Volume 3, Number 3, 2026

Pages 1665-1675

Bioscience Insights

ISSN: 3080-3489 (Print)
ISSN: 3080-3497 (Online)

BioSciInsights.com

ARTIFICIAL INTELLIGENCE AND DIGITAL HEALTH EQUITY IN PREVENTIVE HEALTHCARE: DEVELOPING AN EXPLAINABLE FRAMEWORK TO IDENTIFY UNDERSERVED POPULATIONS THROUGH INTEGRATION OF HEALTHCARE ACCESSIBILITY, SOCIOECONOMIC CONDITIONS, DIGITAL LITERACY, ENVIRONMENTAL EXPOSURE, AND COMMUNITY HEALTH INDICATORS

Humaira Mehar

Emergency Operation Theatre, Pediatrics, The Children's Hospital & University of Child Health Sciences, Lahore, Pakistan

humairamehar9@gmail.com

https://doi.org/10.5281/zenodo.24126894

Article Information

Article History

Received:	June 19, 2026
Revised:	August 19, 2026
Accepted:	September 07, 2026
Available Online:	October 30, 2026

Keywords:

Explainable Artificial Intelligence (XAI), Digital Health Equity, Preventive Healthcare, Algorithmic Bias, Social Determinants of Health, Digital Determinants of Health, Federated Learning, Health Disparities.

Abstract

The fast arrival of intelligence (AI) and machine learning (ML) in preventive health gives new ways to spot disease early and give the right help. These tools can also make health gaps bigger when they learn from old data that is unfair. In this review we show an idea that helps find people who are left out by looking at five big areas: how easy it is to get care how well people are doing money-wise how good people are at using technology what the environment looks like and how strong the local health is. Looking at all the studies, we see that just the usual medical risk marks do not give fair results. We need to add the Digital Determinants of Health (DDOH) small environmental dangers and the big gaps that keep people down to make predictions. The review also looks closely at the challenge of making models that work well. Are easy to understand. We check open-box models after-the-fact tools such as SHAP and LIME and counter-story explanations while keeping in mind the "Table 2 Fallacy" that mistakes chance links for causes. To fix these method and moral problems we suggest an Integrated Explainable Equity Framework. It brings together kinds of data safely keeps fairness all the way through the model's life and uses causal XAI rules to check how the model acts. The framework puts people in charge lets communities help design it and uses low-energy computers so that big health AI does not hurt the planet. We finish by saying that real digital health fairness needs to move from fixing bias after the fact, to guarding fairness from the start. That way explainable AI helps many experts decide together not act on who gets care.

1. INTRODUCTION

The integration of intelligence and machine learning into public health and clinical systems marks a major shift in how we prevent disease. By spotting patterns in many types of data—biomedical, environmental and socio-demographic—predictive algorithms can find early signs of disease predict how many people will get sick and help us plan targeted prevention (Kelly et al. 2019).

However this fast growth in technology happens in a world where health inequities still exist. If we train algorithms on old data that contain social biases, unequal resources or systematic exclusion those predictive algorithms can repeat and even increase those problems while claiming to be objective data science (Rajkomar et al. 2018).

Digital health equity means every person and community should have a chance to benefit from digital health innovations no matter their race, where they live how much money they have or how tech-savvy they are. If we do not keep equity in care we see delays in diagnosis wasted prevention resources and higher avoidable death rates in historically underserved groups (Crawford & Serhal 2020). To fix these problems we need to make predictive artificial intelligence look beyond standard clinical risk factors and include broader social, digital and environmental determinants of health (Joseph, 2025).

At the time using complex black-box machine learning models creates serious problems with transparency and validation in public health policy. Big models like neural networks and ensemble algorithms often give up clear interpretability for small boosts in prediction accuracy. When these algorithms sort health risk in a population without decision logic, clinicians and policy makers cannot tell if the results come from real physiological mechanisms, actionable environmental factors or hidden data biases (Agarwal et al., 2023). Explainable Artificial Intelligence gives tools to look at how these models behave. Post-hoc XAI methods are often used as if they explain cause when they really only describe statistics, which can lead to misinterpretation, in allocating resources (Lundberg & Lee 2017).

This review offers a framework that brings together five key domains: healthcare accessibility, socioeconomic conditions, digital literacy, environmental exposure and community health indicators. By looking at health equity as a combined social and technical problem this report examines how explainable artificial intelligence can find underserved groups accurately reduce algorithmic bias throughout the software life cycle and help create data-driven preventive care plans (Suresh & Guttag 2021).

2. Multidimensional Determinants of Health Inequity in the Digital Era

Achieving health equity through preventive care requires moving beyond individual clinical markers to analyze the broader structural conditions that shape health outcomes. In modern digital public health, disparities stem from complex interactions between physical, economic, technological, and environmental factors (Amann et al., 2020).

