npj | precision oncology

Perspective

Published in partnership with The Hormel Institute, University of Minnesota

https://doi.org/10.1038/s41698-024-00656-0

Radiology and multi-scale data integration for precision oncology

Hania Paverd[1,2,3], Konstantinos Zormpas-Petridis[4], Hannah Clayton[2,3], Sarah Burge[3] & Mireia Crispin-Ortuzar[2,3]

In this Perspective paper we explore the potential of integrating radiological imaging with other data types, a critical yet underdeveloped area in comparison to the fusion of other multi-omic data. Radiological images provide a comprehensive, three-dimensional view of cancer, capturing features that would be missed by biopsies or other data modalities. This paper explores the complexities and challenges of incorporating medical imaging into data integration models, in the context of precision oncology. We present the different categories of imaging-omics integration and discuss recent progress, highlighting the opportunities that arise from bringing together spatial data on different scales.

The advent of modern machine learning has completely shifted the landscape of predictive modelling for precision oncology. Applications within medical imaging in particular have boomed, achieving notable success. In parallel, there has been a growing interest in data integration, thanks to the development of datasets that span from the molecular scale of genes all the way to the macroscopic scale of medical scans. These datasets offer a holistic view of the tumour and its surrounding environment, with the potential of enabling tailor-made therapies for individual patients. However, applications that integrate radiological imaging are still comparatively underdeveloped. Instead, most of the attention has focused on the fusion of multi-omic molecular profiles, which already provide very rich, high-dimensional information. Multi-omic datasets also have the advantage of sharing the same subcellular physical scale. Sometimes, thanks to the availability of the data, studies have been extended to include digital pathology, bringing in a new scale—cells and their microenvironment—and introducing the need to perform two-dimensional spatial analysis[1–3].

Including radiological images, however, is a completely different challenge. They capture the entirety of the disease, macroscopically and in three dimensions. If the cancer is metastatic, usually the majority of the lesions will be included in the scan, introducing a source of heterogeneity that is inaccessible for biopsies. The resolution of a typical computed tomography (CT) or magnetic resonance imaging (MRI) scan is of the order of a millimetre, orders of magnitude larger than the pixels in a microscopy image. Moreover, scanners and image acquisition protocols can result in noticeable and highly variable noise and image artefacts, threatening the generalisability of any result. These differences are challenging, but they are also the reason why including radiological imaging in fusion models can add significant value, as no other data type contains comparable information of such critical importance for the patient's journey. In addition, advanced imaging methods can provide a wealth of additional data: from in-vivo functional imaging using positron emission tomography[4], to measures of blood flow with MRI perfusion techniques[5], physical properties of tissues, such as liver stiffness with MR elastography[6], and molecular displacement with diffusion weighted imaging[7]. Due to its non-invasive nature, imaging is far more often performed at regular intervals in clinical practice compared to other data modalities (e.g. biopsies), bringing temporal dimension to data integration.

At a time of rapid development of artificial intelligence (AI) models, we want to shed light on the complexities of integrating medical imaging data in data integration models for precision oncology. The majority of the work in this space relies on well-established machine learning (ML) methods that can provide the robustness required. We illustrate this point in Section "Data fusion", which reviews the literature on the prediction of clinical endpoints using imaging-omics fusion. We also discuss current approaches and challenges in data translation (Section "Data translation") and data aggregation (Section "Data aggregation"), which are parallel but equally critical elements of the data integration paradigm for precision oncology.

Types of data integration

Data integration is a broad term that can refer to different aspects or objectives within the process of bringing different data types together. In this review we explore three broad types of data integration; namely data fusion, data aggregation, and data translation, as illustrated in Fig. 1a.

We use the term *data fusion* to refer to the combined use of data from diverse sources in a single model, with the aim to predict clinical endpoints such as survival or treatment response with higher accuracy than would be

[1]Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK. [2]Department of Oncology, University of Cambridge, Cambridge, UK. [3]Cancer Research UK Cambridge Centre, University of Cambridge, Cambridge, UK. [4]Fondazione Policlinico Universitario Agostino Gemelli IRCCS, Rome, Italy. e-mail: mc973@cam.ac.uk

