Journal of Translational Medicine

REVIEW Open Access

# Medical laboratory data-based models: opportunities, obstacles, and solutions

Jiaojiao Meng[1†], Moxin Wu[2†], Fangmin Shi[3,4], Ying Xie[1], Hui Wang[5] and You Guo[1,3,4*]

**Abstract**

Medical Laboratory Data (MLD) models, which combine artificial intelligence with big medical data, have great potential in disease screening, diagnosis, personalized medicine, and health management. This study thoroughly examines the opportunities, challenges, and solutions in this field. The use of large-scale MLD improves diagnostic accuracy and allows for real-time disease monitoring. Additionally, integrating social and environmental data enables the analysis of disease mechanisms and trends. Despite these benefits, challenges such as data quality, model optimization, computational requirements, and limited interpretability remain, along with concerns about data privacy, fairness, and security. Proposed solutions include establishing standardized data formats, utilizing deep learning frameworks, employing distributed computing, improving interpretability, and implementing techniques like federated learning and algorithm optimization to address bias and safeguard privacy. Future directions will focus on enhancing performance in specific scenarios, expanding applications across different domains, increasing transparency, enabling real-time processing, and building a supportive ecosystem. It is essential to strengthen policy oversight and promote collaboration among governments, medical institutions, and academia to ensure that technological advancements align with societal progress.

**Keywords** Medical laboratory data, Large models, Opportunities, Obstacles

## Introduction

The advancement of artificial intelligence and big data technologies is leading to the growing importance of predictive models based on MLD in medicine and healthcare [1–4]. These lab-data based models utilize deep learning, machine learning, and other technologies by combining vast health, disease, environment, and social data [5–7]. They show great promise in disease screening [8–10], diagnosis [11–14], personalized medicine [15–17], early warning systems [18–20], and health management [21–24]. Nevertheless, the broad adoption of these models in various possible cases encounters a range of obstacles concerning technology, privacy, and security. This article will systematically explore the following aspects: the characteristics of data sources for MLD models, their emerging value in medical practice, the technical challenges they face, privacy protection, fairness issues, and

†Jiaojiao Meng and Moxin Wu contributed equally to this work.

*Correspondence:
You Guo
gy@gmu.edu.cn
[1]The First School of Clinical Medicine, Gannan Medical University, Ganzhou 341000, China
[2]Department of Medical Laboratory, Affiliated Hospital of Jiujiang University, Jiujiang 332000, China
[3]Medical Big Data and Bioinformatics Research Centre, First Affiliated Hospital of Gannan Medical University, Ganzhou 341000, China
[4]School of Public Health and Health Management, Gannan Medical University, Ganzhou 341000, China
[5]State Key Laboratory of Oncogenes and Related Genes, Center for Single-Cell Omics, School of Public Health, School of Medicine, Shanghai Jiao Tong University, Shanghai 200025, China

© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

security considerations, as well as the key assessment factors that must be considered when evaluating the potential deployment of these models, and look ahead to their future development paths.

## Source and characteristics of MLD

A foundational understanding of the sources and characteristics of MLD is crucial for constructing robust, clinically relevant models, enhancing clinical diagnosis and treatment, and advancing modern medical research. The sources of MLD are diverse, reflecting the advancements in medical technology and research, as well as the multifaceted nature of patient health. For example, the novel biomarker t6A can accurately identify sepsis patients [25]. IWRAP device enables real-time monitoring of vital signs in patients with venous thromboembolism, allowing for assessment of life-threatening pulmonary embolism risk [26]. Such data offer valuable insights into clinical endpoints, aiding in accurate diagnosis, effective treatment, and timely intervention.

Firstly, as shown in Fig. 1, the sources of MLD are diverse, consisting mainly of clinical tests, laboratory biomolecular omics data and physiological monitoring from portable devices. Data from hospitals or clinics, where bodily fluid samples such as blood, urine, and exudates are examined to support disease diagnosis and treatment choices [27–29]. Laboratory biomolecular omics data, which is centered on domains like gene sequencing, metabolomics, and proteomics [4, 30–32], provides insights into the intricate molecular landscape in the human body during both wellness and illness [33]. Moreover, the rise of modern smart health devices and mobile medical tools has incorporated physiological monitoring data, such as blood glucose, heart rate, blood pressure, and sleep data [34–37], into the scope of medical laboratory data, further expanding the range of data sources.

Secondly, MLD is known for its multidimensionality, diverse formats, complexity, and dynamics (Fig. 1). Multidimensionality is evident in the wide array of data types, such as physiological indicators and molecular features, collected at a single time point and over the long term, as well as including test results and details of testing methods and instruments [38]. Diverse formats include quantitative data like complete blood counts [39], qualitative data such as microbiological culture results [24], image data like coagulation curves [40], and waveform data such as mass spectrometry graphs [41]. The complexity arises from the potential interconnections among the data and the biological variations among individuals [42], which are driven bygenetic backgrounds, lifestyle habits and comorbidities, and therefore account for the increased difficulty of analysis [43]. Furthermore, laboratory data displays unique time series features, including precise and systematic documentation of timestamps for sample collection, testing, and reporting, which facilitates the tracking of dynamic changes in patient test indicators [44]. Given these characteristics of MLD, advanced data management and integration becomes crucial. Specifically, the multidimensionality and diverse formats of MLD data, which often coexist at different temporal resolutions, necessitate sophisticated data harmonization and multi-modal analysis approaches. These efforts are essential to align and integrate data from various sources, such as quantitative data, image data, and waveform data, ensuring compatibility and meaningful analysis [38–41]. To support these processes, best practices in data formatting and annotation are vital. These include adopting standardized data formats (such as HL7 for clinical test results and FASTQ for omics data) to enhance interoperability and consistency [39, 41], providing detailed metadata annotations covering data provenance, collection methods, and quality metrics for each dataset, implementing rigorous data quality checks to identify and correct errors or inconsistencies in the data, and maintaining version control of datasets to track changes and ensure the reproducibility of analyses [38–41, 43]. Moreover, the complexity and individual variability of the data demand robust data management systems to handle inconsistencies and ensure data quality [42, 43]. Collectively, these measures ensure the usability and reliability of the data, laying a solid foundation for subsequent multi-modal analysis and applications of artificial intelligence. These characteristics not only enhance the potential value of lab data but also present new challenges for data analysis techniques.

The rapid accumulation of medical laboratory data is notably pronounced. This accumulation confers distinct advantages apparent because it facilitates data standardization through consistent and automated collection processes, supports rapid growth by providing a continuous stream of structured data, ensures timeliness for real-time clinical decision-making, enhances application value through reliable and actionable insights, and enables the potential for intelligence through advanced analytics and machine learning techniques. Collectively, these characteristics provide the essential foundation for big data analysis and predictive modeling, making automated analysis and intelligent interpretation more accessible through advanced techniques like machine learning and artificial intelligence [45].

## Rapidly accumulating MLD has generated new applications

The integration of medical lab data (MLD) with advanced modeling technologies, particularly artificial intelligence (AI), holds significant promise for enhancing diagnostic accuracy, enabling real-time disease monitoring, and improving patient outcomes. For example, the First

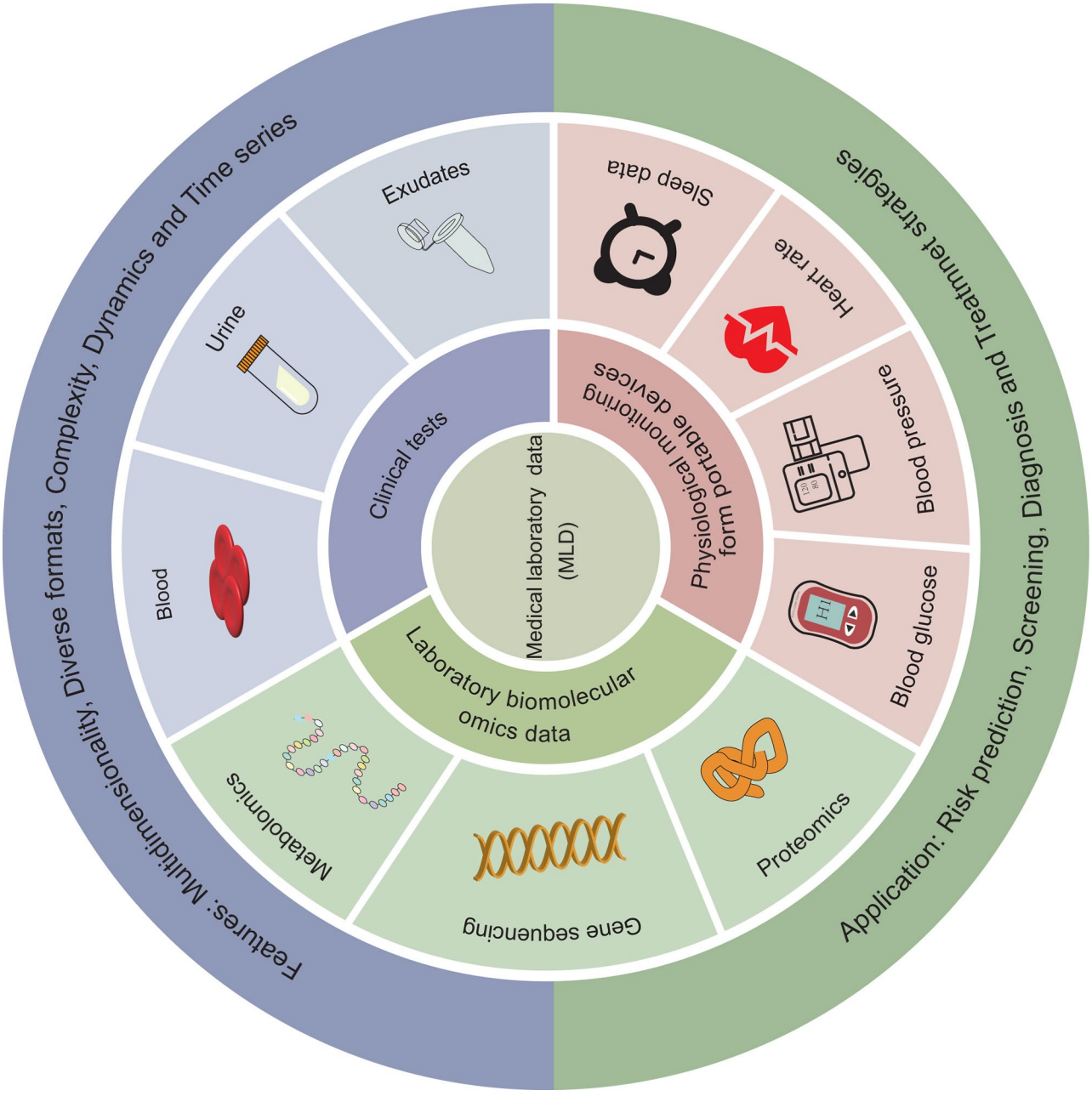



**Fig. 1** Sources, Features, and Applications of MLD. This figure outlines the different sources and characteristics of MLD, emphasizing their multidimensionality, diverse formats, complexity, dynamics, and time series features. It classifies MLD into clinical tests, laboratory biomolecular omics data and physiological monitoring from portable devices, demonstrating how these data types are used in risk prediction, screening, diagnosis, and treatment strategies in various medical fields.

Affiliated Hospital of Zhengzhou University successfully integrated MLD with the electronic health record (EHR) system to develop a model for early detection of sepsis. Trained on over 4,449 patient records, the model achieved a sensitivity of 87% and a specificity of 89%, significantly outperforming traditional methods [46]. Nevertheless, despite these technological advancements, there is currently a lack of a unified framework for applying artificial intelligence (AI) to clinical decision-making processes such as diagnosis. Although AI has demonstrated great potential from a technical standpoint, its deployment in clinical settings remains limited.

There is a significant difference between AI models that have been successfully deployed and those that are still in the development or testing phase. Deployed AI models must undergo rigorous clinical validation and regulatory approval processes to ensure their efficacy and safety. The process of clinical validation includes extensive testing to

confirm that the models perform as intended and do not introduce new risks to patient care [47]. Regulatory bodies, such as the FDA in the United States, require substantial evidence of a model's efficacy and safety before granting approval for clinical use. This includes validation studies, risk assessments, and sometimes even post-market surveillance [48]. Deploying these models in healthcare institutions not only requires technical integration with existing systems, but also necessitates consideration of ethical, legal, and social implications. Data privacy and security are of utmost importance, and ensuring that patient data is handled in compliance with regulations such as the General Data Protection Regulation (GDPR) and the Health Insurance Portability and Accountability Act (HIPAA) is essential. Additionally, adherence to the FAIR principles (Findability, Accessibility, Interoperability, Reusability) is required to ensure that data is usable and shareable in a responsible manner.

Our study focuses primarily on the theoretical aspects of integrating MLD with AI in healthcare. Through in-depth theoretical analysis, we aim to explore the potential opportunities, address the underlying challenges, and propose viable solutions. This approach is intended to provide valuable insights and guidance for future research and practice, and to facilitate the broader application of MLD and AI integration in healthcare.

