Cancer Patients

by

Kevin M Boehm

A Dissertation

Presented to the Faculty of the Louis V. Gerstner, Jr.

Graduate School of Biomedical Sciences,

Memorial Sloan Kettering Cancer Center

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

New York, NY

September 2021

_____ _____

Sohrab Shah, PhD
Dissertation Mentor

Date

To my parents, DeAnn and Michael.

ABSTRACT

Advances in quantitative biomarker development have accelerated new forms of data-driven insights for cancer patients. However, most approaches are limited to a single modality of data, leaving integrated approaches across modalities relatively underdeveloped. Multimodal integration of advanced molecular diagnostics, radiologic and histologic imaging, and codified clinical data presents opportunities to advance precision oncology beyond genomics. Patients with high-grade serous ovarian cancer demonstrate poor prognosis and variable response to treatment. Homologous recombination deficiency status, patient age, pathologic stage, and residual disease status after cytoreductive surgery are important prognostic factors, with recent work also highlighting prognostic information captured in computed tomography and histopathologic specimens. However, little is known about the capacity of combined features to discriminate between patients and explain clinical outcomes. Herein, we developed and integrated histopathologic, radiologic, and clinico-genomic machine learning models to determine their combined impact on risk stratification. We assembled a multimodal dataset of 409 high-grade serous ovarian cancer patients and showed that human-interpretable features, such as necrosis on H&E and omental textural complexity on CT, are associated with worse prognosis. We then integrated these models and demonstrated that the imaging models contain complementary—rather than purely mutual—prognostic information to clinico-genomic prognostic factors, as evidenced by improved risk stratification and stronger association with pathological chemotherapy response score than

unimodal models. This work empirically supports multimodal machine learning approaches as a promising path toward improved risk stratification of cancer patients.

BIOGRAPHICAL SKETCH

Kevin was born and raised in Northern Virginia, where he graduated from Thomas Jefferson High School for Science and Technology in 2011. He then moved to New Haven, Connecticut in to attend Yale University, during which time he interned at the National Institutes of Health and studied abroad in Kanazawa, Japan. After earning a bachelor's degree in biomedical engineering in 2015, Kevin moved to New York to pursue his MD at Weill Medical College of Cornell University and his PhD at Gerstner Graduate School of Memorial Sloan Kettering Cancer Center.

ACKNOWLEDGMENTS

Sohrab Shah, Olaf Andersen, Sam Bakhoum, Ingo Mellinghoff, Christina Iacobuzio-Donahue, Florian Markowetz, Yulia Lakhman, Nicole Rusk, Mike Overholtzer, Linda Burnley, David McDonagh, Ushma Neill, Christina Leslie, Catharine Boothroyd, Renee Horton, Hanna Silvast, Concynella Graham-Wright, Margie Mendoza, Olivier Elemento, John Chodera, Yoon Kang, John Ng, Mohammed Alghamdi, Lora Ellenson, Robert Soslow, Dmitriy Zamarin, Kara Long Roche, Ying Liu, Pier Selenica, Justin Jee, Pegah Khosravi, Rami Vanguri, Jia Luo, JianJiong Gao, Arfath Pasha, Andrew Aukerman, Doori Rose, Druv Patel, Ignacio Vázquez-García, Emily Aherne, Ines Nikolovski, Pamela Causa Andrieu, Junting Zheng, Marinela Capanu, Jorge Reis-Filho, Natalie Gangai, Ramon Sosa, Anika Begum, Samantha Leung, Harini Veeraraghavan, Andrew McPherson, Wesley Tansey, Viki Bojilova, Nick Ceglia, Tyler Funnell, Diljot Grewal, Minsoo Kim, Jamie Lim, Hongyu Shi, Maryam Pourmaleki, Adam Weiner, Marc Williams, Douglas Abrams, Eli Havasov, Florian Uhlitz, Yessie Werner, Cristina Radu, Cynthia Berry, the MSK-MIND Program, the Weill Cornell/Rockefeller/Sloan Kettering Tri-Institutional MD-PhD Program, The Tri-Institutional Computational Biology and Medicine PhD Program, Gerstner Sloan Kettering Graduate School of Biomedical Sciences, Award T32GM007739 of the National Institute of General Medical Sciences, Cycle for Survival, the Jonathan Grayer Fellowship, and Award F30CA257414 of the National Cancer Institute.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AI	artificial intelligence
C-index	concordance index
CAM	class activation map
CE-CT	contrast-enhanced computerized tomography
CGR	complete gross resection
CNA	copy number aberration
CNN	convolutional neural network
COSMIC	catalog of somatic mutations in cancer
CPH	Cox Proportional Hazards
CRS	chemotherapy response score
CT	computerized tomography
DICOM	Digital Imaging and Communications in Medicine
DL	deep learning
DNA	deoxyribonucleic acid
DNN	deep neural network
GDC	Genomic Data Commons
GE	General Electric
GLCM	gray level co-occurrence matrix
GLDM	gray level dependence matrix
GLRLM	gray level run length matrix
GLSZM	gray level size zone matrix
H&E	hematoxylin and eosin
HGSOC	high-grade serous ovarian cancer
HRD	homologous recombination deficiency
HRD	homologous recombination proficiency
HRD-DDR	homologous recombination deficiency DNA damage response
ICB	immune checkpoint blockade
KM	Kaplan-Meier
LST	large-scale state transition
ML	machine learning
MRI	magnetic resonance imaging
MSI	microsatellite instability
MSKCC	Memorial Sloan Kettering Cancer Center
NACT-DPS	neoadjuvant chemotherapy / delayed primary surgery
NGS	next-generation sequencing
NGTDM	neighboring gray tone difference matrix
NtAI	num. subchrom. regions with allelic imbal. extending to telomere
OHDSI	Observational Health Data Sciences and Informatics

OS	overall survival
PACS	picture archiving and communication system
PARP	poly-ADP ribose polymerase
PDS	primary debunking surgery
PET	positron emission tomography
PFS	progression-free survival
PHI	protected health information
RD	residual disease
RNA	ribonucleic acid
RNN	recurrent neural network
SaMD	software as a medical device
SBS	single base substitution
SNV	single-nucleotide variant
TCGA	The Cancer Genome Atlas
TCIA	The Cancer Imaging Atlas
TFN	tensor fusion network
TIL	tumor-infiltrating lymphocyte
TMB	tumor mutational burden
UMAP	uniform manifold approximation and projection
VIF	variance inflation factor
WES	whole-exome sequencing
WGS	whole-genome sequencing
WSI	whole-slide image
WSL	weakly supervised learning

CHAPTER ONE: Introduction

As cancer patients traverse diagnostic, treatment, and monitoring processes, physicians order a suite of diagnostics across distinct modalities to guide management. A significant opportunity thus emerges to aggregate, integrate, and analyze these complementary digital assets across large patient populations to discover multimodal prognostic features, learning from the collective history of large cohorts of patients to inform better management of future patients. For example, genomic profiling of tumor tissue has significantly enhanced clinical decision-making, and the genomic data produced in turn yield a rich molecular repository for further study [1]. This leads to further understanding of the cancer genome, drug sensitivity [2] and resistance mechanisms, [3] and prognostic associations [4,5]. During and after treatment, serial radiologic imaging, such as positron emission tomography (PET) and computerized tomography (CT), quantifies tumor burden in response to intervention, yielding digital archives for large-scale machine learning. Pathology specimens depicting cell morphology, tissue architecture, and tumor-immune interfaces also are increasingly digitized [6]. Other modalities in development, such as cell-free DNA analysis and serial laboratory medicine tests of biochemical and metabolic analytes, provide longitudinal read-outs of tumor progression and recurrence [7–11].

We contend that integrated anatomic, histologic, and molecular measurements approach a comprehensive description of the state of a cancer, resulting in an effective “digital biobank” [12] for each patient. At present however, even when these data are available, they are rarely integrated, and few

advances have been reported that computationally exploit the research discovery potential of large-scale, multi-modal integration. Artificial intelligence (AI) and machine learning (ML) techniques have enormous potential to convert data into a new generation of diagnostic and prognostic models and to drive clinical and biological discovery, but the potential of these techniques often goes unrealized in biomedical contexts, where research-ready datasets are sparse. Cultural and infrastructural changes toward scaled research-ready data archives and development of multimodal ML methods will advance our understanding of the statistical relationships among diagnostic modalities and the contextual relevance of each. Repurposing aggregated, multimodal data—the digital biobanks—therefore presents opportunities to develop next-generation, data-driven biomarkers to advance patient stratification and personalized cancer care.

The central premise of multimodal data integration is that orthogonally derived data complement one another, thereby augmenting information content beyond that of any individual modality. Concretely, modalities with fully mutual information would not yield improved multimodal performance compared to each modality alone. Modalities with fully orthogonal information, conversely, would dramatically improve inference. For example, radiologic scans and pathologic specimens describe tumors spatially at different scales and thus are expected to describe disparate elements of tumor biology. Each modality is incomplete and often noisy, but integrating weak signals across modalities can overcome noise in any one modality and more accurately infer response variables of interest, such as risk of relapse or treatment failure.

To exemplify this premise, we will focus on four major modalities in cancer data: histopathology, radiology, genomics, and clinical information (**Figure 1**). While rapid progress using deep learning (DL) and other ML methods has been made in each of these individual modalities, major unresolved questions about multimodal data integration remain. What are the latent relationships and underlying causal mechanisms at the molecular, cellular, and anatomic scales? Can rational multimodal predictive models enhance clinical outcomes for cancer patients? Can cancer research exploit advances in computational methods and AI models to realize new insights from multimodal data integration? How much data is enough to realize such generalizable predictive models? How can annotations produced during routine clinical care and focused research studies be repurposed to train robust models? How can we fully engage and academically credit both clinicians and data scientists in collaborative studies? How do we establish data infrastructures to enable meaningful and rapid scientific advances while preserving the integrity of patient consent? Herein, we explore these questions through literature review and by developing a blueprint for navigating the infrastructural, methodological, and cultural challenges along the path to achieving robust multimodal data integration in cancer research.

Unimodal machine learning methods to stratify patients

Cancer imaging data have been exploited to predict molecular features of tumors and to discover new prognostic associations with clinical outcomes, and we refer readers to a number of excellent reviews in these areas [13–15]. In

radiology specifically, previous work analyzed features manually extracted by radiologists, such as the VASARI set of imaging features for glioma, and their association with clinical outcomes and molecular biomarkers [16]. However, such features are highly prone to inter-reader variability, and the laborious nature of extraction limits cohort size. As radiology data are digital by construction, automatically extracting deterministic, quantitative features is tractable.[17] These features have been associated with clinical outcomes, such as response to immune checkpoint blockade (ICB) in pan-cancer analyses [18], residual tumor volume after resection in ovarian cancer [19], and progression of disease in pediatric optic pathway glioma [20]. Furthermore, when cohorts are sufficiently large, convolutional neural networks (CNNs), a type of deep neural network (DNN) [**Appendix 2**] have been shown to predict *IDH1* mutational status of glioma from magnetic resonance imaging (MRI), pathologic grade of prostate cancer from MRI, *EGFR* mutational status of lung adenocarcinoma from CT, and BRCA1/2 mutational status of breast cancer from full-field digital mammography [21–25]. Three-dimensional models incorporating axial context have more parameters, and thus require additional controls for overfitting [26], but have stratified NSCLC patients by overall survival (OS) [27] and empirically outperformed two-dimensional models in other radiology tasks, such as diagnosing appendicitis [26]. The relative performance of DL versus conventional ML-based methods on human-defined (“engineered”) features is largely determined by cohort size.

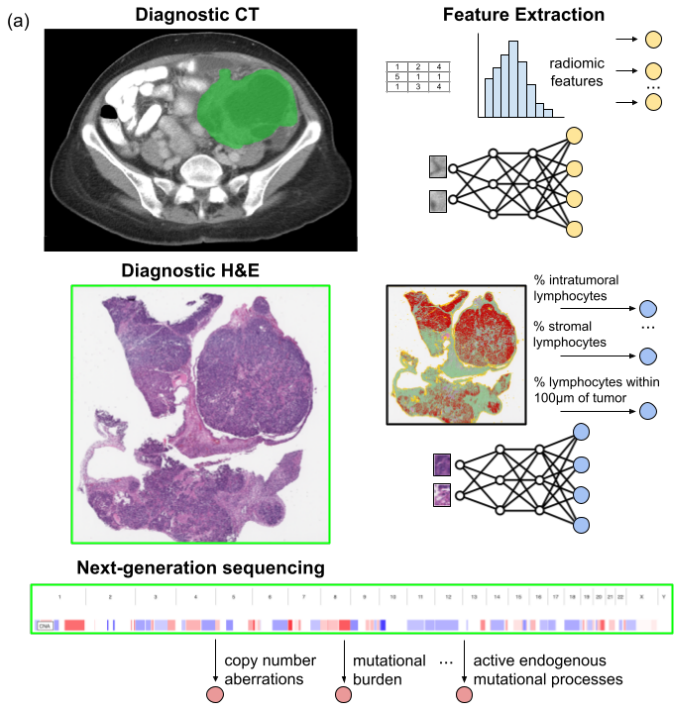
Similar computational models have advanced biomarker inference from histologic imaging, particularly hematoxylin and eosin (H&E)-stained whole slide images (WSIs) [28–32], beyond the previously dominant practice of using pathologist-extracted features [33]. One notable multi-center example in colorectal cancer showed that H&E WSIs contain information predictive of microsatellite instability (MSI) status as a biomarker for response to immune checkpoint blockade [34,35]. However, these DL analyses suffer from poor interpretability and depend heavily on large training cohorts (depending on the task and data complexity, generally thousands of labeled examples for excellent, generalizable performance). Interpretable quantitative analyses of histological images can be conducted using expert-guided cellular and tissue annotations, identifying biological features such as tumor-infiltrating lymphocytes (TILs) and their correlation with molecular features [36]. A recent pan-cancer analysis found that annotation-guided interpretable features predict endogenous mutational processes and features of the tumor microenvironment [37], and other studies have linked biologically interpretable features with clinical outcomes [38,39,40].

Molecular features are the true targets of intervention, either directly or through synthetic lethality, and they are thus the most direct measure for predicting drug response. Examples include mutations in *BRAF* in melanoma [41], *EGFR* in NSCLC [42], ERBB2 (HER2) in breast cancer [43], *IDH1* in acute myeloid leukemia (AML) [44], BRCA1/2 in ovarian [45] and prostate cancer [46] and even rare events such as NTRK fusions [47] for solid tumors, among many others. Targeted cancer therapies are continually being added, for example, ongoing clinical trials

of KRAS (G12C) [48] and [49] and recently PIK3CA [50] in lung and breast cancer, respectively. Higher-order genomic properties such as tumor mutational burden (TMB) [51], endogenous mutational processes such as MSI [52] and homologous recombination deficiency (HRD), and large-scale features such as whole genome duplication [53] are also clinically meaningful. In a recent study, Vöhringer et al. present an algorithm (TensorSignatures) to characterize transcription-associated mutagenesis in seven cancer types [54]. Copy number signatures from low-pass whole genome sequencing [55] and integrated ML models across single nucleotide variant and structural variant scales have also effectively stratified patients into prognostic subgroups [56]. Both studies find that patients with HRD tumors have better prognosis, but further granularity is needed to better resolve clinically meaningful subgroups. Emerging spatial genomics techniques [57,58,59] and complementary clinical and imaging modalities are opportunities to enrich these data and refine prognostication.

Multimodal machine learning methods to stratify patients

We suggest that such unimodal models across radiology, histopathology, molecular, and clinical domains will become the building blocks of integrated multimodal models (**Figure 1**). A major design choice for multimodal approaches is the extent to which each data input should be modeled before encoding joint representations (**Figure 2**). In *early fusion* architectures, features are simply concatenated at the outset and used to train a single model (**Figure 2a**). At the other extreme, *late fusion* architectures model unimodal data fully individually,


(a)
Diagnostic CT
Feature Extraction
radiomic features
Diagnostic H&E
% intratumoral lymphocytes
% stromal lymphocytes
% lymphocytes within 100μm of tumor
Next-generation sequencing
copy number aberrations
mutational burden
active endogenous mutational processes

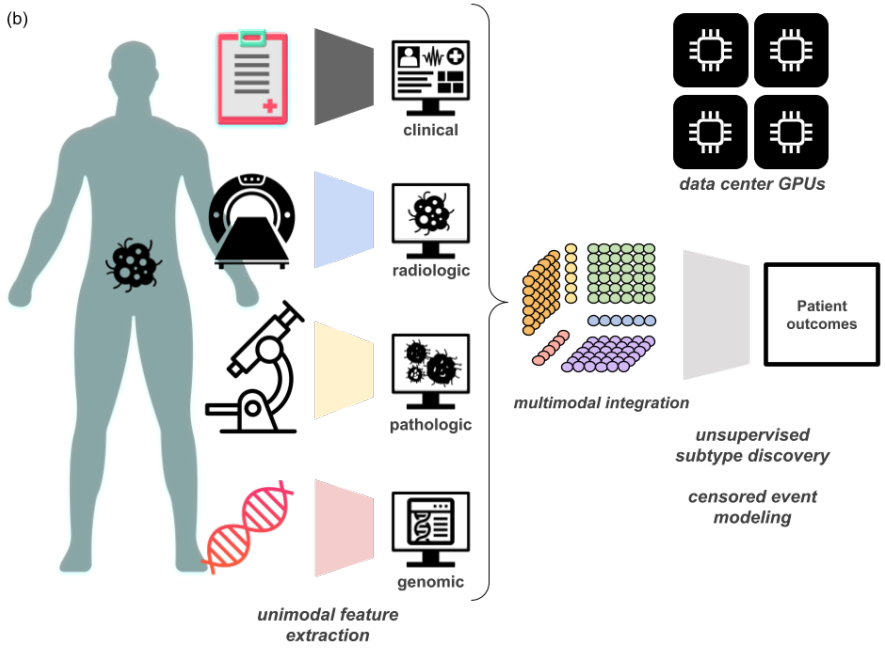

(b)
clinical
radiologic
pathologic
genomic
unimodal feature extraction
data center GPUs
multimodal integration
Patient outcomes
unsupervised subtype discovery
censored event modeling

Figure 1. Multimodal machine learning data modalities and schemata. (**a**) Example data modalities for integration include radiology, histopathology, and genomic information. Image feature extraction involves choosing deep learning or engineered features. (**b**) Sub-models extract unimodal features from each data modality. Next, a multimodal integration step generates intermodal features—a Tensor Fusion Network (TFN) is indicated here [60] . A final sub-model infers patient outcomes.

and then aggregate learned parameters or derived scores (**Figure 2b**).

Intermediate fusion architectures develop a representation of each modality and then model intermodal interactions before joint modeling [61] (**Figure 2c**). Most multimodal architectures have more parameters to be fit than their unimodal counterparts, making them prone to overfitting, which paradoxically can result in worse performance in the supervised setting [62]. One mechanism to address this is incorporating the estimated generalization error in the training objective, using techniques such as gradient blending, to directly account for overfitting [62]. A related design choice is unimodal sub-model complexity. Though overparameterized DL models can outperform traditional ML, their performance is highly dependent on the size of the training dataset. This data size requirement often precludes DL application in biomedical multimodal studies, where missingness of individual data modalities, and requirement of laborious curation of multiple data modalities limits studies to the very small data regime, defined loosely as ~5000 or fewer data points [63]. This makes ML on engineered features an essential approach in the field and suggests that studies with resource constraints requiring very large cohorts, such as those in cancers with high heterogeneity, or those where a single modality overwhelmingly carries the important discriminative features, may opt for a unimodal study.

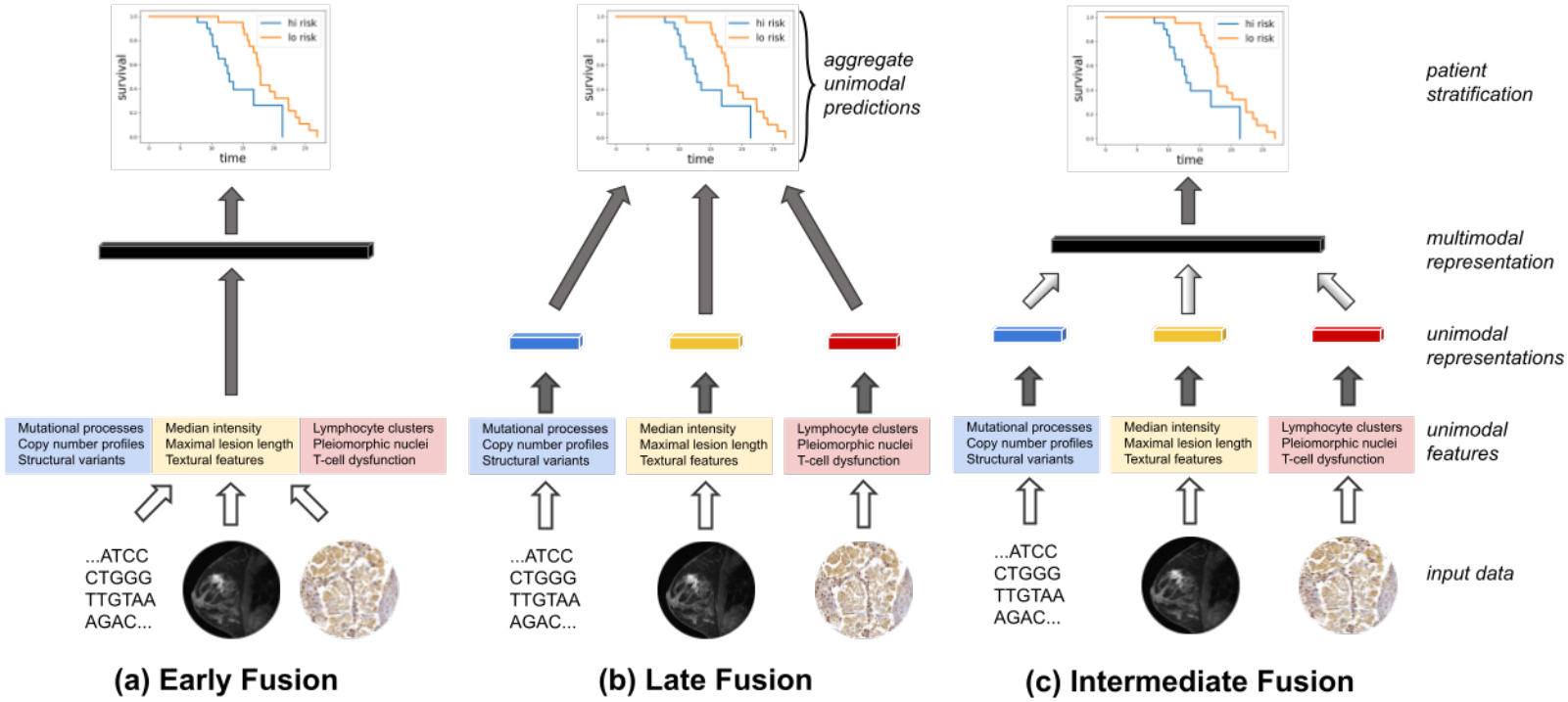

survival
time
hi risk
lo risk
aggregate
unimodal
predictions
patient
stratification
multimodal
representation
unimodal
representations
unimodal
features
input data
Mutational processes
Copy number profiles
Structural variants
Median intensity
Maximal lesion length
Textural features
Lymphocyte clusters
Pleiomorphic nuclei
T-cell dysfunction
...ATCC
CTGGG
TTGTAA
AGAC...
(a) Early Fusion
(b) Late Fusion
(c) Intermediate Fusion

Figure 2. Fusion architectures of multimodal models. Design choices for multimodal models with genomics, radiology, and histopathology data. Filled arrows indicate stages with learnable parameters (linear or otherwise), transparent arrows indicate stages with no learnable parameters, and partially filled arrows indicate the option for learnable parameters, depending on model architecture. (**a**) In early fusion, features from disparate modalities are simply concatenated at the outset. (**b**) In late fusion, each set of unimodal features is separately and fully processed to generate a unimodal score before amalgamation by a classifier or simple arithmetic. (**c**) In intermediate fusion, unimodal features are initially processed separately prior to a fusion step, which may or may not have learnable parameters, and subsequent analysis of the fused representation. All schemata shown are for DL on engineered features: for CNNs directly on images, unimodal features and unimodal representations are synonymous. For linear ML on engineered features, no representations are learned between features and stratification.

Preliminary applications of multimodal machine learning models to stratify cancer patients

Multimodal patient stratification using complementary multi-omics cancer data is well developed [64–69]. The Cancer Genome Atlas (TCGA) catalogues of genomic, transcriptomic, epigenomic and proteomic data enabled integrated, multimodal inference. For example, integrating bulk transcriptomics, miRNA sequencing, and promoter methylation status with early fusion autoencoders showed enhanced ability to stratify hepatocellular carcinoma patients by OS [65]. A similar approach identified distinct survival subtypes in the majority of TCGA cancer types, outperforming existing stratification methods [66]. Joint dimensionality reduction techniques, such as integrative non-negative matrix factorization, learn unsupervised representations of multi-omic profiles for downstream association with outcomes and biomarkers [70]. As experimental and computational techniques advance, these data will more completely characterize

the molecular state of patients' disease [71,72], yet they still only capture a fraction of the informative data.