Table 1: Key Dimensions and Indicator Domains for Integrated Digital Health Equity Assessment

Domain Dimension	Primary Data Variables & Indicators	Data Source Paradigms	Mechanism of Equity Impact
Healthcare Accessibility	Travel time to primary care, clinician-to-patient ratios, insurance coverage type, facility capacity, wait times	GIS mapping, claims databases, electronic health records (EHR)	Geographical and structural barriers delay preventive screenings, leading to late-stage interventions.
Socioeconomic Conditions	Median household income, educational attainment, poverty index, housing stability, employment status	Census data, longitudinal cohort studies (e.g., HRS, SHARE, ELSA)	Financial constraints limit access to nutritious food, preventive therapies, and stable living environments.
Digital Literacy & DDOH	Broadband availability, smart device ownership, digital skills proficiency, portal usage rates	Municipal telecom surveys, user engagement logs, broadband maps	Lower digital literacy and poor connectivity block access to telemedicine,

			digital portals, and remote monitoring.
Environmental Exposure	$PM_{2.5}$ levels, ground-level ozone, extreme heat days, industrial proximity, built environment walkability	Satellite imagery, IoT environmental sensors, EPA monitoring feeds	Micro-level ecological hazards trigger chronic pulmonary, cardiovascular, and metabolic conditions.
Community Health Indicators	Local chronic disease prevalence, age-adjusted mortality, vaccination rates, localized social capital	Public health registries, municipal surveillance, community surveys	Population-level disease burdens signal systemic shortfalls in local public health infrastructure.

2.1 Healthcare Accessibility and Systemic Friction

Healthcare access is defined as physical, structural, and operational barriers to treating a person in a timely manner for preventive care. Geographical distance to primary care facilities has remained the largest cause of missed appointments particularly in rural or resource-poor urban settings (Richardson et al., 2022). However, systemic friction also exists in the administrative operations of the health care system. These algorithms that speed up clinic operations via automated scheduling are designed for optimization often LW patients with historical non-attendance can get less valuable appointment slots. Non-attendance is often due to causes beyond the individual, such as inflexible hour work schedules or lack of reliable public transportation; these algorithms can compound systemic exclusion, resulting in higher dropout rates for racial minority and low-income populations (Chin et al., 2023).

2.2 Socioeconomic Conditions and Structural Deprivation

Socioeconomic status, which includes income, education, occupation, and distribution of wealth, is a fundamental driver of population health. Structural deprivation directly affects a person and its capabilities to maintain their health through prevention socioeconomic deprivation has a direct impact on an person's ability to take preventive measures to maintain some measure of health. A lower level of educational attainment is often associated with weaker health literacy which may hinder individuals from navigating high cognitive burden care guidelines, meaning making risk communicating and chronic disease self-management strategies (Obermeyer et al., 2019). Several longitudinal and multi-country studies show that socioeconomic predictors not merely universal baseline factors for mortality, such as age and sex determine patterns of avoidable disease burden among middleaged and older cohorts (Paulus & Kent, 2020).

2.3 Digital Literacy and the Digital Determinants of Health

In the shift towards digital-first delivery models of healthcare systems, Digital Determinants of Health (DDOH) have become pivotal factors influencing the overall outcomes of population health. DDOH comprise elements key to technology, including infrastructure like broadband, hardware access, and digital literacy on an individual level. Digital literacy is the ability to benefit from health information delivered through digital interfaces in terms of access, interpretation, evaluation, and action (Wiens et al., 2019).

Health care providers who put patient portals, remote monitoring systems, or app-based preventative programs in place without assessing the underlying DDOH have the potential to construct a selective benefit pattern. These tools are more likely to be adopted by the technologically well-off and digitally literate populations (known as the "worried well"); however, digitally excluded communities face increased difficulties. These variables transform digital tools from equity enablers into mechanisms that amplify widespread health disparities (Nazer et al., 2023).

2.4 Environmental Exposure and Micro-Level Ecological Hazards

Environmental exposure represents a critical driver of health disparities, particularly in underserved communities disproportionately located near industrial zones, high-traffic corridors, or urban heat

islands. Continuous exposure to fine particulate matter ($PM_{2.5}$), nitrogen dioxide (NO_2), and ambient heat stress increases the risk of chronic conditions, including chronic obstructive pulmonary disease (COPD), cardiovascular disease, and adverse metabolic outcomes (Rudin, 2019).