#### Improved the accuracy and timeliness of diagnoses

Primarily, the aggregation of large-scale medical testing data has made it possible for artificial intelligence to assist in diagnosis, resulting in improved accuracy and timeliness of diagnoses. By integrating the analysis of biomarkers in patients' blood, such as circulating tumor DNA, tumor marker CA-125, and cardiac marker hs-TnI, with individual clinical phenotypes, the model can effectively predict the risk of cancer or endocrine diseases [30, 49, 50]and provide real-time monitoring and forecasting of complex diseases like cardiovascular(such as coronary artery disease and heart failure) and neurodegenerative disorders(such as Alzheimer's disease and Parkinson's disease) [51, 52]. This integrated approach significantly enhances the efficiency and accuracy of early screening for major diseases such as cancer and cardiovascular diseases [53], thereby improving diagnostic sensitivity and specificity, and effectively increasing patients' chances of recovery [54].

Additionally, AI models can rapidly implement real-time monitoring of complex disease changes by analyzing the dynamic alterations in the concentration of biomarkers in patients' blood. For example, the THEMIS platform identifies cancer by examining alterations in cfDNA levels and different epigenetic characteristics, including whole methylome sequencing data and fragmentation patterns [30]. It can also continuously monitor cfDNA in the bloodstream, enabling real-time monitoring of cancer advancement. This capability represents a significant step forward in the field of AI-driven diagnostics.

To further demonstrate the practical impact of these technological advancements, we conducted a comprehensive analysis of ovarian cancer diagnostic models based on blood tests. Table 1 provides a quantitative analysis of these models' performance, highlighting their accuracy, sensitivity, and specificity. Sensitivity refers to the proportion of true positives correctly identified by the model, specificity refers to the proportion of true negatives correctly identified, and AUC (Area Under the Curve) measures the model's ability to distinguish between positive and negative samples.From Table 1, it can be seen that the model by Medina, Jamie E. et al. achieved the highest sensitivity and specificity on both the training set (0.91 and 0.96, respectively) and the external validation set (0.89 and 0.94), indicating a high level of accuracy and discriminative power. However, it may require more complex detection techniques and higher costs. In contrast, the model by Katoh, Kanoko et al. showed excellent specificity (0.94), which is advantageous for reducing false positives, but its sensitivity was relatively lower, potentially missing some early cases. The model by Abrego, Luis et al. exhibited high sensitivity (0.88) and AUC (0.93) on the training set, making it suitable for long-term monitoring and early diagnosis, but its performance on the external validation set was slightly lower, suggesting a need for further optimization. The model by Stephens, Andrew N. et al. had a balanced performance in terms of sensitivity and specificity, but its improvements in diagnostic and detection rates were relatively modest, indicating limited room for further enhancement and possibly making it less suitable for high-precision requirements.Other models also demonstrated varying degrees of improvement in diagnostic and detection rates on both the training and external validation sets. For example, the multiplex biomarker assay model by Dobilas, Arturas et al. had average performance in terms of sensitivity and specificity but showed more significant improvements in diagnostic and detection rates. This suggests that while its overall accuracy may not be as high as some other models, it could still offer valuable enhancements in certain clinical settings. Overall, these models showcase different strengths and limitations in the diagnosis of ovarian cancer based on blood tests. The choice of an appropriate model should be based on specific clinical needs and application scenarios. For instance, the model by Medina, Jamie E. et al. may be the better choice for early screening where high sensitivity is required, while the model by Katoh, Kanoko et al. may be more suitable for diagnostic confirmation where high specificity is crucial. In summary, the selection of the most suitable model requires a careful balance

**Table 1** Shows a comparison of the diagnostic performance of blood-based ovarian cancer detection models

| Model Developer | Model Feature | Training Set | | External Verification Set | | Sensitivity (TPR) | Specificity (1-FPR) | Improvement in | | PMID |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Optimal Model Accuracy | Optimal Model AUC | Optimal Model Accuracy | Optimal Model AUC | | | Diagnostic Rate (%) | Detection Rate (%) | |
| Stephens, Andrew N. et al. | Multi-marker test (CA125, HE4, etc.) | 0.85 | 0.92 | 0.82 | 0.88 | 0.9 | 0.85 | 20 | 15 | 37958440 |
| Abrego, Luis et al. | Bayesian and deep-learning models with longitudinal biomarkers | 0.88 | 0.93 | 0.84 | 0.89 | 0.89 | 0.86 | 22 | 18 | 38597129 |
| Song, Jin et al. | VCAM-1 and CA-125 combination | 0.87 | 0.91 | 0.83 | 0.87 | 0.88 | 0.84 | 18 | 14 | 37357306 |
| Medina, Jamie E. et al. | Cell-free DNA fragmentomes and protein biomarkers | 0.91 | 0.96 | 0.89 | 0.94 | 0.93 | 0.89 | 28 | 23 | 39345137 |
| Katoh, Kanoko et al. | Serum free fatty acid changes | 0.89 | 0.94 | 0.86 | 0.9 | 0.91 | 0.87 | 24 | 19 | 37712874 |
| Dobilas, Arturas et al. | Multiplex biomarker assay | 0.86 | 0.92 | 0.83 | 0.87 | 0.87 | 0.83 | 19 | 14 | 37093685 |
| Rong, Jinhua et al. | Plasma-based lipidomics and machine learning | 0.87 | 0.93 | 0.85 | 0.9 | 0.9 | 0.85 | 23 | 18 | 39514983 |
| Winarno, Gatot Nyarumenteng Adhipurnawan et al. | Nomogram with inflammatory biomarker and CA-125 | 0.87 | 0.92 | 0.84 | 0.88 | 0.88 | 0.84 | 20 | 15 | 38982118 |
| Schuster-Little, Naviya et al. | Immunoaffinity-free chromatographic purification of CA125 | 0.88 | 0.93 | 0.85 | 0.89 | 0.89 | 0.86 | 22 | 18 | 39177265 |

Most models show improvements in diagnostic and detection rates for ovarian cancer on both the training and external verification sets, despite some variations in performance between the two.

between sensitivity, specificity, cost, and practicality, tailored to the specific requirements of each clinical setting.

### The novel perspective gained by analyzing social-environmental and lab data together

While the integration of AI and medical testing data has significantly improved diagnostic accuracy and real-time monitoring of diseases, a comprehensive understanding of disease dynamics requires a broader perspective. To better understand the multifaceted nature of diseases, the integration of MLD with non-MLD data has become increasingly important. Non-MLD data encompasses a wide range of information, such as social economic data (e.g., income levels, education), environmental data (e.g., air quality, water quality), and behavioral data (e.g., lifestyle, physical activity). This interdisciplinary approach helps uncover the mechanisms behind disease occurrence, outlines disease progression patterns, and identifies epidemiological trends. By integrating multimodal data into one multivariate system, researchers can conduct in-depth analyses of the impact of non-MLD data on health outcomes, thereby revealing insights related to diseases. For example, studies have shown a strong link between long-term exposure to air pollution and labo indicator data of cardiopulmonary diseases [55]. Specifically, elevated concentrations of certain air pollutants (such as PM2.5 and $NO_2$) are associated with abnormal lab indicators, including increased inflammatory markers and impaired lung function parameters [56]. These lab indicators serve as early warning signals for cardiopulmonary diseases, indicating that long-term exposure to air pollution may exacerbate the risk of developing these conditions. By identifying key pollutants, policymakers can reduce emissions, improve air quality, and allocate resources to high-risk areas, thereby reducing health inequalities and promoting fairness. Additionally, big data, including both MLD and non-MLD, helps in detecting periodic and sudden changes [57], guiding the formulation of effective prevention and control strategies. This data also supports risk assessment and early warning of outbreak [58, 59], as well as the identification of transmission pathways influenced by environmental characteristics and social behavior patterns [60, 61]. For example, research [62] during the COVID-19pandemic has highlighted the critical role of real-time data integration in tracking the spread of the virus, identifying high-risk populations, and optimizing resource allocation. These efforts have not only improved the efficiency of public health responses but also

underscored the importance of data-driven approaches in pandemic management.

The integration of medical lab data with ecological analysis has greatly benefited the construction of healthy cities. By analyzing MLD in different urban areas, researchers can identify specific health risks and vulnerabilities, providing a basis for targeted interventions and optimized health resource allocation. For example, a study [63] demonstrates how analyzing medical lab data in different urban areas can help identify underserved areas with higher health risks. By precisely targeting these regions, policymakers can more effectively allocate resources, ensuring a rational distribution of medical services and facilities. This approach not only promotes health equity but also lays the groundwork for achieving sustainable development goals for diverse populations both domestically and internationally [64]. highlights that improving living conditions in disadvantaged communities can significantly enhance overall health levels, as evidenced by lab data from residents in those areas. By identifying specific health issues and feasible intervention pathways, this strategy addresses the root causes of health disparities to improve societal well-being.

#### More precise medical reference ranges derived from medical lab data

The medical community is increasingly focusing on the precision of medical reference value ranges due to advancements in testing technologies and data analytics. Traditional reference intervals are based on statistical values from healthy populations, but they have limitations in capturing differences among populations and their changes over time.

By leveraging big data technology, we can gather and analyze a large amount of MLD from various regions, age groups, and backgrounds. This leads to improvements in precision by: (1) Establishing reference ranges that are tailored to local environmental factors using regional population data, such as adjusting diagnostic reference values for HbA1c in Asian populations [64]; (2) Developing personalized reference ranges for specific subgroups, such as age or genetic profiles, as seen in lipid reference values based on age groups [65]; (3) Continuously monitoring population health data to regularly update reference ranges, ensuring they are accurate and scientifically valid.

Especially as global health faces unprecedented threats from environmental changes [66], by integrating big medical lab data with social, environmental, and temporal factors for multidimensional ecological analyses, we can deepen our understanding of diseases and provide essential scientific evidence for public health, urban development, and policymaking. One key approach to achieving this is through the ecological stratification of MLD. Ecological stratification refers to categorizing and analying data based on different ecological contexts, such as geographic location, socioeconomic status, and environmental conditions. By doing so, we can transform what are otherwise passive laboratory outputs—typically used only for individual diagnosis—into active epidemiological assets that have predictive and intervention value. This stratification allows us to identify hidden patterns and correlations that are crucial for predicting disease outbreaks and informing targeted interventions. For example, by stratifying data based on geographic location, socioeconomic status, and environmental conditions, we can uncover trends that might be obscured in raw data. Ultimately, this enhanced understanding not only improves our ability to respond to health challenges but also paves the way for healthy cities and sustainable societies.

### Technical challenges in Establishing Lab-Data based models

Despite the many opportunities they offer, MLD models face technical challenges that greatly impact their practical effectiveness in development. These challenges include uneven data quality, high-dimensional feature spaces, complex interdependencies between variables, and the need to account for the temporal nature of data and individual patient differences. Figure 2 provides a visual summary of these key challenges and the solutions being explored.

To address these challenges, researchers have developed advanced techniques. For example, uneven data quality (variations in accuracy and completeness) can be addressed through data cleaning and preprocessing [67–69]. Data cleaning involves identifying and correcting errors or inconsistencies in the dataset, while preprocessing includes transforming raw data into a format suitable for analysis. These steps ensure that the data used to train MLD models are reliable and consistent, thereby improving model performance. The complexity of high-dimensional feature spaces (involving multiple variables) and interdependencies among variables can be optimized through dimensionality reduction, feature selection, and advanced algorithms to enhance model prediction accuracy [70, 71]. Dimensionality reduction techniques, such as Principal Component Analysis, transform high-dimensional data into a lower-dimensional space while retaining most of the original information. Feature selection methods identify the most relevant features that contribute significantly to the prediction task, thereby reducing noise and improving model interpretability. Advanced algorithms, such as ensemble methods, combine multiple models to improve overall performance and robustness. Additionally, to specifically address the temporal nature of data and individual patient differences