Several multi-omic models also incorporate traditional clinical features [73,74]. For example, dimensionality reduction, early fusion (**Figure 2**), and a deep Cox Proportional Hazards (CPH) model to integrate multi-omics with age and hormone receptor status stratified breast cancer patients by OS more accurately than unimodal models [74]. Adding additional modalities can paradoxically fail to increase performance, with most clinico-genomic models in the study slightly underperforming the genomic model alone, except when tumor mutational and copy number burdens were integrated [74]. Further work is needed to determine when and why adding modalities is useful. CPH models also are limited by their assumption of linear dependence on each variable and challenges with handling tied samples (when events occur at the same time). Deep binned time survival[75] overcomes these limitations by discretizing follow-up times and predicts risk of NSCLC recurrence from 30 clinical and histopathologic features. Recurrent neural networks (RNNs), a leading method for time series prediction [**Appendix 2**], have not yet been widely applied in oncology but have been shown to accurately predict clinical events from multimodal serological and clinical data [76,77].

Though under-developed relative to clinical and 'omics integration, multimodal models including histopathology imaging features have recently emerged. One such model uses deep highway networks [**Appendix 2**] to integrate H&E images with mRNA-seq and miRNA-seq data to learn the

importance of individual genomic features rather than perform *a priori* dimensionality reduction [78], embedding the individual data modalities in the same, shared information space by minimizing the similarity loss. The model achieves a c-Index of 0.78 to stratify patients by OS[79] and is robust to missingness, but it conceptually encourages *mutual* information, potentially at the expense of *complementary* information gained via fusion methods (**Figure 2**), though this remains to be tested in a head-to-head comparison.

Similarly, Imaging-AMARETTO [80], a framework developed on TCGA glioma data, advances associations between imaging phenotypes and molecular multi-omics, but it does not integrate information explicitly for prognostication. Other examples of multimodal ML studies using histopathology include cellular morphologic features and mRNA-seq data integration in NSCLC [81], combined histologic and gene expression features in breast cancer [82], and genomic survival convolutional neural networks[83] and TFNs [84] in glioma. TFNs are intermediate fusion architectures using the outer product of deep unimodal embeddings [60], which enables the model to learn intermodal dynamics and outperform models based only on grade and molecular subtype (c-Index 0.83 vs 0.78) or any individual modality [84]. It also outperforms simpler multimodal models, such as genomic survival convolutional networks (c-Index 0.83 vs 0.78). In general, these studies demonstrate that multimodal integration with histopathologic imaging improves outcome predictions and stratification over unimodal and molecular methods alone.

Few multimodal models include radiologic imaging. However, a model to diagnose breast cancer using digital mammography and diffusion contrast-enhanced MRI achieved an AUROC of 0.87, higher than the respective unimodal AUROC values of 0.74 and 0.78 [85]. Another study found that the combination of deep features from histologic imaging and engineered features from MRI outperformed unimodal classifiers for stratification of brain tumor subtypes [86]. MRI radiomic features also refine survival stratification beyond IDH1 mutational status and WHO classifications alone, demonstrating the potential of multi-scale information to improve stratification [87]. Multiple kernel learning has been used on small, noisy datasets to integrate clinical factors with MRI- and PET-derived imaging features [88,89]. PET imaging is a particularly promising area for multimodal integration, providing spatial profiles of metabolic activity [90]. Similarly, MRI sequences such as dynamic contrast enhanced images depicting vasculature and diffusion weighted images, whose voxel intensities are influenced by cellularity, provide rich physical profiles with potentially complementary prognostic information.

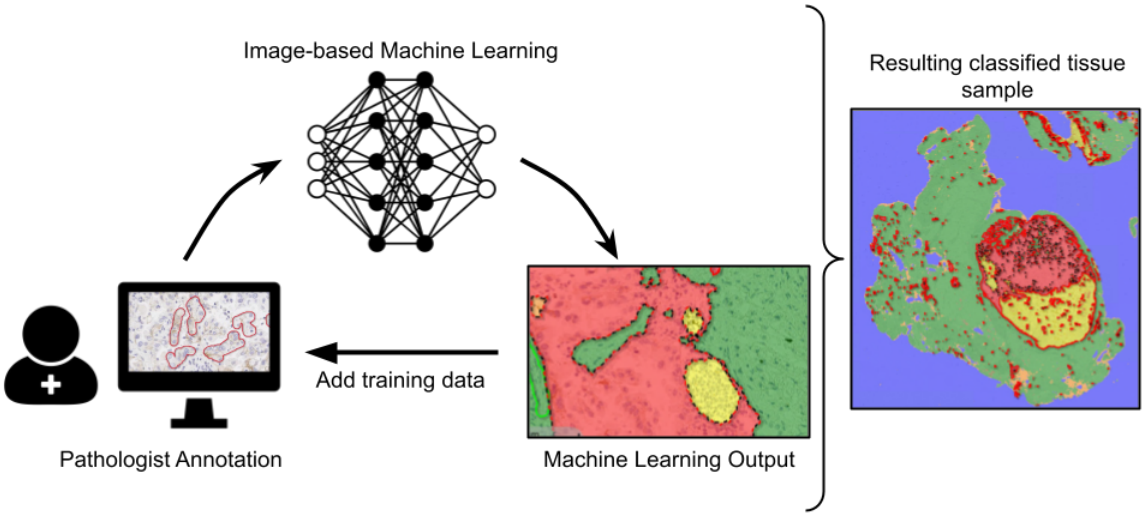


Figure 3. Active learning reduces data annotation burden. In active ("human in the loop") learning, a pathologist first annotates small training areas representing tissue areas (e.g., tumor, stroma, lymphocytes). Next, a machine learning classifier is trained from these expert annotations. Finally, the resulting labeled sample can be examined for misclassified regions, and the pathologist adds targeted additional training areas. This process is repeated until the classification is accurate and can be applied to multiple samples.

Promising methodologic frontiers for multimodal integration

Multimodal ML in the medical setting is most limited by the tension between data availability and amount of data needed to fit multimodal models. Hence, many methodologic frontiers involve increasing robustness to overfitting and dealing rationally with missingness. For example, transfer learning in unimodal models involves pre-training a model on a large, tangentially related dataset and then fine-tuned on the actual dataset of interest, which is typically small. Some example datasets used are ImageNet [91], a database of more than 14 million labeled images used to train image classification algorithms for two-dimensional CNNs, and Kinetics, a curated collection of approximately 650,000 YouTube videos depicting human actions for three-dimensional CNNs [92].

However, recent evidence shows that small models without pretraining can perform comparably to pre-trained large models, such as ResNet-50, for small medical imaging datasets [63]. This is consistent with the hypothesis that the benefits of pretraining for small medical imaging datasets are related to low-level feature reuse and feature-independent weight scaling [63]. It remains an open question whether pretraining multimodal fusion models can combat overfitting through similar weight scaling of the parameters involved in fusing unimodal representations. Both prospective clinical trials and highly curated retrospective cohorts often have low numbers of patients, highlighting the importance of studying how to use DL techniques appropriately to discover patient strata in the very small data regime.

One of the root causes of data scarcity is the need for extensive annotation: tumors need to be localized on CT scans or H&E images, and survival outcomes typically require manual review of medical records. Harnessing data at scale requires reducing this burden of annotation, especially in multimodal studies. Automated annotation approaches could provide solutions. For example, RetinaNet, an object detection CNN, has been used to localize lung nodules on CT, enabling use of 42,290 CT cases for training [93]. Analogously, an ML-based model to automatically delineate representative tumor tissue from colorectal carcinoma histology slides enabled training on 6,406 specimens [35].

Weakly supervised learning (WSL) also helps reduce the burden of annotation by using informative-yet-imperfect labels for the training dataset. While weak labels may be incomplete, inexact, or inaccurate [94], WSL

applications in computational pathology have resulted in robust models to infer genomic alterations [31] and diagnose cancer on WSIs [95]. Weaknesses of this approach include the absence of a ground-truth dataset for model evaluation when all labels are inexact or inaccurate and its dependence on large dataset sizes.

Active learning solicits strong labels for targeted instances, selected using either informativeness or representativeness of an instance [94]. For example, it can be used to prioritize expert annotations in real time for pathology tissue-type labeling (**Figure 3**). These strategies are essential in clinical contexts, where most data elements possess only weak labels, and are a leading strategy to learn robust models from large, information-poor datasets. Therefore, WSL is a useful strategy to augment annotations, dramatically increasing the size and robustness of usable multimodal datasets for clinical oncology.

As more such datasets become annotated and integrated, oncology will benefit from multimodal recommender systems, analogous to inferring cancer drug response based on unimodal gene expression data [96]. Retrospective observational studies contain no matched controls, which biases training data and requires methods such as counterfactual ML to learn accurate recommendation policies from logged interventions and resultant outcomes [97]. In oncology, a counterfactual recommender system (**Figure 4**) would learn policies to recommend future therapies for new patients based on historical patient records of administered treatments, patient contexts (e.g. a pre-treatment CT scan and H&E-stained biopsy), and survival outcomes [97]. In general, this is not

currently possible because patient data are not accessible and annotated at the scale required, but such methods have great potential as datasets are assembled and prospective data collection methods improve.

Finally, unsupervised learning [**Appendix 2**] continues to develop in general, with potential to both facilitate discovery of new cancer phenotypes and probe multimodal associations. For example, deep probabilistic canonical correlation analysis jointly learns parameters for two DNNs and a transformation to embed them in the same information space, all with Bayesian inference suitable for small datasets [98]. This method is especially well suited for probing the *mutual* information to generate hypotheses for experimental biology, such as genomic drivers of cellular morphological heterogeneity. At the patient level, an unsupervised Bayesian topic model has been applied to learn multimodal topics that stratify patients by risk of mortality [99] and in deriving mutational process activities in genomic datasets [56]. Surprisingly, progress in this area demonstrates statistical power across feature spaces from data measuring signals at vastly disparate scales (e.g., histologic-genomic, or radiomic-molecular). We therefore anticipate that generative methods have potential to discover new phenotypes and to generate hypotheses to guide experimental biology.

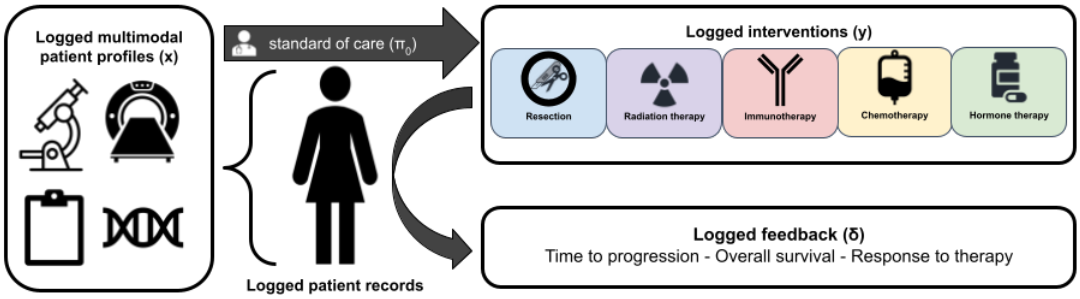


Figure 4. Multimodal recommender systems applied to clinical oncology. Logged healthcare data comprises multimodal patient contexts *x*, interventions *y* based on the standard of care, and feedback δ based on the outcome of the intervention. Learning from such data is challenging because of the lack of two-arm design and the biased data based on the changing standard of care. Counterfactual recommender systems learn theoretically guaranteed unbiased policies from these data. Then, the validated policy *π* can be applied prospectively to support physicians' management decisions.

Challenges in multimodal data integration and analysis

The challenges in multimodal integration of clinical cancer data fall into three broad categories: data engineering and curation, ML methods, and data access and governance provisions. These challenges extend to both retrospective studies seeking to discover biomarkers from standard-of-care data and prospective studies focused on bespoke or advanced data types. The field at large also shares some challenges with unimodal ML studies in medicine, such as interpreting results and ensuring their reproducibility. Here, we describe some of these challenges along with potential solutions to address them.

Data silos

Medical data storage systems are designed for computational research, especially multimodal research. Archived data reside in disparate data silos, often in formats unsuited for computing, and lack well-structured metadata. The limitations of conducting research using data from the electronic medical record, for example, are widely discussed [100]. Chart review, or manually reviewing patient records to extract specific features into spreadsheets, is error prone and variable, and repeated review is often required to capture new clinical events [101,102]. Efforts are underway to add structured ontologies and computationally interpret the healthcare record, but a true reform of the healthcare record structure is required to make data accessible for research [103]. Other data modalities face similar challenges: radiologic images are typically stored in the picture archiving and communication system (PACS) with limited clinical annotation. Similarly, stained tissue specimens are not typically digitized and must be located manually and scanned. Furthermore, any available tissue imaging is often stored by researchers on their own file systems or in digital pathology file systems with limited metadata. Integrating data for multimodal studies results in disjoint patient identifier spaces, complicating alignment and analysis.

The data lake approach, which organizes original data and tracks their use during subsequent analysis, is an appealing solution for integrating and presenting data in a research-ready format [104]. Data lake technologies such as Delta Lake present a scalable solution to bring together siloed data into a single

home for structured, semi-structured, and unstructured data, accommodating both known and unforeseen file types. Ideally, these data for research use should be stripped of protected health information (PHI) to protect patient privacy and facilitate inter-institutional sharing. However, data lakes alone do not address the lack of structured data in the medical record. Some existing solutions involve adding research-friendly structure to certain clinical notes, hiring full-time data curators to extract data *post hoc*, or building models to codify information automatically from the archive. An additional complication in multimodal research is relating extracted data points spatially, which is especially important in studies of biological correlation among modalities. For example, the logistical challenges of colocalizing transcriptomics or H&E imaging from matched specimens limit studies of intratumoral heterogeneity and clonal evolution.

Another challenge is co-registering tissue specimens with corresponding lesions on radiology: image-guided biopsies or 3D-printed molds based on tumor morphology [105,106] are possible solutions, but scaling these approaches for prospective research is far from realized. All these solutions are ultimately insufficient stopgap measures, and presenting data generated during the course of care in truly research-ready forms with mappings across modalities remains elusive. Data infrastructures to automatically detect and annotate lesions, track them over time in the radiologic workflow and then integrate them in other clinical workflows are aspirational solutions.

To promote benchmarking and collaboration among institutions, standard multi-institutional data sharing models are required. Platforms such as the

database of Genotypes and Phenotypes (dbGaP), the European Genome-phenome Archive (EGA), The Cancer Imaging Archive (TCIA), the Genomic Data Commons (GDC), and other resources in the NCI Cancer Research Data Commons have been indispensable for inter-institutional benchmarking and reproducibility. MIMIC-III is another example, providing critical care data for more than 50,000 admissions in a research-ready format [107]. However, beyond matched genomic data and H&E WSIs of TCGA and METABRIC, public resources contain only small patient cohorts with multiple data modalities.

Observational Health Data Sciences and Informatics (OHDSI), a common data model supporting observational studies and integrating controlled vocabularies to standardize data infrastructure [108] will help to enable cross-institutional resources. The American Association for Cancer Research project Genomics Evidence Neoplasia Information Exchange (AACR project GENIE) is another model to integrate inter-institutional sets of matched genomic sequencing and clinical outcomes toward advancing precision oncology: further efforts are needed to extend such architectures to incorporate additional data modalities [1]. However, given the logistical challenges of anonymizing data, such as DICOM headers for radiology containing PHI, and institutional policies hindering data sharing, federated learning is a potential solution [109]. Federated learning shares the model to be trained among institutions rather than centrally amalgamating multi-institutional data [110]. Depending on the choice of model, federated learning can require novel training approaches [111] but enables training on multi-institutional cohorts without data having to leave local networks.

Data integration and analysis

As integrated datasets mature, challenges will shift to data analysis. Complete data on all patients of a study of interest is rare, and this missingness complicates multimodal data integration. Most traditional multivariate models, such as Cox Models, cannot handle this directly and thus require either exclusion of patients without all data modalities or overly simplistic interpolation (e.g., by median). Both of these strategies fail to harness all available data to train effective models. To circumvent this, one simple solution is to use late fusion (**Figure 2b**), where each unimodal model can be trained separately to infer the outcome of interest, which can then be integrated. Bayesian approaches [112] also offer analytical solutions for missingness.

Data modeling will also be complicated by institution-specific biases in the data, such as staining and scanning particularities in histopathology [113–115], scanner parameters in MRI, and differing ontologies in clinical data. Preprocessing techniques in MRI [116] and H&E [117,118] address this heterogeneity, and with large cohorts, DL is somewhat robust to noise [28,119], but such heterogeneity is a major reason that AI systems fail when trialed in the clinic [120]. An additional complexity in multimodal studies is that unimodal biases are likely to be correlated. For example, biasing factors such as MRI manufacturers and H&E staining artifacts likely differ more between institutions than within an institution. This will make it more challenging to model general intermodal relationships, motivating greater cross-institutional data representation and potentially motivating methods that explicitly model these multimodal biases,

and/or normalize against them. Different modalities with different levels of heterogeneity may require different training dataset sizes—in this case, training the overall model may involve pre-training the unimodal sub-model using the enlarged unimodal cohort.

Another analytical challenge is overfitting. Multimodal ML is more prone to overfitting because, in most cases, multimodal datasets are smaller and multimodal models have more parameters to fit. Traditional ML models enable investigators to calculate the necessary dataset size for a tolerable generalization error before analysis. Black box models such as DNNs do not offer such analytical forms. Instead, target dataset size is decided empirically by comparing performance when the model is trained on different proportions of the full data set [35,95]. Some evidence suggests that early fusion strategies can perform comparably to unimodal results using less training data [121], but in general, highly parameterized fusion models are likely to require more training data to fit the additional parameters.

Hence, in many settings, multimodal approaches cannot yet fully harness the performance benefits of deep learning. The most important response to this is to advance clinical data collection to assemble large datasets and better support methods development and benchmarking (see Data Silos). Meanwhile, smaller datasets curated at single institutions require less complex models to avoid spurious results due to overfitting. Each unimodal model can thus be formulated using ML on engineered features, such as radiomic features from MRI and nuclear morphology features from H&E. One major drawback is the need for

laborious annotation, such as segmentation on MRI and tissue type delineation on H&E, which can be reduced using weakly supervised and active learning (see Promising methodologic frontiers). For all model types, cross-validation and external testing cohorts are critical to demonstrate generalizability.

Infrastructurally, multimodal analytic workflows present hardware and software challenges. Centralized data lakes and workflow management tools minimize duplicated computation, such as image pre-processing, among multiple investigators' workflows. Computational needs also differ during different parts of the workflow, with a much higher demand during model training than during cohort curation. This is especially true for multimodal models such as TFNs, which generate intermodal representations that scale exponentially with the number of data modalities. Elastic cloud computing resources and the distributed data parallelism of modern DL-based frameworks handle these computational bursts appropriately, but the use of off-premises cloud computing requires robust de-identification of patient data, data security certifications, and measures to control data ingestion and egress costs.

Reproducibility

Reproducibility and benchmarking are major challenges in AI, with many published biomedical AI studies failing to provide source code, test data, or both [122]. Several recent seminal works do not provide source code, claiming that internal code dependencies prevent code sharing and that textual descriptions are sufficient to reproduce the results [93,123,124]. However, a recent investigation of

one of these studies[124] found that significant information needed to actually reproduce the study was missing, greatly reducing the impact and ability of the field at large to scrutinize [125] and improve upon it. To foster transparency, scientific reproducibility, and measurable progress, investigators should be encouraged to deposit new multimodal architectures and preprocessing regimens in standardized repositories such as modelhub.ai [126]. Furthermore, to promote benchmarking and multicenter validation, journals should require investigators to make available published deidentified datasets on platforms such as the dbGaP, EGA, GDC, and TCIA. Beyond center-specific confounders, the clinical environment has unpredictable effects on model performance, often leading to substantial performance decrements [127].

Hence, prospective clinical validation is the most relevant measure of a model's performance [128]. This is because directly comparing clinical outcomes with and without the AI system, where both arms are exposed to the inherent noise such as varying image quality and user error, provides an objective, quantitative assessment of a model's value. SPIRIT-AI and CONSORT-AI are consensus guidelines for clinical trial protocols and reports, respectively, that extend the SPIRIT and CONSORT guidelines for randomized clinical trials [128–130]. In broad terms, these guidelines improve reporting transparency and ensure that readers can evaluate practical factors that may impact AI system performance in clinical contexts, such as required training, error handling, and output data format.

Balancing the need for interpretability with empiric efficacy

The nature of DL architectures creates a limiting paradox. While often outperforming standard, interpretable models, users are left to explain improved results without the benefit of drawing from model assumptions encoded in more traditional approaches such as hierarchical Bayes. We argue that investigators should seek to understand learned models from biological and clinical perspectives in order to realize rational multi-modal implementation. Depending on the goals of a study, understanding a model is arguably as important as improving its predictive capacity and will lead to greater mechanistic insight and testable hypotheses. For example, post-hoc explanation methods, which seek to interpret model predictions in terms of input feature values, have been applied to probe medical algorithms [131]. However, post-hoc explanations are prone to misinterpretation and cannot supplant true interpretability [132] to elucidate a mechanism or generate hypotheses for experimental biology.

Yet when the main purpose of an algorithm is to improve patient outcomes, understanding models mechanistically at the expense of denying patients empirically improved quality of life is unethical. Many empirically beneficial medical interventions, such as general anesthesia, have incompletely understood mechanisms [133]. Hence, the most important threshold for using these models in the clinic is the same as for a drug: robust, prospective, multi-center empiric evidence of benefit for patients and an understanding of cases in which the model fails. Given our limited understanding of black-box models, pilot

studies must demonstrate that the model is effective and equitable for all patient subpopulations it will encounter before deployment at scale [134].

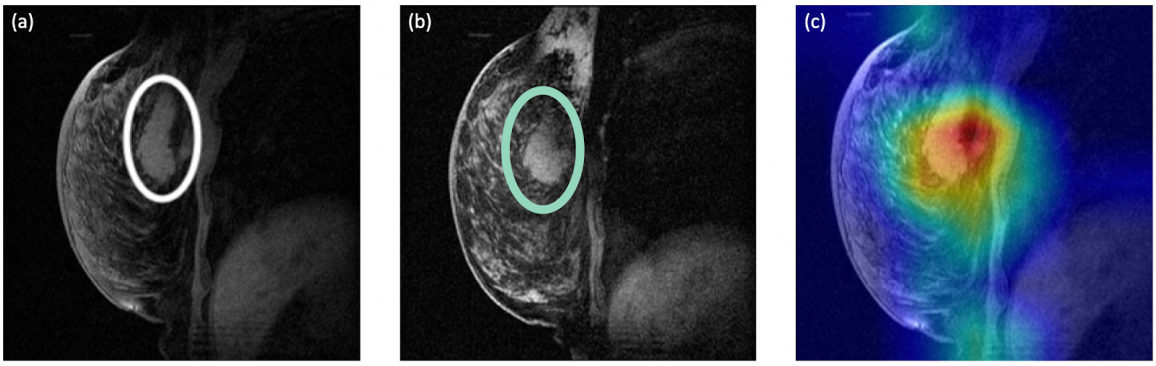


Figure 5. Class activation maps provide some interpretability. Breast MRI images (**a**) before neoadjuvant chemotherapy with region of interest circled by a radiologist, (**b**) after neoadjuvant chemotherapy with circled region of interest by a radiologist, and (**c**) before neoadjuvant chemotherapy with class activation mapping by a neural network trained to predict response to therapy. Warmer colors indicate higher saliency. Images from [135,136]

Truly causal models are a frontier of AI research, and in the future such models will be highly valuable in this field [137]. Less challenging than interpretability, explicability is also useful for black-box models. For example, class activation maps (CAM) [138] (**Figure 5**) depicts which parts of the image are most important for the model to arrive at its decision. The saliency, or dependence of the output on a specific region, is shown for prediction of response to chemotherapy in **Figure 5c**. This technique is limited by seeking explicability rather than interpretability [132], but it can be useful to rule out obviously spurious determinants of model output. For example, if the CAM in **Figure 5c** showed highest saliency in the area outside the breast, it would raise serious concern about the validity of the model. Lucid is another method for

explicability which uses the learned model to generate example images for each class [139]. For example, it has been applied to visualize what a CNN is looking for in breast H&E images to distinguish tumor from benign tissue [140]. For DNNs with definable input variables, layer-wise relevance propagation is widely used and has been applied to clinical data [131]. However, these methods were developed for unimodal ML, and interpreting multimodal ML is more challenging.

Future work must quantify the relative contribution of each modality and their interactions. Uninformative feature counterfactuals also have been used to probe feature importance with guaranteed false discovery rates [141], and such a method might similarly quantify the performance of a modality in a late fusion architecture, for example. Yet feature importance is only an early step toward interpretability: probing a model with potentially informative data counterfactuals (e.g. “How would the inferred genomic subtype change if the tumor texture were more coarsely heterogeneous on CT?”) would further our understanding of black-box multimodal models [137,141].

Data governance and data stewardship

Progress will require appropriate data governance and stewardship. Patient consent lies at the core of appropriate use of data and dictates terms of use as stipulated by institutional review boards. Beyond patient consent, high quality, curated and annotated datasets require the expertise and domain knowledge of clinician scientists or clinical fellows. As such, terms of use for these valuable datasets are likely to be set by those who invested the expertise and time required for curation. Success will therefore depend heavily

on collaborative models coupling expert clinical annotations with the expertise of data scientists for advanced analyses [142]. Furthermore, cross-departmental coordination, access provisions, and governance structures will be required to achieve large scale multimodal data integration.