The compound relationship between environmental hazards, socioeconomic vulnerability, and healthcare accessibility can be formulated as an integrated risk function:

$$P_{\text{risk}} = f(E) \cdot g(S) \cdot \frac{1}{h(A)}$$

where E represents environmental exposure metrics, S captures socioeconomic vulnerability vectors, and A measures healthcare and digital accessibility capacity. Integrating fine-grained environmental sensor data and satellite imagery into population health AI models allows systems to identify ecological hazard clusters that traditional clinical trials often overlook (Vayena et al., 2018).

2.5 Community Health Indicators and Population Dynamics

Community health indicators such as co-morbidity rates, vaccination coverage, disease incidence, and social capital indicators at the local level provide macro-level context for individual's health trajectory. High community disease burdens are a result of years of underinvestment in local public health infrastructure. The use of community-level indicators in machine learning frameworks contextualizes individual risk scores where the lack of public health resources is incorrectly blamed on the individual patient behaviours (Wiessing,2017).

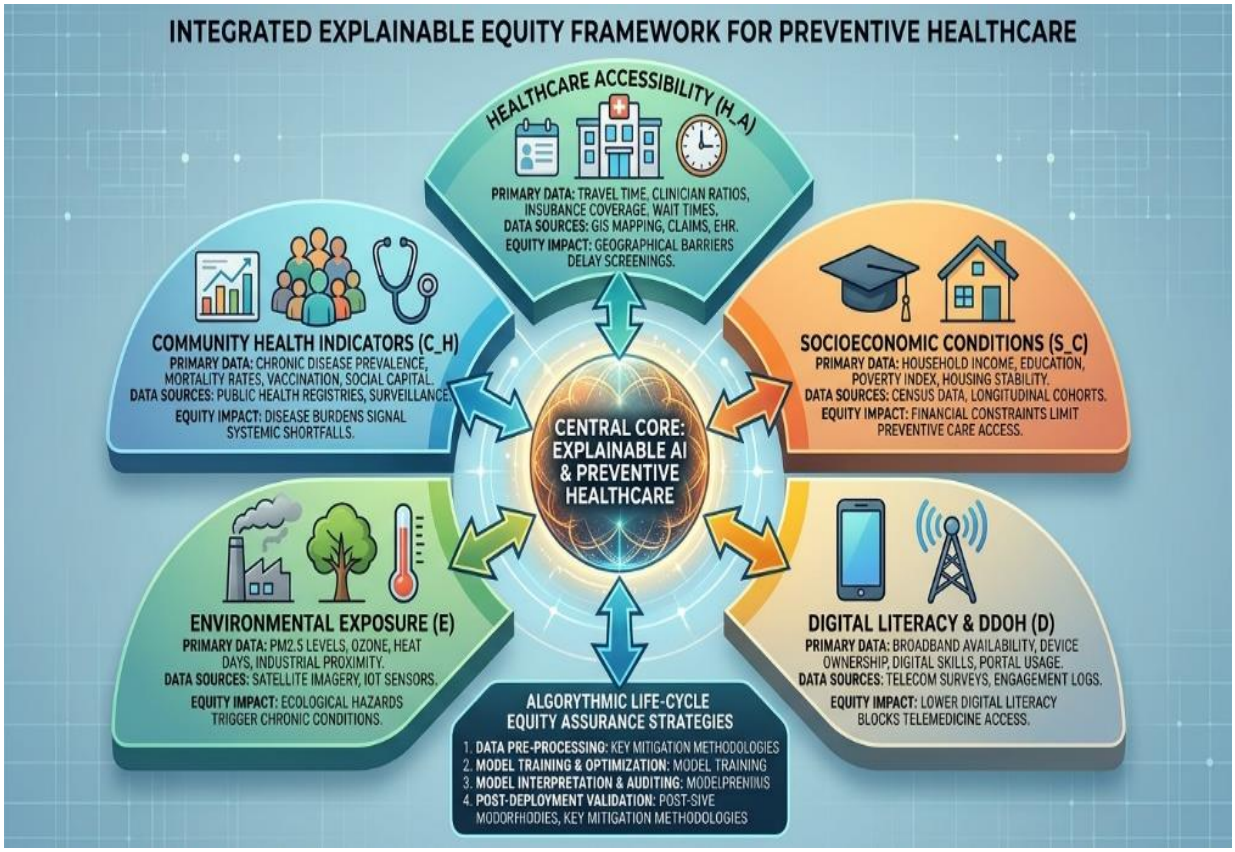


Figure 1: The Five Vector Domains of Digital Health Equity

3. Artificial Intelligence in Preventive Healthcare and Risk Stratification

The core objective of AI in preventive healthcare is to shift intervention timelines from reactive management to proactive prevention. Machine learning models excel at identifying multi-dimensional interactions across heterogeneous datasets, enabling early risk stratification for chronic conditions, oncology, and acute disease exacerbations (Dekkers & Desmet, 2013).