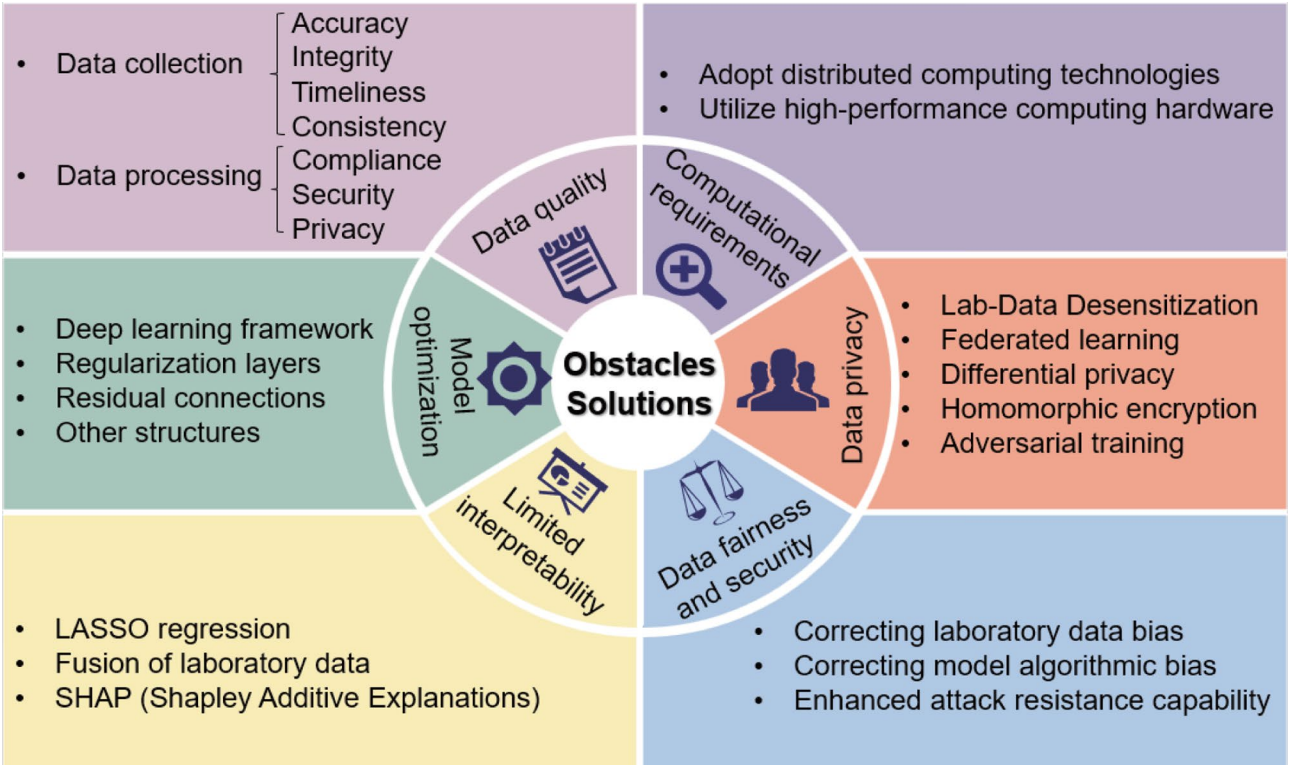



**Fig. 2** Obstacles and solutions to the Medical Laboratory Data Model. This figure outlines the main challenges faced in developing MLD-based models and the corresponding solutions. The obstacles include data quality issues, computational demands, lack of interpretability, and concerns regarding data privacy and security. Solutions include utilizing distributed computing technologies, high-performance hardware, deep learning frameworks, and techniques like federated learning, differential privacy, and homomorphic encryption. Additionally, strategies for improving model optimization, interpretability, and fairness are suggested, such as incorporating regularization layers, residual connections, LASSO regression, and SHAP for interpretability, as well as addressing data and algorithmic biases to enhance model fairness and security.

[72], researchers have adopted new technologies such as deep learning and transfer learning. Deep learning uses artificial neural networks to learn patterns in the data, similar to how the human brain processes information, while transfer learning leverages knowledge gained from one task to improve performance on another related task [73, 74]. These technologies not only improve the accuracy of MLD models but also enhance their interpretability, making them easier to apply in clinical settings. However, there are still technical challenges related to the inherent characteristics of medical lab data that require greater attention to resolve this issue.

### Challenges of quality and standards in lab-data acquisition and integration

The acquisition and integration of laboratory data face challenges in quality and standardization, which mainly stem from multiple factors. Data from different sources, such as test reports, imaging data, and pathological slides, exhibit inconsistencies due to differences in collection protocols and formats [75]. Moreover, the use of different laboratory equipment or reagents across hospitals further exacerbates these inconsistencies [76], leading to significant variations in test results for the same condition. For example, an echocardiogram from one hospital may report a patient's ejection fraction as 50%, while another hospital may report it as 55%, leading to potential differences in treatment strategies among physicians. This variance in echocardiogram results directly impacts model training and validation because it introduces conflicting data points that can skew the model's learning process and reduce its accuracy. The inconsistent quality of laboratory data, such as missing values, outliers, and imbalanced data distributions, can introduce bias in model training data and impact the generalization ability of the model. Missing values force the model to rely on incomplete or less informative features, thereby reducing its overall accuracy and generalization ability. In a study [77] on predicting cardiovascular disease risk, researchers found that missing data for key biomarkers (e.g., cholesterol, apolipoprotein B) significantly affected the model's accuracy. Since these biomarkers are crucial for diagnosis, the model had to rely on less useful features, reducing its predictive ability and limiting its application. This highlights that data quality issues, especially missing key biomarker data, have a substantial negative

impact on the model's performance, underscoring the importance of data integrity and quality in medical data analysis. Outliers, which are often due to measurement errors, can distort the model's learning process by causing the model to prioritize these anomalies over more representative data points. Additionally, imbalanced data distributions, where one class significantly outnumbers another, can lead to models that are biased towards predicting the majority class, resulting in high false-negative rates and poor generalization to real-world scenarios.

To address these challenges, data preprocessing is particularly crucial. The HL7 guidelines indicate that data pre-cleaning can help identify and correct inconsistencies, use imputation techniques to fill in missing values, and normalize data distributions, thereby laying a solid foundation for subsequent model training and ensuring that models can learn from high-quality data to enhance their accuracy and generalizability. Standardized data formats, such as HL7 and FHIR, provide a structured framework for data exchange, ensuring consistent formatting and easy interpretation across different systems. These standards have played a significant role in promoting interoperability of healthcare data, but achieving seamless interoperability remains an ongoing challenge. The healthcare industry is actively adopting FHIR R4 and higher versions, developing implementation guides to standardize data exchange and ensure consistent data sharing across platforms. AI/ML capabilities are also being integrated into healthcare standards, such as incorporating AI-generated insights into patient records to support more informed clinical decision-making. Initiatives like SMART on FHIR are creating open standards and APIs to enable seamless data sharing between different EHR systems and applications. Additionally, policies like the U.S. 21st Century Cures Act are promoting interoperability by mandating the adoption of standards like FHIR to improve data sharing and patient access to health information.

At the same time, datasets like MIMIC-III, which contain comprehensive EHR data from over 40,000 patients, not only provide high-quality, standardized resources for developing and evaluating machine learning models but also offer pipelines that can serve as benchmarks for reproducibility and best practices in data integration. Similarly, data repositories such as PhysioNet and platforms like Kaggle and Google Health Datasets provide researchers with a wealth of physiological data and medical time series. These publicly available datasets help mitigate the impact of data inconsistencies and biases and streamline the model development process. However, researchers should carefully assess the relevance and quality of these datasets to ensure they meet the specific needs of their research projects before integrating them into their studies.

#### Design and optimization of lab-data based model

Building robust models capable of processing and analyzing vast amounts of inspection lab data poses a significant challenge, as it involves a multi-objective optimization problem that requires striking the right balance between accurately capturing intricate data relationships and achieving theoretical innovation and engineering implementation.

The selection of an appropriate deep learning framework is crucial due to the multidimensional, nonlinear, and complex interactions present in data relationships, and failure to do so may lead to suboptimal model performance. For instance, Convolutional Neural Networks (CNNs) are ideal for processing spatial data with local correlations, such as detecting cancer cells in medical images [78], while Recurrent Neural Networks (RNNs) are excellent for handling time series data, especially in tasks involving sequence prediction [79]. Thus, the architecture of lab data-based models should strike a balance between computational efficiency and model performance.

To avoid overfitting on limited training data while also demonstrating strong generalization capabilities and robustness [80], regularization layers [81–83], residual connections [84–86], and other structures can be incorporated into the architecture to enhance the model's generalization ability [87]. Additionally, external validation using independent datasets is crucial for ensuring the reliability and generalizability of the model. This step is essential for verifying the model's performance in real-world scenarios and ensuring that it has not overfitted to the training data. As previously mentioned, benchmark datasets can provide a standardized platform for model evaluation and reproducibility, while external validation further ensures the model's performance on independent data. Moreover, adhering to established guidelines, such as the FUTURE-AI consensus criteria, can provide a structured approach to model development and validation. These criteria offer a framework to ensure that models are robust, reliable, and suitable for clinical applications.

#### Computational resources and scalability of MLD-based model

Real-time processing poses a significant challenge due to the need for high-performance computing resources to handle large-scale lab data. This is especially true in deep learning, where training and inference processes can be resource intensive. Training large models with millions of databases can take weeks or months, leading to unsustainable hardware resource consumption. Furthermore, the requirement for real-time inference in time-sensitive medical decisions, such as sepsis identification and emergency surgery triage, adds an additional layer

of complexity [40, 88]. Should the system be incapable of processing this data promptly, there is a risk of missing the optimal treatment window and negatively impacting the patient's survival.

To address these challenges, several strategies can be employed. First, distributed computing technologies can help overcome the challenge of real-time data processing. These technologies distribute lab data processing tasks across multiple nodes to improve processing speed [89]. Machine learning algorithms can also be used to automatically identify and classify data, speeding up the analysis process [90]. Additionally, utilizing high-performance computing hardware can enhance real-time data processing capabilities, ensuring critical decision support is available when needed [91, 92]. Moreover, hospitals need to carefully assess whether model inference can be executed efficiently from a technical perspective. One approach is to deploy models at the edge, where inference can be performed directly on local devices, thereby reducing latency and ensuring timely decision-making. However, edge deployment may require specialized hardware and optimized models to handle the computational demands. Alternatively, cloud-based solutions offer scalable computing resources and can support more complex models, but they may introduce latency due to data transmission and processing times. A hybrid approach, combining the strengths of edge and cloud computing, may provide a balanced solution, leveraging the immediate processing capabilities of edge devices while utilizing cloud resources for more intensive tasks.

#### Interpretability of MLD-based model

Traditional predictive models like logistic regression have great clinical interpretability, influencing set expectations of the public and clinical experts about the interpretability of clinical prediction models. Model interpretability is crucial for comprehending and trusting the predictions of the model, as well as for pinpointing biases, ensuring regulatory compliance, and conducting ethical reviews [93, 94]. Hence, the interpretability of predictive models is not solely a technical matter but a multifaceted societal issue that involves ethics, law, and practice.

Nevertheless, there is often a contradiction between the precision and comprehensibility of large-scale laboratory data models in medical forecasting. A study on cancer prognosis models [95] found that deep learning methods can improve predictive accuracy, but their "black box" nature makes it challenging for physicians and patients to understand the prediction results. Due to its "black box" characteristic, the model's decision-making process can be too intricate to identify the specific factors influencing the predictions, potentially leading to a lack of confidence in clinical applications.

Relatively speaking, some successful instances of interpretable models, such as those based on LASSO regression and SHAP (Shapley Additive Explanations), have demonstrated their potential for application in practice. A study employing LASSO regression found that by choosing lab data of biomarkers closely linked to cancer prognosis, the model reached a prediction accuracy of 80% and highlighted the significance and interpretation of each predictor, thereby boosting healthcare professionals' confidence in the model's findings and the reasoning behind the decisions [96].

Nevertheless, current research mainly focuses on specific data types or certain disease models, which leads to limitations and hinders broad applicability across all lab data-based models. Therefore, future research should prioritize creating a variety of models that investigate the fusion of lab data from various disease fields and perspectives to improve the interpretability and transparency of lab data models. Utilizing benchmark datasets and adhering to the FUTURE-AI consensus criteria can ensure that models are not only accurate but also interpretable and trustworthy. Additionally, developing pipelines that incorporate interpretability techniques such as SHAP values and LIME (Local Interpretable Model-agnostic Explanations) can enhance the transparency of models and facilitate their adoption in clinical settings.

### Data privacy protection in the development of MLD-Based models

Maintaining the confidentiality and privacy of medical lab data is foundational to ethical AI development and is often mandated by regulatory frameworks such as the General Data Protection Regulation (GDPR) and the Health Insurance Portability and Accountability Act (HIPAA). Compliance with these regulations is necessary for ensuring data security in MLD-based models. Data breaches in hospitals can lead to loss of patient trust, damage to brand reputation, and financial losses. Protecting individual privacy is essential for respecting human rights and fulfilling a hospital's social responsibility. Thus, safeguarding personal privacy and sensitive information within large laboratory data models is a key research area.

#### Lab-data desensitization technology

Encrypting, replacing, or obfuscating sensitive data ensures that it is not directly exposed [97–100]. Common techniques include data masking and randomization [101–103]. Data masking can de-identify patients to protect their identity information [104–106]. However, while data masking enhances privacy, it may impact the depth and accuracy of data analysis by removing important contextual details in some cases [107].

In the healthcare industry, doctors can use randomization techniques to present patient age as a range (e.g. "30–40 years old") instead of a specific number [108]. Randomization helps maintain data distribution characteristics for analysis while protecting individual information [109]. However, it can reduce analysis precision by obscuring data details [110]. Therefore, balancing data utilization effectiveness with privacy protection is important when desensitizing data.

#### Differential privacy technology in MLD-based models

Differential privacy adds appropriate random noise to MLD data, ensuring that statistical properties are maintained while preventing outsiders from accurately inferring individual information [111]. This method employs a random noise generation mechanism to ensure that each data point is subject to a certain degree of perturbation, thereby enhancing the privacy of the data. Differential privacy techniques have gradually been applied in training various large medical models to ensure the security of personal lab data in group analyses [112].