We argue that open data models are the most productive approaches to fully leverage data for discovery and promote reproducibility. This has been demonstrated in the cancer genomics community with TCGA, and community data standards promoting multi-institutional clinical data integration, such as AACR Project GENIE, are now gaining traction [143]. Moreover, as the clinical journey for a cancer patient plays out over time, technology systems and governance structures to capture relevant events and new data in real-time will enhance efforts for data integration and computational discovery. Effective stewardship plans, including accuracy of data, collaborative access provisions, imposition of data standards, and longitudinal data updates, are therefore critical to managing and deploying appropriate use of data for large scale multimodal data integration.

Regulatory challenges

During model development, the main legal challenges surround protection of patient privacy by limiting access to protected health information (PHI). Per HIPAA, researchers must have the minimum amount of access necessary to conduct the research, which often is greater for data-driven projects than for hypothesis-driven projects given the unspecified amount and types of data required to discover new patterns. Deidentifying data promotes compliance with

the minimum access stipulation of HIPAA. This also enables multi-institutional collaboration, wherein data must be deidentified prior to transmission. However, this process presents a data engineering challenge. In radiology, DICOM image headers contain PHI that can be removed electronically, but the process often includes a manual validation step for liability reasons. This is infeasible at the large scales required for modern machine learning, and medical centers will need to invest in developing and validating more reliable automatic anonymization software. In pathology, slides can include disparate markings or labels containing PHI, which can be solved with conservatively calibrated automatic exclusion tools, then followed by manual review. For the healthcare record, schemata such as REDCap already exist to present anonymized data [144]. For analysis, if cloud computing is used for data containing PHI, the Health Insurance Portability and Accountability Act of 1996 requires that the cloud computing providers enter into both business associate and service level agreements with the medical center [145]. Stripping training data of PHI protects patient privacy, thereby reducing these regulatory burdens during model training and validation.

After the model is trained and robustly validated, investigators may look to deploy the model for clinical use. The current FDA regulations for software as a medical device (SaMD) are insufficient for AI models in precision oncology, which inherently learn and adapt after initial deployment. In a 2020 workshop, the FDA proposed amending SaMD regulations to account for this inherent strength while managing its associated risk [146]. The key elements of this proposed total product lifecycle approach are an algorithm change protocol and SaMD pre-specifications

(SPS). The algorithm change protocol would delineate plans for managing data, iteratively training the model, evaluating performance, and deploying updates, and the SPS would draw a “region of potential changes” around the original specifications and uses of the algorithm [146]. Further precautions to reduce risk would include a pre-review of organizational software quality and culture. Updates such as these would control risks while enabling AI algorithms to iteratively improve after deployment.

After deployment to clinical care in the medical center, another concern arises: legal culpability. Current liability in medicine is established by tort doctrines, wherein civilians claim compensation from physicians, healthcare organizations, or manufacturers [147]. AI models will likely initially be used in concert with other clinical tests, and the final decision on treatment will continue to rest with physicians. Hence, physician liability is currently established based on the standard of care, an AI model’s recommendation for or against the standard of care, and a physician’s adherence to or rejection of the recommendation [148]. Interpretable models are thus highly advantageous in the current regulatory framework. However, it is not clear that this paradigm focused on physician gatekeeping will be sufficient as algorithms that demonstrably outperform clinicians advocate for an intervention that subsequently causes harm. Yet in these cases of harm, patients must have mechanisms for recompense. Potential directions for legal reform include treating the AI model itself as a culpable independent entity to be insured similar to the no-fault compensation for vaccine injuries [148], or a common enterprise liability, wherein

the model itself is viewed as faultless and all developers and non-patient users of the algorithm jointly share liability [147].

Perspectives

Multimodal cancer biomarker discovery occurs at the interface of clinical oncology, ML research, and data engineering, which typically operate separately. To advance the field, collaborative research programs must unify and promote clear communication among these stakeholders through platform design, model development, and the publication lifecycle [149]. These programs will enable clinical investigators to ask questions centering on patient stratification and ultimately produce predictive models by integrating multimodal data. A team science approach with appropriately shared attribution of credit and agreed-upon data stewardship provisions is essential for progress.

The main roadblock to progress in this field is the lack of usable data. Advances in multimodal ML methods have been impressive in other fields, such as sentiment analysis [60,150–155] , with large benchmark datasets, but the largest multimodal oncologic dataset, the TCGA, contains limited data modalities and only a few hundred patients per cancer type. This data scarcity largely prevents investigators from using advanced data-hungry models and, critically, hampers benchmarking of new methods in the field, required for rational development of multi-model biomarkers.

Institutional datasets must be assembled and shared, but current data infrastructure typically necessitates months of laborious extraction and

annotation before analysis begins. This is perhaps the most well-known issue of conducting ML research for healthcare applications, and a general solution is not imminent. To address this in specific cases, imposed structures on certain notes and full-time data curators have hastened chart review. Automatic annotation strategies relying on weakly supervised and active learning conserve scarce expert annotations and have begun to reduce annotation burdens for large imaging cohorts.

Until these fundamental challenges are addressed, multimodal ML models must often operate in the very small data regime. Simple ML models should be used in place of DL methods for small cohorts. DL models should be used judiciously for tasks with large statistical sample size and with strategies to combat overfitting, such as gradient blending, early stopping, data augmentation, and weight decay. Investigators must be wary of spurious results due to institutional biases and small sample sizes, with cross-validation, retrospective external validation, prospective validation, and clinical trials serving as key measures to assess algorithm effectiveness.

Ultimately, as biomedical data infrastructures develop, the goal of this line of inquiry is to refine cancer prognosis and rational management by integrating multiple data modalities. Genomic biomarkers have improved upon traditional staging and have begun to implement personalized cancer care, promoting targeted therapies. We predict new classes of multimodal biomarkers will further harness information content from various sources, thereby leading to improved predictive models for therapeutic response. Validated models will be deployed to

the electronic medical record, providing near-real-time risk stratification and recommendations for individual patients for clinicians to integrate with other factors to inform management.

While we focused on genomics, histology, radiomic and clinical outcomes in our discussion, we expect additional measurements such as microbiome, metabolic analytes, longitudinal cell free DNA analysis, and deep immune profiling will become integrated as informative determinants of clinical trajectories. In summary, we project that as data access challenges are overcome, multimodal computational techniques will play important roles in clinicians' decisions around disease management. Developing multimodal ML methods, usefully logging and annotating patient data, and advancing data engineering infrastructures are outstanding hurdles that remain in the field. As these challenges are met, the field is poised for a reimagined class of rational, multimodal biomarkers and predictive tools to refine evidence-based cancer care and precision oncology.

Risk stratification in high-grade serous ovarian cancer

High-grade serous ovarian cancer (HGSOC) is the most common cause of death from gynecologic malignancies, with a five-year survival rate of less than 30% for metastatic disease[156]. Initial clinical management relies on either primary debulking surgery (PDS) or neoadjuvant chemotherapy followed by delayed primary surgery (NACT-DPS). Endogenous mutational processes are an established determinant of clinical course, with improved response of

homologous recombination deficient (HRD) disease to platinum-based chemotherapy and poly-ADP ribose polymerase (PARP) inhibitors [157–159]. More nuanced genomic analyses integrating point mutation and structural variation patterns further refine this stratification into four biologically and prognostically meaningful subtypes [55,160] including distinct sub-groups of HRD, foldback inversion enriched tumors and those with distinctive accrual of large tandem duplications.

Beyond genomic factors, clinical indicators such as patient age, pathologic stage, and residual disease (RD) status after debulking are also prognostic [161]. However, these clinico-genomic factors alone fail to adequately account for the heterogeneity of outcomes. Identifying patients at risk of poor response to standard chemotherapy remains a critical unmet need, and improved risk stratification models to identify such patients will aid medical oncologists in planning monitoring frequency and administration of maintenance therapy and may help patients in considering clinical trials of investigative agents.

Beyond clinico-genomic features, multi-scale clinical imaging is routinely acquired during the course of care, including contrast-enhanced computed tomography (CE-CT) at the mesoscopic scale and hematoxylin and eosin (H&E)-stained slides of diagnostic biopsies at the microscopic scale. Digital forms of these diagnostics present opportunities to develop computationally driven quantitative models for improved risk-stratification. Accordingly, we tested whether integrated multi-modal data from quantitative imaging, genomics and clinical data could improve identification of risk groups for HGSOC.

At the mesoscopic scale, recent radiologic studies have uncovered quantitative CE-CT features that are predictive of early progression, time to recurrence, and overall survival in HGSOC [19,162,163]. Most studies to date have analyzed the prognostic information captured within adnexal lesions [162,164,165] or the whole burden of disease [166–168]. Deep learning-based radiologic models [165] have the potential for higher performance than engineered features, but given that they are even more challenging to interpret and prone to overfitting in the typical clinical data regime of hundreds of examples, radiomic features from the Imaging Biomarker Standardization Initiative [17] . Furthermore, we opted to develop and validate a radiomic prognostic model based on omental lesions because, in comparison to adnexal lesions, omental implants are easier to delineate and are very common in advanced stage HGSOC. They also constitute only a portion of the total burden of disease and thus require less time and experience to outline than the whole tumor volume.

At the microscopic scale, H&E-stained tissue biopsies enable pathologic diagnosis and are routinely acquired before the start of therapy. A quantitative histopathologic study of HGSOC identified patterns of immune infiltration on H&E slides that correlate with mutational subtypes [160]. In other cancer types, studies of whole slide images (WSIs) have advanced our ability to quantify the histopathologic architecture of tumors using deep [30,123] and interpretable [169] features. Apart from stage, HGSOC lacks independent pre-treatment pathologic factors by which to stratify patients [161], and as such quantitative approaches present an opportunity to systematically search for them at scale beyond

qualitative human interpretation. Interpretable features are less prone to overfitting in small cohorts and can be more easily interrogated by human pathologists [132,169], and thus, we opt for this strategy in our analysis.

Multimodal machine learning models integrating information from distinct measurements of a system have empirically outperformed unimodal approaches in fields as diverse as sentiment analysis [60,62] and multi-omic molecular models to stratify cancer patients [73,74]. Initial results demonstrate that multimodal models that include clinical imaging can outperform their unimodal constituents. For example, a tensor fusion network integrating H&E-based features with features derived from RNA-seq and whole-exome sequencing stratifies glioma patients better than models based on genomics or histology alone [84]. In the pan-cancer context, jointly embedding histopathologic imaging representations with clinical covariates and features from miRNA and mRNA-seq improves stratification beyond these molecular characterizations [79]. Conceptually, disaggregated genomic sequencing omits spatial context, and we thus hypothesize that multiscale imaging contains complementary prognostic information, rather than merely recapitulating genomic prognostic information.

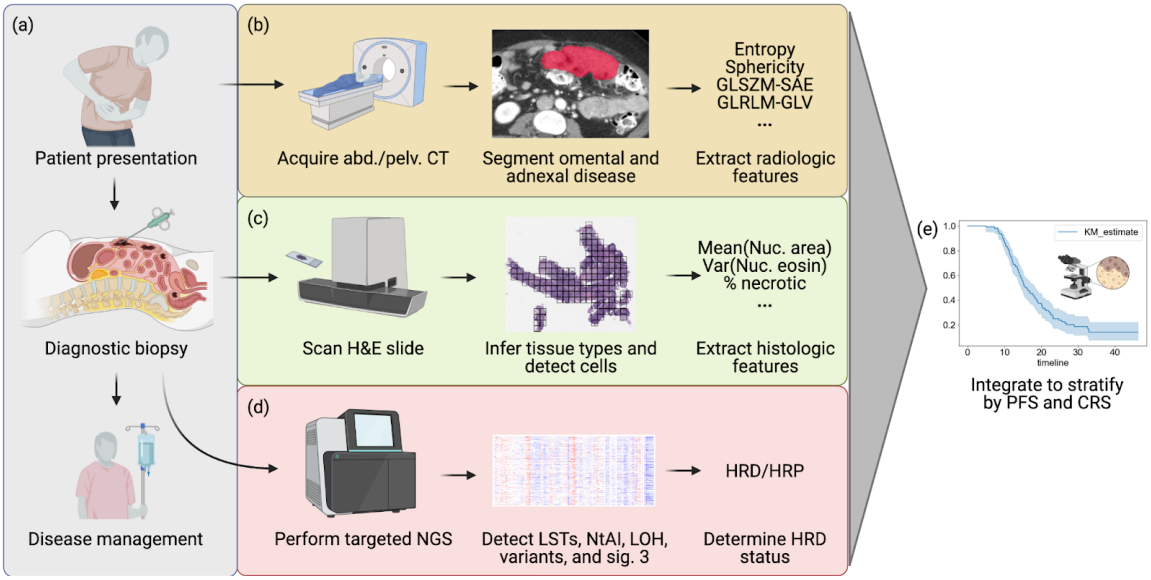


Figure 6. Schematic outline of the study. (**a**) As patients proceed through diagnosis and therapy, treating clinicians acquire multiple data modalities to inform management. This study generates prognostic models based on (**b**) pre-treatment abdominal/pelvic contrast-enhanced CT studies, (**c**) pre-treatment H&E-stained diagnostic biopsies, and (**d**) HRD status inferred from hybridization-capture based targeted sequencing or clinical HRD-DDR gene panels. (**e**) We integrated these unimodal features to stratify patients by progression-free survival and compare with pathologic chemotherapy response score. Please refer to list of abbreviations as needed.

Objectives of the thesis

In this work, we set out to study the complementary prognostic information of multimodal features derived from clinical, genomic, histopathologic, and radiologic data obtained during the routine diagnostic workup of HGSOC patients (**Figure 6a**). We developed a radiomic model based on CE-CT-derived quantitative omental features (**Figure 6b**) and a histopathologic model based on pre-treatment biopsies to risk stratify patients (**Figure 6c**). The models were validated on an internal test cohort and an external TCGA test cohort and were integrated with clinical and genomic information (**Figure 6d)** using a late fusion

multimodal statistical framework (**Figure 6e**). Our results revealed the empirical advantages of cross-modality integration and demonstrate the ability of multimodal machine learning models to improve risk-stratification of HGSOC patients.

CHAPTER TWO: Multimodal machine learning strengthens inference of high-grade serous ovarian cancer patients

Summary

Patients with high-grade serous ovarian cancer demonstrate poor prognosis and variable response to treatment. Homologous recombination deficiency status, patient age, pathologic stage, and residual disease status after cytoreductive surgery are important prognostic factors, with recent work also highlighting prognostic information captured in computed tomography and histopathologic specimens. However, little is known about the capacity of combined features to discriminate between patients and explain clinical outcomes. In this work, we developed and integrated histopathologic, radiologic, and clinico-genomic machine learning models to determine their combined impact on risk stratification. We assembled a multimodal dataset of 409 high-grade serous ovarian cancer patients and showed that human-interpretable features, such as necrosis on H&E and omental textural complexity on CT, are associated with worse prognosis. We then integrated these models and demonstrated that the imaging models contain complementary—rather than purely mutual—prognostic information to clinico-genomic prognostic factors, as evidenced by improved risk stratification and stronger association with pathological chemotherapy response score than unimodal models. This work empirically supports multimodal machine learning approaches as a promising path toward improved risk stratification of cancer patients.

Cohort characteristics

We analyzed 409 patients, 262 HGSOC internal cases at MSKCC for discovery (222 for training, 40 for internal testing) and 147 TCGA cases for external testing (**Figure 7a**). The 40 internal test cases were randomly sampled from the discovery cohort before analysis. The entire discovery cohort contained 143 Stage IV, 115 Stage III, 3 Stage II, and 1 Stage I patients, while the external test cohort contained 31 Stage IV, 103 Stage III, 7 Stage II, and 6 Stage I patients (**Figure 7b**) [170]. Median age at diagnosis was 65 years [IQR 58-72] for discovery and 60 years [IQR 51-68] for external test sets (**Figure 7c**). In the discovery cohort, 188 patients received neoadjuvant chemotherapy followed by delayed primary surgery, and the remaining 74 underwent primary debulking surgery. Treatment regimens are not annotated for TCGA patients. Median PFS was 15.9 months [IQR 12-22] for discovery patients and 14.9 months [IQR 9-25] for TCGA testing patients. 86 discovery patients and 37 testing patients had censored PFS outcomes (**Table 1**).

The discovery cohort was composed of H&E WSIs from 142 patients, combined with adnexal and omental lesions on CT from 192 and 216 patients (**Figure 8a**). All 40 patients in the internal test cohort had omental lesions, H&E, and available sequencing by construction (**Figure 8b**). The TGCA test cohort was composed of H&E WSIs from 84 patients, combined with 62 patients with at least one adnexal lesion on CT, and 54 had with an omental implant on CT (**Figure 8c**). Three radiologists segmented adnexal lesions and representative

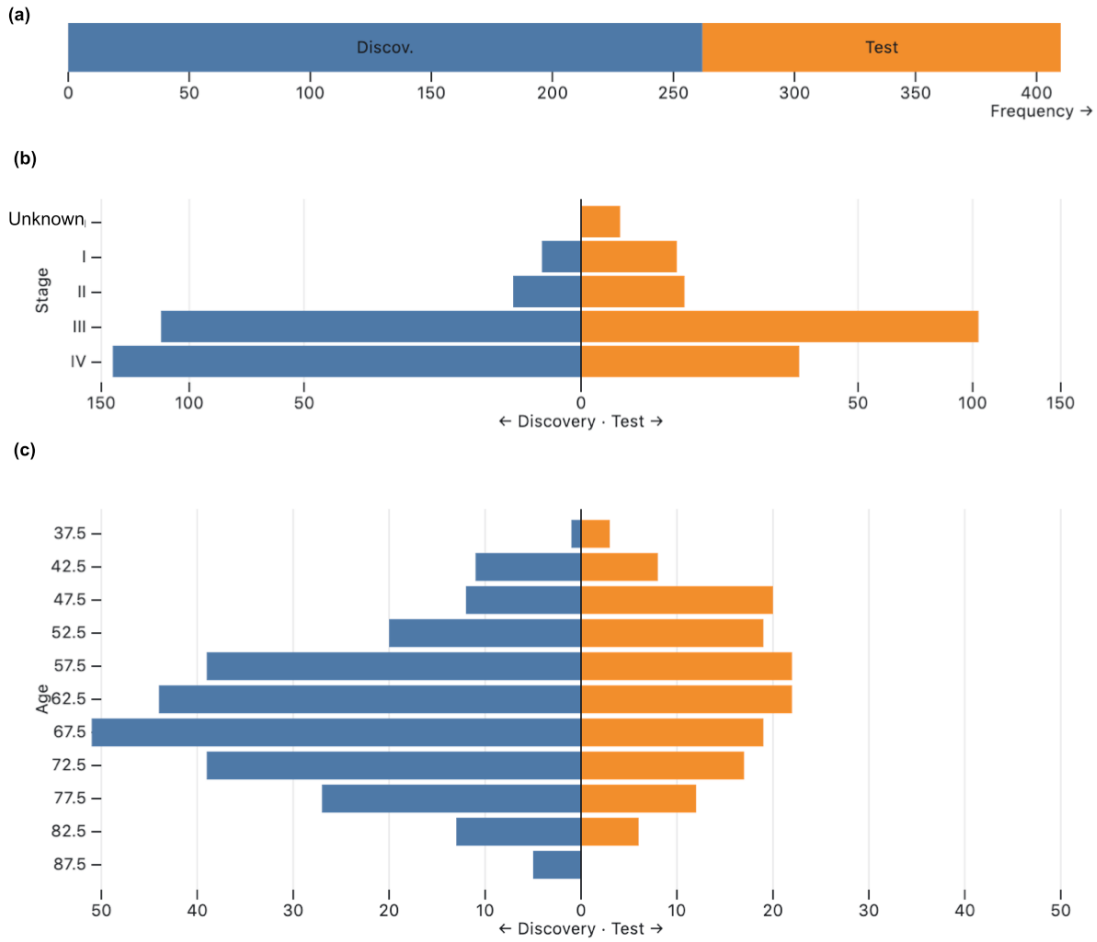


Figure 7. Age and stage distributions of discovery and test cohorts. (**a**) We analyzed 262 discovery patients and 147 external TCGA patients. (**b**) The age distributions are similar for the discovery and test cohorts, with the discovery cohort being slightly older. (**c**) The stage distributions are similar for the discovery and test cohorts, with both cohorts containing primarily advanced disease, and the discovery cohort containing proportionally more stage IV disease. Visualization by Samantha Leung; used with permission.

omental lesions in three dimensions (**Figure 9a**). The discovery and testing data were obtained with similar CT scanners (**Figure 9b**).

We used clinical sequencing to infer HRD status, in particular variants in genes associated with HRD DNA damage response (DDR) [171,172] such as *BRCA1* and *BRCA2*, and those specific to disjoint tandem duplicator- and

	Discovery	**Test**
Patients included	262 (222 training, 40 test)	147 (all test)
Median age at diagnosis	65 years [IQR 58-72]	60 years [IQR 51-68]
Stage		
I	1 (0%)	6 (4%)
II	3 (1%)	7 (5%)
III	115 (44%)	103 (70%)
IV	143 (55%)	31 (21%)
Treatment		
NACT-DPS	188 (72%)	--
Complete gross resect.	126/188 (67%)	--
Number of NACT cycles	4 cycles [IQR 3-5]	--
Received neoadj. PARPi	11/188 (6%)	--
PDS	74 (28%)	--
Unknown	0	147 (100%)
Progression-free survival		
Duration	15.9 months [IQR 12-22]	14.9 months [IQR 9-25]
Censored	86 (33%)	37 (25%)
Overall survival		
Duration	24.5 months [IQR 17-36]	34.8 months [IQR 19-54]
Censored	115 (44%)	53 (36%)

Table 1. Clinical characteristics of cohorts. Continuous values are described by the median. Censored durations are also included in survival descriptions.

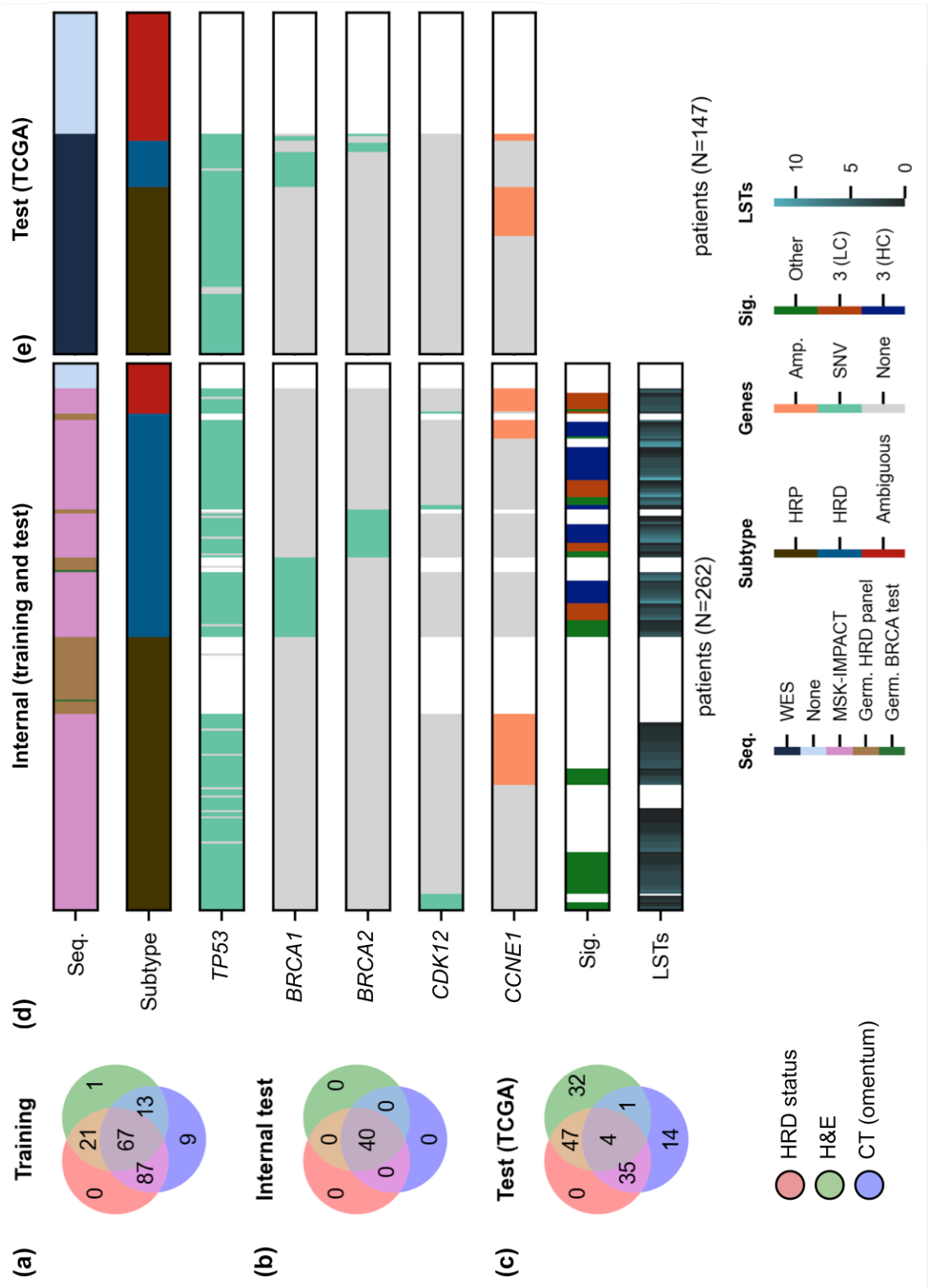


Figure 8. Overview of cohorts and data types acquired. We assembled outcomes, multi-scale imaging data, and HRD status from 409 patients, divided into **(a)** training, **(b)** internal test, and **(c)** TCGA test splits. Patients with available

sequencing but conflicting evidence for HRD status are counted as not having known HRD status in both Venn diagrams. **(d)** HRD status in the discovery cohort was inferred using primarily MSK-IMPACT sequencing to identify OncoKB-annotated variants of known significance in HRD-DDR-associated genes [171] and HRP-associated gene *CDK12*, along with copy number amplifications in *CCNE1* [56], LSTs [173] and Signature 3 activity. (**e**) For the TCGA test cohort, WES was used to infer HRD status using the same methodology, except with an expanded list of HRD-DDR-associated genes [172]. We excluded any cases without sequencing or with conflicting evidence from both training and testing. Only genes with five or more variants in the discovery cohort are shown in this figure. Gray represents tested genes without the aberrations shown, and white represents an untested gene.

