General Psychiatry

Thorny but rosy: prosperities and difficulties in 'AI plus medicine' concerning data collection, model construction and clinical deployment

Yujia Xia ,[1,2] Zhangsheng Yu [1,2,3]

To cite: Xia Y, Yu Z. Thorny but rosy: prosperities and difficulties in 'AI plus medicine' concerning data collection, model construction and clinical deployment. *General Psychiatry* 2024;**37**:e101436. doi:10.1136/gpsych-2023-101436

► Additional supplemental material is published online only. To view, please visit the journal online (https://doi.org/10.1136/gpsych-2023-101436).

Received 26 December 2023
Accepted 11 November 2024

[1]Department of Bioinformatics and Biostatistics, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, China
[2]SJTU-Yale Joint Center for Biostatistics and Data Science, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, China
[3]Clinical Research Institute, Shanghai Jiao Tong University School of Medicine, Shanghai, China

Correspondence to
Professor Zhangsheng Yu; yuzhangsheng@sjtu.edu.cn

INTRODUCTION

Artificial intelligence (AI) has brought about revolutionary changes in the medical field, including clinical practice, basic research and health monitoring. The powerful computing capabilities and intelligent algorithms enable it to process and analyse large-scale medical data, thereby assisting clinicians and researchers in better understanding and addressing complex medical problems. In recent years, significant progress has been made in AI-aided medical image interpretation, clinical decision support and personalised medicine. Furthermore, the development of large language models has further expanded the prospective applications of AI models. However, there are still several issues that need to be addressed in integrating AI into clinical practices. We have reviewed the existing achievements (figure 1) and possible challenges (figure 2) in the three stages: data preparation, modelling and prediction as well as model deployment and application. Finally, we discussed the differences between mental health and other domains as well as the special considerations in 'AI plus psychiatry'.

A TIP OF THE ICEBERG: EFFECTIVELY USE THE MASSIVE MEDICAL DATA

It was estimated that a hospital produces roughly 50 petabytes of electronic health data daily on average,[1] including electronic health records (EHRs), radiological imaging, genomic sequencing, pathological images and other related information.

Traditional approaches often face challenges in fully utilising and analysing the vast and complex healthcare data, while the development of AI has injected new vitality into the paradigm of 'data-driven discovery'. For instance, in psychiatric disorders, natural language processing (NLP) techniques can analyse patients' speech, text and social media data to assist in the early diagnosis and monitoring of mental illnesses.[2] In neurodegenerative diseases, AI models can automatically detect and locate brain abnormalities and monitor disease progression through brain imaging.[3] In oncology, deep learning algorithms can use genomic data to predict tumour susceptibility, personalised treatments and drug responses.[4] With the development of large language models, AI techniques have further enhanced their capability to integrate and analyse complex data more generally.

Despite the abundance of medical data, its utilisation remains a mere tip of the iceberg. There are three key reasons for the insufficient usage: data quality, data security and privacy as well as data sharing and collaboration.

The quality of data determines the model performance. AI functions as a processing tool rather than a creator that can beautify the data. If the data quality is poor, the model will ultimately face the dilemma of 'garbage in, garbage out'. An extreme example is that any model would lose its efficacy when faced with totally randomly generated data.

Nevertheless, improving medical data quality is not an easy task. The sheer volume of medical data and various data acquisition approaches make it impossible to achieve perfection. Some algorithm-based approaches are proposed to efficiently process the imperfections in data. Recently, deep learning methods have shown their advantages in detecting and removing artefacts from radiological[5 6] and pathological imaging.[7] The trained convolutional neural network (CNN), when applied to medical