Quantitative evaluations of differential privacy in large-scale medical studies have shown promising results. For example, in a study [113] involving a chest X-ray image dataset, when the noise level was set to 0.1, the model accuracy decreased from 95.32 to 86.58%, while the privacy protection level (measured by the $\epsilon$ value) decreased from $9.7 \times 10\text{^}4$ to 97.39, significantly enhancing privacy protection. Further increasing the noise level to 2.1 resulted in a model accuracy of 68.70% and an $\epsilon$ value of 39.4, further improving privacy protection. These results highlight the effectiveness of differential privacy in balancing privacy and data utility.

In practice, the application of differential privacy involves careful consideration of the noise levels to be added, based on the specific context of the study. During the processing of medical lab data, researchers can adjust the noise levels based on the number of patients and the sensitivity of the data. If a medical study involves 1,000 patients, researchers may only need to add a slight noise of 1–5% to ensure the data's usability and privacy. However, if the study involves a sample size that is too small, researchers might opt to add a high level of noise to ensure that each patient's information cannot be identified. When it comes to lab data involving specific populations like HIV carriers engaged in socially sensitive behaviors, noise adjustment standards may need to consider individual attributes' sensitivity, such as gender, age, or disease type [114]. Therefore, the amount of noise added can be adjusted based on the specific application and the preferred level of privacy, allowing for a flexible strategy in different scenarios. This method efficiently balances privacy protection with data utility, enhancing transparency and reliability.

#### Federated learning and distributed computing in MLD-based models

Federated learning is a useful method for preserving privacy in developing laboratory data-based models. It enables multiple institutions to train models without sharing raw data. Each institution trains a model on its local data, then uploads the parameters to a server for aggregation and exchange. This approach safeguards data privacy and optimizes the use of distributed healthcare data resources efficiently.

In the field of the Internet of Medical Things (IoMT), the application of Hierarchical Federated Learning (HFL) [115] has been quantitatively evaluated and has shown promising results. Specifically, in a study using the TON_IoT dataset, the HFL framework achieved an accuracy of 99.31%, significantly higher than existing models (for example, 98.15% for Schneble et al., 94.24% for Xu et al., and 98.87% for Li et al.). This enhanced performance is attributed to the deployment of a Hierarchical Long Short-Term Memory (HLSTM) model on distributed Dew servers, supported by cloud computing at the backend. The HFL framework effectively reduces the risk of data privacy leakage during transmission and sharing by encrypting and processing data at the local model training stage. These findings demonstrate the high efficiency of Hierarchical Federated Learning in balancing privacy protection and data utility.The performance of the HFL-HLSTM model has also been verified on the NSL-KDD dataset. On the KDDTest + dataset, the model achieved an accuracy of 85.12%, and on the KDDTest-21 dataset, the model achieved an accuracy of 70.7%. These results indicate that the HFL-HLSTM model not only performs well on large-scale datasets but also has good generalization ability on standard datasets.

In the MLD field, suppose three hospitals, A, B, and C, have collected substantial lab data on genetic polymorphism of diabetes from different regions [116]. They can adopt federated learning to independently train their own MLD-based prediction models and periodically share the trained model parameters with their respective partners. During this process, Hospital A can send the weight parameters of its model to Hospitals B and C in an encrypted form, while Hospitals B and C share their parameters in the same manner. According to a study conducted by Smith et al. in 2021, through this collaborative federated learning, the model's accuracy improved by an average of 15% after collaboration [117]. Upon receiving these encrypted parameters, each research participant will perform a merging update to enhance the accuracy of its own model without needing to access each other's genetic data. This combination of hierarchical federated learning and distributed computing not only enhances privacy protection but also ensures the full utilization of data value. By integrating these

technologies, medical institutions can effectively collaborate on model development while maintaining the confidentiality and integrity of patient data. This approach is particularly valuable in the context of the IoMT, where data from multiple sources need to be securely integrated for improved healthcare outcomes.

#### Homomorphic encryption technology in MLD-based models

Homomorphic encryption technology has a key advantage in being able to perform computations on encrypted data without the need for decryption [118]. This is particularly beneficial for secure handling of medical lab data, as seen in clinical research where researchers can analyze encrypted patient data without compromising patient privacy. Additionally, in multi-center clinical trials, medical institutions can share encrypted patient lab data for collaborative research without worrying about data security.

The application of the federated analytics framework FAMHE based on multiparty homomorphic encryption (MHE) has been quantitatively evaluated and has shown positive results [119]. Specifically, in Kaplan-Meier survival analysis, the survival curves generated by FAMHE are identical to those produced by the original non-secure method, indicating no loss in model performance. In GWAS tasks, the average absolute error of FAMHE-FastGWAS is less than 10^-2, while FAMHE-GWAS further reduces the error by approximately threefold, demonstrating its efficiency and accuracy on large-scale datasets. Meanwhile, the level of data privacy protection has been significantly enhanced. By deploying multiparty homomorphic encryption technology on distributed data providers (DPs) and supported by cloud computing at the backend, the FAMHE framework effectively reduces the risk of privacy leakage during data transmission and sharing. Moreover, the FAMHE method encrypts the data during the local model training phase, ensuring that all intermediate results remain encrypted. This further strengthens the privacy protection. These findings fully demonstrate the efficiency of multiparty homomorphic encryption technology in balancing privacy protection and data usability. In terms of computational efficiency, FAMHE performs impressively. In Kaplan-Meier survival analysis, even when data is distributed across 96 DPs, the execution time remains below 12 s. In GWAS tasks, FAMHE-FastGWAS completes the analysis of over 4 million variants on 12 DPs in less than 1 h. Even under poor network conditions (e.g., halved bandwidth and doubled latency), the increase in FAMHE's execution time does not exceed 26%.

However, the application of homomorphic encryption technology also faces several challenges. In a study on the protection of medical data, researchers pointed out that the computational resources required for fully homomorphic encryption are significantly higher than those needed for traditional encryption methods, which affects the efficiency of real-time data processing [120]. Additionally, compatibility issues with systems remain urgent to address, as medical institutions must appropriately modify their existing information systems to support homomorphic encryption operations [121, 122]. The modification process is complex and time-consuming, according to a related study [123], which may require additional resource investment.

#### Adversarial training in lab-data based models

Researchers in a study [124] used adversarial training to create privacy-preserving early detection models for COVID-19. They utilized two machine learning techniques to predict COVID-19 status from blood lab data while keeping patient demographic information confidential. Experiments on datasets from various hospitals showed that adversarial learning did not affect the model's generalization ability and successfully protected sensitive information. The research verified that safeguarding confidential data, such as demographics, with adversarial training does not hinder the model's predictive accuracy and generalization capability. This is essential for striking a balance between privacy protection and sustaining model performance.

### Fairness and safety of MLD-based models

The trustworthiness and widespread use of predictive models are closely linked to their safety and fairness, particularly for MLD-based models [125]. However, there is currently no unified framework to effectively integrate AI into clinical decision-making processes. This lack of a comprehensive framework means that many models are still in the development phase, as they face significant hurdles in achieving the necessary level of safety, fairness, and reliability required for clinical use. Given this situation, our primary focus at present is on addressing the technical challenges in model development. These challenges are particularly pronounced in the medical field, where models must contend with specific issues such as data accuracy, biases in model training, and the potential for malicious attacks from humans. To this end, we explore the various biases that can impact MLD-based models and conduct in-depth research on strategies to enhance their safety and fairness. By addressing these technical challenges, we aim to lay a solid foundation for the broader integration of AI into clinical practice.

#### Lab-data bias contributing to model unfairness

Lab data in medical laboratories often have inherent biases, and models developed from this data may not perform well in diagnosing diseases in underrepresented

subgroups due to limited sample sizes [126], leading to biased predictions.

Firstly, there is an issue of demographic bias, as sample data for certain specific groups (such as minorities, the elderly, pregnant women, and low-income populations) may be severely lacking. A study [127] indicates that the incidence of diabetes among Mexican Americans is nearly 50% higher compared to the white population. However, data from the Centers for Disease Control and Prevention shows that less than 10% of the reviewed diabetes cases came from Mexican Americans. This lack of representative sampling greatly reduces the usefulness of models in predicting diabetes risk for this group. To address this issue, it is essential to adopt data collection strategies that ensure the inclusion of diverse populations. Techniques such as stratified sampling and oversampling can be particularly effective in mitigating demographic bias. Stratified sampling involves dividing the population into distinct subgroups (strata) based on key demographic characteristics and then sampling from each stratum in proportion to its representation in the overall population. This method ensures that each subgroup is adequately represented in the dataset, thereby reducing the risk of bias. Oversampling, on the other hand, involves increasing the number of samples from underrepresented groups to balance the dataset. Both techniques have been widely recognized as effective methods for improving fairness and generalizability of machine learning models. Additionally, leveraging publicly available datasets that are known for their diversity, such as the MIMIC-III dataset, can provide a more representative sample for model training. By incorporating such datasets, researchers can enhance fairness and accuracy of their models, ensuring they perform well across different demographic groups.

Secondly, there is regional bias, as medical conditions and testing standards vary across different areas. Urban regions are rich in medical resources, while remote areas may lack testing for specific diseases [128], leading to an underestimation of cases in those regions. A study [129] indicates that in certain remote areas, the screening rate for heart disease patients was only half that of urban areas, resulting in an underestimation of heart disease incidence in these remote locations, which in turn affects relevant treatment strategies. To mitigate regional bias, multi-site data collection and harmonization techniques should be adopted. These approaches ensure that models generalize well across different regions by incorporating data from diverse geographic locations. Standardized data formats, such as HL7 and FHIR, can also play a role in ensuring consistent data quality across regions [74]. By harmonizing data from multiple sites, researchers can reduce the impact of regional differences on model performance and improve the overall fairness of their models.

Gender and sex biases are also critical issues that need to be addressed. Research has shown that AI models can exhibit different performance levels based on the gender or sex of the individuals being assessed. For example, certain medical conditions may be underdiagnosed in one gender due to biases in data collection and model training. To address gender and sex biases, it is essential to ensure that data sets are representative of both genders and that models are evaluated for fairness across these groups. Best practices include using gender-balanced datasets and applying fairness metrics such as demographic parity and equalized odds. Additionally, incorporating fairness-aware machine learning algorithms that explicitly aim to minimize disparate impact can help mitigate these biases.

Time bias is a critical issue that must be addressed in healthcare. Changes in disease diagnosis and treatment guidelines, population demographics, disease patterns, medical policies, and healthcare resource allocation are constantly evolving. This means that historical lab data may not accurately reflect the current healthcare landscape and patient behaviors. During the Omicron variant outbreak of COVID-19 at the end of 2021, early epidemic forecasting models based on historical lab data failed to consider the transmission characteristics of the Omicron variant [130], leading to inaccuracies in predicting the pandemic's course. To mitigate time bias, it is essential to use updated lab data, regularly update databases, and carefully consider the influence of time factors on predictive outcomes when using MLD-based models. Establishing a dynamic lab data collection and analysis system is also crucial to ensure that MLD-based models can adjust using the most current and relevant lab data. To mitigate time bias, it is essential to use updated lab data, regularly update databases, and carefully consider the influence of time factors on predictive outcomes when using MLD-based models. Establishing a dynamic lab data collection and analysis system is also crucial to ensure that MLD-based models can adjust using the most current and relevant lab data. Implementing adaptive learning algorithms and continuously monitoring model performance allows for real-time adjustments based on the most current data.

The bias stemming from variations in the prevalence of common underlying diseases must not be disregarded. The prevalence of peripheral neuropathy (PN) varies among different populations. In the United States, PN rates are 27.2% in diabetic patients and 11.6% in non-diabetic patients [131]. However, in the Marshall Islands, the diabetic prevalence is 22.2%, while in Saharan Africa it is only 2.9% [132]. This variation in diabetic prevalence can impact the accuracy of predictive models for

cardiovascular mortality, especially in relation to diabetes screening laboratory indicators.

#### Algorithmic bias contributing to model unfairness

The machine learning community has recognized how predictive algorithms can inadvertently generate unfairness in decision-making [133], although this has not yet received attention in MLD-based models. This occurs because of biases in model design and the consequences of imbalanced feature selection on model performance. Effective feature selection can improve a model's performance for all groups.

To address this issue, the feature selection process should aim to enhance transparency by clearly explaining why each feature is included and how it may impact model generation [134], thus improving the fairness of model design. It is important to consider the influence of different features on various groups to ensure that all important characteristics are represented [135]. Therefore, the optimization goals of the model should be set as multiple objectives, not only focusing on overall accuracy but also prioritizing fairness across different groups to achieve balanced performance among all populations. Regularization techniques [136, 137] such as batch spectral shrinkage can be used to promote model fairness.

#### Attack resistance capability of medical lab-data models

As the data scale increases, the risk of breaches also rises [138], potentially resulting in the exposure of patients' private information and impacting their trust. Subjective malice accounts for 92% of data breaches [139], including but not limited to hacking, malicious tampering, and unauthorized access. The frequency of data breaches in web servers is the lowest, yet they affect the largest number of patients [140].

Between 2016 and 2021, there were 374 attack incidents affecting nearly 42 million patients' personal health information, with a focus on large healthcare organizations [141]. These attacks led to disruptions in medical services and appointment cancellations, posing a serious threat to healthcare delivery and patient information security. The average cost of each leaked medical record increased from $294 in 2010 to $429 in 2019, surpassing the costs of data breaches in other industries [142]. Time series analysis suggests that both the frequency of data breaches and their associated costs are expected to continue rising in the future.