The Hormel Institute University of Minnesota

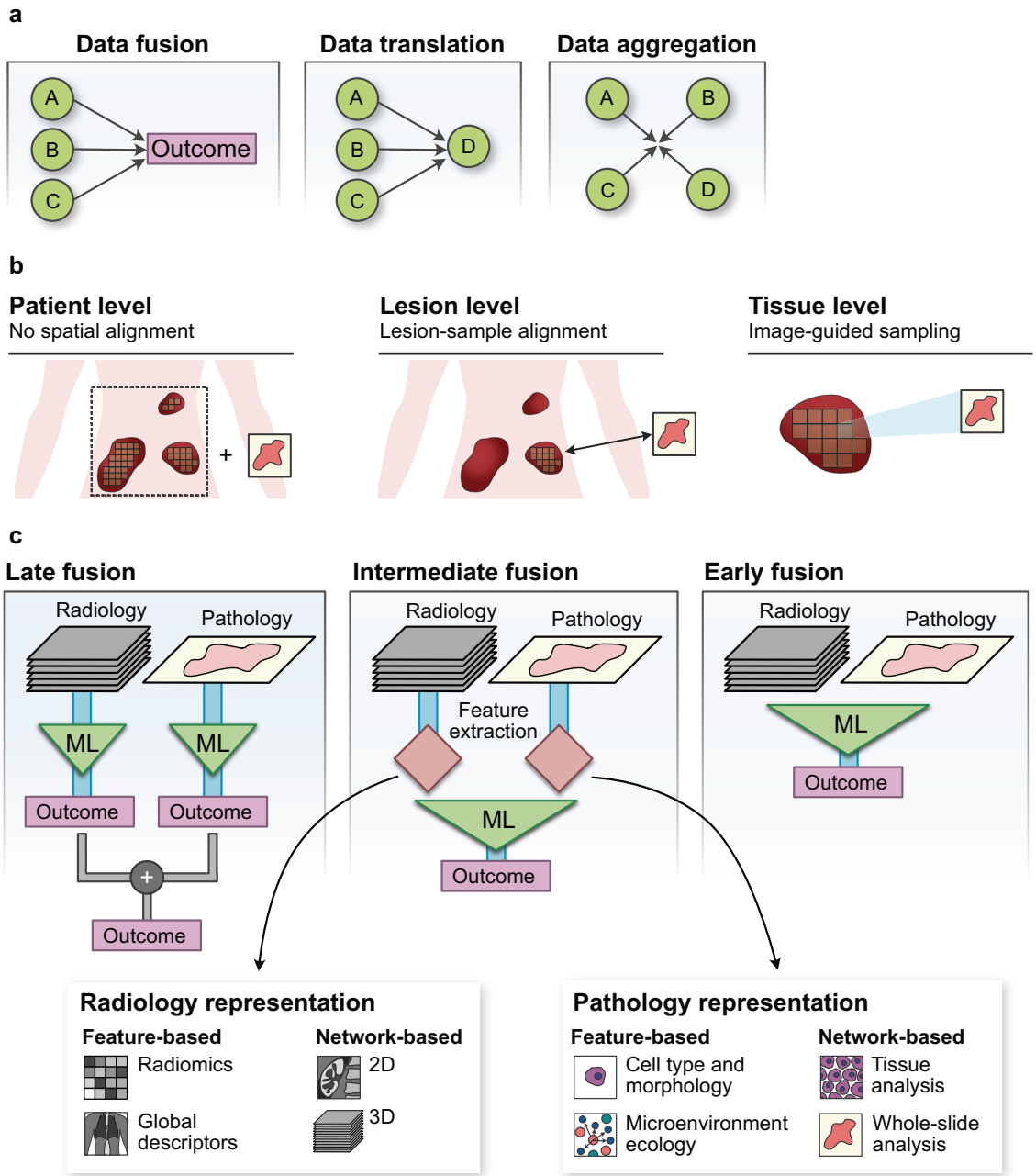

Fig. 1 | Types of data integration. a Simplified illustration of the three types of data integration processes considered in this review, classified as a function of the task and its objective. Data sources of different modalities are depicted in the figure with the letters A-D. Clinically-relevant endpoints are indicated by the label 'Outcome'. **b** Representative examples of the different levels of data fusion based on the physical scale at which the data is aligned, including patient-level (data sources are treated as independent), lesion-level (imaging features are matched to corresponding samples), or tissue-level (tissue samples are taken from specific locations using image guidance) fusion. **c** Different types of data fusion architectures in terms of the stage at which data streams are combined. The figure uses radiology and pathology images as an illustration. Clinically-relevant endpoints are indicated by the label 'Outcome'.

possible using each source separately. For example, a data fusion model may combine MRI scans, genomics and digital pathology to predict progression-free survival. The main challenges for data fusion are related to the disparity of the data modalities, which makes modelling difficult; and the lack of large enough datasets that contain all the data sources of interest. Data fusion is the main focus of this review, and we will explore it in detail in Section "Data fusion".

The term *data translation* is closely related to data fusion, and refers to models that predict the information from a given data type by using all the others jointly in a predictive model. For example, a data translation model may use clinical data and radiomics to predict histopathologic subtype. Data translation is particularly useful as an explainability tool, or for creating surrogate markers of data types which may be difficult to obtain. We explore data translation briefly in Section "Data translation".

Finally, *data aggregation* refers in this context to the systematic process of collecting, preparing, and presenting diverse datasets for subsequent analysis. For example, a data aggregation framework may link electronic health records with digitised histopathology slides, to enable the deployment of computational pathology models. Although data aggregation is a necessary step prior to other data integration tasks (i.e. data fusion or data translation), it can also be a goal in itself, for example enabling clinicians to access multiple data modalities seamlessly to improve direct patient care, or

to create online databases and biobanks. The main challenges of data aggregation are typically related to systems and processes, and include data integrity, system interoperability, and anonymisation. We provide an overview of the goals and challenges of data aggregation in Section "Data aggregation".

Data fusion

In data fusion models, the information from multiple modalities is combined to obtain a better predictor of a given clinical endpoint, which could be tumour detection, diagnosis, treatment response, or outcome. They represent the essence of the personalised precision oncology paradigm—in other words, the idea that the treatment journey of a cancer patient can be tailored using their unique, individual, multimodal data profile.

One of key decisions for data fusion models is the level at which the different modalities will be integrated, namely at the patient, lesion, or tissue levels (Fig. 1b). Typically, each data modality is treated as providing an independent patient-level representation, with no attempt at spatial co-registration. When the data is sufficiently annotated, it is also possible to provide lesion-level alignment, by matching specific samples with the images corresponding to the lesion they were extracted from. Tissue-level alignment, in which samples can be traced to specific locations within the lesion, normally requires advanced spatial techniques or careful sampling using image guidance, as discussed in Section "Data translation".

In addition, fusion models are typically divided into late, intermediate or early fusion[8,9], as illustrated in Fig. 1c. In late fusion approaches, single-modality models are developed independently and merged in the final layers. This reduces the complexity of the training process and may also make it possible to re-utilise previously trained models. In early fusion, all modalities are processed simultaneously within a single deep learning architecture. Models where data fusion happens earlier typically require more data for learning but may be able to capture cross-modal interaction better[10–12]. More work is needed to develop and benchmark performance of such methods in the biomedical domain. Intermediate fusion models take early representations of the data for further learning. Fusion models that integrate medical images often opt for late or intermediate fusion approaches, in which imaging information is first collapsed down to a reduced, unidimensional set of features. One option is to generate lists of handcrafted features based on fixed definitions, which can be motivated by clinical or biological expectations (e.g. presence of disease in a certain area), or by mathematical characterisations of the texture of the image (e.g. Haralick features)[13].

Alternatively, these representations can be based on pre-trained deep learning models that act as feature extractors, which is particularly useful if the primary dataset is limited in size, as is often the case in clinical studies. Pre-training can also be used as a general approach, whereby models are trained on an auxiliary task and then fine-tuned on the dataset of interest, in a process known as transfer learning[14]. Radiomics feature extraction through transfer learning has been successfully applied in the development of models for computer-aided diagnosis as well as outcome prediction in a variety of cancer types, including lung[15,16], breast[15], gastric[17] and brain tumours[18]. Transfer learning can also be used for medical image segmentation, with the potential to reduce the need for manual annotation of the datasets and even to increase the segmentation accuracy, although consideration must be given to the pre-training and target dataset domains[19–21]. Self-supervised learning, a different type of pre-training that does not require annotated labels, has also been more recently shown to be effective[22,23].

Literature review

We conducted a PubMed search to identify predictive data fusion studies including medical imaging (18 June 2023). We focused our search on models predicting clinical outcome only, such as overall survival and progression-free survival[24]; and treatment response metrics, such as RECIST[25]. The exact query terms can be found in the Supplementary Table 1. Other related tasks such as diagnosis, which are closer to the data translation category, were not included in our search.

Our search identified a total of 27 studies. Their details are summarised in the Supplementary Table 2. Upon manual review, we excluded a total of 11 studies (one review, one abstract book, one database report, and eight studies that did not include an outcome prediction model). Of the remaining 16 papers that met the inclusion criteria, the majority (10) only integrated two data modalities, as shown in Fig. 2. Three studies integrated three data modalities, and three integrated four data modalities. Of the 16 studies that met the criteria, only six used deep learning, of which only two integrated more than two data modalities. The limitations of our review are discussed in Section "Discussion and outlook".

Machine learning vs deep learning. The majority of studies we reviewed use a late fusion approach, with separate models being trained for each data modality, followed by an integration layer of typically minimal complexity. Two studies conduct separate analyses (including ML models) for each single modality and then combine the derived signatures by statistical analysis[26,27]. For example, Wang et al.[26] use four different data modalities (haematoxylin and eosin slides, radiomics, immunohistochemistry and clinical data) to predict survival in colon cancer, but they derive a separate risk score for each modality and then integrate them into a combined nomogram by multivariate analysis.

