Trends in Cancer

CellPress

Review

Big data and artificial intelligence in cancer research

Xifeng Wu,[1,2,*] Wenyuan Li,[1,3] and Huakang Tu[1,4]

The field of oncology has witnessed an extraordinary surge in the application of big data and artificial intelligence (AI). AI development has made multiscale and multimodal data fusion and analysis possible. A new era of extracting information from complex big data is rapidly evolving. However, challenges related to efficient data curation, in-depth analysis, and utilization remain. We provide a comprehensive overview of the current state of the art in big data and computational analysis, highlighting key applications, challenges, and future opportunities in cancer research. By sketching the current landscape, we seek to foster a deeper understanding and facilitate the advancement of big data utilization in oncology, call for interdisciplinary collaborations, ultimately contributing to improved patient outcomes and a profound understanding of cancer.

Introduction to big data and AI

In 2020, approximately 19.3 million new cancer cases were reported globally, and nearly 10 million cancer-related deaths [1]. Over the past few decades, cancer prevention and targeted therapy have made progress in controlling and managing the disease. However, the heterogeneity and complexity of cancer types still pose enormous challenges. Cancer heterogeneity refers to the genetic, molecular, and phenotypic diversity within a single tumor or among different tumors of the same type, resulting in individual differences in environmental exposure reactions, susceptibility, treatment responses, and clinical outcomes. It is therefore imperative to adopt the precision medicine approach, which essentially pertains to the efficient collection and utilization of big data. Advances in laboratory technology, unique population-based studies, and clinical practices based on **electronic health records (EHRs)** (see Glossary) have accumulated an enormous number of various types of data. In the past, it would be difficult and even practically impossible to link these data and extract meaningful information. Now, with **artificial intelligence (AI)** becoming reality, big data and AI have shown superior advantages in our efforts to conquer cancer.

The alliance of big data and AI holds immense promise for revolutionizing our understanding of cancer, from its genesis to screening, diagnosis, treatment, response, toxicity, recurrence, and survival [2]. AI has been highly integrated into many aspects of cancer research such as standardizing large datasets and biobanks, clarifying the roles of modifiable risk factors, discovering new biomarkers or drug targets, creating prediction models and **knowledge graphs**, and establishing novel service platforms. These essential components pertain to the efficient collection and utilization of cancer big data. However, many challenges remain in areas including harmonization, missing data handling, and management (Table 1). This review aims to emphasize the transformative impact of big data and AI in oncology, outline the framework of collecting and utilizing big data in precision oncology, highlight current challenges and solutions, and review the application of these technologies, propelling advances in precision oncology.

Highlights

The integration of big data and artificial intelligence (AI) is transforming precision oncology from early diagnosis to personalized treatment, and innovative methodologies. We provide a comprehensive overview of advances in the application of big data and AI technologies in cancer research.

We discuss key challenges in data curation and utilization for cancer research, offering strategic solutions.

We detail the role and application of AI methodologies in processing cancer big data, with a special emphasis on multimodal data fusion analysis.

We introduce a framework for multiomics analysis, outlining its potential applications in identifying new biomarkers, understanding mechanisms, and developing therapies.

We propose a machine learning based intelligent service platform, designed to integrate cancer big data and employ AI algorithms for personalized health management.

[1]Department of Big Data in Health Science, School of Public Health, Center of Clinical Big Data and Analytics of The Second Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, Zhejiang, China
[2]National Institute for Data Science in Health and Medicine, Zhejiang University, Hangzhou, Zhejiang, China
[3]The Key Laboratory of Intelligent Preventive Medicine of Zhejiang Province, Hangzhou, Zhejiang, China
[4]Cancer Center, Zhejiang University, Hangzhou, Zhejiang, China

*Correspondence:
xifengw@zju.edu.cn (X. Wu).

https://doi.org/10.1016/j.trecan.2023.10.006

Table 1. List of challenges and potential solutions to the efficient curation and utilization of big data

Challenges	Explanations	Solutions
Data acquisition	**Volume and complexity:** the sheer volume and complexity of data, including genetic, clinical, and lifestyle information, can be overwhelming to process and analyze.	Implementing scalable computational infrastructure and using advanced algorithms.
	Data quality variation: variations in data ranges from meticulous quality control in some datasets to potential inaccuracies in others.	Instituting standardized data quality assessment procedures.
	Quality and standardization: inconsistencies in data quality and lack of standardization due to diverse variable definitions, measurements, and temporal fluctuations arising from changes in clinical guidelines and data recording practices can lead to unreliable results.	Establishing clear variable definitions and measurements, implementing dynamic algorithms to address temporal variability, and adopting universal standards for data collection.
	Integration: integrating various types of data from multiple sources, such as genomic, imaging, and EHRs, is a complex task.	Developing integrative platforms and employing interoperable data models.
Data management	**Privacy:** ensuring the privacy and security of sensitive patient data is a major concern.	Implementing robust encryption techniques and strict privacy policies.
	Collaboration: barriers in data sharing between institutions can hinder the overall progress in cancer research.	Establishing data-sharing consortiums and collaborative agreements.
	Ethics: navigating the complex regulatory landscape and ethical considerations surrounding the use of patient data can be difficult.	Clear guidelines, ethical oversight, and regular consultation with legal experts.
Model interpretation	**Understanding:** lack of metadata and advanced data analysis methods may lead to uninterpretable models or poor robustness.	Utilizing comprehensive metadata standards to enhance data context, combining AI with expert knowledge, and employing interpretable models.
	Application: AI interpretations need to seamlessly fit into the existing clinical workflow.	Developing user-friendly interfaces and providing real-time decision support.
	Algorithmic bias: algorithms adopt biases from training data may result in unjust or inaccurate outcomes.	Employing diverse training datasets, assessing regularly and adjusting algorithms for equitable results.

Big data curation and management

Data curation generally involves data acquisition, quality control, and validation to ensure that the data are accurate, complete, and reliable, compliant with legal and ethical requirements. Despite the absence of a uniform definition, cancer big data typically refers to the vast amounts of data derived from multiple sources.

Data sources and types

Sources of big data include epidemiology questionnaires, EHRs, imaging data, omics data, and mobile health device data. Epidemiological questionnaires typically include questions regarding demographics, medical history, lifestyle factors (dietary pattern, alcohol intake, smoking, physical activity, and sleep, etc.), environmental exposure, family history, medication use, disease outcome, psychological and cognitive function, reproductive information, and quality of life. EHRs constitute comprehensive repositories encompassing patient demographics, clinical history, medication records, laboratory outcomes, treatment plans, progress notes, billing data, and referrals, serving as fundamental references for clinicians. Imaging data are generated through various types of imaging modalities, such as magnetic resonance imaging (MRI), computed tomography (CT), and positron emission tomography (PET). These data provide rich visual

Glossary

Artificial intelligence (AI): computer systems that can perform tasks typically requiring human intelligence, such as understanding language, recognizing patterns, and making decisions.

Blockchain: a decentralized, secure digital technology that provides secure, transparent, and auditable data transactions and storage.

Convolutional neural network (CNN): a specialized deep learning architecture employing convolutional layers to automatically learn hierarchical features, enabling tasks like image recognition, classification, and object detection.

Deep learning (DL): a subset of artificial intelligence technologies using neural networks to process and learning from vast data for image recognition, and decision-making.

Distributed and federated learning: distributed learning uses various devices to collectively train a centralized model. Federated learning, a subset, enables devices to train locally and share updates without sharing raw data.

Electronic health records (EHRs): digitalized collections of individuals' health information, including patient demographics, clinical history, and treatment processes.

Framingham Heart Study (FHS): a project that investigates cardiovascular disease trends and risk factors over generations, contributing extensively to heart health knowledge.