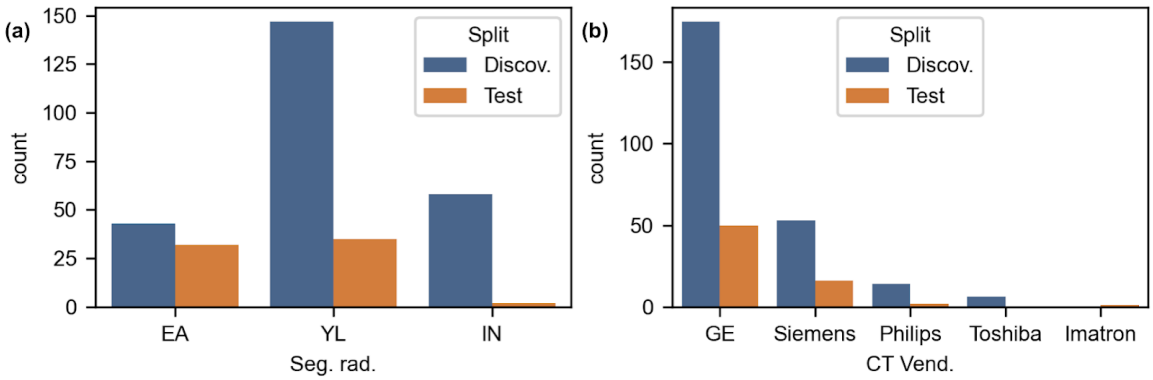


Figure 9. Segmenting radiologist and CT vendor in discovery and test cohorts. (**a**) The same three expert radiologists segmented the discovery and test cases. (**b**) The most common scanner vendors were General Electric and Siemens for both cohorts, with other vendors being less represented. Only the discovery cohort contained scans acquired on Toshiba hardware, and the test cohort contained one scan acquired on an Imatron device.

foldback inversion-enriched mutational subtypes (*CDK12* and *CCNE1* [56,160] respectively, **Figure 6d**, **Figure 8d-e**). We also examined the genomes of 184 discovery patients for direct evidence of homologous recombination deficiency, namely double stranded breaks in the form of large-scale state transitions (LSTs), the median number of subchromosomal regions with allelic imbalance extending to the telomere (NtAI), and COSMIC single base substitution (SBS) signature 3, which is associated with defective HRD-DDR. The median number

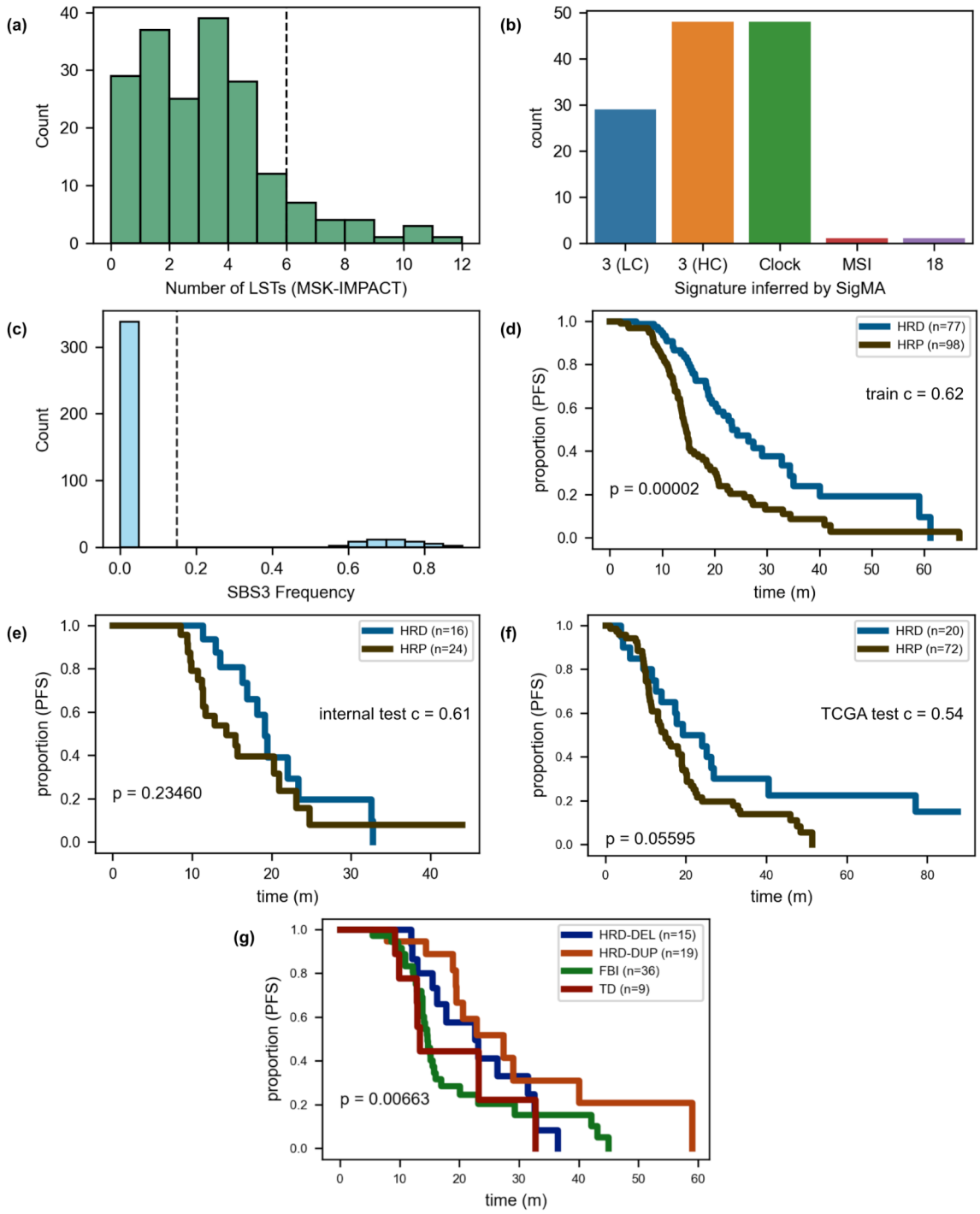


Figure 10. Genomic features and stratification of the discovery and test cohorts. (**a**) The distribution of large-scale state transitions in the discovery cohort is depicted. We set the threshold for LST_{high} versus LST_{low} at 6 LSTs, which is lower than previously reported thresholds for whole-exome sequencing [174]. This is because MSK-IMPACT is a targeted gene panel, and LSTs occurring at the same rate will measure lower on targeted panels compared to more comprehensive sequencing. (**b**) Signature three was detected by SigMA as the

dominant signature with high confidence (HC) and low confidence (LC) in a significant number of cases, and the next most prevalent was the clock signature. (**c**) The COSMIC SBS3 frequencies for all TCGA-OV cases with sequencing from [175] are shown, and the distribution is clearly bimodal but imbalanced. This was not used for HRD status assessment due to poor prognostic association in our cohort. (**d, e, f**) Patients with HRD-type disease have longer PFS than those with HRD-type disease in the training, internal test, and TCGA test cohorts. In **(d)**, only the subset of patients associated with an H&E slide or CT with omental lesion are included. (**g**) Using *BRCA2* SNVs, *BRCA1* SNVs, *CCNE1* CNAs, and *CDK12* SNVs, we categorized a subset of patients into the following mutational subtypes: HRD-Deletion (HRD-DEL), HRD-Duplication (HRD-DUP), Foldback Inversion (FBI), and Tandem Duplications (TD), respectively. The patients stratify as expected, with HRP-type patients suffering earlier progression of disease (p value for log-rank test between aggregated HRD patients and aggregated HRP patients).

of LSTs [173] was 3 [IQR 1-4; max 12] (**Figure 10a**), and NtAI [176] was 5 [IQR 2-7; max 14]. Signature 3 was detected by SigMA [177] in the discovery cohort for 74 cases (45 high confidence and 29 low confidence); it was found not to be the dominant signature in 49 cases (**Figure 10b**). In the TCGA test set, nine patients had COSMIC SBS signature three with a frequency greater than 15%, and 89 were measured that did not (**Figure 10c**) [175]. Patients without sequencing were excluded from all multimodal analyses involving HRD status: no interpolation was used. Patients with conflicting points of evidence were also excluded from these analyses. In total, the discovery cohort had 107 HRD, 131 HRP, and 24 missing or ambiguous cases (**Figure 8d**). No patients in the internal test set within the discovery cohort were of ambiguous status: 16 were HRD, and 24 were HRP. The TCGA test cohort had 20 HRD, 72 HRP, and 55 missing or ambiguous cases (**Figure 8e**). HRD status alone (excluding ambiguous) stratified patients

with a c-Index of 0.62 in the training cohort, 0.61 in the internal test set, and 0.54 in the TCGA test set (**Figure 10d-f**). Aberrations specific to distinct endogenous mutational processes also stratified patients as expected: that is, patients with HRP disease had worse outcomes than those with HRD disease (p=7e-3; **Figure 10g**).

CT imaging feature selection and stratification

We began by studying the prognostic relevance of features derived from radiology scans. Pre-treatment CE-CT scans (**Figure 11a**), were segmented by

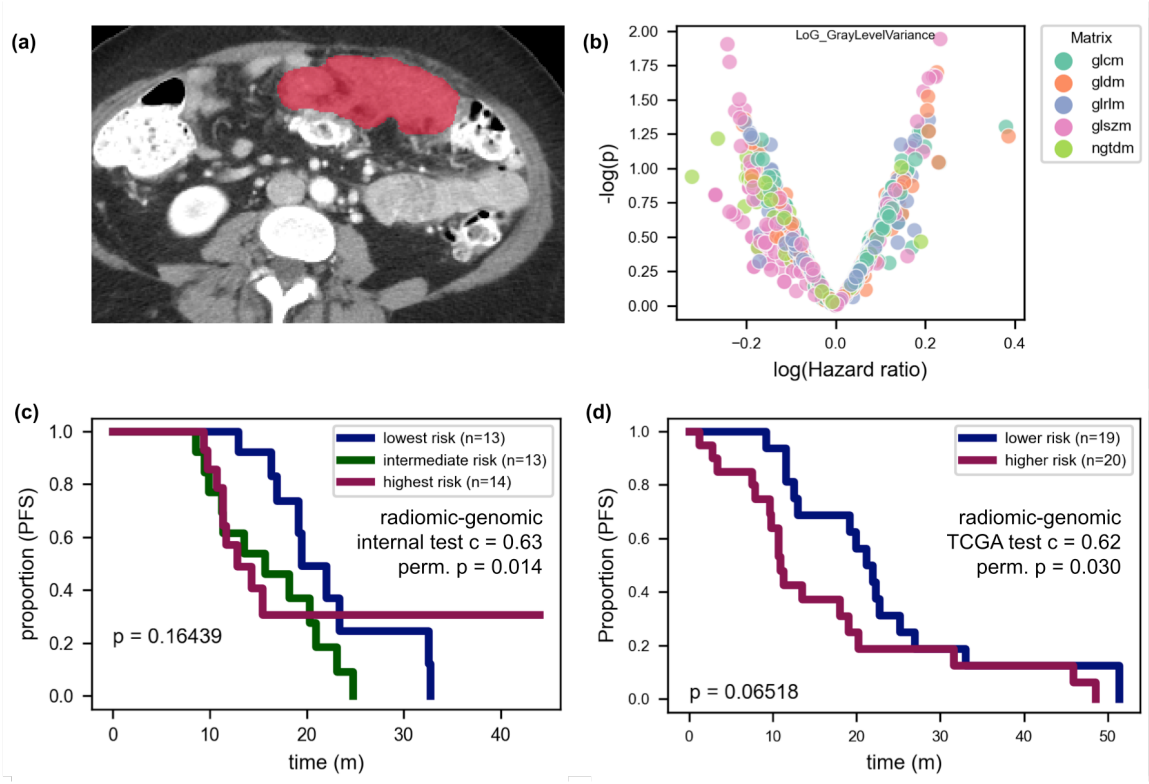


Figure 11. High-density omental zones are associated with shorter progression-free survival. (**a**) Expert radiologists segmented omental lesions in

3D on CT. (**b**) The logarithm of the univariate hazard ratio is depicted for each radiomic feature, with the gray level variance derived from the gray level size zone matrix for the Laplacian of Gaussian-filtered 3D image highlighted as an example prognostic feature. (**c**) A late fusion radiologic-genomic model stratifies patients in the internal test set. (**d**) The model also stratifies patients in the TCGA test set.

fellowship-trained radiologists, focusing on omental implants and adnexal lesions (**Figure 6b, 11a**). Using the training cohort, we identified omental (**Figure 11b**) and ovarian (**Figure 12**) radiomic features associated with progression-free survival using univariate Cox proportional hazards models [162] with bootstrapping of the training cohort [162]. From the top 25 (of 750) omental features, we reduced multicollinearity by iteratively removing features [19]. This yielded a five-feature radiomic signature based on features derived from the

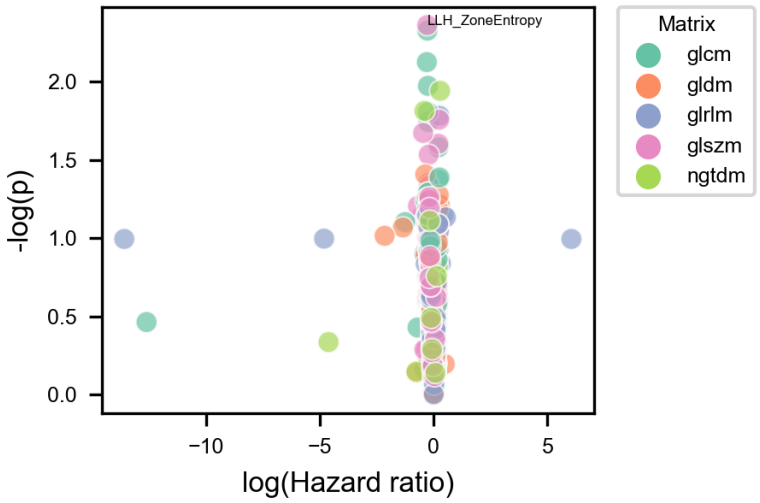


Figure 12. Radiologic ovarian feature discovery. The logarithm of the univariate hazard ratio is depicted for each radiomic feature, with the maximal correlation coefficient derived from the gray level size zone matrix for the Coif wavelet-filtered 3D image highlighted as an example prognostic feature.

gray level size zone [178], gray level run length [179], and gray level dependence [180] matrices calculated on Coif wavelet-transformed [181] images and from the gray level size zone matrix of the Laplacian of Gaussian-transformed image. We found that radiomics features derived from omental implants stratified patients better than ovarian lesions (**Figure 13**) and thus, going forward, we only considered the omental lesions.

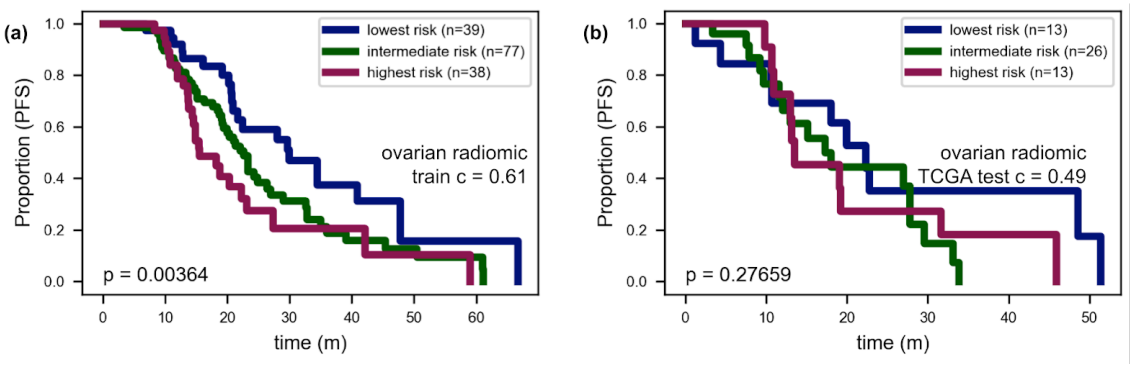


Figure 13. Ovarian radiologic features do not stratify TCGA test set by PFS. (**a**) Ovarian features selected using the training set stratify the training set as expected. (**b**) However, they do not stratify the TCGA test set by PFS.

This five-dimensional signature was invariant to CT scanner manufacturers and segmenting radiologists (**Figure 14**). One feature positively correlated with higher-density voxel zones (in Hounsfield units) was associated with higher risk, while another feature describing large lower-density zones corresponded to lower risk (**Table 2**). Unimodal radiomics achieved a training c-index of 0.55, an internal test c-index of 0.55, and a TCGA test c-index of 0.61 (**Figure 15a-b**). The small size of the test sets should be taken into

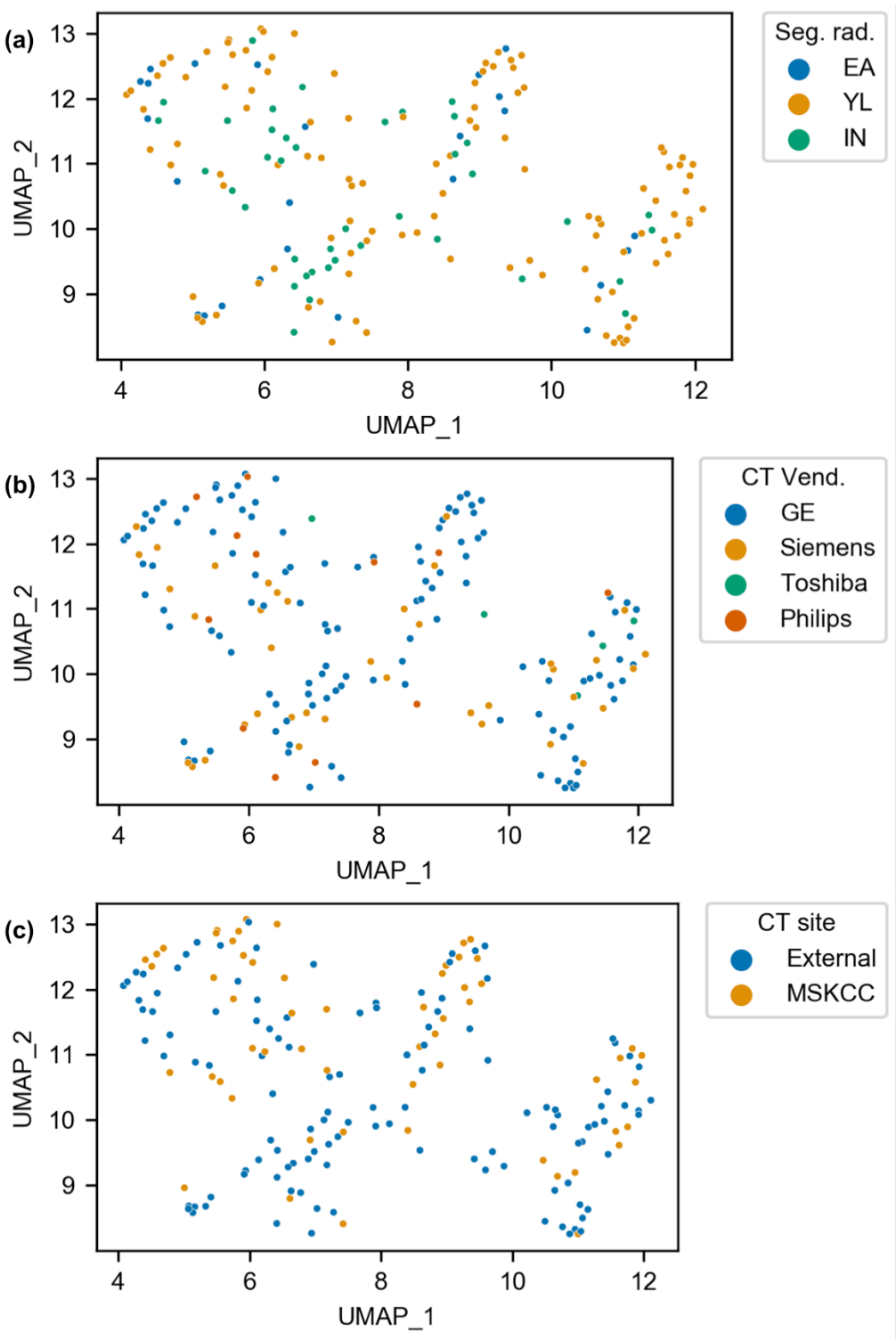


Figure 14. Radiomic embeddings by segmenting radiologist, CT scanner manufacturer, and acquisition site. The radiomic embeddings for the signature we identified in the discovery cohort do not appear significantly confounded in UMAP space by (**a**) segmenting radiologist, (**b**) CT vendor, or (**c**) whether the scan was acquired at our institution (MSKCC) or elsewhere.

consideration, particularly for the test c-index higher than the training c-index. Kaplan-Meier analysis of the high risk, intermediate risk, and low risk groups (as determined by inferred partial hazard) showed correct ordering and separation of the groups in the TCGA test set (p=0.14).

Variable	**Coef.**
wavelet-LHH_glrlm_LongRunHighGrayLevelEmphasis	0.33
wavelet-LHL_gldm_LargeDependenceLowGrayLevelEmphasis	-0.54
log-sigma-1-0-mm-3D_glszm_SizeZoneNonUniformityNormalized	-0.49
wavelet-LLH_glszm_ZonePercentage	-0.58
wavelet-LHL_glszm_GrayLevelVariance	0.45

Table 2. Omental radiomic Cox model parameters.

Combining this radiomic model with HRD status using a late fusion strategy yielded a good stratification in the internal test set, with a c-Index of 0.63 (p=0.014) and correct ordering of the highest-, intermediate-, and lowest-risk groups with modest separation (Kaplan-Meier analysis, p=0.16, (**Figure 11c**). Median PFS was 12.8, 15.7, and 19.4 months for the highest, intermediate, and lowest risk groups, respectively, in the internal test set, compared to the HRD-based risk groups of 14.3 and 19.1 months (p=0.23). In the TCGA test set, the combined model achieved a higher training c-index of 0.65, and the test set performance increased slightly to c=0.62 (p=0.030; **Figure 15e-f**). The highest-risk group separated well from the intermediate- and lowest-risk groups, whereas the intermediate and low risk groups were similar (median PFS 19.2, 20.2, and

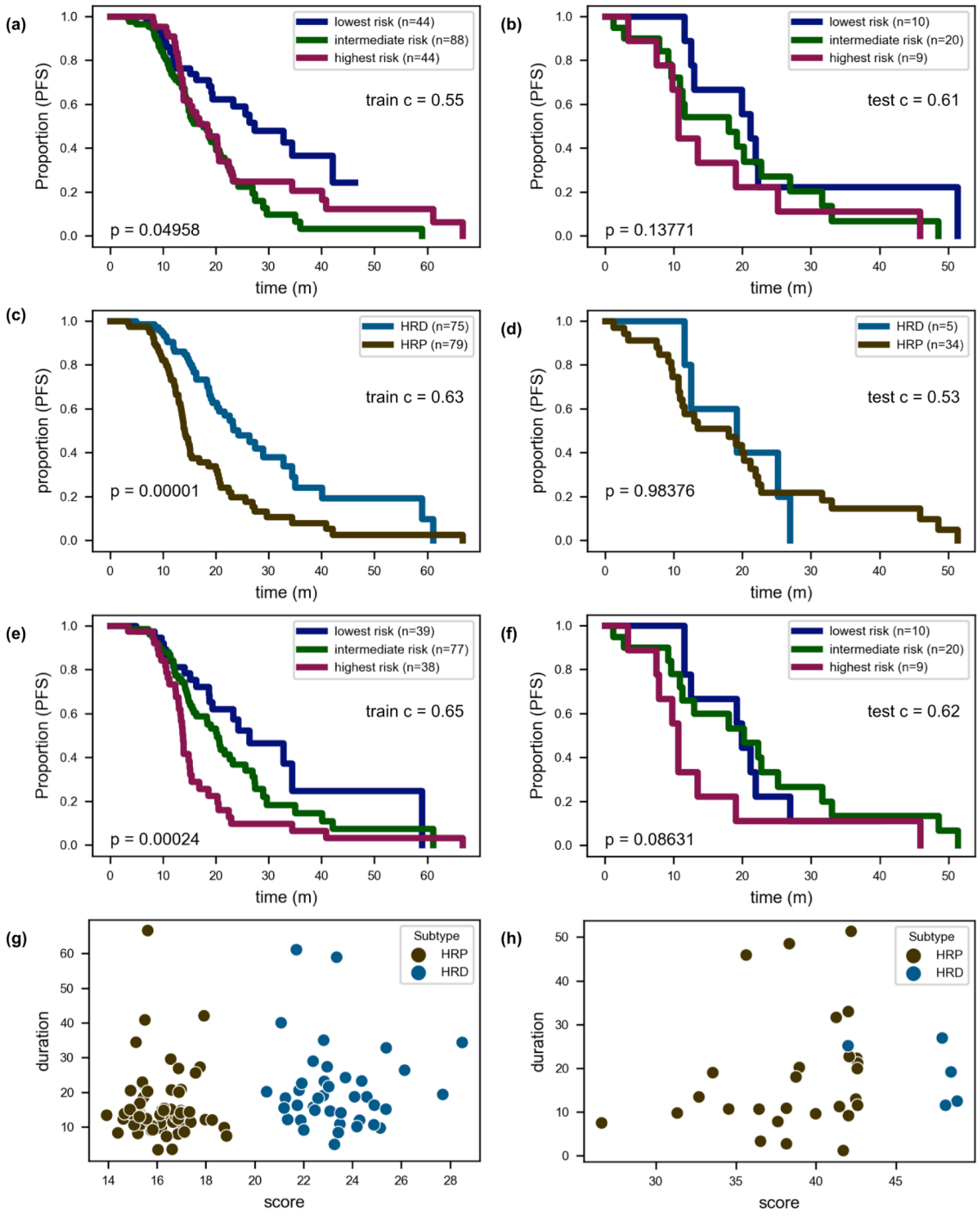


Figure 15. Additional KM analyses of radiologic-genomic models on the training cohort and the TCGA test cohort. (**a, b**) The unimodal radiologic model stratifies the training cohort used for feature selection and withheld TCGA

test cohort. (**c, d**) HRD status alone stratifies the discovery cohort well but does not stratify this subset of patients with omental lesions within the testing cohort well, perhaps due to the low numbers of HRD patients represented. (**e, f**) Combining HRD status and the radiomic model improves stratification of the training cohort and the TCGA test set. (**g, h**) The inferred score (here, expected PFS times) for uncensored patients are displayed. In the test cohort, the radiomic score improves stratification for HRP patients with a good prognosis, effectively softening the categorical distinction. However, a group of patients designated as HRD are given a high score (expected PFS time), whereas they actually suffer worse outcomes, which explains the low-risk group curve crossing over the intermediate risk curve in panel (f).