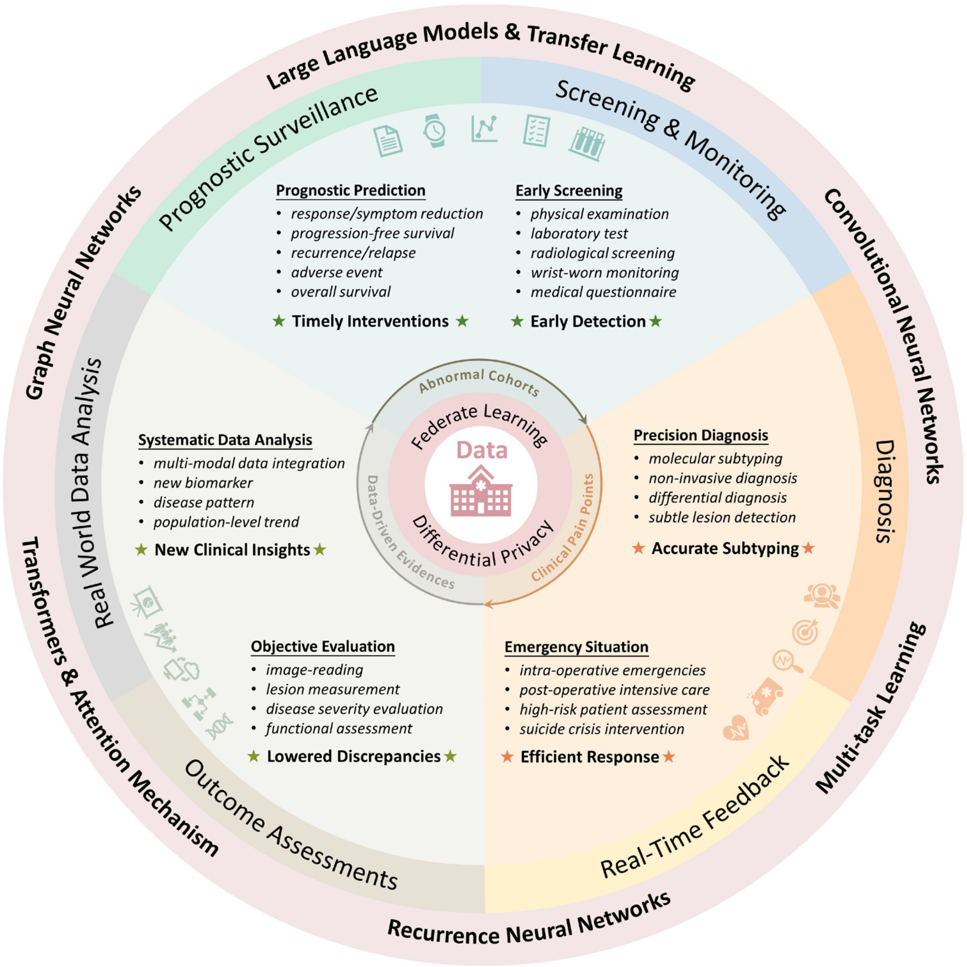
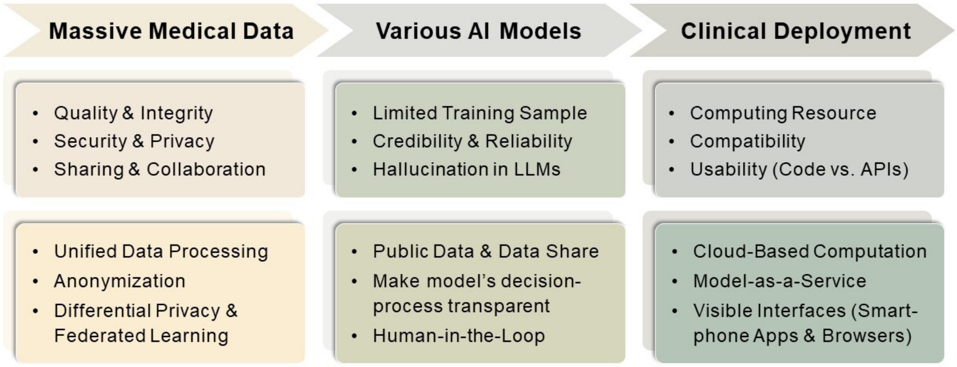


Figure 1 Prosperities and various applications of artificial intelligence in the medical domain.