Genome-wide association studies (GWASs): analyze genetic variations across a population to identify links between specific genetic markers and traits.

Knowledge graph: a structured representation of information, linking entities and their relationships.

National Institutes of Health (NIH): a US federal agency that funds and conducts biomedical and health research.

Natural language processing (NLP): a field of artificial intelligence that focuses on the interaction between computers and humans, involving tasks like language understanding, sentiment analysis, and language generation.

Virtual screening (VS): a technique using computer simulations to identify potential drug compounds for testing, accelerating the drug discovery process.

insights into different aspects of tumors, including their growth, spread, and response to treatments. Omics data encompass extensive datasets derived from diverse omics technologies, including genomics, transcriptomics, proteomics, microbiomics, and metabolomics. These datasets are derived from diverse biological samples and collectively provide a comprehensive perspective on the molecular constituents of cells, tissues, or organisms, enabling a holistic comprehension of the intricate biological processes underlying cancer. Mobile health device data come from wearable devices and mobile health applications, which offer real-time monitoring of patients' vital signs, activity levels, symptoms, and even treatment responses. They have the potential to enhance cancer therapy adherence, manage treatment-related symptoms, boost physical activity levels, and offer insights into behavior patterns. The landscape of data types has significantly expanded, encompassing datasets from chronic disease surveillance, cancer screening records, routine physical examinations, and medical insurance details. Their integration undeniably amplifies the precision of risk prediction models and propels the advancement of cancer research.

Data harmonization

Having outlined the diverse sources from which cancer big data can be collected, the subsequent challenges cannot be overlooked. Challenges stem from the heterogeneity of data sources, inconsistencies in formats, and variable data quality. These issues are compounded by semantic differences, temporal variability, and ethical constraints. To tackle these challenges, we propose a few key strategies including standardizing protocols, using advanced algorithms for missing data, ensuring secure and compliant data sharing, and implementing version control and cloud-based solutions.

The issue of missing data arises from incomplete records, inconsistent data entry, and gaps in longitudinal studies. These challenges are amplified by the high dimensionality of the data, the need for time-sensitive analyses, and ethical considerations. To mitigate these issues, potential strategies may include robust data validation checks, machine learning for imputing missing values, secure protocols for data handling, and real-time monitoring systems. EHRs can also be used for cross-verification.

Compared with traditional data, the advent of big data poses new challenges due to the rapid speed of data generation and updates, necessitating the development of pioneering storage systems. The main components of a data storage system may include various components such as data dictionaries, ID tracking and consent data, epidemiological data, biospecimen data, clinical data, biomarker modules, genetic modules, query tools, and reporting tools. Also, the volume of the database is increasing substantially. For example, the UK Biobank contains over 11 petabytes of data and is expected to exceed 40 petabytes by 2025. To effectively manage the substantial data volumes, distributed storage systems like the Hadoop Distributed File System (HDFS) can disperse data across multiple servers or nodes to ensure high availability and scalability. In addition, distributed structured query language (SQL) databases such as Google Spanner and NoSQL databases such as MongoDB are also utilized to manage structured and semistructured data respectively, offering additional layers of flexibility and efficiency. The application of data compression techniques and the optimization of storage structures can also be helpful in mitigating storage resource consumption.

The aggregation of extensive patient data amplifies concerns over data security and privacy (Table 1). These challenges are further complicated by the need for secure sharing and legal compliance. A multilayered approach including using **distributed and federated learning** for local data training [3,4], integrating **blockchain** for secure transactions [5], establishing robust governance for data access and compliance [6], and implementing real-time monitoring with regular security audits may be helpful.

Multimodal data analysis

Medical imaging

The convergence of medical big data with AI is revolutionizing radiomics and digital pathology. **Deep learning (DL)** algorithms excel in image analysis and pattern recognition, often surpassing human performance. Radiomics uses advanced mathematical algorithms such as a gray level co-occurrence matrix, histogram-based features, and support vector machines for quantitative analysis of high-dimensional features in MRI, CT, and PET scans. It adopts a comprehensive workflow that includes image acquisition, preprocessing, tumor segmentation, feature extraction, and ultimately, model validation. These algorithms enable the identification and quantification of various textural, shape-based, and intensity-based features within the images, providing a comprehensive understanding of tumor heterogeneity, severity, and other clinically relevant feature. Digital pathology, the gold standard for tumor diagnosis, is also evolving due to AI. Unlike traditional pathologic methods, AI-enhanced digital pathology mitigates human biases by enabling digital capture and comprehensive analysis of specimens at both the cell and regional levels, refining the diagnostic process but also alleviating the workload of the pathologist. The fusion of AI with radiomics and digital pathology is thus creating a synergistic effect that holds considerable promise for the advancement of diagnostic accuracy and efficiency in oncology.

Fusion analysis

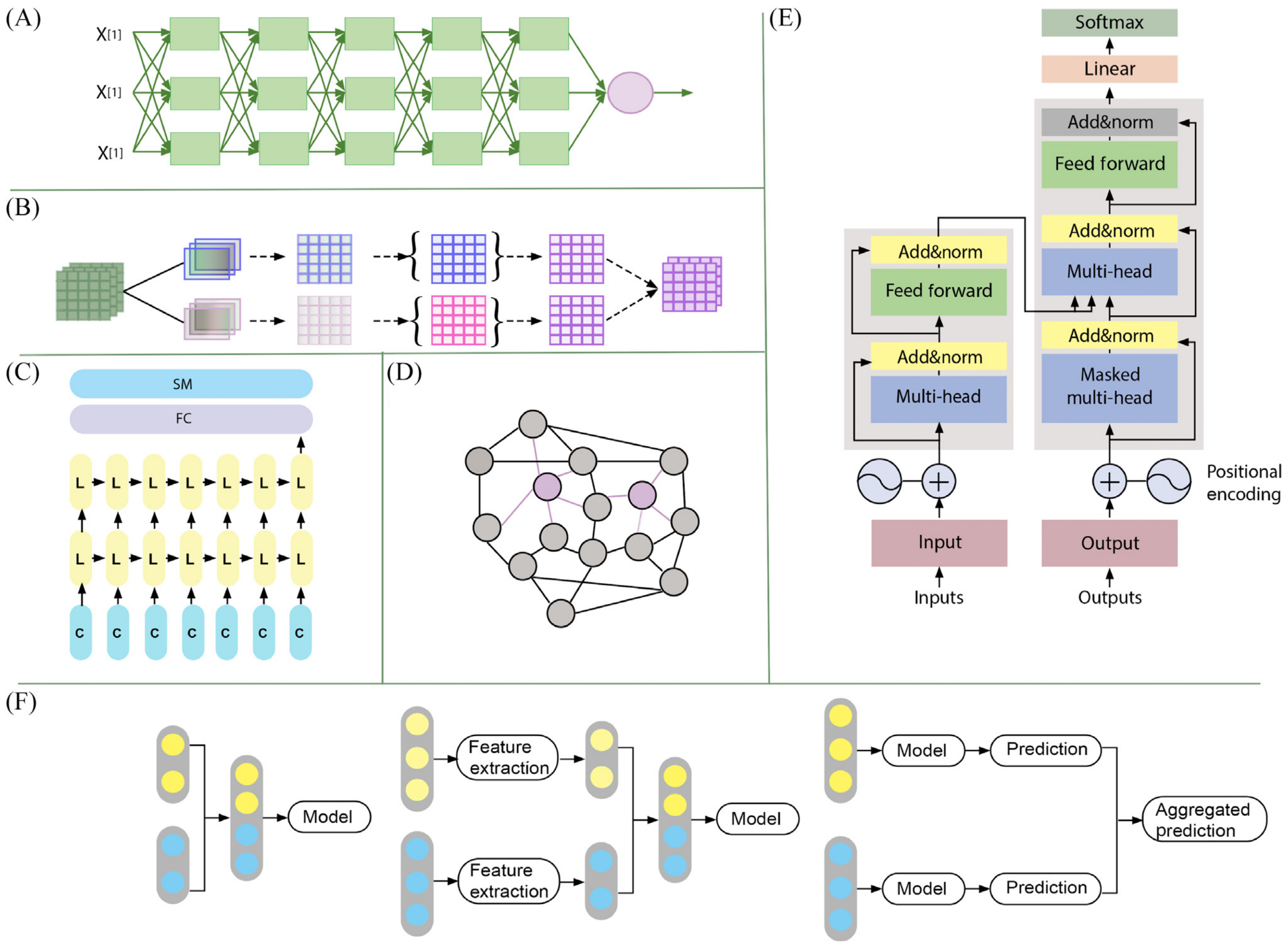
The multiscale, multimodal, and high-dimensional data can be harnessed through fusion analysis [7,8]. Commonly used algorithms are illustrated in Figure 1. The IRENE model [9] uses embedding layers to convert images, unstructured text, and structured clinical data into visual and text tokens, then uses bidirectional blocks with both intramodal and intermodal attention to learn holistic representations, outperforming traditional and image-only models in identifying pulmonary disease and predicting outcomes. By using various fusion strategies, DL-fused histopathology images with gene expression profile models outperformed single-data models and identified more relevant biological pathways in glioma patients [10].

