



Harnessing the power of artificial intelligence to combat infectious diseases: Progress, challenges, and future outlook

Hang-Yu Zhou,^{1,3,*} Yaling Li,^{2,3} Jia-Ying Li,¹ Jing Meng,¹ and Aiping Wu^{1,*}

*Correspondence: zhy@ism.cams.cn (H.Z.); wap@ism.cams.cn (A.W.)

Received: May 25, 2024; Accepted: August 6, 2024; Published Online: September 4, 2024; <https://doi.org/10.59717/j.xinn-med.2024.100091>

© 2024 The Author(s). This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

GRAPHICAL ABSTRACT



PUBLIC SUMMARY

- Artificial intelligence technologies show significant promise in preventing and controlling infectious diseases.
- Advances made across all outbreak stages: pre-pandemic, early pandemic, pandemic, and periodic epidemics.
- Deep learning facilitates early detection, risk assessment, policy formulation, and vaccine development.
- Challenges arise from data quantity/quality, model complexity/interpretability, and individual privacy concerns.
- Promising directions lie in the deep integration of deep learning models with specific biological knowledge.

Harnessing the power of artificial intelligence to combat infectious diseases: Progress, challenges, and future outlook

Hang-Yu Zhou,^{1,3,*} Yaling Li,^{2,3} Jia-Ying Li,¹ Jing Meng,¹ and Aiping Wu^{1,*}

¹State Key Laboratory of Common Mechanism Research for Major Diseases, Suzhou Institute of Systems Medicine, Chinese Academy of Medical Sciences & Peking Union Medical College, Suzhou 215123, China

²Zhejiang Lab, Hangzhou 311121, China

³These authors contributed equally

*Correspondence: zhy@ism.cams.cn (H.Z.); wap@ism.cams.cn (A.W.)

Received: May 25, 2024; Accepted: August 6, 2024; Published Online: September 4, 2024; <https://doi.org/10.59717/j.xinn-med.2024.100091>

© 2024 The Author(s). This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Citation: Zhou H., Li Y., Li J., et al. (2024). Harnessing the power of artificial intelligence to combat infectious diseases: Progress, challenges, and future outlook. *The Innovation Medicine* 2(4): 100091.

The rapid emergence and global spread of infectious diseases pose significant challenges to public health. In recent years, artificial intelligence (AI) technologies have shown great potential in enhancing our ability to prevent, detect, and control infectious disease outbreaks. However, as a growing interdisciplinary field, a gap exists between AI scientists and infectious disease biologists, limiting the full potential of AI in this field. This review provides a comprehensive overview of the applications of AI in infectious diseases, focusing on the progress along the four stages of outbreaks: pre-pandemic, early pandemic, pandemic, and periodic epidemic stages. We discuss AI methods in early detection and risk assessment, outbreak surveillance, diagnosis and control, and understanding pathogenic mechanisms. We also propose the primary limitations, challenges, and potential solutions associated with AI tools in public health contexts while examining crucial considerations for future enhanced implementation. By harnessing the power of AI, we can develop more precise and targeted strategies to mitigate the burden of infectious diseases and improve global health.

INTRODUCTION

Infectious diseases continue to pose a significant threat to global health in the 21st century. The COVID-19 pandemic has demonstrated this threat, causing significant excess mortality and morbidity worldwide.¹ Several factors, including climate change, urbanization, and increased human mobility, have heightened the risk of infectious disease emergence and spread. These factors raise serious concerns about future outbreaks of unknown diseases, often referred to as "Diseases X".² To address these escalating challenges, we urgently need to integrate innovative technologies like AI into disease surveillance, prevention, and control strategies.

In recent years, AI, especially machine learning and deep learning, has revolutionized biomedical research.³⁻⁵ We now have access to an unprecedented amount of data from various sources (Figure 1A) including pathogen genomics, host multi-omics, clinical records, and environmental monitoring. The wealth of data presents new opportunities for AI applications in infectious disease research (Figure 1B). AI techniques are particularly effective in handling complex datasets with high dimensionality, heterogeneity. These methods can uncover hidden patterns and making accurate predictions, which can inform targeted interventions at both individual and population levels (Figure 1C).

This review focuses on cutting-edge AI applications in infectious disease prevention and control, following the timeline of disease outbreaks. We start by discussing AI-based early warning systems that integrate internet data, environmental monitoring, and host-pathogen interaction analysis to detect potential threats prior to outbreaks (Figure 1D). Next, we highlight the role of AI in enhancing outbreak surveillance through evolutionary analysis, transmission risk assessment, and inference of variants of concern (Figure 1E). We then focus on the use of AI in disease diagnosis, clinical risk prediction, digital contact tracing, and evaluation of public health interventions during pandemics (Figure 1F). Lastly, we explore AI's potential in revealing host susceptibility and pathogenic mechanisms (Figure 1G), which can inform the development of targeted prevention and treatment strategies during periodic

epidemics. We conclude by discussing the challenges of implementing AI in infectious disease contexts and exploring future directions for this field.

PRE-PANDEMIC: EARLY WARNING OF EMERGING INFECTIOUS DISEASES

Emerging infectious diseases are typically identified through clinical observation of usual symptoms. However, this method has a significant drawback: by the time these disease are detected in clinics, they have often been surreptitiously spreading for a long time. In the short term, we often lack effective vaccines, drugs and control measures. This makes early warning systems for emerging infectious crucial. AI technology provide new insights into the early response to infectious diseases. AI can help in three key ways including identifying early transmission signals of infectious diseases from Internet data, detecting pathogens in environmental surveillance data, as well as predicting zoonotic pathogens with spillover risk.

Detection of infectious disease signals based on internet data

Data collection technology and computational power has advanced rapidly. This progress allows us to use vast amounts of Internet data to detect infectious diseases early, offering promising prospects for public health surveillance. AI algorithms can parse, filter, and analyze Internet data with high precision and speed, enabling cost-effective and timely detection of infectious disease outbreaks.

Internet data like search queries, social media posts, and news reports contain valuable information about disease transmission.⁶ These user-generated digital traces have the potential to capture fluctuations in human behavior that mirror disease activity. AI methods, including linear regression and Bayesian models, can be applied to these data to nowcast and forecast outbreaks. For example, Ginsberg et al.⁷ employed linear regression to convert Google search query data into a real-time influenza surveillance system on a regional level. This system could estimate flu activity in different regions (Figure 2A). However, these methods still possess certain limitations. They may overfit data, be affected by changes in search algorithms, or be biased by data sources.⁸

During the COVID-19 pandemic, researchers have explored the integration of multiple data streams to improve the accuracy and timeliness of outbreak predictions. Kogan et al.⁹ developed a Bayesian model that incorporated diverse health and behavioral data, including Google Trends, Twitter microblogs, clinician searches, mobility data, smart thermometer measurements, and Global Epidemic Model (GLEAM) predictions. By integrating these data sources, they were able to predict upcoming COVID-19 outbreaks in near real-time, approximately weeks earlier than traditional approaches. The Bayesian model leverages the strengths of each data source while mitigating their individual limitations, resulting in more robust and reliable predictions.

More recently, Stoleran et al.¹⁰ proposed an early warning system based on machine learning. This system uses a dynamic training approach, updating the prediction model upon observing new outbreak events. It can adapt to select the most informative predictors. The study analyzed data from 97 counties in the United States from January 2020 to January 2022 and extended the method to 50 U.S. states. The experimental results demonstrated the system's strong predictive performance, successfully providing early or synchronous warnings for most of the observed outbreaks. Moreover, compared to single-source methods, the proposed multi-source

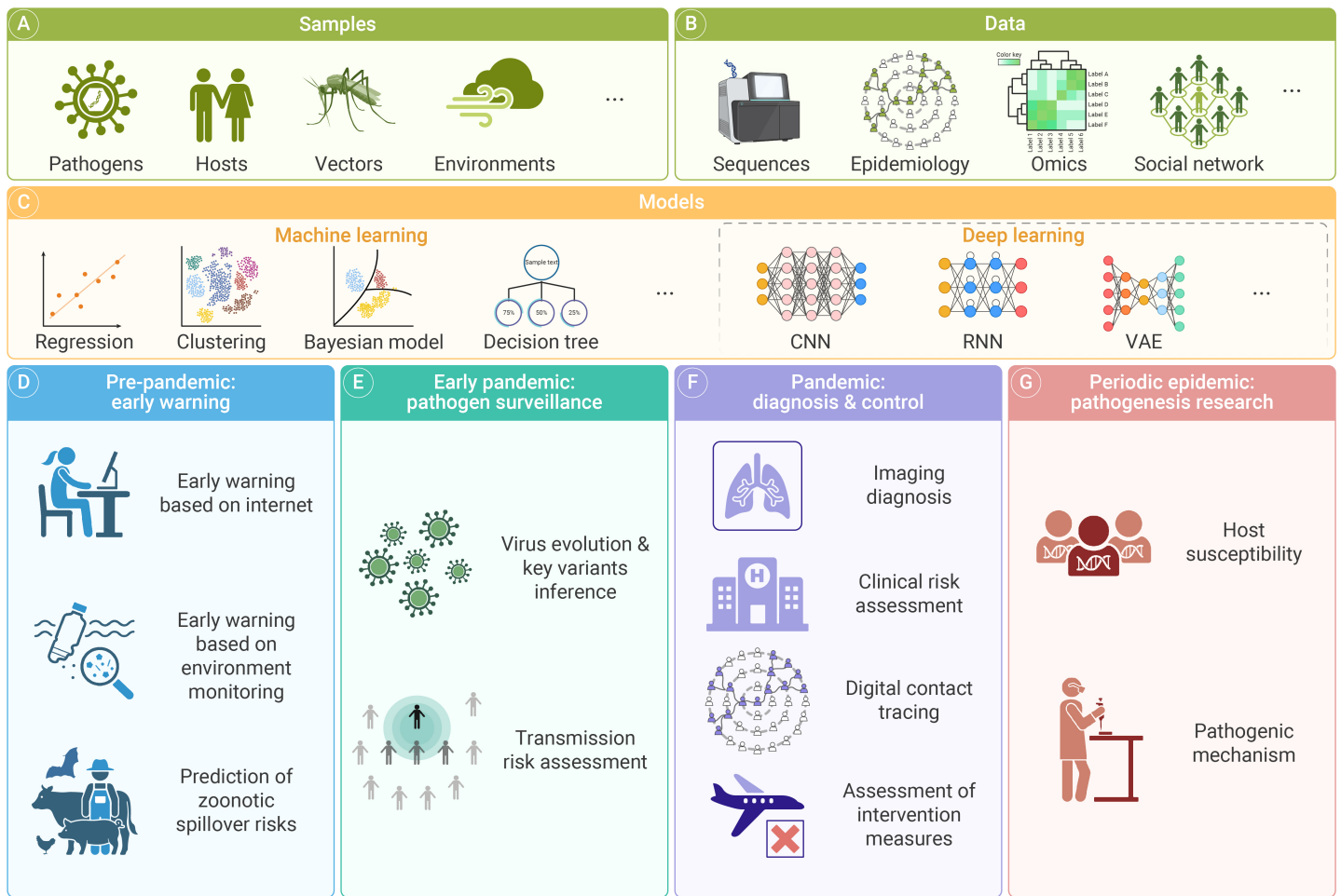


Figure 1. The schematic representation of AI utilization in infectious disease research (A) Specimen collection spans a diverse array including pathogen isolates, host tissues, vector samples, and environmental probes. (B) These collected specimens are subsequently processed, transforming into datasets such as gene sequences, epidemiological figures, and a variety of 'omics' data types. (C) The fusion of these biological datasets with advanced machine learning and deep learning algorithms enhances the construction of AI models, precisely engineered to meet distinct biological research aims. (D) In the lead-up to pandemics, AI methods have been harnessed for preemptive outbreak alerts, incorporating internet-based disease transmission signals, environmental monitoring through wastewater analysis, and zoonotic infection risk prediction. (E) At the onset of a pandemic, AI has played a pivotal role in pathogen monitoring, including tracking evolutionary patterns of viruses and projecting infectious threat levels. (F) Amidst the pandemic phase, AI has been widely integrated into disease diagnostics, clinical risk evaluation, digital exposure tracking, and analysis of non-pharmacological intervention impacts. (G) Following the pandemic and during recurrent epidemic surges, AI's application extends to the appraisal of population vulnerability and the dissection of pathogen virulence pathways.

approach significantly reduced the false alarm rate.

Infectious disease signal identification based on environmental surveillance

New, affordable high-throughput sequencing technologies has paved the way for innovative environmental surveillance of infectious diseases. This approach offer several advantages: it's non-invasive, cost-effective, and accessible to the public. Methods include monitoring wastewater,¹¹ sampling air,¹² and testing contact surface.¹³ Unlike traditional epidemiological tracking of individuals, environmental surveillance can detect circulating pathogens even before patients develop symptoms. This allows public health experts to take preventive measures prior to an outbreak.

Wastewater surveillance has become particularly popular among environmental monitoring methods. It provides comprehensive insights into environmental chemicals and biomarkers. For example, Jahn et al.¹¹ utilized generalized linear models to predict local transmission patterns of the Alpha, Beta, and Gamma variants. They analyzed SARS-CoV-2 RNA from wastewater samples in two Swiss municipalities. However, this method faces a major challenge: obtaining high-quality sequences from wastewater samples. These samples often have low viral loads and high levels of PCR inhibitors. To tackle this issue, Karthikeyan et al.¹⁴ employed a pedigree deconvolution algorithm. This algorithm achieved close to 95% genome coverage, even in samples with low viral loads (Figure 2B).