Two papers with a late fusion approach use traditional ML methods for the integration step. For example, Feng et al.[28] derive separate radiomics and pathomics signatures to predict response to neoadjuvant chemotherapy in rectal cancer, and combine them into a predictive model using a support vector machine (SVM). Wang et al.[29] also use SVM in a late fusion model

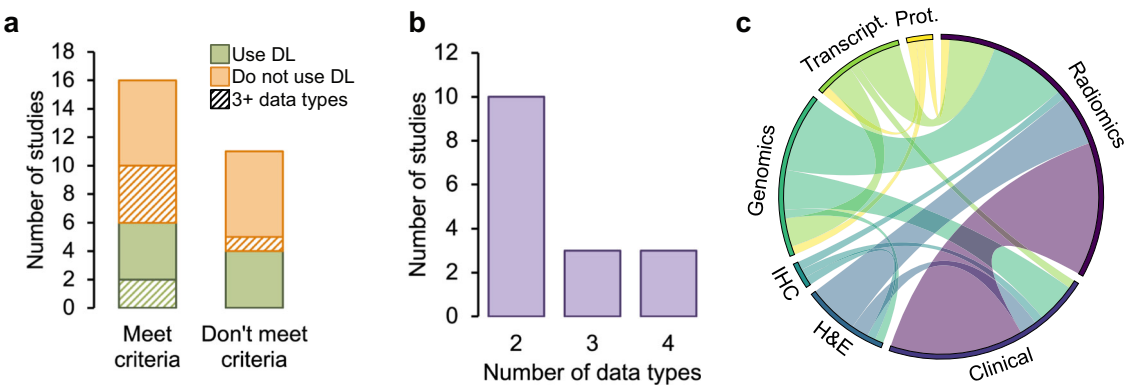

Fig. 2 | Review of data fusion. a Breakdown of the papers identified according to whether they satisfy the inclusion criteria; use deep learning or not; and integrate three or more data types or not. **b** Number of data types integrated in each of the studies that satisfied all the inclusion criteria. **c** Circos plot illustrating the co-occurrence of the different data types in the studies that satisfied all the inclusion criteria. Prot. = Proteomics, Transcript. = Transcriptomics.

combining a radiomics score (derived by a deep neural network) and clinical data to predict EGFR genotype in lung cancer. Five other studies use traditional ML methods for data integration with an intermediate fusion approach, including logistic regression[30], random forest[31–33], SVM[33,34] and LASSO[33].

Several studies take an indirect approach to data integration. Iwatate et al.[35] use a two-step approach for survival prediction: they predict molecular status (p53 and PD-L1) of pancreatic ductal adenocarcinoma from radiomics using an XGBoost algorithm, and then show a difference in survival between the predicted groups by conventional statistical analysis. Unsupervised ML is utilised in some studies, e.g. Hoivik et al.[36] perform unsupervised clustering of radiomics features extracted from MRI scans of endometrial cancer, and then show a statistical difference between the groups in several domains including survival, clinicopathological, transcriptomic and genomic features. Liang et al.[37] also use an indirect approach: they derive two clusters of patients with renal cell carcinoma based on their transcriptomics data, show a statistical difference in survival between the two clusters, and then build a machine learning model from radiomics features to predict these two clusters.

Only three studies use deep learning approaches for data integration, among which two use representation learning. Zhang et al.[38] use representation learning to extract deep radiomics features from MRI scans and integrate them with clinical data to predict microsatellite instability in rectal cancer. In contrast, Vanguri et al.[39] use radiomics features as input to a multiple-instance learning model to predict immunotherapy response in non-small cell lung cancer, integrating the radiomics with genomics and immunohistochemistry data. Interestingly, Li et al.[40] use an intermediate fusion approach, in which the algorithm characterises the multimodal interactions between histology and radiology features (both handcrafted and convolutional neural network (CNN) generated) by learning hierarchical co-attention mappings for the two modalities. This integrated embedding as well as the original radiomics data are then passed on together to the transformer layers for survival prediction in gastric cancer.

Deep learning was more commonly used as an isolated part of the image processing pipelines, including image segmentation and radiomics feature extraction. Image segmentation is potentially a major barrier to implementation of radiomics models in clinical practice. Manual delineation of all disease sites is necessary for radiomics feature extraction in traditional ML approaches, but it is time-consuming and requires radiologists' expertise. Although the majority of studies in this systematic review include a manual segmentation step[26–28,31–35,38,39], four papers successfully utilise deep learning models to automate the segmentation task[29,36,40,41]. Moreover, Hoivik et al.[36] compare both approaches, concluding that an automated segmentation model reproduces the same results as manual segmentation. The growing success of deep learning based segmentation algorithms for both tumours and healthy organs is lowering the entry barrier for the inclusion of imaging in data integration studies, as well as making these multimodal models easier to implement in clinical practice.

A further issue when considering traditional ML versus deep learning approaches is the method of radiomics feature extraction. Only deep learning allows for autonomous extraction of features from raw radiological data - otherwise the model depends on a set of handcrafted features. Among papers using traditional ML for data integration, all but one[29] use a handcrafted set of radiomics features, often extracted with open-source software such as pyradiomics. Only two papers use deep learning (CNN) for radiomics feature extraction[29,38], and one utilises both handcrafted and deep learning (CNN) features[40]. Deep learning is also used for pathology feature extraction in five studies (CNN[26–28,39], vision transformer[40]).

Using traditional ML versus deep learning also has implications on interpretability techniques which can be applied. Simple models such as logistic regression are usually very transparent and can provide a quantitative explanation of the model (e.g. feature coefficients). However, the implementation, or at least reporting, of such techniques is variable. In this review, most studies using traditional ML identify the list of radiomics features selected by the model, but further exploration and biological interpretation of the selected features are often limited. Some interpretability approaches include quantification of the importance of radiomics features for the model's predictive power[31], and analysis of a correlation heatmap between radiomics features and data from other modalities (including hallmark genes and tumour microenvironment)[34]. In contrast, deep learning approaches can be less transparent, however new interpretability approaches are being developed to shine light into the model's processes. For example, of the three papers using deep learning for the data integration step, two[38,40] include saliency or attention maps superimposed on the original radiological image to gain insight into the decision-making process. The development of such interpretability approaches for deep learning techniques is particularly important if these models, often more powerful than traditional approaches, are to be implemented in clinical practice.

Integration with computational pathology. Five studies in our literature review include integration with computational pathology, which is an evolving technique based on the analysis of digitised histopathology images, including haematoxylin and eosin (H&E) and immunohistochemistry (IHC) slides[42]. This number excludes studies in which pathology-based features are included as part of a list of clinicopathological variables[29,30,33,35,36,38], instead of using the pathology slides as input to a computational method.

Boehm et al.[27] use a weakly-supervised ResNet-18 CNN for tissue classification and feature extraction from whole slide images (WSIs). These histopathological features are then integrated with radiomic, genomic and clinical data by a Cox model in a late fusion approach.

Late fusion is also used by Wang et al.[26], who develop a pathomics signature from WSIs using a weakly-supervised patch-level CNN. This pathomics signature is then integrated by multivariate analysis with three other single-modality signatures (radiomics, IHC and clinical data) to develop a combined nomogram patient survival prediction in colorectal cancer.