Knowledge graph

A knowledge graph integrates interconnected data from multiple sources to provide a comprehensive view of entities like genes, proteins, and patient outcomes, offering a navigable snapshot of individual health status. REMAP [11], a multimodal machine learning approach for extracting disease relations from both structured knowledge graphs and unstructured text, improved accuracy and F1 score by aligning multimodal data sources, and outperforms graph-based methods in recommending new disease relationships. Another work applied multimodal reasoning by reverse-hyperplane projection based on structure, category and description embeddings, and demonstrated the versatility of embedding models in classifying biomolecular interactions [12]. A recommendation system integrating preclinical, clinical, and literature data was built on a heterogeneous biomedical knowledge graph to addresses the challenge of resistance to epidermal growth factor receptor (EGFR) inhibitors in non-small cell lung cancer [13]. The system successfully narrows down potential resistance markers from >3000 genes to 57.

Multiomics analysis

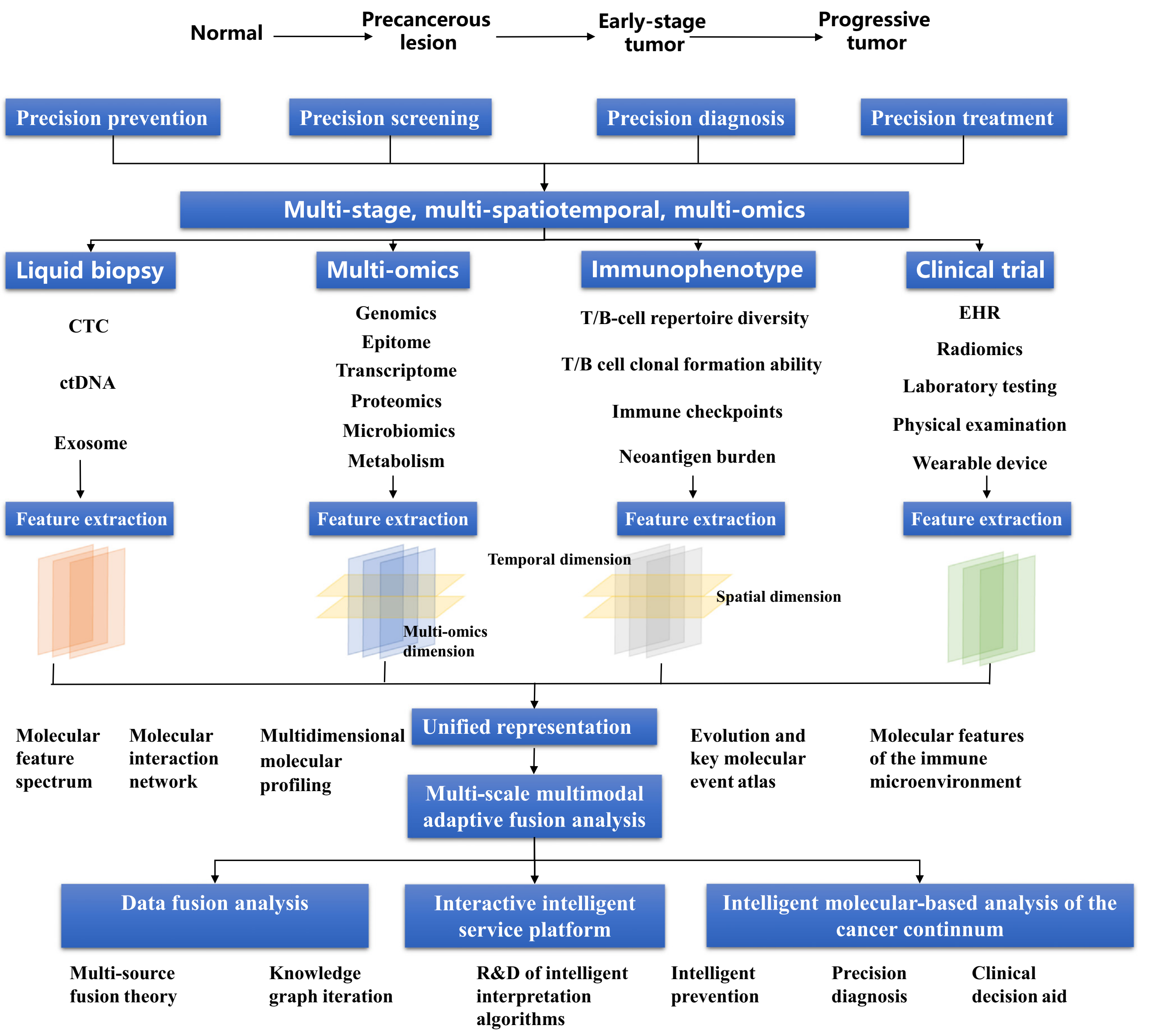
Data sets of different omics groups such as genomics, transcriptomics, epigenomics, proteomics, microbiomics, and metabolomics, can be combined during analysis. The heterogeneity of data types and high dimensionality require substantial computational resources and



Trends in Cancer

Figure 1. Common machine-learning models and fusion strategy. (A) Multilayer perceptron is a deep learning model that utilizes multiple fully connected layers to obtain feature vectors, and it finds application in cancer research for data mining and classification tasks. (B) Convolutional neural networks operate by using cascading layers of convolution and pooling to progressively extract features from medical images, thereby facilitating enhanced diagnostic and prognostic capabilities in the context of cancer risk assessment for patients. (C) Short-term memory is a recurrent neural network designed to capture sequential dependencies in data and is commonly used for time-series analysis of patient data to help predict disease progression and treatment outcomes. (D) Graph neural networks operate on graph-structured data, allowing them to model relationships and interactions; they can integrate information about different treatments, including efficacy, risks and individual patient characteristics, to assist physicians and patients in making more informed treatment choices. (E) Transformer architecture features a self-attention mechanism that facilitates the capture of contextual information from data and can be used in cancer research to analyze genome sequences and gene expression patterns, thereby improving our understanding of cancer biology and potential therapeutic targets. (F) According to the communication and aggregation mechanism between different modalities, feature fusion can be principally categorized into early fusion, middle fusion, and late fusion. Abbreviations: C, sequence's element; FC, fully connected layer; L, network's layer; SM, SoftMax activation function.

specialized algorithms. We have illustrated a framework for multiomics analysis in Figure 2. Histopathology images have been leveraged to predict multi-omics aberrations and prognoses in cancer patients [14]. Utilizing weakly supervised DL models, integrative multiomics–histopathological analysis for breast cancer classification explores the link between histopathological images and genetic statuses [15]. Employing the Multi-omics Multicohort Assessment platform, a study identified interpretable pathology patterns predictive of gene expression profiles, microsatellite instability status, and clinically actionable genetic alterations [16]. Shifting to transcriptomics, by combining a CRISPR interference (CRISPRi) screen with orthogonal multiomics approaches, the long noncoding RNA, DARS1-AS1, was shown to play a pivotal role in glioblastoma [17].



