HHS Public Access

Author manuscript

Nat Mach Intell. Author manuscript; available in PMC 2023 October 06.

Published in final edited form as:

Nat Mach Intell. 2023 April ; 5(4): 351–362. doi:10.1038/s42256-023-00633-5.

Multimodal data fusion for cancer biomarker discovery with deep learning

Sandra Steyaert[1], **Marija Pizurica**[1], **Divya Nagaraj**[2], **Priya Khandelwal**[2], **Tina Hernandez-Boussard**[1,3], **Andrew J Gentles**[1,3], **Olivier Gevaert**[1,3]

[1)]Stanford Center for Biomedical Informatics Research (BMIR), Department of Medicine, Stanford University.

[2)]Department of Computer Science, Stanford University.

[3)]Department of Biomedical Data Science, Stanford University.

Abstract

Technological advances now make it possible to study a patient from multiple angles with high-dimensional, high-throughput multi-scale biomedical data. In oncology, massive amounts of data are being generated ranging from molecular, histopathology, radiology to clinical records. The introduction of deep learning has significantly advanced the analysis of biomedical data. However, most approaches focus on single data modalities leading to slow progress in methods to integrate complementary data types. Development of effective multimodal fusion approaches is becoming increasingly important as a single modality might not be consistent and sufficient to capture the heterogeneity of complex diseases to tailor medical care and improve personalised medicine. Many initiatives now focus on integrating these disparate modalities to unravel the biological processes involved in multifactorial diseases such as cancer. However, many obstacles remain, including lack of usable data as well as methods for clinical validation and interpretation. Here, we cover these current challenges and reflect on opportunities through deep learning to tackle data sparsity and scarcity, multimodal interpretability, and standardisation of datasets.

Introduction

Over the past decades, technological innovations have transformed the healthcare domain with ever-growing availability of clinical data supporting diagnosis and care. Medicine is moving towards gathering multimodal patient data, especially in the context of age-related chronic diseases such as cancer[1, 2]. Integrating different data modalities can enhance our understanding of cancer[3, 4], but also paves the way for precision medicine which promises individualised diagnosis, prognosis, treatment and care[1, 5, 6].

Increasingly, we are moving from the traditional one-size-fits-all approach to more targeted testing and treatment. While molecular pathology revolutionised precision oncology, the first FDA-cleared companion diagnostic (CDx) assays relied on simpler molecular methods, and most assays focused on a single-gene of interest[7, 8]. However, advances in next-generation sequencing (NGS) now allow for multi-target CDx assays which are becoming more prevalent[8, 9]. The continuing cost reduction make it possible to simultaneously profile thousands of genomic regions hinting that soon multi-target panels could be run at a similar

price point to that of testing 5 to 10 targets individually[10]. Multi-target tests not only conserve time and tissue, but also have the potential to identify complex genetic interactions, thereby enhancing our understanding of tumour biology. While NGS is still in full swing, a third wave of technologies featuring single-molecule, long-read and real-time sequencing is already on the rise. Pacific Biosciences (PacBio) and Oxford Nanopore Technologies allow to assemble and explore genomes at unprecedented resolution and speed[11]. This technology was recently used in a clinical setting to diagnose rare genetic diseases with a turnaround rate of only 8 hours[12]. Since cancer often is multicausal, the area of precision oncology greatly benefits from these developments.

At the same time, histopathology and radiology have been critical tools in clinical decision-making during cancer management[13, 14]. Histopathological evaluation enables the study of tissue architecture and remains the gold standard for cancer diagnosis[15]. More recently, significant progress in whole slide imaging (WSI) has led to a transition from traditional histopathology methods towards digital pathology[16]. Digital pathology, the process of "digitising" conventional glass slides to virtual images, has many practical advantages over more traditional approaches, including speed, more straightforward data storage and management, remote access and shareability, and highly accurate, objective, and consistent readouts. On the other end of the spectrum is radiographic imaging, a non-invasive method for detecting and classifying cancer lesions. In particular Computed Tomography (CT) and Magnetic Resonance Imaging (MRI) scans are useful for generating 3D images of (pre)malignant lesions.

Ongoing improvements in artificial intelligence (AI) and advanced machine learning (ML) techniques have had major impacts on these cancer imaging ecosystems, especially in diagnostic and prognostic disciplines[17]. Current annotation of histopathological slides heavily relies on specialised pathologists. Leveraging image-based AI applications would not only alleviate the pathologists' workload but also has the potential for more efficient, reproducible, and accurate spatial analysis capturing information beyond visual perception[17–19]. Radiomics and pathomics refer to fields focusing on the quantitative analysis of radiological or histopathological digital images, respectively, with the aim of extracting quantitative features that can be used for clinical decision-making[20]. This extraction used to be done with standard statistical methods, but more advanced deep learning (DL) frameworks like convolutional neural networks (CNN), deep autoencoders (DAN) and vision transformers (ViTs) are now available for automated, high-throughput feature extraction[21–24]. Automatic assessment of deterministic objective features has enabled the quantification of tumour microenvironments (TME) at unprecedented speed and scale. In addition to the quantification of known hand-crafted salient features without inter-observer variability, DL also has the ability to discover unknown features and relationships that can provide biological insights and improve disease characterisation[25]. A notable radiomics study in lung cancer found that DL features captured prognostic signatures, both within and beyond the tumour region, that correlated with cell cycle and transcriptional processes[26]. Despite DL's diverse capacity, one of the main challenges is the need for large datasets to train, test and validate its algorithms. But, due to ethical restrictions and the labour intensity to annotate clinical images, most studies only have limited access to large cohorts that contain ground truth labelled data[27].

Under the 21st Century Cures Act[28], the FDA set a goal to advance precision medicine where the patient is at the centre of care. This act defines timelines for discovery, development, and delivery, and requires the fusion of evidence across modalities, with the provision that this must include real-world data and patient experience. Technological advances initiated an era where clinical data is being captured from multiple sources at unprecedented pace, ranging from medical images to genomics data and patient-generated health data (PGHD). Together with successes in AI, this opens the opportunity and necessity to analyse many data types with these advanced tools to better inform decision-making and improve patient care. To date, the FDA has cleared and approved several AI-based software as a medical device (SaMD)[29]. Together with the publication of their recent AI/ML white paper[30] the FDA wants to highlight their intention to develop a regulatory framework for these highly iterative, autonomous, and continuously learning algorithms as well as for the specific data types necessary to assure safety and effectiveness. Some proposed considerations for data inclusion are (i) relevance to the clinical problem and current clinical practice, (ii) data acquisition in a consistent, generalisable, and clinically relevant manner, (iii) appropriate definition and separation of training, tuning and test sets, and (iv) appropriate level of transparency of the algorithm and its output to users.

Integration of AI functionalities in medical applications has increased in recent years[31]. However, so far most methods focused on only one specific data type at a time, leading to slow progress in approaches to integrate complementary data types with many remaining questions about the technical, analytical and clinical aspect of multimodal integration[32–35]. To advance precision oncology, healthcare AI should not only inform about cancer incidence and tumour growth, but must identify the optimal treatment path, accounting for treatment-related side effects, socioeconomic factors, and care goals. Precision medicine can therefore only be achieved by merging complex and diverse multimodal data that span space and time. Single data modalities can be noisy or incomplete, but when combined with redundant signals from other modalities they can be more sensitive and robust to diagnose, prognose and assign treatments. Multimodal data are now being collected, providing a resource for biomarker discovery[36–39]. For cancer, both prognostic and predictive biomarkers are of interest. While the former provides information on the patient's diagnosis and overall outcome, the latter informs about treatment decisions and response[40].

Here, we argue that several sources of routinely collected medical data are not used to their full potential for diagnosing and treating cancer patients, because they are studied mostly in isolation instead of in an integrated fashion. These are: (i) electronic health records (EHR), (ii) molecular data, (iii) digital pathology and (iv) radiographic images. When combined, these data modalities provide a wealth of complementary, redundant, and harmonious information that can be exploited to better stratify patient populations and provide individualised care (Fig. 1). In the next sections, we discuss both challenges and opportunities for multimodal biomarker discovery as it applies to cancer patients. We cover strategies for data fusion and examine approaches to address data sparsity and scarcity, data orchestration and model interpretability.

