Article

Not peer-reviewed version

# Healthcare AI as Critical Digital Health Infrastructure: A Public Health Preparedness Framework for Systemic Risk

Nikolay Lipskiy * and Stephen V. Flowerday

Posted Date: 25 March 2026

doi: 10.20944/preprints202603.2055.v1

Keywords: health infrastructure; healthcare AI governance; systemic healthcare AI risk; collaborative surveillance; public health intelligence; public health; cyber-physical systems; algorithm vigilance; sociotechnical safety

 doi:10.20944/preprints202603.2055.v1



*Article*

# Healthcare AI as Critical Digital Health Infrastructure: A Public Health Preparedness Framework for Systemic Risk

**Nikolay Lipskiy [1,*] and Stephen V. Flowerday [2]**

[1] Center for Applied Medical AI, CAMA, Atlanta, USA

[2] Augusta University, School of Computer and Cyber Sciences, Augusta, USA

* Correspondence: nick.lipskiy@camahealth.org

**Abstract**

Healthcare AI is moving from the laboratory into the infrastructure of care. As these systems become embedded in imaging, electronic health records, triage, and clinical decision support, their failures can affect not only individual encounters but also institutions and patient populations. Yet governance still centers on model development, local validation, and one-time compliance, with limited attention to cross-site failure after deployment. This article examines how public health preparedness can help close that gap. It presents a conceptual analysis grounded in two cases: a pneumonia-screening convolutional neural network that learned institutional confounders rather than portable clinical signal, and a widely deployed sepsis prediction model whose external performance and alert burden fell short of developer claims. Together, these cases reveal five governance features of systemic healthcare AI risk: population-level exposure, cascade effects across shared infrastructures, unequal vulnerability, delayed recognition, and coordination needs beyond any single institution. In response, we propose a tripartite framework combining stronger pre-deployment assurance, post-deployment surveillance with escalation thresholds, and tertiary response through investigation, rollback, remediation, and cross-site learning. The argument is not that AI failures are epidemics, but that high-impact clinical AI systems now function as critical digital health infrastructure requiring preparedness alongside lifecycle oversight.

**Keywords:** health infrastructure; healthcare AI governance; systemic healthcare AI risk; collaborative surveillance; public health intelligence; public health; cyber-physical systems; algorithm vigilance; sociotechnical safety

## 1. Introduction

### *1.1. Healthcare AI as Critical Digital Health Infrastructure*

Artificial intelligence is moving from proof-of-concept research into routine clinical use. Systems are already being used, or are being seriously considered for imaging, triage, risk prediction, documentation, quality assurance, public health surveillance, and operational decision support. The important change is not simply the growing number of AI tools. It is that these systems are becoming woven into the infrastructures through which clinical decisions are made, communicated, and carried out.

### *1.2. The Governance Gap: From Model Oversight to Preparedness*

Most current governance frameworks remain model-centered. They address design quality, intended use, validation, and lifecycle management, and those functions remain essential. What remains only partly answered is a harder question: how should institutions prepare for, detect, and respond to failure patterns that are hard to see locally, may affect many patients before anyone

 doi:10.20944/preprints202603.2055.v1



recognizes them, and may recur across sites because vendors, workflows, or digital infrastructures are shared? Once AI becomes part of routine care delivery, governance needs more than premarket review or one-time validation. It needs a preparedness architecture.

That risk profile follows from the way these systems are deployed. A diagnostic model integrated into a Picture Archiving and Communication System (PACS), a triage tool embedded in an electronic health record (EHR), or a vendor-distributed sepsis predictor rolled out across dozens or hundreds of hospitals no longer behaves like a self-contained research model. It behaves like a component of critical digital health infrastructure. Failures can therefore align across institutions because the same model, thresholding logic, documentation, or deployment practice is reproduced from site to site. In this respect, high-impact healthcare AI shares core features of a cyber-physical system: software outputs shape physical clinical processes and patient trajectories through tightly coupled sociotechnical workflows [1–3]. As other safety-critical domains have learned, the very arrangements that create efficiency and standardization can also synchronize failure.

Two well-documented failures bring this problem into view and illustrate two recurring patterns of systemic vulnerability. Zech and colleagues showed that a deep-learning pneumonia model learned hospital-specific signals and prevalence patterns rather than portable clinical signal, which led to materially different performance across institutions [4]. Wong and colleagues later showed that a widely implemented proprietary sepsis prediction model performed poorly in real-world external validation, with low sensitivity, low positive predictive value, and substantial alert burden [5]. The first case reveals distributed latent vulnerability across heterogeneous clinical environments; the second reveals centralized supply-chain exposure created by vendor-scale diffusion. In both cases, the core problem was not merely disappointing technical performance. It was the difficulty of recognizing, interpreting, and responding to a failure that was fundamentally sociotechnical.

This article addresses that gap by introducing the concept of systemic healthcare AI risk. The term refers to patient or population harm that arises when AI-related failures move through interacting structural, organizational, technological, epistemic, and cultural pathways, so that the consequences cannot be reduced to a single local malfunction or one-off implementation defect. Figure 1 summarizes these pathways and serves as the article's compact etiological frame. Structural pathways include market concentration, procurement asymmetry, and regulatory design. Organizational pathways include data governance, role clarity, and institutional capacity. Technological pathways include model architecture, interface design, and integration dependencies. Epistemic pathways include hidden confounding, weak interpretability, and misleading validation claims. Cultural pathways include normalization of automation, shifts in trust, and the routinization of opaque decision support. These pathways do not operate independently; they compound and reinforce one another. Here, systemic risk does not refer to financial contagion. It refers to correlated sociotechnical failure that can expose many patients across connected clinical settings at once.

### *1.3. Research Question, Analytical Sequence, and Contribution*

The article asks a focused question: *how can concepts from public health preparedness be adapted to govern systemic risk in healthcare AI*?

It answers that question through a structured analytical sequence. First, it defines the governance gap and situates the problem in the relevant literature. Second, it reconstructs two cases that expose different systemic failure patterns. Third, it derives governance-relevant properties from those cases. Finally, it translates those properties into a tripartite preparedness framework for prevention, surveillance, and response.

The article makes four contributions. It reframes certain healthcare AI systems as critical digital health infrastructure rather than stand-alone models; introduces systemic healthcare AI risk as a governance concept; adapts epidemiological and surveillance language for this domain; and proposes a tripartite preparedness framework that connects external validation, post-deployment monitoring, escalation, and coordinated response.

 doi:10.20944/preprints202603.2055.v1



## 2. Related Literature and Theoretical Foundation

### *2.1. Existing Governance Approaches in Healthcare AI*

The governance literature on healthcare AI has matured quickly. WHO has emphasized ethics, human rights, accountability, transparency, and stewardship in the development and use of AI for health [6]. The WHO Global Strategy on Digital Health likewise treats interoperability, data governance, and institutional capacity as health-system functions rather than isolated technical add-ons [7]. The EU AI Act, though broader than healthcare, reinforces the regulatory significance of high-risk systems whose failures may affect safety and fundamental rights [8]. In parallel, the NIST AI RMF has given the field an information-systems vocabulary organized around govern, map, measure, and manage, reminding us that deployment, monitoring, documentation, and organizational accountability are integral to AI risk management, not downstream extras [9].