Trends in Cancer

Figure 2. Framework for multiomics analysis and application. Abbreviations: CTC, circulating tumor cells; ctDNA, circulating tumor DNA; EHR, electronic health record.

Through machine-learning algorithms, specific patterns or biomarkers within the microbiome that are associated with different types of cancer can be discovered [18]. Integrating single-nucleus RNA sequencing and spatial transcriptomics has unveiled the complex cellular architecture of breast cancer tissues and potential therapeutic strategies [19]. Single-cell multiomics analysis generates a comprehensive transcriptional and epigenomic landscape, revealing key transcription factors mediating tumor cell-specific regulatory programs [20]. In a comprehensive approach spanning multiple omics fields, including metabolomics, transcriptomics, proteomics, epigenomics, and genomics, circulating cell-free DNA genomic signatures were integrated, enhancing early-stage lung cancer diagnosis and the detection of minimal residual disease [21].

EHR analysis

Natural language processing (NLP) assists in the extraction and interpretation of unstructured textual data from EHR, medical literature, and clinical notes. PheCAP, a semisupervised system, uses NLP to extract valuable information from EHRs, speeding up phenotyping and enhancing healthcare decision-making [22]. The Multiview Incomplete Knowledge Graph Integration (MIKGI) algorithm combines embeddings from medical code co-occurrence patterns and semantic embeddings from textual strings and synthesizes these into harmonized semantic vectors, thereby achieving high accuracy in tasks like detecting similar or related entity pairs and mapping medical codes across institutions [23]. Federated learning has emerged as a key solution for maintaining data privacy in collaborative model development, allowing institutions to train local models without centralizing patient-level data [24]. This approach not only ensures data security but also improves collective model performance, facilitating cross-institutional research. Advanced techniques like sparse embedding regression efficiently select relevant features from EHRs, offering performance comparable with manually curated features [25].

Integrated big data platform

Cohort, consortium, and omics databases are among the best approaches when integrating big data in cancer (see supplemental information online). Commonly used computer programs to support big data analysis can be found in Table 2.

Table 2. List of commonly used computer programs for big data analysis[a]

Package	Accessibility	Capacity	Advantages	Disadvantages
SimpleITK	Open-source; Python and C++	Medical image processing	Comprehensive APIs for different medical image formats (DICOM, NiFti etc.)	Limited to analytical functions
Nibabel	Open-source; Python	Neuroimaging data processing	Basic operations on common neuroimaging file formats; Pythonic interface	Focus on data structure transformation for neuroimaging and Python proficiency
OpenCV	Open-source; multiple languages	Computer vision and image processing	Real-time optimized Computer Vision library; cross-platform	Not specialized for medical data; multiple programming languages based, requires user to select and combine algorithms adapted to medical imaging
MONAI	Open-source; Python	Medical image analysis	DL-based medical image processing library; PyTorch-based	Python-based, may require DL foundation and Python proficiency
scikit-learn	Open-source; Python	Machine learning (ML)	General-purpose ML library; extensive community support	Not specialized for medical data; limited to traditional ML
Bioconductor	Open-source; R	Bioinformatics and genomics	Rich set of packages for genomics; R community integration	R-based, may require knowledge of genomics; not for all data
mixOmics	Open-source; R	Multi-omics data integration	Integrates multiomics data effectively; statistical methods	R-based, may have a learning curve for beginners
TCGA-Assembler	Open-source; R	TCGA data integration	Simplifies TCGA data integration; R community integration	Focused on TCGA data; limited to cancer genomics
ImageJ	Open-source; Java	Image analysis and processing	Wide range of plugins; extensive user community	May require Java proficiency; GUI-based
DeepPathology	Not specified; likely open-source	Pathology image analysis	Specialized for pathology image analysis; DL	Accessibility and features may vary; relatively new
PathAI	Commercial	Pathology image analysis	AI-assisted pathology diagnostics; commercial support	Paid subscription required; proprietary
HistomicsTK	Open-source; Python and Django	Digital pathology image analysis	Extensive toolkit for digital pathology; web-based interface	Setup and deployment complexity; learning curve for web-based

[a] Abbreviations: API, application programming interface; GUI, graphical user interface; TCGA, The Cancer Genome Atlas.

Large-scale cohort studies are viewed as the best approaches for obtaining high-standard, high-quality, cross-scale, multimodal big data and biological samples. These studies collect not only baseline data like questionnaires, biomarkers, clinical and phenotypic data but also conduct long-term follow-up. The **Framingham Heart Study (FHS)**, launched under the direction of the US **National Institutes of Health (NIH)** in 1948 [26], enrolled >15 000 people of varying ages and backgrounds, and published >3698 research articles by 2018. The All of Us research program, initiated by the NIH in 2018, aims to build a large-scale cohort of at least 1 million participants [27], and collects genome data, including whole genome sequencing and genotyping. It also collects data on lifestyle factors and EHRs, such as physical activity, nutrition, heart rate, and sleep. The UK Biobank, established in 2006, is a large biomedical database, involving over half a million UK participants aged 40–69 years [28] that contains genetic information, blood samples, imaging data, lifestyle and environmental exposure data, and tracks health records that have been regularly updated overtime.

Countries worldwide are increasingly investing in constructing cohorts to identify modifiable risk factors and novel biomarkers of cancer, formulate individualized strategies for cancer screening, diagnosis, treatment, and management, and build intelligent service platforms. However, many cohorts enrolled cancer patients or high-risk individuals only, and had relatively small sample sizes. The establishment of consortia provided a solution. These consortia facilitate the harmonization and integration of collected omics data with clinical phenotypic data and other data types.

The formation of several large databases also provided support for precision medicine. The Cancer Genome Atlas is a landmark collaborative project that has played a pivotal role in advancing our understanding of cancer on a molecular level [29]. It was launched in 2005 by the US National Cancer Institute and the National Human Genome Research Institute in order to comprehensively catalog and analyze the genomic alterations that drive various types of cancer.

Successful use of big data and AI application in cancer research

Discovery of modifiable risk factors

Cancer development is intricately tied to a spectrum of modifiable risk factors; aggregating and analyzing diverse datasets provides the statistical power and robustness necessary for unraveling complex interactions between modifiable risk factors. Studies have consistently revealed positive links between traffic-related air pollution and elevated lung cancer risk in diverse populations [30]. With machine-learning algorithms, researchers were able to construct robust aging biomarkers and explore their contribution to cancer susceptibility. In the UK Biobank, associations of discretionary screen time, Mediterranean lifestyle, physical activity, a composite healthy lifestyle score, and other factors with susceptibility to cancer have been highlighted [31–34]. The pivotal role of nutrition is elucidated through studies examining dietary habits, particularly the consumption of ultraprocessed foods, red meat, and processed meat [35]¸ and iron intake [36,37]. Conversely, research focusing on exercise and cancer risk [38], utilizing large prospective cohorts, demonstrates the potential benefits of resistance training in mitigating cancer susceptibility, notably for bladder and kidney cancers [39]. These findings, derived from extensive cohort studies, illuminated the considerable influence of modifiable risk factors on cancer.