Prediction of zoonotic spillover risks

Zoonotic diseases poses a significant threat to global health. Most emerging infectious diseases originated from zoonotic viruses.¹⁵ While over 260 zoonotic viruses have been identified to date, it is estimated that approximately 1.67 million viral species within key zoonotic virus families residing in mammals and birds remain undiscovered. Nearly half of these could potentially infect humans.¹⁶

The risk of a zoonotic virus spreading to humans depends on two main factors. First, how easily can it transmit from an animal to a human? Second, how likely is it to cause disease in humans?¹⁷ Host range similarity is a key predictor of spillover risk, and identifying potential virus reservoirs can enhance surveillance efforts. Researchers have employed various approaches to predict zoonotic spillover risk. These include regression algorithms and ecological network methods. (Figure 2C) Recent studies have integrated regression and network approaches to improve predictive capability. They look at features of the virus, features of the host, and how viruses and hosts interact in networks. This combined approach has helped predict over 20,000 unknown interactions between viruses and hosts.¹⁸

EARLY PANDEMIC: INFECTIOUS DISEASE SURVEILLANCE

Timely and accurate surveillance is the cornerstone for effective infectious disease control. Traditional pathogen surveillance has several limitations. It typically involves sampling infected individuals and requires coordinating

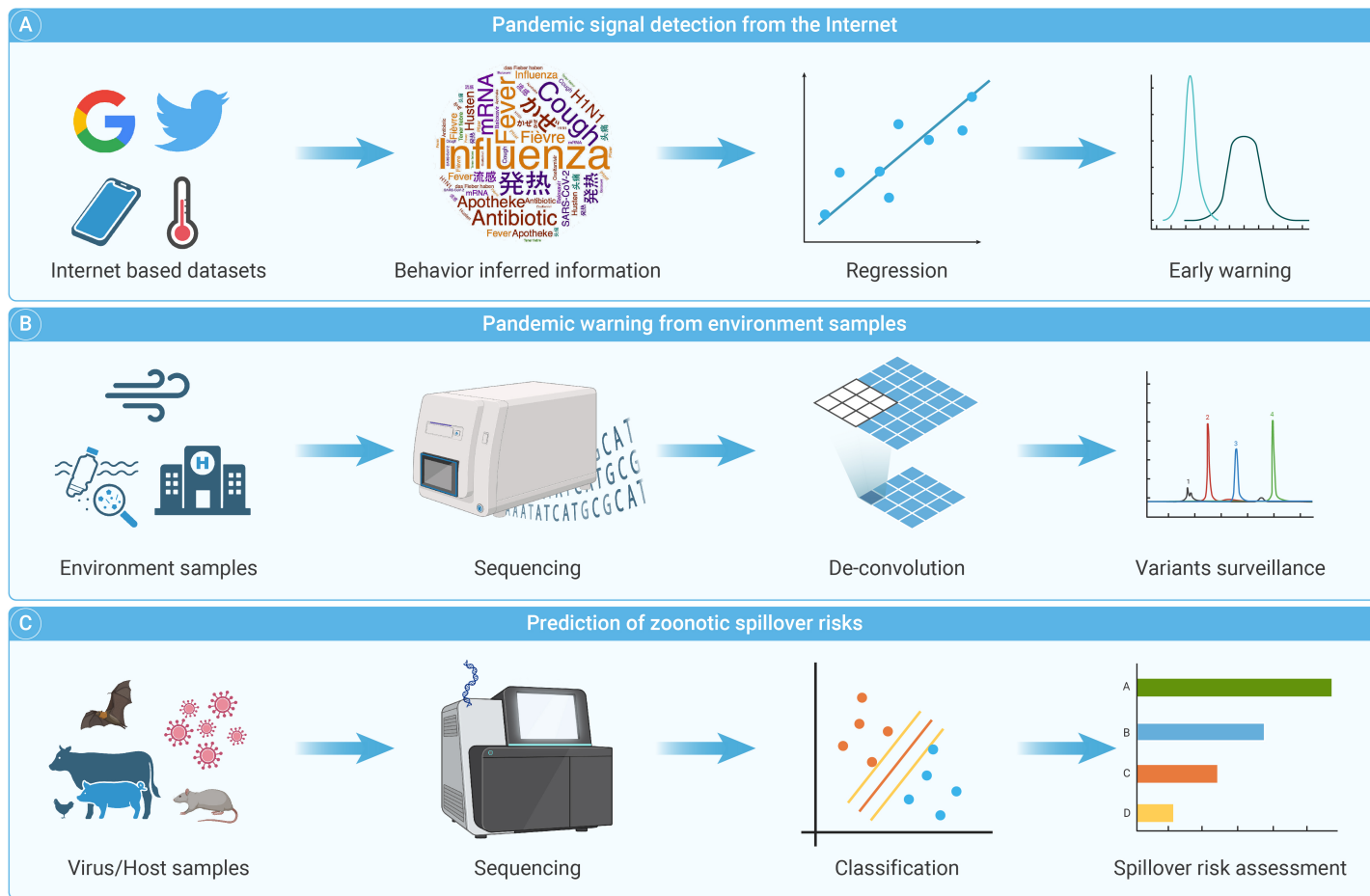


Figure 2. Applications of AI in predicting infectious disease emergence (A) AI systems utilize diverse internet-based datasets, including search trends, social media conversations, geolocation information, and user-reported symptoms. These datasets, analyzed with linear regression and Bayesian methods, enable the generation of precursory alerts for infectious disease outbreaks. (B) The increasing use of environmental samples for epidemic monitoring—air, wastewater, and clinical sources—is facilitated by AI processing. This approach empowers researchers to evaluate viral distribution even in minimally present samples. (C) Integrating machine learning algorithms with extensive collections of pathogen-host samples offers a novel avenue for gauging the risk of zoonotic disease transmission events.

sample collection, clinical testing, and sequencing across multiple sites and laboratories.¹⁹ This approach faces numerous challenges such as inadequate and biased sampling, absence of comprehensive disease reporting systems, as well as incomplete and delayed sharing of epidemiological data. AI techniques offer a novel analytical approach for infectious disease surveillance data, which can integrate multi-source data and explore the hidden correlation within the data. This allows for more agile, accurate, and real-time surveillance capabilities.

Evolutionary analysis and inference of variants of concern

The continuous evolution of pathogens, resulting in the evasion of immune responses at both individual and population levels, presents an ongoing challenge to combat infectious diseases. Accurate prediction of dominant virus strains is crucial for disease diagnosis, treatment, vaccine design, and other response measures. For instance, due to mutations in influenza viruses, it is imperative to annually anticipate the dominant variant for the subsequent year. Incorrect predictions can lead to suboptimal vaccine efficacy, thereby exacerbating the burden of seasonal influenza.²⁰

The establishment of regional sequencing centers, the development of lineage classification systems, and the global sharing of sequence data during the COVID-19 pandemic have facilitated the accumulation of an unprecedented volume of genomic data for a single pathogen (Figure 3A), revolutionizing data-driven research approaches. Over 16 million SARS-CoV-2 genome sequences from 206 countries worldwide have become accessible through the GISAID EpiCoV database.²¹ However, the vast amount of data and its geographical and temporal variability pose challenges for conventional methods. Ensuring data quality control and standardization is crucial

for the reliability of subsequent analyses.

AI techniques, particularly deep learning enable comprehensive analysis of viral sequences distribution over time and space. These methods can track viral evolution dynamics, and accurately predicting mutations that facilitate viral escape to guide epidemic response. For example, Fritz Obermeyer et al.²² developed PyR0, a Bayesian hierarchical regression model, can infer relative prevalence among viral lineages in a specific geographic region, detect lineages with increased prevalence, and identify fitness-associated mutations. In systematic backtest analyses, PyR0 successfully identified experimentally determined S mutations.

Another viable path is to transform viral variant identification into a problem that AI algorithms excel at solving. For instance, M. Cyrus Maher et al.²³ transformed the task of measuring viral fitness into determining the relative predictive significance of amino acid mutational signatures. They selected 14 predictive features based on existing literature, covering the evolution, immunity, epidemiology and other related factors of virus mutation transmission. Subsequently, the researchers applied forward feature selection and logistic regression methods to identify features with the highest predictive performance in forecasting the future prevalence probability of single amino acid mutations in SARS-CoV-2. In a retrospective study, the model demonstrated exceptional accuracy exceeding 92%, enabling early detection of pandemic mutations up to four months in advance.

Viral sequences can also be viewed as textual sequence data. Hence, classical models from natural language processing (NLP) can be applied to viral mutation prediction. Bryan Bryson et al.²⁴ drew inspiration from syntax and semantics in NLP. They showed that mutations capable of evading the immune system preserve virus infectivity while significantly altering its relative

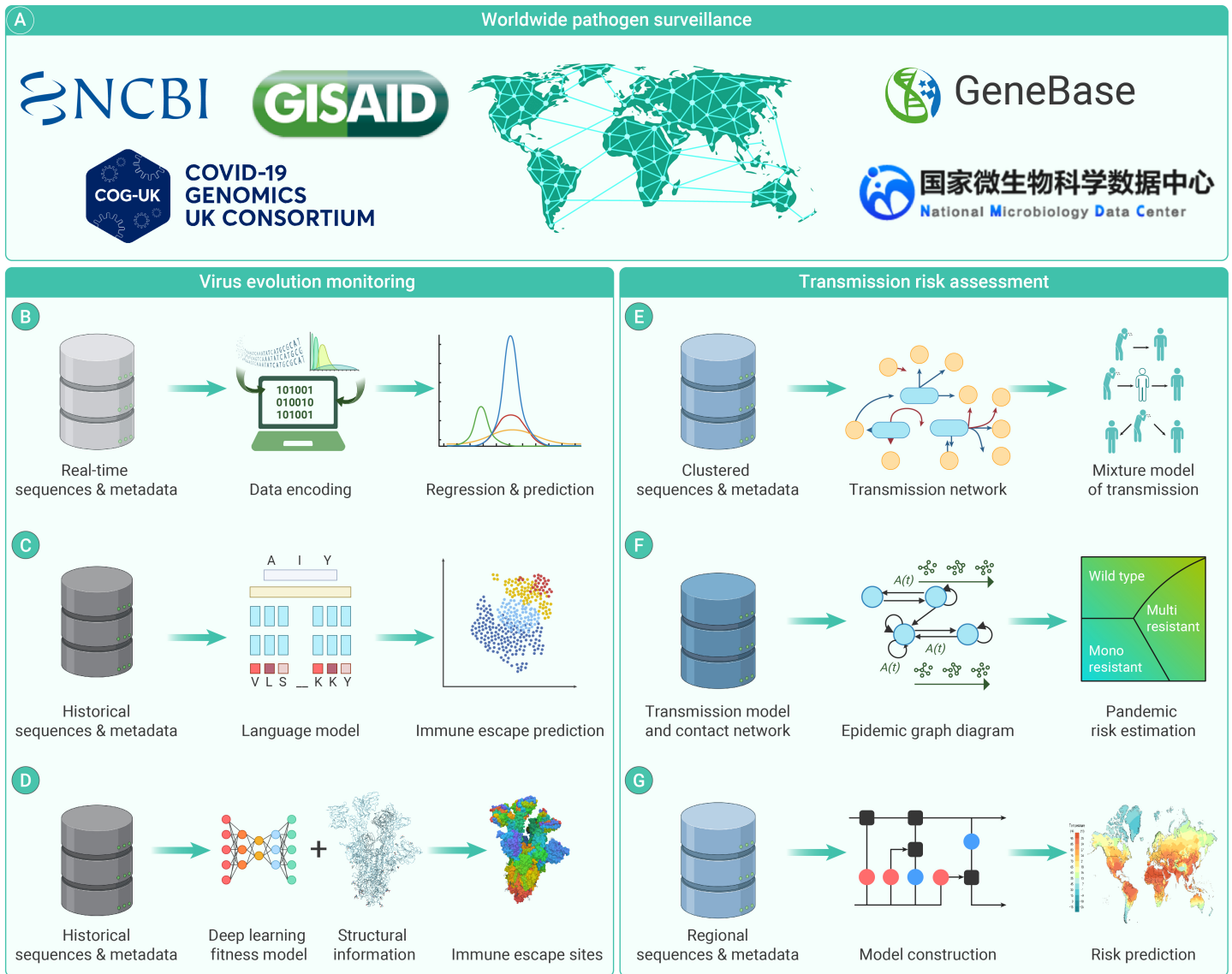


Figure 3. The application of AI technologies in infectious disease surveillance (A) Established global pathogen surveillance networks underpin a robust infrastructure facilitating the aggregation, archival, and dissemination of genetic and epidemiological data critical to infectious disease research. (B) Utilizing genotypic and epidemiologic data acquired in real time, select investigations have encoded these disparate information streams, integrating them within regression frameworks aimed at extrapolating the trajectory of specific viral variants over ensuing intervals. (C) Certain studies have harnessed linguistic computational models to forecast antigenic drift locales, simulating the repercussions of viral mutation on immune evasion through the application of semantic and syntactic paradigms, akin to those found in linguistics, to viral genomic variations. (D) Complementing live surveillance, the archival record of virological evolution is mined for forecasting key antigenic shift sites, particularly by leveraging copious datasets of related viruses augmented by AI techniques, to model potential mutation landscapes, with adjunct protein structural analyses offering vital clues for pinpointing immune escape vectors. (E) Within the domain of viral transmissibility risk evaluation, extant research has amalgamated genetic and epidemiologic evidence to delineate transmission networks, concurrently deploying AI methodologies to refine a composite model predicting the likelihood of various transmission dynamics. (F) Selected research initiatives have embarked on the application of graph-based computational strategies derived from machine learning to quantify epidemic risks, epitomizing the multidisciplinary confluence of computational and epidemiological sciences. (G) The assimilation of AI technologies has enabled sophisticated scrutiny of localized transmission hazards. Investigations leveraging LSTM models have thoroughly appraised the dispersal perils in different locales, underpinned by regional genomic sequences and epidemiological frameworks.

to the immune system. This is analogous to word changes preserving sentence syntax but modifying meaning. Based on this insight, they adopted a neural language model based on bidirectional long short-term memory (BiLSTM) architecture. This model identifies escape mutations under conditions involving constrained semantic changes. (Figure 3B)

To improve predictive performance, researchers often integrate data from multiple sources, such as sequence, structure and energy information. In a recent study, Thadani et al.²⁵ proposed the EVEscape model, which combines deep learning with biophysical constraints. The EVEscape model considers three factors simultaneously: firstly, the possibility of a mutation maintaining virus fitness; secondly, the likelihood of a mutation occurring in the antibody accessible region; and thirdly, the potential for a mutation to disrupt antibody binding. Using these factors, the model calculates the probability of mutation-induced immune escape. (Figure 3C) The key advantage of the EVEscape model lies in its ability to learn from historical sequence data and achieve

higher accuracy predictions. The model addresses the limitations of existing models, which often rely heavily on real-time sequencing and antibody testing. As a result, EVEscape is more suitable for use during early stages of infectious disease outbreaks.

