JOURNAL OF PHARMA INSIGHTS AND RESEARCH ISSN NO. 3048-5428

REVIEW ARTICLE

Calibrated Trust, Rigorous Validation, and the Real Barriers for the Use of Clinical Artificial Intelligence in Antimicrobial Stewardship

Famous Agaga*[1], Adeyinka Moyinoluwa Adejumobi[2], James Momoh[3]

[1] *Department of Diagnostic Services, Geisinger Wyoming Valley Medical Center, Wilkes Barre, Pennsylvania, USA*
[2] *Department of Data, Artificial Intelligence and Modelling, University of Hull, Hull City, England, United Kingdom*
[3] *College of Business and Information Systems, Dakota State University, Madison, South Dakota, USA*

Publication history: Received on 16th April 2026; Revised on 20th May 2026; Accepted on 23rd May 2026

Article DOI: 10.69613/zq94ts47

Abstract: Antimicrobial resistance is a critical threat to global health, driven largely by suboptimal prescribing patterns. Although machine-learning algorithms have high predictive performance for infection risks and resistance profiles in silico, their clinical translation remains severely limited. This practical gap is frequently blamed on a deficit of clinician trust in black-box models, leading to a widespread demand for post-hoc explainable artificial intelligence. However, empirical evidence reveals that post-hoc explanations can mislead, inflating clinician confidence without verifying safety. The core challenge in clinical decision support is not trust itself but calibrated trust: ensuring clinical reliance matches the reliability of a system within a specific clinical environment. Post-hoc explainability methods often fail to reflect actual model mechanisms and can introduce unstable, non-causal associations. Unvalidated predictive systems, even when accompanied by plausible explanations, risk promoting automation bias and severe clinical errors. Clinical safety requires a fundamental shift in priority from post-hoc transparency to rigorous external validation, local calibration, and continuous post-deployment monitoring. Elevating validation and monitoring to primary clinical requirements, while treating explanation as an auxiliary, conditional tool, provides a mathematically and clinically sound model for artificial intelligence-enabled stewardship. Safe clinical adoption depends on establishing clear evidence of real-world utility rather than generating post-hoc rationalizations for unverified predictions.

Keywords: Antimicrobial stewardship; Calibrated trust; Clinical decision support; External validation automation bias.

1. Introduction

Antimicrobial resistance (AMR) has evolved from a prospective global concern into a dominant driver of contemporary clinical mortality. An extensive global evaluation attributed 1.27 million deaths directly to bacterial AMR in 2019, while linking an additional 4.95 million deaths to resistant pathogens [1]. This clinical toll surpasses the mortality rates of HIV/AIDS and malaria combined, with the most severe clinical and economic burdens falling disproportionately on resource-constrained healthcare environments [1]. A substantial and highly actionable portion of this burden is directly attributable to suboptimal prescribing practices, which manifest as unnecessary empiric antibiotic initiation, overly broad-spectrum selection, prolonged durations of therapy, and delayed or absent clinical de-escalation [2].

Traditional antimicrobial stewardship programs (ASPs) were designed to mitigate these precise issues. Utilizing human-centered interventions such as prospective audit and feedback, formulary restriction, and infectious-disease consultations, traditional ASPs have shown success in optimizing therapy and reducing resistance pressure [3, 4]. However, this model suffers from critical scalability limitations. Its efficacy is directly constrained by the availability of specialized infectious-disease clinicians, clinical pharmacists, and microbiologists precisely the workforce that is scarcest in regions experiencing the highest burden of resistance. The volume and velocity of contemporary electronic health record (EHR) data have surpassed the capacity of manual clinical audits. In response to these scalability constraints, artificial intelligence (AI) has been introduced as a potentially transformative solution. Machine-learning models have been designed to estimate clinical infection probabilities, predict specific resistance profiles prior to laboratory culture finalization, detect early physiological deterioration, and recommend optimal therapeutic regimens [5, 6]. These systems frequently display predictive discrimination that matches or exceeds current clinical practice in retrospective in silico testing.

* Corresponding author: Famous Agaga

Despite these impressive retrospective achievements, a profound gap persists between computational accuracy and bedside integration. Advanced algorithms that perform exceptionally well on historic datasets are frequently abandoned, bypassed, or left undeployed. Computational medicine has historically diagnosed this paradox as a deficit of user trust in opaque, "black-box" models. The standard solution has been to mandate post-hoc explainable AI (XAI) as the primary mechanism to cultivate clinician trust, assuming that transparency will automatically lead to clinical adoption and improved patient outcomes.

This review argues that this standard diagnosis is fundamentally incomplete, and the resulting emphasis on post-hoc explainability is misplaced. The central barrier to deployment is not an absolute deficit of trust, but rather a deficit of calibration. Clinicians do not require a generalized willingness to accept algorithmic advice; they require the capability to rely on predictions exactly when they are accurate and to reject them when they are faulty. Explainability is frequently presented as the mechanism to achieve this capability. However, post-hoc explanations can be hazardous because they can artificially elevate clinician confidence without guaranteeing predictive safety. When a clinician relies on an incorrect recommendation due to a highly plausible but inaccurate explanation, patient harm is exacerbated.

This perspective is supported by two key concepts in machine-learning theory. Rudin showed that post-hoc explanations of black-box models, as opposed to inherently interpretable-by-design models, can perpetuate flawed clinical reasoning in high-stakes environments [7]. Ghassemi et al. argued that contemporary explainability methods offer a false hope for clinical decision support, emphasizing that rigorous external and prospective validation represents the only mathematically sound pathway to clinical utility [8]. This systemic vulnerability is illustrated by the widespread implementation and subsequent validation failure of the Epic Sepsis Model. This proprietary tool was deployed across hundreds of hospitals prior to rigorous independent validation. When subjected to rigorous external evaluation, its real-world discriminative capability fell to an area under the receiver operating characteristic curve (AUC) of 0.63, compared to the vendor-reported 0.76 to 0.83 [9]. This failure generated substantial false alarms, contributing to alert fatigue and compromising clinical workflow [9]. No post-hoc feature attribution or visual explanation could have exposed this performance drop. Only systematic, external validation revealed the true performance.

