nature machine intelligence

Perspective

https://doi.org/10.1038/s42256-023-00633-5

Multimodal data fusion for cancer biomarker discovery with deep learning

Received: 28 July 2022

Accepted: 17 February 2023

Published online: 6 April 2023

Check for updates

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]

Technological advances have made it possible to study a patient from multiple angles with high-dimensional, high-throughput multiscale 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 greatly 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 personalized 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 standardization of datasets.

Over recent decades, technological innovations have transformed the healthcare domain with the 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], and paves the way for precision medicine, which promises individualized 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. Although molecular pathology revolutionized precision oncology, the first Food and Drug Administration (FDA)-cleared companion diagnostic 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 multitarget companion diagnostic assays, which are becoming more prevalent[8,9]. The continuing cost reduction makes it possible to simultaneously profile thousands of genomic regions, hinting that multitarget panels could soon be run at a similar price point to that of testing five to ten targets individually[10]. Multitarget 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 and Oxford Nanopore Technologies enable the assembly and exploration of 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 eight hours[12]. As cancer is often 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, notable progress in whole-slide imaging (WSI) has led to a transition

[1]Stanford Center for Biomedical Informatics Research (BMIR), Department of Medicine, Stanford University, Stanford, CA, USA. [2]Department of Computer Science, Stanford University, Stanford, CA, USA. [3]Department of Biomedical Data Science, Stanford University, Stanford, CA, USA.
e-mail: steyaert@stanford.edu; ogevaert@stanford.edu

from traditional histopathology methods towards digital pathology[16]. Digital pathology, the process of 'digitizing' 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 and magnetic resonance imaging (MRI) scans are useful for generating three-dimensional 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 relies on specialized 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 such as convolutional neural networks, deep autoencoders and vision transformers are now available for automated, high-throughput feature extraction[21–24]. Automatic assessment of deterministic objective features has enabled the quantification of tumour microenvironments (TMEs) at unprecedented speed and scale. In addition to the quantification of known handcrafted salient features without inter-observer variability, DL has the ability to discover unknown features and relationships that can provide biological insights and improve disease characterization[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 the diverse capacity of DL, one of the main challenges is the need for large datasets to train, test and validate its algorithms. But, owing to ethical restrictions and the labour intensity to annotate clinical images, most studies have only 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 are being captured from multiple sources at unprecedented pace, ranging from medical images to genomics data and patient-generated health data. 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. So far, the FDA has cleared and approved several AI-based software as a medical device[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 (1) relevance to the clinical problem and current clinical practice, (2) data acquisition in a consistent, generalizable and clinically relevant manner, (3) appropriate definition and separation of training, tuning and test sets, and (4) 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 have 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 aspects of multimodal integration[32–35]. To advance precision oncology, healthcare AI should not only inform about cancer incidence and tumour growth but also must identify the optimal treatment path, accounting for treatment-related side effects, socioeconomic factors and care goals. Precision medicine can therefore be achieved only 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 prognostic biomarkers provide information on the patient's diagnosis and overall outcome, predictive biomarkers inform 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 patients with cancer, because they are studied mostly in isolation instead of in an integrated fashion. These are: (1) electronic health records (EHRs), (2) molecular data, (3) digital pathology and (4) 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 individualized care (Fig. 1). In the next sections, we discuss both challenges and opportunities for multimodal biomarker discovery as it applies to patients with cancer. 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 such as antibody–drug conjugates and adoptive cell therapy (for example, T-cell receptor and chimeric antigen receptor T-cell therapy) have shown to be promising, response rates vary markedly depending on the tumour subtype[43] and the TME[44]. Various TME elements play a role in tumour development and also in the 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 (1) a deeper understanding of the TME and (2) 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 such as The Cancer Genome Atlas (TCGA) and Genomic Data Commons have gathered and standardized 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 (for example, 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