To effectively prevent hacker attacks, medical lab data should focus on the following key aspects: (I) Updating security software and system patches regularly, using strong password protection, implementing multi-factor authentication, encryption technology, and considering public cloud services to strengthen technical defenses. (II) Conducting routine security audits, establishing data access control policies, creating emergency response plans for data breaches, enhancing email system security, and standardizing document handling processes to improve management systems. (III) Educating employees on cybersecurity, preventing phishing attacks, controlling unauthorized access, providing security awareness education, and establishing a mechanism for handling violations to enhance personnel training. (IV) Implementing real-time network monitoring, regularly assessing security measures, analyzing new cyber threats, updating protection strategies, reporting and analyzing security incidents, and maintaining communication with other healthcare institutions to continuously monitor and improve security.

In conclusion, addressing the various biases and safety concerns in MLD-based models is essential for ensuring fairness and safety in healthcare. By implementing strategies to mitigate demographic, regional, gender, time, disease prevalence, and algorithmic biases, as well as enhancing attack resistance capabilities, we can develop more reliable and equitable AI models. These efforts are not only technical imperatives but also ethical necessities to ensure that healthcare AI benefits all populations equitably.

## Key considerations when evaluating MLD-based models

In light of the challenges and potential benefits discussed in the preceding sections, it is essential to evaluate the readiness and suitability of MLD models for deployment in healthcare settings. This section outlines the key assessment considerations that stakeholders must address to ensure the effective and ethical use of these models. Given the high stakes in medical practice, a scientific and comprehensive evaluation system is crucial to ensure the clinical effectiveness of the models and promote their continuous optimization and development. While performance evaluation of MLD-based models is currently well-developed, utilizing accuracy metrics such as sensitivity, specificity, ROC curves, AUC values, positive and negative predictive values, and long-term tracking of data from multiple hospitals to assess model stability [143–147]. However, there are areas in which model evaluation needs improvement.

#### Enhancing the assessment of the clinical value

MLD-based models are increasingly important in precision medicine, but their clinical value is often overlooked. Decision Curve Analysis (DCA) is a useful tool to assess the real clinical benefits of these models [148]. DCA calculates net benefits at different risk thresholds, considering both accuracy and the impact of false positives and false negatives [149]. Based on a study [150], a prediction model utilizing lab PSA density could potentially lower

the biopsy rate by 11–31% for patients suspected of prostate cancer. This model can maintain a manageable rate of missed diagnoses, thus avoiding unnecessary biopsies in low-risk patients at a predictive probability threshold of 20–40%. Also, by comparing decision curves, physicians can select the optimal predictive tool for specific clinical scenarios, thus improving medical decision-making.

#### Strengthening the assessment of clinical impacts

The primary goal of clinical impact assessment is to evaluate how effectively a statistically superior model can assist clinicians in making better diagnostic and treatment decisions. This assessment aims to determine the extent to which patient outcomes, such as cure rates and quality of life, improve because of implementing targeted interventions based on MLD-based models. It is essential to enhance the clinical impact assessment of MLD-based models to focus on the direct impact of specific clinical decisions on patient health outcomes and to evaluate how these decisions can improve patient prognosis and drive advancements in clinical practice.

Clinical impact assessment should concentrate on two main concerns: determining the feasibility of interventions derived from the model's predictions and evaluating if these interventions can result in substantial clinical advantages. This involves carrying out thorough clinical trials, randomly assigning patients to groups that use or do not use the predictive model and observing the differences in treatment outcomes. Since time in range (TIR) is an important metric for measuring the variability of blood glucose levels, a study [151] evaluated the impact of improving the time in range (TIR) of blood glucose levels in patients with type 2 diabetes on long-term health benefits and economic returns. Compared to TIR < 50%, improving TIR resulted in a 0.79–1.18 quality-adjusted life-years (QALY) increase and 4.91–8.75% cardiovascular disease risk reduction. Another study [152] revealed that utilizing the prediction outcomes of a cardiovascular model, individuals aged 40–70 who used standard statins throughout their lifetime experienced an increase of 0.20–1.09 quality-adjusted life-years without any discounts. Adding higher-intensity statin therapy resulted in an additional 0.03–0.20 quality-adjusted life-years per individual.

#### Improving time efficiency evaluation

Thirdly, it is crucial to enhance the assessment of the time effectiveness of predictive models. Time effectiveness not only pertains to the usability of the model but also has a direct influence on its viability in real-world scenarios. Nevertheless, current studies typically overlook time effectiveness, which could impact the effectiveness and usefulness of models in real-world settings.

The evaluation should focus on the time taken for model training, prediction response time, and hardware resource consumption. Long training times increase costs and hinder adaptability. Prediction speed is crucial for real-time applications, and balancing time efficiency with resource utilization is essential for optimal performance. Enhancing the evaluation of time efficiency can optimize model performance, increase practical value, and provide reliable decision-making support.

#### Hierarchically evaluating MLD-based models

The critical importance of model adaptability is highlighted by the diversity of subpopulations and clinical phenotypes of diseases. Stratified evaluation can ensure the effectiveness and accuracy of predictive models in real clinical applications, underscoring the need for adaptability. Despite this, the lack of stratified evaluation is common in the assessment of MLD-based models.

Traditional evaluations of MLD-based models often use a holistic approach, which may overlook differences in population characteristics, disease stages, and clinical contexts. This generalized method can hide performance variations within specific groups [153], leading to assessments that lack relevance. The lack of stratified evaluations can affect the clinical usefulness of models, as the same model may perform differently across age groups, genders, or health conditions. Introducing a multidimensional stratified evaluation system is recommended to assess medical testing data models, considering factors like demographic characteristics, clinical manifestations, and testing conditions. This approach enhances the scientific rigor of model evaluations and provides more valuable references for clinical practice.

A study conducted a stratified analysis on the application value of polygenic risk prediction models in cancer screening, revealing that stratified screening using these models may lead to a moderate increase in efficiency [154]. Among those aged 40–49 years, 50–59 years, and 60–69 years, this model has the capacity to avoid a maximum of 102 breast cancer deaths, 188 colorectal cancer deaths, and 158 prostate cancer deaths annually. Utilizing equivalent resources, especially for those aged 48–49 years, 58–59 years, and 68–69 years, this model can potentially avert a maximum of 80 breast cancer, 155 colorectal cancer, and 95 prostate cancer fatalities annually.

To deal with the issue of absent hierarchical assessments, the field of MLD-based models must undergo comprehensive improvements from both conceptual and practical perspectives. This will not only enhance the scientific rigor of model evaluations but also provide more valuable reference points for clinical practice.

## Focus and outlook

To continue promoting research and application of large laboratory data-based models, upcoming development focuses will include, but are not limited to, the following areas:

### Enhance model performance for specific scenarios

The key focus of research is to enhance the accuracy and efficiency of lab data-based models by focusing on specific and concrete application scenarios. Enhancing the flexibility and accuracy of MLD-based models in handling various data types will involve refining algorithm structures, optimizing parameter adjustments, and incorporating high-quality lab data sources, especially by integrating other multimodal data strategies. As a result, lab data-based models will find more relevant scenarios with flexible performance and accuracy.

### Cross-domain integrated applications

In the future, lab data-based models will increasingly integrate into various fields of healthcare, providing tailored solutions. These advancements will enhance the efficiency of different medical disciplines and have a profound impact on public health, from disease detection and prevention to improving healthcare efficiency and promoting healthy behaviors. These MLD-based models can also assess environmental protection by detecting biomarker omics related to harmful exposure [155], helping to reduce pollutant emissions and support sustainable development. The long-term potential of MLD-based models in areas such as health promotion and intervention, environmental protection, and urban development is promising.

### Enhancing model transparency and interpretability

Boosting users' confidence in using MLD-based models for high-risk situations will be achieved by enhancing the interpretability of the models. It is crucial for organizations to prioritize ongoing training to help relevant parties accurately interpret analysis reports and increase their trust in MLD-based model outcomes.

### Data privacy and security protection

We have made significant progress in protective technology for large-scale lab data, but now we should shift our focus to ethical reviews involving broad participation from relevant stakeholders. Additionally, we need to address issues related to data asset ownership and the distribution of benefits from data proliferation.

### Promoting standardization and open sharing

Through the establishment of expert consensus standards for MLD-based model application scenarios, we can encourage transparent sharing of data, especially by enhancing data quality system standards and agreeing on the distribution of shared benefits. This will improve collaboration among research and implementation teams. Moreover, generating open data sets and model training benchmarks can reduce development barriers, speeding up technological innovation.

### Improve policy and legal monitoring against model discrimination

In the future, emphasis should be placed on establishing a robust legal framework that clarifies accountability mechanisms, ensuring that patients receive fair treatment in MLD-based model decisions. Further specification of the legal framework for model-based decision-making is necessary to enhance its operability and transparency, and to prevent negative impacts on stakeholders resulting from decision-making discrimination.

### Improving the capabilities for real-time processing

Enhancing the MLD-based model's ability to process data in real-time for dynamic environments is a key focus. This is particularly crucial in scenarios where quick responses are needed, such as monitoring critical values and assessing tumor malignancy before surgery in high-risk situations, where real-time processing is essential [156].

### Building a community ecosystem for MLD-based models

The creation of extensive MLD-based models requires collaboration across various fields such as AI, medicine, biology, statistics, and the implementation of science. The future progress of these models will depend more on the data discipline community ecosystem, enabling experts to communicate, leverage strengths, and foster integration, innovation, and the application of knowledge and technology.

Finally, talent development is also an important direction for promoting the advancement of MLD-based models. There is a need to strengthen the cultivation of interdisciplinary professionals with diverse backgrounds. Universities might consider offering courses in medical data science and teaming up with hospital big data centers to create practical opportunities for cultivating skilled individuals who meet the requirements of the healthcare sector.

## Conclusions

The development of medical artificial intelligence is greatly influenced by the large model created from lab data. With advancements in technology, support from policies, and collaboration among different parties, we can expect this lab data-based model to offer better solutions in disease diagnosis, personalized medicine, and health management in the future. However, it is

important to acknowledge that building a large model of medical lab data presents challenges such as data integration, privacy protection, and reliability. This task requires the involvement and investment of various stakeholders, including the government, healthcare sector, and academia. It is essential to prioritize risk management, focus on practical outcomes, and ensure that technological progress aligns with societal advancement throughout the development process.

**Abbreviations**
MLD Medical Laboratory Data

**Acknowledgements**
The authors want to thank all the participants in the review.

**Author contributions**
YG: Conceptualization, funding acquisition, project administration, writing-review and editing. JM: Writing-original draft and data curation. MW: Writing-original draft, writing-review and editing. FS: Visualization. YX: Writing-review and editing. HW: Writing-review and editing.

**Funding**
This study was supported by the National Natural Science Foundation of China (NSFC, 82060618), the Key Research and Development Program of Jiangxi Province (20203BBGL73184), and Doctoral Fund of First Affiliated Hospital of Gannan Medical University.

