

REVIEW OPEN ACCESS

AI and Big Data in Oncology: A Physician-Centered Perspective on Emerging Clinical and Research Applications

Binliang Liu¹  | Qingyao Shang² | Jun Li³ | Shuna Yao⁴ | Meishuo Ouyang⁵ | Yu Wang⁶ | Sheng Luo⁷ | Quchang Ouyang¹

¹Department of Breast Cancer Medical Oncology, Hunan Cancer Hospital/the Affiliated Cancer Hospital of Xiangya School of Medicine, Central South University, Changsha, Hunan, China | ²Department of Breast Surgical Oncology, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China | ³Department of Applied and Computational Mathematics and Statistics, College of Science, University of Notre Dame, Notre Dame, USA | ⁴Department of Internal Medicine, Affiliated Cancer Hospital of Zhengzhou University & Henan Cancer Hospital, Zhengzhou, Henan, China | ⁵Department of Surgery, Duke University School of Medicine, Duke University, Durham, North Carolina, USA | ⁶Department of Mathematics, Louisiana State University, Baton Rouge, Louisiana, USA | ⁷Department of Biostatistics & Bioinformatics, Duke University, Durham, North Carolina, USA

Correspondence: Sheng Luo (sheng.luo@duke.edu) | Quchang Ouyang (oyqc1969@126.com)

Received: 11 June 2025 | **Revised:** 1 December 2025 | **Accepted:** 11 December 2025

Funding: Hunan Provincial Natural Science Foundation of China, Grant/Award Numbers: 2024JJ6289, 2023JJ60464, 2023JJ60334; Changsha City Technology Program, Grant/Award Number: kq2403120; Climb Plan of Hunan Cancer Hospital, Grant/Award Numbers: ZX2021005, QH2023006; High-Level Talent Support Program of Hunan Cancer Hospital, Grant/Award Number: 20250731-1050

Keywords: artificial intelligence | big data | cancer | challenges and solutions | clinical applications | research design

ABSTRACT

The convergence of artificial intelligence (AI) and big data is reshaping contemporary oncology by enabling the integration of multimodal information across imaging, pathology, genomics, and clinical records. From a physician-centered perspective, these technologies can potentially be used to improve diagnostic precision, support individualized treatment planning, enhance longitudinal patient management, and accelerate both clinical and translational research. In this review, we synthesize the core AI methodologies most relevant to oncology—machine learning, deep learning, and large language models—and examine how they interact with established and emerging oncology data platforms. We further highlight practical use cases in clinical workflows and research pipelines, emphasizing opportunities for advancing precision cancer care while also addressing challenges associated with data heterogeneity, model generalizability, privacy protection, and real-world implementation. By underscoring the synergistic value of AI and big data, this review aims to inform the development of clinically meaningful, context-adapted strategies that promote translational innovation in both global and locally resourced healthcare environments.

1 | Current Landscape of Oncology Research

Rapid advancements in oncology have created new challenges for researchers, driven by pronounced tumor heterogeneity, exponential data growth, fragmented data systems, and

fast-paced technological evolution [1]. Cancer heterogeneity—spanning genetic, molecular, and phenotypic variations—has strengthened the demand for precision medicine. For instance, Epidermal Growth Factor Receptor mutations and Anaplastic Lymphoma Kinase rearrangements in non-small cell lung

Abbreviations: AI, artificial intelligence; CNNs, convolutional neural networks; COSMIC, Catalog of Somatic Mutations in Cancer; DL, deep learning; DNNs, deep neural networks; EHRs, electronic health records; FL, Federated Learning; LLMs, large language models; ML, Machine learning; NCCD, National Cancer Center Database; NSCLC, non-small cell lung cancer; TCM, traditional Chinese medicine.

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2026 The Author(s). *Cancer Innovation* published by John Wiley & Sons Ltd on behalf of Tsinghua University Press.

cancer (NSCLC) necessitate tailored targeted therapies, underscoring the complexity of precision diagnostics and treatment [2]. In many countries, including China, oncology data remain scattered across institutions, with electronic health records (EHRs), imaging, pathology images, and genomic data lacking integration. Additionally, access to international databases such as the TCGA and SEER is limited for Chinese researchers due to data governance and regional policy restrictions, which further complicates data utilization [3]. Moreover, the rapid evolution of research technologies, including new experimental and statistical methods, requires continuous updates of knowledge from physicians—an ongoing challenge compounded by uneven resource distribution and high technical barriers [4].

Oncologists face dual pressures from intensive clinical workloads and increasing research expectations. According to Global Cancer Statistics 2020, China reports approximately 4.57 million new cancer cases annually, accounting for 23.7% of the global total [5], requiring physicians to make rapid decisions amid heavy clinical workloads while meeting academic and evaluation expectations. The rapid expansion and multimodal complexity of research data have outpaced the analytical capacity of humans, driving greater reliance on computational tools. In this context, artificial intelligence (AI), which has evolved over the past 70 years, is advancing at an unprecedented pace, offering new opportunities for cancer research [1, 2, 6]. Machine learning (ML) and deep learning (DL) improve diagnostic efficiency, while big data integrates heterogeneous resources to optimize treatment and research workflows, accelerating progress in oncology [1, 6].

From a physician's perspective, this review explores how AI and big data can advance oncology and support Chinese physicians in overcoming current challenges. Being a narrative review, it does not follow PRISMA guidelines; instead, it selectively integrates representative literature, illustrative cases, and physician-centered perspectives to highlight emerging opportunities and challenges associated with AI and big data in oncology. The relevant literature and AI tools were identified through narrative, nonsystematic searches in PubMed, Google Scholar, arXiv, and CNKI, with priority given to clinically impactful and widely adopted applications, particularly those relevant to Chinese healthcare settings.

2 | Efficient Support From AI Tools

The widespread adoption of AI technologies provides oncologists with efficient support in both research and clinical practice. By leveraging ML, DL, and large language models (LLMs), AI tools integrate diverse data sources and analyze multidimensional clinical, imaging, and molecular data to optimize workflows and accelerate research in diagnostics, drug development, and prognosis evaluation [7, 8]. From a physician's perspective, these tools not only increase diagnostic and therapeutic efficiency but also improve research accessibility and cost-effectiveness.

2.1 | Adoption and Impact of AI Technologies in Clinical and Research Settings

The core strength of AI technologies lies in their data processing and pattern recognition capabilities, with ML, DL, and

LLMs each providing distinct and complementary support for oncologists.

2.1.1 | Machine Learning

ML encompasses a range of algorithms that can analyze rapidly expanding datasets and identify patterns for risk stratification and survival prediction [3, 9]. It excels in handling high-dimensional data and uncovering nonlinear relationships, significantly increasing prediction accuracy. ML operates on principles such as supervised learning (training models with labeled data, e.g., predicting patient survival), unsupervised learning (identifying hidden patterns or associations, e.g., tumor subtype clustering), and reinforcement learning (optimizing decisions through trial and error, e.g., treatment plan optimization). Supervised learning is widely applied in survival prediction; for example, models trained with labeled data using radiomic features achieve an AUC of > 0.80 in risk stratification for NSCLC [10]. Unsupervised learning facilitates the identification of cancer subtypes and heterogeneity, such as molecular subtyping in pancreatic cancer [11]. Reinforcement learning is effective for treatment optimization; for example, radiotherapy dose allocation can be refined for NSCLC through iterative learning, thereby improving treatment precision [12].

2.1.2 | Deep Learning

DL employs artificial neural networks with multiple layers of simulated neurons, making it particularly adept at image and video processing. In oncology, DL is extensively used for interpreting CT and MR images and pathology slides [7, 13]. For instance, convolutional neural networks (CNNs) can automatically detect lesions in lung cancer CT images, achieving high diagnostic sensitivity and specificity, with an AUC of 0.949 [14, 15]. In pathology, CNNs can detect lymph node metastases in patients with breast cancer, with an AUC of 0.996 [16]. With respect to breast cancer screening, CNN-driven computer-aided detection systems outperform traditional methods because of their low sensitivity, ability to match human readers, and ability to substantially reduce physicians' workload [17]. Techniques such as lesion annotation and whole-image labeling enable end-to-end training, continuously improving the accuracy of screening for breast cancer [18]. Similar DL models applied to the TCGA dataset can predict survival and recurrence with an AUC of 0.91 and a sensitivity of 98% [19], and they can analyze DNA methylation data to accurately classify disease subtypes [20]. U-Net, for example, processes radiomic data with a screening accuracy of 93% [19]. Furthermore, during surgery, DL can analyze video feeds to identify critical anatomical structures in real time, delineate tumor boundaries, increase surgical precision and safety, and reduce errors [21, 22]. In digitized pathology slide analysis, DL demonstrates superiority in diagnosing breast, lung, and brain tumors [3, 23–25]. Through automated feature extraction and quantitative analysis, these technologies reveal complex gene-driven mutation patterns [24] and identify potential subtypes [25], enabling precise diagnosis and providing oncologists with efficient clinical decision support [3].

2.1.3 | Large Language Models

Based on natural language processing, LLMs can generate text, synthesize literature, and answer questions, and they excel in

literature reviews and physician–patient communication. The widespread adoption of these technologies provides physicians with comprehensive support, from data analysis to decision-making, and has been proven to rapidly generate structured research reviews, increasing research efficiency [4, 26]. Additionally, LLMs can produce health recommendations, offering physicians and patients strategies and plans for treatment and care, and thereby assisting with routine clinical tasks [26].