Feng et al.[28] use two WSI analysis methods: an open-source software (CellProfiler) to extract nucleus features, and VGG-19 to extract microenvironment features. They develop two pathomics signatures (nucleus and microenvironment, respectively), and combine them with the radiomics signature by SVM in another late fusion approach to predict treatment response to neoadjuvant chemoradiotherapy in locally advanced rectal cancer.

Vanguri et al.[39] also use deep learning for tissue classification by means of a DenseNet classifier in combination with the commercially available HALO AI software for feature extraction. These histopathological features are then integrated with radiomics and genomics in an attention-based deep multiple-instance learning model to predict treatment response to immunotherapy in non-small cell lung cancer.

Li et al.[40] use a hierarchical vision transformer to extract the patch- and region-level features in WSIs. As described above, they then apply an intermediate fusion layer which learns hierarchical co-attention mapping between histology and radiology features and subsequently use multiple-instance learning to predict survival in gastric cancer.

Integration with genomics. Most studies that include genomics in their data fusion models do so by focusing on established biomarkers with known predictive power.

Targeted single-gene sequencing is commonly included, e.g. EGFR in Vanguri et al.'s[39] non-small cell lung cancer study or VHL in Zeng et al.'s[32] renal cell carcinoma study. Yi et al.[33] focus on only one gene (SULF1) and investigate its 12 single-nucleotide polymorphisms (SNPs); the SNPs are then included as genomic features in a model predicting platinum resistance in ovarian cancer.

A few studies take a different approach. Boehm et al.[27] determine the homologous recombination deficiency (HRD) status of patients by investigating COSMIC signatures as direct evidence of HRD, in addition to sequencing for known predisposing variants (e.g. in BRCA1 and BRCA2 genes). The HRD status is then combined with computational pathology and radiomics to predict survival in high-grade serous ovarian cancer.

Two papers incorporate a measure of mutation burden of the genome in their predictive models. Veeraraghavan et al.[34] use copy number burden as a measure of genomic instability and combine it with clinical and radiomics features to predict progression-free survival and platinum resistance in high grade serous ovarian carcinoma. Vanguri et al.[39] study both specific genes with known predictive power (such as EGFR) as well as the tumour mutation burden (TMB) to predict response to immunotherapy in non-small cell lung cancer. They show that the performance of a logistic regression model which uses genomics without TMB is inferior to the model trained using genomics with TMB. They also combine TMB and genomics data with radiomics and computational pathology to develop a combined deep learning model.

Data translation

Data fusion leverages the complementarity between data types to produce better overall predictions. However, its focus is not on dissecting the complex interactions between the data types, which could bring additional understanding about the disease. Data translation bridges the gap between different types of data, allowing for insights that might not be apparent when viewing each data type in isolation (Fig. 3).

In particular, data translation underpins the majority of approaches aiming to assign biological explanation to radiomics features. This explainability has been argued to be crucial for clinical application and acceptance of radiomics, with several classes of data suggested as biological correlates (e.g. genomic data, IHC, pathological images and habitat imaging)[43].

Predicting genotype from image-omics

The tumour phenotype, as it is expressed histologically and radiologically, is an indirect reflection of the underlying molecular landscape. In many cases, the genotype directly affects the tumour phenotype giving rise to distinct morphological patterns. This is particularly apparent in histopathology and has led to the development of many computational pathology deep learning algorithms which directly predict key mutations from routinely acquired tissue samples. These algorithms extract morphological features from H&E slides and have shown good accuracy in predicting various types of mutations. Kather et al.[44] use a weakly-supervised method with attention mechanism to predict microsatellite instability in gastrointestinal cancer patients from H&E images with good accuracy. The attention mechanism of the approach allows for the mapping of the network's attention scores showing the areas which are considered important by the model. Similar concepts are also exploited by Fremond et al.[45] where an attention-based deep learning algorithm sorts endometrial cancer patients into four molecular subgroups, and by Lu et al.[46], where a graphical network predicts HER2 mutation in breast cancer patients with high accuracy. Technologies such as spatial transcriptomics can provide full molecular expression analysis at the resolution of few-cells while retaining their spatial context. Recent studies showcase the feasibility of deep learning to predict the spatial transcriptomics profile from H&E samples with high accuracy[47–49].

In the area commonly known as 'radiogenomics', models trained from quantitative features (radiomics) extracted from standard-of-care radiological images are used to predict the underlying genotype, thus offering a non-invasive method to gain insights into the tumour's molecular profile. Researchers have shown promising results of radiomics-based models predicting driver mutations in many cancer types, including lung[29], breast[50], colorectal[51] and clear cell renal cell carcinoma[52]. However, the routine clinical adoption of radiomics approaches remains difficult mainly due to suboptimal robustness in multi-centric unseen data and lack of understanding of the tumour biology that guides the model's prediction. A link between imaging and the underlying histology could be the key in providing confidence and the necessary explainability to the genotypic predictions of radiogenomics-based models.

Predicting histology from imaging

In the UK imaging biomarker roadmap, O'Connor et al.[53] state that imaging biomarkers require stringent validation before they can be clinically deployed. However, histological validation is not a trivial task. Registering the three-dimensional (3D) MRI volume to the two-dimensional histology slides is difficult, due to low out-of-plane resolutions, histology tissue deformations, unknown orientations, and lack of details or landmarks in the radiology images. An elegant solution is to create 3D patient-specific moulds and use a microtome to cut the tumours and acquire slides co-aligned to the radiology slices. 3D moulds have been successfully applied in various cancer types, such as prostate[54,55], ovarian[56] and renal tumours[57] and breast cancer mouse models[58]. After imaging-histology co-alignment is achieved, registration is required to increase the accuracy of the matching between the two modalities.

At that stage, computational pathology can be used to spatially correlate medical imaging information with histopathological features, thereby providing explainability and confidence to the imaging measurements. Jardim-Perassi et al. use 3D printed tumour moulds to obtain co-registered MRI and histology in breast cancer mouse models. Subsequently, they use a commercial digital pathology software platform to process H&E and IHC samples and spatially histologically validate four tumour habitats identified directly from multi-parametric MRI: viable-normoxic, viable-hypoxic, nonviable-normoxic and nonviable-hypoxic[58]. Zormpas-Petridis et al.[59] use an automatic AI cell-classification algorithm on co-registered H&E samples to associate high T1 mapping MRI values with high densities of aggressive