2. Scope and Methodology

This review utilizes a narrative method to synthesize literature across computational medicine, clinical informatics, human factors engineering, and infectious diseases. It deliberately avoids the format of a systematic review to focus on a conceptual integration of highly fragmented domains. The literature analyzed includes English-language studies published between 2015 and 2026, identified via structured searches of PubMed, Scopus, Web of Science, and IEEE Xplore. Key search terms focused on the intersection of healthcare machine learning, explainability metrics, clinical decision support systems, cognitive automation biases, and antimicrobial stewardship programs. This search was supplemented by backward and forward citation tracking of foundational texts in machine-learning validation and human-automation interaction.

The review provides empirical studies that test the behavioral impact of decision support systems over purely in silico performance reports. Special emphasis is placed on clinical environments where algorithmic interventions have been evaluated prospectively. Where clinical evidence is sparse such as in low- and middle-income countries (LMICs) the limitations of the current literature are explicitly detailed rather than minimized.

3. Evolution of Machine Learning in Antimicrobial Stewardship

3.1. Shift from Reactive Machine Learning to Anticipatory Clinical Decision Support

Traditional antimicrobial stewardship is structurally reactive. Clinicians typically initiate empiric therapy based on generalized guidelines and local antibiograms, followed by a retrospective review of the clinical decision forty-eight to seventy-two hours later when microbiology cultures and susceptibility profiles are finalized [3]. This delay often leads to prolonged exposure to broad-spectrum agents or, conversely, delayed escalation of therapy for resistant infections.

3.1.1. Predictive Modeling of Resistance

Machine learning fundamentally shifts this timeline by transforming clinical decision support from a retrospective audit into an anticipatory, real-time intervention. Microbiology databases, and patient-specific longitudinal data, models can predict the likelihood of resistance to specific agents at the exact moment of order entry by training algorithms on historical electronic health records [5, 10].

For example, models leveraging high-dimensional clinical inputs can estimate the probability of a urinary tract pathogen displaying resistance to first-line agents such as trimethoprim-sulfamethoxazole or ciprofloxacin before laboratory confirmation is available

[6]. This temporal acceleration allows clinicians to bypass broad-spectrum empiric selections, narrowing the therapy spectrum from day one without compromising patient safety.

3.1.2. Real-Time Deterioration Warning Systems

Beyond direct resistance prediction, AI applications in stewardship include real-time physiological tracking. Early warning systems continuously ingest high-frequency clinical data such as body temperature, heart rate, white blood cell counts, and respiratory parameters to calculate the probability of clinical sepsis or severe infection [11].

When these systems function correctly, they alert clinicians to physiological deterioration hours before traditional threshold-based scoring systems trigger. This early warning enables timely specimen collection, targeted diagnostic testing, and the rapid administration of appropriate antimicrobials. This proactive approach helps mitigate systemic inflammatory cascades and improves clinical outcomes.

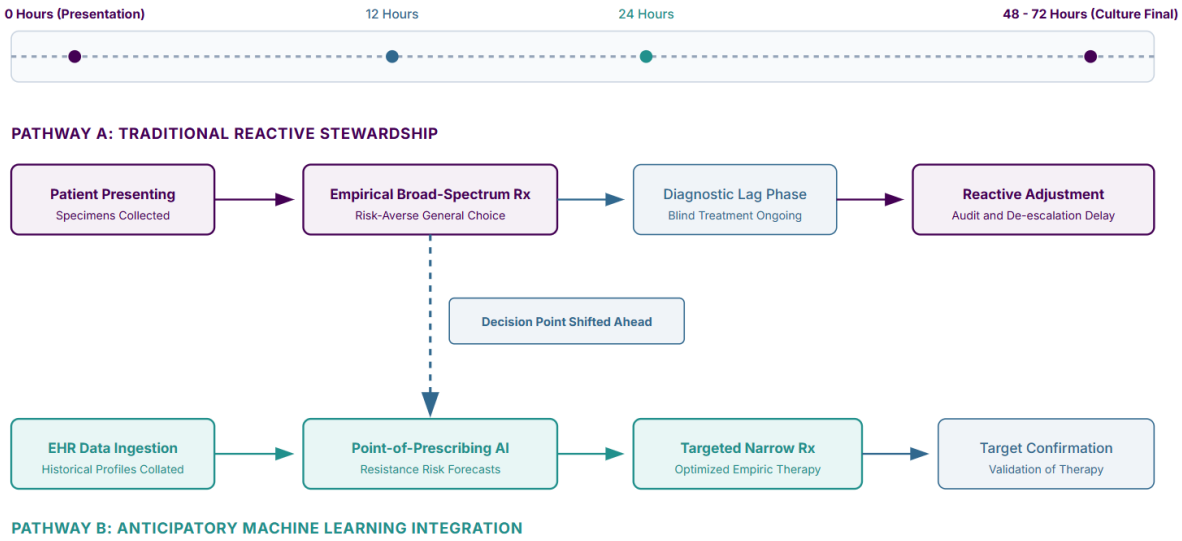

Figure 1. Traditional reactive pathways vs. anticipatory machine learning integration

3.2. Why Retrospective Performance Fails at the Bedside?

The mathematical performance of a machine-learning model in silico rarely translates directly to clinical efficacy at the bedside. This discrepancy arises from a fundamental validation gap: the reliance on retrospective performance metrics calculated on highly curated, static datasets that do not reflect the dynamic realities of clinical deployment.

3.2.1. Institutional Distribution Shifts and Local Epidemiology

A primary driver of this validation gap is the phenomenon of distribution shift. Machine-learning models are highly sensitive to the statistical properties of their training data. In clinical medicine, these properties are determined by local epidemiology, institutional testing protocols, patient demographics, and electronic health record configurations.

When a model developed at a tertiary academic center is deployed at a community hospital, it encounters a different patient population and varying baseline prevalence rates of resistant pathogens. This mismatch can lead to a substantial drop in predictive accuracy. For instance, a model designed to predict carbapenem resistance in an environment with high endemicity will yield unacceptably high false-positive rates when deployed in an environment with low baseline resistance.

Clinical practices themselves shift over time. A change in standard diagnostic guidelines or the introduction of a new rapid diagnostic platform can alter the data-generation process. This can invalidate a model's historical assumptions and cause its performance to decay.

3.2.2. The Epic Sepsis Model Failure as a Systematic Paradigm

The failure of the Epic Sepsis Model serves as a clear illustration of the validation gap in clinical decision support. This proprietary, deep-learning-based tool was integrated directly into a widely used EHR platform and deployed across hundreds of healthcare facilities. The vendor reported high discriminative capabilities, citing an AUC between 0.76 and 0.83.