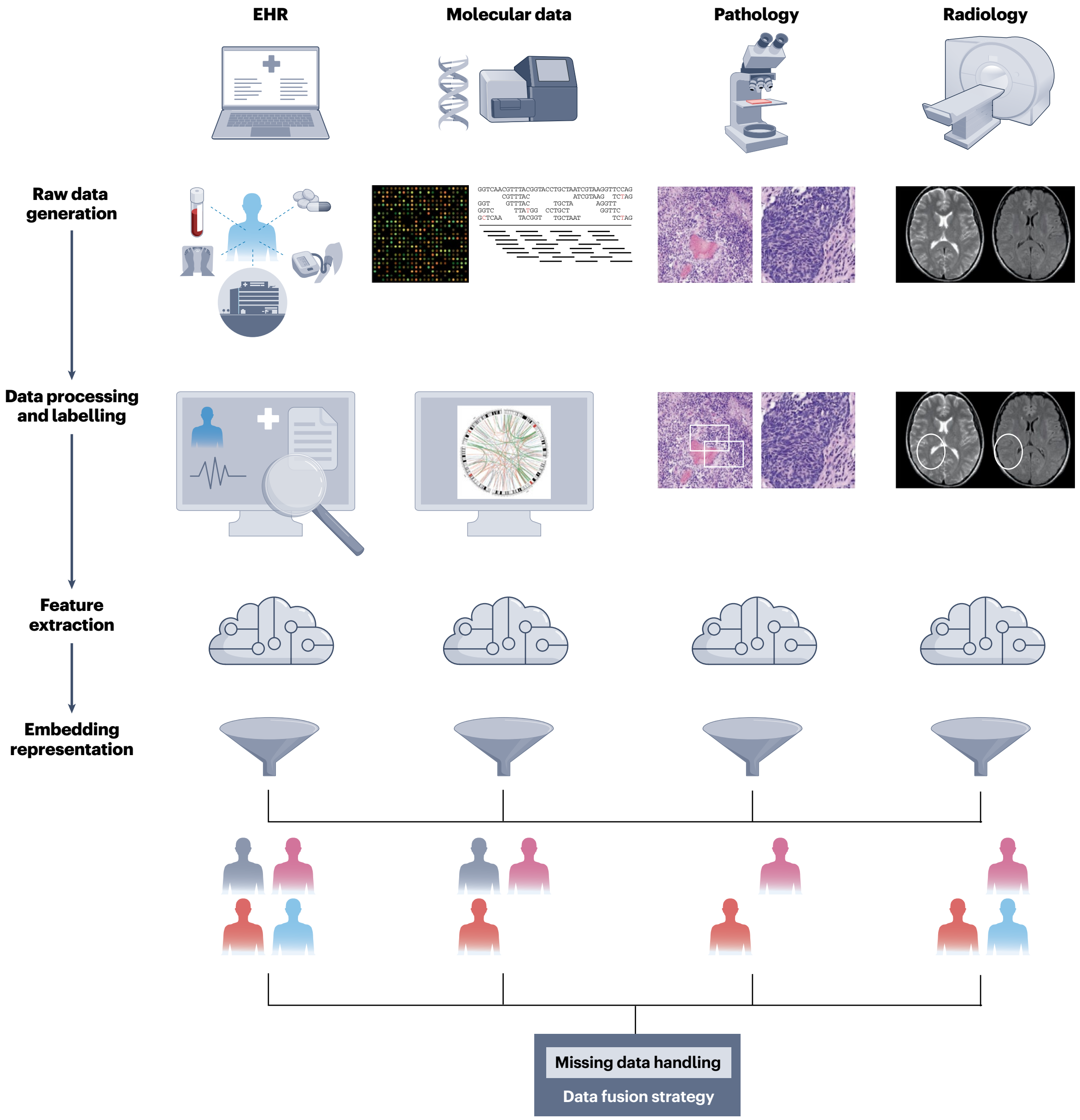


Fig. 1 | Generation and processing of routinely collected biomedical modalities in oncology. Before data fusion, different steps are needed to go from the raw data to workable data representations for each modality—for example, EHRs, molecular data and medical images. Icon credits: microarray, Guillaume Paumier, under a Creative Commons licence CC BY-SA 3.0; EHR, data processing, DNA and encoder icons, the Noun Project (https://thenounproject.com/); DNA sequencer, MRI machine and stethoscope, created with Biorender.com; genomic circular circus plot, ref. 166, Cold Spring Harbor Laboratory Press; tissue and brain slices, created using TCGA data originally published by the National Cancer Institute.

about the TME's spatial heterogeneity using image analysis and spatial statistics[52]. Similarly, radiographic images such as MRI or computed tomography scans provide visual data of the tissue morphology and three-dimensional 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 World Health Organization classification of tumours of the central nervous system revisited the guidelines to classify diffuse gliomas, recommending histopathological diagnosis in combination with molecular markers (for example, *isocitrate dehydrogenase 1 and 2* (*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 central nervous system tumours[34,35].

The need for integrative modelling is increasingly emphasized. In 2015, a report from Ritchie et al.[57] highlighted that "approaches to combine multiple data types provide a more comprehensive understanding of complex genotype–phenotype associations than analysis of one dataset". In recent years, there have been several attempts to develop multimodal approaches, to a great degree stimulated by community-driven competitions, such as DREAM and Kaggle (that is, 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 two-dimensional in histopathology or three-dimensional 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 convolutional neural networks[59–61]. Continuous improvement in detection, segmentation, classification and spatial characterization 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, these data are embedded into a lower-dimensional vector space and fed as input to a recurrent neural network (RNN). Long short-term memory and gated recurrent unit 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 to mine the unstructured narratives of patient records. One way to process these data is to convert free text to medical concepts and create lower-dimensional 'concept embeddings'. Older methods such as Word2Vec[67] and global vectors for word representations (GloVe)[68] have almost been overtaken by 'contextualized embeddings' such as embeddings from language models (ELMo)[69] and bidirectional encoder representations from transformers (BERT)[70–72]. While ELMo uses RNNs, BERT is based on transformers, a neural architecture that has revolutionized the NLP field since its inception[73]. To unlock the full potential of EHRs, 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 unvectorized 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 neural networks are needed before 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: (1) early, (2) intermediate or (3) late (Fig. 2)[81–83]. Early fusion is characterized 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). Late fusion not only allows the use of a different, often more suitable, model for each modality but 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 develop 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 for modeling 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 increasing 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 has also published numerous papers investigating the integration of genomic, transcriptomic, epigenomic and proteomic data for multiple cancer types[94–96]. Additionally, for therapy response and drug combination predictions, multi-omics ML methods have 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].

Multiscale data fusion

Similar efforts as for multi-omics data fusion have been explored for multiscale data[89,102–107]. For example, Cheerla and Gevaert[48] 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 RNA sequencing, microRNA sequencing, 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, owing to its high dimensionality and computational demands, so far most studies have avoided its inclusion[112].