Discovery of biomarkers

Biomarkers of susceptibility

Identifying individuals at higher risk for certain diseases based on their genetic profiles, enables medical practitioners to implement personalized preventive measures at an early stage, reducing the overall disease burden. **Genome-wide association studies (GWASs)** provide a new way to identify genetic risk factors associated with tumors [40]. Over the past two decades, GWASs have identified approximately 40 sites associated with lung cancer susceptibility [41–43], >160 common loci associated with prostate cancer susceptibility [44], and 48 sites associated with breast cancer susceptibility [45]. Establishing polygenic models and helping to calculate polygenic risk scores in cancer can improve the prediction of genetic diseases [46]. A risk prediction model was developed using data from 16 633 prostate cancer families [47]. This model offers personalized validated predictions of prostate cancer risk by considering known intermediate- and high-risk pathogenic variants, low-risk common genetic variants, and a well-defined family history of cancer.

Biomarkers for diagnosis and prognosis

Diagnostic and prognostic biomarkers can be molecular, histological, radiographic, or physiological characteristics that indicate the presence of cancer. Molecular biomarkers have become crucial in the prevention and diagnosis of cancer, as they enhance our understanding of its causes and improve the accuracy of diagnosis and prognosis. RNA sequencing and methylation have contributed to the identification of new biomarkers for various types of cancer, such as esophageal [48], colorectal [49], gastric [50], and pancreatic [51] cancer. The advancement of imaging technology has also played a significant role in the discovery of these biomarkers [52]. Specific gut microbiome signatures are identified to predict lung cancer and colorectal cancer, assisting doctors in detecting cancer at an earlier stage and thereby improving treatment success rates [53,54]. By combining the interpretations of radiologists, pathological factors, imaging metrics, and machine learning techniques, higher diagnostic accuracy were achieved, which greatly benefits patient management [55]. Some studies have established connections between the characteristics of medical images and molecular phenotypes, giving rise to a new field known as radiogenomics [56].

Drug discovery and repurposing

AI is overcoming limitations of traditional techniques such as **virtual screening (VS)** and molecular docking, specifically in improving drug–target interaction, structure-based VS, and toxicity characterization [57,58], enhanced drug design and mass-production capabilities. Computational pipelines can predict new drug interactions within heterogeneous networks [59]. Additionally, deep generative models have shown promise in designing molecules that inhibit specific receptors with favorable pharmacokinetics [60]. AI has also been instrumental in streamlining drug–target interaction prediction, expanding opportunities in drug reuse and combination therapies [61]. One study used a systems biology approach using genome-wide microarray data and machine learning models to identify potential molecular drugs for diseases [62]. Deep neural network models, along with experimental approaches, have identified new drug combinations for diseases like leukemia, increasing the therapeutic options [63].

Biomarkers for therapeutic response and adverse events

A predictive biomarker is a tool used to predict the outcome of a specific therapeutic intervention, including both therapeutic benefits and possible side effects of chemotherapy, radiotherapy, and immunotherapy. Although immunotherapy has proven to be effective against many cancer types, the presence of primary or secondary resistance still leaves most immunotherapy-eligible patients without significant benefits. Therefore, the tumor microenvironment needs to be assessed with appropriate biomarkers to determine the best therapy to use in a specific patient population and predict resistance. Analysis of tumor tissue samples [64] (such as tumor mutation burden and tumor immune microenvironment), gene expression [65], gut microbiome features [66], and noninvasive plasma-derived biomarkers such as α-fetoprotein (AFP) can provide information on tumor biology in order to assess the response of cancer patients to immunotherapy [67]. However, overactivation of the immune system caused by immunotherapy often leads to a range of toxicities, namely immune-related adverse events (irAEs). It is therefore critical to investigate appropriate biomarkers to timely detect and manage irAEs. Most studies to date have used many biological specimens for biomarker discovery, such as peripheral blood (serum, plasma, or whole blood) and stool samples [68]. Gut microbes in stool samples were also found to be associated with irAEs [66].

Drug dose adjustment

By integrating patient-specific factors such as age, weight, genetics, and kidney/liver function, describing how drugs are absorbed, distributed, metabolized, and eliminated in different patient groups, pharmacokinetic models can be developed to guide the calculation of optimal drug or radiation doses tailored to each patient. In a case involving metastatic castration-resistant prostate cancer, the AI-guided dosing was both effective and well tolerated, significantly reducing prostate-specific antigen concentrations [69]. In radiation therapy, AI can consider variances in tumor biology and the geometric relationships between tumors and nearby organs, predict tumor radiation sensitivity, and formulate optimal dose prescriptions, tailored to the unique aspects of the tumor and surrounding organs [70,71].

Medical imaging

The application of AI through radiomics image analysis has seen outstanding advances. Some machine-learning models have been crafted for gland segmentation and tumor classification, demonstrating remarkable detection and grading accuracies [72]. Computer-aided detection systems were utilized in a study to detect flat polyps on CT images, with a high success rate [73]. Studies have predicted risks and constructed radiology scores for prognosis in various cancers [74,75]. There have also been explorations into the relationship between radiological features and tumor transcriptomics [76], and platforms that integrate multi-omics data to aid in decision-making in patients with lung cancer [77].

Risk prediction modeling

Population risk stratification

DL models are being increasingly used for risk prediction to provide more accurate risk scores for cancers, resulting in a shift towards more personalized and precise cancer risk stratification [78–80]. MeScore, a machine-learning-based prediction model developed by a binational study between Israel and the UK, have achieved promising results for detecting high-risk patients [81]. Machine learning methodologies are being extended to predict cancer risk from different modalities like chest X-rays and MRIs, and some models are designed to provide detailed visual insights, such as heatmaps, to indicate where cancer is most likely to develop.

Models for response

Precise response prediction is of great clinical importance for providing evidence for clinical decision on choosing the appropriate treatment, and precluding the need for surgery. A recent study applied DL models to pairs of ultrasonography images to predict response to neoadjuvant chemotherapy (NAC) in breast cancer [82]. Another study innovatively built a DenseNet model to assess programmed death-ligand 1 (PD-L1) expression in non-small cell lung cancer [83], enabling noninvasive prediction of response to immunotherapy.