However, when an independent team of researchers performed an external validation of the model using data from a single large health system involving approximately 38,000 hospitalizations, the real-world performance fell to an AUC of 0.63 [9].

This performance drop had severe clinical consequences. The model failed to identify 67% of patients who went on to develop sepsis, while simultaneously generating an immense volume of false alerts [9]. This high false-positive rate contributed significantly to clinical alert fatigue.

The critical insight from this failure is that the model's performance drop was not due to an algorithmic bug, but rather to a fundamental difference in how sepsis was defined and documented in the real-world validation cohort versus the vendor's proprietary training cohort [11].

The clinical documentation patterns, order-set timings, and patient demographics in the active clinical environment differed substantially from the static datasets used to train the model. This failure shows that high retrospective performance on historical data is insufficient to guarantee clinical safety and effectiveness at the bedside.

4. The Trust Problem

4.1. The Cognitive Mechanics of Calibration and Reliance

In contemporary clinical informatics, trust is frequently treated as a binary or linear variable that must be maximized to ensure clinical adoption. However, human-factors engineering shows that the primary objective of clinical decision support integration is not the maximization of trust, but its precise calibration. Trust calibration is defined as the alignment between a user's trust in an automated system and the system's objective capabilities [12]. When clinical trust is calibrated, a physician's reliance on algorithmic output matches the actual accuracy of that output across different clinical scenarios. The dynamic between human clinicians and clinical algorithms can be categorized into distinct behavioral states based on this alignment.

4.1.1. Automation Bias and Over-reliance

The first primary failure of calibration is automation over-reliance, often referred to as automation bias. This occurs when a clinician's trust exceeds the actual reliability of the automated system, leading to the uncritical acceptance of erroneous recommendations [13]. In antimicrobial prescribing, where clinical decisions are frequently made under high cognitive load, time constraints, and diagnostic uncertainty, automation bias represents a critical patient safety hazard.

For example, if an EHR alert recommends a narrow-spectrum antibiotic for a patient displaying clinical signs of sepsis, and the clinician accepts this recommendation without verifying that the pathogen is susceptible, the resulting delay in effective therapy can lead to rapid physiological deterioration. Automation bias is often exacerbated by highly integrated clinical systems that present recommendations as institutional defaults, making them cognitively easier to accept than to challenge.

4.1.2. Automation Disuse and Under-reliance

The second calibration failure is automation disuse, or under-reliance. This occurs when a clinician's trust falls below the system's objective reliability, leading to the rejection of accurate recommendations [13]. In antimicrobial stewardship, under-reliance typically stems from high false-alarm rates or perceived clinical irrelevance, which drive systemic alert fatigue. When a model continuously generates alerts for low-risk patients or repeatedly recommends redundant therapeutic de-escalation, clinicians develop a generalized skepticism toward the system. Consequently, they are likely to ignore or bypass valid, life-saving alerts, such as those flagging a critical drug-pathogen mismatch or predicting a highly resistant infection. This under-reliance neutralizes the potential benefits of the technology, rendering the algorithmic investment ineffective.

Table 1. Human-AI Calibration States in Antimicrobial Prescribing

Behavioral Alignment State	Trust Level vs. Objective Capability	Prescribing Manifestation	Potential Clinical Outcome	Recommended Mitigation Measures
Calibrated Trust	Dynamic mathematical alignment	Prescriber accepts valid guidance and overrides erroneous predictions based on localized clinical indicators	Optimal clinical efficacy, minimized selection pressure, and sustained system utility	Continuous performance feedback and exposure to historical model failure modes
Automation Bias (Over-reliance)	Pathological inflation of trust (Trust > Capability)	Prescriber reflexively adopts inappropriate narrow-spectrum or broad-spectrum alerts without diagnostic verification	Delays in effective therapy for resistant pathogens, patient deterioration, and systemic safety hazards	Cognitive forcing functions, active safety checklists, and visualization of prediction uncertainty
Automation Disuse (Under-reliance)	Pathological suppression of trust (Trust < Capability)	Prescriber bypasses valid resistance warnings or de-escalation alerts due to generalized alert fatigue	Prolonged broad-spectrum exposure, escalated institutional resistance rates, and wasted technological investment	Alert rate optimization, reduction of low-utility triggers, and transparent validation metrics
Miscalibrated Rejection	Inverse correlation of trust and capability	Prescriber rejects highly accurate predictions while accepting erroneous recommendations due to confirmation bias	High-variability clinical outcomes and exposure of vulnerable populations to empirical prescribing errors	Human-factors interface design focused on collaborative clinical reasoning rather than binary directives

4.2. Dynamic Trust and Heterogeneity Across Clinical Environments

Trust calibration is not a static state achieved at deployment, but rather a dynamic, evolving cognitive process. A clinician's initial trust in an AI tool is shaped by institutional framing, perceived risk, and prior experience with technology.

As clinicians interact with the system over time, their trust is continuously updated based on the perceived accuracy of the recommendations they encounter [14]. This temporal evolution of trust is highly sensitive to early errors. A single highly visible, incorrect recommendation early in the deployment phase can cause a catastrophic collapse in trust that is difficult to rebuild, even if the model maintains high overall statistical accuracy. Trust calibration is highly heterogeneous across clinical specialties and levels of seniority. Experienced infectious disease specialists often display lower baseline trust in automated recommendations, relying instead on their extensive clinical heuristic models.

In contrast, junior residents or non-specialist clinicians working in high-pressure environments, such as the emergency department, may display higher baseline susceptibility to automation bias. These clinicians may lack the specialized knowledge required to recognize when an algorithmic recommendation is clinically inappropriate.

Clinical decision support systems must be evaluated and implemented with a precise understanding of the specific clinical cohort interacting with the technology, ensuring that trust calibration interventions are tailored to the users' baseline expertise and cognitive environment.

5. Limitations of Post-hoc Explainability Methods

5.1. The Epistemic Gap Between Interpretability and Post-hoc Explainability

A common error in clinical machine learning is the conflation of interpretable-by-design models with post-hoc explainability methods. This distinction is critical because they carry fundamentally different epistemic guarantees.

An inherently interpretable model is one whose internal mechanism is directly transparent and inspectable. Classic examples include sparse linear regression, decision trees, and rule lists. In these models, the relationship between input variables and the final prediction is mathematically explicit, leaving no gap between the model's actual calculations and the human's comprehension [7].