10.7 months; p=0.086). Separating the patients into two risk groups based on multimodal risk scores (**Figure 11d**) revealed median PFS of 10.8 months and 21.2 months (log-rank p=0.086), which is substantially greater separation than with either the HRD status alone (median PFS 19.0 for HRP, 19.2 months for HRD) or the radiomic features alone (13.4 and 19.9 months). In this subgroup of the TCGA test set, where HRD status-based stratification achieved only c=0.52, the radiomic score corrected the ordering for HRD-designated patients with poor outcomes and also identified HRP patients with better outcomes (**Figure 15h**). In the training set, HRD status-based stratification was stronger than in the TCGA test set, and the radiomic score stratified patients within each mutational subgroup (**Figure 15g**). The same radiomic-genomic model also stratified the TCGA test cohort by overall survival with c=0.60 (high- and low-risk median PFS 44.5 and 57.1 months; **Figure 16**). A clinical submodel trained on patient age and pathologic stage (unimodal train c-Index 0.53, TCGA test c-index 0.49) did not improve stratification beyond the radiologic-genomic model (trimodal train c-index 0.64, test c-index 0.59).

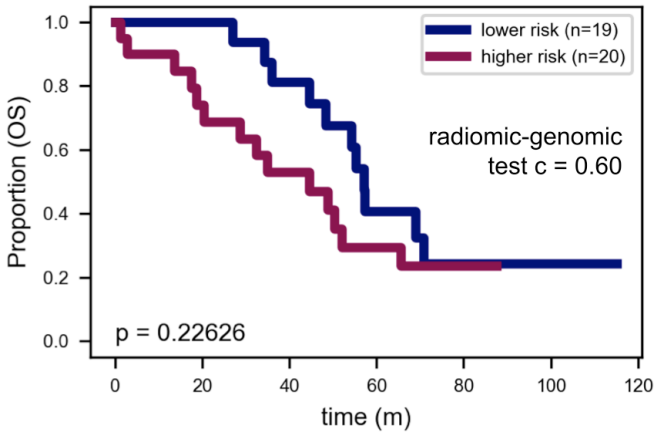


Figure 16. TCGA test performance of radiologic-genomic model using overall survival. The same radiologic-genomic model, with parameters estimated using PFS in the discovery cohort, also stratifies test patients by overall survival (c=0.60). The times to median overall survival are 45 and 57 months.

Histopathologic tissue type classifier for interpretable features

We next trained a tissue type classifier from histology images using a weakly supervised approach. We annotated tissue types on 65 H&E WSIs, yielding more than 1.4 million partially overlapping tiles, each containing 4096 μm^2 of tissue (**Figure 17a**) and trained a ResNet-18 convolutional neural network pretrained on ImageNet using these data (**Figure 17b**). We evaluated performance by four-fold slide-wise cross-validation. The model achieved a balanced classification accuracy of 0.81 ± 0.05 on pathologist-annotated areas labeled as fat, stroma, necrosis, and tumor (**Figure 17c)**. Moreover, the model correctly identified small stromal regions at the edge of the fat and necrotic regions within the tumor, supporting the suitability of weakly supervised deep learning for this task and refining annotations into more granular classifications.

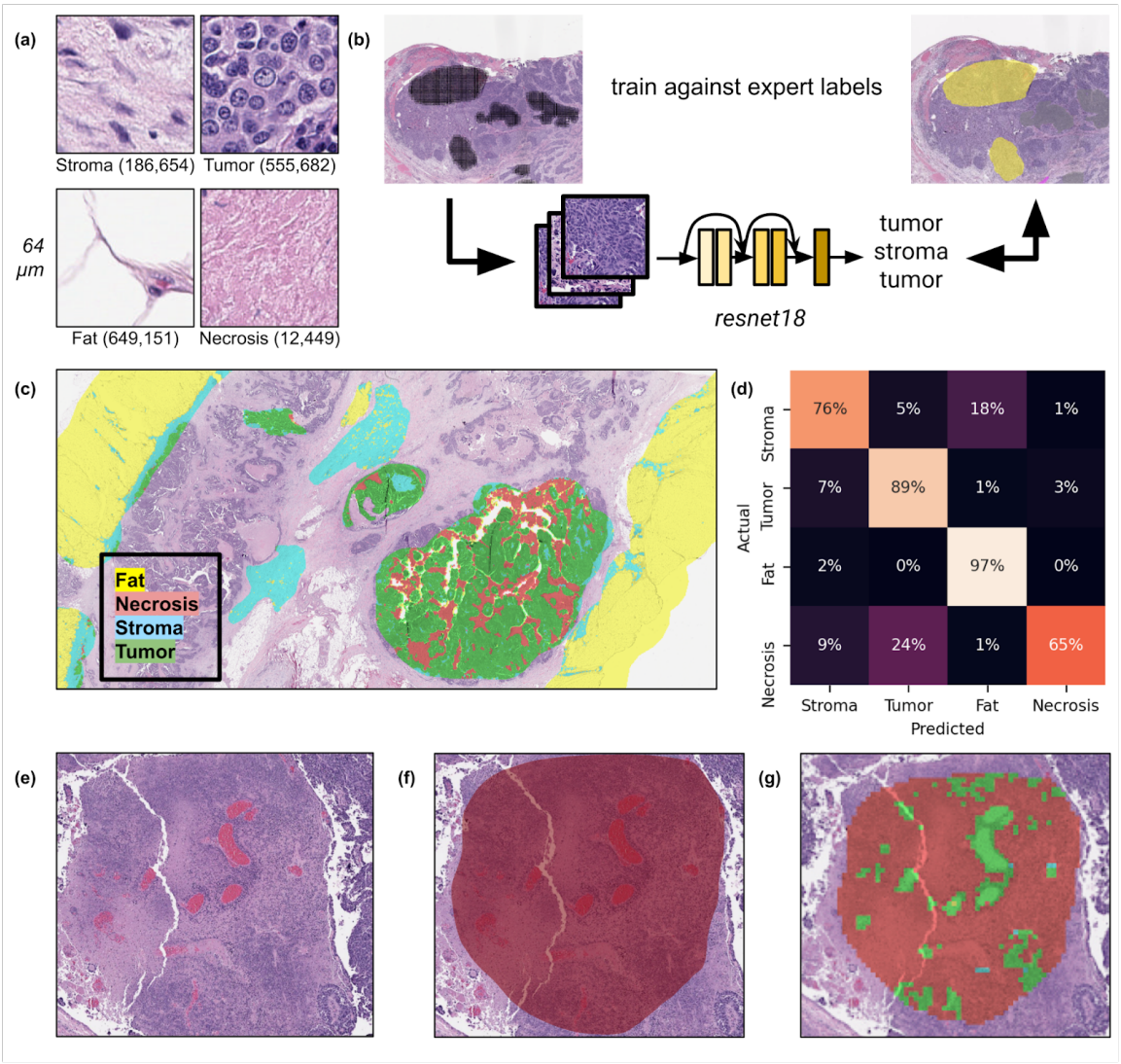


Figure 17. Weakly supervised deep learning accurately infers HGSOC tissue type on H&E. (**a**) A training dataset of labeled tiles was generated with 50% overlap from coarse labels by experienced gynecologic pathologists. (**b**) Annotated regions were tiled, and a ResNet-18 model was trained using weak supervision with the labels. (**c**) The model recapitulates histological structures: it correctly identifies necrosis within a tumor-labeled region, stroma at the edges of fat-labeled regions, and an island of stroma within a tumor-labeled region despite noise in the training data. (**d**) The confusion matrix aggregated across folds of cross validation shows high accuracy but confusion particularly between necrosis and tumor. Consider the region shown in (**e**). The pathologist labels this region as necrosis (**f**), and the model's prediction on cross validation is mainly necrosis, with the vasculature and surrounding intact cells labeled as tumor within (**g**). This is closer to correct than the ground truth but counted as incorrect in the confusion matrix. This supports the efficacy of weakly supervised deep learning for this task.

The cross-validation confusion matrix integrated across folds showed good performance (**Figure 17d**), with the most significant confusion being necrotic tiles predicted to be tumor and stroma tiles predicted to be fat. However, one disadvantage of weakly supervised learning is that neither the training data nor the validation data are exactly labeled: hence, the cross-validation metrics are not computed against the exact truth. Visual inspection of the predictions revealed excellent qualitative performance (**Figure 17c**). For example, in a necrotic region with vessels and perivascular cellularity (**Figure 17e)**, the pathologists labeled the entire region as necrosis (**Figure 17f**), while the classifier assigned labels of necrosis to most tiles, but classified vessels and their surrounding intact cells as tumor (**Figure 17g**). Those tiles classified as tumor were counted as incorrect in the confusion matrix given the coarseness of the labels; the prediction in this case was more accurate than the label.

Histopathologic stratification

We applied the tissue type classifier to our 141 H&E WSIs of soft tissue lesions from pretreatment biopsies (**Figure 6c**). We combined these inferred tissue type maps with detected cellular nuclei, yielding labeled nuclei (**Figure 18a**). Subsequently, we extracted cell-type features from these nuclei and tissue-type features from the tissue-type maps based on the methods of Diao et al. [169]. We

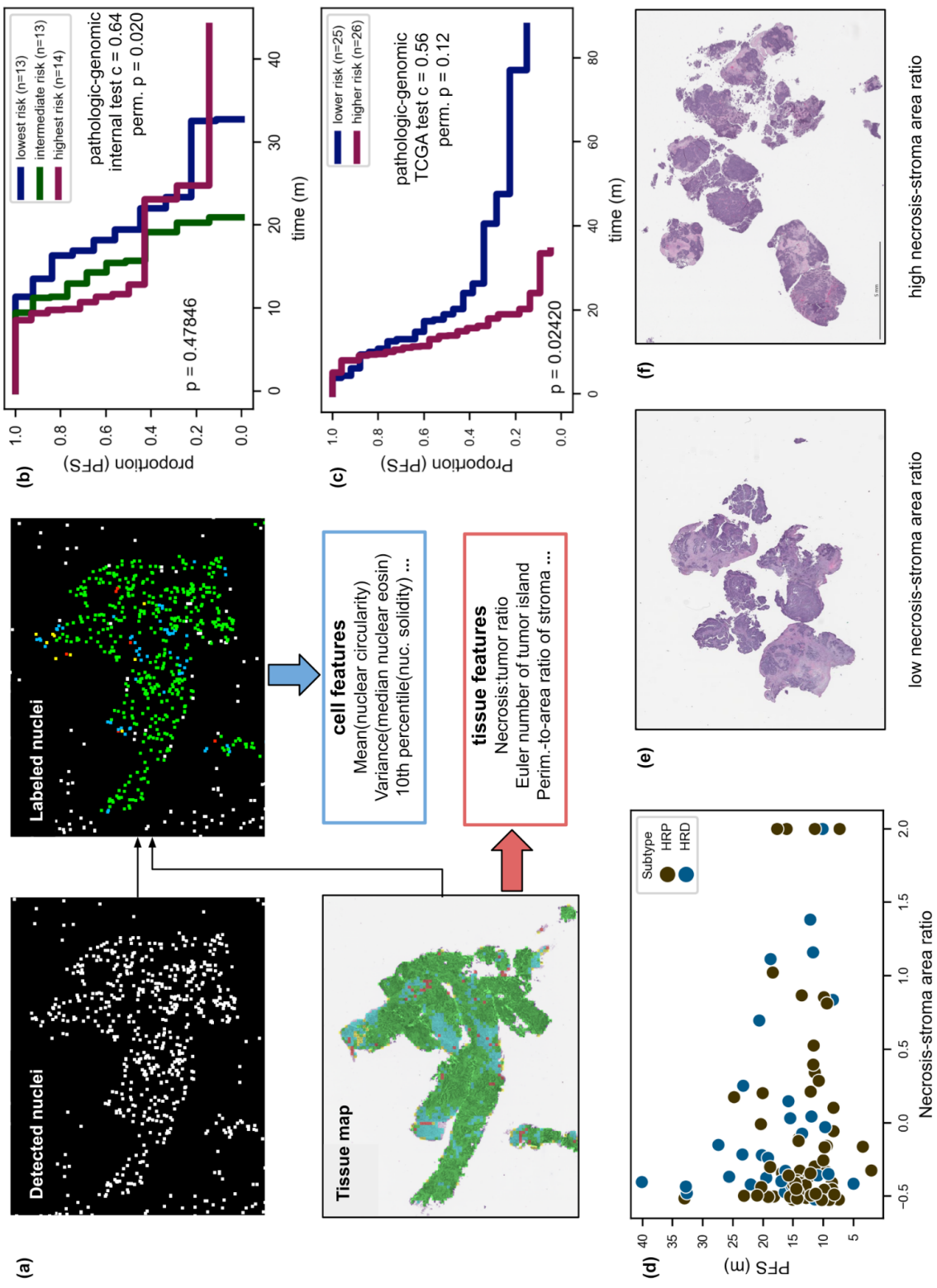


Figure 18. Interpretable histopathologic features stratify HGSOC patients by PFS. (a) For each H&E slide, the tissue map is generated and combined with

StarDist nuclear detections to yield both tissue-type features and cell-type features, examples of which are shown. (b) A late fusion pathologic-genomic model stratifies patients in the internal test cohort. (c) The model also stratifies patients in the TCGA test dataset. (d) One prognostic feature is the ratio of necrotic area to stromal area for a slide. Patients with higher values of this feature have shorter PFS. Censored patients are not plotted; truncated at 2 std. dev. for visualization. (e) This slide was inferred to have a low necrotic to stromal area ratio: note stroma amidst tumor but no prominent necrosis. (f) This slide was inferred to have a high necrotic to stromal area ratio, as verified by the pink soupy material with total loss of cellular architecture adjacent to the serous carcinoma cells.

next selected features using univariate Cox models on features derived from slides in the training cohort (**Figure 19**). Several tissue-type features, such as overall necrotic area, were partially determined by specimen sizes, and we thus controlled for this during selection. Using cross-validation (**Figure 20**), we chose to use the top four most significant features associated with PFS for the model: the ratio of necrotic area to stromal area, the perimeter-to-area ratio of the largest tumor component, the whole-slide skew of maximal hematoxylin of tumor nuclei, and the whole-slide kurtosis of median eosin of tumor nuclei (**Table 3**).

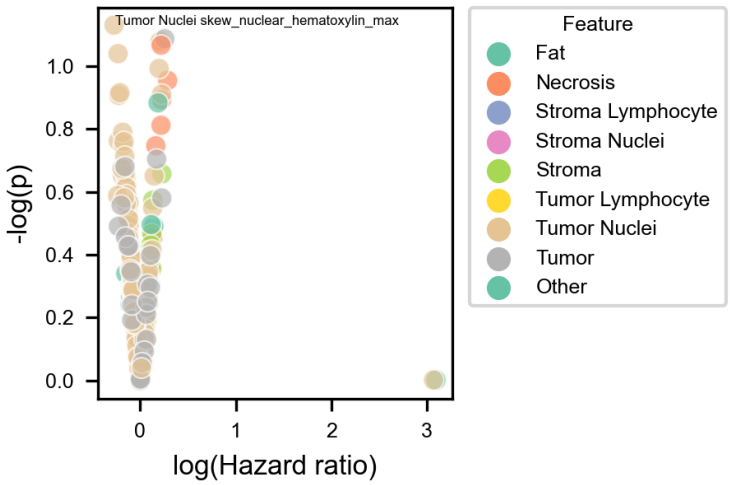


Figure 19. Histopathologic feature discovery. The logarithm of the univariate hazard ratio is depicted for each histopathologic feature, with the skew of the

maximal hematoxylin values of each tumor nucleus highlighted as an example prognostic feature.

This histopathologic signature was not significantly confounded by specimen size (**Figure 21**). The training c-Index was 0.63, the TCGA test c-index was 0.54, and the internal test c-index was 0.54 (**Figure 22a-b**). For comparison, the HRD status-based model achieved a unimodal train c-Index of 0.59 and test c-Index of 0.53 (**Figure 22c-d**).

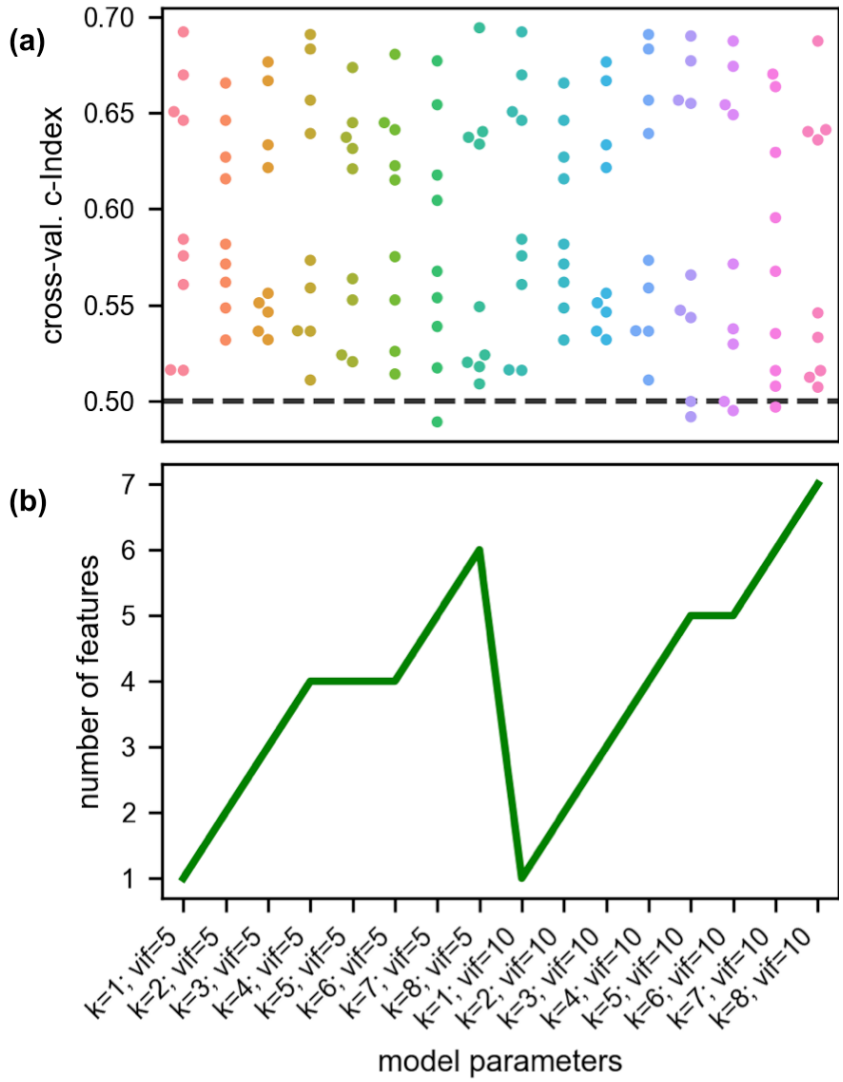


Figure 20. Histopathologic feature selection hyperparameters and resultant cross-validation performance. (**a**) For feature selection, we varied *k*, the number of top prognostic features to consider, and *VIF*, the variance inflation factor at which iterative feature removal halts. We chose k=4 (either tested value of VIF yields the same signature) because the model has comparable performance to the other models and none of the top four features are yet collinear with the VIF thresholds we used. (**b**) The numbers of resultant features after iterative feature removal are shown.

Variable	**Coefficient**
ratio_necrosis_to_stroma	0.17
Tumor_largest_component_PA_ratio	0.21
Tumor_Other_skew_nuclear_hematoxylin_max	-0.23
Tumor_Other_kurtosis_nuclear_eosin_median	0.14

Table 3. Histopathologic Cox model parameters.

Integrating genomic and histopathologic (GH) models increased the c-indices (0.65, 0.64, and 0.56 for training, internal test, and TCGA test sets, respectively (**Figure 18b, Figure 22e, f**) (permutation p = 0.020, 0.12). On multivariate regression, the histopathologic and HRD status-based sub-models were both significant (HRD p=0.02, H&E p=0.01). Kaplan-Meier analysis of the highest, mid and lowest risk quartiles yielded correct ordering but statistically insignificant separation in both the internal test and TCGA test cohorts (p=0.48, 0.35; **Figure 22f**). Median score-based dichotomization yielded significant separation in the TCGA test cohort (p=0.02; **Figure 18c**). The median time to progression for the risk halves were 18.9 and 13.7 months, slightly less than the genomic HRD and HRP risk groups with 24.0 and 14.8 months (p=0.13).