2.2 | Key AI Tools and Their Applications

To illustrate the diversity of AI applications in oncology, we summarize representative tools frequently used in clinical practice and research.

2.2.1 | ChatGPT

ChatGPT, which was developed by OpenAI, is a multimodal LLM with broad capability that can process text, images, and structured clinical data. It is particularly useful for literature review and hypothesis generation, enabling physicians to rapidly construct research ideas and study frameworks. Website: <https://chat.openai.com>.

2.2.2 | Grok

Grok, which was launched by xAI, is a real-time inference model designed for rapid validation of clinical hypotheses. It can assist in evaluating antitumor drug efficacy, assessing treatment data reliability, and supporting real-time clinical decision-making, such as chemotherapy regimen optimization. Website: <https://x.ai/grok>.

2.2.3 | DeepSeek

DeepSeek, an open-source LLM developed in China, offers low training costs and strong adaptability to Chinese clinical data. It is particularly suitable for processing Chinese EHRs and tumor registry data, and its strong performance in Chinese-language generation supports oncology research in China.

2.2.4 | Claude

Claude, which was developed by Anthropic, is an inference-optimized LLM with strong interpretability that can analyze complex treatment plans, generate scientific text, and increase academic writing efficiency. It is well-suited for drafting clinical trial designs in oncology. Website: <https://www.anthropic.com/claude>.

2.2.5 | Gemma and Med-Gemma

Gemma, which was developed by Google DeepMind, is an open-source LLM designed for lightweight, privacy-preserving, and scalable deployment. While Gemma itself is a general-purpose model, it provides the architectural foundation for specialized healthcare models such as Med-Gemma. Med-Gemma, which was released in 2024, is a multimodal medical LLM optimized for the joint interpretation of clinical text and medical imaging data. It supports advanced tasks, including radiology report generation and visual question answering,

demonstrating superior performance in multimodal medical reasoning. Website: <https://ai.google.dev/gemma>.

2.2.6 | Other LLMs Developed in China

These models demonstrate similar potential in applications in oncology, including literature synthesis, clinical data processing, and patient communication support, thereby improving both clinical and research efficiency. They generally excel at processing Chinese text. The following are several representative models and their applications.

2.2.6.1 | Qwen (Tongyi Qianwen, Alibaba Cloud)

Qwen, which is an open-source multimodal LLM, excels at processing Chinese text, images, and data, which makes it suitable for generating literature reviews in oncology, analyzing EHRs, and optimizing clinical trial designs. Its low cost and high performance support a wide range of enterprise applications. Website: <https://www.aliyun.com/product/qwen>.

2.2.6.2 | ERNIE Bot (Wenxin Yiyan, Baidu)

ERNIE Bot, which is a Chinese-optimized LLM, supports intelligent consultations and literature processing, providing health advice to cancer patients (e.g., addressing chemotherapy side effects) or generating cancer research reports, thereby improving diagnostic and therapeutic communication efficiency. Website: <https://yiyan.baidu.com/>.

2.2.6.3 | KIMI (Moonshot AI)

KIMI, which is an efficient Chinese LLM, focuses on long-text processing and knowledge retrieval, which makes it suitable for rapidly accessing oncology research updates or generating structured research documents, such as reviews on targeted therapies. Website: <https://kimi.moonshot.cn/>.

2.2.6.4 | Tencent Yuanbao (Tencent)

Tencent Yuanbao, which is a multimodal LLM based on the Hunyuan model, supports literature summarization, patient education, and multimodal data analysis. These uses make it suitable for generating oncology patient guidance materials or analyzing treatment outcomes. Website: <https://yuanbao.tencent.com/>.

2.2.6.5 | Other Distinctive LLMs

2.2.6.5.1 | DISC-MedLLM [27]

DISC-MedLLM, which was designed specifically for medical dialog scenarios, is an LLM based on the Baichuan-13B model. It is fine-tuned with high-quality medical knowledge graphs and real dialog datasets. It excels at single- and multiturn medical consultations and can be used for patient inquiries, generating clinical records, or supporting telemedicine. Website: <https://www.baichuan-ai.com/home> (company website).

ZhongJing, which is a domain-specific LLM designed for traditional Chinese medicine (TCM), is fine-tuned in the TCM field to support TCM diagnosis, treatment assistance, and training. In oncology, ZhongJing can be used to analyze TCM literature, generate cancer-related TCM treatment plans (e.g., optimizing herbal prescriptions), or support patient health education. Website: Not available (academic project).

3 | Big Data Platforms: Physicians' Perspective on Data Resources

Big data platforms are designed to store, integrate, and analyze large-scale medical data, covering molecular omics (genomics, transcriptomics, epigenetics, proteomics, and metabolomics), perturbed phenotypes, molecular interactions, imaging, and clinical text data [3]. For instance, EHRs document patient histories and treatment responses, multiomics data reveal tumor molecular profiles, and imaging data aid in lesion localization [3]. From a physician's perspective, these platforms increase diagnostic precision and expand research resources, while localized platforms are particularly valuable in China considering the access to many international databases is restricted. Big data platforms can be broadly categorized into traditional and emerging types, each serving distinct research and clinical applications (Table 1).

3.1 | Traditional Big Data Platforms

Traditional platforms typically contain regional or institution-based datasets with relatively standardized data collection and long-term follow-up, which makes them foundational resources for oncology research. Notable examples include the SEER, TCGA, GEO, and Catalog of Somatic Mutations in Cancer (COSMIC) databases and the National Cancer Center Database (NCCD), which provide standardized data for epidemiological and molecular studies but often lack global representativeness or detailed clinical annotations (see Table 1 for details).

3.2 | Emerging Big Data Platforms

Emerging platforms leverage cloud computing, cross-institutional data sharing, and interactive analytical tools to provide dynamic and scalable support, representing the future direction of data integration in oncology. Platforms such as Vivli, GBD, cBioPortal, ICGC, UK Biobank, CKB, and ImageNet offer expanded data coverage and cloud-based analytical capabilities but may be limited by access policies, population bias, or insufficient clinical depth (see Table 1 for details).

4 | Empowering Clinical Practice and Research With AI and Big Data

The synergistic integration of AI and big data technologies, particularly through multimodal data analysis, significantly increases the precision and efficiency of oncology practice while simultaneously advancing research. Big data platforms such as COSMIC, TCGA, and the UK Biobank, in addition to high-quality localized

institutional databases, provide vast standardized datasets, while AI algorithms extract insights via DL and ML. Together, they enable comprehensive optimization across the entire continuum—from tumor screening and diagnosis to treatment decision-making, as well as from research design to outcome evaluation [3].

4.1 | Empowering Clinical Practice

4.1.1 | Cancer Screening and Imaging Diagnosis Support

Advances in diagnostic imaging and biomedical image analysis have enabled unprecedented access to structural and metabolic information. AI integrates multimodal data, including imaging (e.g., X-ray, ultrasound, and MRI), EHRs, and molecular profiles (e.g., gene expression and mutation status), thereby significantly improving diagnostic accuracy [1, 10, 48]. AI demonstrates a core advantage in image analysis, leveraging standardized imaging data from platforms such as CBIS-DDSM and the UK Biobank and genomic insights from the TCGA. When DL is used, AI algorithms can rapidly detect lesions, reduce misdiagnosis rates, and achieve diagnostic performance comparable to that of experienced specialists [4, 49, 50].

AI excels at breast cancer screening, particularly in detecting microcalcifications, which are early lesion markers [48]. Studies using the Transpara AI system to analyze mammography images from the CBIS-DDSM dataset achieved automated detection of microcalcifications and soft tissue lesions, with an AUC of 0.84, outperforming the radiologists by 61.4% [50, 51]. Shen et al. [18] further validated the utility of AI in ultrasound imaging and developed an “end-to-end” CNN based on the CBIS-DDSM dataset, achieving an AUC of 0.91. Fine-tuning further increased diagnostic accuracy and reduced the number of unnecessary biopsies. Additionally, AI combined with digital breast tomosynthesis data has been shown to improve early breast cancer detection sensitivity. Using institutional datasets, Geras et al. [52] reported an AUC of 0.87 with AI models in DBT, minimizing the number of false positives. Dynamic contrast-enhanced MRI, a DL model developed by Ayatollahi et al. [53] that is based on the Radboud University Medical Center breast MRI dataset, analyzed the spatiotemporal patterns of contrast uptake, enabling ultrafast lesion segmentation with an AUC of 0.85.

These results underscore the dependence on standardized imaging data from CBIS-DDSM or other high-quality databases, highlighting the synergy between AI algorithms and high-quality data in yielding superior models.