Transmission risk assessment

Traditional epidemiological surveillance often focuses on detecting disease cases within specific geographical areas. However, this approach has limitations, especially for COVID-19. The presence of asymptomatic infections and the incubation period can lead to biased predictions. With decreasing sequencing costs, improving accuracy, and increasing portability of sequencing instruments, surveillance of the spread of infectious diseases becomes more timely and effective. AI technology enhances this process by integrating and analyzing data from diverse sources. This integration allows for better predictions of COVID-19's spatial transmission trends at various

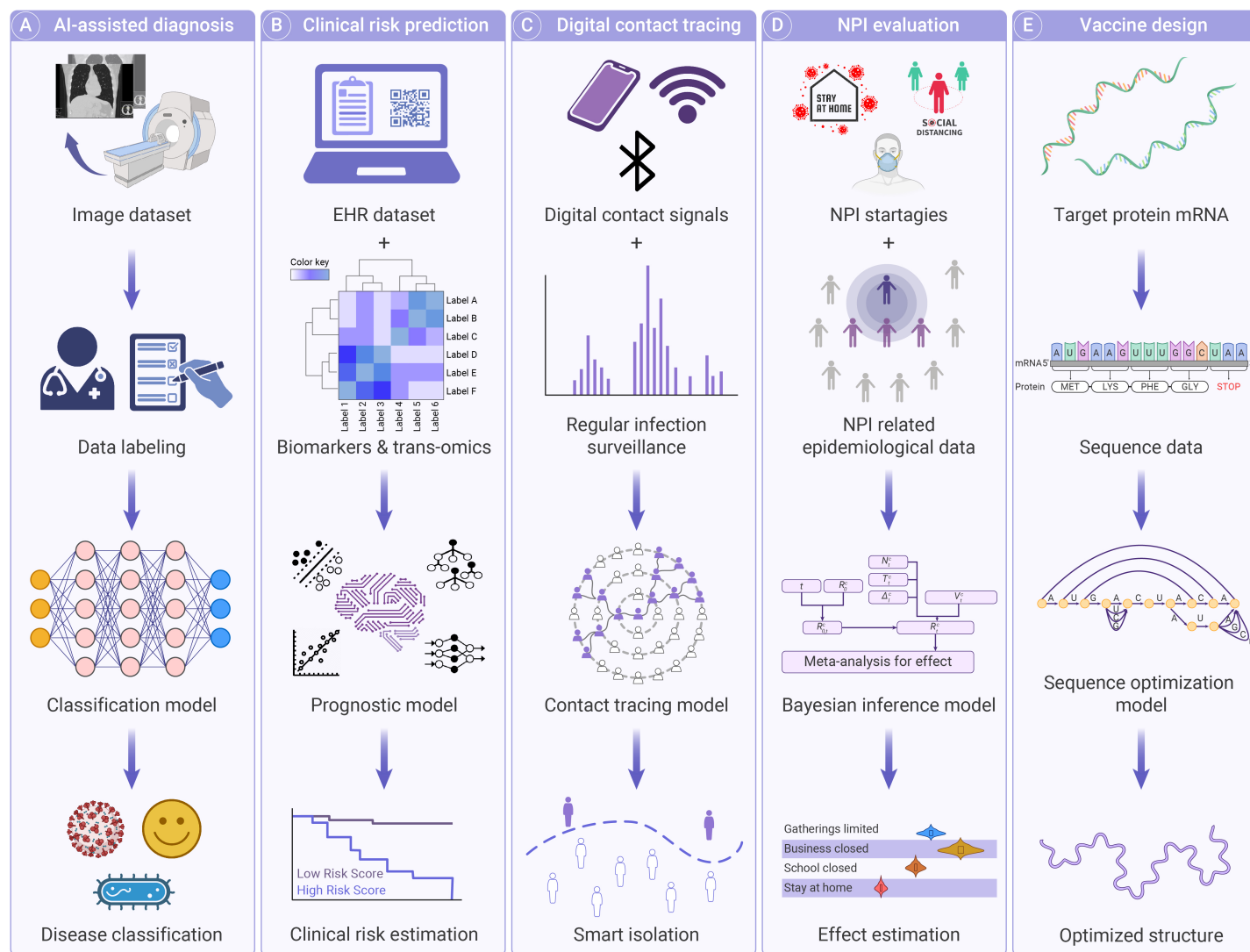


Figure 4. The role of AI technologies in disease diagnosis and epidemic control (A) This panel displays the synergistic fusion of advanced AI methodologies with image recognition, signalling a new frontier in computed tomography (CT)-based disease detection. It depicts the sophistication of image recognition systems, which, leveraging clinician-annotated datasets, offer improved disease identification and categorization. (B) The evolution of AI now permits the extrapolation and forecasting of salient factors tied to clinical risk, harnessed from electronic medical records, biomolecular markers, and omics insights, thus underpinning the judicious allocation of medical resources. (C) The proliferation of mobile technology introduces an innovative approach to infectious disease surveillance and localization. By amalgamating signals of digital interaction stemming from cellular and Bluetooth data with traditional infection monitoring, we can craft AI-informed models of infection propagation, furnishing an empirical basis for more sophisticated isolation protocols. (D) Models for appraising the effectiveness of non-pharmacological interventions incorporate relevant infectious disease epidemiologic intelligence, assessing their utility in mitigating the spread of infectious agents. (E) The introduction of concepts from computational linguistics and other areas of AI technology has greatly facilitated the design of vaccine sequences.

scales. As a result, it provides valuable insights for developing effective epidemic intervention measures.

Spatial-temporal risk assessment. Modeling future trends in infectious disease transmission is a valuable tool. It enhances situational awareness, reveals potential transmission pathways, and evaluates the risk associated with disease spread. AI models integrating epidemiological, genomic, and mobility data can capture the fine-grained dynamics of disease spread, providing actionable insights for outbreak response. Currently, infectious disease modeling methods can be divided into three categories: traditional compartment models, AI methods such as machine learning, and a combination of the two approaches.

Traditional compartment models, like the susceptible-exposed-infectious-removed (SEIR) model, are based on transmission dynamics and utilize known exposure information to predict key epidemiological parameters and transmission trends. However, these simple models are prone to bias due to asymptomatic infections and the incubation period of COVID-19. To address this limitation, researchers have developed scalable compartmental models. These integrate aggregated mobility and physical interaction data with traditional SEIR models, enhancing real-time tracking and short-term prediction

capabilities for infectious disease spread.²⁶ For instance, Chang et al.²⁷ proposed an integrated SEIR model using fine-grained dynamic mobile networks. This model simulates SARS-CoV-2 transmission in the ten largest US metropolitan areas. By leveraging mobile phone geolocation data to capture population movements, this model effectively fits observed case numbers and enables detailed analysis of high-risk infection populations and locations. Similarly, Leung et al.²⁸ utilized Octopus transport card transactions as a digital proxy for population mixing. They integrated age-specific digital mobility information into the traditional SIR model, enabling real-time tracking of local COVID-19 transmission. (Figure 3E)

Machine learning methods have also been employed for risk assessment of infectious disease transmission at different spatial scales. (Figure 3F) These methods included generalized linear models, long short-term memory (LSTM) algorithms, clustering,²⁹ and graph computation.³⁰ Cooper et al.³¹ utilized generalized linear models to quantify factors associated with the risk and magnitude of hospital-acquired transmission. Their study found that approximately 1-2% of admitted patients are infected with SARS-CoV-2 during their hospital stay, with the risk often associated with fewer single rooms and lower heating per bed. Ward et al.³² used a spatially integrated

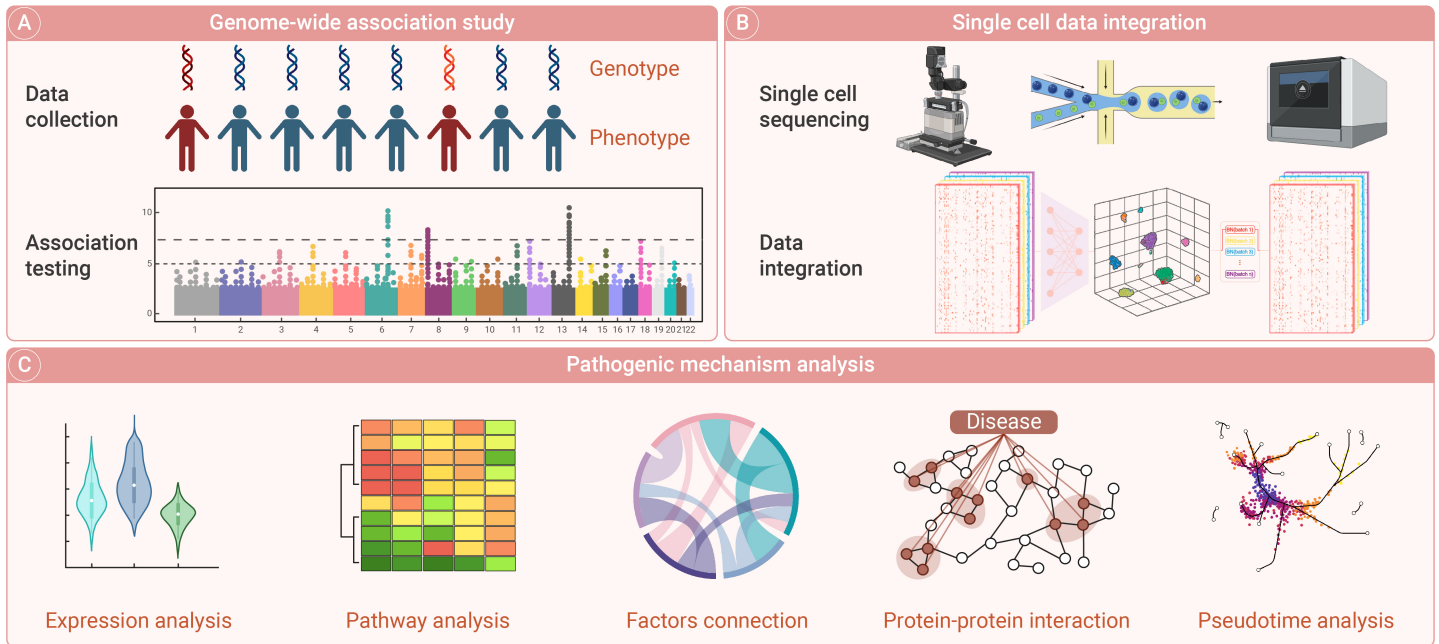


Figure 5. Exploiting AI technologies to unravel host susceptibility and pathogenic mechanisms (A) Established GWAS methodology is extensively applied to elucidate associations between genetic markers and phenotypic characteristics. (B) AI-centric techniques are principally leveraged for the synthesis of single-cell data, advancing the understanding of pathogenic mechanisms. (C) Enriched biological datasets are subjected to comprehensive analyses integral to decoding pathogenic processes, encompassing expression profiling, pathway elucidation, and proteomic interaction mapping.

LSTM (SI-LSTM) framework for monitoring early indicators of local COVID-19 outbreaks at the local authority district scale in the UK (Figure 3G). By integrating diverse high-resolution spatial data sources, this approach accurately identified hot spot locations with over 99% accuracy for periods of one month or longer.

Some studies have combined traditional SEIR models with machine learning techniques to reduce prediction bias. Nicholson et al.³³ proposed a customized causal debiased-based model that merges the conventional SIR method with Bayesian inference and Markov Chain Monte Carlo methods. This approach achieves unbiased estimation of crucial epidemiological parameters by combining data from traditional epidemiological surveillance with UK National Randomised Surveillance Survey (REACT) data, enabling debiased estimates at finer spatio-temporal scales. Yang et al.³⁴ developed a model inference system based on the susceptible-exposed-infectious-recovered-susceptible (SEIRS) model. Their system incorporates a Kalman adjustment filter algorithm and spatial reprobating techniques. It accurately captures SARS-CoV-2 transmission dynamics in different regions by fitting weekly incidence and mortality data while accounting for various factors. These factors include insufficient testing, delayed reporting, disease seasonality, vaccination, interventions, and other variables.