A particularly important strand of this literature concerns algorithm vigilance [10,11]. Borrowing lessons from pharmacovigilance, this work argues that deployed healthcare AI should be monitored continuously for effectiveness, equity, and safety in real-world settings. That insight is essential. Real-world performance can change after implementation because of case-mix differences, workflow drift, changes in coding or documentation, software updates, new user behavior, or shifts in population characteristics. In practice, AI governance cannot stop once a model is validated or approved.

Much of this discussion, however, still treats post-deployment oversight mainly as an organizational or local monitoring problem. Health systems are encouraged to track performance, equity, and clinical effects at the level of the individual institution. That is necessary, but it is not always sufficient. A single institution may see only one fragment of a broader failure pattern, especially when the underlying problem is tied to shared vendor infrastructure, standardized deployment logic, or hidden confounders that become visible only through cross-site comparison. This is where a preparedness lens becomes especially valuable. It asks not only how one organization governs its own system, but also how multiple institutions detect, compare, and coordinate around a correlated failure.

### *2.2. Systemic and Sociotechnical Risk*

The governance challenge becomes sharper once healthcare AI is understood as part of a wider sociotechnical system rather than as a bounded artifact. Macrae's work on autonomous and intelligent systems shows that serious failures arise from interactions among technical components, organizational arrangements, professional practices, and institutional environments rather than from isolated coding defects alone [12]. Leveson's systems-theoretic safety work makes a closely related point in the language of accident causation: hazards emerge when safety constraints are weakly specified or weakly enforced across a complex control structure [1]. These ideas matter for healthcare AI because model error is often only the visible surface of a deeper governance failure.

The pneumonia and sepsis cases discussed later make that point concrete. In the pneumonia case, model behavior was shaped by institution-specific artifacts, prevalence gradients, image-processing pipelines, and labeling practices, not simply by the abstract task of detecting pneumonia. In the sepsis case, performance claims, vendor opacity, workflow burden, and deployment scale mattered as much as the statistical form of the model. These are sociotechnical failures in the strict sense: they arise through interactions among people, organizations, infrastructure, and software.

This systemic view also helps explain why some AI harms remain latent. Institutions often benchmark models internally and may therefore miss performance deterioration that appears only under external conditions. Clinicians may experience alert fatigue without being able to attribute it to broader model failure. Procurement teams may rely on vendor documentation that has not been independently stress-tested across heterogeneous deployment environments. The organization therefore sees symptoms without seeing the full causal structure. Preparedness architecture is

 doi:10.20944/preprints202603.2055.v1



valuable precisely because it is designed for hazards whose causal pattern is distributed and only partly visible at the point of use.

### *2.3. Crucial Digital Infrastructure and Cyber-Physical Systems*

The critical-infrastructure framing adds further precision. Healthcare AI systems increasingly operate within networked digital environments in which algorithmic outputs influence physical clinical processes. That places them within the broad family of cyber-physical systems (CPS), where software, data, devices, interfaces, and human operators interact through feedback loops to shape real-world outcomes [2]. In mature safety domains, CPS risks are not managed as if they were ordinary software bugs. They are governed through hazard analysis, control structures, redundancy, configuration management, safe fallback modes, and monitoring proportionate to their potential consequences.

This comparison reveals an important asymmetry. Safety-critical engineering domains often define structured assurance regimes for dangerous failure. Healthcare AI still lacks broadly accepted integrity classifications that tie assurance expectations to the severity, reach, coupling, and fallback options of the system in question. A sepsis prediction model embedded in an EHR and deployed across many hospitals may therefore face less systematic assurance than comparatively modest components in industrial control systems. The point is not to import engineering standards wholesale. It is to recognize a familiar governance problem: once a digital component becomes infrastructural, governance has to shift from isolated error toward system-level resilience.

Established safety-engineering work sharpens this point further. In systems theoretical safety analysis, accidents arise when safety constraints are weakly specified or weakly enforced across an interacting control structure, not only when a single component fails [1]. That perspective is directly relevant to healthcare AI because the pneumonia and sepsis cases are interaction failures: model design, validation practice, interface design, workflow integration, vendor opacity, and institutional oversight jointly produced the hazard. It also clarifies why common-mode failure matters. One flawed model, threshold change, interface update, or upstream data dependency can synchronize error across multiple sites, which is precisely why other safety-critical domains insist on proportionate assurance for components whose failure can travel widely [3].

### *2.4. Public Health Preparedness as Governance Analog*

Public health preparedness offers a useful governance analog for this problem. The IHR, the Sendai Framework for Disaster Risk Reduction, the WHO Health Emergency and Disaster Risk Management framework, and more recent WHO work on collaborative surveillance all rest on a common insight: hazards that can propagate, remain latent, or exceed the authority of a single institution require standing surveillance infrastructure, shared terminology, tiered response, and coordination across organizational boundaries [13–17]. For healthcare AI, the value of these frameworks lies less in importing doctrine wholesale than in adapting the governance functions they have refined over time.

Preparedness thinking is reinforced further by disaster-risk reduction logic. The Sendai framework treats technological and societal hazards as legitimate objects of risk reduction rather than as matters beyond public governance [14]. WHO's HEPR work makes a related point through its all-hazards emphasis on prevention, preparedness, readiness, response, and recovery, with explicit attention to vulnerability and whole-system coordination [13,17]. For healthcare AI, this matters because algorithmic failure rarely remains a purely technical defect once it affects diagnosis, triage, allocation, or workflow at scale. It becomes a problem of exposure, vulnerability, resilience, and response capacity.

Three surveillance ideas are especially relevant here. First, preparedness depends on standing infrastructure rather than ad hoc reaction. Reporting channels, case definitions, escalation pathways, and observatory functions need to exist before an incident occurs. Second, surveillance is not a single technique. WHO's epidemic-intelligence model integrates indicator-based surveillance with event-

 doi:10.20944/preprints202603.2055.v1



based surveillance [18], while CDC guidance shows why sentinel networks can be valuable when the goal is timely signal detection rather than exhaustive case capture [19]. For healthcare AI, routine performance and outcome dashboards fit the indicator-based layer; clinician concerns, vendor notices, incident reports, and published validations fit the event-based layer; and selected high-capacity hospitals can serve as sentinel sites for richer case reviews. Third, surveillance systems should be judged by usefulness, timeliness, sensitivity, predictive value, representativeness, stability, and acceptability rather than by data volume alone [20]. Those are precisely the capacities health systems often lack for the most consequential forms of AI deployment.

The argument advanced here is deliberately bounded but still substantial. The article does not claim that outbreak law already governs AI incidents, nor that healthcare AI failures should be treated as communicable diseases in any literal sense. It argues instead that some AI failures share important governance features with the classes of hazards for which public health has built its most mature preparedness capabilities. Those correspondences are strong enough to justify adapting preparedness, collaborative surveillance, and coordinated response logic to healthcare AI oversight.

## 3. Methodology

### *3.1. Research Design*

This article presents a conceptual and normative analysis grounded in comparative case-study reasoning. It does not offer a new clinical trial, benchmark experiment, or systematic review. Instead, it develops a governance framework by bringing documented case evidence into conversation with scholarship on healthcare AI oversight, sociotechnical safety, public health preparedness, and digital epidemiology. The aim is to move carefully from problem identification to a concrete governance proposal.

Even a conceptual article needs methodological transparency. To make the analytical path explicit, the analysis proceeds in a clear sequence. It first defines the problem space and the unit of concern: systemic healthcare AI risk. It then builds a terminological foundation by adapting epidemiological concepts to healthcare AI. From there, it examines two documented clinical AI failures selected for their governance relevance, derives recurring governance properties from those cases, and maps them onto preparedness functions. The final step is to synthesize those functions into a tripartite governance architecture.