4.1.2 | Support for Tumor Pathology Diagnosis

AI applies deep neural networks (DNNs) to digitized pathology slides, demonstrating high efficiency and consistency in diagnosing cancers such as breast, lung, and brain tumors [1, 23–25]. For instance, Bejnordi et al. [23] confirmed that DNNs reduce the diagnosis time and improve the classification consistency when large pathology datasets are used. In lung cancer, AI utilizes whole-slide images from the TCGA to accurately classify cancer types and detect specific gene-driven mutation patterns that are undetectable by traditional pathologists [24]. Additionally, AI leverages tumor DNA methylation

TABLE 1 | Comparison of big data platforms in oncology.

| Platform | Data coverage | Sample size | Cancer types covered | Access level | Population representation | Clinical linkage | Strengths | Limitations | Website |
|-------------------|--|---------------------|----------------------------------|---------------------|----------------------------------|-------------------------|---|--|---|
| SEER [29] | U.S. cancer epidemiology | ~10 M cases | All major cancers (U.S.) | Open | U.S. population | Yes | High-quality, standardized, survival analysis | Western-focused, differs from Chinese profiles | https://seer.cancer.gov |
| TCGA [30] | Multi-omics | ~11,000 samples | 33 cancer types | Open | Primarily U.S. | Partial | Core resource for molecular research | Slow updates, limited sample size (11,000) | https://cancergenome.nih.gov |
| GEO [31] | Multi-omics | ~6 M samples | Multiple (user-submitted) | Open | Global | No | Diverse data, frequent updates, user uploads | Low standardization, inconsistent annotations | https://www.ncbi.nlm.nih.gov/geo |
| COSMIC [32] | Somatic mutations | > 37,000 genomes | Various somatic mutation data | Open | Global | No | Large data volume, molecular target research | Limited clinical information | https://cancer.sanger.ac.uk/cosmic |
| NCCD/NCDL [33] | Chinese cancer registry | Nationwide registry | All major cancers (China) | Restricted | Chinese | Limited | Convenient access, localized research | Macro-level, lacks multi-omics details | https://www.ncc.org.cn |
| OncoKB [34] | Genomic variants, therapeutic implications | Variant-based | Therapeutically relevant cancers | Open | Global | Yes | Precision oncology, actionable insights | Limited to annotated variants | https://www.oncokb.org |
| GDC [35] | Cancer genomics | TCGA, TARGET, etc. | Multiple cancer types | Open | Primarily U.S. | Partial | Standardized, supports data sharing | Primarily U.S.-based data | https://gdc.cancer.gov |
| ArrayExpress [36] | Functional genomics | Millions | Broad (user-submitted) | Open | Global | No | Diverse data, user submissions | Inconsistent annotations | https://www.ebi.ac.uk/arrayexpress |
| Viviti [37] | Global clinical trial data | Thousands of trials | Varies by trial | Controlled | Global | Yes | High-quality, cloud-based analysis | Strict compliance, high entry barrier | https://viviti.org |
| GBD [38] | Global disease burden | Global estimates | All major cancers | Open | Global | No | Authoritative, macro-level research | Lacks individualized clinical details | https://www.healthdata.org/gbd |
| cBioPortal [39] | Cancer genomics | ~200 studies | Broad (mostly TCGA-linked) | Open | Primarily Western | Yes | Interactive visualization, mutation analysis | Western-focused, limited global representation | https://www.cbioportal.org |

(Continues)

TABLE 1 | (Continued)

| Platform | Data coverage | Sample size | Cancer types covered | Access level | Population representation | Clinical linkage | Strengths | Limitations | Website |
|-----------------|---|-----------------------|------------------------------------|--------------|---------------------------|------------------|---|--|---|
| ICGC [40] | Cancer genomics | ~20,000 samples | 50+ cancer types | Open | Global (incl. China) | Partial | Dynamic updates, includes Chinese data | Limited clinical annotations | https://dcc.icgc.org |
| UK Biobank [41] | Genomics, phenotypes | 500,000 individuals | Various (population study) | Controlled | UK | Yes | Open-access, cloud analysis (UKB-RAP) | Western-focused | https://www.ukbiobank.ac.uk |
| CKB [42] | Chinese health data | 500,000 individuals | Various (Chinese population) | Restricted | Chinese | Yes | Rich localized data | Limited multi-omics depth | https://ckbiobank.pku.edu.cn |
| ImageNet [43] | Annotated images | 14 M images | Not cancer-specific | Open | Non-medical | No | Large volume, supports AI pre-training | Non-medical, requires transfer learning | https://www.image-net.org |
| CIViC [44] | Clinical interpretations of cancer variants | Thousands of variants | Cancer variants (clinical) | Open | Global | Yes | Community-driven, actionable insights | Limited scale, ongoing curation | https://civcdb.org |
| CCLLE [45] | Cancer cell line multi-omics | ~1000 cell lines | Many cancer types (preclinical) | Open | Cell lines | No | Drug response, molecular profiling | Cell line-based, lacks patient context | https://portals.broadinstitute.org/cclle |
| EGA [46] | Genomic and phenotypic data | Hundreds of studies | Various (Europe-focused) | Controlled | European | Partial | Secure access, supports cancer research | Restricted access, complex application | https://ega-archive.org |
| DepMap [47] | Cancer dependency genes, drug targets | > 1000 cell lines | Cancer dependency/target discovery | Open | Cell lines | No | Functional genomics, drug screening | Cell line-focused, limited clinical data | https://depmap.org |
| PCAWG [30] | Pan-cancer whole genomes | ~2,800 genomes | Pan-cancer | Open | Global | Partial | Comprehensive genomic insights | Complex data, limited clinical integration | https://dcc.icgc.org/pcawg |

patterns from large sequencing datasets for ML, markedly improving the accuracy of brain tumor subtype classification over that of traditional histological methods [25].

4.1.3 | Precision Therapy and Personalized Treatment Decision-Making

By integrating multimodal data from big data platforms, along with clinical and treatment response data, AI significantly optimizes breast cancer treatment decision-making, increases clinical precision, and improves patient management efficiency.

In surgical planning, AI uses DL to analyze MR and CT images and automatically segments tumor boundaries to assist in preoperative planning and intraoperative navigation for patients with breast cancer. For example, Zhang et al. achieved tumor segmentation with an AUC of > 0.90 using MRI data, reducing the reliance on manual annotations [54]. CNNs can process intraoperative videos to identify critical anatomical structures in real time during colorectal surgery or liver resection, increasing surgical precision and safety while minimizing errors [21]. In precision breast cancer subtyping, Jiang et al. [55] integrated multidimensional data—genomics, transcriptomics, proteomics, metabolomics, and radiomics—from 1226 breast cancer patients, proposed molecular subtypes for luminal breast cancer, and achieved accurate predictions using AI models. In treatment response prediction, AI integrates multiomics data (e.g., mutation burden and immune infiltration) to predict the pathological complete response to neoadjuvant chemotherapy. For example, Sammut et al. [56] used data from 168 patients to develop an integrated ML model that achieved an AUC of 0.87 in external validation (75 patients), outperforming traditional models. DL architectures such as U-Net architectures precisely delineate lesion boundaries in breast MR images, reducing the reliance on manual expert annotations and improving the efficiency of radiotherapy planning [57]. Through feature selection and multiomics integration, AI has demonstrated substantial potential in breast cancer biomarker discovery and application, providing robust support for precision diagnosis and personalized treatment. A model combining the gray wolf optimization algorithm with support vector machines improved diagnostic accuracy by 27.68% on the TCGA breast cancer dataset by optimizing gene expression features. Similarly, the multiomics graph convolutional network, which integrates TCGA gene expression, DNA methylation, and miRNA data, achieved an F1 score of 0.89 for breast cancer subtype classification, supporting targeted therapies such as HER2-directed drugs [58]. The CURATE AI platform used in chemotherapy for advanced solid tumors can dynamically adjust capecitabine doses on the basis of patient biomarkers (e.g., PSA levels in prostate cancer), improving quality of life and disease control rates and offering a promising approach for managing adverse reactions in the future [59].

AI also facilitates rapid retrieval of the latest treatment guidelines, using natural language processing to access breast cancer protocols, assisting physicians in developing standardized plans, predicting adverse drug reactions, and identifying rare drug interactions to increase treatment safety [26].

4.1.4 | Patient Management

LLMs enhance oncology patient management by providing online consultations, health education, and emotional support,

thus significantly improving patient self-management and reducing clinical burdens. LLMs provide personalized information on disease, treatment, and prevention for breast cancer patients, equipping them with essential knowledge before clinical consultations [60]. During treatment, LLMs facilitate effective physician–patient communication and health education. For instance, Bibault et al. [61] reported that AI chatbots provide breast cancer information comparable to that provided by physicians, alleviating clinical workloads. ML-based health management platforms integrate multimodal data (e.g., heart rate and sleep data from wearable devices and EHRs) to create customized care plans, such as exercise recommendations or medication adherence reminders [2, 62, 63]. LLMs also generate health recommendations, offer personalized care advice, develop individualized care plans [64], and provide emotionally supportive dialogs to reduce patients' psychological stress and promote adherence [65, 66], thereby enabling effective self-management and substantially lowering hospital management and care costs.

4.2 | Empowering Research

In research, AI and big data provide efficient support for physicians from study design to outcome generation, with particular strengths in localized data processing:

4.2.1 | Study Design, Writing, and Figure Generation

LLMs such as ChatGPT rapidly acquire multidisciplinary data and efficiently retrieve the latest studies from medical literature to assist researchers in synthesizing literature, drafting trial protocols, and constructing manuscript frameworks, thereby facilitating research [4, 26, 67]. LLMs such as Grok leverage real-time literature validation to track research trends and emerging evidence, helping ensure that studies remain current and meet publication standards. Additionally, the ability of AI to automatically generate statistical figures and mechanistic diagrams increases manuscript visual quality and supports physicians in conducting high-quality research, overcoming knowledge barriers, and promoting innovative outputs across research institutions.