imaging data, can significantly reduce image artefacts and improve the visualisation of critical anatomical structures. Apart from imaging quality, accurate labelling is another crucial part. It is reported that the proportion of incorrectly labelled data ranges from 8% to 38% across several real-world datasets.[8] Therefore, methods for label quality control and cleaning should also be given attention. For instance, researchers have proposed Annotation Quality Assessment (AQuA),[9] a benchmarking tool for label quality assessment, which can be easily integrated into labelling workflows, enabling flexible, versatile and comprehensive data annotation quality evaluation. Besides, 'active label cleaning',[10] which ranks instances according to estimated label correctness and labelling difficulty of each sample, has been reported to aid with model training and enhance performance. Conclusively, efforts should be made to promote a virtuous cycle between AI model development and data management (ie, high-quality data aid the construction of reliable AI models, and AI models facilitate better data management).



Massive Medical Data	Various AI Models	Clinical Deployment
Quality & Integrity Security & Privacy Sharing & Collaboration	Limited Training Sample Credibility & Reliability Hallucination in LLMs	Computing Resource Compatibility Usability (Code vs. APIs)
Unified Data Processing Anonymization Differential Privacy & Federated Learning	Public Data & Data Share Make model's decision-process transparent Human-in-the-Loop	Cloud-Based Computation Model-as-a-Service Visible Interfaces (Smart-phone Apps & Browsers)

Figure 2 Existing difficulties in 'AI plus medicine' and possible solutions. API, application programming interface. LLM, large language model.

In addition to high-quality data, protecting data privacy and ensuring compliance are crucial, especially in healthcare. The Governance Model for AI in Healthcare (GMAIH)[11] provides principles for fairness, transparency, trustworthiness and accountability in model use and deployment. Researchers should standardise data collection and model development, including anonymisation and minimisation, and establish a robust management process based on GMAIH or similar guidelines. Third-party audits are also necessary to regulate AI system compliance and security.

Concerns for data privacy have led to a dilemma known as 'data silos',[12] that is, the isolation of data among different medical institutions. Anonymisation can conceal identities but is limited to personally identifiable information. Differential privacy, by adding noise to data queries, balances deidentification and utility, enabling meaningful analysis while protecting sensitive information. For instance, researchers have proposed a differential privacy deep learning framework called '*deepee*',[13] which can be integrated with the PyTorch deep learning framework. The framework maintains the excellent performance of the AI model for pneumonia classification and liver tumour segmentation while ensuring strict privacy protection.

Moreover, federated learning (FL) makes cross-centre data communication possible. In contrast to traditional approaches that require centralised data storage for model training, FL ensures the data are stored locally, but the model parameters are shared among centres. For example, a study[14] trained an FL model using data from 20 global medical institutions to forecast the future oxygen demand of patients with COVID-19 using vital signs, laboratory data and chest X-rays. Compared with single-site models, the FL model achieved a 16% higher average area under the curve and 38% better generalisability across sites.

BEGINNING AI TRAINEE: BUILD TRUSTWORTHY MODELS FOR SURVEILLANCE, DIAGNOSIS AND PROGNOSIS

Owing to the extensive medical data, deep learning has flourished in the field of healthcare. A series of diverse backbone networks (ie, universal architecture) have emerged, each of which excels in handling specific types of medical data. CNN is a widely used automatic feature extractor that can be applied to different modalities, such as X-ray, computed tomography (CT), magnetic resonance imaging (MRI) and ultrasound. Transformers have demonstrated remarkable efficacy in multiple tasks, such as EHR processing, healthcare question-answering (QA), image classification and object detection. Recurrent Neural Network (RNN) is designed for sequential data, such as time series data (eg, electrocardiograms, electroencephalograms and longitudinal monitoring data), multi-phased data (eg, contrasted MRI and CT series), as well as video medical imaging (eg, endoscopic videos). Graph Neural Network (GNN) is a powerful tool for learning complex relationships and interactions between nodes in a graph structure, such as medical knowledge graphs, pathology images and gene interaction networks.