However, the multimodal model achieved statistically significant separation while the genomic model did not. In addition, stratification was refined, resulting in nearly double the number of patients at low risk while maintaining this clear separation of curves, and expanding the low-risk group from 14 HRD patients to 25 patients with a low multimodal risk score. In both the training and TCGA test

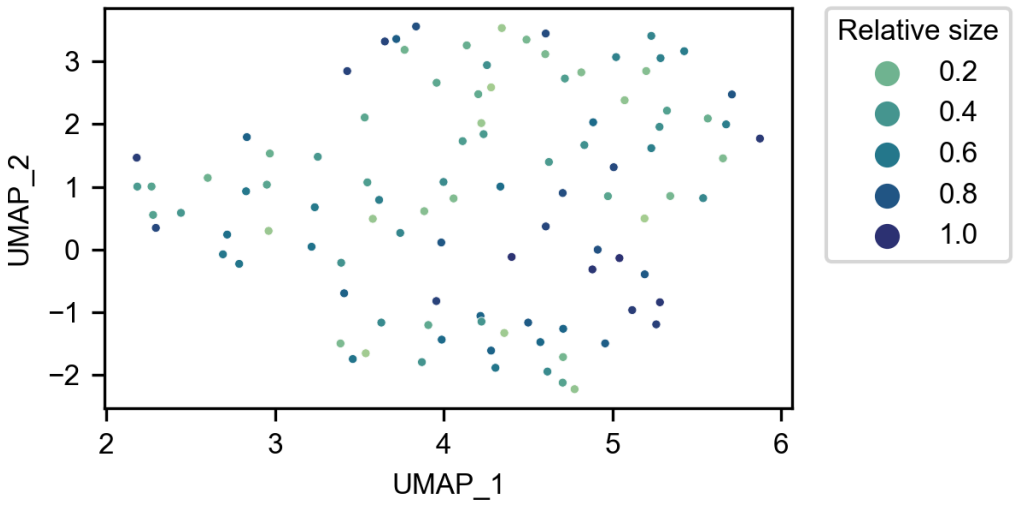


Figure 21. Histopathologic embeddings do not vary by specimen size. The embeddings in UMAP space of the four-feature histopathologic signature do not appear influenced by the relative specimen size (here depicted as the quantile of the number of foreground tiles detected). The larger specimens appear relatively evenly distributed.

sets, the histopathologic model identified HRP patients with a better prognosis and adjusted prognostication accordingly (**Figure 22g, h**). Using overall survival, the genomic-pathologic model also stratified patients with c-index of 0.63 (high- and low-risk median OS 33.3, 49.1 months; **Figure 23**). Integrating a clinical sub-model based on patient age and pathologic stage did not improve stratification by

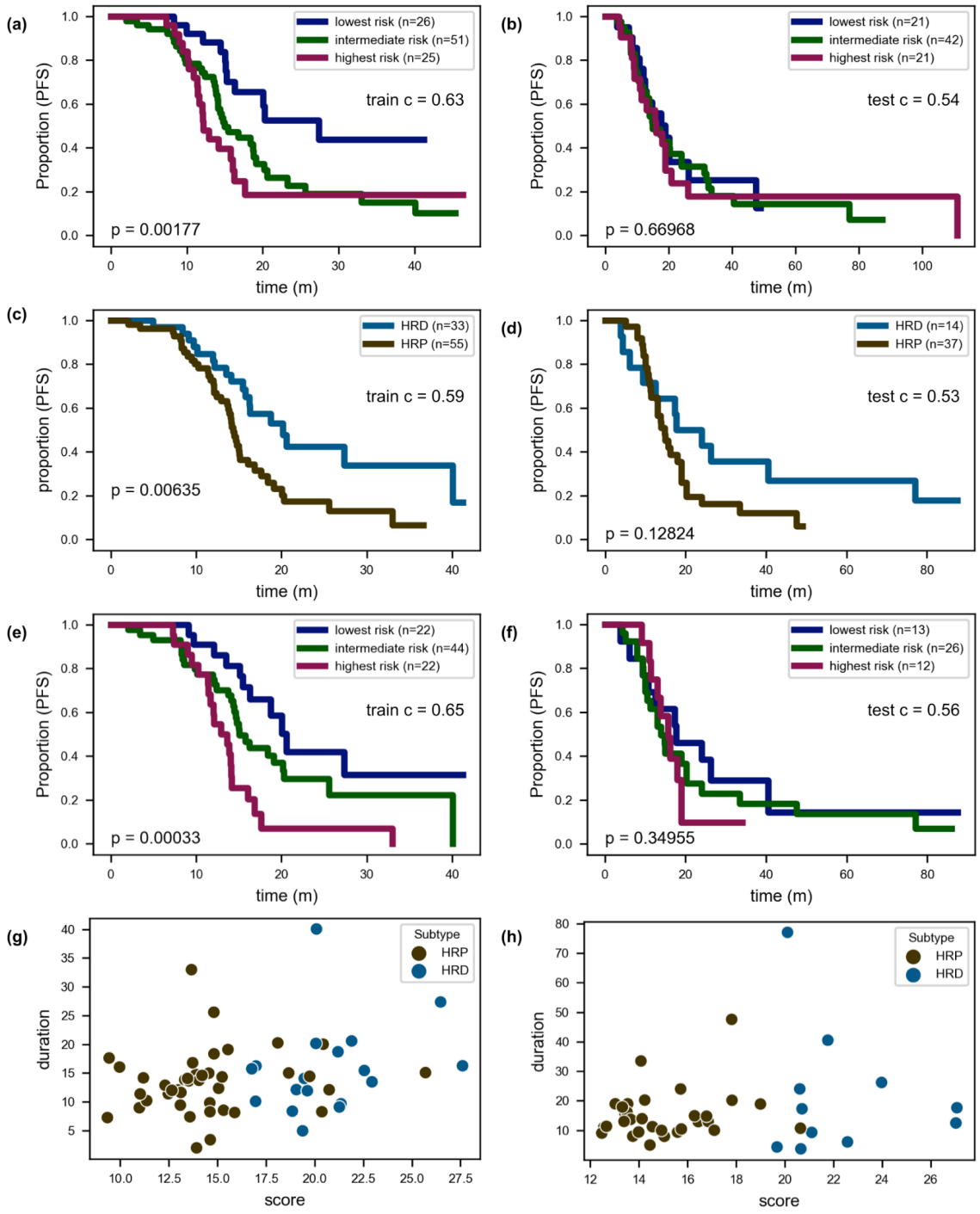


Figure 22. Additional KM analyses of histopathologic-genomic models on training cohort and TCGA test cohort. (**a**) The unimodal histopathologic model stratifies the training set used for feature discovery well as expected. (**b**) On the TCGA test set, the model stratifies patients with c=0.54 but does not delineate distinct risk groups on KM analysis. (**c, d**) HRD status alone stratifies these subsets of the training and TCGA test cohort. (**e**) Combining HRD status and the histopathologic model improves stratification of patient risk groups in the training

cohort. (**f**) The histopathologic-genomic model stratifies the test cohort: note that the high-risk curve is determined by the HRD status in (d). The high-risk curve has a steeper slope than the HRP curve in (d), but the intermediate and high-risk groups are very similar. An adaptive threshold would likely improve the separation. (**g**, **h**) The inferred expected PFS times for uncensored patients are displayed. In the discovery cohort, HRP patients with better prognosis and HRD patients with worse prognosis are shunted away from the dichotomous genomic risk groups. In the test cohort, the histopathologic score primarily stratifies HRP disease, adjusting genomic prognostication correctly.

PFS (train c-Index of 0.65, test c-Index of 0.55) but was significant on multivariate regression (p=0.04) and helped slightly separate the low and intermediate risk groups (**Figure 24**). To probe the interpretability of the histopathologic features, we investigated the necrosis-to-stroma area ratio (**Figure 18d**). We show examples of low (**Figure 18e**) and high (**Figure 18f**) values, respectively, associated with better and worse prognosis.

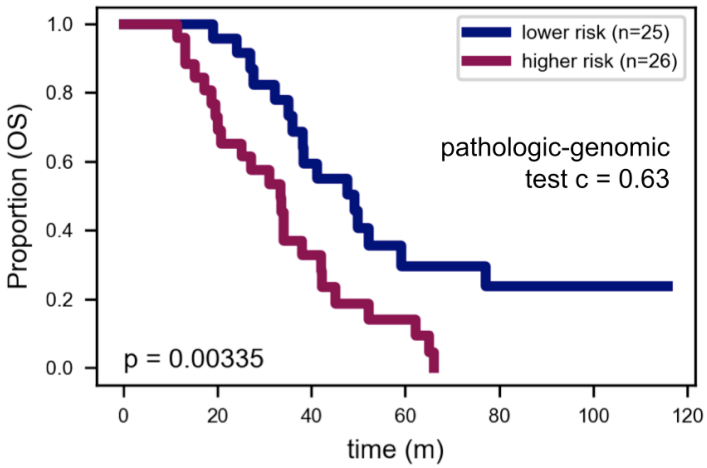


Figure 23. TCGA test performance of histopathologic-genomic model using overall survival. The same histopathologic-genomic model, with parameters estimated using PFS in the discovery cohort, also stratifies test patients by overall survival (c=0.63). The times to median survival are 33 and 49 months.

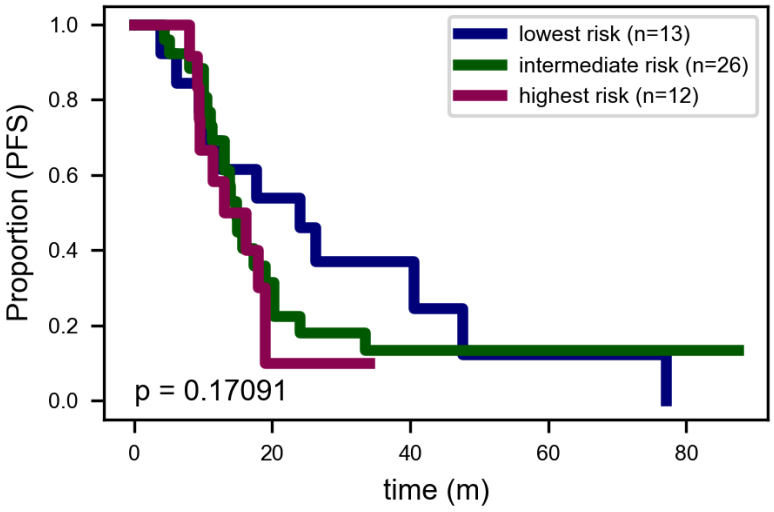


Figure 24. Adding a clinical sub-model slightly improves separation of low- and intermediate-risk groups. The clinical sub-model does not appreciably improve the concordance but does slightly increase separation between the intermediate- and low-risk groups.

Multimodal prognostication

We then integrated histopathologic, radiologic, genomic, and clinical data into a single model (**Figure 6e**), finding that in the internal test set, the model performance rose and significantly outperformed the HRD status-based model, clinical model, and individual imaging models. Kaplan-Meier analysis of the HRD status-based model stratified patients into two separate risk groups with median PFS of 14.3 and 19.1 months (log-rank p=0.23; **Figure 10e**). The same risk stratification using the multimodal model showed the potential value of multimodal stratification, with median PFS of 12.8 months, 15.7 months, and 19.4 months (log-rank p=0.21). The histopathologic-radiologic-genomic (GHR) model had a concordance index of 0.70 on the training set (**Figure 25a**) and 0.66 (**Figure 25b;** permutation p=6.0e-3) on the internal test set, superior to 0.54, 0.55, and 0.61 achieved by unimodal histopathologic, radiologic, and genomic

models, respectively. The GHR model also outperformed the histopathologic-genomic and radiologic-genomic models (**Figure 25c**; c=0.64 and 0.63, respectively). A clinical model including patient age, pathologic stage, residual disease status after cytoreductive surgery, and number of NACT cycles (**Table 4**) achieved a unimodal c=0.47 and did not improve stratification for any model.

The mean absolute Kendall rank correlation coefficient values were low between individual modalities (<0.25) (**Figure 25d**), demonstrating that the radiologic and histopathologic models use distinct information to stratify patients, as compared to both the genomic model and to one another. The GHR model ordered patients in partial accordance with all included modalities, and excluding either imaging modality reduced the correlation and performance as expected. Individual imaging modalities achieve similar unimodal c-indices but identify distinct patient subgroups with good prognosis (**Figure 25e**). That is, the modalities in effect tempered one another: some patients with good outcomes were identified as high risk by the radiologic sub-model but correctly assigned a lower risk score by the histopathologic sub-model, and vice versa. Patients with HRD and HRP disease were distributed relatively evenly, agnostic to unimodal imaging risk scores. Finally, the GHR-estimated partial hazard associated with pathologic chemotherapy response score (**Figure 25f**) with increased discriminatory power over the clinico-genomic risk score (**Figure 25g**): patients demonstrating worse chemotherapy response scores received higher risk scores. This was also true for the GHRC, R, and GRC models, but not for the G, H, C, GR, GC, GH, HR, and GHC models.

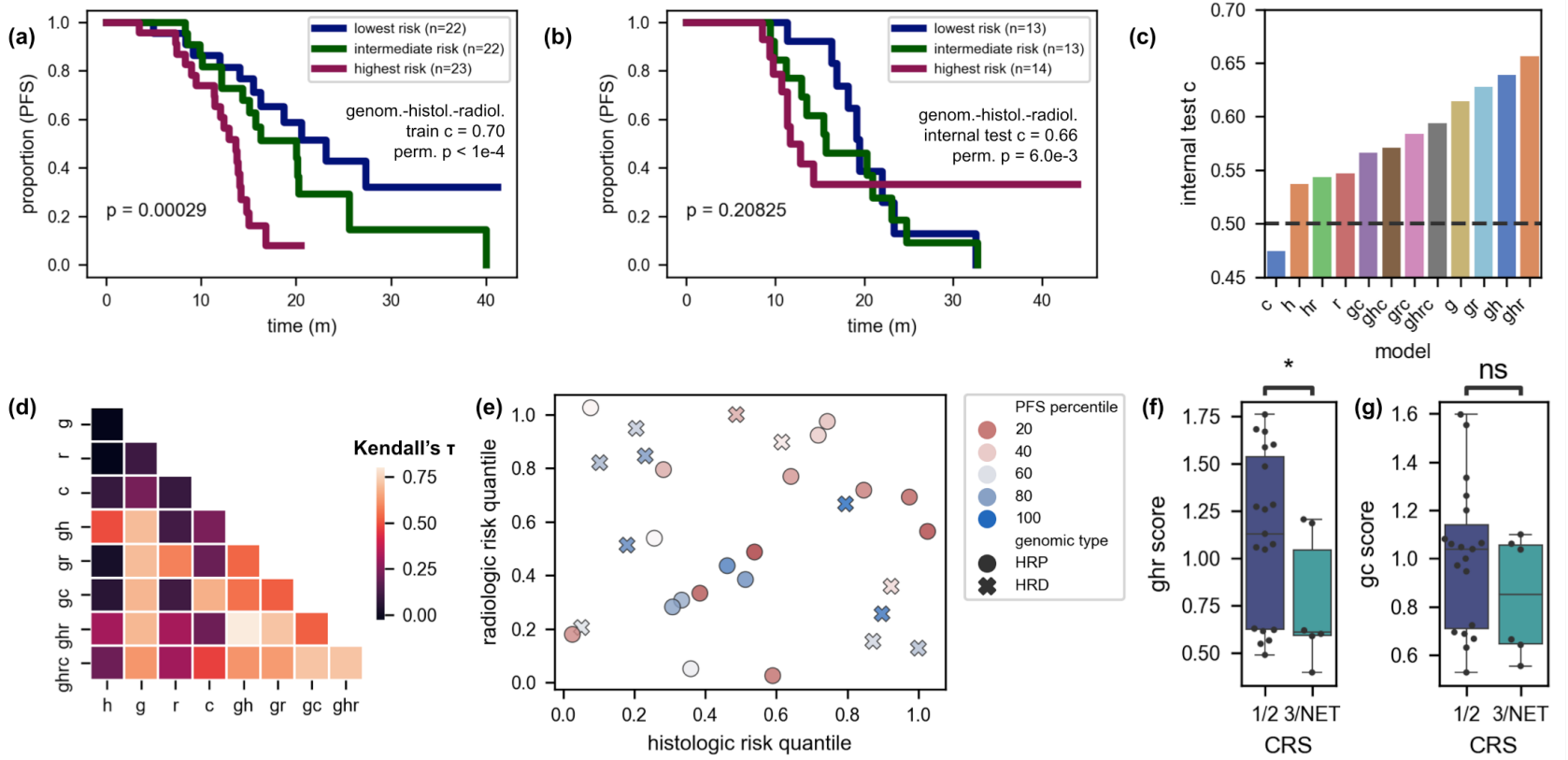

(a)
lowest risk (n=22)
intermediate risk (n=22)
highest risk (n=23)
genom.-histol.-radiol.
train c = 0.70
perm. p < 1e-4
p = 0.00029
proportion (PFS)
time (m)
(b)
lowest risk (n=13)
intermediate risk (n=13)
highest risk (n=14)
genom.-histol.-radiol.
internal test c = 0.66
perm. p = 6.0e-3
p = 0.20825
proportion (PFS)
time (m)
(c)
internal test c
model
c h hr r gc ghc grc ghrc g gr gh ghr
(d)
Kendall's τ
g r c gh gr gc ghr ghrc
h g r c gh gr gc ghr
(e)
radiologic risk quantile
histologic risk quantile
PFS percentile
20
40
60
80
100
genomic type
HRP
HRD
(f)
*
ghr score
1/2 3/NET
CRS
(g)
ns
gc score
1/2 3/NET
CRS

Figure 25. Multimodal integration identifies clinically significant subgroups and improves stratification by response to therapy in the internal test cohort. (**a, b**) Fusing histopathologic and radiologic models with HRD status improves the overall training and testing stratification to c=0.70 and c=0.66, respectively. (**c**) The best performing model integrates both imaging modalities with HRD status. (**d**) The Kendall rank correlation coefficient of the risk quantile is near-zero between any two of the individual modalities, demonstrating low mutual ordering information between individual modalities. (**e**) Corroborating this, unique patients at risk of early progression are identified by radiologic, histopathologic, and genomic modalities. Many patients with longer PFS (in blue) are categorized as higher risk by one imaging model but correctly identified as lower risk by the other. Only patients with uncensored outcomes are shown. (**f**) Higher inferred multimodal risk is associated with a worse chemotherapy response score ($p<0.05$ by the Mann-Whitney U Test). (**g**) The clinico-genomic score alone does not associate significantly with chemotherapy response score. Boxes denote interquartile range, and whiskers denote the entire distribution excluding any outliers. Significance assessed by one-sided Mann-Whitney U test.

Variable	**Coefficient**
dps_cgr [1=True]	-0.36
parp_nact [1=True]	-0.31
patient_age	0.39
pathologic_stage [0=III; 1=IV]	-0.09
cycles_nact	0.06

Table 4. Clinical Cox model parameters.

CHAPTER THREE: Discussion

Machine learning in cancer prognostics is a growing field with great potential, but how multimodal integration across common diagnostic modalities can contribute to risk stratification is still poorly understood. In this work, we have presented two new unimodal models to stratify HGSOC patients using routine clinical imaging, validated these models on two separate test sets, and studied the relative contributions of each modality to risk-stratifying HGSOC patients. For radiologic imaging, we developed a simple model based on omental implants. Patients with more heterogeneous omental implants as indicated by large higher-density zones in Hounsfield Units on CE-CT had worse outcomes. We focused on the omentum over the adnexa because it is the most common site of tropism in HGSOC [182], omental implants are generally easier to identify and delineate by observers with variable levels of experience, and models of omental CT features outperformed those based on ovarian features. Thus, our model may be useful when ovarian masses are not present (as in primary peritoneal high-grade serous cancer) or challenging to delineate due to adjacent structures, such as the uterus. To our knowledge, prior HGSOC radiomic models have not explored the prognostic information captured within omental implants, relying instead on more demanding segmentations of adnexal lesions or the entire tumor burden.

For histopathologic imaging, we developed an H&E WSI-based model to stratify HGSOC patients. We discovered histopathologic features that are associated with PFS in HGSOC, namely the ratio of necrotic area to stromal area, the perimeter-to-area ratio for the largest island of tumor, and features

describing nuclear staining. This result validated on two test sets and supports the presence of pathologic prognostic factors beyond the pathologic stage in HGSOC. Necrosis is associated with rapidly growing, aggressive tumors that outpace their vasculature [183]. The perimeter-to-area ratio is difficult to interpret for a two-dimensional slice of tissue but may reflect distinct patterns of disease infiltration into surrounding stroma. It is also challenging to interpret nuclear staining distributions at 20x magnification, but future studies could be conducted at 40x magnification to establish whether differences in chromatin conformation or nucleolar activity [184] are correlated. We included the trained weights for our HGSOC model and the source code for extension to other cancer types.

Integrating clinical imaging data with HRD status improved stratification beyond clinico-genomic models and any unimodal model, with the best model incorporating histopathologic, radiologic, and genomic information. These results, in addition to the low correlation between individual modalities, support that clinical imaging contains complementary prognostic information rather than merely recapitulating clinico-genomic information. Histopathologic and radiologic imaging characterize the tumor architecture at microscopic and mesoscopic scales, respectively. Therefore, it stands to reason that these data complement HRD status, which is derived from spatially agnostic sequencing. In the radiologic-genomic model applied to the TCGA cohort, data integration only slightly improved stratification. This may signal that multimodality is not a universal guarantee of improved performance [185]. In this case, the most likely reason is that the HRD stratification was weaker than the radiology stratification

(perhaps due to imperfect HRD status assessment). With larger datasets, mechanisms such as attention can be explored to adaptively adjust unimodal contributions.

Though our results demonstrate the empiric benefits of multimodal analysis, our concordance indices remained well below 1.0. This is consistent with metrics from other HGSOC studies, where c-Indices are between 0.60 and 0.70 for various OS models including miRNA-seq and mRNA-seq [79], and approximately 0.6 for H&E WSI-based models [30]. A previously published adnexal radiomic model for overall survival yielded a c-Index of 0.68 [162]. Ultimately, larger multi-institutional cohorts and homogeneous outcomes definitions will help reduce overfitting and improve stratification.

This lack of usable large datasets is one of the main challenges for multimodal machine learning in oncology. We have made our dataset of 409 HGSOC patients available to enable future work toward improving upon the models presented here. Though relatively sizable by clinical standards, the dataset remains in the very small data regime for machine learning [186]. This limits the utility of highly flexible machine learning techniques such as deep learning, which are likely to overfit with so few independent samples. It also makes handling missingness especially pertinent: we used a late fusion strategy to train each unimodal submodel on all available data and estimated integrative parameters only on patients with all available modalities. As larger multi-institutional cohorts coalesce, deep models and more advanced intermediate

fusion strategies [60] offer potential performance improvements and ought to be explored.

We intentionally mined data that are routinely available from the standard of care. This has the advantage of drastically reducing adoption costs in the clinical workflow for any potential resultant models, but the data were not collected specifically with computational modeling in mind. A study readable by a diagnostic radiologist for clinical purposes is not necessarily suitable for quantitative feature extraction, and we excluded many studies due to quality issues. Similarly, many candidate H&E slides contained only a few serous carcinoma slides, which may be enough for a pathologist to register suspicion for HGSOC but insufficient for quantitative analysis. We made every effort to include imperfect data to improve the generalizability of our results. Similarly, we included some patients in our discovery cohort with only germline panels of HRD-DDR genes, a clinically relevant but biologically imperfect measure of HRD status. Our test cohort included only patients with WES, and the advent of clinical whole-genome sequencing will enable more nuanced retrospective genomic analyses.

In summary, we have assembled a multimodal dataset in HGSOC patients and used this to develop and integrate combined radiologic, histopathologic model, and clinico-genomic models to risk-stratify patients. We show that these modalities are demonstrably orthogonal, and their computational integration improves stratification beyond previously known clinico-genomic factors in two test cohorts. Our results motivate further large-scale studies driven by multimodal

machine learning to stratify cancer patients, both in HGSOC and other cancer subtypes.

CHAPTER FOUR: Methods

Discovery cohort curation

with Kara Long Roche, Ying Liu, Dmitriy Zamarin, Emily Aherne, and Yulia Lakhman

Most of the discovery cohort was sourced from a retrospective clinical database of patients who underwent diagnostic workup and NACT-DPS at our institution. We reviewed the EHR to find associated pathology cases with intraperitoneal soft tissue lesions (primarily omental), and expert pathologists reviewed the slides to select high-quality specimens for digitization. To expand the cohort, we also searched the institutional data warehouse for patients with MSK-IMPACT sequencing and available CT studies, then filtered these patients to those with unambiguous mutational subtype based on annotated variants. We subsequently reviewed the associated CE-CT scans and excluded patients with poor quality studies (artifacts, low signal-to-noise ratio, or poor intravenous contrast bolus timing). We extracted cytoreductive status, number of cycles of chemotherapy, pathologic stage, diagnostic biopsy accession numbers, and patient age at diagnosis from the electronic medical record. We reviewed the institutional data repository for scanned slides associated with the diagnostic biopsy and included those containing tumors. Pathologic stage was recorded for all except three patients for whom it was unavailable: for these patents, clinical stage was used instead. Cytoreductive status was unknown for one patient who underwent external debulking surgery: this patient was treated as not having undergone a complete gross resection. It could not be definitively determined from the medical record whether three patients received neoadjuvant PARP inhibitors: since no

mention was made of them and these patients did not undergo any HRD-DDR sequencing, these patients were assumed not to have received PARP inhibitors. We only included histopathologic specimens and clinical covariates for patients receiving neoadjuvant chemotherapy. Only patients receiving neoadjuvant chemotherapy were included in the internal test set, which was sampled at random from the internal cohort patients with known HRD status, omental lesion on CT, and H&E specimen.

TCGA Test cohort selection

From the TCGA-OV project [187], we selected patients with clinical data annotated in the TCGA Clinical Data Resource [170] and pathologic grade 3 and at least one of either a diagnostic FFPE H&E WSIs or contrast-enhanced abdominal/pelvic CT study in the TCIA [188]. Patients with scans judged to be of low-quality by radiologists were excluded before analysis. Only diagnostic WSIs of formalin-fixed, paraffin-embedded H&E-stained specimens from the TCGA-OV project were included.

Inferring HRD status

with Pier Selenica

In the discovery cohort, we used MSK-IMPACT clinical sequencing [189], when available, to infer HRD status. Variant calling for these genes and copy number analysis of *CCNE1* was performed using the standard MSK-IMPACT clinical pipeline (https://github.com/mskcc/Innovation-IMPACT-Pipeline). We also inferred COSMIC SBS3 activity using SigMA (for cases with at least five mutations across all 505 genes) [177] and searched for large-scale state transitions

[173] using our own pipeline (https://github.com/jrflab/modules/) [174]. We used OncoKB and Hotspot annotations for variant significance [190–192]. variants of significance in genes involved in HRD-DDR to assign patients to the HRD subtype. Patients with high-confidence dominant signature 3 or with six or more LSTs were assigned to the HRD group in all cases. Patients with low-confidence dominant signature 3 or at least one significant variant in the HRD-DDR genes [171] were assigned to the HRD subtype, except when there was evidence that patients belonged to the foldback inversion- or tandem duplicator-enriched subgroups (via *CCNE1* amplification or *CDK12* SNVs, specifically) [56,160]: these patients with conflicting evidence were assigned to the ambiguous subtype. Patients with available results from clinical HRD-DDR panels (n=47) or *BRCA1/2* sendout panels (n=2) were assigned HRP unless there were variants of known significance (as determined by the test provider) in at least one reported gene.