3.1 Machine Learning Modalities and High-Dimensional Risk Modeling

Modern risk stratification harnesses a range of algorithm-based techniques, from interpretable regression and decision trees to sophisticated deep neural networks and ensemble methods, such as Gradient Boosted Decision Trees. Multiple say this (Fridkin & Bendersky, 2026) but popular algorithms for big population health prediction tasks, such as LightGBM achieve higher discrimination (PR-AUC) performance even on mixed tabular data sets with features including hundreds socio-demographic, environmental, and clinical factors (Fridkin & Bendersky, 2026). Ensemble algorithms can learn complex, non-linear interactions such as the additive contributions from air pollution, low income and lack of primary care access without manually creating and engineering features (Abbas et al., 2026).

3.2 Data Bias, Representation Shifts, and the Mechanics of Algorithmic Harm

Nonetheless, the predictive ability of AI systems comes with the danger of instigating heavy bias if the training data underlying the model does not accurately reflect the target population. Data representation bias arises when a protected group or demographic segment (e.g. racial minorities, rural dwellers, or low-income populations) are underrepresented in training cohorts relative to their relative incidence in the real world. This leads to predictive models learning statistical patterns optimized for majority groups, which create greater levels of error and misclassification when used in minoritized populations (Norori et al., 2021).

For example, two studies show that for models trained on clinical data, non-English language preference is underrepresented and consequently there are large misclassification errors. In particular, in a primary care deployment with known

disparities, 20% of patients who declared a preference for Spanish were incorrectly classified as English-only based on the unbalanced training set (Kumar et al., 2026). Likewise, a managerial approach that equates health care utilization with health care need for development of an algorithmic policy to direct efficient health care spending inevitably favours wealthier organizations: historical underinvestment in low-income and minority communities due to funding gaps leads the algorithms to think that these individuals are healthier than they are, and as a result consistently under-allocate resources aimed at prevention (Ntoutsi et al., 2020). Moreover, current analyses of multi-ethnic cohorts indicate that adverse clinical quality outcomes is shaped strongly by both race and place, a fact that indicates the need for building models that target equity (Chinta et al., 2025).

4. Explainable AI (XAI) and Algorithmic Fairness

To ensure AI models contribute to health equity rather than deepening existing disparities, model predictions must be interpretable, transparent, and auditable. Explainable AI (XAI) encompasses methodologies designed to clarify how complex algorithms generate outputs, offering mechanisms to audit models for demographic bias and unmeasured confounding (Alghamdi et al., 2025).

Table 2: Comparative Analysis of Explainable AI (XAI) Methodologies and Causal Limitations in Health Inequity Modeling

XAI Approach	Methodology	Primary Advantages	Critical Vulnerabilities & Equity Risks
Glass-Box Models (*e.g., Decision Trees, Generalized Additive Models [GAMs], Explainable Boosting Machines [EBMs]*)	Model architectures designed to be inherently transparent, mapping input features directly to risk scores without post-hoc approximations.	Full intrinsic transparency; simple auditability; eliminates post-hoc explanation uncertainty.	May sacrifice marginal predictive accuracy on ultra-high-dimensional, non-linear data structures.
Post-Hoc Feature Attribution (*e.g., SHAP [Shapley*	Additive feature attribution techniques derived from cooperative game	Applicable to complex, black-box architectures (e.g., LightGBM, Deep	**Table 2 Fallacy**: Recovers statistical patterns, not causal mechanisms. High sensitivity

Additive exPlanations], LIME)	theory; calculates marginal contribution of features to specific model outputs.	Neural Networks); localized risk attribution.	to feature correlation/collinearity.
Counterfactual Explanations	Computes the minimum shift in feature values required to flip a prediction outcome (e.g., moving a patient from high-risk to low-risk).	Direct operational guidance; identifies actionable targets for individual or public health intervention.	May suggest unrealistic or non-actionable feature changes (e.g., changing age or race) if non-modifiable features are unconstrained.

4.1 Intrinsic Interpretability versus Post-Hoc Attribution

The XAI landscape is broadly split between inherently interpretable ("glass-box") models and black-box models to which post-hoc attribution methods are applied. Glass-box algorithms, e.g., Explainable Boosting Machines (EBMs) or sparse decision trees, restrict functional interactions so that human experts can directly explore decision boundaries. In contrast most post-hoc attribution techniques most notably SHAP (Shapley Additive exPlanations) use model training to estimate feature contributions by evaluating feature permutations (Linardatos et al., 2020).

SHAP beeswarm plots are a common way to demonstrate global feature importance among population cohorts but capture only feature importance in association with other, unrelated features rather than establishing the decision processes. Most post-hoc techniques work poorly when confronted by highly correlated features as when income is totally dominated by zip code, digital literacy, or health care access metrics they together explain health disparities and the factors (xai contributors) that cause health disparities (Fridkin & Bendersky, 2026).