Several works have provided strong support for the powerful capabilities of AI in medical prediction problems. An AI model for pancreatic cancer early screening via non-contrast CT reached a sensitivity of 92.9% and a specificity of 99.9% through over ten thousand training and validation samples.[15] Besides, AI algorithms specialised for Gleason grading have been validated to demonstrate pathologist-level performance on independent, cross-continental cohorts, where the algorithms achieved agreements of over 0.85 with expert uropathologists.[16] Furthermore, AI-based biomarkers extracted from H&E slides have also demonstrated the ability to stratify stage II and III colorectal cancer patients into distinct prognostic groups within large, independent patient cohorts.[17] These exciting results enthuse researchers to believe that AI-based precise medicine is entering its prime time.

Though quite a few AI models are being developed, most of them have not been applied in clinical practice, that is, they are not qualified enough to be trainee doctors. We consider that this is mainly attributable to two reasons: one is the insufficient number of training samples, which affects the model's accuracy; the other is the lack of transparency and interpretability of the model.

The lack of extensive training data results in instability and unreliability of the AI model, making it difficult to reach the same level as clinicians who have grown and trained through years of dedicated study and practice. Most AI models are trained on small sample sizes (typically only a few hundred, with a small proportion of a few thousand) from retrospective cohorts. Consequently, the models tend to suffer from overfitting to specific datasets, resulting in a lack of robustness and generalisability when applied to external validation or prospective cohorts. We encourage researchers to fully use public datasets in the process of training AI models. Additionally, if appropriate, researchers should consider adopting data-sharing approaches such as FL, which allows them to integrate sample data from multiple institutions to increase the sample size.

While high accuracy is crucial, transparent decision-making processes and interpretability are also vital components. Most AI models are black-box structured, with their decision-making processes being opaque. Heatmaps are commonly used to visualise the attention regions of a deep learning model, helping researchers better understand the model's decision-making process. Moreover, integrating medical prior knowledge can be another approach to improve the reliability of the AI models. This approach can use the rich experience and knowledge accumulated by experts and convert it into a form that AI can learn and apply in order to make up for the limitations of relying solely on data-driven training.

Furthermore, we should carefully integrate AI tools into the medical field, seeking a balanced collaboration between doctors and AI rather than completely relying on AI's decisions. Google has proposed a framework called 'Complementarity-Driven Deferral to Clinical Workflow' (CoDoC),[18] which is an AI system that learns to decide between the prediction results from deep learning models and clinicians. This helps AI systems identify their own limitations, thereby improving clinical reliability. For example, in the breast cancer X-ray identification task, the AI model that integrated CoDoC reduced the false positive rate by 25% while maintaining the same true positive rate. This demonstrates the potential benefits of a complementary approach, where AI and clinicians work together to enhance the accuracy and reliability of medical diagnoses.

PANDORA'S BOX: CAREFULLY EMBRACE THE EMERGING FOUNDATION MODELS

The development of Generative Pre-trained Transformer (GPT) models has been a significant milestone in deep learning, demonstrating the potential of large-scale pretraining and fine-tuning techniques for understanding massive real-world data. Foundation models like GPT have significantly advanced the field of NLP and computer vision and subsequently have rapidly expanded into the medical domain.

In healthcare, compared with 'specialist' models that focus on single tasks, foundation models can integrate multi-modal medical data, including imaging, EHRs and multimedia health consultation data, to be a 'generalist' to assist clinicians in various multitask prediction scenarios, such as diagnosis, treatment assessment and disease surveillance. Through iterative optimisation of trillions of parameters on extensive medical datasets, foundation models can grow from an 'AI-medical trainee' to an 'AI-general doctor'.[19]

These trained AI doctors have achieved remarkable performance in various aspects, including medical exam QA, real-life clinical dialogues, report identification and visual understanding. For instance, Med-Pathways Language Model (Med-PaLM), a general instruction-prompt large language model, yields promising performance on long-form QA scenarios that have been rated a recognition score of 92.6% by a group of clinicians, which is comparable to real-world clinician-generated answers (92.9%).[20] Alongside its precise general medical QA capabilities, Med-PaLM exhibits a remarkable ability to assess psychiatric functioning in various mental disorders. Notably, it achieves an accuracy exceeding 0.8 in predicting depression scores based on standardised assessments, with no significant difference compared with clinical assessors.[21] Besides, RadBERT, a family of bidirectional encoder representations from transformers (BERT)-based language models tailored to radiology, demonstrate appreciable performances on abnormality identification, report coding and content summarisation (with accuracies all above 95%).[22]