In contrast, post-hoc explainability methods are applied to complex, opaque "black-box" models, such as deep neural networks or gradient-boosted trees. These methods do not reveal the actual internal reasoning of the model. Instead, they generate a secondary, simplified model often called a surrogate designed to approximate the black box's behavior around a specific prediction [8].

Table 2. Mathematical Foundations and Epistemic Liabilities of Explainability Methods

Post-hoc Method	Operational and Mathematical Basis	Core Epistemic Assumptions	Critical Vulnerability in Clinical EHR Datasets	Ideal Stewardship Application
SHAP (Shapley Additive exPlanations)	Game-theoretic attribution calculating marginal contributions (ϕ_i) across all possible feature subsets	Input features act as independent players in a cooperative game	Violates physiological collinearity; distributes attribution arbitrarily among coupled laboratory markers	Retrospective clinical audit and model debugging by development teams
LIME (Local Interpretable Model-agnostic Explanations)	Local linear surrogate model approximation fit via perturbed inputs in a localized distance neighborhood (π_x)	The complex black-box decision boundary is locally linear and stable	High instability; random perturbations yield divergent explanations for the same clinical state	Post-hoc multi-disciplinary review of borderline cases
Attention Maps	Softmax-normalized weights within transformer architectures indicating computational focus	Spatial or temporal attention weights correspond directly to clinical feature importance	Attention weights can be altered or optimized without changing the underlying clinical prediction	Supplementary visualization tool for natural language processing of clinical notes
Counterfactual Explanations	Optimization of distance metrics to find the closest input vector change that alters the classification output	Clinical variables can be modified independently and continuously	Generates clinically impossible states such as reversing stage-four renal disease or reducing patient age	Deliberative educational programs and clinical simulation training
Rule-Based Surrogates	High-level decision trees or rule lists approximating global black-box decision boundaries	Complex multidimensional decision boundaries can be compressed into discrete logic lists	Trade-off between accuracy and interpretability; complex rules become human-unreadable	Static clinical protocols and high-level guideline alignment dashboards

This process introduces a fundamental epistemic gap: the post-hoc explanation is a simplified approximation, meaning it is inherently incomplete and can be misleading. The explanation may present a logical, clinically intuitive justification for a prediction, while the underlying black-box model is actually relying on non-causal correlations or clinical noise within the dataset.

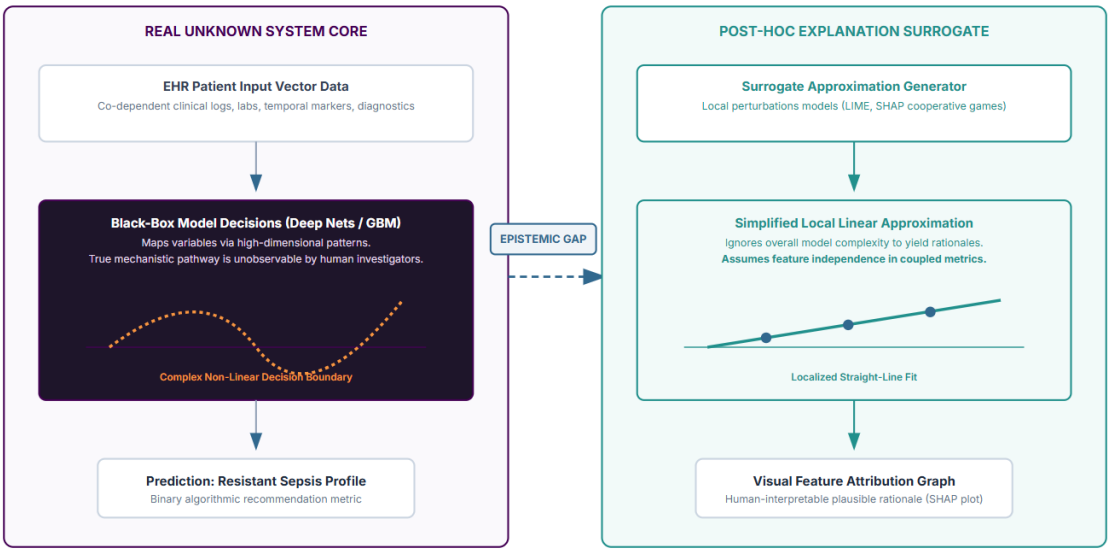

Figure 2. The Epistemic Disconnect of Post-hoc Explanation Models

5.2. Local and Global Attribution Models: SHAP and LIME

5.2.1. SHAP: Assumptions of Feature Independence and Game-Theoretic Attributions

SHAP (Shapley Additive exPlanations) is a widely used post-hoc explainability model in clinical machine learning. It is based on cooperative game theory, where the "players" are the input features and the "payout" is the model's prediction. SHAP calculates the marginal contribution of each feature across all possible feature combinations, providing a mathematically principled attribution score (ϕ_i) for each input variable [5].

$$\phi_i(x) = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} [f_x(S \cup \{i\}) - f_x(S)]$$

While SHAP provides strong local properties, such as local accuracy and consistency, its mathematical formulation relies on assumptions that are routinely violated by clinical EHR data. Specifically, SHAP assumes feature independence when calculating marginal contributions.

In clinical medicine, variables are highly correlated; for example, physiological markers like serum creatinine, blood urea nitrogen, and urine output are physiologically coupled. When these features are highly collinear, SHAP can distribute attribution scores arbitrarily across the correlated variables, generating explanations that are clinically misleading or highly sensitive to minor data perturbations.

5.2.2. LIME: Instability, Surrogate Approximations, and Local Perturbation Weaknesses

LIME (Local Interpretable Model-agnostic Explanations) takes a different approach by fitting a simple, interpretable model typically a sparse linear regressor locally around the prediction of interest. It does this by drawing random samples in the neighborhood of the input instance, querying the black-box model for these samples, and weighting them based on their distance from the original instance.

$$L(f,g,\pi_x) = \sum_{z,z' \in Z} \pi_x(z)\left(f(z)-g(z')\right)^2 + \Omega(g)$$

LIME suffers from several structural limitations that complicate its use in high-stakes clinical decision support. First, LIME's explanations are notoriously unstable. Because the local surrogate is constructed using random perturbations of the input data, different runs of LIME on the exact same prediction can yield highly divergent feature attributions.

Second, the definition of the "local neighborhood" (π_x) is arbitrary and must be manually configured. In high-dimensional clinical spaces, minor changes in the neighborhood radius can result in drastically different surrogate models. This can lead to situations where a clinician is presented with different explanations for the same clinical recommendation based on arbitrary parameter choices within the explainability software.