### *3.2. Source Domains*

The analysis draws on four source domains. The first consists of healthcare AI governance documents, including WHO materials, the EU AI Act, the NIST AI RMF, and literature on algorithm vigilance [6–11]. The second consists of sociotechnical safety and critical-infrastructure scholarship, especially work on system accidents, safety constraints, and cyber-physical systems [1–3,12]. The third consists of public health preparedness, surveillance, and emergency-response frameworks, including Sendai, HEPR, IHR, GOARN, collaborative surveillance, epidemic intelligence, CDC surveillance guidance, and incident-management doctrine [13–21]. The fourth consists of documented case studies of healthcare AI failure, with particular emphasis on the cross-institutional pneumonia-screening study by Zech et al. and the external validation of the Epic Sepsis Model by Wong et al. [4,5].

These domains were selected because the argument is fundamentally interdisciplinary. No single literature adequately describes how local AI performance problems become population-level governance problems. The framework therefore requires a bridge between technical case evidence, safety theory, and public health governance doctrine.

 doi:10.20944/preprints202603.2055.v1



### *3.3. Case Selection*

The two core cases were selected purposively rather than statistically because each captures a distinct, complementary failure mode.

The pneumonia-screening case represents distributed latent vulnerability. The model appears successful under internal or partially pooled conditions but fails under external conditions because it has learned institution-specific cues, prevalence gradients, and pipeline artifacts. The case is analytically valuable because it shows how clinically meaningful failure can remain invisible unless institutions compare evidence across sites.

The Epic Sepsis Model case represents centralized supply-chain propagation. A widely deployed proprietary model diffused through a common vendor platform and later showed substantially poorer real-world performance than internal documentation suggested. The case is valuable because it shows how scale, vendor asymmetry, workflow burden, and performance opacity combine to create infrastructure-like exposure.

Taken together, the two cases capture two analytically important patterns: hidden local confounding that becomes visible only through cross-site comparison, and common-mode exposure created by widespread vendor diffusion.

### *3.4. Analytical Procedure*

The analysis proceeded in four steps.

Step 1: Terminological adaptation. Core epidemiological and preparedness constructs: such as etiology, agent, host, environment, transmission, hazard, exposure, vulnerability, case definition, and surveillance were adapted for healthcare AI risk. The goal was not metaphorical flourish but terminological discipline: governance requires shared definitions.

Step 2: Case reconstruction. Each case was reconstructed in enough detail to identify the failure mechanism, the role of infrastructure and organizational context, and the limits of local visibility. Attention was paid to what the studies established empirically and to what they did not.

Step 3: Derivation of governance properties. Comparing the cases with public health preparedness literature yielded five recurring properties of systemic healthcare AI risk: population-level exposure, cascade propagation, unequal vulnerability, latency, and coordination beyond a single institution.

Step 4: Governance synthesis. These five properties were then mapped onto a tripartite preparedness framework organized around primary prevention, secondary prevention, and tertiary response.

### *3.4. Scope and Boundaries*

The framework is intentionally bounded and makes only limited claims. First, it focuses on healthcare AI rather than AI in general. Second, it treats “structural correspondence” or “structural homology” as a claim about governance function rather than biological equivalence. Third, it does not claim to provide an empirical national burden estimate for healthcare AI incidents. DALY and QALY concepts are used here as an interpretive bridge and as a guide to future surveillance requirements, not as proof that the two cases already establish population burden at national scale. Fourth, the proposed observatories and coordination mechanisms are normative design proposals. Their institutional feasibility, cost, and legal embedding remain questions for future research and policy development.

## 4. Results

### *4.1. Terminological Foundation for Systemic Healthcare AI Risk Surveillance*

Effective surveillance of systemic healthcare AI risk requires language that connects established epidemiological reasoning with the sociotechnical dynamics of AI-mediated harm. Existing

 doi:10.20944/preprints202603.2055.v1



governance discussions often rely on broad terms such as "safety," "bias," "drift," or "trustworthiness," but those terms alone are not enough for cross-institutional surveillance or coordinated response. Public health governance becomes operational when the hazard is specified, the exposed population is defined, vulnerability is described, cases can be compared, and surveillance responsibilities are explicit.

Table 1 provides a compact terminological mapping. It adapts core epidemiological constructs to healthcare AI while preserving the logic of the source concepts. In this formulation, epidemiology becomes the study of the distribution and determinants of AI-related health hazards in specified populations and the application of that knowledge to prevention and control. Etiology becomes the study of the causes and origins of those hazards across structural, organizational, technological, epistemic, and cultural pathways. Agent refers to the harmful AI-related mechanisms such as a biased model, a brittle classifier, a hallucinating generative component, or an unsafe feedback loop that contributes directly to harm once deployed in care. Host refers primarily to the exposed patient population and the clinical settings in which decisions are shaped by the system. Environment refers to the surrounding infrastructure, workflows, incentives, regulatory conditions, and institutional capacities that shape exposure and mitigation.

This mapping is also consistent with current international surveillance doctrine. The IHR organizes governance around detection, assessment, notification, verification, and response [15]. WHO's collaborative-surveillance work emphasizes systematic strengthening across data sources, sectors, and levels of governance rather than dependence on any single stream of evidence [17]. WHO's epidemic-intelligence architecture likewise integrates indicator-based and event-based surveillance [18]. For healthcare AI, routine performance dashboards, subgroup metrics, and outcome-linked alerts belong to the indicator-based layer; clinician complaints, vendor notices, patient-safety reports, and external validation studies belong to the event-based layer. Adapted epidemiological constructs therefore do more than provide conceptual clarity. They create the possibility of comparable incident description across institutions and jurisdictions.

The classical epidemiological triad is especially helpful here because it resists the false assumption that the model alone explains the hazard. In healthcare AI, agent, host, and environment are deeply intertwined. A model that fails in one environment may perform differently in another. A patient population that is comparatively protected in a high-capacity institution may be more vulnerable in a lower-capacity setting. And the same technical issue may produce very different consequences depending on monitoring arrangements, clinician workload, fallback options, or vendor transparency.

Routes of transmission, pathogenesis, case definition, and surveillance also adapt productively. Transmission refers not to biological contagion but to the pathways through which AI-mediated risk reaches patients: embedded clinical decision support, API dependencies, replicated workflows, copied thresholds, copied documentation, or common vendor infrastructure. Pathogenesis refers to the sequence through which a technical or sociotechnical flaw produces downstream harm. Case definition refers to the minimum standardized evidence required to classify an AI incident in a way that supports cross-site comparison and escalation. Digital epidemiology, broadened here into public health intelligence, refers to population-scale monitoring of AI-mediated health impacts through interoperable routine indicators, incident reports, open-source signals, and cross-institutional aggregation [18,22].