The need for multimodal data fusion in oncology

Despite huge investments in cancer research and improved diagnosis and treatments, cancer prognosis is still bleak. Predictive models based on single modalities offer a limited view of disease heterogeneity and might not provide sufficient information to stratify patients and capture the full range of events that take place in response to treatments[41, 42]. For example, although immunotherapeutic methods like antibody-drug conjugates (ADCs) and adoptive cell therapy (ACT) (e.g. T-cell receptor (TCR) and chimeric antigen receptor T-cell (CAR-T) therapy) have shown to be very promising, response rates vary dramatically depending on the tumour subtype[43] and the TME[44]. Various TME elements play a role in tumour development, but also in therapeutic response. Furthermore, the cellular composition of the TME dynamically evolves with tumour progression and in response to anticancer treatments[45, 46]. The increasing application of immunotherapy underlines the need for (i) a deeper understanding of the TME and (ii) multimodal approaches that allow longitudinal TME monitoring during disease progression and therapeutic intervention[47].

Currently, biomarker discovery is mainly based on molecular data[48]. Increasing implementation of genomics and proteomic technologies in a clinical setting has led to growing availability, but also growing complexity, of molecular data[8]. Large consortia like The Cancer Genome Atlas (TCGA) and Genomic Data Commons (GDC) have gathered and standardised large datasets, accumulating petabytes of genomic, expression and proteomics data[37, 49, 50]. Barriers for NGS assay development, validation, and routine implementation remain due to many factors, such as tumour heterogeneity, sampling bias and interpretation of the results. Clinically accepted performance requirements are also often cancer-specific and depend on where in the care trajectory and for what specific purpose (e.g. diagnostic, stratification, drug response or treatment decision) tests are used[51]. As relevant as molecular data are for precision medicine, they discard tissue architecture, spatial and morphological information.

Although lower in resolution than genomic information, both WSI and radiographic images potentially harness orthogonal and complementary information. Digital pathology with WSIs provides data about the cellular and morphological architecture in a visual way for pathologists to interpret and can provide key information about the TME's spatial heterogeneity using image analysis and spatial statistics[52]. Similarly, radiographic images like MRIs or CT scans provide visual data of the tissue morphology and 3D structure[53].

Integration of data modalities that cover different scales of a patient has the potential to capture synergistic signals that identify both intra- and inter-patient heterogeneity critical for clinical predictions[54–56]. For example, the 2016 WHO classification of tumours of the central nervous system (CNS) revisited the guidelines to classify diffuse gliomas recommending histopathological diagnosis in combination with molecular markers (e.g. IDH1/2 mutation status), as each modality alone is insufficient to explain patient outcome variance[32, 33]. Of late, some reports also suggest the use of DNA-methylation-based classification of CNS tumors[34, 35].

The need for integrative modelling is increasingly emphasised. In 2015, a report from Ritchie *et al.* highlighted that "*approaches to combine multiple data types provide a more*

comprehensive understanding of complex genotype-phenotype associations than analysis of one dataset"[57]. The last years, there have been several attempts to develop multimodal approaches, to a great degree stimulated by community-driven competitions such as DREAM and Kaggle (i.e. http://dreamchallenges.org/ and https://www.kaggle.com/). But more work is needed to integrate routinely collected data modalities into clinical decision systems.

Data fusion strategies for multimodal biomarker discovery

The age of precision medicine demands powerful computational techniques to handle high-dimensional multimodal patient data. Each data source has strengths and limitations in its creation, analysis, and interpretation that must be addressed.

Medical images, whether 2D in histopathology or 3D in radiology, contain dense information that is encoded at multiple scales. Importantly, they contain high spatial correlation and any successful approach needs to take this into account[58]. So far, the best performing methods have been based on DL, and specifically CNNs[59–61]. Continuous improvement in detection, segmentation, classification, and spatial characterisation means that these methods are becoming a crucial part of cancer biomarker algorithms.

EHRs comprise various data types ranging from structured data such as medications, diagnosis codes, vital signs, or lab tests, to unstructured data in the form of clinical notes, patient emails, and detailed clinical processes. Natural language processing (NLP) algorithms that can extract useful clinical information from structured and unstructured EHR data are being developed. A recent study showed the feasibility and power of such ML tools in a lung cancer cohort to reliably extract important prognostic factors embedded in the EHRs [62]. Structured EHR sources are the easiest to process. Usually, this data is embedded into a lower dimensional vector space and fed as input to a recurrent neural network (RNNs). Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) are the most popular RNN architectures for this purpose[63–65]. While structured EHR data have obvious value, integration with insights from unstructured clinical data has shown to greatly improve clinical phenotyping[66]. Fortunately, advances in NLP now make it possible mine the unstructured narratives of patient records. One way to process this data is to convert free text to medical concepts and create lower dimensional "concept embeddings". Older methods such as *Word2Vec*[67] and *GloVe*[68] have almost been overtaken by "contextualised embeddings" like *ELMo*[69] and *BERT*[70–72]. While *ELMo* uses RNNs, *BERT* is based on transformers, a neural architecture that has revolutionised the NLP field since its inception[73]. To unlock EHRs' full potential, more appropriate techniques are needed combining structured and unstructured information, while accounting for the noise and inaccuracies that are common to these data[74]. In this regard, the concept of transfer learning for extracting clinical information from EHRs has gained a lot of traction[75].

Effective fusion methods must integrate high-dimensional multimodal biomedical data, ranging from quantitative features to images and text[76]. Representing raw data in a workable format remains challenging as ML methods do not readily accept unvectorised data. A multimodal representation thus poses many difficulties. Different modalities measure distinct unmatched features with different underlying distributions and dimensionalities.

Also, not all modalities and observations have the same level of confidence, noise, or information quality[77]. Multimodal fusion often suffers from dealing with wide feature matrices originating from very few samples with many features across modalities. Often advanced feature extraction methods such as kernel-based methods, graphical models, or NNs are needed prior to or as part of the data fusion process to reduce the dimensionality while preserving most of the salient biological signals[77–80]. Meaningful feature descriptions are the critical backbone of any model.

A major decision that must be made is at what specific modelling stage the data fusion takes place: (i) early, (ii) intermediate or (iii) late (Fig. 2)[81–83]. Early fusion is characterised by concatenating feature vectors of different data modalities and only requires the training of a single model (Fig. 2a). In contrast, late fusion is based on developing models on each data modality separately and integrating their single predictions with specific averaging, weighting, or other mechanisms (Fig. 2c). Not only does late fusion allow the use of a different, often more suitable, model for each modality but it also makes it more straightforward to handle situations when some modalities are missing in the data. However, fusion at the late stage ignores possible synergies between different modalities[84].

While both early and late fusion approaches are model-agnostic, they are not specifically designed to cope with or take full advantage of multiple modalities. Anything between early and late fusion is defined as intermediate or joint data fusion[84]. Intermediate fusion does not merge input data, nor develops separate models for each modality, but instead involves the development of inference algorithms to generate a joint multimodal low-level feature representation that retains the signal and properties of each individual modality (Fig. 2b). Although dedicated inference algorithms must be developed for each model type, this approach attempts to exploit the advantages of both early and late fusion[79, 83]. One key difference with early fusion is that the loss is propagated back to the inference algorithms during training, thus creating updated feature representations per training iteration[84]. Although this allows to model complex interactions between modalities, techniques need to be in place to prevent overfitting on the training cohort. Importantly, there is currently no decisive evidence that one fusion strategy is superior, and the choice of a specific approach is usually empirically based on the available data and task[84].

Advances in multimodal biomarkers for patient stratification

Multi-omics data fusion—Although a single omics technology provides insights into the profile of a tumour, one technique alone does not fully capture the underlying biology. The rising collection of large cohorts of multi-omics cancer data has spurred several efforts to fuse multi-omics data to fully grasp the tumour profile and several models for survival and risk prediction have been proposed[4, 6, 56, 85–93]. The TCGA research network also published numerous papers investigating the integration of genomic, transcriptomic, epigenomic and proteomic data for multiple cancer types[94–96]. Also for therapy response and drug combination predictions, multi-omics ML methods proved their value over traditional unimodal models[97–100]. Although various multi-omics fusion strategies now exist, one single method will not be optimal for all research questions and data types, and sometimes adding more omics layers can even negatively impact performance[101]. Each

strategy has its own strengths and weaknesses, and careful selection of effective approaches should be based on the purpose and available data types[57].