Apart from language models such as Med-PaLM and RadBERT, large vision models and multimodal foundation models also have exhibited versatile medical prediction capabilities and superb generalisation ability with no-new (ie, zero-shot learning) or limited (ie, few-shot learning) training sample on a new task. Localize and Segment Anything Model for 3D Medical Image—a two-stage methodology that first performs prompt-based (ie, clinical cues like organ names or specific anatomical structures) position identification and then delineates the boundary—achieves precise automated segmentation with zero-shot learning on 38 distinct organs.[23] Pathology language-image pretraining (PLIP), goes beyond unimodal text or image data and adopts a multimodal approach by combining image and text understanding. PLIP exhibits superb zero-shot learning capability and achieves precise performance in various types of inferences, such as histopathological tissue classification, text-to-image and image-to-image retrieval.[24]

In conclusion, these medical foundation models, with large parameters trained on extensive medical data, can exhibit emergent intelligence to achieve favourable performances compared with small-parameter single-task models. Consequently, they can better provide healthcare consultations and decision-making.

Though prospects are rosy, applying large models into clinical settings remains fraught with challenges. In the healthcare setting, there are more specialised terminologies and a lower tolerance for errors compared with other application domains. Large AI models should serve as objective medical assessors rather than casual storytellers. Therefore, evaluating the reliability becomes crucial to minimise potential security risks to ensure their positive contribution to clinical practice. It is particularly important to address the potential issue of 'hallucination' in large language models, where the models may 'unconsciously' generate fabricated, inconsistent or erroneous information.[25]

The mitigation of hallucinations in healthcare is a multipronged process, where both inner (ie, the model itself) and outer (ie, human intervention) approaches should be adopted. For inner ways, fine-tuning for specific tasks can aid the model in becoming an expert in a particular domain, thereby reducing the generation of fabricated false information. Besides, in situations where the model's output is uncertain, providing prompts such as 'I don't know' can help minimise the occurrence of hallucinations. For outer ways, human-in-the-loop can assist. During the model development process, domain experts can monitor and then make timely corrections to erroneous outputs. This facilitates a feedback loop that allows the model to enhance its performance through reinforcement learning.

While these approaches offer certain avenues to enhance the reliability of large medical models, there is still a long way to go before they are truly applied in practice. Once the Pandora's box is opened, it will be difficult to close it again. Therefore, we should recognise the hope brought by large models, but we must treat them with caution and actively address any potential issues that may arise.

CASTLE IN THE AIR: LAND AI IN CLINICAL PRACTICE

Despite the rapid advancements of AI models in the academic domain, their practical implementation in clinical settings still lags. Apart from the accuracy, robustness and stability, deployment and usability of AI models are crucial considerations when applying AI in hospitals. Deploying AI models necessitates a comprehensive evaluation of factors such as time efficiency, computing resources, storage capacity and compatibility with the existing medical systems. Furthermore, visual interfaces are essential to improve usability, that is, clinicians can easily use the model, making AI a desirable assistant rather than a burden for doctors.

Model-as-a-Service (MaaS) has simplified the process of applying AI models in clinical settings. It adopts a plug-and-play approach, enabling hospitals or clinical institutions to access and use AI models through simple application programming interfaces without manually configuring and maintaining extensive hardware infrastructure. In traditional local deployments, high-performance servers or graphics processing units are required to meet the computational demands of AI models, whereas, as MaaS is cloud-based, all hardware would be well-settled by service providers.

In addition, most AI models are only available in code format, which is hard for clinicians who lack programming expertise to use. Encapsulating AI models into web-based interfaces or client programmes provides a user-friendly approach to ensure clinicians can directly access and use the models through browsers, which could largely enhance usability and operability.