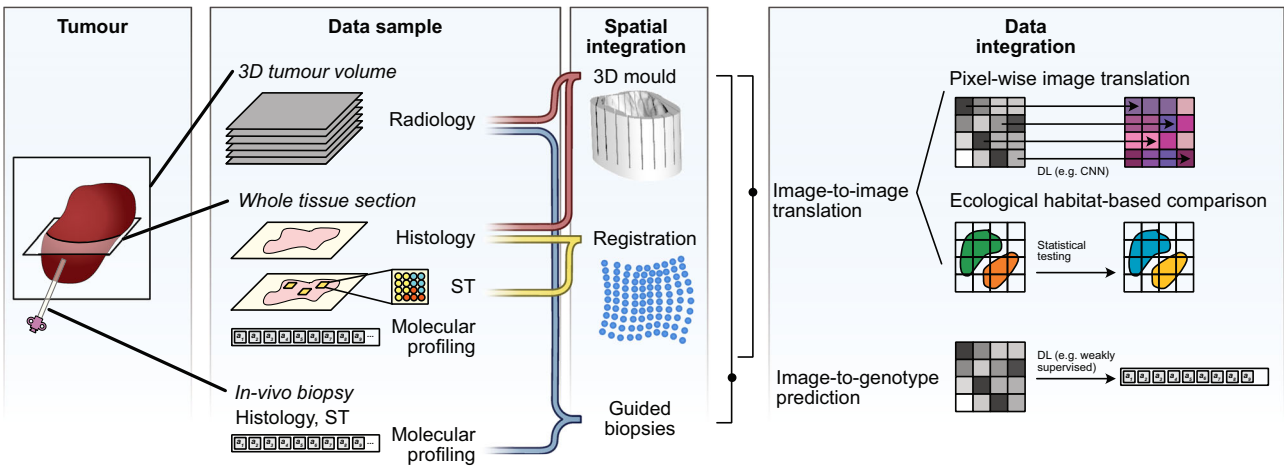

Fig. 3 | Schematic of data translation process. From left to right: types of data samples extracted from tumours, spatial integration techniques for sample co-registration, and data integration techniques for cross-modal translation, including image-to-image translation and image-to-genotype prediction. ST spatial transcriptomics.

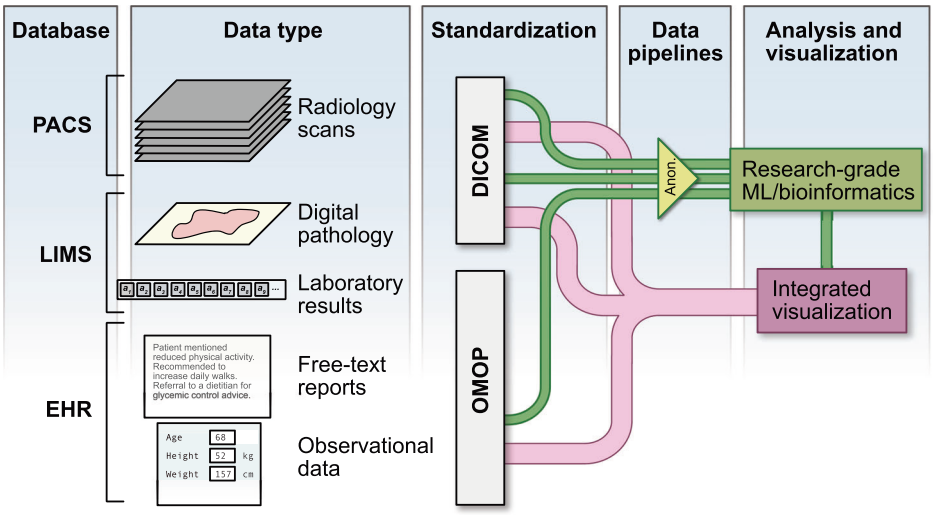

Fig. 4 | Data aggregation. Schematic of the data aggregation process to connect clinical information from Electronic Health Records (EHRs), imaging modalities from Picture Archiving and Communication Systems (PACS), and molecular markers from Laboratory Information Management Systems (LIMS), including a standardisation step for interoperability. Data can be processed offline after anonymisation (green pipelines) or displayed and visualised in real time (pink pipelines). Models trained on anonymised data can be deployed and integrated as part of live visualisation tools.

undifferentiated neuroblasts in a transgenic mouse model of high-risk neuroblastoma.

Despite these promising results, tumour mould development is a complex process, with more research needed before it can be included in standard clinical practice. On a technical level, ensuring the precision of radiological image segmentation, tumour positioning in the mould and histology-to-radiology slice co-registration is challenging, particularly in tumours without clear anatomical landmarks. Some of the proposed solutions include annotating the mould base position at the segmentation stage and cutting an additional 'orientation incision' at a pre-defined location relative to the lesion base[56], or, in the case of highly deformable organs such as the liver, employing a mould that completely surrounds the sample, which can help prevent changes in shape during slicing[60]. More recently, alternative approaches have been proposed that eliminate the need for moulds by letting a deep learning based co-registration model find the optimal alignment between tissue sections and 3D scan slices[61].

Virtual biopsies

Despite the crucial information about the histological, molecular and genetic properties of the tumour that is obtained from surgical biopsies, a biopsy sample represents only a "snapshot" of the tumour at a given time or location. Monitoring tumour evolution would require multiple repeated biopsies, which is often not feasible and is limited by multiple complications. Additionally, the highly heterogeneous nature of many tumour types offers no guarantees that the biopsy sample is representative of the entire tumour. This was clearly demonstrated in the multi-region TRACERx cohort, where the existence of "immune-cold" regions identified by computational pathology was independently prognostic for survival despite other regions of the tumour being "immune-hot"[62].

With advancements in imaging techniques and computational methodologies, the concept of "virtual biopsies" has gained traction[63]. By interconnecting imaging data at a spatial level with both molecular and histological information, it is possible to infer the internal characteristics of a tumour without physically extracting tissue samples. This approach is not only less invasive but can also provide longitudinal real-time insights into tumour heterogeneity and evolution. Although the length scales of radiology do not allow the imaging of single cells, they can characterise the regions ("habitats") in which they reside, potentially revealing the spatial variations in the density of different cell population[59,64] and inform on the underlying conditions, such as tumour metabolism and blood flow[65,66]. Virtual biopsies can also be used to guide radiotherapy treatment to target the most aggressive tumour areas and inform on more meaningful sites for surgical biopsies[67]. Identifying and monitoring tumour habitats can also lead to the early detection of emergence to treatment resistance[68].

The successful deployment of virtual biopsies requires a spatial streamline link between the radiological images, the histological samples and genomic/transcriptomic data. Beer et al.[67] incorporate a CT/ultrasound fusion to guide biopsy sampling from radiomics-defined tumour habitats. Additionally, Weigelt et al.[69] demonstrate in a patient with high-grade ovarian cancer that radiomics-defined image habitats present with distinct histological and molecular patterns.

Data aggregation

The third and final type of data integration, namely data aggregation, refers to the systematic process of collecting, preparing, and presenting diverse datasets for analysis (Fig. 4). Even though it is a step that is often left out of the data integration discussion, without it clinical models that rely on data integration, particularly if they contain images, could not be deployed. The objective is to unify various data streams, including clinical information from Electronic Health Records (EHRs), imaging modalities from Picture Archiving and Communication Systems (PACS), or molecular biomarkers from Laboratory Information Management Systems (LIMS).

Data frameworks

One of the foundational steps in data aggregation involves establishing structured frameworks that standardise data formats, ensuring uniformity and simplifying subsequent integration processes.

Of particular interest for healthcare-focused data integration is the adoption of data models, acting as a universal translator to facilitate cross-organisation data exchange and utility. For example, the Observational Medical Outcomes Partnership (OMOP) Common Data Model[70] seeks to standardise the structure and content of observational data to facilitate robust and reproducible downstream analyses, no matter which organisation was responsible for generating the data. For imaging, the DICOM format is widely used for radiological imaging[71], and is also gaining traction in digital pathology[72]. One of the strengths of DICOM format is the incorporation of metadata in the files, ensuring that this additional layer of information is directly linked to the pixel data. Metadata can be critical for accurate analysis and comparison across different datasets, for example to ensure that there is no systematic bias arising from differences in equipment settings or choice of machine manufacturer[73]. More research is being conducted on automated metadata interpretation, in order to more easily identify desired image characteristics, e.g. MRI sequences[74].