Multi-scale data fusion—Similar efforts as for multi-omics data fusion have been explored for multi-scale data[89, 102–107]. For example, Cheerla *et al.* used an intermediate fusion strategy to integrate histopathology, clinical, and expression data to predict patient survival for multiple cancer types. For each modality, an unsupervised encoder compressed the data into a single feature vector per patient. These feature vectors were aggregated into a joint representation allowing possible absence of one or more modalities[48]. Similarly, another study proposed a late fusion strategy to classify lung cancer. Using RNAseq, miRNAseq, WSI, copy number variation, and DNA-methylation they achieved better performance than obtained by each individual modality[108]. A few examples exist that show the potential of radiology to further refine patient stratification[109–111]. However, due to its high dimensionality and computational demands, so far most studies have avoided its inclusion[112].

Imaging genomics & Radiogenomics—When possible, molecular tumour information is nowadays used in cancer prognosis and treatment decisions. Interestingly, multiple studies have shown that phenotypes derived from medical images can act as proxies or biomarkers of molecular phenotypes like an EGFR mutation in lung cancer[113–115]. This discovery immediately gave rise to an emerging field called “radiogenomics”, the study of directly linking image features to underlying molecular properties[116]. For example, Itakura *et al.* used MRI phenotypes to define subtypes of glioblastoma associated with molecular pathway activity[117]. Also for breast cancer, the value of radiogenomics for risk prediction and better subtype stratification has been shown[118–120].

Current challenges and future directions for multimodal data fusion

Use of multimodal data models is likely the only way to advance precision oncology, but many challenges exist to realise their full potential. Although data availability is the main driver of multimodal data fusion, it also poses its major barrier. DL requires large amounts of data and data sparsity and scarcity both present serious challenges, especially for biomedical data. In clinical practice there are often different types of data missing between patients as not all patients might have all modalities due to cost, insurance coverage, material availability and lack of systemic collection procedures amongst others. To become relevant in an oncology setting, methods need to be able to handle different patterns of missing modalities. Fortunately, various interpolation, imputation and matrix completion algorithms have already been successfully applied for clinical data. These can range from basic methods including mean/median substitution, regression, k-nearest neighbour, and tree-based methods to more advanced algorithms like multiple imputation, multivariate imputation by chained equations or NN like RNN, LSTM and GANs[121–123]. Also, with the recent successes in DL techniques, dedicated fusion approaches are becoming available that allow joint representations that can handle incomplete or missing modalities[48, 124–129].

However, there are two major hurdles to advance these efforts. Firstly, the depth of data per patient, i.e. many observables per patient are routinely generated and stored, but typical

cohort sizes of patients are relatively small. Emerging evidence highlights that these cohorts are often biased, representing patients from higher socioeconomic status with continuous access to care and high levels of patient engagement[130, 131]. Limiting analyses to patients with complete data will lead to model overfitting, bias, and poor generalisation. Secondly, the lack of large 'golden labelled' cohorts with matched multimodal data, mainly due to the intense labour to annotate cancer datasets combined with privacy concerns. Luckily, also here DL algorithms are starting to be developed. One popular approach is data augmentation[132–135], which can include basic data transformations as well as generation of synthetic data, but also other strategies such as semi-supervised learning[136–139], active learning[140, 141], transfer learning[139, 142–144], and automated annotation[145, 146] have shown to be promising avenues to overcome labelled data scarcity.

Despite its potential, a critical roadblock for the widespread adoption of DL in a clinical setting is the lack of well-defined methods for model interpretation. While DL can extract predictive features from complex data, these are usually abstract, and it is not always apparent if they are clinically relevant[147]. To be useful in clinical decision-making, models need to undergo extensive testing, be interpretable, and their predictions need to be accompanied by confidence or uncertainty measures[148, 149]. Only then will they be relevant for and adopted by clinical practitioners.

Interpretation of black box models is a heavily investigated topic and some methods for post-hoc explanations have been proposed[147, 150]. In histopathology, most work focuses on extracting the most informative tiles by selecting those with the highest model confidence or by visualising tiles that are most relevant to the final prediction (Fig. 3a). For interpreting model predictions at higher resolution, the most relevant regions can be highlighted using gradient-based interpretation methods like Grad-CAM (Fig. 3b)[151]. Similarly, for molecular data, predictive features can be determined and visualised via Shapley Additive Explanation (SHAP)-based methods (Fig. 3d,e)[150, 152–154].

Multimodal data adds additional complexity and needs careful evaluation of appropriate methods before scaling to multimodal interpretability. However, multimodal approaches are starting to emerge with encouraging solutions not only for interpretability but also for discovery of associations between modalities[147, 150]. Note that the aforementioned methods specify why a model makes a specific decision, but do not explain the used features. Additional strategies could be leveraged to further unravel biological insights. For example, selected tiles could be overlayed with Hover-Net[155] to segment and classify nuclei to evaluate predominant cell types (Fig. 3c, unpublished data).

Standardisation will lead to more uniform and complete datasets, which are easier to process and fuse with other sources and will be much more interpretable on their own. TCGA is probably the best known and most used resource[37], but many other initiatives are underway to structurally capture clinical, genomics, imaging, and pathological data for oncology, such as The Cancer Imaging Archive (TCIA)[36] and the Genomics Pathology Imaging Collection (GPIC)[38]. Together, these efforts have the shared aim to process, analyse and share data using a community-embraced standard in a FAIR (findable, accessible, interoperable, and reusable) way[156]. This will not only promote reproducibility and transparency, but also

encourages reutilisation and optimisation of existing work. However, the volume and complexity of multimodal biomedical data makes it increasingly difficult to produce and share FAIR data and current solutions often require specific expertise and resources[157]. Furthermore, some modalities such as EHRs are not only extremely difficult to standardise and share, but also very expensive to obtain by researchers[158, 159]. Efforts like OMOP aim at tackling this issue by harmonising EHR data across institutes and countries[160, 161]. To make progress in multimodal studies, there is a dire need for data orchestration platforms[157], but also appropriate regulatory frameworks to preserve patients' privacy[162].

The importance of biomedical multimodal data fusion becomes increasingly apparent as more clinical and experimental data becomes available. To tackle the multimodal-specific obstacles, multiple methods and frameworks have been proposed and are currently heavily explored. While often still problem-specific and experimental, the field is gaining knowledge to evaluate and define what methods excel given specific conditions and data modalities. DL approaches have only touched a limited range of potential applications, mainly because of the challenges inherent to the current state of health care data as discussed above, again emphasising the need for large collaborative data standardisation and sharing efforts. In this space, competitions such as DREAM and Kaggle have been an effective concept for making standardised multimodal data available. Importantly, these initiatives also facilitate exchange of ideas and code, reproducibility, innovation, and unbiased evaluation[163, 164]. It is our expectation that such efforts will significantly advance development of robust multimodal approaches.

Ultimately, the goal is to advance precision oncology by rigorous clinical validation of successful models in larger independent cohorts to prove any clinical utility. So far, most efforts have focused on multimodal cancer biomarkers to refine risk stratification, but with dedicated strategies multimodal data fusion could also assist in treatment decision or drug response. However, outcomes in real-world patients often lag relative to clinical trials thereby hindering to evaluate efficacies due to lack of follow-up data. Fortunately, efforts are underway to capture treatment response in automated scalable ways using NLP from clinical notes[165]. With careful study design, ongoing improvements in data collection and sharing methods, and decreasing cost and/or availability of disease monitoring technologies, DL algorithms present a promising choice to further accelerate the field of precision oncology in this direction.

Acknowledgements

We would like to thank Marie Humbert-Droz for discussions during the early stages of this manuscript. We are grateful for her insightful ideas and comments about these topics. We would also like to express our great appreciation to Christophe Sadée and Yuanning Zheng for their valuable and constructive suggestions during the write-up of this manuscript.