5.2.3. Attention Mechanisms

With the rise of deep learning and Transformer-based architectures in clinical natural language processing and electronic health record modeling, attention weights are frequently highlighted as intuitive explanations of model behavior. The assumption is that if a model assigns a high attention weight to a specific clinical parameter, such as an elevated white blood cell count or a previous microbiology result, then that parameter must be the primary driver of the model's prediction. This assumption has been challenged by methodological evaluations. Research has showed that attention weights do not reliably correspond to feature importance [15]. Specifically, alternative attention distributions can be constructed that yield identical model outputs, indicating that the relationship between attention weights and prediction is not unique [16]. Attention weights represent a mathematical component of the model's internal optimization architecture rather than a causal explanation of its decision-making logic. Treating attention maps as direct trust signals can lead clinicians to infer clinical reasoning that the model does not possess.

5.3. Counterfactual and Rule-Based Post-hoc Explanations

Counterfactual explanations attempt to provide intuitive insights by identifying the minimum changes in input variables required to alter the model's prediction. For example, a counterfactual explanation for an elevated sepsis alert might state: "Had the patient's lactate level been 1.8 mmol/L instead of 2.4 mmol/L, the alert would not have triggered."

While counterfactuals align well with clinical diagnostic reasoning, they suffer from two major limitations in clinical deployment. First, counterfactuals are non-unique. For any given clinical prediction, there are multiple, often conflicting, sets of input modifications that can flip the model's output. A system designer can select which counterfactual to present to the clinician, introducing a risk of cherry-picking explanations to make the model appear more clinically intuitive than it is.

Second, many mathematically valid counterfactuals are clinically impossible. A model might generate a counterfactual suggesting that a patient's age be reduced or that their baseline chronic kidney disease stage be reversed to change a resistance prediction.

Rule-based surrogate models, which approximate a black-box model's global decision boundary using a set of "if-then" statements, face a similar trade-off. They are either highly simplified, which compromises their accuracy in representing the true black-box behavior, or they are highly complex, which defeats their purpose as interpretable explanations for a clinical audience under time constraints.

6. The Empirical Evidence of Explainability on Decision Calibration

6.1. Improved Engagement and Diagnostic Support

The clinical literature contains several studies showing the positive impact of explainability methods on clinician engagement and initial technology acceptance. In areas such as diagnostic radiology, pathology, and dermatology, presenting visual explanations alongside algorithmic predictions has been shown to improve clinician confidence and reduce diagnostic error rates [17].

When visual explanations such as saliency maps highlighting suspicious regions in a chest radiograph are coupled with accurate predictive models, they can help guide a clinician's attention to subtle pathological features that might otherwise be overlooked.

In antimicrobial stewardship, a similar benefit has been reported when models predicting specific resistance profiles are accompanied by clear feature-attribution plots. These plots highlight key patient risk factors, such as previous antibiotic exposure or recent hospitalization duration.

This transparency can lower the cognitive barrier to adopting automated recommendations, prompting clinicians to consider narrow-spectrum empiric regimens they might have otherwise avoided due to generalized risk aversion. In these scenarios, explanations can act as an educational tool, encouraging clinicians to engage with the system rather than dismissing it out of hand.

6.2. Deceptive Explanations and Unwarranted Confidence

While explanations can increase clinician engagement and self-reported confidence, empirical evaluations suggest that this increased confidence does not always correlate with improved decision quality. Evidence suggests that post-hoc explanations can foster unwarranted confidence, leading clinicians to accept incorrect algorithmic advice.

6.2.1. The Propagation of Algorithmic Errors via Rationalization

A critical study by Gaube et al. evaluated how radiologists interacted with clinical decision aids of varying quality. The researchers found that clinicians were susceptible to incorrect advice, and the presence of post-hoc explanations did not protect them from accepting these errors [18]. More concerningly, research has indicated that presenting post-hoc explanations alongside machine-learning predictions can increase clinician acceptance of both correct and incorrect recommendations [19].

This phenomenon occurs because post-hoc explanations are designed to look plausible. When a clinician is presented with an incorrect prediction accompanied by a highly professional, clinically logical explanation, they are likely to rationalize the error, attributing it to a complex clinical interaction rather than a system failure. In a high-pressure clinical environment, clinicians rarely have the time or the specialized technical training to audit the mathematical validity of a post-hoc explanation. As a result, explainability can function as a persuasive marketing tool rather than an objective safety mechanism. This can degrade calibration and increase the risk of automation-induced clinical errors.

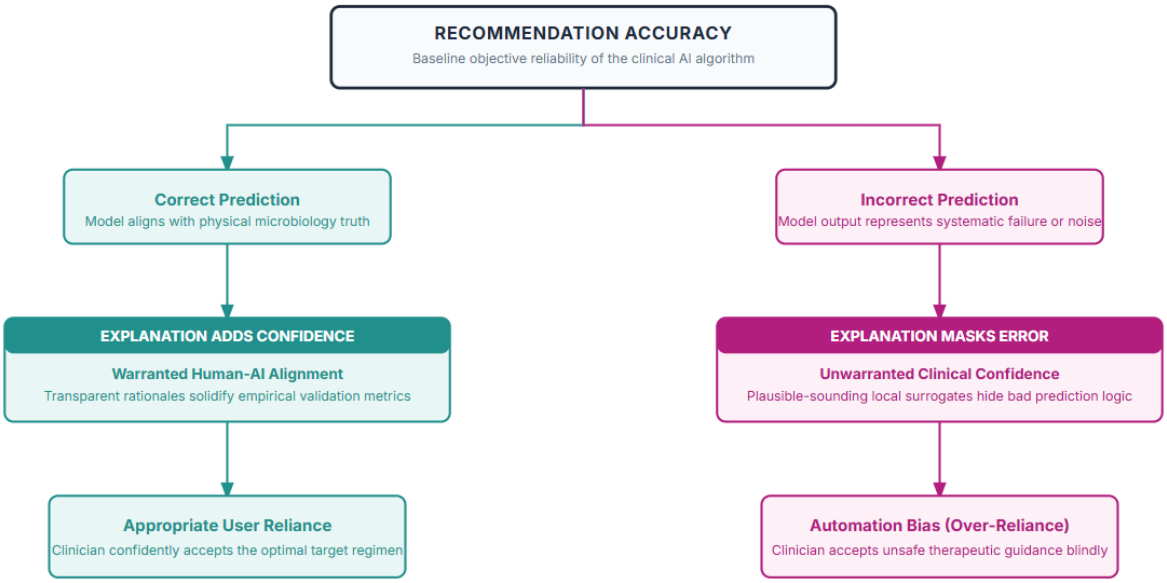

Figure 3. The Bifurcated Influence of Explanations on Clinician Reliance

7. Applications of Stewardship

7.1. Antibiotic Selection Support Systems

Antibiotic selection support systems assist clinicians in choosing optimal antimicrobial regimens by integrating patient-specific clinical characteristics with institutional susceptibility patterns [5, 6]. These systems represent a direct intervention at the point of prescribing, aiming to reduce the use of unnecessarily broad-spectrum agents.