In the test cohort, we downloaded SNV data from the TCGA-OV project on cBioPortal for an expanded set of genes implicated in HRD-DDR [172] (because these genes are profiled in WES) and *CCNE1* amplifications, again filtering to variants deemed significant by OncoKB Using these criteria, patients with at least one SNV in HRD-DDR genes or SBS3 frequency greater than 15% were assigned the HRD subtype. Patients without aberrations in these HRD-DDR-associated genes were assigned the HRP subtype. Patients with an SNV in *CDK12* or CNA in *CCNE1* and also with an SNV in at least one of the HRD-DDR genes or SBS3 frequency greater than 15% were assigned the ambiguous subtype. Patients without available SNV and CNA data in cBioPortal were

assigned to the ambiguous subtype and excluded. We also downloaded COSMIC SBS3 frequencies [175] from Synapse (syn11801889), which is clearly bimodal (S. Fig. 3c), but we found that only 9 patients in our imaging-associated cohort were in the SBS3-high group, and 89 patients were found to be in the SBS3-low group. This distribution was more unbalanced than expected, and stratification based on SBS3 status was weak (PFS c=0.52, OS c=0.51). Hence, SBS3 activity was not used for HRD status assignment.

Adnexal and omental lesions segmentation

by Yulia Lakhman, Emily Aherne, and Ines Nikolovski

Three expert radiologists manually segmented ovarian lesions and representative omental implants on each pretreatment CE-CT for all patients in the internal discovery and in the external TCGA test cohorts. Using the Insight Segmentation and Registration Toolkit–SNAP version 3.8.0 software, each radiologist traced the outer contour of ovarian and omental lesions on every tumor-containing axial slice. All questions that arose during segmentation were resolved via joint review and consensus.

Radiologic feature extraction and selection

We converted all DICOM series to volumetric images in Hounsfield Units and applied an abdominal window (level 50, width 400). Using PyRadiomics [193], we resampled images to isotropic 1mm^3 voxels using the Simple ITK B-spline interpolator and binned images with bin size of 25 HU. We extracted features in 3D from Coif wavelet- and Laplacian of Gaussian (with standard deviations of 1

and 3)-transformed images. We extracted features from the gray level size zone [178], neighboring gray tone difference [194], gray level run length [179], gray level dependence [180], and gray level co-occurrence [195] matrices, yielding a 750-dimensional representation of each study's representative omental lesion(s). For each feature, we fit a univariate Cox Proportional Hazards model to the full discovery cohort using the Python Lifelines package without regularization, and we plotted the univariate coefficient and significance confidence. For features whose model failed to converge, we re-attempted fitting with L2 regularization C=0.2, and any model still failing to converge was assigned a log Hazard Ratio of 0 and p value of 1. Given the goal of identifying top features associated with PFS in the cohort for a prognostic model (to be tested) rather than making claims about confidence intervals for individual features from the discovery cohort, no correction for multiple testing was used. We repeated this process ten times using bootstrapping (95% of the training set) to reduce the impact of uncommon patient phenotypes. We next chose the top 35 (of 750) features based on these average univariate log hazard ratios and calculated the variance inflation factor (VIF), iteratively removing the feature with the highest-valued VIF until no VIF exceeded 3, yielding a radiomic signature with low multicollinearity.

Histopathologic annotation

by Lora Ellenson and Rob Soslow

Expert pathologists partially annotated 65 H&E WSIs using the MSK Slide Viewer [196]. The approach was to label example regions of necrosis, lymphocyte-rich tumor, lymphocyte-poor tumor, lymphocyte-rich stroma, lymphocyte-poor stroma,

veins, arteries, and fat with reasonable but imperfect accuracy. We exported these annotations as bitmaps and converted them to GeoJSON objects. We amalgamated lymphocyte-rich/poor tumor labels and lymphocyte-rich/poor stroma labels for training and omitted vessels from the training data for the models presented in this work. We next used these annotations to generate tissue-type tiles.

Training the histopathologic tissue type classifier

We generated tiles measuring 64μm x 64μm with 50% overlap, using the annotations to delineate regions to be tiled. Putative tile squares within an annotation but with <20% foreground as assessed by Otsu's method were not tiled. No computational stain normalization was used. We trained a ResNet-18 model (pretrained on ImageNet) for 30 epochs with a learning rate of 5e-4, 1e-4 L2 regularization, and the Adam optimizer. The objective function was class-balanced cross entropy, and we used mini batches of 96 tiles on a single NVIDIA Tesla V100 GPU. We used four-fold, slide-wise cross-validation for model evaluation and hyperparameter tuning. We selected the number of epochs to train the final model using the epoch with the highest lower 95% C.I. bound estimated using the mean and standard deviation of the cross-validation F1 scores. We trained the model on tiles from all 65 slides for 18 epochs.

Histopathologic feature extraction and selection

We tiled the 142 WSIs associated with the patients in this cohort without overlap, performing inference using mini batches of 800 across four NVIDIA Tesla V100 GPUs. We used Macenko stain normalization for external slides because

staining intensity differences from our MSKCC-based training cohort confounded inference. We assembled tile predictions into downscaled bitmaps, which were then used to calculate tissue-type features in an approach based on [169]. We included the region properties from scikit-image [197] for both the largest connected component and the entirety of each tissue type. We also calculated features such as the area ratio of one tissue type to another and the entropy of tumor and stroma. Using the StarDist method [198] for QuPath [199], we segmented and characterized individual nuclei, using nuclei with a detection probability greater than 0.5. We used a lymphocyte classifier trained iteratively using manual annotations to distinguish lymphocytes from other cells. We assigned a tissue parent type to each nucleus using the inferred tissue type maps and calculated aggregative statistics by tissue type and cell type of the QuPath-extracted nuclear morphologic and staining features, such as variance in eosin staining or circularity. Together, these cell type features and tissue type features constitute the histopathologic embedding for each slide. To mitigate the effect of extreme outliers, we replaced feature values with a Z-score greater than five with the median of the respective feature. To select features, we again modeled each feature as for the radiomic features using Cox Proportional Hazards models, and we controlled for the relative specimen size by including the scaled number of foreground tiles per slide. We chose the top four prognostic features, based on good performance on cross validation in the discovery cohort.

Survival modeling

We used linear Cox Proportional Hazards models with L2 regularization (C=0.5) and no L1 regularization for all multimodal models and for all unimodal models except for histopathologic, where we used no regularization. We chose a late fusion approach to increase unimodal sample sizes available for parameter estimation. Parameters for unimodal sub-models were estimated using all available unimodal data, and late fusion parameters were estimated for a multivariate Cox model integrating each unimodal sub-model's score using only the intersection set of patients. Radiologic and histopathologic features were chosen using the discovery cohort. No sub-model was fit for the genomic modality: patients assigned to the HRP subtype were designated high risk (risk score=1.0), and patients assigned to the HRD subtype were designated low risk (risk score=0.0). Interaction terms were used between each imaging score and the genomic score for the GHR model. The clinical sub-model for the internal test set was fit on the binary variable representing whether complete gross resection was achieved during delayed primary surgery, whether neoadjuvant PARP inhibitors were administered, the scaled number of cycles of neoadjuvant chemotherapy administered, scaled patient age at diagnosis, and pathologic stage by encoding stage into supergroups III vs IV. The clinical sub-model for testing was fit only on pathologic stage and patient age because the other variables were not available for the TCGA cases. We used Kaplan Meier analysis to determine whether each model stratified patients into clinically significant groups, examining both a three-group (highest risk 33%, middle risk 33%, lowest

risk 33%) and two-group (highest and lowest 50%) splitting strategy. P values for concordance indices were calculated using 1000-fold permutation tests. All p values for Kaplan-Meier analysis were calculated by comparing the highest risk group (as determined by the model's inferred risk score) to the lowest risk group using the log-rank test, except in **Figure 10g**, where it was calculated by comparing the aggregated HRD curve to the aggregated HRP curve. P values for covariate significance in Cox Proportional Hazards models are reported for models fit with C=0.0.

APPENDICES

Appendix 1. Glossary

Artificial intelligence: Artificial intelligence (AI) is a broad field of computer science concerned with developing computational tools to carry out tasks historically requiring human-level intelligence.

AUROC: The area under the receiver operating characteristic curve (AUROC) measures the ability of a binary classifier to separate the populations of interest. It describes the increase in true positive rate relative to the increase in false positive rate over the range of score thresholds chosen to separate the two classes.

Autoencoder**:** An autoencoder is an unsupervised neural network architecture trained to represent data in a lower dimensional space. It is a form of lossy compression that can be used to uncover latent structure in the data or reduce computational needs before further analysis.

Bayesian inference: Bayesian inference is a statistical method that refers to the application of Bayes' Theorem in determining the updated probability of a hypothesis given new information. Bayesian inference allows the posterior probability to be calculated given the prior probability of a hypothesis and a likelihood function.

Biomarker: A measurement which indicates a biological state. Cancer biomarkers can be categorized into diagnostic (disease progression), predictive (treatment response), and prognostic (survival).

C-Index: The concordance index, or c-Index, generalizes the AUROC to measure the ability of a model to separate censored data [200]. As with the AUROC, the baseline value for a model with arbitrary predictions is 0.5, and the ceiling value for a perfect prediction model is 1.0.

Client server model: The client server model describes a framework for computer network communications in which a computer system called a server provides requested services to more than one computer or program, called a client. These two components interact with each other through a unidirectional network connection using a given protocol. The client-server model has become the predominant framework for providing services like email and internet access to multiple clients.

Convolutional neural networks: Convolutional neural networks (CNNs) are a form of DNNs typically used to analyze images. CNNs are named for their use of convolutions, a mathematical operation involving the input data and a smaller matrix known as a kernel. This parameter sharing reduces the number of parameters to be learned and encourages the learning of features which are invariant to image shifts.

Computer vision**:** Computer vision is an interdisciplinary scientific field which attempts to find high-level representations from a series or individual digital images. Computer vision models often attempt to perform tasks normally performed by a human visual system.

Counterfactual machine learning: Counterfactual ML is a set of techniques for interpretable ML. For example, a counterfactual analysis could involve using a

model developed to predict a disease outcome using a set of measurements to predict scenarios where the input measurements are perturbed to study their causal relationship. This paradigm has also been harnessed to learn unbiased recommenders from logged data, such as user purchases on online marketplaces, despite changes in how products are recommended over time and the lack of a controlled experimental setup.

Cox proportional hazards models: Cox proportional hazards (CPH) models are regression models used to associate censored temporal outcomes, such as time to survival, and potential predictor variables, such as age or cancer stage. It is the most common method to evaluate prognostic variables in cancer patient survival analyses.

Deep Learning: Deep learning (DL) comprises a class of ML methods based on artificial neural networks (ANN), which use multiple non-linear layers to derive progressively higher-order features from data.

Data lake: A data lake stores relational and non-relational data from a vast pool of raw data. The structure of the data or schema is not defined when data is captured. Different types of analytics on data like SQL queries, big data analytics, full text search, real-time analytics, and machine learning can be used to uncover insights.

Data parallelism**:** Data parallelism is the approach of performing a computing task in parallel utilizing multiple processors. It focuses on distributing data across various cores and enabling simultaneous sub-computations.

DICOM headers: The Digital Imaging and Communications in Medicine (DICOM) was developed jointly by American College of Radiology (ACR) and National Electrical Manufacturers Association (NEMA) to aid the distribution and viewing of medical images, such as CT scans, MRIs, and ultrasound. A single DICOM file contains both a header that stores information about the patient, scan parameters as well as all of the image data that contain information in three dimensions.

Deep neural networks: Deep neural networks (DNNs) are a form of deep learning, namely artificial neural networks with more than one hidden layer between the input and output layers.

Graph convolutional networks: Graph convolutional networks (GCN) learn convolutions on data structures known as graphs. Graphs are defined as a set of objects (nodes) and their corresponding relationships (edges).

K-means clustering: K-means clustering is an unsupervised machine learning algorithm which aims to partition data into k clusters.

Layer-wise relevance propagation: Layer-wise Relevance Propagation (LRP) is one of the most prominent techniques in explainable machine learning. LRP decomposes the network's output score into the individual contributions of the input neurons using model parameters (i.e., weights) and neuron activations.

Machine learning: Machine learning (ML) is a type of artificial intelligence which aims to discover patterns in data which are not explicitly programmed. ML models typically use a dataset for pattern discovery, known as “training”, to make predictions on unseen data, known as “inference”.

Natural language processing**:** Natural Language Processing (NLP) is a branch of artificial intelligence that deals with the analysis of natural language data. NLP techniques rely on machine learning algorithms to identify and extract the natural language rules to process unstructured language data. NLP adds useful numeric structure to the data for many downstream applications, such as speech recognition or text analytics.

Recommender systems: Recommender systems aim to predict relevant items to users by building a model from past behavior. In precision medicine, recommender systems can be used to predict the preferred treatment for a disease based on multiple patient measurements.

Recurrent neural networks: Recurrent neural networks (RNNs) are a form of DNNs optimized for time series data. An RNN analyzes each element of the input sequence in succession and updates its representation of the data based on previous elements.

TRACERx: TRAcking Cancer Evolution through therapy Rx (TRACERx) is a prospective cohort study designed to integrate clinical and genomic data to assess intratumor heterogeneity (ITH) and its evolution from diagnosis through relapse in early-stage non-small cell lung cancer.

Unsupervised learning**:** Unsupervised learning is a class of machine learning algorithms which aim to identify patterns in a dataset without assigning predefined labels or categories.

Voxel: A voxel, or “volume element” is the 3D equivalent of a picture element (pixel) in 3D space used by medical imaging modalities. Its dimensions are given by the pixel, together with the thickness of the slice.

Appendix 2. Deep learning architectures

ML can be divided broadly into unsupervised learning and supervised learning. Unsupervised learning seeks to discover intrinsic patterns in data, sometimes without known labels for each data point, while supervised learning seeks to predict a label of interest from the input data. DL is a subtype of ML that has the potential to learn more informative features than engineered features, but there is difficulty in model interpretability and performance is notoriously dependent on the amount of training data available [14]. No ML algorithm is universally superior to another, but the data and targets to be related motivate the choice of model [201,202]. With sufficient training data, DNNs have become a leading approach to capture salient patterns within data. DNNs are universal function approximators that learn a distributed representation of given data, with deep features often describing data better than competing human-defined features [203]. Though these methods are limited by the need for large training datasets and the difficulty of interpreting their learned features, they are indispensable for discovering highly informative features in clinical datasets.

Specific variants of DNNs exist for different data modalities. For example, CNNs learn sliding window-like kernels to detect textural patterns within images, often achieving or exceeding human performance in image classification. Some of the most popular variants, available off the shelf in modern DL frameworks, are ResNet, Inception, DenseNet, and SqueezeNet [204–207]. For sequential data such as time series of lab values, RNNs can be the architecture of choice. The RNN uses each data point to update its understanding of the data, building an

amalgamated representation that is then used to predict the outcome of interest, such as risk of disease recurrence. The most successful variants are LSTM and gated recurrent unit (GRU) networks [208,209]. Though RNNs have not yet been widely applied in oncology, preliminary studies of RNNs for longitudinal medical event prediction have yielded promising results [76,77,210]. For high-dimensional data such as transcriptomic profiles, the attention gating mechanisms inherent in deep highway networks [78] have helped identify salient features amidst potentially uninformative background [79].

BIBLIOGRAPHY

1. AACR Project GENIE Consortium. AACR Project GENIE: Powering Precision Medicine through an International Consortium. *Cancer Discov.* **7**, 818–831 (2017).
2. Vasan, N. *et al.* Double PIK3CA mutations in cis increase oncogenicity and sensitivity to PI3Kα inhibitors. *Science* **366**, 714–723 (2019).
3. Razavi, P. *et al.* The Genomic Landscape of Endocrine-Resistant Advanced Breast Cancers. *Cancer Cell* **34**, 427-438.e6 (2018).
4. Jonsson, P. *et al.* Genomic Correlates of Disease Progression and Treatment Response in Prospectively Characterized Gliomas. *Clin. Cancer Res.* **25**, 5537–5547 (2019).
5. Soumerai, T. E. *et al.* Clinical Utility of Prospective Molecular Characterization in Advanced Endometrial Cancer. *Clin. Cancer Res.* **24**, 5939–5947 (2018).
6. Cui, M. & Zhang, D. Y. Artificial intelligence and computational pathology. *Lab. Invest.* (2021) doi:10.1038/s41374-020-00514-0.
7. Shen, S. Y. *et al.* Sensitive tumour detection and classification using plasma cell-free DNA methylomes. *Nature* **563**, 579–583 (2018).
8. Cristiano, S. *et al.* Genome-wide cell-free DNA fragmentation in patients with cancer. *Nature* **570**, 385–389 (2019).
9. Klupczynska, A. *et al.* Study of early stage non-small-cell lung cancer using Orbitrap-based global serum metabolomics. *J. Cancer Res. Clin. Oncol.* **143**, 649–659 (2017).
10. Helland, T. *et al.* Serum concentrations of active tamoxifen metabolites predict long-term survival in adjuvantly treated breast cancer patients. *Breast Cancer Res.* **19**, 125 (2017).
11. Luo, P. *et al.* A Large-scale, multicenter serum metabolite biomarker identification study for the early detection of hepatocellular carcinoma: Luo, Yin, et al. *Hepatology* **67**, 662–675 (2018).
12. Medina-Martínez, J. S. *et al.* Isabl Platform, a digital biobank for processing multimodal patient data. *BMC Bioinformatics* **21**, 549 (2020).
13. Bhinder, B., Gilvary, C., Madhukar, N. S. & Elemento, O. Artificial Intelligence in Cancer Research and Precision Medicine. *Cancer Discov.* **11**, 900–915 (2021).
14. Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H. & Aerts, H. J. W. L. Artificial intelligence in radiology. *Nat. Rev. Cancer* **18**, 500–510 (2018).
15. Bera, K., Schalper, K. A., Rimm, D. L., Velcheti, V. & Madabhushi, A. Artificial intelligence in digital pathology — new tools for diagnosis and precision oncology. *Nature Reviews Clinical Oncology* vol. 16 703–715 (2019).
16. Gutman, D. A. *et al.* MR imaging predictors of molecular profile and survival: multi-institutional study of the TCGA glioblastoma data set. *Radiology* **267**, 560–569 (2013).

17. Zwanenburg, A. *et al.* The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. *Radiology* **295**, 328–338 (2020).
18. Sun, R. *et al.* A radiomics approach to assess tumour-infiltrating CD8 cells and response to anti-PD-1 or anti-PD-L1 immunotherapy: an imaging biomarker, retrospective multicohort study. *Lancet Oncol.* **19**, 1180–1191 (2018).
19. Rizzo, S. *et al.* Radiomics of high-grade serous ovarian cancer: association between quantitative CT features, residual tumour and disease progression within 12 months. *Eur. Radiol.* **28**, 4849–4859 (2018).
20. Pisapia, J. M. *et al.* Predicting pediatric optic pathway glioma progression using advanced magnetic resonance image analysis and machine learning. *Neurooncol Adv* **2**, vdaa090 (2020).
21. Chang, K. *et al.* Residual Convolutional Neural Network for the Determination of IDH Status in Low- and High-Grade Gliomas from MR Imaging. *Clin. Cancer Res.* **24**, 1073–1081 (2018).
22. Li, Z., Wang, Y., Yu, J., Guo, Y. & Cao, W. Deep Learning based Radiomics (DLR) and its usage in noninvasive IDH1 prediction for low grade glioma. *Scientific Reports* vol. 7 (2017).
23. Lu, C.-F. *et al.* Machine Learning-Based Radiomics for Molecular Subtyping of Gliomas. *Clin. Cancer Res.* **24**, 4429–4436 (2018).
24. Wang, S. *et al.* Predicting EGFR mutation status in lung adenocarcinoma on computed tomography image using deep learning. *Eur. Respir. J.* **53**, (2019).
25. Khosravi, P., Lysandrou, M., Eljalby, M. & Li, Q. A Deep Learning Approach to Diagnostic Classification of Prostate Cancer Using Pathology–Radiology Fusion. *J. Magn. Reson.* (2021).
26. Rajpurkar, P. *et al.* AppendiXNet: Deep Learning for Diagnosis of Appendicitis from A Small Dataset of CT Exams Using Video Pretraining. *Sci. Rep.* **10**, 3958 (2020).
27. Hosny, A. *et al.* Deep learning for lung cancer prognostication: A retrospective multi-cohort radiomics study. *PLoS Med.* **15**, e1002711 (2018).
28. Khosravi, P., Kazemi, E., Imielinski, M., Elemento, O. & Hajirasouliha, I. Deep Convolutional Neural Networks Enable Discrimination of Heterogeneous Digital Pathology Images. *EBioMedicine* **27**, 317–328 (2018).
29. Coudray, N. *et al.* Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nature Medicine* vol. 24 1559–1567 (2018).
30. Fu, Y. *et al.* Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *Nature Cancer* **1**, 800–810 (2020).
31. Kather, J. N. *et al.* Pan-cancer image-based detection of clinically actionable genetic alterations. *Nature Cancer* **1**, 789–799 (2020).
32. Ding, K. *et al.* Feature-Enhanced Graph Networks for Genetic Mutational Prediction Using Histopathological Images in Colon Cancer. in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020* 294–304 (Springer International Publishing, 2020).

33. Rutledge, W. C. *et al.* Tumor-infiltrating lymphocytes in glioblastoma are associated with specific genomic alterations and related to transcriptional class. *Clin. Cancer Res.* **19**, 4951–4960 (2013).
34. Kather, J. N. *et al.* Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nat. Med.* **25**, 1054–1056 (2019).
35. Echle, A. *et al.* Clinical-grade Detection of Microsatellite Instability in Colorectal Tumors by Deep Learning. *Gastroenterology* (2020) doi:10.1053/j.gastro.2020.06.021.
36. 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).
37. Diao, J. A. *et al.* Dense, high-resolution mapping of cells and tissues from pathology images for the interpretable prediction of molecular phenotypes in cancer. 2020.08.02.233197 (2020) doi:10.1101/2020.08.02.233197.
38. Corredor, G. *et al.* Spatial Architecture and Arrangement of Tumor-Infiltrating Lymphocytes for Predicting Likelihood of Recurrence in Early-Stage Non-Small Cell Lung Cancer. *Clin. Cancer Res.* **25**, 1526–1534 (2019).
39. AbdulJabbar, K. *et al.* Geospatial immune variability illuminates differential evolution of lung adenocarcinoma. *Nat. Med.* **26**, 1054–1062 (2020).
40. Kong, J. *et al.* Machine-based morphologic analysis of glioblastoma using whole-slide pathology images uncovers clinically relevant molecular correlates. *PLoS One* **8**, e81049 (2013).
41. Flaherty, K. T. *et al.* Improved survival with MEK inhibition in BRAF-mutated melanoma. *N. Engl. J. Med.* **367**, 107–114 (2012).
42. Maemondo, M. *et al.* Gefitinib or chemotherapy for non-small-cell lung cancer with mutated EGFR. *N. Engl. J. Med.* **362**, 2380–2388 (2010).
43. Slamon, D. J. *et al.* Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. *N. Engl. J. Med.* **344**, 783–792 (2001).
44. DiNardo, C. D. *et al.* Durable Remissions with Ivosidenib in IDH1-Mutated Relapsed or Refractory AML. *N. Engl. J. Med.* **378**, 2386–2398 (2018).
45. Mirza, M. R. *et al.* Niraparib Maintenance Therapy in Platinum-Sensitive, Recurrent Ovarian Cancer. *N. Engl. J. Med.* **375**, 2154–2164 (2016).
46. de Bono, J. *et al.* Olaparib for Metastatic Castration-Resistant Prostate Cancer. *N. Engl. J. Med.* **382**, 2091–2102 (2020).
47. Drilon, A. *et al.* Efficacy of Larotrectinib in TRK Fusion-Positive Cancers in Adults and Children. *N. Engl. J. Med.* **378**, 731–739 (2018).
48. Canon, J. *et al.* The clinical KRAS(G12C) inhibitor AMG 510 drives anti-tumour immunity. *Nature* **575**, 217–223 (2019).
49. Hallin, J. *et al.* The KRASG12C Inhibitor MRTX849 Provides Insight toward Therapeutic Susceptibility of KRAS-Mutant Cancers in Mouse Models and Patients. *Cancer Discov.* **10**, 54–71 (2020).
50. André, F., Ciruelos, E. & Rubovszky, G. Alpelisib for PIK3CA-Mutated, Hormone Receptor–Positive Advanced Breast Cancer. *England Journal of …* (2019).