4.2 The Causality Dilemma and the Table 2 Fallacy in Health AI

The danger of applying XAI to health equity. Attributing a statistical explanation in terms of causal inference. In classical epidemiology, treating every multivariable regression coefficient as a causal effect estimate is known as the "Table 2 Fallacy". A similar fallacy of attribute causes and effects occurs in explainable AI when policy makers interpret SHAP and LIME attributions as causal drivers of outcome disparities (Coupland et al., 2025). For example, in a disease prediction model built with explainable AI that identifies kidney disease risk and vents a high SHAP value to hypertension and 0 SHAP to smoking, a noncausal interpretation might conclude that smoking is not a risk factor for kidney disease in this population (Zhang et al., 2026). However, this is not true since there's something about hypertension that is in the causal chain of effects downstream of smoking, or hypertension status has maybe overlapped information, that the predictive algorithm just statistically bets to the greatest extent on hypertension as the short-hand explanatory variable. In the management of preventive healthcare services, interpreting post-hoc model attribution should be treated with caution as a causal target, otherwise health systems and policymakers may mistakenly redirect investments away from the upstream structural, environmental, and digital drivers of inequality to target downstream biological symptoms (Rafe & Das, 2025).

5. Conceptual Architecture of the Integrated Explainable Equity Framework

Here, we propose an Integrated Explainable Equity Framework to overcome these methodological and structural issues. Using multi-source data streams, it adopts privacy-preserving model training, incorporates fairness constraints during model development, and through causal principles of XAI, identifies populations in need of preventive care interventions (Hartung et al., 2026).

5.1 Data Ingestion and Multimodal Vector Integration

The framework continuously ingests data across five core domains, linking patient-level clinical records with community-level socio-environmental data:

- **Healthcare Accessibility Vector**: Electronic Health Records (EHR), claims data, clinic travel time maps, and scheduling system records.

- **Socioeconomic Conditions Vector**: Geocoded census metrics, housing instability indexes, income levels, and longitudinal cohort social risk indicators.

- **Digital Literacy Vector**: Area-level broadband access metrics, device ownership metrics, patient portal engagement histories, and digital literacy evaluations.

- **Environmental Exposure Vector**: High-resolution atmospheric air quality feeds ($PM_{2.5}, O_3$), land surface temperature maps, proximity to industrial sites, and built-environment walkability scores.

- **Community Health Indicators Vector**: County- and tract-level chronic disease registries, localized mortality rates, and social vulnerability scores (Straw, 2024).

5.2 Privacy-Preserving Collaborative Architecture

Combining confidential socio-demographic, geographic, and clinical information from various institutions demands a high level of data privacy protection. The adopted methodological structure utilizes a decentralized federated learning of a continual weight transfer architecture (Shah et al., 2025). Rather than funneling patient data into centralized databases, models locally train within separate health systems, municipal health departments, or community health centers. Model updates (gradients) instead of raw personal health information are protected with differential privacy protocols and aggregated filing through a secure server. This preserves regulatory compliance with laws like HIPAA and GDPR and permits models, the acquisition of knowledge from heterogeneous cross-institutional datasets, which captures varied geographies and demographics (Sankar et al., 2025).

5.3 Algorithmic Life-Cycle Equity Assurance Framework

Mitigating bias requires integrating fairness controls throughout every stage of the machine learning life cycle, rather than relying solely on post-hoc evaluation (Towett et al., 2023).