Imaging genomics and 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 such as an *epidermal growth factor receptor* (*EGFR*) mutation in lung cancer[113–115]. This discovery immediately gave rise to an emerging field called 'radiogenomics',

https://doi.org/10.1038/s42256-023-00633-5

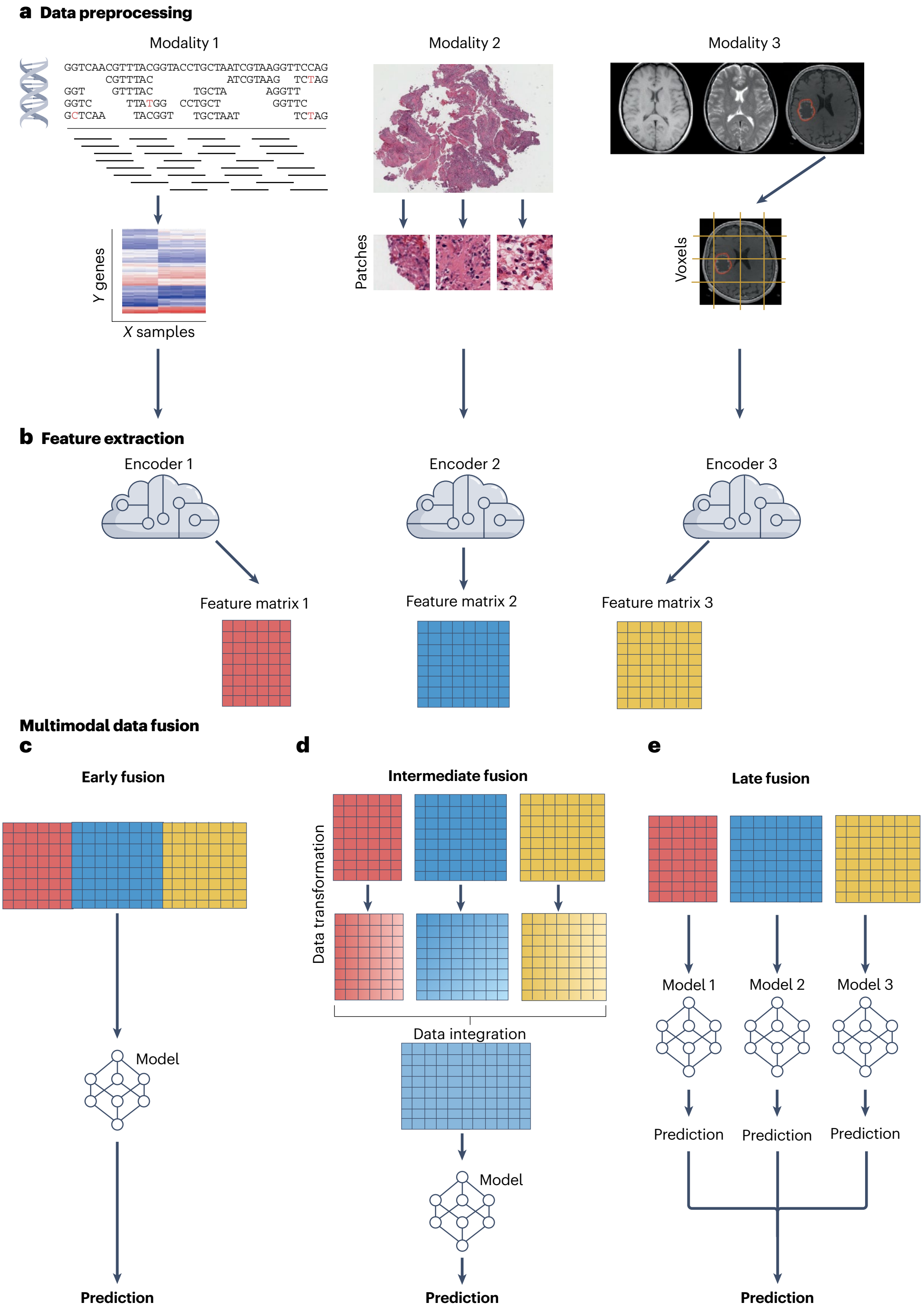


Fig. 2 | Overview of different fusion strategies for multimodal data. a, Raw data are processed into workable formats. **b**, For each modality, features are extracted using dedicated encoder algorithms. **c**, Early fusion. **d**, Intermediate fusion. **e**, Late fusion. Icon credits: **a**, DNA icon, the Noun Project (https://thenounproject.com/); tissue and brain slices, created using TCGA data originally published by the National Cancer Institute; **b**, encoder, the Noun Project (https://thenounproject.com/); **c**, model icon, the Noun Project (https://thenounproject.com/).

the study of directly linking image features to underlying molecular properties[116]. For example, Itakura et al.[117] used MRI phenotypes to define subtypes of glioblastoma associated with molecular pathway activity. 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 probably the only way to advance precision oncology, but many challenges exist to realize their full potential. Although data availability is the main driver of multimodal data fusion, it also poses the major barrier. DL requires large amounts of data, and both data sparsity and scarcity 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 owing to cost, insurance coverage, material availability and lack of systemic collection procedures, among 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 such as multiple imputation, multivariate imputation by chained equations or neural networks such as RNNs, long short-term memory and generative adversarial networks[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. First, the depth of data per patient, that is, 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 generalization. Second, 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 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 visualizing 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 such as gradient-weighted class activation mapping (Grad-CAM) (Fig. 3b)[151]. Similarly, for molecular data, predictive features can be determined and visualized via Shapley additive explanation (SHAP)-based methods (Fig. 3d,e)[150,152–154]. Multimodal data add additional complexity and need 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 results on TCGA data).