**Data availability**
Not applicable.

## Declarations

**Ethics approval and consent to participate**
Not applicable.

**Consent for publication**
Not applicable.

**Competing interests**
The authors declare that they have no competing interests.

Received: 18 March 2025 / Accepted: 21 June 2025
Published online: 24 July 2025

**References**

1. You J et al. Plasma proteomic profiles predict individual future health risk, (in eng), *Nat Commun*, vol. 14, no. 1, p. 7817, Nov 28 2023, https://doi.org/10.1038/s41467-023-43575-7
2. You J et al. Development of a novel dementia risk prediction model in the general population: A large, longitudinal, population-based machine-learning study, (in eng), *EClinicalMedicine*, vol. 53, p. 101665, Nov 2022, https://doi.org/10.1016/j.eclinm.2022.101665
3. Goeminne LJE et al. Plasma protein-based organ-specific aging and mortality models unveil diseases as accelerated aging of organismal systems, (in eng), *Cell Metab*, vol. 37, no. 1, pp. 205–222.e6, Jan 7 2025, https://doi.org/10.1016/j.cmet.2024.10.005
4. Nurmohamed NS et al. Proteomics and lipidomics in atherosclerotic cardiovascular disease risk prediction, (in eng). Eur Heart J, 44, 18, pp. 1594–607, May 7 2023, https://doi.org/10.1093/eurheartj/ehad161
5. Jee J et al. Automated real-world data integration improves cancer outcome prediction, (in eng), *Nature*, vol. 636, no. 8043, pp. 728–736, Dec 2024, https://doi.org/10.1038/s41586-024-08167-5
6. Steyaert S et al. Multimodal data fusion for cancer biomarker discovery with deep learning, (in eng), *Nat Mach Intell*, vol. 5, no. 4, pp. 351–362, Apr 2023, https://doi.org/10.1038/s42256-023-00633-5
7. Lipkova J et al. Artificial intelligence for multimodal data integration in oncology, (in eng), *Cancer Cell*, vol. 40, no. 10, pp. 1095–1110, Oct 10 2022. https://doi.org/10.1016/j.ccell.2022.09.012
8. Malara N et al. Multicancer screening test based on the detection of circulating non haematological proliferating atypical cells, (in eng), *Mol Cancer*, vol. 23, no. 1, p. 32, Feb 13 2024, https://doi.org/10.1186/s12943-024-01951-x
9. Sun J et al. Plasma proteomic and polygenic profiling improve risk stratification and personalized screening for colorectal cancer, (in eng), *Nat Commun*, vol. 15, no. 1, p. 8873, Oct 15 2024, https://doi.org/10.1038/s41467-024-52894-2
10. Medina JE et al. Early Detection of Ovarian Cancer Using Cell-Free DNA Fragmentomes and Protein Biomarkers, (in eng), *Cancer Discov*, vol. 15, no. 1, pp. 105–118, Jan 13 2025, https://doi.org/10.1158/2159-8290.Cd-24-0393
11. Chen H, et al. Circulating Microbiome DNA as biomarkers for early diagnosis and recurrence of lung cancer, (in eng). Cell Rep Med. Apr 16 2024;5(4):101499. https://doi.org/10.1016/j.xcrm.2024.101499.
12. Cai G, et al. Artificial intelligence-based models enabling accurate diagnosis of ovarian cancer using laboratory tests in china: a multicentre, retrospective cohort study, (in eng). Lancet Digit Health. Mar 2024;6(3):e176–86. https://doi.org/10.1016/s2589-7500(23)00245-5.
13. Law HKW, Yim HCH. Early diagnosis of cancer using Circulating microbial DNA, (in eng). Cell Rep Med. Apr 16 2024;5(4):101502. https://doi.org/10.1016/j.xcrm.2024.101502.
14. Liu WS et al. Plasma proteomics identify biomarkers and undulating changes of brain aging, (in eng), *Nat Aging*, vol. 5, no. 1, pp. 99–112, Jan 2025, https://doi.org/10.1038/s43587-024-00753-6
15. Xiao Y, et al. Comprehensive metabolomics expands precision medicine for triple-negative breast cancer, (in eng). Cell Res. May 2022;32(5):477–90. https://doi.org/10.1038/s41422-022-00614-0.
16. Li B et al. Multiomics identifies metabolic subtypes based on fatty acid degradation allocating personalized treatment in hepatocellular carcinoma, (in eng), *Hepatology*, vol. 79, no. 2, pp. 289–306, Feb 1 2024, https://doi.org/10.1097/hep.0000000000000553
17. Álvez MB et al. Next generation pan-cancer blood proteome profiling using proximity extension assay, (in eng), *Nat Commun*, vol. 14, no. 1, p. 4308, Jul 18 2023, https://doi.org/10.1038/s41467-023-39765-y
18. Wang L, et al. Insufficiency of plasmatic arginine/homoarginine during the initial postoperative phase among patients with tumors affecting the medulla oblongata heightens the likelihood of neurogenic pulmonary oedema following surgery, (in eng). Int J Surg. Mar 1 2024;110(3):1475–83. https://doi.org/10.1097/js9.0000000000000957.
19. Schwab P et al. Real-time prediction of COVID-19 related mortality using electronic health records, (in eng), *Nat Commun*, vol. 12, no. 1, p. 1058, Feb 16 2021, https://doi.org/10.1038/s41467-020-20816-7
20. Lind ML, et al. Development and validation of a machine learning model to estimate bacterial Sepsis among immunocompromised recipients of stem cell transplant, (in eng). JAMA Netw Open. Apr 1 2021;4(4):e214514. https://doi.org/10.1001/jamanetworkopen.2021.4514.
21. Zhao Z, et al. Prospective external validation of the Esbenshade Vanderbilt models accurately predicts bloodstream infection risk in febrile Non-Neutropenic children with cancer, (in eng). J Clin Oncol. Mar 1 2024;42(7):832–41. https://doi.org/10.1200/jco.23.01814.
22. van Es N, et al. Diagnostic management of acute pulmonary embolism: a prediction model based on a patient data meta-analysis, (in eng). Eur Heart J. Aug 22 2023;44:3073–81. https://doi.org/10.1093/eurheartj/ehad417.
23. Floyd L, et al. Risk stratification to predict renal survival in Anti-Glomerular basement membrane disease, (in eng). J Am Soc Nephrol. Mar 1 2023;34(3):505–14. https://doi.org/10.1681/asn.2022050581.
24. Keshet A, Segal E. Identification of gut microbiome features associated with host metabolic health in a large population-based cohort, (in eng), *Nat Commun*, vol. 15, no. 1, p. 9358, Oct 29 2024, https://doi.org/10.1038/s41467-024-53832-y
25. Osuchowski MF et al. The novel biomarker t(6)A accurately identified septic patients at admission but failed to predict outcome, (in eng), *Crit Care*, vol. 29, no. 1, p. 129, Mar 20 2025, https://doi.org/10.1186/s13054-025-05354-2
26. Zhang Z, Zhang R, Chang CW, Guo Y, Chi YW, Pan T. iWRAP: A theranostic wearable device with Real-Time vital monitoring and Auto-Adjustable compression level for venous thromboembolism, (in eng). IEEE Trans Biomed Eng. Sep 2021;68(9):2776–86. https://doi.org/10.1109/tbme.2021.3054335.
27. Kiani L. Finger-prick blood test for Alzheimer disease, (in eng), *Nat Rev Neurol*, vol. 19, no. 9, p. 507, Sep 2023, https://doi.org/10.1038/s41582-023-00857-4

28. Park SM, Ge TJ, Won DD, Lee JK, Liao JC. Digital biomarkers in human excreta, (in eng). Nat Rev Gastroenterol Hepatol. Aug 2021;18(8):521–2. https://doi.org/10.1038/s41575-021-00462-0.
29. Saleem M, et al. Exosome-based therapies for inflammatory disorders: a review of recent advances, (in eng). Stem Cell Res Ther. Dec 18 2024;15(1):477. https://doi.org/10.1186/s13287-024-04107-2.
30. Bie F et al. Multimodal analysis of cell-free DNA whole-methylome sequencing for cancer detection and localization, (in eng), *Nat Commun*, vol. 14, no. 1, p. 6042, Sep 27 2023, https://doi.org/10.1038/s41467-023-41774-w
31. Mutz J, Iniesta R, Lewis CM. Metabolomic age (MileAge) predicts health and life span: A comparison of multiple machine learning algorithms, (in eng), *Sci Adv*, vol. 10, no. 51, p. eadp3743, Dec 20 2024, https://doi.org/10.1126/sciadv.adp3743
32. Argentieri MA et al. Proteomic aging clock predicts mortality and risk of common age-related diseases in diverse populations, (in eng), *Nat Med*, vol. 30, no. 9, pp. 2450–2460, Sep 2024, https://doi.org/10.1038/s41591-024-03164-7
33. Deng YT et al. Atlas of the plasma proteome in health and disease in 53,026 adults, (in eng), *Cell*, vol. 188, no. 1, pp. 253–271.e7, Jan 9 2025, https://doi.org/10.1016/j.cell.2024.10.045
34. Narasaki Y et al. Accuracy of Continuous Glucose Monitoring in Hemodialysis Patients With Diabetes, (in eng), *Diabetes Care*, vol. 47, no. 11, pp. 1922–1929, Nov 1 2024. https://doi.org/10.2337/dc24-0635
35. Matusik PS, Matusik PT, Stein PK. Heart rate variability and heart rate patterns measured from wearable and implanted devices in screening for atrial fibrillation: potential clinical and population-wide applications, (in eng). Eur Heart J. Apr 1 2023;44(13):1105–7. https://doi.org/10.1093/eurheartj/ehac546.
36. Donkor R, Jammal AA, Greenfield DS. Relationship between Blood Pressure and Rates of Glaucomatous Visual Field Progression: The Vascular Imaging in Glaucoma Study, (in eng), *Ophthalmology*, vol. 132, no. 1, pp. 30–38, Jan 2025, https://doi.org/10.1016/j.ophtha.2024.07.026
37. Nôga DA et al. Habitual Short Sleep Duration, Diet, and Development of Type 2 Diabetes in Adults, (in eng), *JAMA Netw Open*, vol. 7, no. 3, p. e241147, Mar 4 2024, https://doi.org/10.1001/jamanetworkopen.2024.1147
38. Zhang BB et al. Monitoring long-term cardiac activity with contactless radio frequency signals, (in eng), *Nat Commun*, vol. 15, no. 1, p. 10598, Dec 5 2024, https://doi.org/10.1038/s41467-024-55061-9
39. Poisner H, Faucon A, Cox N, Bick AG. Genetic determinants and phenotypic consequences of blood T-cell proportions in 207,000 diverse individuals, (in eng), *Nat Commun*, vol. 15, no. 1, p. 6732, Aug 7 2024, https://doi.org/10.1038/s41467-024-51095-1
40. Ghetmiri DE, Venturi AJ, Cohen MJ, Menezes AA. Quick model-based viscoelastic clot strength predictions from blood protein concentrations for cybermedical coagulation control, (in eng), *Nat Commun*, vol. 15, no. 1, p. 314, Jan 5 2024, https://doi.org/10.1038/s41467-023-44231-w
41. Hein MY, et al. Global organelle profiling reveals subcellular localization and remodeling at proteome scale, (in eng). Cell Dec. 2024;26. https://doi.org/10.1016/j.cell.2024.11.028.
42. Embedding AI. in biology, (in eng), *Nat Methods*, vol. 21, no. 8, pp. 1365–1366, Aug 2024, https://doi.org/10.1038/s41592-024-02391-7
43. Pan L et al. Association of accelerated phenotypic aging, genetic risk, and lifestyle with progression of type 2 diabetes: a prospective study using multistate model, (in eng), *BMC Med*, vol. 23, no. 1, p. 62, Feb 4., 2025, https://doi.org/10.1186/s12916-024-03832-y
44. Longato E et al. Time-series analysis of multidimensional clinical-laboratory data by dynamic Bayesian networks reveals trajectories of COVID-19 outcomes, (in eng), *Comput Methods Programs Biomed*, vol. 221, p. 106873, Jun 2022, https://doi.org/10.1016/j.cmpb.2022.106873
45. Jamarani A, Haddadi S, Sarvizadeh R, Haghi Kashani M, Akbari M, Moradi S. Big data and predictive analytics: A systematic review of applications. Artif Intell Rev. 2024;57(7). https://doi.org/10.1007/s10462-024-10811-5.
46. Wang D, et al. A machine learning model for accurate prediction of Sepsis in ICU patients, (in eng). Front Public Health. 2021;9:754348. https://doi.org/10.3389/fpubh.2021.754348.
47. Mansmann U, Ön BI. The validation of prediction models deserves more recognition, (in eng). BMC Med. Mar 18 2025;23(1):166. https://doi.org/10.1186/s12916-025-03994-3.
48. Matheny ME et al. Enhancing Postmarketing Surveillance of Medical Products With Large Language Models, (in eng), *JAMA Netw Open*, vol. 7, no. 8, p. e2428276, Aug 1 2024, https://doi.org/10.1001/jamanetworkopen.2024.28276
49. Glasgow CG et al. CA-125 in Disease Progression and Treatment of Lymphangioleiomyomatosis, (in eng), *Chest*, vol. 153, no. 2, pp. 339–348, Feb 2018, https://doi.org/10.1016/j.chest.2017.05.018
50. Collister D et al. Variability in Cardiac Biomarkers during Hemodialysis: A Prospective Cohort Study, (in eng), *Clin Chem*, vol. 67, no. 1, pp. 308–316, Jan 8 2021. https://doi.org/10.1093/clinchem/hvaa299
51. Torres-Soto J, Ashley EA. Multi-task deep learning for cardiac rhythm detection in wearable devices, (in eng). NPJ Digit Med. 2020;3:116. https://doi.org/10.1038/s41746-020-00320-4.
52. Oliveira-Silva R, Sousa-Jerónimo M, Botequim D, Silva NJO, Paulo PMR, Prazeres DMF. Monitoring Proteolytic Activity in Real Time: A New World of Opportunities for Biosensors, (in eng), *Trends Biochem Sci*, vol. 45, no. 7, pp. 604–618, Jul 2020, https://doi.org/10.1016/j.tibs.2020.03.011
53. Singh M et al. Artificial intelligence for cardiovascular disease risk assessment in personalised framework: a scoping review, (in eng), *EClinicalMedicine*, vol. 73, p. 102660, Jul 2024, https://doi.org/10.1016/j.eclinm.2024.102660
54. D'Adderio L, Bates DW. Transforming diagnosis through artificial intelligence, (in eng), *NPJ Digit Med*, vol. 8, no. 1, p. 54, Jan 24 2025, https://doi.org/10.1038/s41746-025-01460-1
55. Shi H, et al. Air pollution associated with cardiopulmonary disease and mortality among participants with preserved ratio impaired spirometry, (in eng). Sci Total Environ. Nov 10 2024;950:175395. https://doi.org/10.1016/j.scitotenv.2024.175395.
56. Fang J et al. Personal PM(2.5) Elemental Components, Decline of Lung Function, and the Role of DNA Methylation on Inflammation-Related Genes in Older Adults: Results and Implications of the BAPE Study, (in eng), *Environ Sci Technol*, vol. 56, no. 22, pp. 15990–16000, Nov 15 2022. https://doi.org/10.1021/acs.est.2c04972
57. Imran S, Mahmood T, Morshed A, Sellis T. Big data analytics in healthcare– A systematic literature review and roadmap for practical implementation. IEEE/CAA J Automatica Sinica. 2020;8(1):1–22.
58. Pang X et al. Oct., Early warning COVID-19 outbreak in long-term care facilities using wastewater surveillance: correlation, prediction, and interaction with clinical and serological statuses, (in eng), *Lancet Microbe*, vol. 5, no. 10, p. 100894, 2024, https://doi.org/10.1016/s2666-5247(24)00126-5
59. Wang C et al. Integrating electronic health records and GWAS summary statistics to predict the progression of autoimmune diseases from preclinical stages, (in eng), *Nat Commun*, vol. 16, no. 1, p. 180, Jan 2 2025, https://doi.org/10.1038/s41467-024-55636-6
60. Pickering AJ et al. Fecal Indicator Bacteria along Multiple Environmental Transmission Pathways (Water, Hands, Food, Soil, Flies) and Subsequent Child Diarrhea in Rural Bangladesh, (in eng), *Environ Sci Technol*, vol. 52, no. 14, pp. 7928–7936, Jul 17 2018, https://doi.org/10.1021/acs.est.8b00928
61. O'Brien SJ, Halder SL. GI Epidemiology: infection epidemiology and acute gastrointestinal infections, (in eng), *Aliment Pharmacol Ther*, vol. 25, no. 6, pp. 669– 74, Mar 15 2007, https://doi.org/10.1111/j.1365-2036.2007.03245.x
62. Hunter RF et al. City mobility patterns during the COVID-19 pandemic: analysis of a global natural experiment, (in eng), *Lancet Public Health*, vol. 9, no. 11, pp. e896-e906, Nov 2024, https://doi.org/10.1016/s2468-2667(24)00222-6
63. Wen X et al. Clinlabomics: leveraging clinical laboratory data by data mining strategies, (in eng), *BMC Bioinformatics*, vol. 23, no. 1, p. 387, Sep 24 2022, https://doi.org/10.1186/s12859-022-04926-1
64. Heredia NI, Xu T, Lee M, McNeill LH, Reininger BM. The neighborhood environment and hispanic/latino health, (in eng). Am J Health Promot. Jan 2022;36(1):38–45. https://doi.org/10.1177/08901171211022677.
65. Li Y, et al. Identifying reference values for serum lipids in Chinese children and adolescents aged 6–17 years old: A National multicenter study, (in eng). J Clin Lipidol. May-Jun 2021;15(3):477–87. https://doi.org/10.1016/j.jacl.2021.02.001.
66. Romanello M et al. The 2024 report of the Lancet Countdown on health and climate change: facing record-breaking threats from delayed action, (in eng), *Lancet*, vol. 404, no. 10465, pp. 1847–1896, Nov 9 2024. https://doi.org/10.1016/s0140-6736(24)01822-1
67. Schwabe D, Becker K, Seyferth M, Klaß A, Schaeffter T. The METRIC-framework for assessing data quality for trustworthy AI in medicine: a systematic review. Npj Digit Med. 2024;7(1). https://doi.org/10.1038/s41746-024-01196-4.
68. Veedhi BK, Das K, Mishra D, Mishra S, Behera MP. Balancing data imbalance in biomedical datasets using a stacked augmentation approach with STDA, DAGAN, and pufferfish optimization to reveal AI's transformative impact. Int J Inform Technol,vol. 17, no. 1, pp. 455-480, 2025. https://doi.org/10.1007/s41870-024-02234-w