The primary constraint on antibiotic selection models is the highly dynamic nature of local resistance patterns. A model developed using historical susceptibility data from one institution cannot easily generalize to another facility, even within the same geographic region, due to differences in patient demographics, prescribing cultures, and baseline resistance rates.

Post-hoc explainability methods do not address this issue. A SHAP plot displaying the patient features that drove a narrow-spectrum recommendation will not flag when the underlying local prevalence of a resistant pathogen has shifted, or if the model's training data failed to reflect a recent outbreak of resistant organisms.

Continuous local revalidation and performance monitoring are required to ensure the model's recommendations remain clinically safe. Treating post-hoc explanations as a surrogate for this systematic validation risks exposing patients to ineffective empiric therapy.

7.2. Real-Time Sepsis Alerting Networks

Sepsis alerting networks utilize continuous electronic health data to identify patients displaying early signs of systemic infection, prompting clinicians to initiate rapid diagnostic and therapeutic bundles [20]. These systems are highly sensitive to the temporal dynamics of clinical documentation.

The primary clinical challenge with sepsis warning networks is the significant cognitive load generated by false-positive alerts. Sepsis is a complex, heterogeneous clinical syndrome, and many non-infectious conditions can mimic its early physiological signs. When models are optimized for high sensitivity to avoid missing true cases, they inevitably generate a high volume of false alarms [9].

This creates severe alert fatigue, leading clinicians to ignore or bypass alerts. Adding post-hoc explanations to these alerts does not solve this problem and can sometimes exacerbate it.

If a clinician must review a complex feature-attribution diagram for every alert to determine its validity, the resulting cognitive friction can slow down clinical workflows. In high-stakes, time-sensitive environments like the emergency department, explanations must be designed to minimize cognitive load rather than simply justifying the algorithm's output.

7.3. Pathogen Resistance Forecasting

Pathogen resistance forecasters utilize patient history, previous hospitalizations, and prior microbiology results to estimate the probability that a current infection is caused by a resistant organism [5, 10]. This information is valuable when selecting empiric therapy for suspected multidrug-resistant infections.

Pathogen resistance patterns are subject to substantial temporal drift. Over time, the introduction of new antibiotics, changes in clinical guidelines, and natural evolutionary selection pressures alter the baseline distribution of resistant strains. Consequently, a resistance forecasting model's performance will degrade if it is not continuously updated and recalibrated.

Post-hoc explainability methods cannot compensate for this temporal degradation. If a model's predictive accuracy has decayed due to temporal drift, its explanations will continue to reflect historical associations, presenting them as logical justifications for what has become an inaccurate prediction. Clinical safety therefore depends on establishing institutional protocols for continuous model recalibration and prospective performance auditing, ensuring that the tool's output remains aligned with active clinical realities.

7.4. Surveillance Dashboards and Program-Level Analytics

Surveillance dashboards aggregate patient-level data to provide antimicrobial stewardship teams with real-time insights into institutional prescribing patterns, resistance trends, and clinical outcomes [21]. These platforms support high-level decision-making and strategic planning. The clinical utility of explainability differs significantly between retrospective surveillance and real-time bedside decision support. During program-level reviews, clinicians have the time to critically analyze the variables driving a specific epidemiological signal.

Table 3. Application-Specific Validation Metrics and Monitoring Protocols

Stewardship Application Class	Primary Decision Horizon	Critical Distribution Shift Hazards	Mandatory Validation and Monitoring Protocol	Role of Post-hoc Explanations
Antibiotic Selection Support	Real-time, point-of-prescribing empiric regimen selection	Fluctuations in local pathogen prevalence and changes in institutional formulary access	Institutional validation on a rolling 90-day cohort with prospective audits of clinical outcomes	Auxiliary tool to present patient-specific risk factors to clinicians during active decision-making
Real-Time Sepsis Alerting	Continuous, physiological monitoring in critical care units	Modifications in nursing documentation frequency and electronic health record flow-sheet configurations	Prospective evaluation of alert-to-intervention latency and tracking of cumulative alert burden	Cognitive optimization aids designed to display physiological trajectories rather than model justification
Pathogen Resistance Forecasters	Diagnostic phase prior to laboratory culture finalization	Temporal evolutionary drift in resistance mechanisms and introduction of novel antimicrobial agents	Monthly calibration curve tracking and external validation on demographically distinct populations	Informative tools during multidisciplinary infectious disease consultations
Surveillance Dashboards	Retrospective, program-level strategic planning	Administrative coding changes and variations in diagnostic microbiology testing volumes	Annual retrospective comparison with manual infectious disease clinical chart audits	High-utility diagnostic signals used to identify institutional prescribing anomalies and systemic trends

In this deliberative environment, post-hoc explanations can be highly valuable, helping stewardship teams identify potential data anomalies, systemic shifts in prescribing habits, or unexpected pockets of resistance. Because these decisions are made outside of active patient care workflows, the risk of automation bias is minimized. This highlights a key concept: the value of explainability is highly dependent on the decision-making environment, with its utility rising as decisions become more deliberative and falling as they become more reflexive and time-sensitive.

8. Accountability, Bias, and Systemic Governance

8.1. The Legal and Ethical Landscape of Shared Responsibility

The clinical integration of machine learning complicates traditional models of medical liability. Historically, malpractice litigation has focused on the individual clinician, who is presumed to exercise independent clinical judgment.

However, as clinical decision support systems become more complex, the line between clinician autonomy and algorithmic direction begins to blur [22]. When a clinician follows an incorrect algorithmic recommendation that leads to patient harm, the responsibility is distributed across several stakeholders, including the prescribing clinician, the healthcare institution that deployed the tool, the development team, and the software vendor.