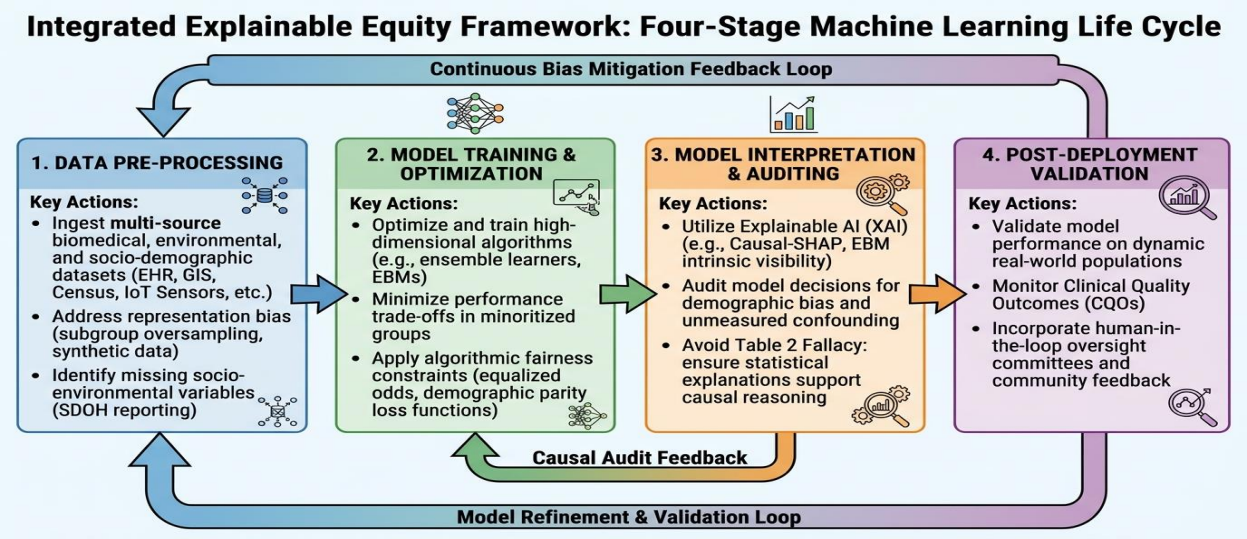


Figure 2: Algorithmic Life-Cycle Equity Assurance Framework

Table 3: Algorithmic Life-Cycle Strategies for Equity Assurance and Bias Mitigation

Life-Cycle Phase	Core Equity Risk	Mitigation Methodology	Technical & Operational Enablers
1. Data Pre-processing	Historical underrepresentation; skewed	Subgroup oversampling; synthetic data generation;	Open non-identifiable

	class distributions; missing socio-environmental variables.	standardized SDOH reporting schemas.	databases; multi-institutional data integration schemas.
2. Model Training & Optimization	Performance trade-offs prioritizing overall accuracy at the expense of minoritized group precision.	Adversarial debiasing; fair-representation learning; inclusion of fairness penalty terms in loss functions.	Equalized odds constraints; demographic parity optimization functions.
3. Model Interpretation & Auditing	Misinterpreting statistical associations as causal mechanisms (Table 2 Fallacy); black-box opacity.	Causal graph integration (Directed Acyclic Graphs [DAGs]); domain-level SHAP aggregation; intrinsically interpretable EBMs.	Causal discovery algorithms; sensitivity analysis across collinear features.
4. Post-Deployment Validation	Performance degradation over time; temporal data shifts; real-world behavioral feedback loops.	Prospective subgroup validation; continuous monitoring of clinical quality outcomes (CQOs); algorithmic auditing.	Real-time EHR monitoring dashboards; human-in-the-loop oversight committees.

6. Ethical Governance, Environmental Sustainability, and Policy Imperatives

Deploying explainable AI frameworks for population-level preventive care requires addressing broader ethical, ecological, and policy challenges. AI tools operate within dynamic socio-technical environments, where operational decisions impact real-world public health resources (Abi & Adetunji, 2023).

6.1 Regulatory Compliance, Data Governance, and User Empowerment

Combining very detailed spatial and socio-demographic indicators with clinical data presents a number of privacy challenges. Even when traditional identifiers are omitted, advanced spatial geocoding can result in patient re-identification. Data governance schemes need to provide explicit user rights management, allowing patients to decide how to use their socio-environmental data in predictive algorithms (Amann et al., 2020). These systems have to be compliant with already-established legal frameworks, that use anonymisation technologies, dynamic consent protocols and role-based access limits to secure confidential patient information (Amann et al., 2020).

6.2 The Eco-Ethical Paradox of Medical Computational AI

The ever-increasing challenge in biomedical AI is that of the environmental footprint associated with extreme high-performance computing infrastructure. As you train deep neural networks and process multi-terabyte datasets, huge amounts of energy are consumed. The direct result is a large increase in global greenhouse gas emissions, raising an eco-ethical paradox: The use of computationally heavy AI models to predict and quantify ambient climate, and environmental health risks can lead to an increase in global computational carbon emissions, which will contribute to climate change affecting the very populations models were built to protect (Lundberg & Lee, 2017).

In order to resolve this dilemma, public health AI systems require energy efficient architectural designs, distilled model structures and green computing protocols that combine predictive performance with environmental sustainability (Joseph, 2025).

6.3 Human-in-the-Loop Governance and Participatory Community Design

AI models should not be the only decision-makers for allocating public health resources. Human-in-the-loop operational frameworks are required for effective bidder. These explainable model outputs are decision support tools to enhance the collective efforts of the entire health sector, including multidisciplinary care teams, public health officials and community health workers (Kelly et al., 2019).

Equitable AI also needs inclusivity of marginalized communities throughout the AI development cycle. Community members, local health workers and advocacy groups input into feature selection, model auditing, and validation to help ensure algorithms reflect an accurate representation of persons in the real world and surface actionable, community-most-needed health interventions (Rajkomar et al. 2018).