69. Mujahid M, et al. Data oversampling and imbalanced datasets: an investigation of performance for machine learning and feature engineering. J Big Data. 2024;11(1):87.
70. Sun X, Liu Y, An L. Ensemble dimensionality reduction and feature gene extraction for single-cell RNA-seq data, (in eng). Nat Commun. Nov 17 2020;11(1):5853. https://doi.org/10.1038/s41467-020-19465-7.
71. Wu Y, Burch KS, Ganna A, Pajukanta P, Pasaniuc B, Sankararaman S. Fast Estimation of genetic correlation for biobank-scale data, (in eng). Am J Hum Genet. Jan 6 2022;109(1):24–32. https://doi.org/10.1016/j.ajhg.2021.11.015.
72. Chen G, Zhang J, Fu Q, Taly V, Tan F. Integrative analysis of multi-omics data for liquid biopsy. Br J Cancer. 2023;128(4):505–18.
73. Jiang Y et al. Biology-guided deep learning predicts prognosis and cancer immunotherapy response, (in eng), *Nat Commun*, vol. 14, no. 1, p. 5135, Aug 23 2023, https://doi.org/10.1038/s41467-023-40890-x
74. Xu Y et al. Integrating Machine Learning in Metabolomics: A Path to Enhanced Diagnostics and Data Interpretation, (in eng), *Small Methods*, vol. 8, no. 12, p. e2400305, Dec 2024, https://doi.org/10.1002/smtd.202400305
75. Lewis AE et al. Electronic health record data quality assessment and tools: a systematic review, (in eng). J Am Med Inf Assoc, 30, 10, pp. 1730–40, Sep 25 2023, https://doi.org/10.1093/jamia/ocad120
76. Song XD et al. Ensure the accuracy and consistency of biochemical analyzer test results: Chemometrics for instrument and inter-instrument item comparison in Chinese hospital laboratory, (in eng), *Heliyon*, vol. 10, no. 1, p. e24306, Jan 15 2024, https://doi.org/10.1016/j.heliyon.2024.e24306
77. Sandhu PK, et al. Lipoprotein biomarkers and risk of cardiovascular disease: A laboratory medicine best practices (LMBP) systematic review, (in eng). J Appl Lab Med. Sep 1 2016;1(2):214–29. https://doi.org/10.1373/jalm.2016.021006.
78. Kather JN et al. Pan-cancer image-based detection of clinically actionable genetic alterations, (in eng), *Nat Cancer*, vol. 1, no. 8, pp. 789–799, Aug 2020, https://doi.org/10.1038/s43018-020-0087-6
79. Yuan AE, Shou W. Data-driven causal analysis of observational biological time series, (in eng), *Elife*, vol. 11, Aug 19 2022, https://doi.org/10.7554/eLife.72518
80. Ming W, et al. Early prediction model for disease progression of COVID-19 patients based on xgboost: establishment and evaluation. J Army Med Univ. 2022;44(3):195–202.
81. Chen K, Qin T, Lee VH, Yan H, Li H. Learning robust shape regularization for generalizable medical image segmentation, (in eng). IEEE Trans Med Imaging. Mar 2024;4. https://doi.org/10.1109/tmi.2024.3371987. Pp.
82. Linder J, Srivastava D, Yuan H, Agarwal V, Kelley DR. Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation, (in eng). Nat Genet Jan. 2025;8. https://doi.org/10.1038/s41588-024-02053-6.
83. Yang P et al. Spatial integration of multi-omics single-cell data with SIMO, (in eng), *Nat Commun*, vol. 16, no. 1, p. 1265, Feb 1 2025, https://doi.org/10.1038/s41467-025-56523-4
84. Sobahi N, Sengur A, Tan RS, Acharya UR. Attention-based 3D CNN with residual connections for efficient ECG-based COVID-19 detection, (in eng). Comput Biol Med. Apr 2022;143:105335. https://doi.org/10.1016/j.compbiomed.2022.105335.
85. Chattopadhyay S, Dey A, Singh PK, Oliva D, Cuevas E, Sarkar R. MTRRE-Net: A deep learning model for detection of breast cancer from histopathological images, (in eng). Comput Biol Med. Nov 2022;150:106155. https://doi.org/10.1016/j.compbiomed.2022.106155.
86. Khan RA, Fu M, Burbridge B, Luo Y, Wu FX. A multi-modal deep neural network for multi-class liver cancer diagnosis, (in eng), *Neural Netw*, vol. 165, pp. 553–561, Aug 2023, https://doi.org/10.1016/j.neunet.2023.06.013
87. Sarafraz G, Behnamnia A, Hosseinzadeh M, Balapour A, Meghrazi A, Rabiee HR. Domain adaptation and generalization of functional medical data: A systematic survey of brain data. ACM-CSUR. 2024;56(10):1–39.
88. Choi A et al. A novel deep learning algorithm for real-time prediction of clinical deterioration in the emergency department for a multimodal clinical decision support system, (in eng), *Sci Rep*, vol. 14, no. 1, p. 30116, Dec 3., 2024, https://doi.org/10.1038/s41598-024-80268-7
89. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement, (in eng), *Bmj*, vol. 350, p. g7594, Jan 7 2015, https://doi.org/10.1136/bmj.g7594
90. Grim S, Kotz A, Kotz G, Halliwell C, Thomas JF, Kessler R. Development and validation of electronic health record-based, machine learning algorithms to predict quality of life among family practice patients, (in eng). Sci Rep. Dec 3 2024;14(1):30077. https://doi.org/10.1038/s41598-024-80064-3.
91. Li J, Wang S, Rudinac S, Osseyran A. High-performance computing in healthcare: an automatic literature analysis perspective. J Big Data. 2024;11(1). https://doi.org/10.1186/s40537-024-00929-2.
92. Wang C, Luo Y, Du W, Wang K, Gu N, Yu J. Faster and stronger: unleashing data processing potential through hardware heterogeneity. IEEE Internet Things J, vol. 12, no. 10, pp. 14559-14576, 15 May15, 2025. https://doi.org/10.1109/JIOT.2025.3526662
93. La Cava WG et al. A flexible symbolic regression method for constructing interpretable clinical prediction models, (in eng), *NPJ Digit Med*, vol. 6, no. 1, p. 107, Jun 5., 2023, https://doi.org/10.1038/s41746-023-00833-8
94. Nasarian E, Alizadehsani R, Acharya UR, Tsui K-L. Designing interpretable ML system to enhance trust in healthcare: A systematic review to proposed responsible clinician-AI-collaboration framework. Inform Fusion, Vol. 108, pp. 102412, 2024. https://doi.org/10.1016/j.inffus.2024.102412
95. Jiang L, et al. Autosurv: interpretable deep learning framework for cancer survival analysis incorporating clinical and multi-omics data, (in eng). NPJ Precis Oncol. p. 4, Jan 5 2024;8(1). https://doi.org/10.1038/s41698-023-00494-6.
96. Hamedi SZ, et al. Application of machine learning in breast cancer survival prediction using a multimethod approach. Sci Rep. 2024;14(1). https://doi.org/10.1038/s41598-024-81734-y.
97. Ravichandran D, Jebarani WSL, Mahalingam H, Meikandan PV, Pravinkumar P, Amirtharajan R. An efficient medical data encryption scheme using selective shuffling and inter-intra pixel diffusion IoT-enabled secure E-healthcare framework. (in eng) Sci Rep. Feb 3 2025;15(1):4143. https://doi.org/10.1038/s41598-025-85539-5.
98. Kaabachi B et al. A scoping review of privacy and utility metrics in medical synthetic data, (in eng), *NPJ Digit Med*, vol. 8, no. 1, p. 60, Jan 27 2025, https://doi.org/10.1038/s41746-024-01359-3
99. Abouelmehdi K, Beni-Hessane A, Khaloufi H. Big healthcare data: preserving security and privacy. J Big Data 5 (1)(2018), ed. https://doi.org/10.1186/s40537-017-0110-7
100. Xu J et al. Multi-layer encryption of medical data in DNA for highly-secure storage, (in eng), *Mater Today Bio*, vol. 28, p. 101221, Oct 2024, https://doi.org/10.1016/j.mtbio.2024.101221
101. Fang C et al. Decentralised, collaborative, and privacy-preserving machine learning for multi-hospital data, (in eng), *EBioMedicine*, vol. 101, p. 105006, Mar 2024, https://doi.org/10.1016/j.ebiom.2024.105006
102. Chen W, et al. Mask-aware transformer with structure invariant loss for CT translation, (in eng). Med Image Anal. Aug 2024;96:103205. https://doi.org/10.1016/j.media.2024.103205.
103. Froelicher D, et al. Truly privacy-preserving federated analytics for precision medicine with multiparty homomorphic encryption. Nat Commun. 2021;12(1). https://doi.org/10.1038/s41467-021-25972-y.
104. Jeon S et al. Proposal and Assessment of a De-Identification Strategy to Enhance Anonymity of the Observational Medical Outcomes Partnership Common Data Model (OMOP-CDM) in a Public Cloud-Computing Environment: Anonymization of Medical Data Using Privacy Models, (in eng), *J Med Internet Res*, vol. 22, no. 11, p. e19597, Nov 26 2020, https://doi.org/10.2196/19597
105. Kushida CA, Nichols DA, Jadrnicek R, Miller R, Walsh JK, Griffin K. Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies, (in eng). Med Care. Jul 2012;50:S82–101. https://doi.org/10.1097/MLR.0b013e3182585355. Suppl, no. Suppl.
106. Kaissis G, et al. End-to-end privacy preserving deep learning on multi-institutional medical imaging. Nat Mach Intell. 2021;3(6):473–84.
107. Zhou J, et al. PPML-Omics: A privacy-preserving federated machine learning method protects patients' privacy in omic data, (in eng). Sci Adv. Feb 2 2024;10(5):eadh8601. https://doi.org/10.1126/sciadv.adh8601.
108. Kumar-M P, Mishra A. Conducting randomization in clinical trials. In: Srinivasan, A., Mishra, A., Kumar-M, P. (eds)R for basic biostatistics in medical research. Springer; 2024. pp. 269–88. https://doi.org/10.1007/978-981-97-6980-3_14
109. Gadotti A, Rocher L, Houssiau F, Creţu AM, de Montjoye YA. Anonymization: The imperfect science of using data while preserving privacy, (in eng), *Sci Adv*, vol. 10, no. 29, p. eadn7053, Jul 19 2024, https://doi.org/10.1126/sciadv.adn7053
110. Kaissis GA, Makowski MR, Rückert D, Braren RF. Secure, privacy-preserving and federated machine learning in medical imaging. Nat Mach Intell. 2020;2(6):305–11.