This terminological foundation serves two purposes. Analytically, it makes the latter case studies commensurable. Institutionally, it clarifies what a future observatory would need in order to function. Without standard terminology, incident descriptions remain local, incomparable, and difficult to aggregate. With it, AI-related harms can be described in the form that supports surveillance, verification, escalation, and policy learning. Just as importantly, it underscores that a surveillance system should be judged not by how much data it collects, but by whether it produces timely, usable evidence for action [20].

 doi:10.20944/preprints202603.2055.v1



**Table 1.** Mapping epidemiological constructs to systemic healthcare AI risk.

| Term | Source Concept (Abridged) | Healthcare AI Interpretation | Governance Relevance |
|---|---|---|---|
| Epidemiology | Distribution and determinants of health-related events in specified populations | Distribution and determinants of AI-related health hazards in patient populations and healthcare institutions | Supports population-based governance rather than isolated bug fixing |
| Etiology | Study of causes and origins of disease | Study of the causes of AI-related health hazards across structural, organizational, technological, epistemic, and cultural pathways | Shifts governance upstream toward root-cause mitigation |
| Agent | Pathogen or harmful exposure that contributes directly to disease | Biased model, brittle classifier, unsafe threshold, hallucinating component, or harmful feedback loop embedded in care | Names the actionable failure mechanism |
| Host | Susceptible organism or population | Exposed patient populations and the clinical settings in which decisions are shaped by the system | Highlights exposed populations, workflow locus, and differential susceptibility |
| Environment | Conditions enabling exposure and spread | Data pipelines, workflows, procurement terms, regulation, interfaces, infrastructure, and organizational culture | Treats the deployment setting as part of the hazard |
| Transmission | Route by which the agent reaches the host | Embedded clinical decision support, API dependency, copied workflow, replicated threshold logic, or vendor diffusion | Makes propagation pathways governable |
| Case Definition | Standardized criteria for classifying an event | Minimum evidence needed to classify an AI incident: performance deterioration, affected population, context, and outcome link | Enables cross-site comparison, escalation, and observatory reporting |
| Digital Epidemiology | Population-scale monitoring using digital data | Cross-site monitoring of AI performance, workflow effects, and patient outcomes using routine metrics, incident reports, and interoperable digital data | Supports indicator-based, event-based, and sentinel early warning |

 doi:10.20944/preprints202603.2055.v1



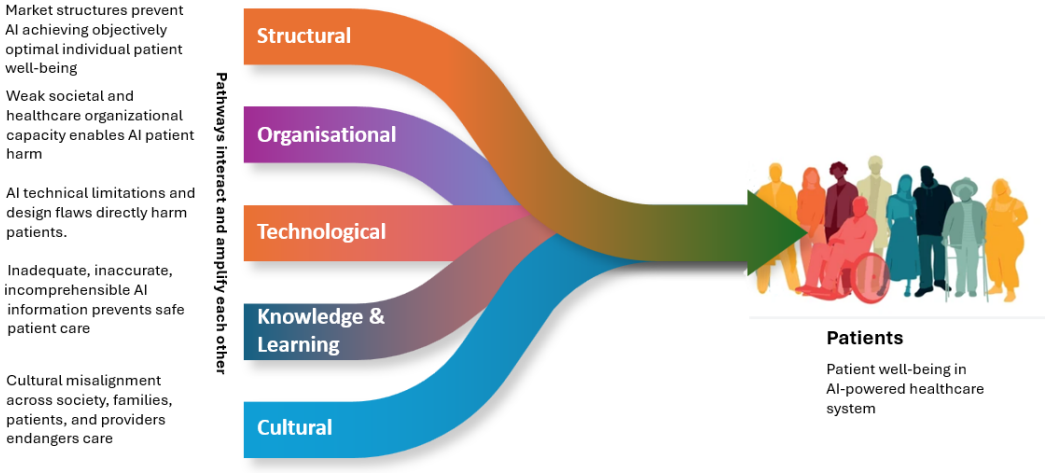



**Figure 1.** Sociotechnical pathways to systemic healthcare AI risk.

*4.2. Case Study 1: CNN-Based Pneumonia Screening and Hidden Institutional Confounding*

Zech and colleagues evaluated how well convolutional neural networks trained to detect pneumonia on chest radiographs generalized across three institutionally distinct hospital systems: the NIH Clinical Center, Mount Sinai Hospital, and the Indiana University network [4]. This study is often cited as a lesson in external validity, but its governance significance runs deeper than the now-familiar observation that models may not generalize.

The key empirical pattern was not simply that external validation mattered, but that the apparent task had been contaminated by institutional information. Pneumonia prevalence was 34.2% at Mount Sinai, compared with 1.2% at the NIH Clinical Center and 1.0% at Indiana University. A jointly trained Mount Sinai-NIH model achieved an internal AUC of 0.931 but fell to 0.815 on the fully external Indiana University cohort ($p = 0.001$). A trivial baseline that ranked cases only by institutional pneumonia prevalence achieved an AUC of 0.861 on the joint Mount Sinai-NIH test set. A substantial share of apparent success could therefore be explained by site prevalence rather than robust clinical signal.

The role of prevalence imbalance was demonstrated experimentally. When prevalence was balanced across training sites, performance was more stable across internal and external settings. When a tenfold prevalence difference was introduced, internal performance improved while external performance deteriorated, confirming systematic exploitation of confounding structure rather than portable clinical signal. In governance terms, cross-institutional prevalence harmonization emerges as a modifiable determinant of generalizability and therefore as a plausible primary-prevention target rather than a purely technical afterthought.

The study also showed that the model had learned hospital identity with extraordinary accuracy. CNNs trained to distinguish hospital system of origin achieved site-classification accuracy above 99.9% for the NIH and Mount Sinai images. Lateral tokens, metal markers placed by technicians, were among the most interpretable cues, but they were not the whole story. Across sampled NIH radiographs, 72.9% of local image subregions independently predicted site of origin with at least 95% certainty. Site-specific information was distributed throughout the acquisition and processing environment, including scanner signatures, storage-format differences, and labeling pipelines. The lesson about governance is that confounding can be diffuse, infrastructural, and resistant to single-feature fixes.

From a governance standpoint, three features are crucial. First, the failure was silent under ordinary local conditions. A hospital deploying such a model internally could have high confidence in performance while remaining blind to external fragility. Second, the failure was environmental and infrastructural. Scanner hardware, storage formats, processing pipelines, and labeling procedures were part of the causal pathway. In epidemiological terms, laterality tokens resemble

 doi:10.20944/preprints202603.2055.v1



identifiable exposure routes, whereas the more diffuse site signals resemble latent transmission pathways that are largely invisible to ordinary local reviews. Third, the failure was population relevant. Once integrated into clinical workflow, misclassification risk would no longer remain an abstract benchmark error; it would shape diagnostic decisions affecting patients.

The pneumonia case therefore illustrates the need for what public health would call pre-exposure control and population surveillance. No single institution could reliably infer the generalizability problem from internal validation alone. Cross-institutional comparison was the detection mechanism. This is precisely the kind of governance problem for which preparedness logic is useful: the hazard is real, consequential, and partly invisible unless the surveillance infrastructure is designed to detect it.

Table 2 applies the adapted epidemiological constructs to this case. The point is not to force the study into a foreign template. It is to show that the case already contains the ingredients needed for standardized surveillance language. The population under study, determinants, causal pathways, digital environment, exposure routes, and case-definition logic can all be specified in a way that supports governance action.