Risk assessment of superspreading events. Superspreading events, where a single individual disproportionately contributes to a significant number of secondary infections,³⁵ often play a crucial role in regional transmission dynamics and the overall trajectory of pandemics. Understanding these events is pivotal for the subsequent development of preventive and control measures against infectious diseases. Lemieux et al.³⁶ employed genomic epidemiology to investigate the introduction and dissemination of SARS-CoV-2 in Boston between March and May 2020. The findings suggest that superspreading events involving younger, healthier, and more mobile populations impose greater costs on society. Similarly, Lau et al.³⁷ utilized Bayesian inference and Markov Chain Monte Carlo methods in their study. They proposed that individuals infected at a younger age (under 60) tend to be more infectious and are more likely to contribute to supertransmission. Notably, their research highlighted the significant role of funerals as superspreading events, particularly in driving outbreaks in rural areas.

PANDEMICS: DIAGNOSIS AND CONTROL OF INFECTIOUS DISEASES

During the COVID-19 pandemic, AI has been successfully applied across all

aspects of infectious disease diagnosis and control. It has played a pivotal role in curbing disease transmission and reducing the burden of infectious diseases. Considerable literature has summarized the utilization of AI in drug research and development.³⁸⁻⁴¹ A recent comprehensive review by Wong et al.⁴ analyzed AI application in drug discovery for infectious diseases. This review covers areas such as prediction drug activity, understanding drug-target interaction and mechanism of action, and assessing therapeutic effects. Thus, in this chapter, we focused on typical applications of AI in auxiliary diagnosis of infectious diseases, clinical risk assessment, digital contact tracing, evaluation of public health epidemic prevention measures, and vaccine design.

AI-assisted disease diagnosis

Early identification and diagnosis of infectious diseases are crucial for rapid treatment, patient isolation, and disease containment.⁴² In the case of COVID-19, the diagnosis is confirmed by molecular PCR testing, which is time-consuming and has a high false positive rate. Chest CT is faster and provides more detailed pathological information, serving as an important basis for evaluating and diagnosing pulmonary infections caused by infectious diseases. During pandemics, the exponential increase in patient numbers and limited availability of specialized radiologists hinder effective triage of suspected patients. AI algorithms present valuable tools to address these challenges.^{43,44} With large labeled datasets and advanced computing resources such as GPUs, AI algorithms can achieve or even surpass human experts' performance in certain medical image diagnostic tasks. Additionally, they also offer advantages such as high efficiency, good repeatability, and easy large-scale deployment.⁴⁵

Accurate AI systems based on CT scans contribute to the early diagnosis, monitoring, and treatment of infectious diseases, while also providing guidance for medical resource allocation. In the initial stage of COVID-19, Mei et al.⁴⁶ developed a model integrating algorithms like deep convolutional neural networks (CNNs), support vector machines (SVMs), random forests (RFs), and multilayer perceptrons (MLPs). By utilizing chest CT results, clinical symptoms, exposure history, and laboratory tests, rapid identification of SARS-CoV-2 infection was achieved. The model exhibited an area under the receiver operating characteristic curve (AUC) of 0.92, with a sensitivity comparable to senior thoracic radiologists at 84.3%. In practical AI-assisted diagnosis of infectious diseases, successful outcomes rely on effective

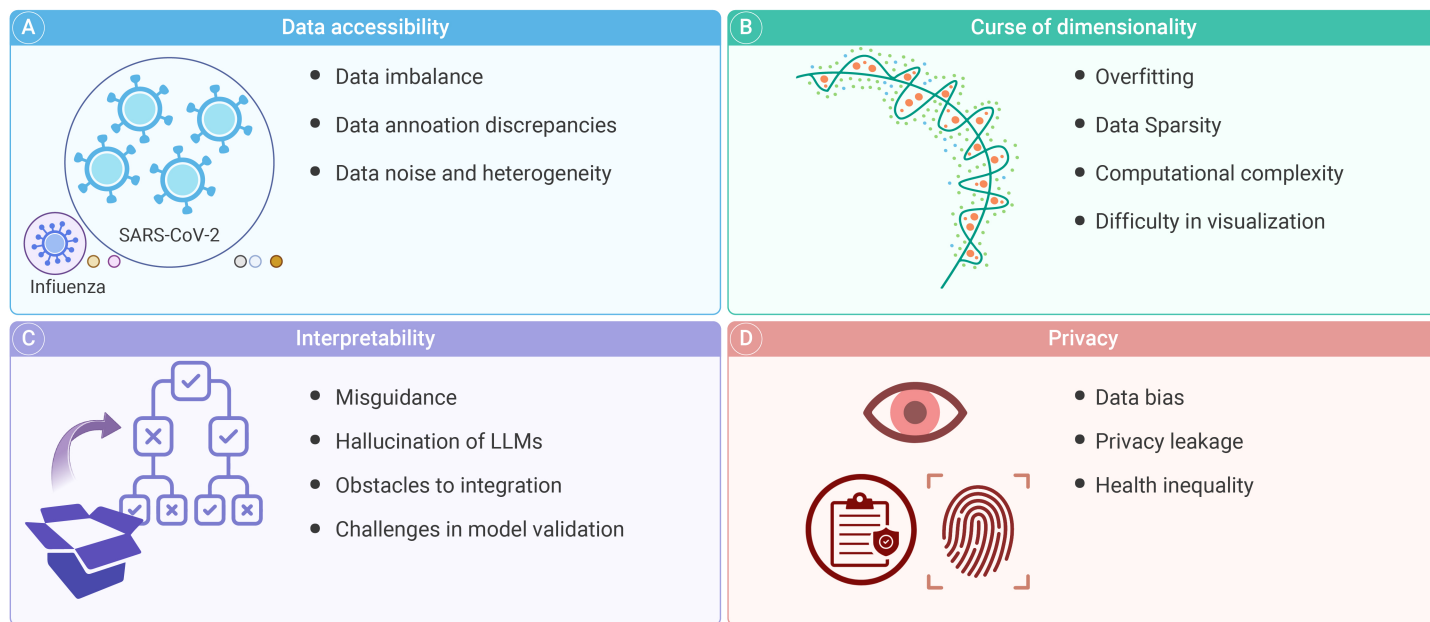


Figure 6. The challenges of implementing AI in infectious disease research (A) A historical bottleneck to AI utilization has been the lack of ample datasets and the deficiency of high-quality training resources, restricting the broader application of AI in the infectious disease field. (B) Furthermore, the complexity of biological systems, alongside data paucity, risks invoking the curse of dimensionality. This phenomenon may lead to overfitting of models on one front and the intensification of model intricacy and visualization obstacles on the other. (C) Model interpretability stands as a cornerstone in infectious disease research, impacting the robustness of model validation, the universality of their applicability, and public reception of the derived conclusions. (D) Privacy and information security are intricately woven into all aspects of AI, presenting an enduring dilemma in balancing data confidentiality against the need for information availability – a challenge that must be diligently navigated as AI finds its footing in the domain of infectious disease research.

collaboration between clinicians and AI systems. One viable approach is adopting a “human-in-the-loop” strategy for human-computer interaction. (Figure 4A) For example the AI tool CoDoC⁴⁷ employed a prediction model to diagnose imported cases and calculate confidence score. These confidence scores are then used to determine whether a case should be handled by a clinician or the prediction model. Research showed that CoDoC can reduce the false positive rate by 5–15% in tuberculosis classification.

Responding to global infectious disease crises often requires inter-regional data sharing due to insufficient and potentially biased high-quality medical image data.⁴⁸ Concerns about data security and trustworthiness have led to the emergence of federated learning as a solution. The UCADI framework, developed by Bai et al.,⁴⁹ is a multinational collaborative AI framework based on federated learning. It allows distributed training and independent execution without the need for data sharing. By utilizing chest CT data from patients in China and the UK, the model achieved better sensitivity and specificity than local models, while performing comparably to professional radiologists’ assessments. Another approach is to utilize the latest generative models to generate medical image data, improve the model’s predictive performance for underrepresented groups. Recent studies⁵⁰ have shown that incorporating real samples into chest X-ray image data through diffusion models can effectively enhance the robustness of the model.

Clinical risk prediction

During infectious disease outbreaks, the surge in patient numbers often significantly strains medical resources. Early prediction of patient clinical outcomes can support data-driven medical resource allocation, alleviate the burden on healthcare systems, and serve as a reference for comprehensive decision-making.

Traditional patient prognostic prediction models are typically based on rule-based scoring systems developed from expert experience. Such models usually utilize limited temporal or trend information, have restricted input features, and suffer from high false positive rates and challenges in real-time prediction. In recent years, the introduction of AI methods has partially addressed these issues.^{51–53}

The widespread adoption of electronic health records (EHRs) provides researchers with high-resolution data for further analysis. Recurrent neural network (RNN) models, represented by Long Short-Term Memory (LSTM)

networks, can effectively incorporate previous time series information from EHRs along with newly acquired data. This enables automatic and continuous assessment of disease severity in patients, guiding for hierarchical diagnosis and treatment strategies.^{54,55}

Based on EHR data, Gao et al.⁵⁶ developed an ensemble model called MRPMC. This model consists of four machine learning algorithms: logistic regression (LR), support vector machines (SVM), gradient boosting decision trees (GBDT), and neural networks (NN). It accurately predicts the risk of physiological deterioration and death in COVID-19 patients up to 20 days in advance. The model achieved high performance in the validation data test, with area under the receiver operating characteristic curve (AUC) values ranging from 0.9186 to 0.9762.

Apart from EHR data, biomarkers, transcriptome information, and CT images can also be utilized in various prognostic prediction tasks (Figure 4B) to enhance the predictive performance of models.^{57,58} Yan et al.⁵⁹ developed a machine learning model based on a supervised multi-tree XGBoost classifier. They identified three biomarkers, namely lactate dehydrogenase (LDH), lymphocytes, and high-sensitivity C-reactive protein (hs-CRP), as the most discriminative indicators for predicting patient mortality more than 10 days in advance, achieving an accuracy rate of over 90%. In a recent study, Devaux et al.⁶⁰ proposed that a specific fragment within the long non-coding RNA, LEF1-AS1, also serves as a robust predictor of in-hospital mortality among COVID-19 patients.

Transparency and interpretability are crucial for the application of AI models in clinical practice. Clinicians need to understand the underlying reasoning processes of AI models to establish trust in the predictions and detect false ones. Furthermore, it is also essential to disclose the clinical factors highly correlated with predicted outcomes. To enhance model interpretability, Lauritsen et al.⁶¹ proposed an explainable AI-based early warning score system (xAI-EWS) to predict acute critical illness from EHRs. The system demonstrated exceptional predictive performance while effectively presenting the most relevant input data associated with prediction outcomes, thereby improving model transparency and interpretability.

Beyond short-term effects, growing scientific and clinical evidence indicates that SARS-CoV-2 infection may have potential post-acute and long-term effects on multiple organ systems, referred to as post-acute sequelae of COVID-19 (PASC). Recently, some studies have introduced AI techniques for

Table 1. Summary of methods used in each stage of pandemics

Stage	Method	Common Usage	References	Illustrations
Pre-epidemic	Linear regression	Detecting influenza epidemics using search engine query data	[7], [8]	Utilizing Google search query data to estimate influenza activity in different regions
	Bayesian approach	Integrating multiple data streams for early warning of COVID-19 outbreaks	[9], [10]	Combining diverse health and behavioral data sources to predict COVID-19 outbreaks weeks in advance
	Generalized linear models, Random forest	Predicting local transmission patterns of SARS-CoV-2 variants and AMR genes from wastewater samples	[11], [14]	Analyzing SARS-CoV-2 RNA sequencing data and antimicrobial resistance in urban sewage to track variant transmission and hospital environments
	Support vector machine, class weight classification	Predicting virus-mammal associations by integrating viral, host, and network features	[18]	Combining viral features, host features, and virus-host network topological features to predict unknown virus-mammal associations
Early pandemic	Bayesian hierarchical regression, forward feature selection, logistic regression	Inferring relative prevalence of viral lineages, identifying fitness-associated mutations, and forecasting future prevalence	[22], [23]	Analyzing SARS-CoV-2 genome sequences to track viral evolution, detect mutations associated with increased fitness, and predict mutation prevalence
	Bidirectional long short-term memory (BiLSTM), deep learning with biophysical constraints	Identifying immune escape mutations and predicting their probabilities	[24], [25]	Applying NLP and integrating sequence, structure, and energy data to predict viral mutations that evade immune responses
	Network-based SEIR models	Simulating SARS-CoV-2 transmission in metropolitan areas using mobile phone data	[27], [28]	Incorporating population mobility data into epidemiological models to capture transmission dynamics across locations
	Spatially integrated long short-term memory (SI-LSTM)	Monitoring early indicators of local COVID-19 outbreaks	[32], [54]	Integrating high-resolution spatial data sources to identify COVID-19 hotspots with high accuracy
	Bayesian inference, Markov chain Monte Carlo, Kalman filter, spatial reprobing	Estimating unbiased epidemiological parameters and capturing transmission dynamics of variants	[33], [34], [77]	Combining surveillance data and fitting incidence and mortality data to estimate key parameters and dynamics of different variants
	Genomic epidemiology	Investigating the impact of superspreading events on SARS-CoV-2 transmission	[36], [37]	Analyzing SARS-CoV-2 genome sequences to elucidate the role of superspreading events in driving outbreaks
	Deep CNN, SVM, RF, MLP	Rapid identification of SARS-CoV-2 infection from chest CT scans and clinical data	[45], [46]	Developing AI systems that achieve high accuracy in diagnosing COVID-19 from medical imaging and clinical features
Pandemics	Federated learning	Advancing COVID-19 diagnosis with privacy-preserving collaboration	[49]	Enabling distributed training and execution of AI models without the need for data sharing across institutions
	Ensemble models (LR, SVM, GBDT, NN), XGBoost	Predicting risk of physiological deterioration, death, and early mortality in COVID-19 patients	[53], [54], [59]	Combining multiple machine learning algorithms to accurately predict clinical outcomes and identify discriminative biomarkers
	Interpretable AI-based early warning score system (xAI-EWS)	Predicting acute critical illness from electronic health records	[55], [61]	Developing explainable AI models that provide insights into the most relevant input data associated with prediction outcomes
	Bayesian hierarchical model	Estimating the effect of non-pharmaceutical interventions on COVID-19 transmission	[26], [76]	Quantifying the impact of various non-pharmaceutical interventions on reducing COVID-19 transmission rates across countries
	Mathematical models, case studies	Quantifying the impact of digital contact tracing on epidemic control	[65], [66]	Assessing the effectiveness of mobile app-based contact tracing in controlling the spread of fast-spreading epidemics
	Deterministic finite automata (DFA) algorithm	Compress the search space of massive mRNA sequences	[79]	Finding the best mRNA vaccine sequence with high speed
	Top-down reinforcement learning-based approach	Generate novel protein architectures for vaccine design	[80]	Creating an ultra-compact icosahedral architecture that displays immunogens at exceptionally high densities, enhancing vaccine responses