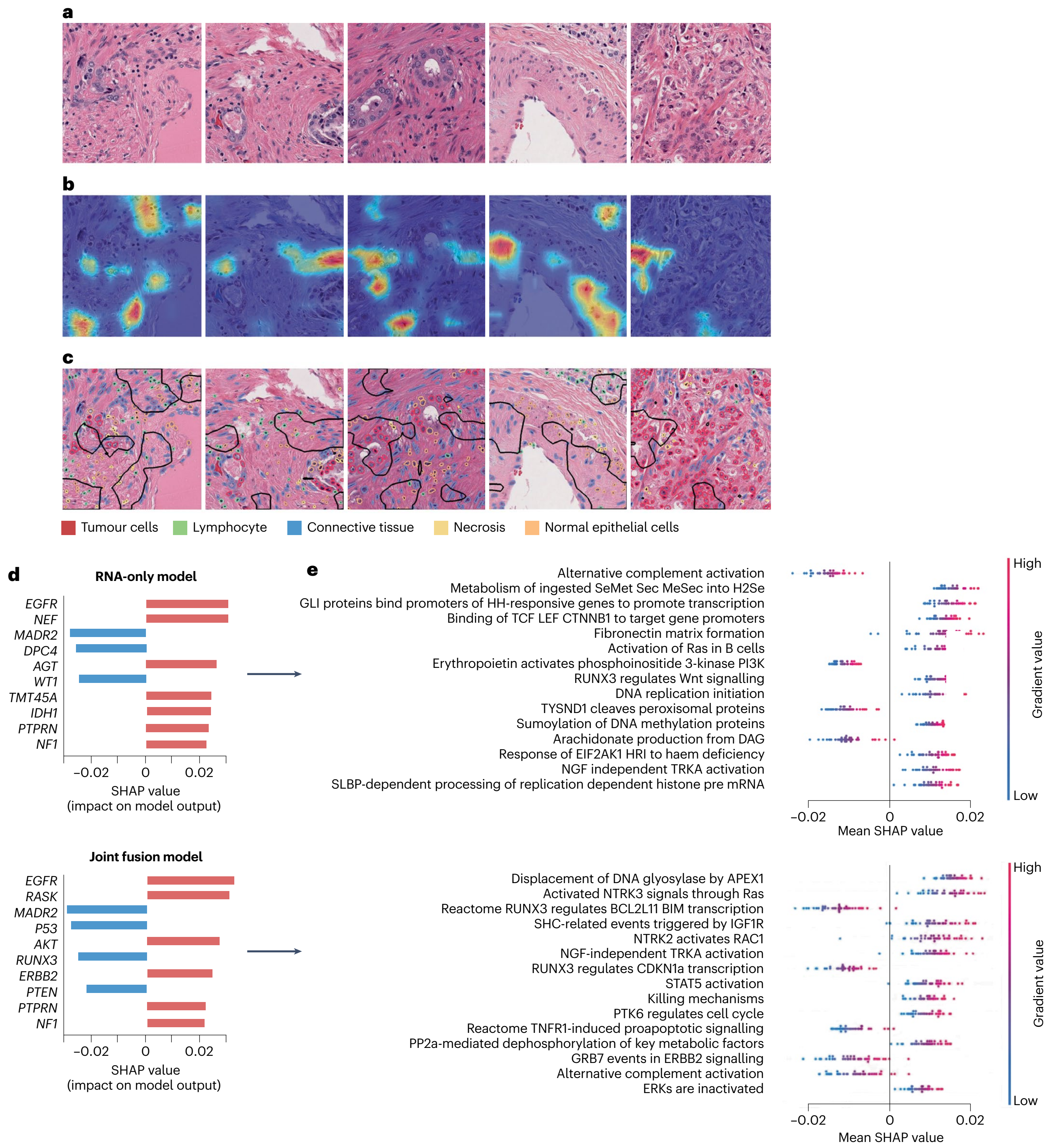
Standardization will lead to more uniform and complete datasets, which are easier to process and fuse with other sources and will be much more interpretable on their own. TCGA is probably the best known and most used resource[37], but many other initiatives are underway to structurally capture clinical, genomics, imaging and pathological data for oncology, such as The Cancer Imaging Archive[36] and the Genomics Pathology Imaging Collection[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 encourage reutilization and optimization 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 standardize and share but also very expensive to obtain by researchers[158,159]. Efforts such as the Observational Medical Outcomes Partnership (OMOP) aim at tackling this issue by harmonizing 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 become 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