Offline analysis

Aggregation and transformation of multimodal data into coherent and linked sets of data ready for further use requires robust data processing pipelines. In the research setting, a common challenge is the de-identification of datasets and production of 'minimum viable datasets' to

ensure that participant anonymity is preserved. Some modalities, like free text, can be very challenging to de-identify, with results that are often inaccurate and unreliable[75]. Deep learning based natural language processing methods have also been proposed in recent years for de-identification purposes[76] and beyond. They are of particular interest in imaging given the duality of the data—images coupled with free text reports—and have showed promise in various tasks which could facilitate data aggregation. For example, large language models (LLMs) have been successful in transforming free text radiology reports into impression statements[77] and structured reports[78], both of which are more efficient data formats for further analysis. Furthermore, LLMs have also been used to extract labels from radiology reports, such as labels of oncological progression or diagnosis of metastatic disease[79], which shows their potential in automating the step of patient classification in data aggregation workflows. This step will likely also benefit from developments in a related method: vision-language processing, where the model includes both a text encoder and an image encoder; this allows for self-supervised learning from image-report pairs, with new research incorporating a temporal aspect to the process[80].

There are a number of unique challenges to de-identification of radiology data. Within the DICOM format, metadata need to be anonymised without affecting the diagnostic relevance of the images, which might be ambiguous, for example when dealing with scan dates, manufacturers' private tags or institution names. The reliability of de-identification tools has also been questioned in literature—in 2015, Aryanto et al.[81] reviewed 10 free toolkits and found that full anonymisation was achieved by only one of them with default settings, and four with manual adjustment of settings. However, a more recent review of available techniques is needed. A further challenge of radiological image anonymisation is posed by the process of embedding identifiable data into the pixel image as "burned-in" text, which might be difficult to detect and anonymise without removing the full image from the file; new techniques based on optical character recognition are being developed to tackle this issue[82].

Live deployment

Data integration models designed to be deployed in clinical practice rely on all the data streams being available simultaneously and in a timely fashion. However, historically EHR and PACS systems were developed independently, and most radiology departments do not have PACS-EHR integration[83] even though existing data suggest that it significantly improves productivity, even in the absence of novel AI-based support tools[84].

Discussion and outlook

The inclusion of medical imaging in multi-omic models for precision oncology brings challenges and opportunities.

Our review of the literature in the imaging-omics data fusion space showed that the field is still young and has not yet exploited the latest developments in AI research. Most studies integrate few data sources and use a late fusion approach in which most of the analysis is done for each data type independently. Many studies still rely on handcrafted features and the manual selection of interesting biomarkers. This is not as suboptimal as it may sound; a few well selected features guided by a domain expert could summarise information that would otherwise be very expensive for an unbiased deep learning framework to learn. Similarly, given the practical challenges of data aggregation for live deployment of data fusion models in the clinical setting, simplified models may be generally preferable. Nevertheless, our literature search has limitations. We limited ourselves to papers that studied response, prognosis, or outcome, and excluded other areas of data fusion such as diagnosis. Even though we did not actively exclude them, we did not specifically search for integration with clinical reports and medical records either—a field that is now receiving considerable attention with the recent success of large language models. In addition, we looked for specific terms to identify the different data modalities integrated in the study; as a result, we may have missed papers with a broader remit, such as those more interested in the method development than the application.

A unique feature of the integration of medical images is the introduction of a new physical dimension that provides context to the other data modalities. Data translation studies are making significant progress in this area, with the ultimate vision of generating a true multi-scale atlas, spanning from macroscopic radiology features down to the cellular microenvironment. Data fusion could also benefit from similar ideas of spatial co-localisation. For example, intra- and inter-tumour heterogeneity is a well known feature of many cancers, and yet data fusion models do not encode spatial alignment between biopsies and radiological location, which could lead to significant confounding effects.

Finally, deploying any of these models in a clinical scenario will generally require significant technological investment, as clinical systems are not necessarily designed for effective data aggregation. As the popularity of AI-assisted digital tools grows, this problem is likely to get progressively resolved.

Received: 19 January 2024; Accepted: 15 July 2024;
Published online: 26 July 2024

References

1. Chen, R. J. Pan-cancer integrative histology-genomic analysis via multimodal deep learning. *Cancer Cell* **40**, 865–878 (2022).
2. Mobadersany, P. Predicting cancer outcomes from histology and genomics using convolutional networks. *Proc. Natl. Acad. Sci.* **115**, E2970–E2979 (2018).
3. Sammut, S.-J. Multi-omic machine learning predictor of breast cancer therapy response. *Nature* **601**, 623–629 (2022).
4. Aloj, L. The emerging role of cell surface receptor and protein binding radiopharmaceuticals in cancer diagnostics and therapy. *Nucl. Med. Biol.* **92**, 53–64 (2021).
5. Sourbron, S. Technical aspects of MR perfusion. *Eur. J. Radiol.* **76**, 304–313 (2010).
6. Venkatesh, S. K., Yin, M. & Ehman, R. L. Magnetic resonance elastography of liver: technique, analysis, and clinical applications. *J. Magn. Reson. Imaging* **37**, 544–555 (2013).
7. Hagmann, P. Understanding diffusion MR imaging techniques: from scalar diffusion-weighted imaging to diffusion tensor imaging and beyond. *Radiographics* **26**, S205–S223 (2006).
8. Lipkova, J. Artificial intelligence for multimodal data integration in oncology. *Cancer Cell* **40**, 1095–1110 (2022).
9. Boehm, K. M., Khosravi, P., Vanguri, R., Gao, J. & Shah, S. P. Harnessing multimodal data integration to advance precision oncology. *Nat. Rev. Cancer* **22**, 114–126 (2022).
10. Barnum, G., Talukder, S. & Yue, Y. On the benefits of early fusion in multimodal representation learning. *arXiv preprint arXiv:2011.07191* (2020).
11. Gadzicki, K., Khamsehashari, R. & Zetzsche, C. Early vs late fusion in multimodal convolutional neural networks. In *2020 IEEE 23rd international conference on information fusion (FUSION)*, 1–6 (IEEE, 2020).
12. Hemker, K., Simidjievski, N. & Jamnik, M. HEALNet – Hybrid Multi-Modal Fusion for Heterogeneous Biomedical Data. *arXiv preprint arXiv:2311.09115* (2023).
13. Van Timmeren, J. E., Cester, D., Tanadini-Lang, S., Alkadhi, H. & Baessler, B. Radiomics in medical imaging—"how-to" guide and critical reflection. *Insights into Imaging* **11**, 1–16 (2020).
14. Mei, X. RadImageNet: an open radiologic deep learning research dataset for effective transfer learning. *Radiology: Artif. Intell.* **4**, e210315 (2022).
15. Sun, W., Zheng, B. & Qian, W. Automatic feature learning using multichannel ROI based on deep structured algorithms for computerized lung cancer diagnosis. *Computers Biol. Med.* **89**, 530–539 (2017).
16. Paul, R. Deep feature transfer learning in combination with traditional features predicts survival among patients with lung adenocarcinoma. *Tomography* **2**, 388–395 (2016).