Models for recurrence

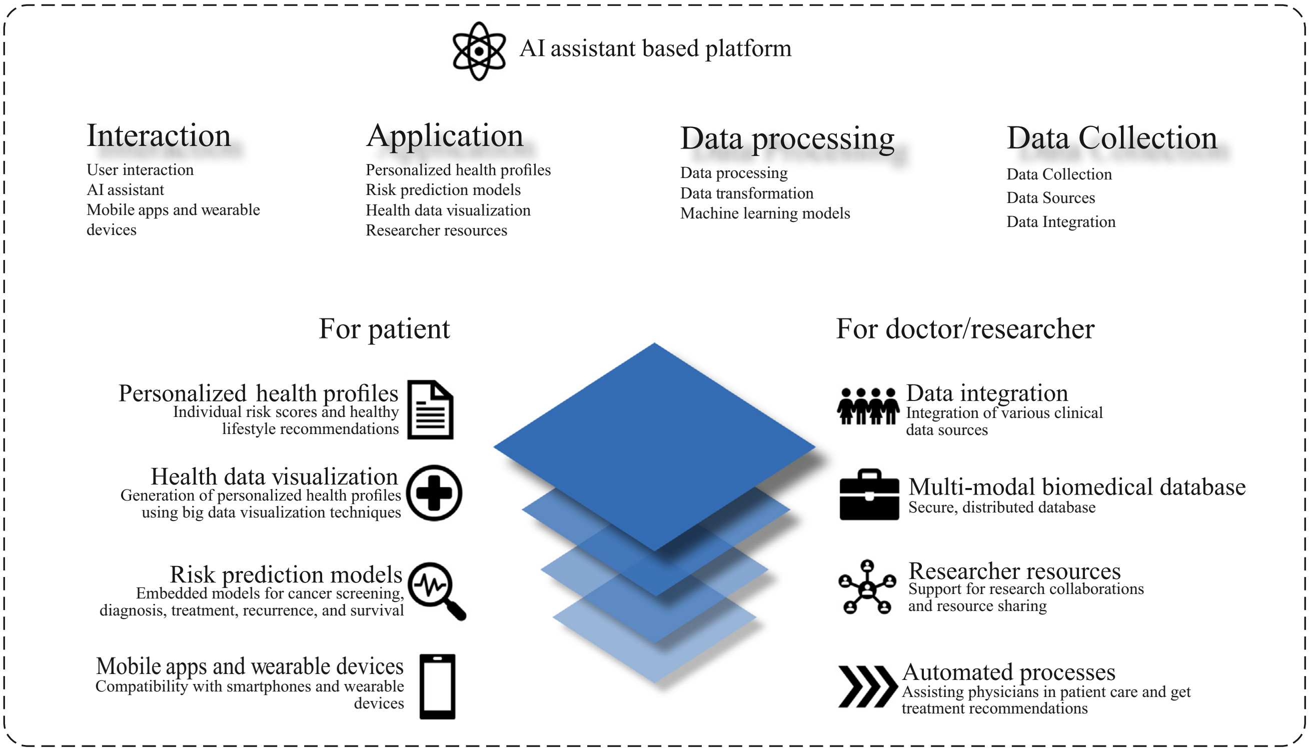
Estimating recurrence is central to cancer staging and treatment planning. Current models utilize various clinical parameters such as age, gender, cancer stage, genetic alterations, circulating molecular markers, and a multitude of histology risk factors [84,85]. However, higher-level features also carry prognostic information, like the spatial arrangement of lymphocytes and chromatin texture. A previous study [86] developed a nomogram for predicting recurrence after nonmetastatic colorectal cancer surgery, and **convolutional neural network (CNN)** models using PET/CT data were applied to predict local tumor recurrence, demonstrating better predictive ability compared with traditional models [87].

Models for survival

Survival predictive models have become essential tools in cancer prognosis, aiding clinicians in evaluating the prognosis and tailoring individualized interventions [84,85,88]. AI presents a promising alternative, potentially harnessing this data more effectively for estimating patient viability and survival time. Recent studies have showcased the ability of CNN to automate the extraction of prognostic factors. A CNN was trained on >100 000 hand-delineated image patches from 86 colorectal cancer tissue slides, achieving a nine-class accuracy of >94% on an independent dataset, and generated a deep stroma score that served as an independent prognostic factor for overall survival [88].

Intelligent service platform

There is a demand to develop an intelligent service platform specifically tailored to clinical scenarios and patient needs, integrating various forms of data. The platform (Figure 3) will feature embedded risk prediction models for cancer screening, diagnosis, treatment, recurrence, and survival, and will generate personalized health profiles using big data visualization techniques. It will provide individuals with risk scores, healthy lifestyle recommendations, and real-time updated screening and treatment plans derived from reinforcement learning algorithms. For health promotion, individuals will receive recommendations regarding modifiable risk factors including


AI assistant based platform
Interaction
User interaction
AI assistant
Mobile apps and wearable devices
Application
Personalized health profiles
Risk prediction models
Health data visualization
Researcher resources
Data processing
Data processing
Data transformation
Machine learning models
Data Collection
Data Collection
Data Sources
Data Integration
For patient
Personalized health profiles
Individual risk scores and healthy lifestyle recommendations
Health data visualization
Generation of personalized health profiles using big data visualization techniques
Risk prediction models
Embedded models for cancer screening, diagnosis, treatment, recurrence, and survival
Mobile apps and wearable devices
Compatibility with smartphones and wearable devices
For doctor/researcher
Data integration
Integration of various clinical data sources
Multi-modal biomedical database
Secure, distributed database
Researcher resources
Support for research collaborations and resource sharing
Automated processes
Assisting physicians in patient care and get treatment recommendations

Trends in Cancer

Figure 3. Artificial intelligence (AI) assistant-based platform.

smoking, drinking, dietary intake, and sleep habits. An AI assistant, backed by advanced large language models, will be integrated into the platform. To facilitate this, a secure, distributed multimodal biomedical database is essential. The platform will also include a section for researchers, providing resources and guidelines to encourage future collaborations. Additionally, the service platform can be compatible with smartphone apps and wearable devices. It uses automated processes to aid physicians in patient care and boosts patients' self-management capabilities, aligning with health management and cost control goals. The platform is also scalable, with the capacity to extend its services to support clinical decision-making systems and manage other diseases, thereby setting a robust foundation for future health management efforts.

Concluding remarks

The integration of big data and AI in cancer research offers unprecedented discovery and application in precision oncology practices. However, this transformation is not without its hurdles (see Outstanding questions). These challenges demand robust solutions, which can be achieved through interdisciplinary collaborations among researchers, clinicians, data scientists, and policymakers. By maintaining a focus on innovation considerations, there is promise for more precise, effective, and individualized cancer treatment, ultimately improving patient outcomes and contributing to a deeper understanding of the disease.

Acknowledgments

This study was supported by grants 2020E10004 from the Key Laboratory of Intelligent Preventive Medicine of Zhejiang Province, 2019R01007 from the Leading Innovative and Entrepreneur Team Introduction Program of Zhejiang, 2020C03002 from the Key Research and Development Program of Zhejiang Province, and K20230085 from the Healthy Zhejiang One Million Cohort (all to X. W.). The funding source had no role in study design, data collection, data analysis, data interpretation, writing of the report, or the decision to submit the article for publication.

Declaration of interests

No potential conflicts of interest relevant to this article were reported.

Supplemental information

Supplemental information associated with this article can be found online https://doi.org/10.1016/j.trecan.2023.10.006

Outstanding questions

Multiomics data are inherently complex and diverse. How can we address this challenge and develop accurate and reliable AI models that cover the continuum of cancer?

The sheer volume of cancer big data, including high-resolution medical images and complex genomic sequences, requires substantial computational power and storage capacity. What strategies can be employed to overcome these limitations and facilitate efficient data analysis?

Incomplete or missing data can significantly impact the quality of AI models, leading to inaccurate predictions and compromising findings. How can we develop a unified framework for handling missing data?

With the integration of various data types, ensuring data privacy and security becomes increasingly challenging. What are the best practices for maintaining data privacy and security, especially when data is sourced from multiple institutions?

Cancer data often include time-series elements, such as longitudinal patient records and real-time monitoring data. These pose unique challenges in data integration and analysis. How can these challenges be effectively managed to enhance predictive modeling in oncology?

These different types of data have their own unique formats and structures, making it challenging to create a unified representation for analysis. How can AI and big data technologies be leveraged to create a unified data representation that facilitates more effective and comprehensive analysis?

The effective utilization of big data and AI necessitates specialized skills in both the medical and computational domains. What strategies can academic and research institutions adopt to enhance talent training for the application of big data and AI in precision oncology?

The complexity of big data demands a collaborative approach involving oncologists, data scientists, and computational biologists. How can academic institutions and healthcare organizations foster an environment that encourages multidisciplinary collaboration?