4.2.2 | Breakthroughs in Cross-Modal Integration and Cross-Cohort Aggregation

The integration of AI with big data enables breakthroughs in elucidating cancer molecular mechanisms and optimizing diagnostics by cleaning heterogeneous data and fusing multimodal datasets. AI models integrate genomic, transcriptomic, and proteomic data to elucidate key signaling pathways. Single-cell multiomics technologies (e.g., scRNA-seq combined with scATAC-seq), analyzed via AI, significantly increase the accuracy of cell lineage classification in the breast cancer microenvironment [68]. Studies have shown that integrating scRNA-seq and scATAC-seq data with AI methods, such as graph neural networks, effectively identifies cancer-associated fibroblast subpopulations (e.g., inflammatory and matrix cancer-associated fibroblast), outperforming traditional methods. Cross-cohort, multicenter data integration mitigates the noise and bias inherent in single datasets, markedly improving

the identification of key driver genes and disease-associated mutations [69].

4.2.3 | AI Prediction of Biological Responses

AI effectively predicts the activity, toxicity, and pharmacokinetic properties of new drug compounds by analyzing multiomics data, including genomics, metabolomics, and proteomics data, offering early screening potential that significantly shortens development timelines and reduces costs [70, 71]. AlphaFold3 surpasses existing specialized tools in predicting protein–ligand, protein–nucleic acid, and antibody–antigen interactions, demonstrating increased accuracy. Additionally, AlphaFold3 can predict the structural impact of post-translational modifications on molecular systems, offering a powerful tool for drug design, biotechnology, and fundamental biological research. This model achieves high-precision modeling across biomolecular spaces through a unified DL framework [72]. In breast cancer drug repurposing, AI integration of multiomics data also has notable advantages: the NCI-DREAM challenge employs multikernel learning and combines genomic, transcriptomic, and proteomic data from breast cancer cell lines to predict drug responses, outperforming single-omics models [4, 73]. Furthermore, random forests and support vector machines achieved an AUC of 0.90 in predicting chemotherapy responses on the GDSC dataset, whereas DNNs predicted drug sensitivity on the CCLE dataset with an AUC of

0.91 [19, 74]. These findings highlight the substantial potential of AI in predicting biological responses (Figure 1).

5 | Challenges and Solutions for AI and Big Data in Oncology

A key emerging opportunity emphasized in this review is the shift from physicians serving as passive users of AI tools to becoming active participants in model development, evaluation, and ongoing optimization. In clinical oncology, physicians possess irreplaceable contextual knowledge regarding disease progression, treatment responses, and patient-specific factors. Integrating this expertise into AI model refinement establishes a physician-centered feedback loop, enabling AI systems to better align with real-world clinical decision-making. This collaborative model has the potential to substantially improve model interpretability, clinical relevance, and patient outcomes, representing a meaningful evolution in the implementation of AI in oncology practice.

Despite their transformative potential, the adoption of AI and big data technologies in oncology is hindered by multiple challenges. Current barriers include heterogeneous data quality, stringent privacy regulations, technical limitations, limited interpretability of models, and uncertainty regarding reproducibility [3]. To mitigate these issues, several strategies have

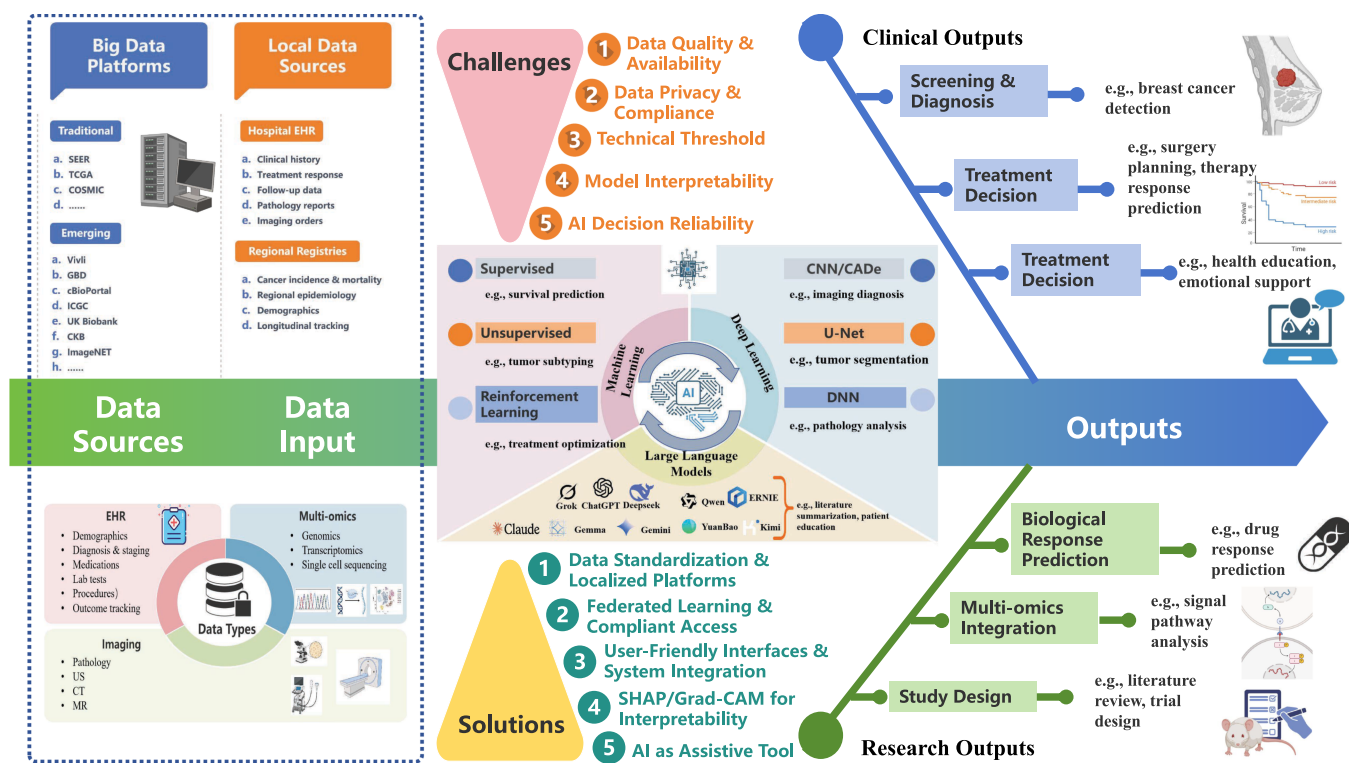


FIGURE 1 | Physician-centered closed-loop framework for the integration of AI and big data into oncology practice. In this model, multimodal patient information—including structured clinical records, radiological and pathological imaging, and genomic or molecular profiling—is systematically collected and standardized through hospital information and big-data platforms. AI models—including machine learning (ML), deep learning (DL), and large language models (LLMs)—trained on these datasets generate diagnostic support, risk stratification outputs, and therapeutic recommendations. Rather than functioning as passive end users, physicians critically evaluate AI outputs, contextualize them within clinical judgment, and adapt decision-making accordingly. Their clinical feedback and outcome observations are then iteratively incorporated into model retraining and refinement. This bidirectional interaction forms a dynamic learning cycle, ensuring that the AI system remains clinically aligned, achieves progressively improved performance, and evolves in response to real-world oncological needs.

emerged, including federated learning (FL) to enable privacy-preserving data sharing, standardized data pipelines, user-centered interface design, and explainable AI frameworks. Collectively, these approaches can foster physician trust and facilitate the integration of AI into precision oncology.

5.1 | Challenges

5.1.1 | Data Quality and Availability

Traditional databases such as the COSMIC focus on somatic mutation data across diverse cancer types but lack comprehensive clinical and multiomics information, limiting their utility in precision medicine [32]. Emerging platforms such as cBioPortal offer interactive genomic analysis but are based primarily on Western samples, potentially underrepresenting global populations, such as Asians with a high prevalence of EGFR mutations [75]. Additionally, data from different centers are often fragmented and insufficiently standardized, undermining the reliability of data integration and analysis [1]. Moreover, the overall scale of cancer data remains limited, with measurement inconsistencies and batch effects further impacting analytical accuracy [3].

5.1.2 | Data Privacy and Compliance

The biomedical field maintains limited data openness, with privacy regulations posing a major obstacle [3]. China's Data Security Law [76] and Personal Information Protection Law prohibit the cross-border transfer of sensitive data without approval [77]; similarly, the U.S. Executive Order 14117 (issued in February 2024) restricts access to sensitive personal U.S. data and government-related data for “countries of concern”, including China, increasing the difficulty of accessing databases such as SEER or the TCGA and hindering international research collaboration and publication. The unique nature of the medical field also demands caution when new technologies are introduced [78]. Despite the robust data processing and integration capabilities of AI, prediction errors persist in clinical settings [79].

5.1.3 | Technical Barriers

Oncologists often lack a data science background, making it challenging to navigate complex platforms and thus limiting the adoption of technology [80]. For instance, cBioPortal's genomic visualization requires programming skills [39], and Vivli's cloud-based analysis demands familiarity with data processing workflows, restricting direct use by clinicians. Additionally, many AI tools operate as standalone systems and struggle to integrate seamlessly with EHRs or picture archiving and communication systems (PACSs). This often requires manual data transfer or uploading, increasing physicians' workload and further hindering widespread adoption [81].