Fig. 3 | Examples of model interpretability methods for histopathology and gene expression. a–**c**, Histopathology. **a**, Examples of informative tiles for predicting the presence of *TP53* mutations from histopathology images in prostate cancer (unpublished results on TCGA data). **b**, Visualization 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. **d**,**e**, Gene expression. **d**, Examples of SHAP visualization[152] of hypothetical gene importance according to a unimodal model (top) and a joint multimodal model (bottom) for cancer survival prediction. **e**, Example of pathway importance visualization based on the respective gene SHAP values in unimodal (top) versus joint multimodal (bottom) models with respect to cancer survival prediction[154]. SeMet, selenomethionine; Sec, selenocysteine; MeSec, methylselenol; H2Se, hydrogen selenide; GLI, glioma-associated oncogene family zinc finger 1; HH, hedgehog; TCF, T cell factor; LEF, lymphoid enhancer factor; CTNNB1, catenin beta 1; Ras, rat sarcoma; PI3K, phosphatidylinositol 3-kinase; RUNXC3, runt-related transcription factor 3; Wnt, wingless/integrated; TYSND1, trypsin-like peroxisomal matrix peptidase 1; DAG, diacylglycerol; EIF2AK1, eukaryotic translation initiation factor 2 alpha kinase 1; HRI, heme-regulated inhibitor; NGF, nerve growth factor; TRKA, tropomyosin receptor kinase A; SLBP, stem-loop binding protein; APEX1, apurinic/apyrimidinic endodeoxyribonuclease 1; NTRK2/3, neurotrophic receptor tyrosine kinase 2/3; BCL2L11, B cell lymphoma 2-like 11; BIM, B cell lymphoma 2 interacting mediator of cell death; SHC, src homology 2 domain containing transforming protein; IGF1R, insulin-like growth factor 1 receptor; RAC1, ras-related C3 botulinum toxin substrate 1; CDKN1a, cyclin-dependent kinase inhibitor 1A; STAT5, signal transducer and activator of transcription 5; PTK6, protein tyrosine kinase 6; TNFR1, tumor necrosis factor receptor 1; PP2a, protein phosphatase 2A; GRB7, growth factor receptor bound protein 7; ERBB2, v-erb-b2 avian erythroblastic leukemia viral oncogene homolog 2; ERK, extracellular signal-regulated kinase.

of the challenges inherent to the current state of healthcare data as discussed above, again emphasizing the need for large collaborative data standardization and sharing efforts. In this space, competitions such as DREAM and Kaggle have been an effective concept for making standardized 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 considerably 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 the evaluation of 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.

References

1. Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* **25**, 44–56 (2019).
2. Riba, M., Sala, C., Toniolo, D. & Tonon, G. Big data in medicine, the present and hopefully the future. *Front. Med.* **6**, 263 (2019).
3. Hanahan, D. Hallmarks of cancer: new dimensions. *Cancer Discov.* **12**, 31–46 (2022).
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).
5. Medina-Martinez, J. S. et al. Isabl platform, a digital biobank for processing multimodal patient data. *BMC Bioinformatics* **21**, 549 (2020).
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).
8. Malone, E. R., Oliva, M., Sabatini, P. J. B., Stockley, T. L. & Siu, L. L. Molecular profiling for precision cancer therapies. *Genome Med.* **12**, 8 (2020).
9. Campbell, M. R. Update on molecular companion diagnostics—a future in personalized medicine beyond Sanger sequencing. *Expert Rev. Mol. Diagn.* **20**, 637–644 (2020).
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, E. L., Jaszczyszyn, Y., Naquin, D. & Thermes, C. The third revolution in sequencing technology. *Trends Genet.* **34**, 666–681 (2018).
12. Gorzynski, J. E. et al. Ultrarapid nanopore genome sequencing in a critical care setting. *N. Engl. J. Med.* **386**, 700–702 (2022).
13. Davidson, M. R., Gazdar, A. F. & Clarke, B. E. The pivotal role of pathology in the management of lung cancer. *J Thorac. Dis.* **5**, S463–S478 (2013).
14. Pomerantz, B. J. Imaging and interventional radiology for cancer management. *Surg. Clin. North Am.* **100**, 499–506 (2020).
15. Yu, K. H. & Snyder, M. Omics profiling in precision oncology. *Mol. Cell. Proteomics* **15**, 2525–2536 (2016).
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).
17. van der Laak, J., Litjens, G. & Ciompi, F. Deep learning in histopathology: the path to the clinic. *Nat. Med.* **27**, 775–784 (2021).
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).
19. Serag, A. et al. Translational AI and deep learning in diagnostic pathology. *Front. Med.* **6**, 185 (2019).
20. Iv, M. et al. MR imaging-based radiomic signatures of distinct molecular subgroups of medulloblastoma. *Am. J. Neuroradiol.* **40**, 154–161 (2019).
21. van Timmeren, J. E., Cester, D., Tanadini-Lang, S., Alkadhi, H. & Baessler, B. Radiomics in medical imaging—'how-to' guide and critical reflection. *Insights Imaging* **11**, 91 (2020).
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).
23. Kim, M. et al. Deep learning in medical imaging. *Neurospine* **16**, 657–668 (2019).
24. Dosovitskiy, A. et al. An image is worth 16x16 words: transformers for image recognition at scale. Preprint at https://arxiv.org/abs/2010.11929 (2020).
25. Gupta, R., Kurc, T., Sharma, A., Almeida, J. S. & Saltz, J. The emergence of pathomics. *Curr. Pathobiol. Rep.* **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, D. C., Walker, I. & Glocker, B. Causality matters in medical imaging. *Nat. Commun.* **11**, 3673 (2020).
28. *21st Century Cures Act. H.R. 34* (114th Congress, 2016); https://www.congress.gov/114/bills/hr134/BILLS-114hr134enr.pdf
29. Artificial intelligence and machine learning (AI/ML)-enabled medical devices. *FDA* (5 October 2022); https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices
30. *Proposed Regulatory Framework for Modification to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD)* (FDA, 2019); https://www.fda.gov/files/medical%20devices/published/US-FDA-Artificial-Intelligence-and-Machine-Learning-Discussion-Paper.pdf
31. Kann, B. H., Thompson, R., Thomas, C. R. Jr., Dicker, A. & Aneja, S. Artificial intelligence in oncology: current applications and future directions. *Oncology* **33**, 46–53 (2019).
32. Louis, D. N. et al. The 2016 World Health Organization classification of tumors of the central nervous system: a summary. *Acta Neuropathol.* **131**, 803–820 (2016).
33. Tateishi, K., Wakimoto, H. & Cahill, D. P. IDH1 mutation and World Health Organization 2016 diagnostic criteria for adult diffuse gliomas: advances in surgical strategy. *Neurosurgery* **64**, 134–138 (2017).
34. Capper, D. et al. DNA-methylation-based classification of central nervous system tumours. *Nature* **555**, 469–474 (2018).
35. Ceccarelli, M. et al. Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma. *Cell* **164**, 550–563 (2016).
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, J. C. The Cancer Genome Atlas: creating lasting value beyond its data. *Cell* **173**, 283–285 (2018).
38. Jennings, C. N. et al. Bridging the gap with the UK Genomics Pathology Imaging Collection. *Nat. Med.* **28**, 1107–1108 (2022).
39. Mo, H., Breitling, R., Francavilla, C. & Schwartz, J. M. Data integration and mechanistic modelling for breast cancer biology: current state and future directions. *Curr. Opin. Endocr. Metab. Res.* **24**, 100350 (2022).
40. Nalejska, E., Maczynska, E. & Lewandowska, M. A. Prognostic and predictive biomarkers: tools in personalized oncology. *Mol. Diagn. Ther.* **18**, 273–284 (2014).
41. Grossman, J. E., Vasudevan, D., Joyce, C. E. & 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).
42. Davis, A. A. & Patel, V. G. 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).
43. van Elsas, M. J., van Hall, T. & van der Burg, S. H. Future challenges in cancer resistance to immunotherapy. *Cancers* **12**, 935 (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).