17. Zeng, Q. Development and validation of a predictive model combining clinical, radiomics, and deep transfer learning features for lymph node metastasis in early gastric cancer. *Front. Med.* **9**, 986437 (2022).
18. Lao, J. A deep learning-based radiomics model for prediction of survival in glioblastoma multiforme. *Sci. Rep.* **7**, 10353 (2017).
19. Van Opbroek, A., Ikram, M. A., Vernooij, M. W. & De Bruijne, M. Transfer learning improves supervised image segmentation across imaging protocols. *IEEE Trans. Med. Imaging* **34**, 1018–1030 (2014).
20. Karimi, D., Warfield, S. K. & Gholipour, A. Transfer learning in medical image segmentation: New insights from analysis of the dynamics of model parameters and learned representations. *Artif. Intell. Med.* **116**, 102078 (2021).
21. Zoetmulder, R., Gavves, E., Caan, M. & Marquering, H. Domain-and task-specific transfer learning for medical segmentation tasks. *Computer Methods Prog. Biomedicine* **214**, 106539 (2022).
22. Tang, Y. et al. Self-supervised pre-training of swin transformers for 3d medical image analysis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 20730–20740 (2022).
23. Huang, S.-C. Self-supervised learning for medical image classification: a systematic review and implementation guidelines. *npj Digital Med.* **6**, 74 (2023).
24. Delgado, A. & Guddati, A. K. Clinical endpoints in oncology-a primer. *Am. J. Cancer Res.* **11**, 1121 (2021).
25. Jaffe, C. C. Measures of response: RECIST, WHO, and new alternatives. *J. Clin. Oncol.* **24**, 3245–3251 (2006).
26. Wang, R. Development of a novel combined nomogram model integrating deep learning-pathomics, radiomics and immunoscore to predict postoperative outcome of colorectal cancer lung metastasis patients. *J. Hematol. Oncol.* **15**, 11 (2022).
27. Boehm, K. M. Multimodal data integration using machine learning improves risk stratification of high-grade serous ovarian cancer. *Nat. Cancer* **3**, 723–733 (2022).
28. Feng, L. Development and validation of a radiopathomics model to predict pathological complete response to neoadjuvant chemoradiotherapy in locally advanced rectal cancer: a multicentre observational study. *Lancet Digital Health* **4**, e8–e17 (2022).
29. Wang, S. Mining whole-lung information by artificial intelligence for predicting EGFR genotype and targeted therapy response in lung cancer: a multicohort study. *Lancet Digital Health* **4**, e309–e319 (2022).
30. Cardoso, M. R. Metabolomics by nmr combined with machine learning to predict neoadjuvant chemotherapy response for breast cancer. *Cancers* **14**, 5055 (2022).
31. Zhou, H. Machine learning reveals multimodal mri patterns predictive of isocitrate dehydrogenase and 1p/19q status in diffuse low-and high-grade gliomas. *J. Neuro-Oncol.* **142**, 299–307 (2019).
32. Zeng, H. Integrative radiogenomics analysis for predicting molecular features and survival in clear cell renal cell carcinoma. *Aging (Albany NY)* **13**, 9960 (2021).
33. Yi, X. Incorporating SULF1 polymorphisms in a pretreatment CT-based radiomic model for predicting platinum resistance in ovarian cancer treatment. *Biomed. Pharmacother.* **133**, 111013 (2021).
34. Veeraraghavan, H. Integrated multi-tumor radio-genomic marker of outcomes in patients with high serous ovarian carcinoma. *Cancers* **12**, 3403 (2020).
35. Iwatate, Y. Radiogenomics for predicting p53 status, PD-L1 expression, and prognosis with machine learning in pancreatic cancer. *Br. J. Cancer* **123**, 1253–1261 (2020).
36. Hoivik, E. A. A radiogenomics application for prognostic profiling of endometrial cancer. *Commun. Biol.* **4**, 1363 (2021).
37. Liang, H. Renal enhanced CT images reveal the tandem mechanism between tumor cells and immunocytes based on bulk/single-cell RNA sequencing. *Funct. Integr. Genomics* **23**, 88 (2023).
38. Zhang, W. Development and validation of mri-based deep learning models for prediction of microsatellite instability in rectal cancer. *Cancer Med.* **10**, 4164–4173 (2021).
39. Vanguri, R. S. Multimodal integration of radiology, pathology and genomics for prediction of response to PD-(L)1 blockade in patients with non-small cell lung cancer. *Nat. Cancer* **3**, 1151–1164 (2022).
40. Li, Z., Jiang, Y., Lu, M., Li, R. & Xia, Y. Survival prediction via hierarchical multimodal co-attention transformer: A computational histology-radiology solution. *IEEE Transactions on Medical Imaging* (2023).
41. Cook, D. Next generation immuno-oncology tumor profiling using a rapid, non-invasive, computational biophysics biomarker in early-stage breast cancer. *Front. Artif. Intell.* **6**, 1153083 (2023).
42. Cui, M. & Zhang, D. Y. Artificial intelligence and computational pathology. *Lab. Investig.* **101**, 412–422 (2021).
43. Tomaszewski, M. R. & Gillies, R. J. The biological meaning of radiomic features. *Radiology* **298**, 505–516 (2021).
44. Kather, J. N. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nat. Med.* **25**, 1054–1056 (2019).
45. Fremond, S. Interpretable deep learning model to predict the molecular classification of endometrial cancer from haematoxylin and eosin-stained whole-slide images: a combined analysis of the PORTEC randomised trials and clinical cohorts. *Lancet Digital Health* **5**, e71–e82 (2023).
46. Lu, W. Slidegraph+: Whole slide image level graphs to predict HER2 status in breast cancer. *Med. Image Anal.* **80**, 102486 (2022).
47. Monjo, T., Koido, M., Nagasawa, S., Suzuki, Y. & Kamatani, Y. Efficient prediction of a spatial transcriptomics profile better characterizes breast cancer tissue sections without costly experimentation. *Sci. Rep.* **12**, 4133 (2022).
48. Levy-Jurgenson, A., Tekpli, X., Kristensen, V. N. & Yakhini, Z. Spatial transcriptomics inferred from pathology whole-slide images links tumor heterogeneity to survival in breast and lung cancer. *Sci. Rep.* **10**, 18802 (2020).
49. He, B. Integrating spatial gene expression and breast tumour morphology via deep learning. *Nat. Biomed. Eng.* **4**, 827–834 (2020).
50. Li, H. 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).
51. Yang, L. Can CT-based radiomics signature predict KRAS/NRAS/BRAF mutations in colorectal cancer? *Eur. Radiol.* **28**, 2058–2067 (2018).
52. Chen, X. Reliable gene mutation prediction in clear cell renal cell carcinoma through multi-classifier multi-objective radiogenomics model. *Phys. Med. Biol.* **63**, 215008 (2018).
53. O'Connor, J. P. Imaging biomarker roadmap for cancer studies. *Nat. Rev. Clin. Oncol.* **14**, 169–186 (2017).
54. Bailey, C. VERDICT MRI validation in fresh and fixed prostate specimens using patient-specific moulds for histological and MR alignment. *NMR Biomed.* **32**, e4073 (2019).
55. Wu, H. H. A system using patient-specific 3D-printed molds to spatially align in vivo MRI with ex vivo MRI and whole-mount histopathology for prostate cancer research. *J. Magn. Reson. Imaging* **49**, 270–279 (2019).
56. Delgado-Ortet, M. Lesion-specific 3D-printed moulds for image-guided tissue multi-sampling of ovarian tumours: A prospective pilot study. *Front. Oncol.* **13**, 1085874 (2023).
57. Crispin-Ortuzar, M. Three-dimensional printed molds for image-guided surgical biopsies: an open source computational platform. *JCO Clin. Cancer Inform.* **4**, 736–748 (2020).
58. Jardim-Perassi, B. V. Multiparametric MRI and coregistered histology identify tumor habitats in breast cancer mouse models. *Cancer Res.* **79**, 3952–3964 (2019).