5.1.4 | Lack of Model Interpretability

DL models often lack interpretability because of their “black box” nature, with opaque decision-making processes that prevent physicians from understanding the basis of predictions [82]. This reduces the degree of trust among clinicians and

patients and limits the translation of AI models into clinical practice.

5.1.5 | Model Reliability and Generalizability

Medicine is an inherently complex field compounded by issues such as outdated data. Current language models exhibit limitations in terms of contextual understanding and causal reasoning for complex medical decisions and require caution to avoid misleading recommendations [83, 84]. While AI diagnostic models perform well in specific research settings, their training data often come from single institutions or datasets, leading to poor generalizability across diverse healthcare settings and populations and resulting in unstable performance. Unvalidated AI predictions may result in errors, as models remain vulnerable to real-world data variability, thereby reducing their robustness [4, 26].

5.2 | Solutions to Address Challenges

5.2.1 | FL and Privacy Protection

FL enables the integration of multicenter data while safeguarding privacy through local training and model parameter sharing, making it well suited for oncology research [85, 86]. It incorporates techniques such as differential privacy, homomorphic encryption, and cryptonets to reduce data leakage risks and facilitate secure multicenter collaboration [87]. Unlike traditional centralized training, FL avoids data aggregation, further reducing the risk of parameter leakage and ensuring patient data security. For example, multiple hospitals can share updates to breast cancer treatment response models, increasing the diagnostic accuracy for rare tumors [63–66]. Additionally, cross-border collaboration and compliant data access (e.g., through anonymized data sharing) ensure that platforms such as Vivli and SEER meet regulatory requirements.

5.2.2 | Data Standardization and Localized Platforms

Standardizing data formats across platforms or research centers and addressing missing values can significantly reduce the complexity of data integration, thereby improving data quality. For instance, AI tools such as DeepSeek utilize natural language processing to automate the cleaning of heterogeneous data, enhancing the research utility of platforms such as COSMIC and cBioPortal [26]. The development of localized platforms, such as ICGC and NCCD, helps address gaps in population representation; ICGC, for example, integrates Asian population data, which makes it ideal for studying regional mutation patterns [33]. Optimizing these platforms to support diverse population data further enhances their global applicability.

5.2.3 | User-Friendly Interfaces and System Integration

Designing simplified data input and analysis interfaces can lower technical barriers for users. For example, Vivli's interactive cloud platform enables nonexpert users to conduct research, whereas cBioPortal's visualization tools allow physicians to analyze genetic mutations easily [75]. DeepSeek's intuitive interface supports chat-based interactions and document uploads, enabling primary care physicians to analyze

EHRs or imaging data without programming skills. Embedding AI tools within EHR and PACS systems provides real-time clinical decision support, such as automated lesion annotation, thereby reducing manual workload [88, 89]. Additionally, lightweight AI models leveraging edge computing can operate on local devices in primary hospitals, analyze ultrasound images in real time and address disparities in medical resources [90].

5.2.4 | *Enhancing Model Interpretability*

To improve model interpretability, tools such as SHapley Additive Explanations quantify feature importance by showing the contribution of specific risk factors to prognosis, helping physicians understand model decisions [91]. Similarly, gradient-weighted class activation mapping generates imaging heatmaps to highlight AI-identified lesion areas, assisting radiologists in validating predictions [92]. These interpretability tools markedly improve clinical acceptance, foster physician trust, and support broader implementation [93, 94].

5.2.5 | *AI Decisions as Assistive Only*

AI should function as an assistive tool in clinical and research decision-making—particularly for initial screening or hypothesis generation—enhancing rather than replacing human judgment. AI predictions are best suited for early-stage exploration or brainstorming, with an emphasis on interpretability and quality monitoring. To improve robustness, subsequent validation using cross-cohort or independent external datasets is essential [4, 95].

6 | **Future Directions: Physicians' Expectations and Technological Outlook**

The rapid advancement of AI and big data technologies presents substantial opportunities for the future of oncology. Clinically, AI increases diagnostic precision and treatment efficiency by aiding diagnosis, optimizing therapeutic decisions, and improving patient management. In research, AI and big data accelerate study design, data analysis, and outcome generation, promoting localized research development. The future of “AI + big data” depends not only on technological breakthroughs but also on balancing societal trust with medical ethics, fostering synergy between localization and internationalization, and redefining the evolving role of physicians [96, 97]. Currently, big data platforms in oncology must increase sample sizes and promote data sharing [3, 16] while increasing the digitization of laboratory tests, imaging, and pathology slides [1, 98, 99]. Multimodal AI, generative AI, and edge computing will be pivotal in advancing precision medicine.

Multimodal AI integrates imaging, genomic, and clinical data to provide comprehensive diagnostic and therapeutic support, substantially increasing clinical precision [4]. Generative AI leverages multisource data and standardized platforms to advance tumor drug repurposing, accelerate novel drug development, and optimize combination therapies through toxicity prediction models [71–73]. Edge computing enables the local deployment of AI on devices, reducing reliance on cloud systems and improving real-time performance, such as enabling immediate lesion detection during ultrasound examinations in

primary hospitals and minimizing data transfer delays—an approach particularly suitable for resource-limited settings [100]. These technologies will markedly improve the precision and efficiency of tumor management, especially in primary care settings.

The parallel development of localization and internationalization will be essential in the future. Localized applications can utilize NCCD and CKB data, with FL ensuring privacy while analyzing Chinese population characteristics. Moreover, localized systems can harness EHRs from Chinese hospitals and the NCCD to develop models tailored to local populations, addressing the limitations of restricted international data access. Intelligent service platforms integrating EHRs, omics, and mobile health data can support patient self-management and clinical decision-making, improve model applicability, and provide accessible solutions for primary hospitals, thus promoting equitable medical resource distribution.

The evolving role of physicians will be central to the integration of AI into clinical oncology. Historically, physicians functioned largely as passive adopters, depending on engineers for technical implementation. Looking ahead, however, clinicians are expected to become active innovators—engaging in interdisciplinary collaborations, shaping trial design, and contributing to model development and validation. This transition represents a paradigm shift in which physicians are not only end-users but also active cocreators of AI-enabled precision oncology. For example, by refining model input features with clinical expertise or reannotating AI outputs, physicians can increase the accuracy and applicability of AI in clinical settings [16, 18]. Leveraging the ability of AI to continuously improve with new data, a closed-loop “data–prediction–feedback” system can be established that incorporates real-time clinical data (e.g., imaging and laboratory results) and physician feedback into iterative model training, creating dynamically updated diagnostic support tools [54]. Achieving this vision requires a collaborative ecosystem in which physicians, engineers, and data scientists work together: engineers develop technical frameworks, physicians provide clinical insights, and data scientists optimize algorithms to create AI systems tailored to medical needs. Cloud collaboration and localized model optimization further support this ecosystem, with tools such as edge computing and DeepSeek enabling lightweight AI deployment in primary hospitals for real-time diagnostics [85].

In summary, the future of oncology will be shaped by multimodal and generative AI, edge computing, and localized applications that integrate national databases such as the NCCD. Physicians will be not only users but also cocreators of these systems, ensuring that technological advances align with patient needs and clinical realities.

7 | **Conclusion**

This review from a physician's perspective highlights how AI and big data are reshaping oncology by accelerating research, improving diagnostic accuracy, and supporting patient-centered care. Localized platforms such as DeepSeek and the NCCD, together with FL and edge computing, offer practical approaches to address China's unique challenges in terms of data security and resource distribution. However, realizing the full

potential of these technologies requires careful and responsible implementation. Without adequate safeguards for privacy, interpretability, and data quality, premature adoption may compromise both patient safety and healthcare equity.

Looking ahead, the dual pathway of localization and internationalization will continue to strengthen China's role in global oncology. As physicians evolve from technology users to active collaborators and innovators, they will remain central to ensuring that AI is responsibly implemented to advance precision oncology and improve patient outcomes.

Author Contributions

Binliang Liu: writing – review and editing, writing – original draft, funding acquisition, conceptualization, data curation. **Qingyao Shang:** conceptualization, writing – original draft, writing – review and editing, data curation. **Jun Li:** writing – original draft, writing – review and editing. **Shuna Yao:** writing – original draft, writing – review and editing. **Meishuo Ouyang:** writing – original draft, writing – review and editing, data curation. **Yu Wang:** writing – original draft, writing – review and editing. **Sheng Luo:** writing – original draft, writing – review and editing, supervision, conceptualization. **Quchang Ouyang:** writing – original draft, writing – review and editing, conceptualization, supervision.

Acknowledgments

During the preparation of this manuscript, artificial intelligence tools were used in a limited manner. Specifically, ChatGPT (version 4.5) was employed to assist with the drafting of preliminary text for selected sections, primarily in the context of testing and comparing AI-assisted writing outputs. All AI-assisted content was subsequently reviewed, revised, and finalized by the authors. No AI tools were used for data analysis, result interpretation, or the formulation of scientific conclusions. The authors take full responsibility for the content of this manuscript.