45. Ott, M., Prins, R. M. & Heimberger, A. B. The immune landscape of common CNS malignancies: implications for immunotherapy. *Nat. Rev. Clin. Oncol.* **18**, 729–744 (2021).
46. Bejarano, L., Jordao, M. J. C. & Joyce, J. A. Therapeutic targeting of the tumor microenvironment. *Cancer Discov.* **11**, 933–959 (2021).
47. Zomer, A., Croci, D., Kowal, J., van Gurp, L. & Joyce, J. A. 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).
49. Grossman, R. L. et al. Toward a shared vision for cancer genomic data. *N. Engl. J. Med.* **375**, 1109–1112 (2016).
50. Hinkson, I. V. et al. A comprehensive infrastructure for big data in cancer research: accelerating cancer research and precision medicine. *Front. Cell Dev. Biol.* **5**, 83 (2017).
51. Putcha, G., Gutierrez, A. & Skates, S. Multicancer screening: one size does not fit all. *JCO Precis. Oncol.* **5**, 574–576 (2021).
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).
54. Lanckriet, G. R., De Bie, T., Cristianini, N., Jordan, M. I. & Noble, W. S. A statistical framework for genomic data fusion. *Bioinformatics* **20**, 2626–2635 (2004).
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–e190 (2006).
56. Daemen, A. et al. A kernel-based integration of genome-wide data for clinical decision support. *Genome Med.* **1**, 39 (2009).
57. Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A. & Kim, D. Methods of integrating data to uncover genotype–phenotype interactions. *Nat. Rev. Genet.* **16**, 85–97 (2015).
58. Panayides, A. S. et al. AI in medical imaging informatics: current challenges and future directions. *IEEE J. Biomed. Health Inform.* **24**, 1837–1857 (2020).
59. George, K., Faziludeen, S., Sankaran, P. & Joseph, K. P. Breast cancer detection from biopsy images using nucleus guided transfer learning and belief based fusion. *Comput. Biol. Med.* **124**, 103954 (2020).
60. Singh, S. P. et al. 3D deep learning on medical images: a review. *Sensors* **20**, 5097 (2020).
61. Sarvamangala, D. R. & Kulkarni, R.V. Convolutional neural networks in medical image understanding: a survey. *Evol. Intell.* **15**, 1–22 (2021).
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).
64. Shickel, B., Tighe, P. J., 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).
65. Ayala Solares, J. R. 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, K. L., Crespo, B. C. & 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).
67. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. & Dean, J. Distributed representations of words and phrases and their compositionality. In *Proc. 26th International Conference on Neural Information Processing Systems* 3111–3119 (Curran Associates, Inc., 2013).
68. Pennington, J., Socher, R. & Manning, C. D. GloVe: global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* **14**, 1532–1543 (2014).
69. Peters, M. E. et al. Deep contextualized word representations. Preprint at 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. In *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* 4171–4186 (Association for Computational Linguistics, 2019).
71. Lee, J. et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**, 1234–1240 (2020).
72. Huang, K., Garapati, S. & Rich, A. S. An interpretable end-to-end fine-tuning approach for long clinical text. Preprint at https://arxiv.org/abs/2011.06504 (2020).
73. Vaswani, A. et al. Attention is all you need. In *Proc. 31st International Conference on Neural Information Processing Systems* 6000–6010 (Curran Associates, Inc., 2017).
74. Jensen, P. B., Jensen, L. J. & Brunak, S. Mining electronic health records: towards better research applications and clinical care. *Nat. Rev. Genet.* **13**, 395–405 (2012).
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).
76. Acosta, J. N., Falcone, G. J., Rajpurkar, P. & Topol, E. J. Multimodal biomedical AI. *Nat. Med.* **28**, 1773–1784 (2022).
77. Jain, M. S. et al. MultiMAP: dimensionality reduction and integration of multimodal data. *Genome Biol.* **22**, 346 (2021).
78. Lahnemann, D. et al. Eleven grand challenges in single-cell data science. *Genome Biol.* **21**, 31 (2020).
79. Baltrusaitis, T., Ahuja, C. & Morency, L. P. Multimodal machine learning: a survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **41**, 423–443 (2019).
80. Yan, K. K., Zhao, H. & Pang, H. A comparison of graph- and kernel-based -omics data integration algorithms for classifying complex traits. *BMC Bioinformatics* **18**, 539 (2017).
81. Pavlidis, P., Weston, J., Cai, J. & Noble, W. S. Learning gene functional classifications from multiple data types. *J. Comput. Biol.* **9**, 401–411 (2002).
82. Serra, A., Galdi, P. & Tagliaferri, R. in *Artificial Intelligence in the Age of Neural Networks and Brain Computing* 265–280 (eds Kozma, R., Alippi, C., Choe, Y., & Morabito, F. C.) (Academic Press, 2019).
83. Stahlschmidt, S. R., Ulfenborg, B. & Synnergren, J. Multimodal deep learning for biomedical data fusion: a review. *Brief Bioinformatics* **23**, bbab569 (2022).
84. Huang, S. C., Pareek, A., Seyyedi, S., Banerjee, I. & Lungren, M. P. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *npj Digit. Med.* **3**, 136 (2020).
85. Picard, M., Scott-Boyer, M. P., Bodein, A., Perin, O. & Droit, A. Integration strategies of multi-omics data for machine learning analysis. *Comput. Struct. Biotechnol. J.* **19**, 3735–3746 (2021).