Post-hoc explainability methods do not resolve these liability challenges and may complicate them. An explanation can show which variables a model weighted most heavily, but it cannot determine whether relying on that prediction met the accepted standard of care. In some scenarios, a highly plausible explanation might make a clinician's reliance on a faulty prediction appear more reasonable, potentially shifting some liability toward the implementing institution or developer.

If an explanation reveals that the model relied on clinically questionable associations, a clinician who followed the recommendation anyway may face increased personal liability for failing to exercise appropriate oversight. Because legal guidelines remain highly clinician-centric, healthcare organizations must establish clear guidelines defining the appropriate use and limitations of clinical AI tools.

8.2. Algorithmic Bias and Historical Inequities in Clinical Datasets

Machine-learning models are inherently reflective of the data used to train them. If the historical training data contains systemic inequities in clinical access, diagnostic testing, or treatment delivery, the model will learn and potentially perpetuate these biased patterns [23]. In antimicrobial stewardship, algorithmic bias can manifest in several ways. For example, if a resistance-prediction model is trained primarily on data from well-resourced tertiary centers, it may underperform when deployed in safety-net hospitals serving marginalized populations with different baseline access to healthcare and varying exposure to infectious pathogens.

Similarly, historical disparities in diagnostic testing such as lower rates of blood culture collection in certain demographic groups can lead models to underestimate infection risk in those populations. Post-hoc explainability methods can sometimes highlight these biases by revealing that a model is heavily weighting demographic variables like race or insurance status. However, explainability alone cannot correct these underlying imbalances. Addressing algorithmic bias requires careful problem formulation, representative dataset curation, and explicit subgroup validation to ensure the model performs equitably across all patient populations.

Table 4. Algorithmic Governance and Bias Auditing

Audit Vector	Manifestation in Infectious Disease Data	Potential Safety and Equity Impact	Operational Governance Mitigation Strategy
Demographic Disparities	Under-representation of marginalized patient populations in baseline model training cohorts	Performance degradation when deployed in safety-net hospitals or resource-limited centers	Mandatory stratified validation across race, ethnicity, socioeconomic status, and geographic zip codes
Diagnostic Access Biases	Disparities in blood and urine culture collection frequencies based on patient insurance status	Model misinterprets lack of diagnostic testing as absence of active clinical infection	Inclusion of healthcare access metrics as potential confounding variables in model training
Temporal Drift Decay	Declining predictive precision due to changes in baseline bacterial resistance and evolutionary pressure	Increased rates of inappropriate empirical therapy recommendations over time	Implementation of automated alerts when model calibration curves drift beyond specified safety bounds
EHR System Mismatch	Variations in data logging practices, nomenclature, and system integration across clinical sites	Total model failure or extreme drop in discriminative performance (AUC) during external transfer	Mandating standardized data mappings prior to clinical model deployment

8.3. Regulatory Standards for Software-as-a-Medical-Device

Regulatory bodies, such as the United States Food and Drug Administration (FDA), are treating clinical machine-learning tools as software as a medical device (SaMD) [24]. These models require developers to show the safety and effectiveness of their algorithms before clinical deployment. While regulatory guidelines often highlight transparency as a desirable characteristic, they are shifting their primary focus toward the concept of algorithmic stewardship [25]. This approach treats model deployment as an ongoing cycle rather than a one-time software installation.

Algorithmic stewardship requires healthcare institutions to establish dedicated oversight committees responsible for monitoring model calibration, tracking performance degradation, and managing local updates. A highly explainable model that lacks this systematic oversight remains a clinical hazard. Consequently, robust governance structures must prioritize continuous clinical validation over post-hoc explanation, ensuring that the systems we deploy are supported by ongoing evidence of safety and utility.

9. A Corrected Model for Explainability, Validation, and Calibrated Trust

The prevailing mental model in clinical AI implementation assumes a linear pipeline where explainability generates clinician trust, which in turn drives technology adoption, ultimately leading to improved clinical outcomes. This linear model is incomplete because it overlooks the critical failure modes of automation bias and clinical miscalibration.

We propose an updated conceptual model that treats rigorous validation as the primary input to calibrated trust, while treating post-hoc explanation as an auxiliary, conditional modifier. In this revised model, an AI-generated stewardship recommendation feeds into two pathways. The primary pathway is validation and monitoring, which includes external validation on local datasets, prospective clinical auditing, and continuous post-deployment performance tracking. This pathway establishes the empirical reliability of the system within its specific clinical environment.

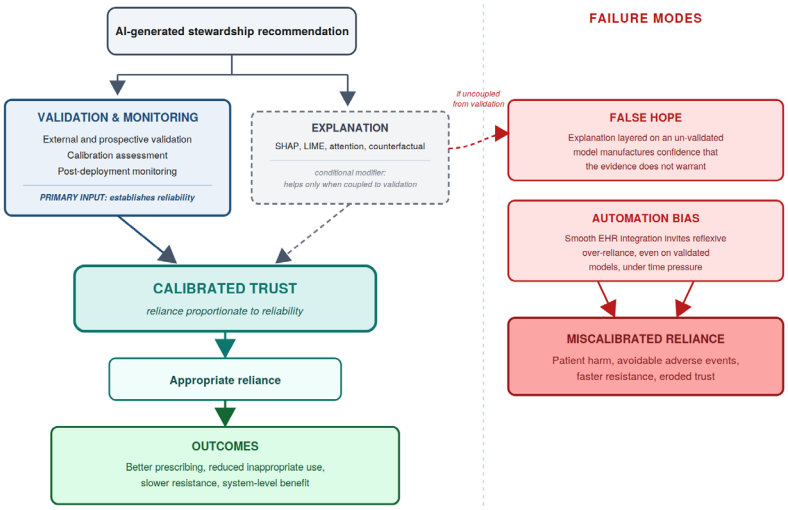

Figure 4. The explainability, validation, and calibrated-trust model.

The secondary pathway is explanation, which provides post-hoc interpretations such as SHAP or LIME attributions. Crucially, the explanation pathway is modeled as a conditional modifier: it helps improve trust calibration only when the underlying system has been validated. If the explanation pathway is uncoupled from the validation pathway, it leaks directly into a failure mode designated as the "False Hope" path, where plausible-looking explanations generate unwarranted clinical confidence in an unvalidated model.