Conclusion

This review has shown that bringing AI into healthcare is a huge chance but also a big problem with ethics. Even though smart computer programs can find signs of disease before it shows up and make health plans better for people using them in places where health care is not equal can make old problems worse by making them seem like they are fixed by numbers and data. The new Integrated Explainable Equity Framework tries to fix this by putting five parts together. These parts are Healthcare Accessibility, Socioeconomic Conditions, Digital Literacy, Environmental Exposure and Community Health Indicators. This creates a way to use technology and people together. Using ways to protect privacy making sure the computer programs are fair all the time and checking the models carefully with special tools the framework goes beyond just trying to fix unfairness. It aims to make sure everyone is treated equally. Some important things come out of this. First there is something called the "Table 2 Fallacy" that can stop people from getting the help. People who make rules and doctors need to know that just because a computer says something is important does not mean it is the reason. They need to use ways to make sure. Second being able to use computers and having internet is not something. It is very important for health. Needs to be part of the plans. This way people who are not sick but worried do not take all the help. Third making big computer systems for health care uses a lot of energy. This creates a problem between helping people and protecting the planet. This needs ways to use less energy. Fourth people need to be involved in making sure the computer systems are fair and right. This is not optional. In the end making sure everyone gets the chance with health care needs a big change. Explainable AI should not make choices on its own. It should help people make choices. It should help teams of experts, people who work in health and people who work in the community to find and help people who need more help. More work should focus on testing this framework in places and with different groups of people. It should also create ways to check if the computer systems are fair. It should create rules that make sure all computer systems used for health care are fair all the time. Without doing all of this the chance that AI can help prevent illness will not be available, to those who need it the most.

REFERENCES

Abbas, G., Ying, Z., & Iqbal, M. (2026). From binary scores to risk tiers: An interpretable hybrid stacking model for multi-class loan default prediction. Systems, 14(1), 78.

Abi, R., & Adetunji, O. (2023). AI-enhanced health informatics frameworks for predicting infectious disease outbreak dynamics using climate, mobility, and population immunization data integration. Int. J. Med. Sci, 5(1), 21-31.

Agarwal, R., Bjarnadottir, M., Rhue, L., Dugas, M., Crowley, K., Clark, J., & Gao, G. (2023). Addressing algorithmic bias and the perpetuation of health inequities: An AI bias aware framework. Health Policy and Technology, 12(1), Article 100702.

Alghamdi, S., Mehmood, R., Alqurashi, F. A., & Alzahrani, A. (2025). Paving the roadmap for XAI and IML in healthcare: data-driven discoveries and the FIXAIH framework. IEEE Access, 13, 174393-174427.

Amann, J., Blasimme, A., Vayena, E., Frey, D., & Madai, V. I. (2020). Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. BMC Medical Informatics and Decision Making, 20, Article 310. https://doi.org/10.1186/s12911-020-01332-6

Chin, M. H., Afsar-Manesh, N., Bierman, A. S., Chang, C., Colón-Rodríguez, C. J., Dullabh, P., Duran, D. G., Fair, M., Hernandez-Boussard, T., Hightower, M., Jain, A., Jordan, W. B., Konya, S., Moore, R. H., Moore, T. T., Rodriguez, R., Shaheen, G., Snyder, L. P., Srinivasan, M., ... Ohno-Machado, L. (2023). Guiding principles to address the impact of algorithm bias on racial and ethnic disparities in health and health care. JAMA Network Open, 6(12), e2345050. https://doi.org/10.1001/jamanetworkopen.2023.45050

Chinta, S. V., Wang, Z., Palikhe, A., Zhang, X., Kashif, A., Smith, M. A., ... & Zhang, W. (2025). AI-driven healthcare: A review on ensuring fairness and mitigating bias. PLOS Digital Health, 4(5), e0000864.

Coupland, H., Scheidwasser, N., Katsiferis, A., Davies, M., Flaxman, S., Hulvej Rod, N., ... & Unwin, H. J. T. (2025). Exploring the potential and limitations of deep learning and explainable AI for longitudinal life course analysis. BMC Public Health, 25(1), 1520.

Crawford, A., & Serhal, E. (2020). Digital health equity and COVID-19: The innovation curve cannot reinforce the social gradient of health. Journal of Medical Internet Research, 22(6), e19361. https://doi.org/10.2196/19361

Dekkers, G., & Desmet, R. (2013, December). The social and budgetary impacts of recent social security reform in Belgium. In General Conference of the International Microsimulation Association, Date: 2013/12/11-2013/12/13, Location: Canberra, Australia.

Fridkin, S., & Bendersky, M. (2026). Interpretable machine learning: a comprehensive review of foundations, methods, and the path forward. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 16(1), e70075.