Despite the potential conveniences MaaS provides, several issues remain to be addressed. In clinical applications, the timeliness and security of data transmission are important. Particularly when dealing with radiological or pathological images (which can reach gigabyte-level file sizes), data transmission becomes a time-consuming task. Furthermore, transferring data to the cloud may introduce risks of privacy breaches; thus, appropriate data protection measures from both technique and policy levels should be adopted to safeguard the confidentiality and integrity of the transmitted data.

DISCUSSION

Compared with other fields, AI faces some unique challenges in mental health. Primarily, the diagnosis and treatment of mental disorders have subjectivity and complexity. Unlike cancer or cardiovascular diseases, which can be objectively diagnosed and monitored through laboratory tests or imaging, mental illnesses need a comprehensive assessment of multiple factors, including feelings, subjective symptoms and the external social environment. Therefore, the gold standard for model training typically requires consensus among senior doctors. Then, the trained model can provide an objective evaluation to reduce the impact of human subjectivity.

Furthermore, in contrast to most physical diseases that primarily involve functional impairments or abnormalities, most mental disorders affect an individual's psychological and emotional state. Some patients may exhibit guardedness and reluctance to disclose their psychological issues and medical history to psychiatrists due to self-esteem concerns. AI models can provide anonymity and impartiality without making judgements or introducing bias on patients' privacy or personal information. Serving as a bridge between psychiatrists and patients, AI can foster a circumstance where patients can feel free to express their concerns. Additionally, AI models are impervious to distractions, stress, fatigue and subjective emotions that can influence human therapists. Therefore, AI may possess certain advantages in assisting with patient treatment.

Though AI models can provide some support, they cannot replace psychologists. The core limitation lies in the inherent absence of sympathy and empathy within AI models.[26] Emotional engagements are vital in comprehending the psychological state of patients with mental disorders. In cases of severe mental health, human psychiatrists with emotional understanding are still rigidly needed to form an emotional connection with the patient. Therefore, it is essential to recognise the limitations of AI models and treat them as auxiliary tools rather than substitutes for psychiatrists.

In conclusion, despite the thorns that lie ahead, AI researchers still choose to forge forward. We strongly believe in the vast prospects of AI models in healthcare. AI will serve as an assistant, enhancing the efficiency of healthcare professionals and providing patients with more precise, personalised and efficient medical services.

Contributors ZY and YX contributed to the idea for the article. YX wrote the draft. ZY critically revised the manuscript.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval Not applicable.

Provenance and peer review Commissioned; externally peer-reviewed.