References

1. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 25, 44–56 (2019). [PubMed: 30617339]
2. Riba M, Sala C, Toniolo D & Tonon G. Big Data in Medicine, the Present and Hopefully the Future. Front Med (Lausanne) 6, 263 (2019). [PubMed: 31803746]

3. Hanahan D. Hallmarks of Cancer: New Dimensions. Cancer Discov 12, 31–46 (2022). [PubMed: 35022204]
4. Lu J et al. Multi-omics reveals clinically relevant proliferative drive associated with mTOR-MYC-OXPHOS activity in chronic lymphocytic leukemia. Nat Cancer 2, 853–864 (2021). [PubMed: 34423310]
5. Medina-Martinez JS et al. Isabl Platform, a digital biobank for processing multimodal patient data. BMC Bioinformatics 21, 549 (2020). [PubMed: 33256603]
6. Chai H et al. Integrating multi-omics data through deep learning for accurate cancer prognosis prediction. Comput Biol Med 134, 104481 (2021).
7. Dietel M et al. Predictive molecular pathology and its role in targeted cancer therapy: a review focussing on clinical relevance. Cancer Gene Ther 20, 211–221 (2013). [PubMed: 23492822]
8. Malone ER, Oliva M, Sabatini PJB, Stockley TL & Siu LL. Molecular profiling for precision cancer therapies. Genome Med 12, 8 (2020). [PubMed: 31937368]
9. Campbell MR. Update on molecular companion diagnostics - a future in personalized medicine beyond Sanger sequencing. Expert Rev Mol Diagn 20, 637–644 (2020). [PubMed: 32167388]
10. Colomer R et al. When should we order a next generation sequencing test in a patient with cancer? EClinicalMedicine 25, 100487 (2020).
11. van Dijk EL, Jaszczyszyn Y, Naquin D & Thermes C. The Third Revolution in Sequencing Technology. Trends Genet 34, 666–681 (2018). [PubMed: 29941292]
12. Gorzynski JE et al. Ultrarapid Nanopore Genome Sequencing in a Critical Care Setting. N Engl J Med 386, 700–702 (2022). [PubMed: 35020984]
13. Davidson MR, Gazdar AF & Clarke BE. The pivotal role of pathology in the management of lung cancer. J Thorac Dis 5 Suppl 5, S463–478 (2013). [PubMed: 24163740]
14. Pomerantz BJ. Imaging and Interventional Radiology for Cancer Management. Surg Clin North Am 100, 499–506 (2020). [PubMed: 32402296]
15. Yu KH & Snyder M. Omics Profiling in Precision Oncology. Mol Cell Proteomics 15, 2525–2536 (2016). [PubMed: 27099341]
16. Rahman A et al. Advances in tissue-based imaging: impact on oncology research and clinical practice. Expert Rev Mol Diagn 20, 1027–1037 (2020). [PubMed: 32510287]
17. van der Laak J, Litjens G & Ciompi F. Deep learning in histopathology: the path to the clinic. Nat Med 27, 775–784 (2021). [PubMed: 33990804]
18. Baxi V, Edwards R, Montalto M & Saha S. Digital pathology and artificial intelligence in translational medicine and clinical practice. Mod Pathol 35, 23–32 (2022). [PubMed: 34611303]
19. Serag A et al. Translational AI and Deep Learning in Diagnostic Pathology. Front Med (Lausanne) 6, 185 (2019). [PubMed: 31632973]
20. Iv M et al. MR Imaging-Based Radiomic Signatures of Distinct Molecular Subgroups of Medulloblastoma. AJNR Am J Neuroradiol 40, 154–161 (2019). [PubMed: 30523141]
21. van Timmeren JE, Cester D, Tanadini-Lang S, Alkadhi H & Baessler B. Radiomics in medical imaging-"how-to" guide and critical reflection. Insights Imaging 11, 91 (2020). [PubMed: 32785796]
22. Liang J, Yang C, Zeng M & Wang X. TransConver: transformer and convolution parallel network for developing automatic brain tumor segmentation in MRI images. Quant Imaging Med Surg 12, 2397–2415 (2022). [PubMed: 35371952]
23. Kim M et al. Deep Learning in Medical Imaging. Neurospine 16, 657–668 (2019). [PubMed: 31905454]
24. Dosovitskiy A et al. An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale. arXiv, https://arxiv.org/abs/2010.11929 (2020).
25. Gupta R, Kurc T, Sharma A, Almeida JS & Saltz J. The Emergence of Pathomics. Current Pathobiology Reports 7, 73–84 (2019).
26. Hosny A et al. Deep learning for lung cancer prognostication: A retrospective multi-cohort radiomics study. PLoS Med 15, e1002711 (2018).
27. Castro DC, Walker I & Glocker B. Causality matters in medical imaging. Nat Commun 11, 3673 (2020). [PubMed: 32699250]

28. 21st Century Cures Act. H.R. 34. 114th Congress, https://www.congress.gov/114/bills/hr134/BILLS-114hr134enr.pdf (2016).
29. FDA. Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices (Accessed Oct 31 2022).
30. FDA. Proposed Regulatory Framework for Modification to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD). White paper, https://www.fda.gov/files/medical%20devices/published/US-FDA-Artificial-Intelligence-and-Machine-Learning-Discussion-Paper.pdf (2019).
31. Kann BH, Thompson R, Thomas CR Jr., Dicker A & Aneja S. Artificial Intelligence in Oncology: Current Applications and Future Directions. Oncology (Williston Park) 33, 46–53 (2019). [PubMed: 30784028]
32. Louis DN et al. The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. Acta Neuropathol 131, 803–820 (2016). [PubMed: 27157931]
33. Tateishi K, Wakimoto H & Cahill DP. IDH1 Mutation and World Health Organization 2016 Diagnostic Criteria for Adult Diffuse Gliomas: Advances in Surgical Strategy. Neurosurgery 64, 134–138 (2017). [PubMed: 28899049]
34. Capper D et al. DNA-methylation-based classification of central nervous system tumours. Nature 555, 469–474 (2018). [PubMed: 29539639]
35. Ceccarelli M et al. Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma. Cell 164, 550–563 (2016). [PubMed: 26824661]
36. Prior F et al. The public cancer radiology imaging collections of The Cancer Imaging Archive. Sci Data 4, 170124 (2017).
37. Hutter C & Zenklusen JC. The Cancer Genome Atlas: Creating Lasting Value beyond Its Data. Cell 173, 283–285 (2018). [PubMed: 29625045]
38. Jennings CN et al. Bridging the gap with the UK Genomics Pathology Imaging Collection. Nat Med (2022).
39. Mo H, Breitling R, Francavilla C & Schwartz JM. Data integration and mechanistic modelling for breast cancer biology: Current state and future directions. Curr Opin Endocr Metab Res 24, None (2022).
40. Nalejska E, Maczynska E & Lewandowska MA. Prognostic and predictive biomarkers: tools in personalized oncology. Mol Diagn Ther 18, 273–284 (2014). [PubMed: 24385403]
41. Grossman JE, Vasudevan D, Joyce CE & Hildago M. Is PD-L1 a consistent biomarker for anti-PD-1 therapy? The model of balstilimab in a virally-driven tumor. Oncogene 40, 1393–1395 (2021). [PubMed: 33500548]
42. Davis AA & Patel VG. The role of PD-L1 expression as a predictive biomarker: an analysis of all US Food and Drug Administration (FDA) approvals of immune checkpoint inhibitors. J Immunother Cancer 7, 278 (2019). [PubMed: 31655605]
43. van Elsas MJ, van Hall T & van der Burg SH. Future Challenges in Cancer Resistance to Immunotherapy. Cancers (Basel) 12 (2020).
44. Dzobo K. Taking a Full Snapshot of Cancer Biology: Deciphering the Tumor Microenvironment for Effective Cancer Therapy in the Oncology Clinic. OMICS 24, 175–179 (2020). [PubMed: 32176591]
45. Ott M, Prins RM & Heimberger AB. The immune landscape of common CNS malignancies: implications for immunotherapy. Nat Rev Clin Oncol 18, 729–744 (2021). [PubMed: 34117475]
46. Bejarano L, Jordao MJC & Joyce JA. Therapeutic Targeting of the Tumor Microenvironment. Cancer Discov 11, 933–959 (2021). [PubMed: 33811125]
47. Zomer A, Croci D, Kowal J, van Gurp L & Joyce JA. Multimodal imaging of the dynamic brain tumor microenvironment during glioblastoma progression and in response to treatment. iScience 25, 104570 (2022).
48. Cheerla A & Gevaert O. Deep learning with multimodal representation for pancancer prognosis prediction. Bioinformatics 35, i446–i454 (2019). [PubMed: 31510656]
49. Grossman RL et al. Toward a Shared Vision for Cancer Genomic Data. N Engl J Med 375, 1109–1112 (2016). [PubMed: 27653561]