Table 1. (Continued)

Stage	Method	Common Usage	References	Illustrations
Periodic epidemics	Genome-wide association study (GWAS)	Identifying genetic susceptibility loci for infectious diseases	[87]	Detecting genetic variants associated with increased risk or severity of infectious diseases like COVID-19
	Polygenic risk score (PRS) models	Predicting individual infection risk using GWAS results	[90], [91]	Constructing risk scores based on GWAS findings to predict susceptibility or severity of infectious diseases
	Single-cell RNA sequencing (scRNA-seq)	Investigating host immune responses and pathogenesis of infectious diseases	[95], [97], [98]	Analyzing single-cell gene expression profiles to uncover cellular heterogeneity and molecular signatures associated with disease progression
	Autoencoders (AE), variational autoencoder (VAE)	Capturing features, enhancing signal-to-noise ratios, and integrating multiple batches of single-cell data	[94], [99], [100], [101]	Applying unsupervised learning to extract meaningful representations, improve clustering accuracy, address batch effects and enable online integration of datasets
	Non-negative matrix factorization (NMF)	Improving interpretability and batch integration of single-cell data	[102], [103]	Combining linear and nonlinear dimensionality reduction techniques to enhance the interpretability of single-cell clustering results

risk prediction of PASC, such as regression analysis,⁶² causal inference,⁶³ and clustering.⁶⁴ These studies have identified potential risk factors and sub-phenotypes associated with PASC, providing valuable insights for understanding and managing this emerging challenge.

Digital contact tracing

Contact tracing is an established part of infectious disease control. During the COVID-19 pandemic, testing individuals and tracing their contacts have become crucial nonpharmaceutical interventions in many countries to break the transmission chain of SARS-CoV-2. This strategy aims to prevent further spread of secondary infections (infections acquired from the primary case) and has proven effective in multiple nations' responses to the COVID-19 pandemic.

Traditionally, contact tracing involves the initially traced individuals recalling their most recent close contacts and activities. Those at risk of infection are then contacted and advised to take measures to reduce further transmission. However, manual contact tracing is slow, labor-intensive, and may not fully reconstruct close contacts. In recent years, digital sensor-based techniques have been developed as a complementary approach to manual tracing. (Figure 4C) Contact tracing applications powered by AI can effectively identify potential hotspots and promptly notify individuals who have been in close contact with infected patients. This helps mitigate the spread of infectious diseases and preventing future outbreaks.

The COVID-19 pandemic highlighted the need for rapid contact tracing. However, the high transmissibility, relatively short incubation periods, and the potential for asymptomatic transmission of SARS-CoV-2 pose challenges for effective isolation and contact tracing. Ferretti et al.⁶⁵ employed mathematical models and case studies to elucidate how mobile phone applications can control fast-spreading epidemics. These apps facilitate rapid and widespread contact tracing and isolation, replacing the time-consuming manual process with instantaneous signals transmitted from a central server. This approach significantly speeds up the contact tracing process, potentially reducing it from a week to mere moments.

Recent studies have further quantified the contribution of digital contact tracing programs in mitigating outbreak spread. Kucharski et al.⁶⁶ investigated the transmission reduction achieved by digital contact tracing interventions across various settings. Leveraging social connection data from an extensive study in the UK encompassing more than 40,000 participants, they found that self-isolation and home isolation combined with app-based contact tracing can reduce transmission by 47%. Wymant et al.^{67,68} revealed that the NHS COVID-19 app prevented between 284,000 and 594,000 cases from occurring through contact tracing. In a recent study, Leng et al.⁶⁹ quantified the transmission risk of different contact patterns by analyzing digital tracing data from 7 million users. The findings revealed that prolonged exposure significantly amplifies the risk of transmission compared to proximity

alone, with a 15-minute close contact carrying a 2.7-fold higher risk than a 5-minute contact. This study not only highlights the potential of infectious disease contact tracing apps as an effective intervention for reducing transmission but also offers novel insights for advancing quantitative epidemiology.

The effectiveness of digital contact tracing programs relies on public trust and utilization. Kretzschmar et al.⁷⁰ showed that mobile app-based contact tracing is less effective when fewer people use the app. However, due to its inherent high efficiency, even with low coverage, it was still more effective than traditional contact tracing. For instance, with only 20% coverage, it could reduce the effective reproductive number R_e (the average number of secondary cases generated by an infectious individual in a population with some level of immunity) by 17.6%.

However, contact tracing has been less effective in countries such as the United States and the United Kingdom, where concerns regarding privacy have hindered the widespread adoption of digital tracking technologies.⁷¹ Therefore, a significant future direction for digital tracking technology lies in the development of a robust privacy protection framework, such as decentralized data storage, encryption, and anonymization techniques, to ensure data security and build public trust. Furthermore, to enhance prediction accuracy, it is essential to incorporate local biometric, pathogen, and environmental data into the modeling process alongside proximity-based contacts.⁷² This data integration can provide a more comprehensive understanding of transmission dynamics and inform targeted interventions.

Evaluation of public health interventions

In the absence of vaccines and antivirals, non-pharmaceutical interventions (NPIs) are the only option to delay and mitigate the spread of emerging infectious diseases in the population. (Figure 4D) Government interventions may incur significant economic and societal costs while impacting individual behavior, mental health, and social stability. Therefore, the development of intervention measures needs to consider the balance between the social impacts and the potential effectiveness in controlling the spread of epidemics. An appropriate combination of NPIs can effectively contain the spread of the virus and reduce the strain on health care systems.

Travel restrictions are particularly effective in the early stages of an outbreak when transmission is often limited to a specific area as the primary source.⁷³ Kraemer et al.⁷⁴ utilized real-time flow data and comprehensive case data (including travel history) from Wuhan to elucidate the role of imported cases in transmission across Chinese cities and assess the impact of control measures. The study demonstrated that the stringent control measures implemented in China significantly impeded the spread of COVID-19. Another study conducted in Brazil also showed the effectiveness of interventions aimed at reducing community and nationwide mobility in mitigating virus transmission and diffusion. Candido et al.⁷⁵ employed a Bayesian semi-

mechanical model to analyze mortality statistics related to severe acute respiratory infections and personnel mobility data. They estimated the daily variation in the effective reproductive number R_e for Sao Paulo and Rio de Janeiro. The study revealed that NPIs, such as school closures and store closures, had the potential to decrease R_e in Sao Paulo and Rio de Janeiro from >3 to 1–1.6.

During the COVID-19 pandemic, governments have implemented various combinations of NPIs to control the virus spread. However, little is known about the effectiveness of different NPIs in reducing transmission. Evidence-based research on the efficacy and societal costs of different NPIs is crucial for improving public health responses during epidemics. Brauner et al.²⁶ employed a Bayesian hierarchical model to estimate the effect of seven NPIs on COVID-19 transmission across 41 countries from January to May 2020, using data from over 34,000 NPIs implementations. Their findings revealed that closing educational institutions, limiting gatherings to 10 or fewer individuals, and shutting down face-to-face businesses significantly reduced transmission rates. Similarly, Klimek et al.⁷⁶ quantified the impact of NPIs implemented in 79 countries or territories on mitigating virus transmission and predicting their efficacy. The most effective NPI included curfews, lockdowns, restrictions on gatherings, etc. Less invasive but more costly measures such as risk communication strategies and assistance programs for vulnerable populations were equally effective. This study serves as a valuable reference for governments to choose the types and timing of NPIs.

The response of China to COVID-19 offers a unique, real-world dataset to evaluate of NPI effectiveness for emerging respiratory infectious diseases in different settings. Ge et al.⁷⁷ developed mathematical and Bayesian inference models to simulate transmission. They assess the impact of NPIs on different variants of SARS-CoV-2, using data from over 1,200 infection clusters across China. The study showed that contact tracing proves more effective than social distancing or mass testing in containing initial outbreaks. However, during pandemics, social distancing measures play a crucial role due to their stronger cumulative effects.

At the end of an epidemic, a systematic and orderly cancellation of NPIs is necessary to reduce interference with people's normal work and life. It is crucial to pre-assess the impact of lifting NPIs on SARS-CoV-2 transmission dynamics. Li et al.⁷⁸ assessed the temporal correlation (measured as R_e) between the lifting of different NPIs and the level of SARS-CoV-2 transmission in 131 countries, using data from over 12,000 NPI relaxations. The effectiveness of NPI elimination does not manifest immediately, with varying timeframes required to achieve specific transmission levels for each NPI. These findings offer crucial evidence for policymakers regarding the optimal timing for implementing significant changes in NPI strategies.

Vaccine design

The conventional process of vaccine development is often time-consuming, posing challenges in responding promptly to highly mutable pathogens or emerging infectious diseases. The integration of AI methods can effectively enhance the efficiency and cost-effectiveness of vaccine design, serving as a crucial tool in mitigating the spread of infectious diseases.

Firstly, the utilization of AI methods could facilitate targeted vaccine design by accurately predicting dominant variants based on pathogen evolutionary patterns (refer to section 3.1). mRNA vaccines have exhibited remarkable potential in combating COVID-19 due to their facile scalability in production and consideration for safety and efficacy. Zhang et al.⁷⁹ introduced the concepts of case representation and case resolution from computational linguistics. This translates the process of finding the most stable mRNA sequence into finding the most likely sentence among many sequences with similar pronunciation. Specifically, the deterministic finite automata (DFA) algorithm was used to compress the search space of massive mRNA (Figure 4E), and the time to find the best mRNA vaccine sequence was only 11 minutes. This method optimized the secondary structure stability and codon usage of mRNA vaccines. Animal experiments showed that the binding antibody level of SARS-CoV mRNA vaccines designed by this method was 128-fold higher than that of the control group.

In addition, the advancements in protein structure prediction models such as AlphaFold3⁸⁰ have significantly deepened our understanding of the three-dimensional structure of proteins. This knowledge provides a novel tool for

antigen recognition. The progress in de novo protein design methods have revolutionized structure-based vaccine design. Employing a cutting-edge top-down reinforcement learning-based approach, Lutz et al.⁸¹ successfully generated an ultra-compact icosahedral architecture that cannot be achieved by current design approaches. This unique structure enables the display of immunogens at exceptionally high densities, thereby enhancing vaccine responses.

AI can also be utilized to identify key determinants of vaccine hesitancy.⁸² This information is imperative to optimize vaccine differentiation and vaccination strategies, and to enhance the vaccination coverage rate.

PERIODIC EPIDEMICS: STUDIES ON PATHOGENIC MECHANISMS

After the initial outbreak, some infectious diseases gradually transition into a periodic epidemic state. The complex and dynamic interaction between pathogens and hosts is one of the most important factors that determine whether an infectious disease outbreak will occur again. As the number of susceptible individuals within a population surpasses a certain level, the disease will break out again. Therefore, revealing the mechanisms of pathogen-host interactions is crucial for the development of new preventive, diagnostic and therapeutic strategies. With the advances in high-throughput and quantitative measurement technologies, large-scale genome and transcriptome data can be collected and analyzed. These data could be utilized to characterize and model the complex molecular networks underlying infection dynamics, thereby providing novel insights into infectious disease pathogenesis.

Host susceptibility

For infectious diseases, host genetics plays a crucial role in affecting disease susceptibility, alongside pathogen genetics. The primary approach for investigating host susceptibility is the genome-wide association study (GWAS). These studies integrate statistical analysis and machine learning methods such as principal component analysis (PCA) and regression models. GWAS has played a pivotal role in identifying susceptibility loci for infectious diseases such as malaria,⁸³ tuberculosis,^{84,85} and HIV.⁸⁶ During the COVID-19 pandemic, researchers have utilized GWAS to identify loci associated with COVID-19 susceptibility, such as ABO blood group,⁸⁷ ACE2,⁸⁸ IFN.⁸⁹ These findings have provided valuable insights into the biological mechanisms underlying COVID-19 susceptibility and severity.