86. 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).
87. Huang, Z. et al. SALMON: Survival Analysis Learning With Multi-Omics Neural Networks on breast cancer. *Front. Genet.* **10**, 166 (2019).
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).
89. Gevaert, O., Villalobos, V., Sikic, B. I. & Plevritis, S. K. 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 Bioinformtics* **20**, 527 (2019).
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).
92. Taskesen, E., Babaei, S., Reinders, M. M. & de Ridder, J. Integration of gene expression and DNA-methylation profiles improves molecular subtype classification in acute myeloid leukemia. *BMC Bioinformatics* **16**, 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).
94. Cancer Genome Atlas Research Network Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. *N. Engl. J. Med.* **372**, 2481–2498 (2015).
95. Cancer Genome Atlas Research Network Integrated genomic and molecular characterization of cervical cancer. *Nature* **543**, 378–384 (2017).
96. Cancer Genome Atlas Research Network Comprehensive and integrated genomic characterization of adult soft tissue sarcomas. *Cell* **171**, 950–965 e928 (2017).
97. Zhang, T., Zhang, L., Payne, P. R. O. & Li, F. Synergistic drug combination prediction by integrating multiomics data in deep learning models. *Methods Mol. Biol.* **2194**, 223–238 (2021).
98. Preuer, K. et al. DeepSynergy: predicting anti-cancer drug synergy with deep learning. *Bioinformatics* **34**, 1538–1546 (2018).
99. Sammut, S. J. et al. Multi-omic machine learning predictor of breast cancer therapy response. *Nature* **601**, 623–629 (2022).
100. Costello, J. C. et al. A community effort to assess and improve drug sensitivity prediction algorithms. *Nat. Biotechnol.* **32**, 1202–1212 (2014).
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, H. R. & Wang, M. D. Multimodal deep learning models for early detection of Alzheimer's disease stage. *Sci. Rep.* **11**, 3254 (2021).
103. Mobadersany, P. et al. Predicting cancer outcomes from histology and genomics using convolutional networks. *Proc. Natl Acad. Sci. USA* **115**, E2970–E2979 (2018).
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).
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, R. J. et al. Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. *IEEE Trans. Med. Imaging* **41**, 757–770 (2022).
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. *J. Pers. Med.* **12**, 601 (2022).
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).
110. Mazzaschi, G. et al. Integrated MRI-immune-genomic features enclose a risk stratification model in patients affected by glioblastoma. *Cancers* **14**, 3249 (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).
114. Gevaert, O. et al. Predictive radiogenomics modeling of EGFR mutation status in lung cancer. *Sci. Rep.* **7**, 41674 (2017).
115. Nair, J. K. R. 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).
116. Pinker, K., Chin, J., Melsaether, A. N., Morris, E. A. & Moy, L. Precision medicine and radiogenomics in breast cancer: new approaches toward diagnosis and treatment. *Radiology* **287**, 732–747 (2018).
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, D. D., Korn, R. L. & Kuo, M. D. Radiogenomic analysis of breast cancer using MRI: a preliminary study to define the landscape. *Am. J. Roentgenol.* **199**, 654–663 (2012).
119. Sutton, E. J. et al. Breast cancer subtype intertumor heterogeneity: MRI-based features predict results of a genomic assay. *J. Magn. Reson. Imaging* **42**, 1398–1406 (2015).
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**, 16012 (2016).
121. Li, J. et al. Imputation of missing values for electronic health record laboratory data. *npj Digit. Med.* **4**, 147 (2021).
122. Luo, Y. Evaluating the state of the art in missing data imputation for clinical data. *Brief Bioinformatics* **23**, bbab489 (2022).
123. Yoon, J., Zame, W. R. & 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).
124. Zhou, T., Liu, M., Thung, K. H. & 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).
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).