50. Hinkson IV et al. A Comprehensive Infrastructure for Big Data in Cancer Research: Accelerating Cancer Research and Precision Medicine. Front Cell Dev Biol 5, 83 (2017). [PubMed: 28983483]
51. Putcha G, Gutierrez A & Skates S. Multicancer Screening: One Size Does Not Fit All. JCO Precis Oncol 5, 574–576 (2021). [PubMed: 34994606]
52. Mi H et al. Digital Pathology Analysis Quantifies Spatial Heterogeneity of CD3, CD4, CD8, CD20, and FoxP3 Immune Markers in Triple-Negative Breast Cancer. Front Physiol 11, 583333 (2020).
53. Fass L. Imaging and cancer: a review. Mol Oncol 2, 115–152 (2008). [PubMed: 19383333]
54. Lanckriet GR, De Bie T, Cristianini N, Jordan MI & Noble WS. A statistical framework for genomic data fusion. Bioinformatics 20, 2626–2635 (2004). [PubMed: 15130933]
55. Gevaert O, De Smet F, Timmerman D, Moreau Y & De Moor B. Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. Bioinformatics 22, e184–190 (2006). [PubMed: 16873470]
56. Daemen A et al. A kernel-based integration of genome-wide data for clinical decision support. Genome Med 1, 39 (2009). [PubMed: 19356222]
57. Ritchie MD, Holzinger ER, Li R, Pendergrass SA & Kim D. Methods of integrating data to uncover genotype-phenotype interactions. Nat Rev Genet 16, 85–97 (2015). [PubMed: 25582081]
58. Panayides AS et al. AI in Medical Imaging Informatics: Current Challenges and Future Directions. IEEE J Biomed Health Inform 24, 1837–1857 (2020). [PubMed: 32609615]
59. George K, Faziludeen S, Sankaran P & Joseph KP. Breast cancer detection from biopsy images using nucleus guided transfer learning and belief based fusion. Comput Biol Med 124, 103954 (2020).
60. Singh SP et al. 3D Deep Learning on Medical Images: A Review. Sensors (Basel) 20 (2020).
61. Sarvamangala DR & Kulkarni RV. Convolutional neural networks in medical image understanding: a survey. Evol Intell, 1–22 (2021). [PubMed: 33425040]
62. Yuan Q et al. Performance of a Machine Learning Algorithm Using Electronic Health Record Data to Identify and Estimate Survival in a Longitudinal Cohort of Patients With Lung Cancer. JAMA Netw Open 4, e2114723 (2021).
63. Rasmy L et al. A study of generalizability of recurrent neural network-based predictive models for heart failure onset risk using a large and heterogeneous EHR data set. J Biomed Inform 84, 11–16 (2018). [PubMed: 29908902]
64. Shickel B, Tighe PJ, Bihorac A & Rashidi P. Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. IEEE J Biomed Health Inform 22, 1589–1604 (2018). [PubMed: 29989977]
65. Ayala Solares JR et al. Deep learning for electronic health records: A comparative review of multiple deep neural architectures. J Biomed Inform 101, 103337 (2020).
66. Hernandez-Boussard T, Monda KL, Crespo BC & Riskin D. Real world evidence in cardiovascular medicine: ensuring data validity in electronic health record-based studies. J Am Med Inform Assoc 26, 1189–1194 (2019). [PubMed: 31414700]
67. Mikolov T, Sutskever I, Chen K, Corrado G & Dean J Distributed representations of words and phrases and their compositionality. Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2, 3111–3119 (2013).
68. Pennington J, Socher R & Manning CD. Glove: Global Vectors for Word Representation. EMNLP 14, 1532–1543 (2014).
69. Peters ME et al. Deep contextualized word representations. arXiv, http://arxiv.org/abs/1802.05365 (2018).
70. Devlin J, Chang M-W, Lee K & Toutanova K BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1, 4171–4186 (2019).
71. Lee J et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 36, 1234–1240 (2020). [PubMed: 31501885]

72. Huang K, Garapati S & Rich AS. An Interpretable End-to-end Fine-tuning Approach for Long Clinical Text. arXiv, https://arxiv.org/abs/2011.06504 (2020).
73. Vaswani A et al. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems, 6000–6010 (2017).
74. Jensen PB, Jensen LJ & Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet 13, 395–405 (2012). [PubMed: 22549152]
75. Rasmy L, Xiang Y, Xie Z, Tao C & Zhi D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. NPJ Digit Med 4, 86 (2021). [PubMed: 34017034]
76. Acosta JN, Falcone GJ, Rajpurkar P & Topol EJ. Multimodal biomedical AI. Nat Med 28, 1773–1784 (2022). [PubMed: 36109635]
77. Jain MS et al. MultiMAP: dimensionality reduction and integration of multimodal data. Genome Biol 22, 346 (2021). [PubMed: 34930412]
78. Lahnemann D et al. Eleven grand challenges in single-cell data science. Genome Biol 21, 31 (2020). [PubMed: 32033589]
79. Baltrusaitis T, Ahuja C & Morency LP. Multimodal Machine Learning: A Survey and Taxonomy. IEEE Trans Pattern Anal Mach Intell 41, 423–443 (2019). [PubMed: 29994351]
80. Yan KK, Zhao H & Pang H. A comparison of graph- and kernel-based -omics data integration algorithms for classifying complex traits. BMC Bioinformatics 18, 539 (2017). [PubMed: 29212468]
81. Pavlidis P, Weston J, Cai J & Noble WS. Learning gene functional classifications from multiple data types. J Comput Biol 9, 401–411 (2002). [PubMed: 12015889]
82. Serra A, Galdi P & Tagliaferri R. Multiview Learning in Biomedical Applications (Chapter 13). Artificial Intelligence in the Age of Neural Networks and Brain Computing, 265–280 (2019).
83. Stahlschmidt SR, Ulfenborg B & Synnergren J. Multimodal deep learning for biomedical data fusion: a review. Brief Bioinform 23 (2022).
84. Huang SC, Pareek A, Seyyedi S, Banerjee I & Lungren MP. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. NPJ Digit Med 3, 136 (2020). [PubMed: 33083571]
85. Picard M, Scott-Boyer MP, Bodein A, Perin O & Droit A. Integration strategies of multi-omics data for machine learning analysis. Comput Struct Biotechnol J 19, 3735–3746 (2021). [PubMed: 34285775]
86. Chaudhary K, Poirion OB, Lu L & Garmire LX. Deep Learning-Based Multi-Omics Integration Robustly Predicts Survival in Liver Cancer. Clin Cancer Res 24, 1248–1259 (2018). [PubMed: 28982688]
87. Huang Z et al. SALMON: Survival Analysis Learning With Multi-Omics Neural Networks on Breast Cancer. Front Genet 10, 166 (2019). [PubMed: 30906311]
88. Wang T et al. MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. Nat Commun 12, 3445 (2021). [PubMed: 34103512]
89. Gevaert O, Villalobos V, Sikic BI & Plevritis SK. Identification of ovarian cancer driver genes by using module network integration of multi-omics data. Interface Focus 3, 20130013 (2013).
90. Xu J et al. A hierarchical integration deep flexible neural forest framework for cancer subtype classification by integrating multi-omics data. BMC Bioinformatics 20, 527 (2019). [PubMed: 31660856]
91. Zhang L et al. Deep Learning-Based Multi-Omics Data Integration Reveals Two Prognostic Subtypes in High-Risk Neuroblastoma. Front Genet 9, 477 (2018). [PubMed: 30405689]
92. Taskesen E, Babaei S, Reinders MM & de Ridder J. Integration of gene expression and DNA-methylation profiles improves molecular subtype classification in acute myeloid leukemia. BMC Bioinformatics 16 Suppl 4, S5 (2015).
93. Argelaguet R et al. Multi-Omics Factor Analysis-a framework for unsupervised integration of multi-omics data sets. Mol Syst Biol 14, e8124 (2018). [PubMed: 29925568]