An important application of GWAS results is the construction of polygenic risk score (PRS) models for predicting infection risk. To improve the accuracy and clinical utility of PRS models, researchers are exploring the integration of AI methods, such as Bayesian regression,⁹⁰ convolutional neural networks (CNNs),⁹¹ and variational autoencoders.⁹² Despite the promise of AI-enhanced PRS models, several challenges remain. These include a lack of diversity in GWAS datasets, the need for rigorous validation in independent cohorts, and the requirement to integrate PRSs with other risk factors.

Infection and pathogenic mechanisms

The Understanding of disease mechanisms is crucial for the diagnosis and management of infectious diseases. The interaction between pathogen infection and the host's immune response affects disease severity, progression and outcome. The perturbation in host immunity is thought to be a key factor in severe symptoms. Single-cell RNA sequencing (scRNA-seq) technology enables simultaneous quantification of thousands of single-cell gene expression profiles. This powerful tool has been widely used to analyze the host immune responses. It has been extensively employed in investigating the pathogenesis of infectious diseases such as severe dengue⁹³ and ebola virus disease.⁹⁴

The scRNA-seq method includes typical steps such as data acquisition, data preprocessing (including representation learning and dimensionality reduction), and subsequent data analysis (involving cell clustering, cell annotation, trajectory inference). These steps integrate various machine learning techniques. (Figure 5B) During the COVID-19 pandemic, single-cell data analysis has been employed to investigate the pathogenesis of COVID-19.^{95,96} In a typical study, Ren et al.⁹⁷ proposed that megakaryocyte and monocyte subsets may contribute to cytokine storm in COVID-19 patients. Their study was conducted based on a single-cell sequencing dataset comprising COVID-

19 patients with varying ages, sexes, symptom severities, and disease progression stages. This study revealed numerous immunological features associated with COVID-19. Moreover, by integrating multi-omics datasets, a comprehensive and in-depth understanding of disease mechanisms can be achieved. Stephenson et al.⁹⁸ used Leiden clustering as well as uniform manifold approximation and projection (UMAP) visualization to reveal a coordinated immune response leading to COVID-19 pathogenesis.

Although scRNA-seq technology has emerged as a mature tool for uncovering cellular heterogeneity, several challenges remain. ScRNA-seq data often exhibit high dimensionality, sparsity, and heterogeneity. Additionally, the influence of data noise and batch effects becomes non-negligible when comparing different samples. Unlike traditional dimensionality reduction methods, deep learning approaches eliminate the need to define or display similarity or distance measures in the original data space and instead uncover the underlying structure in high-dimensional data through learned representations.⁹⁹ This provides a novel approach to address limitations in single-cell data analysis. Autoencoders (AE) have been widely used to capture features and enhance signal-to-noise ratios for precise cell-type clustering, batch correction, and gene imputation in single-cell studies. In a typical study, Amodio et al.¹⁰⁰ applied an autoencoder to a dataset of 11 million T cells from 180 dengue patients and identified cluster-based signatures of acute dengue infection as well as stratified immune responses to dengue. To address the batch effect and online integration of single-cell data, Xiong et al.¹⁰¹ developed a variational autoencoder (VAE)-based model, SCALEX. This model can project different batches of single-cell data into a batch invariant common cell embedding space by generalized projection. The integration performance of the SCALEX model was also validated based on single-cell data from nine typical COVID-19 studies. Enhancing model interpretability is crucial for unlocking the full potential of rich single-cell datasets and translating transcriptional profiles into valuable biological insights.¹⁰² Møller et al.¹⁰³ proposed a joint clustering tool, which combines linear and nonlinear non-negative matrix factorization (NMF), to achieve batch integration of different datasets and improve the interpretability of the model.

CHALLENGES AND FUTURE DIRECTIONS

Despite the promising applications of AI in infectious disease prevention and control, significant challenges remain in terms of data availability, dimensional complexities, model interpretability, privacy concerns, and security considerations.

Data availability and quality

The lack of large-scale, representative, and well-curated datasets is a major bottleneck for developing robust AI models (Figure 6A). Many infectious diseases, particularly neglected tropical diseases, suffer from limited data due to inadequate surveillance and diagnostic infrastructure. Even for diseases with abundant data, such as COVID-19, the data often lacks standardization, annotation, and integration across different sources and modalities. Inconsistent data standards and missing annotation information result in poor data interoperability and reusability. Moreover, data quality issues such as noise, bias, and heterogeneity pose significant challenges for AI model training and evaluation. Coordinated efforts from academia, government, and industry are needed to establish open data repositories, harmonize data standards, and incentivize data sharing. The COVID Moonshot Alliance exemplifies how open science propels early-stage pharmaceutical exploration for infectious diseases.¹⁰⁴ Adopting FAIR (Findable, Accessible, Interoperable, and Reusable) principles and implementing secure data access protocols can maximize data sharing while protecting privacy and confidentiality and accelerate technological innovation.

Curse of dimensionality

The high dimensionality and sparsity of biological data pose computational and statistical challenges for AI models (Figure 6B). Omics data, such as genomic, transcriptomic, and proteomic data, often have a large number of features (e.g., genes, proteins) compared to the number of samples, leading to the "curse of dimensionality". This can result in data sparsity, multicollinearity, multiple testing, and overfitting issues. The exponential growth of potential combinations of features (e.g., genotypes, mutations) further exacerbates

the dimensionality problem. This might make it computationally intractable to exhaustively search the feature space. Traditional statistical methods that rely on distance-based measures become less effective in high-dimensional spaces. To address this, researchers can combine domain knowledge with advanced feature selection and representation learning techniques. Dimensionality reduction methods, such as PCA, t-distributed stochastic neighbor embedding (t-SNE), and UMAP, can be used to extract low-dimensional, informative representations of the data. However, the choice of dimensionality reduction method and hyperparameters requires careful consideration and validation. Rigorous benchmarking and validation using independent datasets are crucial for assessing model generalizability and avoiding overfitting.

Model interpretability and trustworthiness

The black-box nature of many AI models, particularly deep learning models, hinders their interpretability and trustworthiness, which are critical for high-stake applications like infectious disease diagnosis and policy making (Figure 6C). Deep learning models often have millions of parameters and complex architectures that make it difficult to understand how they arrive at specific predictions. The lack of interpretability not only limits the adoption of AI models by domain experts but also poses risks of biased or erroneous predictions that can have severe consequences in public health settings. Developing explainable AI techniques that provide clear and actionable insights into model decisions is an active area of research. Post-hoc interpretation methods, such as attention mechanisms, feature relevance analysis, and causal inference, can be used to illuminate the inner workings of deep learning models. Graph neural networks, which can capture the structural and relational information in biological data, tend to be more interpretable than other deep learning architectures. Engaging domain experts and end-users in the model development and evaluation process can improve transparency and foster trust. Establishing multidisciplinary teams and implementing responsible AI principles, including fairness, accountability, and transparency, are key to ensure the equitable and beneficial deployment of AI technologies.

Privacy and security issues

The use of AI in infectious disease contexts raises important privacy and security issues (Figure 6D). AI models trained on biased or unrepresentative data may perpetuate or amplify existing health disparities and social inequities. The opaque nature of many AI models can lead to unfair or discriminatory decisions that disproportionately affect vulnerable populations. The collection and use of sensitive health data for AI model development also raise concerns about privacy, confidentiality, and data sovereignty. The potential misuse or dual-use of AI technologies for malicious purposes, such as bioweapon development or misinformation campaigns, poses significant biosecurity risks. Establishing clear data governance frameworks, security protocols, and ethical guidelines is critical for the responsible development and deployment of AI in infectious disease research and response. Engaging diverse stakeholders, including public health professionals, policymakers, civil society organizations, and affected communities, in the design and oversight of AI systems can help ensure their alignment with societal values and priorities.

Future directions

The application of AI methods provides innovative tools for addressing infectious diseases. One potential future direction is to explore how biological knowledge can be more effectively integrated into AI algorithms in order to enhance the interpretability and predictive performance of models. Taking the prediction of pathogen evolution as an example, in the mathematical space, the possibility of pathogen sequence change is inexhaustible. However, in practical scenarios, the evolution process of a pathogen is typically constrained by multiple factors including environmental conditions, host characteristics, and its own sequence and structural change potential.²⁵ Therefore, it is crucial to capture the underlying laws governing pathogen evolution in order to accurately comprehend its trajectory. From a methodological perspective, capturing the intrinsic correlations among biological characteristics of pathogens shares similarities and transferability with

language models that learn the inherent logic of languages. This suggests that future AI models are expected to assist us in addressing the challenging problem of determining the direction of pathogen evolution by defining potential evolutionary constraints.

In addition, the highly reusable and deeply personalized nature of the AI approaches makes them exceptionally well-suited to meet the demands of precision medicine. Given the exponential accumulation of genomic data and the rapid decline in computing power costs, personalized diagnosis and treatment based on individual characteristics are increasingly within reach.⁹⁶ Consequently, the widespread applications of AI in infectious disease precision medicine, including personalized diagnosis and tailored treatment development, could potentially herald groundbreaking advancements.

At the macro level, AI can also contribute to our understanding and mitigation of the broader societal impact caused by infectious diseases. By integrating AI with participatory approaches and community-based research, it can be ensured that the solutions are culturally sensitive, context-specific, and aligned with local needs and priorities.

The large language models (LLMs) are the most exciting breakthrough in AI technology in recent years. Some preliminary practices on LLMs have shown great potential for their applications in the domain of infectious diseases.

One potential avenue is to leverage LLMs for optimizing the clinical management of infectious diseases. These LLMs can be utilized for enhancing patient communication and education,¹⁰⁵ thereby improving patients' comprehension of disease knowledge and subsequently enhancing their adherence to treatment protocols. Moreover, LLMs can also aid in summarizing clinical texts,¹⁰⁶ facilitating diagnosis,¹⁰⁷ evaluating medication efficacy,¹⁰⁸ and predicting clinical outcomes,¹⁰⁹ which can effectively address the challenge of limited physician resources during a pandemic outbreak. In a recent study by Lu et al.,¹¹⁰ they proposed a universal multimodal AI assistant for pathology that achieves an accuracy rate of 89.5% and 78.7% on multiple-choice diagnostic questions and open-ended question answering respectively, showcasing the immense potential of collaborative human-machine diagnosis.

Another promising avenue is the utilization of LLMs as a versatile and potent research tool for scientific research in infectious diseases. Boiko et al.¹¹¹ have demonstrated the viability of employing intelligent agents based on LLMs to decompose, plan, and execute intricate scientific tasks proficiently. Moreover, the efficient compression capability of LLMs in training data proves particularly suitable for processing exceedingly lengthy sequence data such as RNA, DNA, and proteins. Taking single-cell omics data as an example, LLMs can effectively extract pivotal biological insights about genes and cells from multi-omics data sets and subsequently accomplish diverse downstream tasks including multi-batch integration, multi-omics integration, and gene network inference.¹¹²

CONCLUSION

Despite the advancements in medicine in the 21st century that have led to improvements in population health, humanity continues to face a constant threat from infectious diseases. A changing world requires changing scientific tools to assess the future risk of infectious diseases. In an increasingly interconnected world, the risk of infectious diseases is shared globally. Applications of AI tools during the COVID-19 pandemic have highlighted the effectiveness in infectious disease research. Therefore, future efforts should focus on establishing robust infectious disease surveillance systems that support cross-agency data sharing within and between countries, thereby expediting the application of AI tools in infectious disease response. By harnessing AI capabilities, scientifically grounded models can be developed for specific infectious diseases, generating results that are applicable in the real world. These modeling outcomes can help reveal key gaps in our understanding of infectious diseases and guide the development of surveillance, diagnosis, and control measures. AI tools offer a viable and cost-effective way to identify influential measures prior to implementing decisions. By employing all of these methods, the precision and targeting of responses to infectious diseases can be enhanced, thereby reducing the disease burden and improving global health.