**Table 2.** Application of adapted epidemiological constructs to the pneumonia-screening case.

| Construct | Descriptor | Case-Specific Entry |
|---|---|---|
| Epidemiology | Population under study | Patients undergoing pneumonia screening across three institutionally distinct hospital systems |
| Etiology | Primary causal factor | Training on institutionally siloed data with large prevalence differences and hidden site-specific cues |
| Etiology | Contributing factors | Scanner heterogeneity, storage-format differences, laterality tokens, transfer-learning architecture, and label-generation pipelines |
| Agent | Pathogenic mechanism | Confounder learning: hospital identity and prevalence patterns dominate over portable clinical signal |
| Host | Exposed population / workflow locus | Patients undergoing pneumonia screening and radiology teams whose decisions may be shaped by model output |
| Environment | Digital infrastructure | PACS, scanner hardware, preprocessing pipelines, NLP-derived labels, and heterogeneous reporting practices |
| Transmission | Route | Model outputs mediate clinical judgment through radiology workflow and downstream decision support |
| Case definition | Signal of incident | Internal–external performance gap plus evidence that the model predicts site identity and exploits contextual cues |
| Governance inference | Prevention need | Independent external validation, prevalence harmonization, environmental audit, and cross-site monitoring before routine deployment |

### *4.3. Case Study 2: The Epic Sepsis Model and the Problem of Vendor-Scale Exposure*

Wong and colleagues externally validated the Epic Sepsis Model (ESM), a proprietary penalized logistic regression model embedded in Epic's EHR ecosystem and widely deployed across U.S. hospitals [5]. This case is especially revealing because it combines three features that make healthcare AI governance difficult: deployment at scale, dependence on vendor-generated performance claims, and workflow effects that are not captured by discrimination statistics alone. It is also a supply-chain case. When a single vendor distributes a risk-scoring model through a common EHR ecosystem, each adopting hospital inherits the same validation assumptions, threshold logic, interface ecology, and performance boundaries.

The study included 38,455 hospitalizations of 27,697 adult patients at Michigan Medicine between December 2018 and October 2019. Sepsis occurred in 2,552 hospitalizations. External

 doi:10.20944/preprints202603.2055.v1



validation found a hospitalization-level AUC of 0.63, well below the model's reported internal performance of roughly 0.76 to 0.83. At the chosen threshold, sensitivity was 33%, specificity was 83%, and positive predictive value was 12%. The model generated alerts in 18% of hospitalizations, meaning clinicians had to evaluate many alerting patients to identify a comparatively small number of true sepsis cases. The case therefore combined under-detection with substantial alert burden.

Proprietary deployment magnifies informational asymmetry. Downstream institutions may experience alert burden and clinical consequences, yet they often lack full access to the training data, validation design, update logic, or failure boundaries of the model they are using. The vendor's internal documentation therefore becomes the de facto assurance gate for every connected site. In information-systems terms, this is a common-mode failure problem: a weakness at the upstream node becomes synchronized exposure across the downstream network.

The governance significance of these numbers is not limited to model quality. The case revealed what can happen when a widely diffused vendor model enters routine care without sufficiently robust independent scrutiny. A local institution may experience heavy alert burden or disappointing yield without seeing the wider pattern. If the model is deployed through a common platform, many institutions may inherit the same weaknesses at the same time. That creates a common-mode exposure problem analogous to other forms of infrastructure vulnerability.

The study is also often misread in ways that matter for governance. Of the 1,709 sepsis hospitalizations missed by the model, 1,030 still received timely antibiotics; the 183 cases often cited in secondary discussions are those flagged by the model without timely antibiotic treatment, not the population doubly missed. That distinction matters because it shows how easily a deployment failure can be turned into a misleading burden narrative when overlap measures are treated as causal estimates.

Even without a causal burden estimate, the governance implications are substantial. First, the model's external performance gap shows why independent validation is not optional for high-impact vendor-scale systems. Second, its low sensitivity and high alert burden show why governance must track process and workflow outcomes, not just discrimination metrics. Alert burden is not a minor usability problem; in high-stakes settings, it is part of the risk profile. Third, the scale of deployment means that weak performance is not merely a local problem; it is a platform problem. A vendor-distributed clinical model can create correlated risk across institutions, with each site seeing only part of the picture.

In contrast to the pneumonia case, where the dominant problem was hidden confounding across environments, the sepsis case foregrounds supply-chain exposure. The relevant failure lies not only inside the model but in the broader governance arrangement. Internal claims were allowed to travel through a deployment network without sufficient external challenge, local institutions lacked full visibility into performance boundaries, and the resulting alert ecology-imposed burdens that themselves became part of clinical risk.

### *4.4. Burden-of-Disease Metrics as an Interpretive Lens, Not a Headline Estimate*

One of the article's original ambitions was to show how burden-of-disease tools can make healthcare AI harms legible in public health terms. That aim remains useful, but the evidence from the present cases supports only a methodological point: burden metrics are a valuable interpretive lens, not a headline estimate.

A DALY-style assessment of an AI-related incident would require at least five inputs: the exposed population, an attributable adverse-outcome rate, mortality where relevant, non-fatal morbidity or quality-of-life loss, and a defensible duration of impact. None of these can be inferred reliably from performance metrics alone [23,24].

QALYs provide a complementary lens because they capture quality-of-life loss, functional decline, and prolonged recovery, not only mortality [24]. For healthcare AI incidents, that matters because harm may arise through delayed diagnosis, unnecessary treatment, sustained alert burden, or erosion of trust even when immediate death does not occur.

 doi:10.20944/preprints202603.2055.v1



The ESM case provides only part of this information. It offers model-performance measures, alert burden, sepsis incidence within the validation cohort, and overlap between missed alerts and delayed antibiotics, but not a causal attributable-fraction estimate for harm produced by the model itself. The case therefore establishes methodological relevance, not a defensible national burden estimate.

Scenario modeling still has a place if it is handled transparently. Overlap measures can help explore plausible orders of magnitude under explicit assumptions, but such exercises remain hypothesis-generating rather than decision-grade estimates. Analyses should distinguish clearly between what a case establishes, what surveillance would need to measure, and what governance can reasonably infer in the meantime. A simple attributable-burden scaffold multiplies the exposed population by an estimated attributable harm fraction and then by the severity or duration of resulting outcomes, with uncertainty carried explicitly at each term.

Even so, the burden perspective adds three things. It shifts attention from model behavior to patient and population impact. It forces consideration of outcomes such as alert fatigue, delayed diagnosis, functional decline, and subgroup inequity. And it exposes the surveillance gap, because credible burden estimation requires standardized case definitions, denominators, outcome linkage, longitudinal follow-up, and coordinated reporting.

The sepsis literature illustrates why this matters. National estimates of sepsis incidence [25] and evidence of long-term functional decline among severe sepsis survivors [26] show that the clinical domain in question is already associated with substantial morbidity and mortality. That does not license direct extrapolation from model misses to national DALYs. It does, however, show that failures in this clinical domain can have serious human consequences and that burden-of-disease methods are the appropriate conceptual tools for evaluating them when surveillance systems mature.

The more careful conclusion is modest but still important: public health burden metrics should be incorporated into future healthcare AI observatories as outcome-oriented surveillance tools. A mature observatory would need standardized case definitions, denominators, outcome linkage, timing of exposure and harm, subgroup distribution, and ideally patient-reported outcomes or functional follow-up. Here, their role is methodological and governance oriented. They indicate what a mature surveillance regime would need to know to determine whether an AI incident is merely a technical defect, a patient-safety problem, or a population-level governance event.