94. Cancer Genome Atlas Research, N. et al. Comprehensive, Integrative Genomic Analysis of Diffuse Lower-Grade Gliomas. N Engl J Med 372, 2481–2498 (2015). [PubMed: 26061751]
95. Cancer Genome Atlas Research, N. et al. Integrated genomic and molecular characterization of cervical cancer. Nature 543, 378–384 (2017). [PubMed: 28112728]
96. Cancer Genome Atlas Research Network. Electronic address, e.d.s.c. & Cancer Genome Atlas Research, N. Comprehensive and Integrated Genomic Characterization of Adult Soft Tissue Sarcomas. Cell 171, 950–965 e928 (2017). [PubMed: 29100075]
97. Zhang T, Zhang L, Payne PRO & Li F. Synergistic Drug Combination Prediction by Integrating Multiomics Data in Deep Learning Models. Methods Mol Biol 2194, 223–238 (2021). [PubMed: 32926369]
98. Preuer K et al. DeepSynergy: predicting anti-cancer drug synergy with Deep Learning. Bioinformatics 34, 1538–1546 (2018). [PubMed: 29253077]
99. Sammut SJ et al. Multi-omic machine learning predictor of breast cancer therapy response. Nature 601, 623–629 (2022). [PubMed: 34875674]
100. Costello JC et al. A community effort to assess and improve drug sensitivity prediction algorithms. Nat Biotechnol 32, 1202–1212 (2014). [PubMed: 24880487]
101. Duan R et al. Evaluation and comparison of multi-omics data integration methods for cancer subtyping. PLoS Comput Biol 17, e1009224 (2021).
102. Venugopalan J, Tong L, Hassanzadeh HR & Wang MD. Multimodal deep learning models for early detection of Alzheimer's disease stage. Sci Rep 11, 3254 (2021). [PubMed: 33547343]
103. 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). [PubMed: 29531073]
104. Cheng J et al. Integrative Analysis of Histopathological Images and Genomic Data Predicts Clear Cell Renal Cell Carcinoma Prognosis. Cancer Res 77, e91–e100 (2017). [PubMed: 29092949]
105. Schulz S et al. Multimodal Deep Learning for Prognosis Prediction in Renal Cancer. Front Oncol 11, 788740 (2021).
106. Zhan Z et al. Two-stage Cox-nnet: biologically interpretable neural-network model for prognosis prediction and its application in liver cancer survival using histopathology and transcriptomic data. NAR Genom Bioinform 3, lqab015 (2021).
107. Chen RJ et al. Pathomic Fusion: An Integrated Framework for Fusing Histopathology and Genomic Features for Cancer Diagnosis and Prognosis. IEEE Trans Med Imaging 41, 757–770 (2022). [PubMed: 32881682]
108. Carrillo-Perez F et al. Machine-Learning-Based Late Fusion on Multi-Omics and Multi-Scale Data for Non-Small-Cell Lung Cancer Diagnosis. Journal of Personalized Medicine 12, 601 (2022). [PubMed: 35455716]
109. 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). [PubMed: 29572492]
110. Mazzaschi G et al. Integrated MRI-Immune-Genomic Features Enclose a Risk Stratification Model in Patients Affected by Glioblastoma. Cancers (Basel) 14 (2022).
111. Wang X et al. Combining Radiology and Pathology for Automatic Glioma Classification. Front Bioeng Biotechnol 10, 841958 (2022).
112. Yamaguchi H et al. Three-Dimensional Convolutional Autoencoder Extracts Features of Structural Brain Images With a "Diagnostic Label-Free" Approach: Application to Schizophrenia Datasets. Front Neurosci 15, 652987 (2021).
113. Liu Y et al. Radiomic Features Are Associated With EGFR Mutation Status in Lung Adenocarcinomas. Clin Lung Cancer 17, 441–448 e446 (2016). [PubMed: 27017476]
114. Gevaert O et al. Predictive radiogenomics modeling of EGFR mutation status in lung cancer. Sci Rep 7, 41674 (2017).
115. Nair JKR et al. Radiogenomic Models Using Machine Learning Techniques to Predict EGFR Mutations in Non-Small Cell Lung Cancer. Can Assoc Radiol J 72, 109–119 (2021). [PubMed: 32063026]

116. Pinker K, Chin J, Melsaether AN, Morris EA & Moy L. Precision Medicine and Radiogenomics in Breast Cancer: New Approaches toward Diagnosis and Treatment. Radiology 287, 732–747 (2018). [PubMed: 29782246]
117. Itakura H et al. Magnetic resonance image features identify glioblastoma phenotypic subtypes with distinct molecular pathway activities. Sci Transl Med 7, 303ra138 (2015).
118. Yamamoto S, Maki DD, Korn RL & Kuo MD. Radiogenomic analysis of breast cancer using MRI: a preliminary study to define the landscape. AJR Am J Roentgenol 199, 654–663 (2012). [PubMed: 22915408]
119. Sutton EJ et al. Breast cancer subtype intertumor heterogeneity: MRI-based features predict results of a genomic assay. J Magn Reson Imaging 42, 1398–1406 (2015). [PubMed: 25850931]
120. Li H et al. Quantitative MRI radiomics in the prediction of molecular classifications of breast cancer subtypes in the TCGA/TCIA data set. NPJ Breast Cancer 2 (2016).
121. Li J et al. Imputation of missing values for electronic health record laboratory data. NPJ Digit Med 4, 147 (2021). [PubMed: 34635760]
122. Luo Y. Evaluating the state of the art in missing data imputation for clinical data. Brief Bioinform 23 (2022).
123. Yoon J, Zame WR & van der Schaar M. Estimating Missing Data in Temporal Data Streams Using Multi-Directional Recurrent Neural Networks. IEEE Trans Biomed Eng 66, 1477–1490 (2019). [PubMed: 30296210]
124. Zhou T, Liu M, Thung KH & Shen D. Latent Representation Learning for Alzheimer's Disease Diagnosis With Incomplete Multi-Modality Neuroimaging and Genetic Data. IEEE Trans Med Imaging 38, 2411–2422 (2019). [PubMed: 31021792]
125. Liu Y et al. Incomplete multi-modal representation learning for Alzheimer's disease diagnosis. Med Image Anal 69, 101953 (2021).
126. Ning Z, Du D, Tu C, Feng Q & Zhang Y. Relation-Aware Shared Representation Learning for Cancer Prognosis Analysis With Auxiliary Clinical Variables and Incomplete Multi-Modality Data. IEEE Trans Med Imaging 41, 186–198 (2022). [PubMed: 34460368]
127. Momeni A, Thibault M & Gevaert O. Dropout-Enabled Ensemble Learning for Multi-Scale Biomedical Data. bioRxiv, 440362 (2018).
128. Mehdipour Ghazi M et al. Training recurrent neural networks robust to incomplete data: Application to Alzheimer's disease progression modeling. Med Image Anal 53, 39–46 (2019). [PubMed: 30682584]
129. Ma Q, Li S & Cottrell GW. Adversarial Joint-Learning Recurrent Neural Network for Incomplete Time Series Classification. IEEE Trans Pattern Anal Mach Intell 44, 1765–1776 (2022). [PubMed: 32997624]
130. Sharrocks K, Spicer J, Camidge DR & Papa S. The impact of socioeconomic status on access to cancer clinical trials. Br J Cancer 111, 1684–1687 (2014). [PubMed: 25093493]
131. Niranjan SJ et al. Perceived Institutional Barriers Among Clinical and Research Professionals: Minority Participation in Oncology Clinical Trials. JCO Oncol Pract 17, e666–e675 (2021). [PubMed: 33974821]
132. Mukherkjee D, Saha P, Kaplun D, Sinitca A & Sarkar R. Brain tumor image generation using an aggregation of GAN models with style transfer. Sci Rep 12, 9141 (2022). [PubMed: 35650252]
133. Qin Z, Liu Z, Zhu P & Xue Y. A GAN-based image synthesis method for skin lesion classification. Comput Methods Programs Biomed 195, 105568 (2020).
134. Huang HH, Rao H, Miao R & Liang Y. A novel meta-analysis based on data augmentation and elastic data shared lasso regularization for gene expression. BMC Bioinformatics 23, 353 (2022). [PubMed: 35999505]
135. Yufei L et al. Wasserstein GAN-Based Small-Sample Augmentation for New-Generation Artificial Intelligence: A Case Study of Cancer-Staging Data in Biology. Engineering 5, 156–163 (2019).
136. Wenqing S, Tzu-Liang T, Jianying Z & Wei Q. Computerized breast cancer analysis system using three stage semi-supervised learning method. Computer Methods and Programs in Biomedicine 135, 77–88 (2016). [PubMed: 27586481]