127. Momeni, A., Thibault, M. & Gevaert, O. Dropout-enabled ensemble learning for multi-scale biomedical data. Preprint at *bioRxiv* https://www.biorxiv.org/content/early/2018/10/11/440362 (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).
129. Ma, Q., Li, S. & Cottrell, G. W. Adversarial joint-learning recurrent neural network for incomplete time series classification. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 1765–1776 (2022).
130. Sharrocks, K., Spicer, J., Camidge, D. R. & Papa, S. The impact of socioeconomic status on access to cancer clinical trials. *Br. J. Cancer* **111**, 1684–1687 (2014).
131. Niranjan, S. J. et al. Perceived institutional barriers among clinical and research professionals: minority participation in oncology clinical trials. *JCO Oncol. Pract.* **17**, e666–e675 (2021).
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).
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, H. H., 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).
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. *Comput. Methods Programs Biomed.* **135**, 77–88 (2016).
137. Dwarikanath, M. Combining multiple expert annotations using semi-supervised learning and graph cuts for medical image segmentation. *Comput. Vision Image Understanding* **151**, 114–123 (2016).
138. Tran, Q. T., Alom, M. Z. & Orr, B. A. Comprehensive study of semi-supervised learning for DNA-methylation-based supervised classification of central nervous system tumors. *BMC Bioinformatics* **23**, 223 (2022).
139. Cheplygina, V., de Bruijne, M. & Pluim, J. P. W. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Med. Image Anal.* **54**, 280–296 (2019).
140. Jie, Y., Xutong, L. & Mingyue, Z. Current status of active learning for drug discovery. *Artif. Intell. Life Sci.* **1**, 100023 (2021).
141. Min, W., Fan, M., Zhi-Heng, Z. & Yan-Xue, W. Active learning through density clustering. *Expert Syst. Appl.* **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 Netw.* **153**, 325–338 (2022).
143. Park, Y., Hauschild, A. C. & 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, W. W. Biologically relevant transfer learning improves transcription factor binding prediction. *Genome Biol.* **22**, 280 (2021).
145. Ganoe, C. H. 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).
147. Lipkova, J. et al. Artificial intelligence for multimodal data integration in oncology. *Cancer Cell* **40**, 1095–1110 (2022).
148. Schaumberg, A. J. et al. Interpretable multimodal deep learning for real-time pan-tissue pan-disease pathology search on social media. *Mod. Pathol.* **33**, 2169–2185 (2020).
149. Begoli, E., Bhattacharya, T. & Kusnezov, D. The need for uncertainty quantification in machine-assisted medical decision making. *Nat. Mach. Intell.* **1**, 20–23 (2019).
150. Chen, R. J. et al. Pan-cancer integrative histology-genomic analysis via multimodal deep learning. *Cancer Cell* **40**, 865–878.e6 (2022).
151. Selvaraju, R. R. et al. Grad-CAM: visual explanations from deep networks via gradient-based localization. Preprint at https://arxiv.org/abs/1610.02391 (2016).
152. Lundberg, S. M. & Lee, S. I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **30**, 4765–4774 (2017).
153. Dickinson, Q. & Meyer, J. G. 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. Preprint at *medRxiv* https://doi.org/10.1101/2022.09.21.22280223 (2022).
155. Simon, G. et al. Hover-Net: simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Med. Image Anal.* **58**, 101563 (2019).
156. Wilkinson, M. D. 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).
158. Mc Cord, K. A. 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, K. A. & Hemkens, L. G. Using electronic health records for clinical trials: where do we stand and where can we go? *CMAJ* **191**, E128–E133 (2019).
160. Makadia, R. & Ryan, P. B. Transforming the Premier Perspective Hospital Database into the Observational Medical Outcomes Partnership (OMOP) common data model. *EGEMS* **2**, 1110 (2014).
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. *Nat. Mach. Intell.* **4**, 669–677 (2022).
163. Costello, J. C. & Stolovitzky, G. Seeking the wisdom of crowds through challenge-based competitions in biomedical research. *Clin. Pharmacol. Ther.* **93**, 396–398 (2013).
164. Saez-Rodriguez, J. et al. Crowdsourcing biomedical research: leveraging communities as innovation engines. *Nat. Rev. Genet.* **17**, 470–486 (2016).
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).
166. Krzywinski, M. et al. Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).

Acknowledgements

We thank M. Humbert-Droz for discussions during the early stages of this Perspective. We are grateful for her insightful ideas and comments about these topics. We also express our great appreciation to C. Sadée and Y. Zheng for their valuable and constructive suggestions during the write-up of this Perspective.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence should be addressed to Sandra Steyaert or Olivier Gevaert.

Peer review information *Nature Machine Intelligence* thanks Peng Jiang, Krishna Bulusu, Po-Hsuan Chen and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

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

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.