111. Bi X, Shen X. Distribution-Invariant Differential Privacy, (in eng), *J Econom*, vol. 235, no. 2, pp. 444–453, Aug 2023, https://doi.org/10.1016/j.jeconom.2022.05.004
112. Ficek J, Wang W, Chen H, Dagne G, Daley E. Differential privacy in health research: A scoping review, (in eng). J Am Med Inf Assoc, 28, 10, pp. 2269–76, Sep 18 2021, https://doi.org/10.1093/jamia/ocab135
113. Ho TT, Tran KD, Huang Y. FedSGDCOVID: federated SGD COVID-19 detection under local differential privacy using chest X-ray images and symptom information, (in eng). Sens (Basel). May 13 2022;22(10). https://doi.org/10.3390/s22103728.
114. Tang Z, et al. High security and privacy protection model for STI/HIV risk prediction, (in eng). Digit Health. Jan-Dec 2024;10:20552076241298425. https://doi.org/10.1177/20552076241298425.
115. Singh P, Gaba GS, Kaur A, Hedabou M, Gurtov A. Dew-Cloud-Based hierarchical federated learning for intrusion detection in IoMT. IEEE J Biomedical Health Inf. 2023;27(2):722–31. https://doi.org/10.1109/jbhi.2022.3186250.
116. Shojima N, Yamauchi T. Progress in genetics of type 2 diabetes and diabetic complications, (in eng). J Diabetes Investig. Apr 2023;14(4):503–15. https://doi.org/10.1111/jdi.13970.
117. Smith M, Sattler A, Hong G, Lin S. From code to bedside: implementing artificial intelligence using quality improvement methods, (in eng). J Gen Intern Med. Apr 2021;36(4):1061–6. https://doi.org/10.1007/s11606-020-06394-w.
118. Yang W, Wang S, Cui H, Tang Z, Li Y. A Review of Homomorphic Encryption for Privacy-Preserving Biometrics, (in eng), *Sensors (Basel)*, vol. 23, no. 7, Mar 29 2023, https://doi.org/10.3390/s23073566
119. .
120. Safa M, Pandian A, Gururaj H, Ravi V, Krichen M. Real time health care big data analytics model for improved QoS in cardiac disease prediction with IoT devices. Health Technol. 2023;13(3):473–83.
121. Munjal K, Bhatia R. A systematic review of homomorphic encryption and its contributions in healthcare industry, (in eng), *Complex Intell Systems*, pp. 1–28, May 3 2022. https://doi.org/10.1007/s40747-022-00756-z
122. Geva R et al. Collaborative privacy-preserving analysis of oncological data using multiparty homomorphic encryption, (in eng), *Proc Natl Acad Sci U S A*, vol. 120, no. 33, p. e2304415120, Aug 15 2023, https://doi.org/10.1073/pnas.2304415120
123. Sheu RK, et al. Adaptive autonomous protocol for secured remote healthcare using fully homomorphic encryption (AutoPro-RHC), (in eng). Sens (Basel). Oct 16 2023;23(20). https://doi.org/10.3390/s23208504.
124. Rohanian O, et al. Privacy-Aware early detection of COVID-19 through adversarial training, (in eng). IEEE J Biomed Health Inf. Pp, no. Dec 20 2022;3:1249–58. https://doi.org/10.1109/jbhi.2022.3230663.
125. Feng J, Xia F, Singh K, Pirracchio R. Not all clinical AI monitoring systems are created equal: review and recommendations. NEJM AI. 2025;2(2):AIra2400657. https://doi.org/10.1056/AIra2400657.
126. Yao S, Dai F, Sun P, Zhang W, Qian B, Lu H. Enhancing the fairness of AI prediction models by Quasi-Pareto improvement among heterogeneous thyroid nodule population, (in eng), *Nat Commun*, vol. 15, no. 1, p. 1958, Mar 4 2024. https://doi.org/10.1038/s41467-024-44906-y
127. Haw JS, Shah M, Turbow S, Egeolu M, Umpierrez G. Diabetes Complications in Racial and Ethnic Minority Populations in the USA, (in eng), *Curr Diab Rep*, vol. 21, no. 1, p. 2, Jan 9 2021, https://doi.org/10.1007/s11892-020-01369-x
128. Mah JC, Stevens SJ, Keefe JM, Rockwood K, Andrew MK. Social factors influencing utilization of home care in community-dwelling older adults: a scoping review, (in eng). BMC Geriatr. Feb 27 2021;21(1):145. https://doi.org/10.1186/s12877-021-02069-1.
129. Aggarwal R, Chiu N, Loccoh EC, Kazi DS, Yeh RW, Wadhera RK. Rural-Urban disparities: diabetes, hypertension, heart disease, and stroke mortality among black and white adults, 1999–2018, (in eng). J Am Coll Cardiol, 77, 11, pp. 1480–1, Mar 23 2021, https://doi.org/10.1016/j.jacc.2021.01.032
130. Cohen JA, et al. The changing health impact of vaccines in the COVID-19 pandemic: A modeling study, (in eng). Cell Rep. Apr 25 2023;42(4):112308. https://doi.org/10.1016/j.celrep.2023.112308.
131. Hicks CW, Wang D, Matsushita K, Windham BG, Selvin E. Peripheral Neuropathy and All-Cause and Cardiovascular Mortality in U.S. Adults: A Prospective Cohort Study, (in eng), *Ann Intern Med*, vol. 174, no. 2, pp. 167–174, Feb 2021, https://doi.org/10.7326/m20-1340
132. Global regional. and national burden of diabetes from 1990 to 2021, with projections of prevalence to 2050: a systematic analysis for the Global Burden of Disease Study 2021, (in eng), *Lancet*, vol. 402, no. 10397, pp. 203–234, Jul 15 2023. https://doi.org/10.1016/s0140-6736(23)01301-6
133. Szepannek G, Lübke K. Facing the challenges of developing fair risk scoring models, (in eng). Front Artif Intell. 2021;4:681915. https://doi.org/10.3389/frai.2021.681915.
134. Paulus JK, Kent DM. Predictably unequal: Understanding and addressing concerns that algorithmic clinical prediction May increase health disparities, (in eng). NPJ Digit Med. 2020;3:99. https://doi.org/10.1038/s41746-020-0304-9.
135. Nwafor CN, Nwafor O, Brahma S. Enhancing transparency and fairness in automated credit decisions: an explainable novel hybrid machine learning approach, (in eng), *Sci Rep*, vol. 14, no. 1, p. 25174, Oct 24 2024, https://doi.org/10.1038/s41598-024-75026-8
136. Guo D, Wang C, Wang B, Zha H. Learning fair representations via distance correlation minimization, (in eng). IEEE Trans Neural Netw Learn Syst. Feb 2024;35(2):2139–52. https://doi.org/10.1109/tnnls.2022.3187165.
137. Saharan SS et al. Logistic Regression and Statistical Regularization Techniques for Risk Classification of Coronary Artery Disease using Cytokines transported by high density lipoproteins, (in eng), *Proc (Int Conf Comput Sci Comput Intell)*, vol. 2023, pp. 652–660, Dec 2023. https://doi.org/10.1109/csci62032.2023.00114
138. Khanijahani A, Iezadi S, Agoglia S, Barber S, Cox C, Olivo N. Factors Associated with Information Breach in Healthcare Facilities: A Systematic Literature Review, (in eng), *J Med Syst*, vol. 46, no. 12, p. 90, Nov 2 2022, https://doi.org/10.1007/s10916-022-01877-1
139. Choi SJ, Johnson ME, Lee J. An event study of data breaches and hospital IT spending. Health Policy Technol. 2020;9(3):372–8.
140. Gabriel MH, Noblin A, Rutherford A, Walden A, Cortelyou-Ward K. Data breach locations, types, and associated characteristics among US hospitals, (in eng). Am J Manag Care. vol. 24, no. 2, pp. 78-84, Feb 2018.
141. Neprash HT et al. Trends in Ransomware Attacks on US Hospitals, Clinics, and Other Health Care Delivery Organizations, 2016–2021, (in eng), *JAMA Health Forum*, vol. 3, no. 12, p. e224873, Dec 2 2022, https://doi.org/10.1001/jamahealthforum.2022.4873
142. Seh AH et al. Healthcare Data Breaches: Insights and Implications, (in eng), *Healthcare (Basel)*, vol. 8, no. 2, May 13 2020, https://doi.org/10.3390/healthcare8020133
143. Vasey B et al. Association of Clinician Diagnostic Performance With Machine Learning-Based Decision Support Systems: A Systematic Review, (in eng), *JAMA Netw Open*, vol. 4, no. 3, p. e211276, Mar 1 2021, https://doi.org/10.1001/jamanetworkopen.2021.1276
144. Hicks SA, et al. On evaluation metrics for medical applications of artificial intelligence, (in eng). Sci Rep. Apr 8 2022;12(1):5979. https://doi.org/10.1038/s41598-022-09954-8.
145. Christiansen F, et al. International multicenter validation of AI-driven ultrasound detection of ovarian cancer, (in eng). Nat Med. Jan 2025;31(1):189–96. https://doi.org/10.1038/s41591-024-03329-4.
146. Schopf CM et al. Artificial Intelligence-Driven Mammography-Based Future Breast Cancer Risk Prediction: A Systematic Review, (in eng), *J Am Coll Radiol*, vol. 21, no. 2, pp. 319–328, Feb 2024, https://doi.org/10.1016/j.jacr.2023.10.018
147. Hu L et al. Enhancing fairness in AI-enabled medical systems with the attribute neutral framework, (in eng), *Nat Commun*, vol. 15, no. 1, p. 8767, Oct 10 2024, https://doi.org/10.1038/s41467-024-52930-1
148. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models, (in eng), *Med Decis Making*, vol. 26, no. 6, pp. 565– 74, Nov-Dec 2006, https://doi.org/10.1177/0272989x06295361
149. Van Calster B et al. Dec., Reporting and Interpreting Decision Curve Analysis: A Guide for Investigators, (in eng), *Eur Urol*, vol. 74, no. 6, pp. 796–804, 2018, https://doi.org/10.1016/j.eururo.2018.08.038
150. Peters M et al. Predicting the Need for Biopsy to Detect Clinically Significant Prostate Cancer in Patients with a Magnetic Resonance Imaging-detected Prostate Imaging Reporting and Data System/Likert ≥ 3 Lesion: Development and Multinational External Validation of the Imperial Rapid Access to Prostate Imaging and Diagnosis Risk Score, (in eng), *Eur Urol*, vol. 82, no. 5, pp. 559–568, Nov 2022, https://doi.org/10.1016/j.eururo.2022.07.022
151. Alkhuzam K, et al. Long-term health benefit and economic return of time in range (TIR) improvement in individuals with type 2 diabetes, (in eng). Diabetes Obes Metab Jan. 2025;8. https://doi.org/10.1111/dom.16168.
152. Mihaylova B et al. Assessing long-term effectiveness and cost-effectiveness of statin therapy in the UK: a modelling study using individual participant data sets, (in eng), *Health Technol Assess*, vol. 28, no. 79, pp. 1-134, Dec 2024, https://doi.org/10.3310/kdap7034

153. El-Hay T, Reps JM, Yanover C. Extensive benchmarking of a method that estimates external model performance from limited statistical characteristics, (in eng). NPJ Digit Med. Jan 27 2025;8(1):59. https://doi.org/10.1038/s41746-024-01414-z.
154. Huntley C et al. Utility of polygenic risk scores in UK cancer screening: a modelling analysis, (in eng), *Lancet Oncol*, vol. 24, no. 6, pp. 658–668, Jun 2023, https://doi.org/10.1016/s1470-2045(23)00156-0
155. Goodrich JA, et al. Integrating Multi-Omics with environmental data for precision health: A novel analytic framework and case study on prenatal mercury induced childhood fatty liver disease, (in eng). Environ Int. Aug 2024;190:108930. https://doi.org/10.1016/j.envint.2024.108930.
156. Wang J, et al. The clinical application of artificial intelligence in cancer precision treatment, (in eng). J Transl Med. Jan 27 2025;23(1):120. https://doi.org/10.1186/s12967-025-06139-5.

## Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.