REFERENCES

1. Msemburi, W., Karlinsky, A., Knutson, V., et al. (2023). The WHO estimates of excess mortality associated with the COVID-19 pandemic. *Nature* **613**: 130–137. DOI: 10.1038/s41586-022-05522-2.
2. WHO. (2024). Prioritizing diseases for research and development in emergency contexts. <https://www.who.int/activities/prioritizing-diseases-for-research-and-development-in-emergency-contexts>.
3. Brownstein, J.S., Rader, B., Astley, C.M., et al. (2023). Advances in artificial intelligence for infectious-disease surveillance. *N. Engl. J. Med.* **388**: 1597–1607. DOI: 10.1056/NEJMra2119215.
4. Wong, F., de la Fuente-Nunez, C., and Collins, J.J. (2023). Leveraging artificial intelligence in the fight against infectious diseases. *Science* **381**: 164–170. DOI: 10.1126/science.adh1114.
5. Topol, E.J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nat. Med.* **25**: 44–56. DOI: 10.1038/s41591-018-0300-7.
6. Aiello, A.E., Renson, A., and Zivich, P.N. (2020). Social media- and internet-based disease surveillance for public health. *Annu. Rev. Public Health* **41**: 101–118. DOI: 10.1146/annurev-publhealth-040119-094402.
7. Ginsberg, J., Mohebbi, M.H., Patel, R.S., et al. (2009). Detecting influenza epidemics using search engine query data. *Nature* **457**: 1012–1014. DOI: 10.1038/nature07634.
8. Lazer, D., Kennedy, R., King, G., et al. (2014). The parable of Google Flu: Traps in big data analysis. *Science* **343**: 1203–1205. DOI: 10.1126/science.1248506.
9. Kogan, N.E., Clemente, L., Liautaud, P., et al. (2021). An early warning approach to monitor COVID-19 activity with multiple digital traces in near real time. *Sci. Adv.* **7**: eabd6989. DOI: 10.1126/sciadv.abd6989.
10. Stolerman, L.M., Clemente, L., Poirier, C., et al. (2023). Using digital traces to build prospective and real-time county-level early warning systems to anticipate COVID-19 outbreaks in the United States. *Sci. Adv.* **9**: eabq0199. DOI: 10.1126/sciadv.abq0199.
11. Jahn, K., Dreifuss, D., Topolsky, I., et al. (2022). Early detection and surveillance of SARS-CoV-2 genomic variants in wastewater using COJAC. *Nat. Microbiol.* **7**: 1151–1160. DOI: 10.1038/s41564-022-01185-x.
12. Chia, P.Y., Coleman, K.K., Tan, Y.K., et al. (2020). Detection of air and surface contamination by SARS-CoV-2 in hospital rooms of infected patients. *Nat. Commun.* **11**: 2800. DOI: 10.1038/s41467-020-16670-2.
13. Doremalen, N.V., Bushmaker, T., Morris, D.H., et al. (2020). Aerosol and surface stability of SARS-CoV-2 as compared with SARS-CoV-1. *N. Engl. J. Med.* **382**: 1564–1567. DOI: 10.1056/NEJMc2004973.
14. Karthikeyan, S., Levy, J.I., De Hoff, P., et al. (2022). Wastewater sequencing reveals early cryptic SARS-CoV-2 variant transmission. *Nature* **609**: 101–108. DOI: 10.1038/s41586-022-05049-6.
15. Jones, K.E., Patel, N.G., Levy, M.A., et al. (2008). Global trends in emerging infectious diseases. *Nature* **451**: 990–993. DOI: 10.1038/nature06536.
16. Carroll, D., Daszak, P., Wolfe, N.D., et al. (2018). The global virome project. *Science* **359**: 872–874. DOI: 10.1126/science.aap7463.
17. Grange, Z.L., Goldstein, T., Johnson, C.K., et al. (2021). Ranking the risk of animal-to-human spillover for newly discovered viruses. *Proc. Natl. Acad. Sci. U. S. A.* **118**: e2002324118. DOI: 10.1073/pnas.2002324118.
18. Wardeh, M., Blagrove, M.S.C., Sharkey, K.J., et al. (2021). Divide-and-conquer: Machine-learning integrates mammalian and viral traits with network features to predict virus-mammal associations. *Nat. Commun.* **12**: 3954. DOI: 10.1038/s41467-021-24085-w.
19. Levy, J.I., Andersen, K.G., Knight, R., et al. (2023). Wastewater surveillance for public health. *Science* **379**: 26–27. DOI: 10.1126/science.ade2503.
20. Luksza, M. and Lassig, M. (2014). A predictive fitness model for influenza. *Nature* **507**: 57–61. DOI: 10.1038/nature13087.
21. Shu, Y. and McCauley, J. (2017). GISAID: Global initiative on sharing all influenza data—from vision to reality. *Eurosurveillance* **22**: 30494. DOI: 10.2807/1560-7917.ES.2017.22.13.30494.
22. Obermeyer, F., Jankowiak, M., Barkas, N., et al. (2022). Analysis of 6.4 million SARS-CoV-2 genomes identifies mutations associated with fitness. *Science* **376**: 1327–1332. DOI: 10.1126/science.abm1208.
23. Maher, M.C., Bartha, I., Weaver, S., et al. (2022). Predicting the mutational drivers of future SARS-CoV-2 variants of concern. *Sci. Transl. Med.* **14**: eabk3445. DOI: 10.1126/scitranslmed.abk3445.
24. Hie, B., Zhong, E.D., Berger, B., et al. (2021). Learning the language of viral evolution and escape. *Science* **371**: 284–288. DOI: 10.1126/science.abd7331.
25. Thadani, N.N., Gurev, S., Notin, P., et al. (2023). Learning from prepandemic data to forecast viral escape. *Nature* **622**: 818–825. DOI: 10.1038/s41586-023-06617-0.
26. Brauner, J.M., Mindermann, S., Sharma, M., et al. (2021). Inferring the effectiveness of government interventions against COVID-19. *Science* **371**: eabd9338. DOI: 10.1126/science.abd9338.
27. Chang, S., Pierson, E., Koh, P.W., et al. (2021). Mobility network models of COVID-19 explain inequities and inform reopening. *Nature* **589**: 82–87. DOI: 10.1038/s41586-020-2923-3.
28. Leung, K., Wu, J.T., and Leung, G.M. (2021). Real-time tracking and prediction of COVID-19 infection using digital proxies of population mobility and mixing. *Nat.*

- Commun. **12**: 1501. DOI: 10.1038/s41467-021-21776-2.
29. Stockdale, J.E., Susvitasari, K., Tupper, P., et al. (2023). Genomic epidemiology offers high resolution estimates of serial intervals for COVID-19. *Nat. Commun.* **14**: 4830. DOI: 10.1038/s41467-023-40544-y.
 30. Valdano, E., Colombi, D., Poletto, C., et al. (2023). Epidemic graph diagrams as analytics for epidemic control in the data-rich era. *Nat. Commun.* **14**: 8472. DOI: 10.1038/s41467-023-43856-1.
 31. Cooper, B.S., Evans, S., Jafari, Y., et al. (2023). The burden and dynamics of hospital-acquired SARS-CoV-2 in England. *Nature* **623**: 132–138. DOI: 10.1038/s41586-023-06634-z.
 32. Ward, T., Johnsen, A., Ng, S., et al. (2022). Forecasting SARS-CoV-2 transmission and clinical risk at small spatial scales by the application of machine learning architectures to syndromic surveillance data. *Nat. Mach. Intell.* **4**: 814–827. DOI: 10.1038/s42256-022-00538-9.
 33. Nicholson, G., Lehmann, B., Padellini, T., et al. (2022). Improving local prevalence estimates of SARS-CoV-2 infections using a causal debiasing framework. *Nat. Microbiol.* **7**: 97–107. DOI: 10.1038/s41564-021-01029-0.
 34. Yang, W. and Shaman, J. (2021). Development of a model-inference system for estimating epidemiological characteristics of SARS-CoV-2 variants of concern. *Nat. Commun.* **12**: 5573. DOI: 10.1038/s41467-021-25913-9.
 35. Lloyd-Smith, J.O., Schreiber, S.J., Kopp, P.E., et al. (2005). Superspreading and the effect of individual variation on disease emergence. *Nature* **438**: 355–359. DOI: 10.1038/nature04153.
 36. Lemieux, J.E., Siddle, K.J., Shaw, B.M., et al. (2021). Phylogenetic analysis of SARS-CoV-2 in Boston highlights the impact of superspreading events. *Science* **371**: eabe3261. DOI: 10.1126/science.abe3261.
 37. Lau, M.S.Y., Grenfell, B., Thomas, M., et al. (2020). Characterizing superspreading events and age-specific infectiousness of SARS-CoV-2 transmission in Georgia, USA. *Proc. Natl. Acad. Sci. U. S. A.* **117**: 22430–22435. DOI: 10.1073/pnas.2011802117.
 38. Gomes, B. and Ashley, E.A. (2023). Artificial intelligence in molecular medicine. *N. Engl. J. Med.* **388**: 2456–2465. DOI: 10.1056/NEJMr2204787.
 39. Sadybekov, A.V. and Katritch, V. (2023). Computational approaches streamlining drug discovery. *Nature* **616**: 673–685. DOI: 10.1038/s41586-023-05905-z.
 40. Mullowney, M.W., Duncan, K.R., Elsayed, S.S., et al. (2023). Artificial intelligence for natural product drug discovery. *Nat. Rev. Drug Discov.* **22**: 895–916. DOI: 10.1038/s41573-023-00774-7.
 41. Pandey, M., Fernandez, M., Gentile, F., et al. (2022). The transformational role of GPU computing and deep learning in drug discovery. *Nat. Mach. Intell.* **4**: 211–221. DOI: 10.1038/s42256-022-00463-x.
 42. Turbe, V., Herbst, C., Mngomezulu, T., et al. (2021). Deep learning of HIV field-based rapid tests. *Nat. Med.* **27**: 1165–1170. DOI: 10.1038/s41591-021-01384-9.
 43. Huang, S.C., Chaudhari, A.S., Langlotz, C.P., et al. (2022). Developing medical imaging AI for emerging infectious diseases. *Nat. Commun.* **13**: 7060. DOI: 10.1038/s41467-022-34234-4.
 44. Rajpurkar, P. and Lungren, M.P. (2023). The current and future state of AI interpretation of medical images. *N. Engl. J. Med.* **388**: 1981–1990. DOI: 10.1056/NEJMr2301725.
 45. Jin, C., Chen, W., Cao, Y., et al. (2020). Development and evaluation of an artificial intelligence system for COVID-19 diagnosis. *Nat. Commun.* **11**: 5088. DOI: 10.1038/s41467-020-18685-1.
 46. Mei, X., Lee, H.-C., Diao, K.-y., et al. (2020). Artificial intelligence-enabled rapid diagnosis of patients with COVID-19. *Nat. Med.* **26**: 1224–1228. DOI: 10.1038/s41591-020-0931-3.
 47. Dvijotham, K.D., Winkens, J., Barsbey, M., et al. (2023). Enhancing the reliability and accuracy of AI-enabled diagnosis via complementarity-driven deferral to clinicians. *Nat. Med.* **29**: 1814–1820. DOI: 10.1038/s41591-023-02437-x.
 48. Roberts, M., Driggs, D., Thorpe, M., et al. (2021). Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nat. Mach. Intell.* **3**: 199–217. DOI: 10.1038/s42256-021-00307-0.
 49. Bai, X., Wang, H., Ma, L., et al. (2021). Advancing COVID-19 diagnosis with privacy-preserving collaboration in artificial intelligence. *Nat. Mach. Intell.* **3**: 1081–1089. DOI: 10.1038/s42256-021-00421-z.
 50. Ktena, I., Wiles, O., Albuquerque, I., et al. (2024). Generative models improve fairness of medical classifiers under distribution shifts. *Nat. Med.* **30**: 1166–1173. DOI: 10.1038/s41591-024-02838-6.
 51. Liang, W., Yao, J., Chen, A., et al. (2020). Early triage of critically ill COVID-19 patients using deep learning. *Nat. Commun.* **11**: 3543. DOI: 10.1038/s41467-020-17280-8.
 52. Hoertel, N., Blachier, M., Blanco, C., et al. (2020). A stochastic agent-based model of the SARS-CoV-2 epidemic in France. *Nat. Med.* **26**: 1417–1421. DOI: 10.1038/s41591-020-1001-6.
 53. Schwab, P., Mehrjou, A., Parbhoo, S., et al. (2021). Real-time prediction of COVID-19 related mortality using electronic health records. *Nat. Commun.* **12**: 1058. DOI: 10.1038/s41467-020-20816-7.
 54. Gao, J., Heintz, J., Mack, C., et al. (2023). Evidence-driven spatiotemporal COVID-19 hospitalization prediction with Ising dynamics. *Nat. Commun.* **14**: 3093. DOI: 10.1038/s41467-023-38756-3.
 55. Tomasev, N., Harris, N., Baur, S., et al. (2021). Use of deep learning to develop continuous-risk models for adverse event prediction from electronic health records. *Nat. Protoc.* **16**: 2765–2787. DOI: 10.1038/s41596-021-00513-5.
 56. Gao, Y., Cai, G.Y., Fang, W., et al. (2020). Machine learning based early warning system enables accurate mortality risk prediction for COVID-19. *Nat. Commun.* **11**: 5033. DOI: 10.1038/s41467-020-18684-2.
 57. Dayan, I., Roth, H.R., Zhong, A., et al. (2021). Federated learning for predicting clinical outcomes in patients with COVID-19. *Nat. Med.* **27**: 1735–1743. DOI: 10.1038/s41591-021-01506-3.
 58. Lassau, N., Ammari, S., Chouzenoux, E., et al. (2021). Integrating deep learning CT-scan model, biological and clinical variables to predict severity of COVID-19 patients. *Nat. Commun.* **12**: 634. DOI: 10.1038/s41467-020-20657-4.
 59. Yan, L., Zhang, H.-T., Goncalves, J., et al. (2020). An interpretable mortality prediction model for COVID-19 patients. *Nat. Mach. Intell.* **2**: 283–288. DOI: 10.1038/s42256-020-0180-7.
 60. Devaux, Y., Zhang, L., Lumley, A.I., et al. (2024). Development of a long noncoding RNA-based machine learning model to predict COVID-19 in-hospital mortality. *Nat. Commun.* **15**: 4259. DOI: 10.1038/s41467-024-47557-1.
 61. Lauritsen, S.M., Kristensen, M., Olsen, M.V., et al. (2020). Explainable artificial intelligence model to predict acute critical illness from electronic health records. *Nat. Commun.* **11**: 3852. DOI: 10.1038/s41467-020-17431-x.
 62. Thompson, E.J., Williams, D.M., Walker, A.J., et al. (2022). Long COVID burden and risk factors in 10 UK longitudinal studies and electronic health records. *Nat. Commun.* **13**: 3528. DOI: 10.1038/s41467-022-30836-0.
 63. Zang, C., Zhang, Y., Xu, J., et al. (2023). Data-driven analysis to understand long COVID using electronic health records from the RECOVER initiative. *Nat. Commun.* **14**: 1948. DOI: 10.1038/s41467-023-37653-z.
 64. Zhang, H., Zang, C., Xu, Z., et al. (2023). Data-driven identification of post-acute SARS-CoV-2 infection subphenotypes. *Nat. Med.* **29**: 226–235. DOI: 10.1038/s41591-022-02116-3.
 65. Ferretti, L., Wymant, C., Kendall, M., et al. (2020). Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. *Science* **368**: eabb6936. DOI: 10.1126/science.abb6936.
 66. Kucharski, A.J., Klepac, P., Conlan, A.J.K., et al. (2020). Effectiveness of isolation, testing, contact tracing, and physical distancing on reducing transmission of SARS-CoV-2 in different settings: A mathematical modelling study. *Lancet Infect. Dis.* **20**: 1151–1160. DOI: 10.1016/S1473-3099(20)30457-6.
 67. Kendall, M., Tsallis, D., Wymant, C., et al. (2023). Epidemiological impacts of the NHS COVID-19 app in England and Wales throughout its first year. *Nat. Commun.* **14**: 858. DOI: 10.1038/s41467-023-36495-z.
 68. Wymant, C., Ferretti, L., Tsallis, D., et al. (2021). The epidemiological impact of the NHS COVID-19 app. *Nature* **594**: 408–412. DOI: 10.1038/s41586-021-03606-z.
 69. Ferretti, L., Wymant, C., Petrie, J., et al. (2024). Digital measurement of SARS-CoV-2 transmission risk from 7 million contacts. *Nature* **626**: 145–150. DOI: 10.1038/s41586-023-06952-2.
 70. Kretzschmar, M.E., Rozhnova, G., Bootsma, M.C.J., et al. (2020). Impact of delays on effectiveness of contact tracing strategies for COVID-19: A modelling study. *Lancet Public Health* **5**: e452–e459. DOI: 10.1016/S2468-2667(20)30157-2.
 71. Colizza, V., Grill, E., Mikolajczyk, R., et al. (2021). Time to evaluate COVID-19 contact-tracing apps. *Nat. Med.* **27**: 361–362. DOI: 10.1038/s41591-021-01236-6.
 72. Pandit, J.A., Radin, J.M., Quer, G., et al. (2022). Smartphone apps in the COVID-19 pandemic. *Nat. Biotechnol.* **40**: 1013–1022. DOI: 10.1038/s41587-022-01350-x.
 73. Flaxman, S., Mishra, S., Gandy, A., et al. (2020). Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature* **584**: 257–261. DOI: 10.1038/s41586-020-2405-7.
 74. Kraemer, M.U.G., Yang, C.-H., Gutierrez, B., et al. (2020). The effect of human mobility and control measures on the COVID-19 epidemic in China. *Science* **368**: 493–497. DOI: 10.1126/science.abb4218.
 75. Candido, D.S., Claro, I.M., de Jesus, J.G., et al. (2020). Evolution and epidemic spread of SARS-CoV-2 in Brazil. *Science* **369**: 1255–1260. DOI: 10.1126/science.abd2161.
 76. Haug, N., Geyrhofer, L., Londei, A., et al. (2020). Ranking the effectiveness of worldwide COVID-19 government interventions. *Nat. Hum. Behav.* **4**: 1303–1312. DOI: 10.1038/s41562-020-01009-0.
 77. Ge, Y., Wu, X., Zhang, W., et al. (2023). Effects of public-health measures for zeroing out different SARS-CoV-2 variants. *Nat. Commun.* **14**: 5270. DOI: 10.1038/s41467-023-40940-4.
 78. Li, Y., Campbell, H., Kulkarni, D., et al. (2021). The temporal association of introducing and lifting non-pharmaceutical interventions with the time-varying reproduction number (R) of SARS-CoV-2: A modelling study across 131 countries. *Lancet Infect. Dis.* **21**: 193–202. DOI: 10.1016/S1473-3099(20)30785-4.
 79. Zhang, H., Zhang, L., Lin, A., et al. (2023). Algorithm for optimized mRNA design improves stability and immunogenicity. *Nature* **621**: 396–403. DOI: 10.1038/s41586-023-06127-z.
 80. Abramson, J., Adler, J., Dunger, J., et al. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**: 493–500. DOI: 10.1038/s41586-024-07487-w.
 81. Lutz, I.D., Wang, S., Norn, C., et al. (2023). Top-down design of protein architectures with reinforcement learning. *Science* **380**: 266–273. DOI: 10.1126/science.adf6591.