References

1. Chen, S. *et al.* (2023) Estimates and projections of the global economic cost of 29 cancers in 204 countries and territories from 2020 to 2050. *JAMA Oncol.* 9, 465–472
2. Swanson, K. *et al.* (2023) From patterns to patients: advances in clinical machine learning for cancer diagnosis, prognosis, and treatment. *Cell* 186, 1772–1791
3. Qi, T. *et al.* (2023) Differentially private knowledge transfer for federated learning. *Nat. Commun.* 14, 3785
4. Kalra, S. *et al.* (2023) Decentralized federated learning through proxy model sharing. *Nat. Commun.* 14, 2899
5. Fatoum, H. *et al.* (2021) Blockchain integration with digital technology and the future of health care ecosystems: systematic review. *J. Med. Internet Res.* 23, e19846
6. Price 2nd, W.N. and Cohen, I.G. (2019) Privacy in the age of medical big data. *Nat. Med.* 25, 37–43
7. Steyaert, S. *et al.* (2023) Multimodal data fusion for cancer biomarker discovery with deep learning. *Nat. Mach. Intell.* 5, 351–362
8. Boehm, K.M. *et al.* (2022) Harnessing multimodal data integration to advance precision oncology. *Nat. Rev. Cancer* 22, 114–126
9. Zhou, H.-Y. *et al.* (2023) A transformer-based representation-learning model with unified processing of multimodal input for clinical diagnostics. *Nat. Biomed. Eng.* 7, 743–755
10. Steyaert, S. *et al.* (2023) Multimodal deep learning to predict prognosis in adult and pediatric brain tumors. *Commun. Med.* 3, 44
11. Lin, Y. *et al.* (2023) Multimodal learning on graphs for disease relation extraction. *J. Biomed. Inform.* 143, 104415
12. Zhu, C. *et al.* (2022) Multimodal reasoning based on knowledge graph embedding for specific diseases. *Bioinformatics (Oxford, England)* 38, 2235–2245
13. Gogleva, A. *et al.* (2022) Knowledge graph-based recommendation framework identifies drivers of resistance in EGFR mutant non-small cell lung cancer. *Nat. Commun.* 13, 1667
14. Tsai, P.C. *et al.* (2023) Histopathology images predict multi-omics aberrations and prognoses in colorectal cancer patients. *Nat. Commun.* 14, 2102
15. Ektefaie, Y. *et al.* (2021) Integrative multiomics-histopathology analysis for breast cancer classification. *NPJ Breast Cancer* 7, 147
16. Mangiante, L. *et al.* (2023) Multiomic analysis of malignant pleural mesothelioma identifies molecular axes and specialized tumor profiles driving intertumor heterogeneity. *Nat. Genet.* 55, 607–618
17. Zheng, C. *et al.* (2023) Multiomics analyses reveal DARS1-AS1/YBX1-controlled posttranscriptional circuits promoting glioblastoma tumorigenesis/radioresistance. *Sci. Adv.* 9, eadf3984
18. Liu, Y. *et al.* (2023) Bioinformatics: advancing biomedical discovery and innovation in the era of big data and artificial intelligence. *Innov. Med.* 1, 100012
19. Liu, S.Q. *et al.* (2022) Single-cell and spatially resolved analysis uncovers cell heterogeneity of breast cancer. *J. Hematol. Oncol.* 15, 19

20. Cao, Z.J. and Gao, G. (2022) Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nat. Biotechnol.* 40, 1458–1466
21. Li, Y. *et al.* (2023) Multi-omics integrated circulating cell-free DNA genomic signatures enhanced the diagnostic performance of early-stage lung cancer and postoperative minimal residual disease. *EBioMedicine* 91, 104553
22. Zhang, Y. *et al.* (2019) High-throughput phenotyping with electronic medical record data using a common semi-supervised approach (PheCAP). *Nat. Protoc.* 14, 3426–3444
23. Zhou, D. *et al.* (2022) Multiview Incomplete Knowledge Graph Integration with application to cross-institutional EHR data harmonization. *J. Biomed. Inform.* 133, 104147
24. Sadilek, A. *et al.* (2021) Privacy-first health research with federated learning. *NPJ Digit. Med.* 4, 132
25. Hong, C. *et al.* (2021) Clinical knowledge extraction via sparse embedding regression (KESER) with multi-center large scale electronic health record data. *NPJ Digit. Med.* 4, 151
26. Mahmood, S.S. *et al.* (2014) The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective. *Lancet (London, England)* 383, 999–1008
27. Ramirez, A.H. *et al.* (2021) Progress with the All of Us research program: opening access for researchers. *Jama* 325, 2441–2442
28. Rusk, N. (2018) The UK Biobank. *Nat. Methods* 15, 1001
29. Blum, A. *et al.* (2018) SnapShot: TCGA-analyzed tumors. *Cell* 173, 530
30. Cheng, I. *et al.* (2022) Traffic-related air pollution and lung cancer incidence: the California multiethnic cohort study. *Am. J. Respir. Crit. Care Med.* 206, 1008–1018
31. Stamatakis, E. *et al.* (2023) Vigorous intermittent lifestyle physical activity and cancer incidence among nonexercising adults: The UK Biobank Accelerometry Study. *JAMA Oncol.* 9, 1255–1259
32. Maroto-Rodriguez, J. *et al.* (2023) Association of a Mediterranean lifestyle with all-cause and cause-specific mortality: a prospective study from the UK Biobank. *Mayo Clin. Proc.* Published online August 16, 2023. https://doi.org/10.1016/j.mayocp.2023.05.031
33. Liang, S. *et al.* (2023) Polygenic risk for termination of the 'healthspan' and its interactions with lifestyle factors: a prospective cohort study based on 288,359 participants. *Maturitas* 175, 107786
34. Celis-Morales, C.A. *et al.* (2018) Associations of discretionary screen time with mortality, cardiovascular disease and cancer are attenuated by strength, fitness and physical activity: findings from the UK Biobank study. *BMC Med.* 16, 77
35. Lin, J. *et al.* (2012) Intake of red meat and heterocyclic amines, metabolic pathway genes and bladder cancer risk. *Int. J. Cancer* 131, 1892–1903
36. Zhang, L. *et al.* (2017) MicroRNA-related genetic variants in iron regulatory genes, dietary iron intake, microRNAs and lung cancer risk. *Ann. Oncol. Off. J. Eur. Soc. Med. Oncol.* 28, 1124–1129
37. Melkonian, S.C. *et al.* (2016) Gene-environment interaction of genome-wide association study-identified susceptibility loci and meat-cooking mutagens in the etiology of renal cell carcinoma. *Cancer* 122, 108–115
38. Wen, C.P. *et al.* (2011) Minimum amount of physical activity for reduced mortality and extended life expectancy: a prospective cohort study. *Lancet (London, England)* 378, 1244–1253
39. Rezende, L.F.M. *et al.* (2020) Resistance training and total and site-specific cancer risk: a prospective cohort study of 33,787 US men. *Br. J. Cancer* 123, 666–672
40. Claussnitzer, M. *et al.* (2020) A brief history of human disease genetics. *Nature* 577, 179–189
41. Byun, J. *et al.* (2022) Cross-ancestry genome-wide meta-analysis of 61,047 cases and 947,237 controls identifies new susceptibility loci contributing to lung cancer. *Nat. Genet.* 54, 1167–1177
42. Amos, C.I. *et al.* (2008) Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. *Nat. Genet.* 40, 616–622
43. Li, Y. *et al.* (2010) Genetic variants and risk of lung cancer in never smokers: a genome-wide association study. *Lancet Oncol.* 11, 321–330
44. Farashi, S. *et al.* (2019) Post-GWAS in prostate cancer: from genetic association to biological contribution. *Nat. Rev. Cancer* 19, 46–59
45. Wu, L. *et al.* (2018) A transcriptome-wide association study of 229,000 women identifies new candidate susceptibility genes for breast cancer. *Nat. Genet.* 50, 968–978
46. Yang, X. *et al.* (2023) Polygenic scores in cancer. *Nat. Rev. Cancer* 23, 619–630
47. Nyberg, T. *et al.* (2023) CanRisk-Prostate: a comprehensive, externally validated risk model for the prediction of future prostate cancer. *J. Clin. Oncol.* 41, 1092–1104
48. Li, K. *et al.* (2023) Salivary extracellular MicroRNAs for early detection and prognostication of esophageal cancer: a clinical study. *Gastroenterology* 165, 932–945.e9
49. Jung, G. *et al.* (2020) Epigenetics of colorectal cancer: biomarker and therapeutic potential. *Nat. Rev. Gastroenterol. Hepatol.* 17, 111–130
50. So, J.B.Y. *et al.* (2021) Development and validation of a serum microRNA biomarker panel for detecting gastric cancer in a high-risk population. *Gut* 70, 829–837
51. Jin, F. *et al.* (2021) A novel class of tsRNA signatures as biomarkers for diagnosis and prognosis of pancreatic cancer. *Mol. Cancer* 20, 95
52. Lomas, D.J. and Ahmed, H.U. (2020) All change in the prostate cancer diagnostic pathway. *Nat. Rev. Clin. Oncol.* 17, 372–381
53. Zheng, Y. *et al.* (2020) Specific gut microbiome signature predicts the early-stage lung cancer. *Gut Microbes* 11, 1030–1042
54. Flemer, B. *et al.* (2018) The oral microbiota in colorectal cancer is distinctive and predictive. *Gut* 67, 1454–1463
55. Hou, Y. *et al.* (2021) Integration of clinicopathologic identification and deep transferrable image feature representation improves predictions of lymph node metastasis in prostate cancer. *EBioMedicine* 68, 103395
56. Chen, M. *et al.* (2023) A novel radiogenomics biomarker for predicting treatment response and pneumotoxicity from programmed cell death protein or ligand-1 inhibition immunotherapy in NSCLC. *J. Thorac. Oncol.* 18, 718–730
57. Kim, J. *et al.* (2021) Comprehensive survey of recent drug discovery using deep learning. *Int. J. Mol. Sci.* 22, 9983
58. Lee, J.W. *et al.* (2022) Big data and artificial intelligence (AI) methodologies for computer-aided drug design (CADD). *Biochem. Soc. Trans.* 50, 241–252
59. Luo, X. *et al.* (2021) A computational framework to analyze the associations between symptoms and cancer patient attributes post chemotherapy using EHR data. *IEEE J. Biomed. Health Inform.* 25, 4098–4109
60. Zhavoronkov, A. *et al.* (2019) Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat. Biotechnol.* 37, 1038–1040
61. Ye, Q. *et al.* (2021) A unified drug-target interaction prediction framework based on knowledge graph and recommendation system. *Nat. Commun.* 12, 6775
62. Su, P.W. and Chen, B.S. (2022) Systems drug design for muscle invasive bladder cancer and advanced bladder cancer by genome-wide microarray data and deep learning method with drug design specifications. *Int. J. Mol. Sci.* 23, 13869
63. He, L. *et al.* (2018) Patient-customized drug combination prediction and testing for t-cell prolymphocytic leukemia patients. *Cancer Res.* 78, 2407–2418
64. Paver, E.C. *et al.* (2021) Programmed death ligand-1 (PD-L1) as a predictive marker for immunotherapy in solid tumours: a guide to immunohistochemistry implementation and interpretation. *Pathology* 53, 141–156
65. Zheng, L. *et al.* (2021) Pan-cancer single-cell landscape of tumor-infiltrating T cells. *Science* 374, abe6474
66. McCulloch, J.A. *et al.* (2022) Intestinal microbiota signatures of clinical response and immune-related adverse events in melanoma patients treated with anti-PD-1. *Nat. Med.* 28, 545–556
67. Greten, T.F. *et al.* (2023) Biomarkers for immunotherapy of hepatocellular carcinoma. *Nat. Rev. Clin. Oncol.* 20, 780–798
68. Jing, Y. *et al.* (2022) Harnessing big data to characterize immune-related adverse events. *Nat. Rev. Clin. Oncol.* 19, 269–280
69. Ngiam, K.Y. and Khor, I.W. (2019) Big data and machine learning algorithms for health-care delivery. *Lancet Oncol.* 20, e262–e273
70. Lou, B. *et al.* (2019) An image-based deep learning framework for individualizing radiotherapy dose. *Lancet Digit. Health* 1, e136–e147
71. Nguyen, D. *et al.* (2019) A feasibility study for predicting optimal radiation therapy dose distributions of prostate cancer patients from patient anatomy using deep learning. *Sci. Rep.* 9, 1076