59. Zormpas-Petridis, K. Noninvasive MRI native T1 mapping detects response to MYCN-targeted therapies in the Th-MYCN model of neuroblastoma. *Cancer Res.* **80**, 3424–3435 (2020).
60. Mikhail, A. S. Liver-specific 3D sectioning molds for correlating in vivo CT and MRI with tumor histopathology in woodchucks (marmota monax). *PLoS One* **15**, e0230794 (2020).
61. Shao, W. RAPHIA: A deep learning pipeline for the registration of MRI and whole-mount histopathology images of the prostate. *Comput. Biol. Med.* **173**, 108318 (2024).
62. AbdulJabbar, K. Geospatial immune variability illuminates differential evolution of lung adenocarcinoma. *Nat. Med.* **26**, 1054–1062 (2020).
63. Martin-Gonzalez, P. Integrative radiogenomics for virtual biopsy and treatment monitoring in ovarian cancer. *Insights Into Imaging* **11**, 1–10 (2020).
64. Jiménez-Sánchez, A. Heterogeneous tumor-immune microenvironments among differentially growing metastases in an ovarian cancer patient. *Cell* **170**, 927–938 (2017).
65. De Feyter, H. et al. Deuterium metabolic imaging (dmi) for mri-based 3d mapping of metabolism in vivo, *Sci. Adv.*, 4, eaat7314 (2018).
66. Wu, J. Intratumoral spatial heterogeneity at perfusion mr imaging predicts recurrence-free survival in locally advanced breast cancer treated with neoadjuvant chemotherapy. *Radiology* **288**, 26–35 (2018).
67. Beer, L. Ultrasound-guided targeted biopsies of CT-based radiomic tumour habitats: technical development and initial experience in metastatic ovarian cancer. *Eur. Radiol.* **31**, 3765–3772 (2021).
68. Gatenby, R. A., Grove, O. & Gillies, R. J. Quantitative imaging in cancer evolution and ecology. *Radiology* **269**, 8–14 (2013).
69. Weigelt, B. et al. Radiogenomics analysis of intratumor heterogeneity in a patient with high-grade serous ovarian cancer. *JCO Precision Oncology* **3**, 1–9 (2019).
70. Bathelt, F. The usage of OHDSI OMOP–a scoping review. *Proceedings of the German Medical Data Sciences (GMDS)* 95–95 (2021).
71. Mustra, M., Delac, K. & Grgic, M. Overview of the DICOM standard. In *2008 50th International Symposium ELMAR*, **1**, 39–44 (IEEE, 2008).
72. Herrmann, M. D. Implementing the DICOM standard for digital pathology. *J. Pathol. Inform.* **9**, 37 (2018).
73. Jha, A. Repeatability and reproducibility study of radiomic features on a phantom and human cohort. *Sci. Rep.* **11**, 2055 (2021).
74. Gauriau, R. Using DICOM metadata for radiological image series categorization: a feasibility study on large clinical brain MRI datasets. *J. digital imaging* **33**, 747–762 (2020).
75. Kushida, C. A. Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies. *Med. Care* **50**, S82 (2012).
76. Sousa, S. & Kern, R. How to keep text private? A systematic review of deep learning methods for privacy-preserving natural language processing. *Artif. Intell. Rev.* **56**, 1427–1492 (2023).
77. Sun, Z. Evaluating GPT-4 on impressions generation in radiology reports. *Radiology* **307**, e231259 (2023).
78. Adams, L. C. Leveraging GPT-4 for post hoc transformation of free-text radiology reports into structured reporting: a multilingual feasibility study. *Radiology* **307**, e230725 (2023).
79. Fink, M. A. Potential of ChatGPT and GPT-4 for data mining of free-text CT reports on lung cancer. *Radiology* **308**, e231362 (2023).
80. Bannur, S. et al. Learning to exploit temporal structure for biomedical vision-language processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15016–15027 (IEEE, 2023).
81. Aryanto, K., Oudkerk, M. & Van Ooijen, P. Free DICOM de-identification tools in clinical research: functioning and safety of patient privacy. *Eur. Radiol.* **25**, 3685–3695 (2015).
82. Osagie, E. & Ayo-Ogbor, S. Challenges in the design of optical character recognitionfor medical image modalities. *Educ. Res. (IJMCER)* **4**, 109–112 (2022).
83. Forsberg, D., Rosipko, B., Sunshine, J. L. & Ros, P. R. State of integration between PACS and other IT systems: a national survey of academic radiology departments. *J. Am. Coll. Radiol.* **13**, 812–818 (2016).
84. Mongan, J. & Avrin, D. Impact of PACS-EMR integration on radiologist usage of the EMR. *J. Digital Imaging* **31**, 611–614 (2018).

Acknowledgements

This work was supported by The Mark Foundation for Cancer Research [RG95043] and the Cancer Research UK Cambridge Centre [CTRQQR-2021/100012]. MCO received support from an NIHR i4i FAST grant [NIHR206092], Academy of Medical Sciences Springboard award [G117526], the EPSRC Tier-2 capital grant [EP/P020259/1], and GE HealthCare. This research was supported by the NIHR Cambridge Biomedical Research Centre (NIHR203312) and the Cambridge Experimental Cancer Medicine Centre (ECMC) [ECMCQQR-2022/100003]. The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care. SWB is supported by the Cancer Research UK Cambridge Centre [CTRQQR-2021/100012]. HP is supported by NIHR BioResource and AstraZeneca. HC is supported by the University of Cambridge Harding Distinguished Postgraduate Scholars Programme.

Author contributions

H.P.: Investigation, Writing—Original Draft, Writing—Review & Editing; K.Z.P.: Writing—Original Draft; H.C.: Writing—Review & Editing; S.B.: Writing—Original Draft; M.C.O.: Conceptualisation, Visualisation, Writing—Original Draft, Writing—Review & Editing, Supervision.

Competing interests

H.P. received research funding from AstraZeneca. M.C.O. received research funding from GE HealthCare and AstraZeneca. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41698-024-00656-0.

Correspondence and requests for materials should be addressed to Mireia Crispin-Ortuzar.