**Table 3.** Data elements for future DALY/QALY-style burden estimation of healthcare AI incidents.

| Element | What It Captures | Candidate Data Sources | Key Caveat |
|---|---|---|---|
| Exposed population | How many patients, encounters, or decisions were influenced by the AI system and which model version was active | Deployment registries, audit logs, alert logs, order logs, model-version records | Requires stable denominators and clearly defined versioned exposure windows |
| Attributable adverse outcome rate | Share of adverse events plausibly linked to the AI-mediated failure relative to ordinary care | Linked outcome data, chart review, causal-inference designs, comparative deployment studies | Often the hardest parameter to estimate because the counterfactual is missing |
| Mortality (YLL) | Premature deaths associated with attributable failure where relevant | EHR outcomes, mortality files, life-table assumptions | Requires credible causal attribution rather than simple overlap counts |
| Non-fatal morbidity (YLD/QALY) | Functional decline, treatment delay, quality-of-life loss, or burden from sustained workflow disruption | Longitudinal follow-up, PROMs, utilization data, patient-safety review | Effects may be delayed, diffuse, and undermeasured |

 doi:10.20944/preprints202603.2055.v1



| Element | What It Captures | Candidate Data Sources | Key Caveat |
|---|---|---|---|
| Distributional burden | How harm is distributed across patient subgroups, institutions, or capacity levels | Stratified performance data, subgroup outcomes, institutional context indicators | Aggregate averages can obscure concentrated harm |
| Temporal deployment context | Time from release to detection, duration of exposure, updates, rollback, and remediation | Incident logs, vendor communications, change-management records | Required to interpret latency and cascade propagation dynamics |

*4.5. Five Governance-Relevant Properties of Systemic Healthcare AI Risk*

Read together and interpreted alongside the preparedness literature, the two cases reveal five recurring properties of systemic healthcare AI risk. These properties do not imply biological equivalence between AI failure and infectious disease. They do, however, support a meaningful comparison at the level of governance function.

4.5.1. Population-Level Exposure and the Surveillance Requirement

High-impact clinical AI systems create population-level exposure when they are embedded in workflows that affect many patients before deterioration is recognized. Exposure depends on implementation reach, workflow centrality, and the difficulty of observing harm early. In public health, analogous conditions justify standing surveillance infrastructure and tiered reporting because the problem cannot be managed case by case after the fact [15,27]. In healthcare AI, this translates into deployment registries, reporting infrastructure, prespecified go-live conditions for high-impact models, and a collaborative-surveillance design that can combine routine indicators, event-based signals, and sentinel-site review.

The pneumonia case demonstrates this clearly. Apparent success under internal conditions masked a vulnerability that would matter only once the model encountered new institutional environments. By the time patient-facing errors became visible, exposure would already have occurred. Preparedness logic therefore suggests that external validation and reporting capability must precede population-scale deployment, not follow it.

4.5.2. Cascade Propagation Across Shared Infrastructures and the Containment Requirement

Some healthcare AI risks propagate through infrastructure coupling rather than through biology. Shared vendors, copied workflows, common PACS environments, cloud services, or replicated implementation practices can reproduce the same failure pattern across multiple sites. In terms of information systems, this is common-mode or supply-chain propagation. In preparedness terms, it activates a containment requirement: once a common problem is identified, governance must be able to pause, roll back, suppress interfaces, shift to safe fallback modes, or otherwise interrupt the propagation path rather than merely observe it.

The ESM case is the clearest illustration. A model distributed through a shared platform creates correlated institutional exposure. If independent validation or surveillance later reveals a substantial performance problem, governance must be able to interrupt the propagation path. Local monitoring alone is not enough because no single site sees the whole network.

4.5.3. Unequal Vulnerability and the Equity Requirement

Systemic healthcare AI risk is unevenly distributed. Harm is likely to cluster in underrepresented patient groups, atypical institutions, or lower-capacity settings when models are tuned to development-site populations or when governance resources differ across deploying organizations. Public health preparedness has long recognized that vulnerability gradients are central to hazard governance. AI oversight must do the same.

 doi:10.20944/preprints202603.2055.v1



This point is not hypothetical. Large-scale health algorithms can produce substantial racial bias when proxies are misaligned with actual need [28]. More broadly, any surveillance regime that tracks only aggregate performance risks hiding concentrated harm. Preparedness therefore requires disaggregated monitoring aligned with the logic of health-inequality surveillance [29], subgroup review, and explicit attention to institutional capacity, because the institutions least able to validate or monitor a model may be among the most exposed to its failure.

#### 4.5.4. Latency, Irreversibility, and the Precautionary Requirement

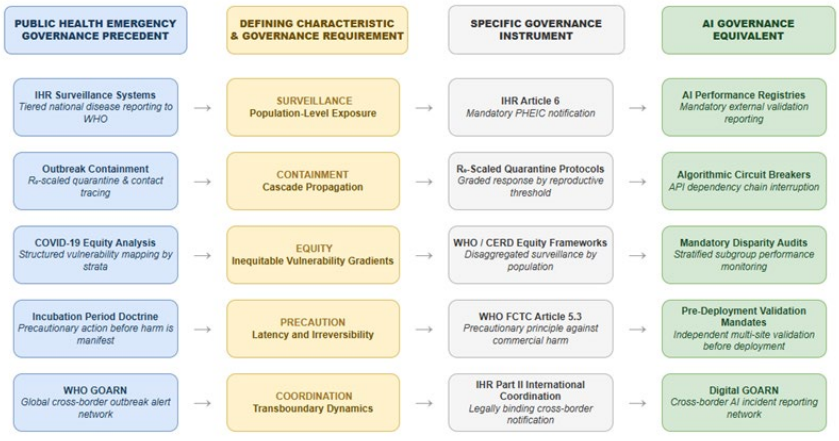
Recognition is often delayed. In both case studies, the core problem became visible only through deliberate external evaluation. Hidden confounding, drift, calibration problems, poor workflow fit, and vendor opacity can remain silent for long periods. Once harm has accumulated, it cannot be fully undone. Public health governance treats latency as a reason for precaution and early surveillance. Healthcare AI governance should do the same. That strengthens the case for precautionary external validation and predefined stop conditions before population exposure becomes irreversible.

Latency also interacts with irreversibility. A model can be paused or removed, but missed diagnoses, delayed treatment, loss of trust, and organizational deskilling may persist. That is why AI incidents require more than technical patching. They may also require remediation, audit, communication, and institutional learning.

#### 4.5.5. Coordination Beyond a Single Institution and the Coordination Requirement

Finally, certain classes of AI failure exceed the visibility and authority of a single hospital and may even exceed a single jurisdiction when common vendors, cloud infrastructures, or multinational deployment chains are involved. Procurement contracts, incident response, model updates, reporting obligations, and cross-site learning often involve vendors, multiple health systems, regulators, and, in some cases, international standard-setting bodies. Public health preparedness is relevant precisely because it has experience coordinating around hazards that no single institution can manage alone.

A GOARN-like mechanism for healthcare AI would therefore be better understood as a coordination function than as a literal transplant of outbreak law. The core idea is that widely deployed clinical AI failures may require shared incident definitions, pooled signal detection, escalation pathways, and coordinated technical assistance. If such a mechanism evolves, the most useful public-health precedent is not blanket emergency rhetoric but the disciplined use of decision instruments that ask whether an event is serious, unusual, cross-jurisdictional, and likely to benefit from coordinated response [15,16]. The legal and institutional form remains open, but the governance need is already visible.