Consent

The authors have nothing to report.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

References

1. E. J. Topol, “High-Performance Medicine: The Convergence of Human and Artificial Intelligence,” *Nature Medicine* 25, no. 1 (2019): 44–56, <https://doi.org/10.1038/s41591-018-0300-7>.
2. X. Wu, W. Li, and H. Tu, “Big Data and Artificial Intelligence in Cancer Research,” *Trends in Cancer* 10, no. 2 (2024): 147–160, <https://doi.org/10.1016/j.trecan.2023.10.006>.
3. P. Jiang, S. Sinha, K. Aldape, S. Hannenhalli, C. Sahinalp, and E. Ruppin, “Big Data in Basic and Translational Cancer Research,” *Nature Reviews Cancer* 22, no. 11 (2022): 625–639, <https://doi.org/10.1038/s41568-022-00502-0>.
4. O. Elemento, C. Leslie, J. Lundin, and G. Tourassi, “Artificial Intelligence in Cancer Research, Diagnosis and Therapy,” *Nature Reviews Cancer* 21, no. 12 (2021): 747–752, <https://doi.org/10.1038/s41568-021-00399-1>.

5. H. Sung, J. Ferlay, R. L. Siegel, et al., “Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries,” *CA: A Cancer Journal for Clinicians* 71, no. 3 (2021): 209–249, <https://doi.org/10.3322/caac.21660>.
6. K. Swanson, E. Wu, A. Zhang, A. A. Alizadeh, and J. Zou, “From Patterns to Patients: Advances in Clinical Machine Learning for Cancer Diagnosis, Prognosis, and Treatment,” *Cell* 186, no. 8 (2023): 1772–1791, <https://doi.org/10.1016/j.cell.2023.01.035>.
7. K. Feng, Z. Yi, and B. Xu, “Artificial Intelligence and Breast Cancer Management: From Data to the Clinic,” *Cancer Innovation* 4, no. 2 (2025): e159, <https://doi.org/10.1002/cai2.159>.
8. P. Rajpurkar and M. P. Lungren, “The Current and Future State of AI Interpretation of Medical Images,” *New England Journal of Medicine* 388, no. 21 (2023): 1981–1990, <https://doi.org/10.1056/NEJMra2301725>.
9. G. S. Collins, M. Chester-Jones, S. Gerry, et al., “Clinical Prediction Models Using Machine Learning in Oncology: Challenges and Recommendations,” *BMJ Oncology* 4, no. 1 (2025): e000914, <https://doi.org/10.1136/bmjonc-2025-000914>.
10. A. Hosny, C. Parmar, J. Quackenbush, L. H. Schwartz, and H. J. W. L. Aerts, “Artificial Intelligence in Radiology,” *Nature Reviews Cancer* 18, no. 8 (2018): 500–510, <https://doi.org/10.1038/s41568-018-0016-5>.
11. M. Sinkala, N. Mulder, and D. Martin, “Machine Learning and Network Analyses Reveal Disease Subtypes of Pancreatic Cancer and Their Molecular Characteristics,” *Scientific Reports* 10, no. 1 (2020): 1212, <https://doi.org/10.1038/s41598-020-58290-2>.
12. H. H. Tseng, Y. Luo, S. Cui, J. T. Chien, R. K. Ten Haken, and I. E. Naqa, “Deep Reinforcement Learning for Automated Radiation Adaptation in Lung Cancer,” *Medical Physics* 44, no. 12 (2017): 6690–6705, <https://doi.org/10.1002/mp.12625>.
13. I. H. Sarker, “Machine Learning: Algorithms, Real-World Applications and Research Directions,” *SN Computer Science* 2, no. 3 (2021): 160, <https://doi.org/10.1007/s42979-021-00592-x>.
14. D. Ardila, A. P. Kiraly, S. Bharadwaj, et al., “End-To-End Lung Cancer Screening With Three-Dimensional Deep Learning on Low-Dose Chest Computed Tomography,” *Nature Medicine* 25, no. 6 (2019): 954–961, <https://doi.org/10.1038/s41591-019-0447-x>.
15. M. A. Thanoon, M. A. Zulkifley, M. Mohd Zainuri, and S. R. Abdani, “A Review of Deep Learning Techniques for Lung Cancer Screening and Diagnosis Based on CT Images,” *Diagnostics (Basel, Switzerland)* 13, no. 16 (2023): 2617, <https://doi.org/10.3390/diagnostics13162617>.
16. S. Benzekry, “Artificial Intelligence and Mechanistic Modeling for Clinical Decision Making in Oncology,” *Clinical Pharmacology & Therapeutics* 108, no. 3 (2020): 471–486, <https://doi.org/10.1002/cpt.1951>.
17. T. Kooi, G. Litjens, B. van Ginneken, et al., “Large Scale Deep Learning for Computer Aided Detection of Mammographic Lesions,” *Medical Image Analysis* 35 (2017): 303–312, <https://doi.org/10.1016/j.media.2016.07.007>.
18. L. Shen, L. R. Margolies, J. H. Rothstein, E. Fluder, R. McBride, and W. Sieh, “Deep Learning to Improve Breast Cancer Detection on Screening Mammography,” *Scientific Reports* 9, no. 1 (2019): 12495, <https://doi.org/10.1038/s41598-019-48995-4>.
19. S. Sohrabei, H. Moghaddasi, A. Hosseini, and S. J. Ehsanzadeh, “Investigating the Effects of Artificial Intelligence on the Personalization of Breast Cancer Management: A Systematic Study,” *BMC Cancer* 24, no. 1 (2024): 852, <https://doi.org/10.1186/s12885-024-12575-1>.
20. K. Danishuddin, S. Khan, and J. J. Kim, “From Cancer Big Data to Treatment: Artificial Intelligence in Cancer Research,” *Journal of Gene Medicine* 26, no. 1 (2024): e3629, <https://doi.org/10.1002/jgm.3629>.