82. Loomba, S., de Figueiredo, A., Piatek, S.J., et al. (2021). Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nat. Hum. Behav.* **5**: 337–348. DOI: 10.1038/s41562-021-01056-1.
83. Band, G., Le, Q.S., Clarke, G.M., et al. (2019). Insights into malaria susceptibility using genome-wide data on 17,000 individuals from Africa, Asia and Oceania. *Nat. Commun.* **10**: 5732. DOI: 10.1038/s41467-019-13480-z.
84. Sveinbjornsson, G., Gudbjartsson, D.F., Halldorsson, B.V., et al. (2016). HLA class II sequence variants influence tuberculosis risk in populations of European ancestry. *Nat. Genet.* **48**: 318–322. DOI: 10.1038/ng.3498.
85. Zheng, R., Li, Z., He, F., et al. (2018). Genome-wide association study identifies two risk loci for tuberculosis in Han Chinese. *Nat. Commun.* **9**: 4072. DOI: 10.1038/s41467-018-06539-w.
86. The International HIV Controllers Study. (2010). The major genetic determinants of HIV-1 control affect HLA class I peptide presentation. *Science* **330**: 1551–1557. DOI: 10.1126/science.1195271.
87. Niemi, M.E.K., Karjalainen, J., Liao, R.G., et al. (2021). Mapping the human genetic architecture of COVID-19. *Nature* **600**: 472–477. DOI: 10.1038/s41586-021-03767-x.
88. Horowitz, J.E., Kosmicki, J.A., Damask, A., et al. (2022). Genome-wide analysis provides genetic evidence that ACE2 influences COVID-19 risk and yields risk scores associated with severe disease. *Nat. Genet.* **54**: 382–392. DOI: 10.1038/s41588-021-01006-7.
89. Zhang, Q., Bastard, P., Liu, Z., et al. (2020). Inborn errors of type I IFN immunity in patients with life-threatening COVID-19. *Science* **370**: eabd4570. DOI: 10.1126/science.abd4570.
90. Ge, T., Chen, C.Y., Ni, Y., et al. (2019). Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nat. Commun.* **10**: 1776. DOI: 10.1038/s41467-019-09718-5.
91. Fizev, P.P., McRae, J., Ulirsch, J.C., et al. (2023). Rare penetrant mutations confer severe risk of common diseases. *Science* **380**: eabo1131. DOI: 10.1126/science.abo1131.
92. Frazer, J., Notin, P., Dias, M., et al. (2021). Disease variant prediction with deep generative models of evolutionary data. *Nature* **599**: 91–95. DOI: 10.1038/s41586-021-04043-8.
93. Zanini, F., Robinson, M.L., Croote, D., et al. (2018). Virus-inclusive single-cell RNA sequencing reveals the molecular signature of progression to severe dengue. *Proc. Natl. Acad. Sci. U. S. A.* **115**: E12363–E12369. DOI: 10.1073/pnas.1813819115.
94. Kotliar, D., Lin, A.E., Logue, J., et al. (2020). Single-cell profiling of ebola virus disease in vivo reveals viral and host dynamics. *Cell* **183**: 1383–1401. DOI: 10.1016/j.cell.2020.10.002.
95. Bost, P., Giladi, A., Liu, Y., et al. (2020). Host-viral infection maps reveal signatures of severe COVID-19 patients. *Cell* **181**: 1475–1488. DOI: 10.1016/j.cell.2020.05.006.
96. Aquino, Y., Bisiaux, A., Li, Z., et al. (2023). Dissecting human population variation in single-cell responses to SARS-CoV-2. *Nature* **621**: 120–128. DOI: 10.1038/s41586-023-06422-9.
97. Ren, X., Wen, W., Fan, X., et al. (2021). COVID-19 immune features revealed by a large-scale single-cell transcriptome atlas. *Cell* **184**: 1895–1913. DOI: 10.1016/j.cell.2021.01.053.
98. Stephenson, E., Reynolds, G., Botting, R.A., et al. (2021). Single-cell multi-omics analysis of the immune response in COVID-19. *Nat. Med.* **27**: 904–916. DOI: 10.1038/s41591-021-01329-2.
99. Wang, J., Ma, A., Chang, Y., et al. (2021). scGNN is a novel graph neural network framework for single-cell RNA-Seq analyses. *Nat. Commun.* **12**: 1882. DOI: 10.1038/s41467-021-22197-x.
100. Amodio, M., van Dijk, D., Srinivasan, K., et al. (2019). Exploring single-cell data with deep multitasking neural networks. *Nat. Methods* **16**: 1139–1145. DOI: 10.1038/s41592-019-0576-7.
101. Xiong, L., Tian, K., Li, Y., et al. (2022). Online single-cell data integration through projecting heterogeneous datasets into a common cell-embedding space. *Nat. Commun.* **13**: 6118. DOI: 10.1038/s41467-022-33758-z.
102. Zhao, Y., Cai, H., Zhang, Z., et al. (2021). Learning interpretable cellular and gene signature embeddings from single-cell transcriptomic data. *Nat. Commun.* **12**: 5261. DOI: 10.1038/s41467-021-25534-2.
103. Moller, A.F. and Madsen, J.G.S. (2023). JOINTLY: Interpretable joint clustering of single-cell transcriptomes. *Nat. Commun.* **14**: 8473. DOI: 10.1038/s41467-023-44279-8.
104. Boby, M.L., Fearon, D., Ferla, M., et al. (2023). Open science discovery of potent noncovalent SARS-CoV-2 main protease inhibitors. *Science* **382**: eabo7201. DOI: 10.1126/science.abo7201.
105. Singhal, K., Azizi, S., Tu, T., et al. (2023). Large language models encode clinical knowledge. *Nature* **620**: 172–180. DOI: 10.1038/s41586-023-06291-2.
106. Van Veen, D., Van Uden, C., Blankemeier, L., et al. (2024). Adapted large language models can outperform medical experts in clinical text summarization. *Nat. Med.* **30**: 1134–1142. DOI: 10.1038/s41591-024-02855-5.
107. Han, T., Adams, L.C., Bressem, K.K., et al. (2024). Comparative analysis of multimodal large language model performance on clinical vignette questions. *JAMA* **331**: 1320–1321. DOI: 10.1001/jama.2023.27861.
108. Pais, C., Liu, J., Voigt, R., et al. (2024). Large language models for preventing medication direction errors in online pharmacies. *Nat. Med.* **30**: 1574–1582. DOI: 10.1038/s41591-024-02933-8.
109. Tayebi Arasteh, S., Han, T., Lotfinia, M., et al. (2024). Large language models streamline automated machine learning for clinical studies. *Nat. Commun.* **15**: 1603. DOI: 10.1038/s41467-024-45879-8.
110. Lu, M.Y., Chen, B., Williamson, D.F.K., et al. (2024). A multimodal generative AI copilot for human pathology. *Nature* **621**: 1076–1083. DOI: 10.1038/s41586-024-07618-3.
111. Boiko, D.A., MacKnight, R., Kline, B., et al. (2023). Autonomous chemical research with large language models. *Nature* **624**: 570–578. DOI: 10.1038/s41586-023-06792-0.
112. Cui, H., Wang, C., Maan, H., et al. (2024). scGPT: Toward building a foundation model for single-cell multi-omics using generative AI. *Nat. Methods* DOI: 10.1038/s41592-024-02201-0.

FUNDING AND ACKNOWLEDGMENTS

A.W. was supported by the National Key Research and Development Program (2021YFC2301300), the CAMS Innovation Fund for Medical Sciences (2021-I2M-1-061), the National Natural Science Foundation of China (92169106), and Non-profit Central Research Institute Fund of Chinese Academy of Medical Sciences (2021-PT180-001); Y. Li was supported by Soft Science Research Project of Department of Science and Technology of Zhejiang Province (2024C35035), National Key Research and Development Program of China (2022YFC3303103) and H.-Y.Z. was supported by the Natural Science Foundation of Jiangsu Province (BK20220278). Icons in all figures were originally created with BioRender.com and polished by editorial office. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

AUTHOR CONTRIBUTIONS

H.-Y. Z. and A. W. supervised and revised the manuscript. Y. L. and H.-Y. Z. wrote the original draft, H.-Y. Z. plot the original figures. Y. L. made the formal analysis. Y. L., J.-Y. L. and J. M. collected the materials. A. W., H.-Y. Z. and Y. L. acquired the fundings. H.-Y. Z. and Y. L. contributed equally to this manuscript. All authors contributed to the article and approved the submitted version.

DECLARATION OF INTERESTS

The authors declare no competing interests.

DATA AND CODE AVAILABILITY

Not applicable.

LEAD CONTACT WEBSITE

Aiping Wu: <https://www.ismsz.cn/Web/KXYJKYTDPage?Id=20&PageId=283>
Hang-Yu Zhou: <https://www.researchgate.net/profile/Hangyu-Zhou>