137. Dwarikanath M. Combining multiple expert annotations using semi-supervised learning and graph cuts for medical image segmentation. Computer Vision and Image Understanding 151, 114–123 (2016).
138. Tran QT, Alom MZ & Orr BA. Comprehensive study of semi-supervised learning for DNA-methylation-based supervised classification of central nervous system tumors. BMC Bioinformatics 23, 223 (2022). [PubMed: 35676649]
139. Cheplygina V, de Bruijne M & Pluim JPW. Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. Med Image Anal 54, 280–296 (2019). [PubMed: 30959445]
140. Jie Y, Xutong L & Mingyue Z. Current status of active learning for drug discovery. Artificial Intelligence in the Life Sciences 1, 100023 (2021).
141. Min W, Fan M, Zhi-Heng Z & Yan-Xue W. Active learning through density clustering. Expert Systems with Applications 85, 305–317 (2017).
142. Nahiyan M & Danilo B. From YouTube to the brain: Transfer learning can improve brain-imaging predictions with deep learning. Neural Networks 153, 325–338 (2022). [PubMed: 35777174]
143. Park Y, Hauschild AC & Heider D. Transfer learning compensates limited data, batch effects and technological heterogeneity in single-cell sequencing. NAR Genom Bioinform 3, lqab104 (2021).
144. Novakovsky G, Saraswat M, Fornes O, Mostafavi S & Wasserman WW. Biologically relevant transfer learning improves transcription factor binding prediction. Genome Biol 22, 280 (2021). [PubMed: 34579793]
145. Ganoe CH et al. Natural language processing for automated annotation of medication mentions in primary care visit conversations. JAMIA Open 4, ooab071 (2021).
146. Krenzer A et al. Fast machine learning annotation in the medical domain: a semi-automated video annotation tool for gastroenterologists. Biomed Eng Online 21, 33 (2022). [PubMed: 35614504]
147. Lipkova J et al. Artificial intelligence for multimodal data integration in oncology. Cancer Cell 40, 1095–1110 (2022). [PubMed: 36220072]
148. Schaumberg AJ et al. Interpretable multimodal deep learning for real-time pan-tissue pan-disease pathology search on social media. Mod Pathol 33, 2169–2185 (2020). [PubMed: 32467650]
149. Begoli E, Bhattacharya T & Kusnezov D. The need for uncertainty quantification in machine-assisted medical decision making. Nature Machine Intelligence 1, 20–23 (2019).
150. Chen RJ et al. Pan-cancer integrative histology-genomic analysis via multimodal deep learning. Cancer Cell 40, 865–878 e866 (2022). [PubMed: 35944502]
151. Selvaraju RR et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. arXiv, https://arxiv.org/abs/1610.02391 (2016).
152. Lundberg SM & Lee SI. A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems 30, 4765–4774 (2017).
153. Dickinson Q & Meyer JG. Positional SHAP (PoSHAP) for Interpretation of machine learning models trained from biological sequences. PLoS Comput Biol 18, e1009736 (2022).
154. Steyaert S et al. Multimodal data fusion of adult and pediatric brain tumors with deep learning. medRxiv, 2022.2009.2021.22280223 (2022).
155. Simon G et al. Hover-Net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. Medical Image Analysis 58, 101563 (2019).
156. Wilkinson MD et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016).
157. Mammoliti A et al. Orchestrating and sharing large multimodal data for transparent and reproducible research. Nat Commun 12, 5797 (2021). [PubMed: 34608132]
158. Mc Cord KA et al. Current use and costs of electronic health records for clinical trial research: a descriptive study. CMAJ Open 7, E23–E32 (2019).
159. Mc Cord KA & Hemkens LG. Using electronic health records for clinical trials: Where do we stand and where can we go? CMAJ 191, E128–E133 (2019). [PubMed: 30718337]
160. Makadia R & Ryan PB. Transforming the Premier Perspective Hospital Database into the Observational Medical Outcomes Partnership (OMOP) Common Data Model. EGEMS (Wash DC) 2, 1110 (2014). [PubMed: 25848597]

161. Papez V et al. Transforming and evaluating electronic health record disease phenotyping algorithms using the OMOP common data model: a case study in heart failure. JAMIA Open 4, ooab001 (2021).

162. Liang W et al. Advances, challenges and opportunities in creating data for trustworthy AI. Nature Machine Intelligence 4, 669–677 (2022).

163. Costello JC & Stolovitzky G. Seeking the wisdom of crowds through challenge-based competitions in biomedical research. Clin Pharmacol Ther 93, 396–398 (2013). [PubMed: 23549146]

164. Saez-Rodriguez J et al. Crowdsourcing biomedical research: leveraging communities as innovation engines. Nat Rev Genet 17, 470–486 (2016). [PubMed: 27418159]

165. Khozin S et al. Real-world progression, treatment, and survival outcomes during rapid adoption of immunotherapy for advanced non-small cell lung cancer. Cancer 125, 4019–4032 (2019). [PubMed: 31381142]

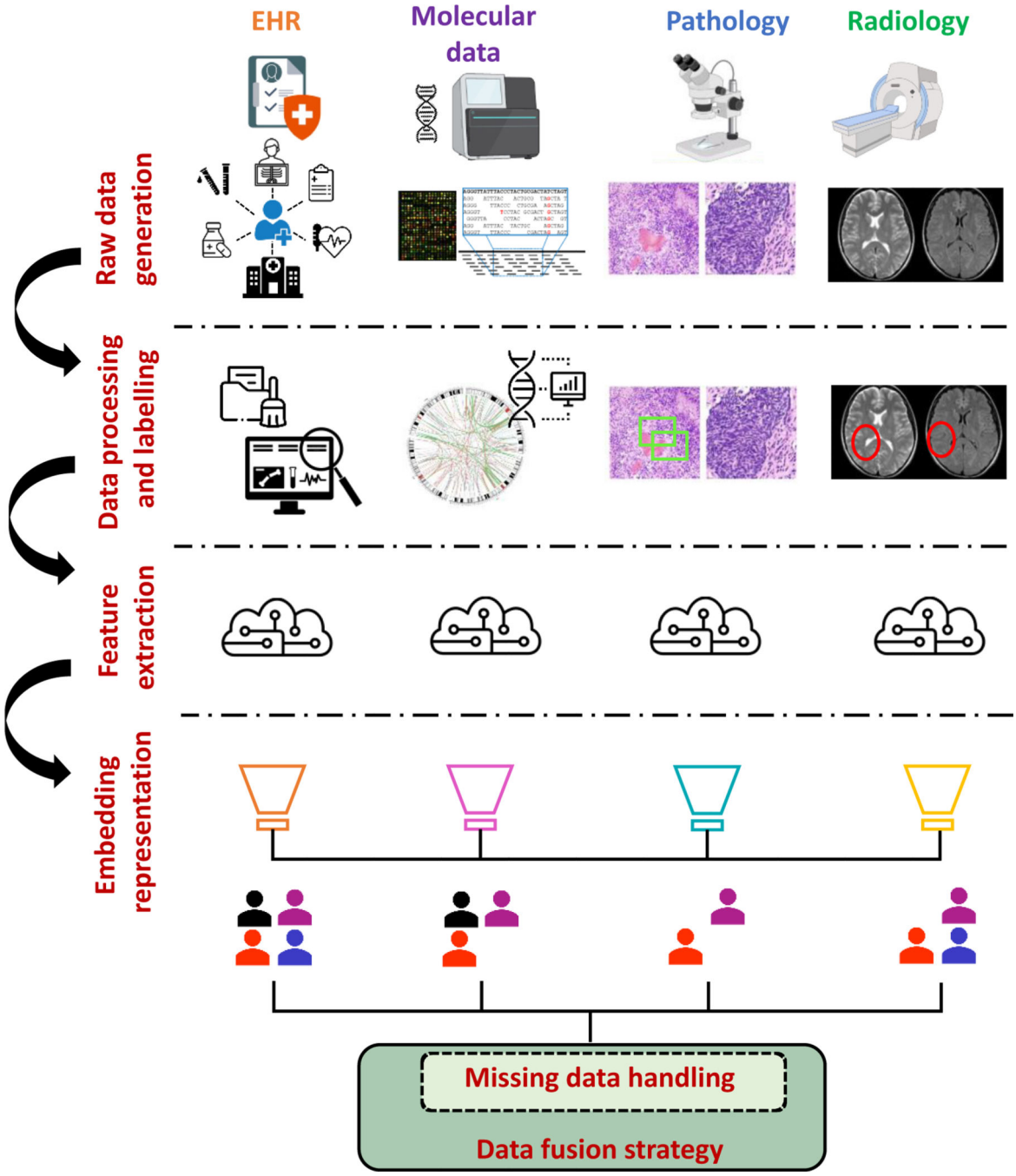

Fig. 1: Generation and processing of routinely collected biomedical modalities in oncology.
Prior to data fusion, different steps are needed to go from the raw data to workable data representations for each modality, e.g. EHRs, molecular data and medical images.