21. H. Li, Z. Han, H. Wu, et al., "Artificial Intelligence in Surgery: Evolution, Trends, and Future Directions," *International Journal of Surgery* 111, no. 2 (2025): 2101–2111, <https://doi.org/10.1097/JS9.0000000000002159>.
22. G. Litjens, T. Kooi, B. E. Bejnordi, et al., "A Survey on Deep Learning in Medical Image Analysis," *Medical Image Analysis* 42 (2017): 60–88, <https://doi.org/10.1016/j.media.2017.07.005>.
23. B. Ehteshami Bejnordi, M. Veta, P. Johannes van Diest, et al., "Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer," *Journal of the American Medical Association* 318, no. 22 (2017): 2199–2210, <https://doi.org/10.1001/jama.2017.14585>.
24. N. Coudray, P. S. Ocampo, T. Sakellaropoulos, et al., "Classification and Mutation Prediction From Non-Small Cell Lung Cancer Histopathology Images Using Deep Learning," *Nature Medicine* 24, no. 10 (2018): 1559–1567, <https://doi.org/10.1038/s41591-018-0177-5>.
25. D. Capper, D. T. W. Jones, M. Sill, et al., "DNA Methylation-Based Classification of Central Nervous System Tumours," *Nature* 555, no. 7697 (2018): 469–474, <https://doi.org/10.1038/nature26000>.
26. X. Meng, X. Yan, K. Zhang, et al., "The Application of Large Language Models in Medicine: A Scoping Review," *iScience* 27, no. 5 (2024): 109713, <https://doi.org/10.1016/j.isci.2024.109713>.
27. S. Zhang, Q. Chu, Y. Li, et al., "Evaluation of Large Language Models Under Different Training Background in Chinese Medical Examination: A Comparative Study," *Frontiers in Artificial Intelligence* 7 (2024): 1442975, <https://doi.org/10.3389/frai.2024.1442975>.
28. Y. Kang, Y. Chang, S. Wu, et al. "ZhongJingGPT: An Expert Knowledge-Guided Language Model for Traditional Chinese Medicine," *Tsinghua Science and Technology* (2025), <https://doi.org/10.26599/TST.2025.9010046>.
29. "National Cancer Institute. Surveillance, Epidemiology, and End Results (SEER) Program," U.S. Department of Health and Human Services, National Institutes of Health, accessed January 11, 2026, <https://seer.cancer.gov/>.
30. L. A. Aaltonen, F. Abascal, A. Abeshouse, et al., "Pan-Cancer Analysis of Whole Genomes," *Nature* 578, no. 7793 (2020): 82–93, <https://doi.org/10.1038/s41586-020-1969-6>.
31. T. Barrett, S. E. Wilhite, P. Ledoux, et al., "NCBI GEO: Archive for Functional Genomics Data Sets: Update," *Nucleic Acids Research* 41, (2013): 991–995, <https://doi.org/10.1093/nar/gks1193>.
32. S. A. Forbes, D. Beare, H. Boutselakis, et al., "COSMIC: Somatic Cancer Genetics at High-Resolution," *Nucleic Acids Research* 45, no. D1 (2017): D777–D783, <https://doi.org/10.1093/nar/gkw1121>.
33. H. Zeng, Y. Liu, L. Wang, et al., "National Cancer Data Linkage Platform of China: Design, Methods, and Application," *China CDC Weekly* 4, no. 13 (2022): 271–275, <https://doi.org/10.46234/ccdcw2022.068>.
34. D. Chakravarty, J. Gao, S. M. Phillips, et al. "OncoKB: A Precision Oncology Knowledge Base," *JCO Precision Oncology* 2017 (2017), <https://doi.org/10.1200/PO.17.00011>.
35. R. L. Grossman, A. P. Heath, V. Ferretti, et al., "Toward a Shared Vision for Cancer Genomic Data," *New England Journal of Medicine* 375, no. 12 (2016): 1109–1112, <https://doi.org/10.1056/NEJMp1607591>.
36. N. Kolesnikov, E. Hastings, M. Keays, et al., "Arrayexpress Update: Simplifying Data Submissions," *Nucleic Acids Research* 43, Database issue (2015): D1113–D1116, <https://doi.org/10.1093/nar/gku1057>.
37. B. E. Bierer, R. Li, M. Barnes, and I. Sim, "A Global, Neutral Platform for Sharing Trial Data," *New England Journal of Medicine* 374, no. 25 (2016): 2411–2413, <https://doi.org/10.1056/NEJMp1605348>.
38. Global Burden of Disease 2019 Cancer Collaboration, J. M. Kocarnik, K. Compton, F. E. Dean, et al., "Cancer Incidence, Mortality, Years of Life Lost, Years Lived With Disability, and Disability-Adjusted Life Years for 29 Cancer Groups From 2010 to 2019: A Systematic Analysis for the Global Burden of Disease Study 2019," *JAMA Oncology* 8, no. 3 (2022): 420–444, <https://doi.org/10.1001/jamaoncol.2021.6987>.
39. E. Cerami, J. Gao, U. Dogrusoz, et al., "The CBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data," *Cancer Discovery* 2, no. 5 (2012): 401–404, <https://doi.org/10.1158/2159-8290.CD-12-0095>.
40. International Cancer Genome Consortium, T. J. Hudson, W. Anderson, et al., "International Network of Cancer Genome Projects," *Nature* 464, no. 7291 (2010): 993–998, <https://doi.org/10.1038/nature08987>.
41. T. J. Littlejohns, J. Holliday, L. M. Gibson, et al., "The UK Biobank Imaging Enhancement of 100, 000 Participants: Rationale, Data Collection, Management and Future Directions," *Nature Communications* 11, no. 1 (2020): 2624, <https://doi.org/10.1038/s41467-020-15948-9>.
42. Z. Chen, J. Chen, R. Collins, et al., "China Kadoorie Biobank of 0.5 Million People: Survey Methods, Baseline Characteristics and Long-Term Follow-Up," *International Journal of Epidemiology* 40, no. 6 (2011): 1652–1666, <https://doi.org/10.1093/ije/dyr120>.
43. J. Deng, W. Dong, R. Socher, L.-J. Li, L. Kai, and F.-F. Li, "Imagenet: A Large-Scale Hierarchical Image Database," *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, USA (2009): 248–255, <https://doi.org/10.1109/CVPR.2009.5206848>.
44. M. Griffith, N. C. Spies, K. Krysiak, et al., "CIVIC Is a Community Knowledgebase for Expert Crowdsourcing the Clinical Interpretation of Variants in Cancer," *Nature Genetics* 49, no. 2 (2017): 170–174, <https://doi.org/10.1038/ng.3774>.
45. J. Barretina, G. Caponigro, N. Stransky, et al., "The Cancer Cell Line Encyclopedia Enables Predictive Modelling of Anticancer Drug Sensitivity," *Nature* 483, no. 7391 (2012): 603–607, <https://doi.org/10.1038/nature11003>.
46. I. Lappalainen, J. Almeida-King, V. Kumanduri, et al., "The European Genome-Phenome Archive of Human Data Consented for Biomedical Research," *Nature Genetics* 47, no. 7 (2015): 692–695, <https://doi.org/10.1038/ng.3312>.
47. A. Tsherniak, F. Vazquez, P. G. Montgomery, et al., "Defining a Cancer Dependency Map," *Cell* 170, no. 3 (2017): 564–576.e16, <https://doi.org/10.1016/j.cell.2017.06.010>.
48. M. Scimeca, N. Urbano, N. Toschi, E. Bonanno, and O. Schillaci, "Precision Medicine in Breast Cancer: From Biological Imaging to Artificial Intelligence," *Seminars in Cancer Biology* 72 (2021): 1–3, <https://doi.org/10.1016/j.semcancer.2021.04.019>.
49. C. Hutter and J. C. Zenklusen, "The Cancer Genome Atlas: Creating Lasting Value Beyond Its Data," *Cell* 173, no. 2 (2018): 283–285, <https://doi.org/10.1016/j.cell.2018.03.042>.
50. P. S. R. C. Murty, C. Anuradha, P. A. Naidu, et al., "Integrative Hybrid Deep Learning for Enhanced Breast Cancer Diagnosis: Leveraging the Wisconsin Breast Cancer Database and the CBIS-DDSM Dataset," *Scientific Reports* 14, no. 1 (2024): 26287, <https://doi.org/10.1038/s41598-024-74305-8>.
51. A. Rodriguez-Ruiz, K. Lång, A. Gubern-Merida, et al., "Stand-Alone Artificial Intelligence for Breast Cancer Detection in Mammography: Comparison With 101 Radiologists," *JNCI: Journal of the National Cancer Institute* 111, no. 9 (2019): 916–922, <https://doi.org/10.1093/jnci/djy222>.
52. K. J. Geras, R. M. Mann, and L. Moy, "Artificial Intelligence for Mammography and Digital Breast Tomosynthesis: Current Concepts and Future Perspectives," *Radiology* 293, no. 2 (2019): 246–259, <https://doi.org/10.1148/radiol.2019182627>.
53. F. Ayatollahi, S. B. Shokouhi, R. M. Mann, and J. Teuwen, "Automatic Breast Lesion Detection in Ultrafast DCE-MRI Using Deep Learning," *Medical Physics* 48, no. 10 (2021): 5897–5907, <https://doi.org/10.1002/mp.15156>.