72. Jiang, P. *et al.* (2022) Big data in basic and translational cancer research. *Nat. Rev. Cancer* 22, 625–639
73. Taylor, S.A. *et al.* (2008) CT colonography: computer-aided detection of morphologically flat T1 colonic carcinoma. *Eur. Radiol.* 18, 1666–1673
74. Liu, X. *et al.* (2021) Deep learning radiomics-based prediction of distant metastasis in patients with locally advanced rectal cancer after neoadjuvant chemoradiotherapy: a multicentre study. *EBioMedicine* 69, 103442
75. Fang, J. *et al.* (2020) Association of MRI-derived radiomic biomarker with disease-free survival in patients with early-stage cervical cancer. *Theranostics* 10, 2284–2292
76. Trivizakis, E. *et al.* (2021) Deep radiotranscriptomics of non-small cell lung carcinoma for assessing molecular and histology subtypes with a data-driven analysis. *Diagnostics (Basel)* 11, 2383
77. Lococo, F. *et al.* (2023) Lung cancer multi-omics digital human avatars for integrating precision medicine into clinical practice: the LANTERN study. *BMC Cancer* 23, 540
78. Boehm, K.M. *et al.* (2022) Multimodal data integration using machine learning improves risk stratification of high-grade serous ovarian cancer. *Nat. Cancer* 3, 723–733
79. Wen, C.P. *et al.* (2012) Hepatocellular carcinoma risk prediction model for the general population: the predictive power of transaminases. *J. Natl. Cancer Inst.* 104, 1599–1611
80. Wu, X. *et al.* (2016) Personalized risk assessment in never, light, and heavy smokers in a prospective cohort in Taiwan. *Sci. Rep.* 6, 36482
81. Labarere, J. *et al.* (2014) How to derive and validate clinical prediction models for use in intensive care medicine. *Intensive Care Med.* 40, 513–527
82. Gu, J. *et al.* (2022) Deep learning radiomics of ultrasonography can predict response to neoadjuvant chemotherapy in breast cancer at an early stage of treatment: a prospective study. *Eur. Radiol.* 32, 2099–2109
83. Tian, P. *et al.* (2021) Assessing PD-L1 expression in non-small cell lung cancer and predicting responses to immune checkpoint inhibitors using deep learning on computed tomography images. *Theranostics* 11, 2098–2107
84. Wu, X. *et al.* (2017) Personalized prognostic prediction models for breast cancer recurrence and survival incorporating multidimensional data. *J. Natl. Cancer Inst.* 109, djw314
85. Kulkarni, P.M. *et al.* (2020) Deep learning based on standard H&E images of primary melanoma tumors identifies patients at risk for visceral recurrence and death. *Clin. Cancer Res.* 26, 1126–1134
86. Weiser, M.R. *et al.* (2008) Individualized prediction of colon cancer recurrence using a nomogram. *J. Clin. Oncol.* 26, 380–385
87. Li, H. *et al.* (2019) Deep convolutional neural networks for imaging data based survival analysis of rectal cancer. *Proc. IEEE Int. Symp. Biomed. Imaging* 2019, 846–849
88. Kather, J.N. *et al.* (2019) Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLoS Med.* 16, e1002730