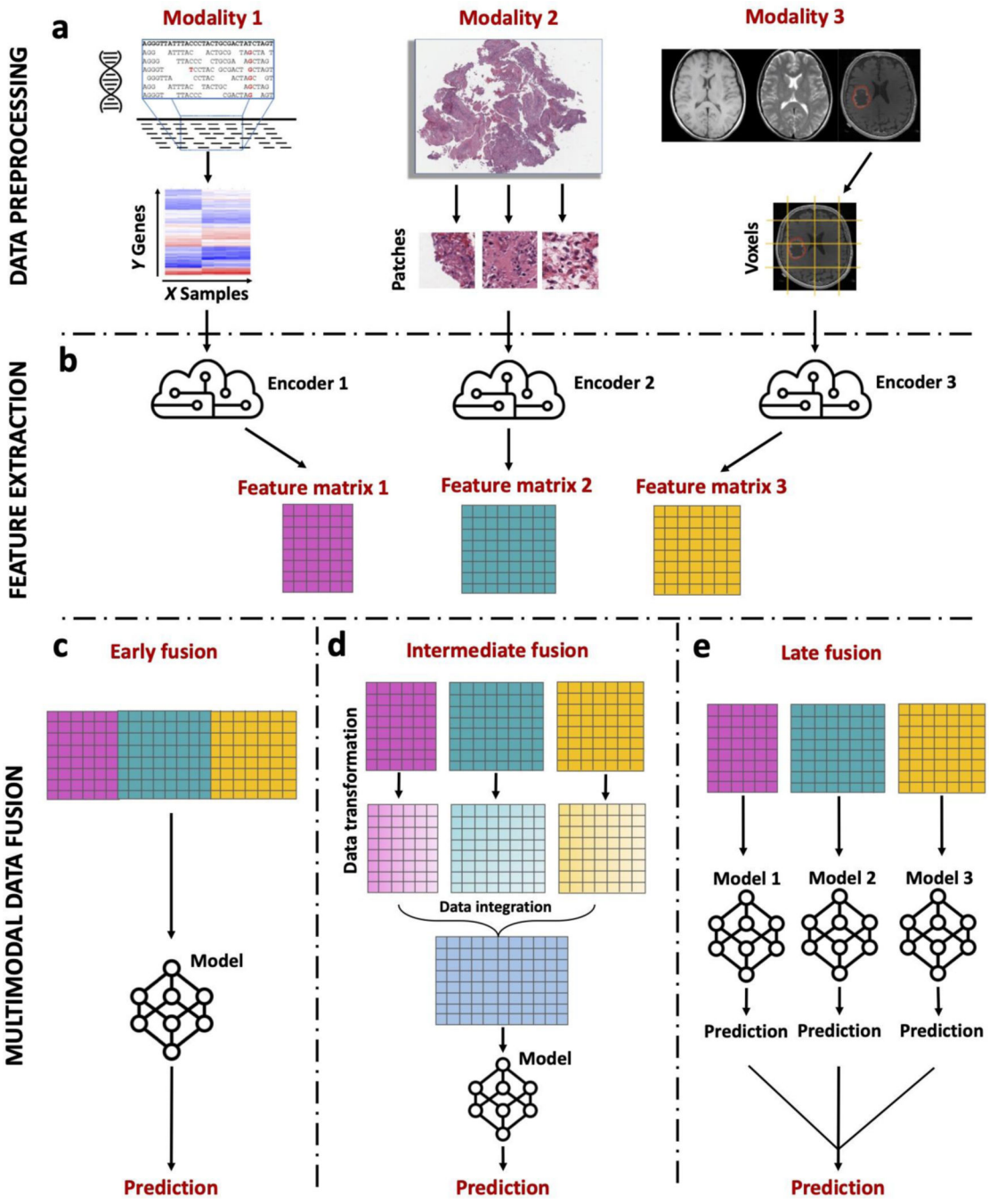

Fig. 2: Overview of different fusion strategies for multimodal data.

a) Raw data is processed into workable formats. b) For each modality features are extracted using dedicated encoder algorithms. c) Early fusion. d) Intermediate fusion. e) Late fusion.

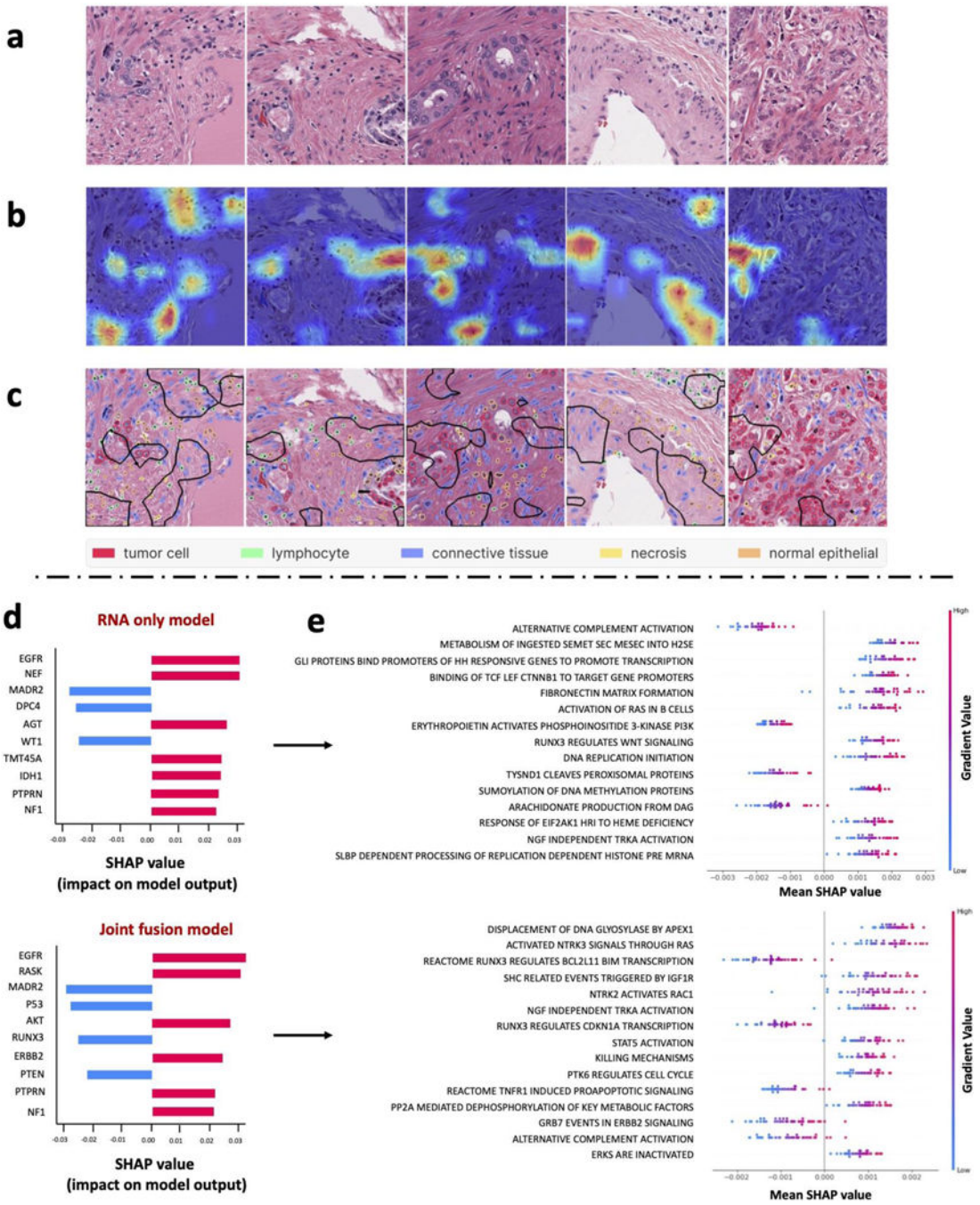

Fig. 3: Examples of model interpretability methods for histopathology and gene expression.

Histopathology: a) Examples of informative tiles for predicting the presence of TP53 mutations from histopathology images in prostate cancer (unpublished data). b) Visualisation of regions within tiles most relevant to the prediction, derived via Grad-CAM[151]. c) Individual cells within informative tiles are segmented and classified by Hover-Net[155]. For a fine-grained interpretation of relevant cells (black annotations), pertinent cells within the tile are encircled by calculating the contours from regions highlighted by Grad-CAM. **Gene Expression**: d) Examples of SHAP visualisation[152] of hypothetical gene

importance according to unimodal model (top) and joint multimodal model (bottom) for cancer survival prediction. e) Example of pathway importance visualisation based on the respective gene SHAP-values in unimodal (top) versus joint multimodal (bottom) models with respect to cancer survival prediction[154].