54. K. Zhang, R. Zhou, E. Adhikarla, et al., “A Generalist Vision-Language Foundation Model for Diverse Biomedical Tasks,” *Nature Medicine* 30, no. 11 (2024): 3129–3141, <https://doi.org/10.1038/s41591-024-03185-2>.
55. Y.-Z. Jiang, D. Ma, X. Jin, et al., “Integrated Multiomic Profiling of Breast Cancer in the Chinese Population Reveals Patient Stratification and Therapeutic Vulnerabilities,” *Nature Cancer* 5, no. 4 (2024): 673–690, <https://doi.org/10.1038/s43018-024-00725-0>.
56. S.-J. Sammut, M. Crispin-Ortuzar, S.-F. Chin, et al., “Multi-Omic Machine Learning Predictor of Breast Cancer Therapy Response,” *Nature* 601, no. 7894 (2022): 623–629, <https://doi.org/10.1038/s41586-021-04278-5>.
57. D. van de Sande, M. Sharabiani, H. Bluemink, et al., “Artificial Intelligence Based Treatment Planning of Radiotherapy for Locally Advanced Breast Cancer,” *Physics and Imaging in Radiation Oncology* 20 (2021): 111–116, <https://doi.org/10.1016/j.phro.2021.11.007>.
58. X. Li, J. Ma, L. Leng, et al., “MoGCN: A Multi-Omics Integration Method Based on Graph Convolutional Network for Cancer Subtype Analysis,” *Frontiers in Genetics* 13 (2022): 806842, <https://doi.org/10.3389/fgene.2022.806842>.
59. A. Blasiak, A. T. L. Truong, N. Foo, et al., “Personalized Dose Selection Platform for Patients With Solid Tumors in the PRECISE CURATE.AI Feasibility Trial,” *NPJ Precision Oncology* 9, no. 1 (2025): 49, <https://doi.org/10.1038/s41698-025-00835-7>.
60. A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting, “Large Language Models in Medicine,” *Nature Medicine* 29, no. 8 (2023): 1930–1940, <https://doi.org/10.1038/s41591-023-02448-8>.
61. J.-E. Bibault, B. Chaix, A. Guillemassé, et al., “A Chatbot Versus Physicians to Provide Information for Patients With Breast Cancer: Blind, Randomized Controlled Noninferiority Trial,” *Journal of Medical Internet Research* 21, no. 11 (2019): e15787, <https://doi.org/10.2196/15787>.
62. T. Xiao, S. Kong, Z. Zhang, D. Hua, and F. Liu, “A Review of Big Data Technology and Its Application in Cancer Care,” *Computers in Biology and Medicine* 176 (2024): 108577, <https://doi.org/10.1016/j.compbmed.2024.108577>.
63. C.-T. Wu, S.-M. Wang, Y.-E. Su, et al., “A Precision Health Service for Chronic Diseases: Development and Cohort Study Using Wearable Device, Machine Learning, and Deep Learning,” *IEEE Journal of Translational Engineering in Health and Medicine* 10 (2022): 1–14, <https://doi.org/10.1109/JTEHM.2022.3207825>.
64. E. Eisenstein, C. Kopacek, S. S. Cavalcante, A. C. Neves, G. P. Fraga, and L. A. Messina, “Telemedicine: A Bridge Over Knowledge Gaps in Healthcare,” *Current Pediatrics Reports* 8, no. 3 (2020): 93–98, <https://doi.org/10.1007/s40124-020-00221-w>.
65. V. Sorin, D. Brin, Y. Barash, et al., “Large Language Models and Empathy: Systematic Review,” *Journal of Medical Internet Research* 26 (2024): e52597, <https://doi.org/10.2196/52597>.
66. K. K. Fitzpatrick, A. Darcy, and M. Vierhile, “Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial,” *JMIR Mental Health* 4, no. 2 (2017): e19, <https://doi.org/10.2196/mental.7785>.
67. B. D. Lund, D. Khan, and M. Yuvaraj, “ChatGPT in Medical Libraries, Possibilities and Future Directions: An Integrative Review,” *Health Information & Libraries Journal* 41, no. 1 (2024): 4–15, <https://doi.org/10.1111/hir.12518>.
68. S. Z. Wu, D. L. Roden, C. Wang, et al., “Stromal Cell Diversity Associated With Immune Evasion in Human Triple-Negative Breast Cancer,” *EMBO Journal* 39, no. 19 (2020): e104063, <https://doi.org/10.15252/embj.2019104063>.
69. K. Zhou, B. S. Kottoori, S. A. Munj, Z. Zhang, S. Draghici, and S. Arslanturk, “Integration of Multimodal Data From Disparate Sources for Identifying Disease Subtypes,” *Biology* 11, no. 3 (2022): 360, <https://doi.org/10.3390/biology11030360>.
70. H. Atas Guvenilir and T. Doğan, “How to Approach Machine Learning-Based Prediction of Drug/Compound-Target Interactions,” *Journal of Cheminformatics* 15, no. 1 (2023): 16, <https://doi.org/10.1186/s13321-023-00689-w>.
71. A. O. Basile, A. Yahi, and N. P. Tatonetti, “Artificial Intelligence for Drug Toxicity and Safety,” *Trends In Pharmacological Sciences* 40, no. 9 (2019): 624–635, <https://doi.org/10.1016/j.tips.2019.07.005>.
72. Y. Shi, “Drug Development in the AI Era: AlphaFold 3 is coming!,” *Innovation (Camb)* 5, no. 5 (2024): 100685, <https://doi.org/10.1016/j.xinn.2024.100685>.
73. Z. Tanoli, M. Vähä-Koskela, and T. Aittokallio, “Artificial Intelligence, Machine Learning, and Drug Repurposing in Cancer,” *Expert Opinion on Drug Discovery* 16, no. 9 (2021): 977–989, <https://doi.org/10.1080/17460441.2021.1883585>.
74. J. Choi, S. Park, and J. Ahn, “RefDNN: A Reference Drug Based Neural Network for More Accurate Prediction of Anticancer Drug Resistance,” *Scientific Reports* 10, no. 1 (2020): 1861, <https://doi.org/10.1038/s41598-020-58821-x>.
75. J. Gao, B. A. Aksoy, U. Dogrusoz, et al., “Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the CBioPortal,” *Science Signaling* 6, no. 269 (2013): pl1, <https://doi.org/10.1126/scisignal.2004088>.
76. Data Security Law of the People’s Republic of China (2021), adopted at the 29th Session of the Standing Committee of the 13th National People’s Congress on June 10, 2021, effective September 1, 2021.
77. Y. Feng, “The Future of China’s Personal Data Protection Law: Challenges and Prospects,” *Asia Pacific Law Review* 27, no. 1 (2019): 62–82, <https://doi.org/10.1080/10192557.2019.1646015>.
78. A. J. Butte, “Artificial Intelligence-From Starting Pilots to Scalable Privilege,” *JAMA Oncology* 9, no. 10 (2023): 1341–1342, <https://doi.org/10.1001/jamaoncol.2023.2867>.
79. Z.-Y. Hu, F. Han, L. Yu, Y. Jiang, and G. Cai, “AI-Link Omnipotent Pathological Robot: Bridging Medical Meta-Universe to Real-World Diagnosis and Therapy,” *Innovation* 4, no. 5 (2023): 100494, <https://doi.org/10.1016/j.xinn.2023.100494>.
80. K. Victor Mugabe, “Barriers and Facilitators to the Adoption of Artificial Intelligence in Radiation Oncology: A New Zealand Study,” *Technical Innovations & Patient Support in Radiation Oncology* 18 (2021): 16–21, <https://doi.org/10.1016/j.tipsro.2021.03.004>.
81. I. S. Chua, M. Gaziel-Yablowitz, Z. T. Korach, et al., “Artificial Intelligence in Oncology: Path to Implementation,” *Cancer Medicine* 10, no. 12 (2021): 4138–4149, <https://doi.org/10.1002/cam4.3935>.
82. K. Juluru, H. H. Shih, K. N. Keshava Murthy, et al., “Integrating AI Algorithms Into the Clinical Workflow,” *Radiology: Artificial Intelligence* 3, no. 6 (2021): e210013, <https://doi.org/10.1148/ryai.2021210013>.
83. E. Harris, “Large Language Models Answer Medical Questions Accurately, but Can’t Match Clinicians’ Knowledge,” *Journal of the American Medical Association* 330, no. 9 (2023): 792–794, <https://doi.org/10.1001/jama.2023.14311>.
84. J. Liu, J. Zheng, X. Cai, D. Wu, and C. Yin, “A Descriptive Study Based on the Comparison of ChatGPT and Evidence-Based Neurosurgeons,” *iScience* 26, no. 9 (2023): 107590, <https://doi.org/10.1016/j.isci.2023.107590>.
85. N. Rieke, J. Hancox, W. Li, et al., “The Future of Digital Health With Federated Learning,” *NPJ Digital Medicine* 3 (2020): 119, <https://doi.org/10.1038/s41746-020-00323-1>.
86. F. Fotouhi, A. Balu, Z. Jiang, Y. Esfandiari, S. Jahani, and S. Sarkar, “Dominating Set Model Aggregation for Communication-Efficient

Decentralized Deep Learning,” *Neural Networks* 171 (2024): 25–39, <https://doi.org/10.1016/j.neunet.2023.11.057>.

87. A. Vizitiu, C. I. Niță, A. Puiu, C. Suciuc, and L. M. Itu, “Applying Deep Neural Networks Over Homomorphic Encrypted Medical Data,” *Computational and Mathematical Methods in Medicine* 2020 (2020): 3910250, <https://doi.org/10.1155/2020/3910250>.

88. A. N. Desai, “Artificial Intelligence: Promise, Pitfalls, and Perspective,” *Journal of the American Medical Association* 323, no. 24 (2020): 2448, <https://doi.org/10.1001/jama.2020.8737>.

89. Y. Peng, Q. Chen, and G. Shih, “Deepseek Is Open-Access and the Next Ai Disrupter for Radiology,” *Radiology Advances* 2, no. 1 (2025): umaf009, <https://doi.org/10.1093/radadv/umaf009>.

90. Z. Ma, M. Zhang, J. Liu, et al., “An Assisted Diagnosis Model for Cancer Patients Based on Federated Learning,” *Frontiers in Oncology* 12 (2022): 860532, <https://doi.org/10.3389/fonc.2022.860532>.

91. F. Tempel, D. Groos, E. A. F. Ihlen, L. Adde, and I. Strümke, “Choose Your Explanation: A Comparison of Shap and Grad-CAM in Human Activity Recognition,” *Applied Intelligence* 54 (2024): e241216003. <https://doi.org/10.1007/s10489-024-05789-3>.

92. H. Jung and Y. Oh, “Towards Better Explanations of Class Activation Mapping,” *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada (2021): 1316–1324, <https://doi.org/10.1109/ICCV48922.2021.00137>.

93. D.-M. Koh, N. Papanikolaou, U. Bick, et al., “Artificial Intelligence and Machine Learning in Cancer Imaging,” *Communications Medicine* 2 (2022): 133, <https://doi.org/10.1038/s43856-022-00199-0>.

94. F. M. Talaat, S. A. Gamel, R. M. El-Balka, M. Shehata, and H. ZainEldin, “Grad-CAM Enabled Breast Cancer Classification With a 3D Inception-ResNet V2: Empowering Radiologists With Explainable Insights,” *Cancers* 16, no. 21 (2024): 3668, <https://doi.org/10.3390/cancers16213668>.

95. D. Froelicher, J. R. Troncoso-Pastoriza, J. L. Raisaro, et al., “Truly Privacy-Preserving Federated Analytics for Precision Medicine With Multiparty Homomorphic Encryption,” *Nature Communications* 12, no. 1 (2021): 5910, <https://doi.org/10.1038/s41467-021-25972-y>.

96. J. A. Omiye, J. C. Lester, S. Spichak, V. Rotemberg, and R. Daneshjou, “Large Language Models Propagate Race-Based Medicine,” *NPJ Digital Medicine* 6, no. 1 (2023): 195, <https://doi.org/10.1038/s41746-023-00939-z>.

97. L. Kolla and R. B. Parikh, “Uses and Limitations of Artificial Intelligence for Oncology,” *Cancer* 130, no. 12 (2024): 2101–2107, <https://doi.org/10.1002/cncr.35307>.

98. H. M. C. Cheung and D. Rubin, “Challenges and Opportunities for Artificial Intelligence in Oncological Imaging,” *Clinical Radiology* 76, no. 10 (2021): 728–736, <https://doi.org/10.1016/j.crad.2021.03.009>.

99. C. Luchini, A. Pea, and A. Scarpa, “Artificial Intelligence in Oncology: Current Applications and Future Perspectives,” *British Journal of Cancer* 126, no. 1 (2022): 4–9, <https://doi.org/10.1038/s41416-021-01633-1>.

100. A. Rancea, I. Anghel, and T. Cioara, “Edge Computing in Healthcare: Innovations, Opportunities, and Challenges,” *Future Internet* 16, no. 9 (2024): 329, <https://doi.org/10.3390/fi16090329>.