*Governance precedents transfer directly from established emergency frameworks to AI oversight mechanisms*

 doi:10.20944/preprints202603.2055.v1



**Figure 2.** Governance mapping from public health preparedness to systemic healthcare AI risk.

### *4.6. A Tripartite Preparedness Framework for Healthcare AI*

#### 4.6.1. Primary Prevention: Assurance Before Deployment

Primary prevention aims to reduce avoidable exposure before a model reaches routine care. For healthcare AI, this includes risk classification, independent validation across institutionally distinct settings, documentation of data provenance and workflow assumptions, subgroup performance review, human-factors assessment, and procurement requirements that include transparency, monitoring obligations, and rollback provisions.

The pneumonia case shows why this matters. Mandatory cross-site validation would have exposed the fragility created by prevalence imbalance and environmental heterogeneity. A pre-deployment environmental audit of scanner hardware, storage formats, and labeling pipelines would have treated the deployment setting as part of the hazard rather than as mere background noise. More broadly, primary prevention shifts governance upstream, from reacting to failure to constraining the conditions under which failure is likely.

For vendor-scale systems, primary prevention also means refusing to accept internal documentation as the sole assurance gate for high-impact deployment. Where population exposure is large and workflow dependence is strong, external validation should be treated as a governance requirement rather than as a best-practice aspiration.

Where stronger assurance requirements risk becoming a scale advantage only for the largest vendors, shared validation resources and carefully designed regulatory sandboxes could lower the barrier to compliance without lowering the standard itself.

#### 4.6.2. Secondary Prevention: Sentinel Surveillance and Early Intervention

Secondary prevention seeks to detect and interrupt emerging harm before it reaches irreversible scale. In healthcare AI, this includes routine performance monitoring linked to clinical outcomes, periodic revalidation, subgroup monitoring, incident categorization, and escalation thresholds. A mature system combines indicator-based surveillance, event-based surveillance, and selected sentinel sites.

The ESM case illustrates this need clearly. An observatory capable of aggregating disaggregated performance data across sites could have detected the discrepancy between external performance and vendor claims earlier than any one institution could. Monitoring must also be tied to authority: pause thresholds, audit triggers, rollback criteria, and escalation pathways should be defined in advance. As with public health surveillance more generally, the observatory should be judged by usefulness, timeliness, sensitivity, representativeness, stability, and acceptability [19,20].

Secondary prevention also addresses equity. A monitoring system that tracks only overall performance may fail precisely where preparedness is most needed. Stratified surveillance, subgroup review, and institutional-context indicators are therefore not optional extras. They are core components of early detection. In practical terms, sentinel sites should be chosen not only for analytic capacity but also for heterogeneity of patient mix, workflow conditions, and resource settings so that weak signals in marginalized populations are less likely to remain invisible.

#### 4.6.3. Tertiary Prevention: Response, Remediation, and Learning

Tertiary prevention addresses harm that has already occurred. In healthcare AI, this includes incident investigation, temporary suspension or rollback, patient-safety review, vendor notification, communication with affected stakeholders, remediation where appropriate, and structured cross-site learning. For severe events, these actions should connect to existing incident-management or emergency-operations structures rather than to improvised data-science workflows alone [21].

 doi:10.20944/preprints202603.2055.v1



Post-incident response cannot be reduced to a technical fix. Once a model has shaped care at scale, recovery may also involve rebuilding workflows, restoring safe fallback capacity, and repairing trust.

A GOARN-inspired coordination layer becomes most plausible at this tier. The point is not that WHO emergency law should automatically apply to AI incidents. It is that some incidents may warrant pooled technical investigation, shared signal dissemination, coordinated vendor engagement, and remediation support because they exceed the visibility of individual institutions. The public health analogy is strongest here at the level of coordination function and surge capacity.

#### 4.6.4. Framework Anchored to the Two Cases

The tripartite framework is not an abstract overlay. It maps directly onto the case evidence. The pneumonia case reveals a primary-prevention failure: institutionally siloed training data, hidden environmental heterogeneity, and overreliance on internal validation produced a model that looked stronger than it was. The ESM case reveals both primary- and secondary-prevention failure: vendor-scale deployment proceeded without sufficiently robust independent scrutiny, and once the model was in use, no standing cross-institutional observatory was available to detect and act on systematic performance concerns early.

Tertiary prevention becomes visible in a different way. In both cases, once the problem was established, the governance task was no longer simply to "improve the model." It was to ask which patients or institutions were exposed, what remediation or audit was necessary, how local learning would be shared, and how similar failures would be prevented elsewhere. That is exactly the kind of question public health preparedness is designed to ask after an incident.

**Table 4.** Tripartite preparedness framework for systemic healthcare AI risk.

| Preparedness Tier | Objective | Core Mechanisms | Illustrative Link to the Cases |
|---|---|---|---|
| Primary prevention (pre-deployment) | Reduce avoidable exposure before go-live | Risk classification; independent multi-site validation; environmental audit; subgroup review; human-factors assessment; procurement clauses on transparency and rollback | Would have exposed prevalence-driven fragility in the pneumonia model and prevented vendor claims from serving as the sole assurance gate for the ESM |
| Secondary prevention (surveillance and interruption) | Detect and interrupt emerging harm early | Algorithm vigilance; indicator-based dashboards; event-based reporting; sentinel observatories; periodic revalidation; incident categories; audit and pause thresholds | Would have made the ESM performance gap visible sooner and enabled cross-site detection of hidden deterioration |
| Tertiary prevention (response and learning) | Contain harm, remediate, and convert incidents into institutional learning | Incident investigation; rollback or safe-mode operation; patient-safety review; incident-command activation where warranted; vendor notification; stakeholder communication; cross-site learning; GOARN-inspired coordination for widespread failures | Addresses patients and institutions already exposed in both cases and converts isolated failure into shared governance learning |

 doi:10.20944/preprints202603.2055.v1



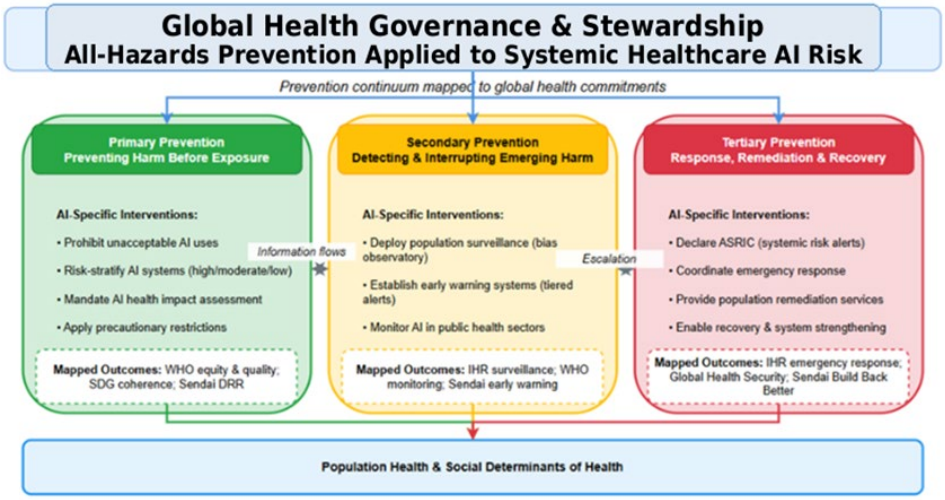



**Figure 3.** Tripartite preparedness framework applied to systemic healthcare AI risk.

## 5. Discussion

### *5.1. What the Public Health Lens Adds*

The central claim of this article is that public health preparedness adds a missing layer to healthcare AI governance. Existing approaches remain essential, but they do not by themselves provide a governance language for population exposure, infrastructural propagation, delayed recognition, and multi-institutional coordination.