The second major failure mode is "Automation Bias," where even a validated model, when integrated into clinical workflows, can invite reflexive over-reliance under high time pressure. This model shifts the implementation focus from simply adding post-hoc explanations to establishing the empirical and organizational infrastructure needed to support appropriate clinical reliance by modeling these pathways.

10. Conclusion

The usage of artificial intelligence in antimicrobial stewardship is a promising opportunity to optimize empiric therapy and curb the rise of resistant infections. However, the current emphasis on post-hoc explainability as a primary driver of adoption is conceptually and clinically flawed. The primary barrier to clinical AI integration is not a general deficit of trust, but rather a lack of trust calibration. Clinicians must be equipped to rely on these tools exactly as much as the empirical evidence warrants, which requires a shift in priority from post-hoc transparency to rigorous external validation, local calibration, and continuous post-deployment monitoring. Post-hoc explanation methods can be valuable tools when coupled with validated systems, but they cannot substitute for empirical evidence of safety and reliability. Offered in place of validation, explanations can generate a false sense of security, mask algorithmic errors and promoting automation-induced clinical mistakes. The failure of the Epic Sepsis Model serves as a reminder that highly integrated, plausible-looking tools can perform poorly in the real world, and that only systematic validation can catch these discrepancies. As the global burden of antimicrobial resistance continues to grow, the clinical value of machine learning in stewardship will be determined by our commitment to building robust, validation-first implementation pathways. This requires a transition from technology-centered systems to human-centered models that prioritize clinical evidence, ongoing monitoring, and calibrated trust. We can build collaborative clinical tools that support clinicians, protect patients, and preserve the long-term efficacy of our antimicrobial arsenal by shifting our implementation measures on empirical safety rather than post-hoc justification.

References

[1] Murray CJL, Ikuta KS, Sharara F, Swetschinski L, Aguilar GR, Gray A, et al. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. Lancet. 2022;399(10325):629-655.

[2] Fleming-Dutra KE, Hersh AL, Shapiro DJ, Bartoces M, Enns EA, File TM Jr, et al. Prevalence of Inappropriate Antibiotic Prescriptions Among US Ambulatory Care Visits, 2010-2011. JAMA. 2016;315(17):1864-1873.

[3] Davey P, Marwick CA, Scott CL, Charani E, McNeil K, Brown E, et al. Interventions to improve antibiotic prescribing practices for hospital inpatients. Cochrane Database Syst Rev. 2017;2(2):CD003543.

[4] Baur D, Gladstone BP, Burkert F, Carrara E, Foschi F, Döbele S, et al. Effect of antibiotic stewardship on the incidence of infection and colonisation with antibiotic-resistant bacteria and Clostridium difficile infection: a systematic review and meta-analysis. Lancet Infect Dis. 2017;17(9):990-1001.

[5] Yelin I, Snitser O, Novich G, Katz R, Tal O, Parizade M, et al. Personal clinical history predicts antibiotic resistance of urinary tract infections. Nat Med. 2019;25(7):1143-1152.

[6] Kanjilal S, Oberst M, Boominathan S, Zhou H, Hooper DC, Sontag D. A decision algorithm to promote outpatient antimicrobial stewardship for uncomplicated urinary tract infection. Sci Transl Med. 2020;12(568):eaay5067.

[7] Rudin C. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. Nat Mach Intell. 2019;1(5):206-215.

[8] Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health. 2021;3(11):e745-e750.

[9] Wong A, Otles E, Donnelly JP, Krumm A, McCullough J, DeTroyer-Cooley O, et al. External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients. JAMA Intern Med. 2021;181(8):1065-1070.

[10] Lewin-Epstein O, Baruch S, Hadany L, Stein GY, Obolski U. Predicting Antibiotic Resistance in Hospitalized Patients by Applying Machine Learning to Electronic Medical Records. Clin Infect Dis. 2021;72(11):e848-e855.

[11] Habib AR, Lin AL, Grant RW. The Epic Sepsis Model Falls Short-The Importance of External Validation. JAMA Intern Med. 2021;181(8):1040-1041.

[12] Lee JD, See KA. Trust in automation: designing for appropriate reliance. Hum Factors. 2004;46(1):50-80.

[13] Parasuraman R, Riley V. Humans and Automation: Use, Misuse, Disuse, Abuse. Hum Factors. 1997;39(2):230-253.

[14] Schulz PJ, Kee KMM, Lwin MO, Goh WW, Chia KY, Cheung MFK, et al. Clinical experience and perception of risk affect the acceptance and trust of using AI in medicine. Front Digit Health. 2025;7.

[15] Jain S, Wallace BC. Attention is not Explanation. arXiv preprint arXiv:1902.10186. 2019.

[16] Wiegreffe S, Pinter Y. Attention is not not Explanation. arXiv preprint arXiv:1908.04626. 2019.

[17] Tschandl P, Rinner C, Apalla Z, Argenziano G, Codella N, Halpern A, et al. Human-computer collaboration for skin cancer recognition. Nat Med. 2020;26(8):1229-1234.

[18] Gaube S, Suresh H, Raue M, Merritt A, Berkowitz SJ, Lermer E, et al. Do as AI say: susceptibility in deployment of clinical decision-aids. Digit Med. 2021;4(1):31.

[19] Bansal G, Wu T, Zhou J, Fok R, Nushi B, Kamar E, et al. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. arXiv preprint arXiv:2006.14779. 2021.

[20] Adams R, Henry KE, Sridharan A, Soleimani H, Zhan A, Rawat N, et al. Prospective, multi-site study of patient outcomes after implementation of the TREWS machine learning-based early warning system for sepsis. Nat Med. 2022;28(7):1455-1460.

[21] Heard KL, Hughes S, Mughal N, Azadian BS, Moore LSP. Evaluating the impact of the ICNET clinical decision support system for antimicrobial stewardship. Antimicrob Resist Infect Control. 2019;8:51.

[22] Price WN, Gerke S, Cohen IG. Potential Liability for Physicians Using Artificial Intelligence. JAMA. 2019;322(18):1765-1766.

[23] Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366(6464):447-453.

[24] Food and Drug Administration. Artificial Intelligence in Software as a Medical Device. Silver Spring, MD: US Food and Drug Administration; 2026. Available from: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-software-medical-device.

[25] Eaneff S, Obermeyer Z, Butte AJ. The Case for Algorithmic Stewardship for Artificial Intelligence and Machine Learning Technologies. JAMA. 2020;324(14):1397-1398.