51. Samstein, R. M. *et al.* Tumor mutational load predicts survival after immunotherapy across multiple cancer types. *Nat. Genet.* **51**, 202–206 (2019).
52. Le, D. T. *et al.* Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade. *Science* **357**, 409–413 (2017).
53. Priestley, P. *et al.* Pan-cancer whole-genome analyses of metastatic solid tumours. *Nature* **575**, 210–216 (2019).
54. Vöhringer, H., van Hoeck, A., Cuppen, E. & Gerstung, M. Learning mutational signatures and their multidimensional genomic properties with TensorSignatures. doi:10.1101/850453.
55. Macintyre, G. *et al.* Copy number signatures and mutational processes in ovarian carcinoma. *Nat. Genet.* **50**, 1262–1270 (2018).
56. Funnell, T. *et al.* Integrated structural variation and point mutation signatures in cancer genomes using correlated topic models. *PLoS Comput. Biol.* **15**, e1006799 (2019).
57. Liu, Y. *et al.* High-Spatial-Resolution Multi-Omics Sequencing via Deterministic Barcoding in Tissue. *Cell* **183**, 1665-1681.e18 (2020).
58. Maniatis, S. *et al.* Spatiotemporal dynamics of molecular pathology in amyotrophic lateral sclerosis. *Science* **364**, 89–93 (2019).
59. Payne, A. C. *et al.* In situ genome sequencing resolves DNA sequence and structure in intact biological samples. *Science* **371**, (2021).
60. Zadeh, A., Chen, M., Poria, S., Cambria, E. & Morency, L.-P. Tensor Fusion Network for Multimodal Sentiment Analysis. *arXiv [cs.CL]* (2017).
61. Liu, K., Li, Y., Xu, N. & Natarajan, P. Learn to Combine Modalities in Multimodal Deep Learning. *arXiv [stat.ML]* (2018).
62. Wang, W., Tran, D. & Feiszli, M. What makes training multi-modal classification networks hard? in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 12695–12705 (2020).
63. Raghu, M., Zhang, C., Kleinberg, J. & Bengio, S. Transfusion: Understanding Transfer Learning for Medical Imaging. *arXiv [cs.CV]* (2019).
64. Zhang, L. *et al.* Deep Learning-Based Multi-Omics Data Integration Reveals Two Prognostic Subtypes in High-Risk Neuroblastoma. *Front. Genet.* **9**, 477 (2018).
65. Chaudhary, K., Poirion, O. B., Lu, L. & Garmire, L. X. Deep Learning-Based Multi-Omics Integration Robustly Predicts Survival in Liver Cancer. *Clin. Cancer Res.* **24**, 1248–1259 (2018).
66. Ramazzotti, D., Lal, A., Wang, B., Batzoglou, S. & Sidow, A. Multi-omic tumor data reveal diversity of molecular mechanisms that correlate with survival. *Nat. Commun.* **9**, 4453 (2018).
67. Poirion, O. B., Chaudhary, K. & Garmire, L. X. Deep Learning data integration for better risk stratification models of bladder cancer. *AMIA Jt Summits Transl Sci Proc* **2017**, 197–206 (2018).
68. Žitnik, M. & Zupan, B. Survival regression by data fusion. *Systems Biomedicine* **2**, 47–53 (2014).

69. Cancer Genome Atlas Research Network *et al.* Comprehensive, Integrative Genomic Analysis of Diffuse Lower-Grade Gliomas. *N. Engl. J. Med.* **372**, 2481–2498 (2015).
70. Cantini, L. *et al.* Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer. *Nat. Commun.* **12**, 124 (2021).
71. Stuart, T. & Satija, R. Integrative single-cell analysis. *Nat. Rev. Genet.* **20**, 257–272 (2019).
72. Hasin, Y., Seldin, M. & Lusis, A. Multi-omics approaches to disease. *Genome Biol.* **18**, 83 (2017).
73. Sun, D., Wang, M. & Li, A. A multimodal deep neural network for human breast cancer prognosis prediction by integrating multi-dimensional data. *IEEE/ACM Trans. Comput. Biol. Bioinform.* (2018) doi:10.1109/TCBB.2018.2806438.
74. Huang, Z. *et al.* SALMON: Survival Analysis Learning With Multi-Omics Neural Networks on Breast Cancer. *Front. Genet.* **10**, 166 (2019).
75. Lee, B. *et al.* DeepBTS: Prediction of Recurrence-free Survival of Non-small Cell Lung Cancer Using a Time-binned Deep Neural Network. *Sci. Rep.* **10**, 1952 (2020).
76. Choi, E., Bahadori, M. T., Schuetz, A., Stewart, W. F. & Sun, J. Doctor AI: Predicting Clinical Events via Recurrent Neural Networks. *JMLR Workshop Conf. Proc.* **56**, 301–318 (2016).
77. Tomašev, N. *et al.* A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature* **572**, 116–119 (2019).
78. Srivastava, R. K., Greff, K. & Schmidhuber, J. Highway Networks. *arXiv [cs.LG]* (2015).
79. Cheerla, A. & Gevaert, O. Deep learning with multimodal representation for pancancer prognosis prediction. *Bioinformatics* **35**, i446–i454 (2019).
80. Gevaert, O. *et al.* Imaging-AMARETTO: An Imaging Genomics Software Tool to Interrogate Multiomics Networks for Relevance to Radiography and Histopathology Imaging Biomarkers of Clinical Outcomes. *JCO Clin Cancer Inform* **4**, 421–435 (2020).
81. Zhu, X. *et al.* Imaging-genetic data mapping for clinical outcome prediction via supervised conditional Gaussian graphical model. *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (2016) doi:10.1109/bibm.2016.7822559.
82. Popovici, V. *et al.* Joint analysis of histopathology image features and gene expression in breast cancer. *BMC Bioinformatics* **17**, 209 (2016).
83. Mobadersany, P. *et al.* Predicting cancer outcomes from histology and genomics using convolutional networks. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E2970–E2979 (2018).
84. 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* **PP**, (2020).
85. Yuan, Y., Giger, M. L., Li, H., Bhooshan, N. & Sennett, C. A. Multimodality computer-aided breast cancer diagnosis with FFDM and DCE-MRI. *Acad. Radiol.* **17**, 1158–1167 (2010).

86. Chan, H.-W., Weng, Y.-T. & Huang, T.-Y. Automatic Classification of Brain Tumor Types with the MRI Scans and Histopathology Images. 353–359 (2020).
87. Rathore, S. *et al.* Radiomic MRI signature reveals three distinct subtypes of glioblastoma with different clinical and molecular characteristics, offering prognostic value beyond IDH1. *Sci. Rep.* **8**, 5087 (2018).
88. Donini, M. *et al.* Combining heterogeneous data sources for neuroimaging based diagnosis: re-weighting and selecting what is important. *Neuroimage* **195**, 215–231 (2019).
89. Gonen, M. & Alpaydin, E. Multiple Kernel Learning Algorithms. https://www.jmlr.org/papers/volume12/gonen11a/gonen11a.pdf (2011).
90. Gillies, R. J., Kinahan, P. E. & Hricak, H. Radiomics: Images Are More than Pictures, They Are Data. *Radiology* **278**, 563–577 (2016).
91. Deng, J. *et al.* ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition* (2009) doi:10.1109/cvpr.2009.5206848.
92. Kay, W. *et al.* The Kinetics Human Action Video Dataset. *arXiv [cs.CV]* (2017).
93. Ardila, D. *et al.* End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nat. Med.* **25**, 954–961 (2019).
94. Zhou, Z.-H. A brief introduction to weakly supervised learning. *Natl Sci Rev* **5**, 44–53 (2018).
95. Campanella, G. *et al.* Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat. Med.* **25**, 1301–1309 (2019).
96. Suphavilai, C., Bertrand, D. & Nagarajan, N. Predicting Cancer Drug Response using a Recommender System. *Bioinformatics* **34**, 3907–3914 (2018).
97. Joachims, T., Swaminathan, A. & de Rijke, M. Deep Learning with Logged Bandit Feedback. (2018).
98. Gundersen, G., Dumitrascu, B., Ash, J. T. & Engelhardt, B. E. End-to-end training of deep probabilistic CCA for joint modeling of paired biomedical observations.
99. Li, Y. *et al.* Inferring multimodal latent topics from electronic health records. *Nat. Commun.* **11**, 2536 (2020).
100. Hersh, W. R. *et al.* Caveats for the use of operational electronic health record data in comparative effectiveness research. *Med. Care* **51**, S30-7 (2013).
101. Allison, J. J. *et al.* The Art and Science of Chart Review. *Jt. Comm. J. Qual. Improv.* **26**, 115–136 (2000).
102. Vassar, M. & Holzmann, M. The retrospective chart review: important methodological considerations. *J. Educ. Eval. Health Prof.* **10**, 12 (2013).
103. Lee, D., de Keizer, N., Lau, F. & Cornet, R. Literature review of SNOMED CT use. *J. Am. Med. Inform. Assoc.* **21**, e11-9 (2014).

104. Stein, B. & Morrison, A. The enterprise data lake: Better integration and deeper analytics. *PwC Technology Forecast: Rethinking integration* **1**, 18 (2014).
105. Weigelt, B. *et al.* Radiogenomics Analysis of Intratumor Heterogeneity in a Patient With High-Grade Serous Ovarian Cancer. *JCO Precision Oncology* 1–9 (2019).
106. Jiménez-Sánchez, A. *et al.* Unraveling tumor-immune heterogeneity in advanced ovarian cancer uncovers immunogenic effect of chemotherapy. *Nat. Genet.* **52**, 582–593 (2020).
107. Johnson, A. E. W. *et al.* MIMIC-III, a freely accessible critical care database. *Sci Data* **3**, 160035 (2016).
108. Hripcsak, G. *et al.* Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. *Stud. Health Technol. Inform.* **216**, 574–578 (2015).
109. Rieke, N. *et al.* The future of digital health with federated learning. *npj Digital Medicine* **3**, 119 (2020).
110. Willemink, M. J. *et al.* Preparing Medical Imaging Data for Machine Learning. *Radiology* **295**, 4–15 (2020).
111. Andreux, M., Manoel, A., Menuet, R., Saillard, C. & Simpson, C. Federated Survival Analysis with Discrete-Time Cox Models. *arXiv [cs.LG]* (2020).
112. Lin, J.-H. & Haug, P. J. Exploiting missing clinical data in Bayesian network modeling for predicting medical problems. *J. Biomed. Inform.* **41**, 1–14 (2008).
113. Khan, A., Atzori, M., Otálora, S., Andrearczyk, V. & Müller, H. Generalizing convolution neural networks on stain color heterogeneous data for computational pathology. in *Medical Imaging 2020: Digital Pathology* vol. 11320 113200R (International Society for Optics and Photonics, 2020).
114. Glatz-Krieger, K., Spornitz, U., Spatz, A., Mihatsch, M. J. & Glatz, D. Factors to keep in mind when introducing virtual microscopy. *Virchows Arch.* **448**, 248–255 (2006).
115. Janowczyk, A., Basavanhally, A. & Madabhushi, A. Stain Normalization using Sparse AutoEncoders (StaNoSA): Application to digital pathology. *Comput. Med. Imaging Graph.* **57**, 50–61 (2017).
116. Lacroix, M. *et al.* Correction for Magnetic Field Inhomogeneities and Normalization of Voxel Values Are Needed to Better Reveal the Potential of MR Radiomic Features in Lung Cancer. *Front. Oncol.* **10**, 43 (2020).
117. Macenko, M. *et al.* A method for normalizing histology slides for quantitative analysis. in *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro* (IEEE, 2009). doi:10.1109/isbi.2009.5193250.
118. Srinidhi, C. L., Ciga, O. & Martel, A. L. Deep neural network models for computational histopathology: A survey. *Med. Image Anal.* **67**, 101813 (2021).
119. Hu, Z., Tang, A., Singh, J., Bhattacharya, S. & Butte, A. J. A robust and interpretable end-to-end deep learning model for cytometry data. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 21373–21380 (2020).

120. Kleppe, A. *et al.* Designing deep learning studies in cancer diagnostics. *Nat. Rev. Cancer* **21**, 199–211 (2021).
121. Lopez, K., Fodeh, S. J., Allam, A., Brandt, C. A. & Krauthammer, M. Reducing Annotation Burden Through Multimodal Learning. *Frontiers in Big Data* **3**, 19 (2020).
122. Gundersen, O. E. & Kjensmo, S. State of the art: Reproducibility in artificial intelligence. in *Thirty-second AAAI conference on artificial intelligence* (2018).
123. Courtiol, P. *et al.* Deep learning-based classification of mesothelioma improves prediction of patient outcome. *Nat. Med.* **25**, 1519–1525 (2019).
124. McKinney, S. M. *et al.* Addendum: International evaluation of an AI system for breast cancer screening. *Nature* **586**, E19 (2020).
125. Haibe-Kains, B. *et al.* Transparency and reproducibility in artificial intelligence. *Nature* **586**, E14–E16 (2020).
126. Hosny, A. *et al.* ModelHub.AI: Dissemination Platform for Deep Learning Models. *arXiv [cs.LG]* (2019).
127. Beede, E. *et al.* A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy. in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* 1–12 (Association for Computing Machinery, 2020).
128. Cruz Rivera, S. *et al.* Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nat. Med.* **26**, 1351–1363 (2020).
129. Moher, D. *et al.* CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ* **340**, c869 (2010).
130. Liu, X. *et al.* Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat. Med.* **26**, 1364–1374 (2020).
131. Lauritsen, S. M. *et al.* Explainable artificial intelligence model to predict acute critical illness from electronic health records. *Nat. Commun.* **11**, 3852 (2020).
132. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* **1**, 206–215 (2019).
133. Pavel, M. A., Petersen, E. N., Wang, H., Lerner, R. A. & Hansen, S. B. Studies on the mechanism of general anesthesia. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 13757–13766 (2020).
134. Wang, F., Kaushal, R. & Khullar, D. Should health care demand interpretable artificial intelligence or accept “black box” medicine? (2020).
135. Clark, K. *et al.* The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *J. Digit. Imaging* **26**, 1045–1057 (2013).
136. Newitt, D. & Hylton, N. Single site breast DCE-MRI data and segmentations from patients undergoing neoadjuvant chemotherapy. *The Cancer Imaging Archive* (2016).

137. Castro, D. C., Walker, I. & Glocker, B. Causality matters in medical imaging. *Nat. Commun.* **11**, 3673 (2020).
138. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A. & Torralba, A. Learning deep features for discriminative localization. in *Proceedings of the IEEE conference on computer vision and pattern recognition* 2921–2929 (2016).
139. Olah, C. *et al.* The Building Blocks of Interpretability. *Distill* **3**, (2018).
140. Graziani, M., Andrearczyk, V. & Müller, H. Visualizing and interpreting feature reuse of pretrained CNNs for histopathology. *Irish Machine Vision and Image Processing (IMVIP)* (2019).
141. Burns, C., Thomason, J. & Tansey, W. Interpreting Black Box Models via Hypothesis Testing. *arXiv [stat.ML]* (2019) doi:10.1145/3412815.3416889.
142. Donoghue, M. T. A., Schram, A. M., Hyman, D. M. & Taylor, B. S. Discovery through clinical sequencing in oncology. *Nature Cancer* (2020) doi:10.1038/s43018-020-0100-0.
143. Kehl, K. L. *et al.* Assessment of Deep Natural Language Processing in Ascertaining Oncologic Outcomes From Radiology Reports. *JAMA Oncol* (2019) doi:10.1001/jamaoncol.2019.1800.
144. Harris, P. A. *et al.* Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. *J. Biomed. Inform.* **42**, 377–381 (2009).
145. Office for Civil Rights (OCR). Cloud Computing. https://www.hhs.gov/hipaa/for-professionals/special-topics/cloud-computing/index.html.
146. U.s, F. Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD)-discussion paper. (2019).
147. Sullivan, H. R. & Schweikart, S. J. Are Current Tort Liability Doctrines Adequate for Addressing Injury Caused by AI? *AMA J Ethics* **21**, E160-166 (2019).
148. Price, W. N., 2nd, Gerke, S. & Cohen, I. G. Potential Liability for Physicians Using Artificial Intelligence. *JAMA* (2019) doi:10.1001/jama.2019.15064.
149. Cosgriff, C. V., Stone, D. J., Weissman, G., Pirracchio, R. & Celi, L. A. The clinical artificial intelligence department: a prerequisite for success. *BMJ Health Care Inform* **27**, (2020).
150. Zadeh, A. *et al.* Memory Fusion Network for Multi-view Sequential Learning. *arXiv [cs.LG]* (2018).
151. Zadeh, A., Liang, P. P., Poria, S., Cambria, E. & Morency, L.-P. Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph. in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* 2236–2246 (Association for Computational Linguistics, 2018).
152. Zadeh, A. *et al.* Multi-attention Recurrent Network for Human Communication Comprehension. *Proc. Conf. AAAI Artif. Intell.* **2018**, 5642–5649 (2018).
153. Kumar, A., Srinivasan, K., Cheng, W.-H. & Zomaya, A. Y. Hybrid context enriched deep learning model for fine-grained sentiment analysis in textual

and visual semiotic modality social data. *Inf. Process. Manag.* **57**, 102141 (2020).
154. Liang, P. P., Zadeh, A. & Morency, L.-P. Multimodal Local-Global Ranking Fusion for Emotion Recognition. in *Proceedings of the 20th ACM International Conference on Multimodal Interaction* 472–476 (Association for Computing Machinery, 2018).
155. Liu, Z. *et al.* Efficient Low-rank Multimodal Fusion with Modality-Specific Factors. *arXiv [cs.AI]* (2018).
156. Cancer Stat Facts. https://seer.cancer.gov/statfacts/.
157. Moore, K. *et al.* Maintenance Olaparib in Patients with Newly Diagnosed Advanced Ovarian Cancer. *N. Engl. J. Med.* **379**, 2495–2505 (2018).
158. Gallagher, D. J. *et al.* Survival in epithelial ovarian cancer: a multivariate analysis incorporating BRCA mutation status and platinum sensitivity. *Ann. Oncol.* **22**, 1127–1132 (2011).
159. Gorodnova, T. V. *et al.* High response rates to neoadjuvant platinum-based therapy in ovarian cancer patients carrying germ-line BRCA mutation. *Cancer Lett.* **369**, 363–367 (2015).
160. Zhang, A. W. *et al.* Interfaces of Malignant and Immunologic Clonal Dynamics in Ovarian Cancer. *Cell* **173**, 1755-1769.e22 (2018).
161. Kobayashi, Y., Banno, K. & Aoki, D. Current status and future directions of ovarian cancer prognostic models. *J. Gynecol. Oncol.* **32**, e34 (2021).
162. Lu, H. *et al.* A mathematical-descriptor of tumor-mesoscopic-structure from computed-tomography images annotates prognostic- and molecular-phenotypes of epithelial ovarian cancer. *Nat. Commun.* **10**, 764 (2019).
163. Rizzo, S. *et al.* Computed Tomography Based Radiomics as a Predictor of Survival in Ovarian Cancer Patients: A Systematic Review. *Cancers* **13**, (2021).
164. Wei, W. *et al.* A Computed Tomography-Based Radiomic Prognostic Marker of Advanced High-Grade Serous Ovarian Cancer Recurrence: A Multicenter Study. *Front. Oncol.* **9**, 255 (2019).
165. Wang, S. *et al.* Deep learning provides a new computed tomography-based prognostic biomarker for recurrence prediction in high-grade serous ovarian cancer. *Radiother. Oncol.* **132**, 171–177 (2019).
166. Vargas, H. A. *et al.* A novel representation of inter-site tumour heterogeneity from pre-treatment computed tomography textures classifies ovarian cancers by clinical outcome. *Eur. Radiol.* **27**, 3991–4001 (2017).
167. Meier, A. *et al.* Association between CT-texture-derived tumor heterogeneity, outcomes, and BRCA mutation status in patients with high-grade serous ovarian cancer. *Abdominal Radiology* vol. 44 2040–2047 (2019).
168. Zargari, A. *et al.* Prediction of chemotherapy response in ovarian cancer patients using a new clustered quantitative image marker. *Physics in Medicine & Biology* vol. 63 155020 (2018).
169. 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).

170. 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).
171. Heeke, A. L. *et al.* Prevalence of Homologous Recombination–Related Gene Mutations Across Multiple Cancer Types. *JCO Precision Oncology* 1–13 (2018).
172. Riaz, N. *et al.* Pan-cancer analysis of bi-allelic alterations in homologous recombination DNA repair genes. *Nat. Commun.* **8**, 857 (2017).
173. Popova, T. *et al.* Ploidy and large-scale genomic instability consistently identify basal-like breast carcinomas with BRCA1/2 inactivation. *Cancer Res.* **72**, 5454–5462 (2012).
174. Mandelker, D. *et al.* The Landscape of Somatic Genetic Alterations in Breast Cancers from CHEK2 Germline Mutation Carriers. *JNCI Cancer Spectr* **3**, kz027 (2019).
175. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
176. Telli, M. L. *et al.* Homologous Recombination Deficiency (HRD) Score Predicts Response to Platinum-Containing Neoadjuvant Chemotherapy in Patients with Triple-Negative Breast Cancer. *Clin. Cancer Res.* **22**, 3764–3773 (2016).
177. Gulhan, D. C., Lee, J. J.-K., Melloni, G. E. M., Cortés-Ciriano, I. & Park, P. J. Detecting the mutational signature of homologous recombination deficiency in clinical samples. *Nat. Genet.* **51**, 912–919 (2019).
178. Thibault, G. *et al.* SHAPE AND TEXTURE INDEXES APPLICATION TO CELL NUCLEI CLASSIFICATION. *Int. J. Pattern Recognit Artif Intell.* **27**, 1357002 (2013).
179. Galloway, M. M. Texture analysis using gray level run lengths. *Computer Graphics and Image Processing* **4**, 172–179 (1975).
180. Sun, C. & Wee, W. G. Neighboring gray level dependence matrix for texture classification. *Computer Vision, Graphics, and Image Processing* **23**, 341–352 (1983).
181. Beylkin, G., Coifman, R. & Rokhlin, V. Fast wavelet transforms and numerical algorithms I. *Communications on Pure and Applied Mathematics* vol. 44 141–183 (1991).
182. Bowtell, D. D. *et al.* Rethinking ovarian cancer II: reducing mortality from high-grade serous ovarian cancer. *Nat. Rev. Cancer* **15**, 668–679 (2015).
183. Bredholt, G. *et al.* Tumor necrosis is an important hallmark of aggressive endometrial cancer and associates with hypoxia, angiogenesis and inflammation responses. *Oncotarget* **6**, 39676–39691 (2015).
184. Chen, H., Klein, R., Arnold, S., Chambers, S. & Zheng, W. Cytologic studies of the fallopian tube in patients undergoing salpingo-oophorectomy. *Cancer Cell Int.* **16**, 78 (2016).
185. Wang, W., Tran, D. & Feiszli, M. What makes training multi-modal classification networks hard? in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2020). doi:10.1109/cvpr42600.2020.01271.

186. Raghu, M., Zhang, C., Kleinberg, J. & Bengio, S. Transfusion: Understanding Transfer Learning for Medical Imaging. in *Advances in Neural Information Processing Systems 32* (eds. Wallach, H. et al.) 3347–3357 (Curran Associates, Inc., 2019).
187. Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609–615 (2011).
188. Prior, F. W. *et al.* TCIA: An information resource to enable open science. in *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* 1282–1285 (2013).
189. Cheng, D. T. *et al.* Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT): A Hybridization Capture-Based Next-Generation Sequencing Clinical Assay for Solid Tumor Molecular Oncology. *J. Mol. Diagn.* **17**, 251–264 (2015).
190. Chakravarty, D. *et al.* OncoKB: A Precision Oncology Knowledge Base. *JCO Precis Oncol* **2017**, (2017).
191. Chang, M. T. *et al.* Accelerating Discovery of Functional Mutant Alleles in Cancer. *Cancer Discov.* **8**, 174–183 (2018).
192. Chang, M. T. *et al.* Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. *Nat. Biotechnol.* **34**, 155–163 (2016).
193. van Griethuysen, J. J. M. *et al.* Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Res.* **77**, e104–e107 (2017).
194. Amadasun, M. & King, R. Textural features corresponding to textural properties. *IEEE Trans. Syst. Man Cybern.* **19**, 1264–1274 (1989).
195. Haralick, R. M., Shanmugam, K. & Dinstein, I. Textural Features for Image Classification. *IEEE Trans. Syst. Man Cybern.* **SMC-3**, 610–621 (1973).
196. Hanna, M. G. *et al.* Whole slide imaging equivalency and efficiency study: experience at a large academic center. *Mod. Pathol.* **32**, 916–928 (2019).
197. van der Walt, S. *et al.* scikit-image: image processing in Python. *PeerJ* **2**, e453 (2014).
198. Schmidt, U., Weigert, M., Broaddus, C. & Myers, G. Cell Detection with Star-Convex Polygons. in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018* 265–273 (Springer International Publishing, 2018).
199. Bankhead, P. *et al.* QuPath: Open source software for digital pathology image analysis. *Sci. Rep.* **7**, 16878 (2017).
200. Harrell, F. E., Jr, Califf, R. M., Pryor, D. B., Lee, K. L. & Rosati, R. A. Evaluating the yield of medical tests. *JAMA* **247**, 2543–2546 (1982).
201. Wolpert, D. H. The Lack of A Priori Distinctions Between Learning Algorithms. *Neural Comput.* **8**, 1341–1390 (1996).
202. Wolpert, D. H. & Macready, W. G. No Free Lunch Theorems for Optimization. *IEEE Trans. Evol. Comput.* **1**, (1997).
203. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).

204. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016) doi:10.1109/cvpr.2016.90.
205. Iandola, F. N. *et al.* SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size. *arXiv [cs.CV]* (2016).
206. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking the Inception Architecture for Computer Vision. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016) doi:10.1109/cvpr.2016.308.
207. Huang, G., Liu, Z., van der Maaten, L. & Weinberger, K. Q. Densely Connected Convolutional Networks. *arXiv [cs.CV]* (2016).
208. Cho, K. *et al.* Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2014) doi:10.3115/v1/d14-1179.
209. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).
210. Lee, G., Kang, B., Nho, K., Sohn, K.-A. & Kim, D. MildInt: Deep Learning-Based Multimodal Longitudinal Data Integration Framework. *Front. Genet.* **10**, 617 (2019).