ORCID iDs

Yujia Xia http://orcid.org/0000-0001-6133-2146

Zhangsheng Yu http://orcid.org/0000-0002-8189-5330

REFERENCES

1 Brian E. How to navigate structured and unstructured data as a healthcare organization? Health Tech; 2023.

2 Malgaroli M, Hull TD, Zech JM, *et al*. Natural language processing for mental health interventions: a systematic review and research framework. *Transl Psychiatry* 2023;13:309.
3 Myszczynska MA, Ojamies PN, Lacoste AMB, *et al*. Applications of machine learning to diagnosis and treatment of neurodegenerative diseases. *Nat Rev Neurol* 2020;16:440–56.
4 Tran KA, Kondrashova O, Bradley AP, *et al*. Deep learning in cancer diagnosis, prognosis and treatment selection. *Genome Med* 2021;13:152.
5 Huang X, Wang J, Tang F, *et al*. Metal artifact reduction on cervical CT images by deep residual learning. *Biomed Eng Online* 2018;17:175.
6 Kyathanahally SP, Döring A, Kreis R. Deep learning approaches for detection and removal of ghosting artifacts in MR spectroscopy. *Magn Reson Med* 2018;80:851–63.
7 Shakhawat H, Hossain S, Kabir A, *et al*. Review of artifact detection methods for automated analysis and diagnosis in digital pathology. In: *Artificial intelligence for disease diagnosis and prognosis in smart healthcare*. Taylor & Francis, 2023: 177–202.
8 Whang SE, Roh Y, Song H, *et al*. Data collection and quality challenges in deep learning: a data-centric AI perspective. *VLDB J* 2023;32:791–813.
9 Mononito G, Vedant S, Arjun C, *et al*. AQuA: a benchmarking tool for label quality assessment. 37th Conference on Neural Information Processing Systems (NeurIPS 2023) Track on Datasets and Benchmarks; 2023.
10 Bernhardt M, Castro DC, Tanno R, *et al*. Active label cleaning for improved dataset quality under resource constraints. *Nat Commun* 2022;13:1161.
11 Reddy S, Allan S, Coghlan S, *et al*. A governance model for the application of AI in health care. *J Am Med Inform Assoc* 2019;27:491–7.
12 Rieke N, Hancox J, Li W, *et al*. The future of digital health with federated learning. *NPJ Digit Med* 2020;3:119.
13 Ziller A, Usynin D, Braren R, *et al*. Medical imaging deep learning with differential privacy. *Sci Rep* 2021;11:13524.
14 Dayan I, Roth HR, Zhong A, *et al*. Federated learning for predicting clinical outcomes in patients with COVID-19. *N Med* 2021;27:1735–43.
15 Cao K, Xia Y, Yao J, *et al*. Large-scale pancreatic cancer detection via non-contrast CT and deep learning. *N Med* 2023;29:3033–43.
16 Bulten W, Kartasalo K, Chen P-HC, *et al*. Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge. *Nat Med* 2022;28:154–63.
17 Skrede O-J, De Raedt S, Kleppe A, *et al*. Deep learning for prediction of colorectal cancer outcome: a discovery and validation study. *Lancet* 2020;395:350–60.
18 Dvijotham KD, Winkens J, Barsbey M, *et al*. Enhancing the reliability and accuracy of AI-enabled diagnosis via complementarity-driven deferral to clinicians. *N Med* 2023;29:1814–20.
19 Qiu J, Li L, Sun J, *et al*. Large AI models in health informatics: applications, challenges, and the future. *IEEE J Biomed Health Inform* 2023;27:6074–87.
20 Singhal K, Azizi S, Tu T, *et al*. Large language models encode clinical knowledge. *Nature* 2023;620:172–80.
21 Galatzer-Levy IR, McDuff D, Natarajan V, *et al*. The capability of large language models to measure psychiatric functioning. 2023. Available: https://arxiv.org/abs/2308.01834 [Accessed 20 Dec 2023].
22 Yan A, McAuley J, Lu X, *et al*. RadBERT: Adapting Transformer-based Language Models to Radiology. *Radiol Artif Intell* 2022;4:e210258.
23 Lei W, Xu W, Zhang X, *et al*. MedLSAM: localize and segment anything model for 3d ct images. arXiv (Cornell University); 2023. Available: https://arxiv.org/abs/2306.14752 [Accessed 20 Dec 2023].
24 Huang Z, Bianchi F, Yuksekgonul M, *et al*. A visual-language foundation model for pathology image analysis using medical Twitter. *N Med* 2023;29:2307–16.
25 van Heerden AC, Pozuelo JR, Kohrt BA. Global mental health services and the impact of artificial intelligence-powered large language models. *JAMA Psychiatry* 2023;80:662–4.
26 Ray A, Bhardwaj A, Malik YK, *et al*. Artificial intelligence and psychiatry: an overview. *Asian J Psychiatr* 2022;70:103021.

Yujia Xia obtained her bachelor's degree in biomedical sciences from the Zhiyuan Honors Program of Shanghai Jiao Tong University (SJTU), Shanghai, China in 2022. Currently, she is a third-year PhD student of Bioinformatics and Biostatistics at the School of Life Sciences and Biotechnology, SJTU. Her current research topics are tumour diagnosis, tumour burden assessment, as well as prognosis and drug response prediction. She is working on developing AI models for renal tumour subtyping, liver tumour burden assessment and progression evaluation, and multimodal-based renal cancer recurrence prediction. Additionally, she has also been exploring medical large language models, especially in the domain of visual models. Her main research interests include the field of artificial intelligence-based precision medicine, mainly focusing on medical imaging such as CT, MRI and pathological images.