Fridkin, S., & Bendersky, M. (2026). Interpretable machine learning: a comprehensive review of foundations, methods, and the path forward. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 16(1), e70075.

Hartung, T., Kopańska, K., Maertens, A., Whaley, P., & Hoffmann, S. (2026). Causality analysis of toxicological mechanisms in networked systems such as adverse outcome pathway networks. Archives of Toxicology, 1-26.

Joseph, J. (2025). Algorithmic bias in public health AI: A silent threat to equity in low-resource settings. Frontiers in Public Health, 13, Article 1643180. https://doi.org/10.3389/fpubh.2025.1643180

Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. BMC Medicine, 17, Article 195. https://doi.org/10.1186/s12916-019-1426-2

Kumar, A., Dhanka, S., Sharma, A., Sharma, A., Nain, M., Kumar, P., ... & Pant, R. (2026). A comprehensive review of bias in AI, ML, and DL models: methods, impacts, and future directions. Archives of Computational Methods in Engineering, 33(4), 5743-5773.

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable ai: A review of machine learning interpretability methods. Entropy, 23(1), 18.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (Vol. 30).

Nazer, L. H., Zatarah, R., Waldrip, S., Ke, J. X. C., Moukheiber, M., Khanna, A. K., Hicklen, R. S., Moukheiber, L., Moukheiber, D., Ma, H., & Mathur, P. (2023). Bias in artificial intelligence algorithms and recommendations for mitigation. PLOS Digital Health, 2(6), e0000278. https://doi.org/10.1371/journal.pdig.0000278

Norori, N., Hu, Q., Aellen, F. M., Faraci, F. D., & Tzovara, A. (2021). Addressing bias in big data and AI for health care: A call for open science. Patterns, 2(10).

Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M. E., ... & Staab, S. (2020). Bias in data-driven artificial intelligence systems An introductory survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10(3), e1356.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447–453. https://doi.org/10.1126/science.aax2342

Paulus, J. K., & Kent, D. M. (2020). Predictably unequal: Understanding and addressing concerns that algorithmic clinical prediction may increase health disparities. npj Digital Medicine, 3, Article 99. https://doi.org/10.1038/s41746-020-0304-9

Rafe, A., & Das, S. (2025). Causal AI and Computational Epistemology-A Survey. Available at SSRN 6045495.

Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G., & Chin, M. H. (2018). Ensuring fairness in machine learning to advance health equity. Annals of Internal Medicine, 169(12), 866–872. https://doi.org/10.7326/M18-1990

Richardson, S., Lawrence, K., Schoenthaler, A. M., & Mann, D. (2022). A framework for digital health equity. npj Digital Medicine, 5, Article 119. https://doi.org/10.1038/s41746-022-00663-0

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1, 206–215. https://doi.org/10.1038/s42256-019-0048-x

Sankar, B. S., Gilliland, D., Rincon, J., Hermjakob, H., Yan, Y., Adam, I., ... & Ping, P. (2024). Building an ethical and trustworthy biomedical AI ecosystem for the translational and clinical integration of foundation models. Bioengineering, 11(10), 984.

Shah, S. T., Ali, Z., Waqar, M., & Kim, A. (2025, October). Federated learning in public health: A systematic review of decentralized, equitable, and secure disease prevention approaches. In Healthcare (Vol. 13, No. 21, p. 2760). MDPI.

Straw, I. (2024). Quantitative Ethics in Healthcare Artificial Intelligence (Doctoral dissertation, UCL (University College London)).

Suresh, H., & Guttag, J. V. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. In Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization. https://doi.org/10.1145/3465416.3483305

Towett, G., Snead, R. S., Grigoryan, K., & Marczika, J. (2023). Geographical and practical challenges in the implementation of digital health passports for cross-border COVID-19 pandemic management: a narrative review and framework for solutions. Globalization and Health, 19(1), 98.

Vayena, E., Blasimme, A., & Cohen, I. G. (2018). Machine learning in medicine: Addressing ethical challenges. PLOS Medicine, 15(11), e1002689. https://doi.org/10.1371/journal.pmed.1002689

Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., Jung, K., Heller, K., Kale, D., Saeed, M., Ossorio, P. N., Thadaney-Israni, S., & Goldenberg, A. (2019). Do no harm: A roadmap for responsible machine learning for health care. Nature Medicine, 25, 1337–1340. https://doi.org/10.1038/s41591-019-0548-6

Wiessing, L. G. (2017). Epidemiology of HIV and viral hepatitis among people who inject drugs in Europe.

Zhang, J., Mu, L., Mai, G., Grundstein, A., Zhou, Z., & Zhang, D. (2026). SpatialCausal: a spatially-aware causal inference deep learning model for out-of-hospital cardiac arrest survival prediction. International Journal of Geographical Information Science, 1-36.