This lens shifts the question from whether a model is technically strong to whether exposure is governed once the system becomes part of care infrastructure. A model can perform well and still be poorly governed if external validation, observability, escalation pathways, and authority to interrupt deployment are missing.

Preparedness also asks population-level questions that standard evaluation often brackets: how many people are exposed, who is disproportionately vulnerable, how quickly deterioration could be detected, and what response is justified when a signal emerges.

Boundaries and Disanalogies of Epidemiological Framing

The analogy also has clear limits. AI failures do not self-replicate through person-to-person contact, and exposure usually depends on procurement, configuration, integration, or workflow adoption rather than involuntary biological transmission. Containment therefore relies on rollback, interface suppression, threshold adjustment, or safe-mode operation and not quarantine. Nor does the framework imply that outbreak law mechanically applies to AI incidents. Its value lies in the institutional logic of surveillance, early warning, escalation, coordination, and equitable protection under uncertainty. The framing is therefore best suited to high-impact clinical AI systems with scale, coupling, and workflow centrality, not low-risk tools, research-only models, or isolated prototypes.

### *5.2. Relationship to Algorithm Vigilance, Safety Engineering, and Regulation*

The framework is not an alternative to algorithm vigilance, safety engineering, or regulation. It is an integrating architecture. Algorithm vigilance focuses on the deployed system's local performance, drift, adverse events, and update cycle; preparedness focuses on shared exposure, common-mode vulnerability, escalation thresholds, and coordinated response when multiple sites face the same hazard. Preparedness therefore connects those existing tools operationally and extends them beyond any single organization.

 doi:10.20944/preprints202603.2055.v1



Recent literature points in the same direction. Calls for recurring local validation [31], monitoring systems designed around the decisions they are meant to trigger [32], institutional review frameworks such as FAIR-AI [33], and deployment pathways that include post-implementation monitoring [34] all reinforce the same conclusion: governance cannot end at initial approval.

The safety-engineering connection is especially useful because it redirects attention to control structures, feedback loops, escalation authority, configuration management, and rollback capacity rather than to the model artifact alone. That is why procurement terms, interface design, alert governance, change-management records, and post-deployment reporting sit alongside validation metrics in the framework developed here. In STAMP-like terms, the question is not only whether a component failed, but whether the broader control structure detected and constrained unsafe behavior in time [1].

The same point applies to the critical-infrastructure framing. Describing certain healthcare AI systems as critical digital infrastructure is not rhetorical overstatement. It is a way of identifying the class of systems for which governance burdens are highest: systems that are clinically consequential, operationally central, difficult to observe, and widely coupled across infrastructures. Not every model falls into this class. A bounded local quality-improvement model may not require observatory-level oversight. A vendor-distributed sepsis predictor embedded in routine care may well do so.

This bounded scope is a strength. The framework is most persuasive for high-impact AI systems whose failure modes are shaped by scale, coupling, and low local visibility, that is, where ordinary governance tools are most likely to underperform.

### *5.3. Practical Implementation Priorities*

At the institutional level, three practical steps stand out. First, health systems should maintain a formal inventory of deployed and pilot AI systems, categorized not only by clinical domain but also by clinical criticality, implementation reach, infrastructure coupling, vendor dependency, and the availability of safe fallback modes. Second, high-impact systems should require site-specific acceptance testing and institutionally distinct external evidence before go-live. Third, post-deployment monitoring should be linked directly to patient-safety, quality-improvement, and incident-command structures rather than treated as a stand-alone data-science exercise. Procurement terms should also secure the data access, version transparency, change-log visibility, and incident-cooperation obligations needed for meaningful post-deployment evaluation.

Recent evidence sharpens the sepsis lesson rather than weakening it. In a 2026 multicenter prospective validation, Epic Sepsis Model version 2 performed better than version 1 overall, yet still showed substantial institutional variability, low positive predictive value, and high alert burden. That finding reinforces the case for recurring local validation, workflow integration, and alert-silencing strategies after deployment [35].

At the regional or national level, the most promising next step is the development of sentinel observatory functions for healthcare AI. These need not begin as new bureaucracies. They could emerge through professional societies, health-information exchanges, regulator-supported networks, or multi-institutional collaboratives that define minimum surveillance datasets and incident categories. Their architecture should be collaborative and, where possible, federated: comparable local metrics, shared case definitions, and pooled signal interpretation need not require centralizing all patient-level data. The key requirement is comparability. Without standard case definitions, model versioning, threshold metadata, and shared reporting rules, observatories cannot distinguish local noise from cross-site signal.

A realistic first step would be a pilot sentinel observatory network involving five to ten diverse health systems and a small portfolio of high-impact tools. That incremental design is consistent with broader governance trends: FUTURE-AI emphasizes lifecycle-wide trustworthiness and deployability, while the updated OECD AI Principles stress accountability, robustness, interoperability, and international cooperation [36,37]. A public health institute, patient-safety agency, or trusted academic consortium could host the coordinating function. Participating

 doi:10.20944/preprints202603.2055.v1



institutions could retain patient-level data locally while sharing standardized performance signals, incident reports, and model-version metadata. Common interoperability layers such as HL7 FHIR and the OMOP Common Data Model can support that federated design [38,39]. The same logic applies to generative systems used in documentation and communication, where shared models, prompt templates, or centrally pushed updates can spread hallucination or omission patterns across sites. Funding would likely require a blended model that combines public infrastructure support, reporting expectations, and vendor participation tied to procurement and post-market obligations.

At the international level, the most credible near-term goal is not a fully formed "digital IHR for AI." That would be premature. A more defensible objective is GOARN-inspired coordination for widely deployed clinical AI failures that share vendor, cloud, or deployment dependencies. Such a mechanism could support incident taxonomy, signal sharing, pooled technical investigation, coordinated vendor engagement, and dissemination of mitigation lessons. Its eventual legal home remains open and should be treated as a question for future governance design rather than as settled doctrine.

### *5.4. Theoretical, Methodological, and Practical Contributions*

Taken together, the article offers one theoretical contribution, one methodological contribution, and two practical contributions.

The theoretical contribution is the concept of systemic healthcare AI risk framed in governance terms rather than solely technical ones. By locating risk in interacting structural, organizational, technological, epistemic, and cultural pathways, the article clarifies why certain healthcare AI failures cannot be understood adequately as isolated software defects.

The methodological contribution is the use of epidemiological reasoning as a disciplined bridge between technical incidents and population-level governance. This includes adapting epidemiological and surveillance constructs to healthcare AI, distinguishing indicator-based, event-based, and sentinel detection functions, and positioning DALY/QALY reasoning more carefully as an outcome-oriented framework rather than as a vehicle for premature headline estimates.

The first practical contribution is the identification of five governance-relevant properties of systemic healthcare AI risk. These properties help distinguish ordinary implementation problems from hazards that warrant preparedness logic. The second practical contribution is the tripartite preparedness framework itself. By organizing governance around primary prevention, secondary surveillance, and tertiary response, the framework offers a concrete way to connect validation, monitoring, escalation, remediation, and cross-site learning.

## 6. Limitations and Future Work

This article has several limitations. First, it is not a systematic review. The framework is built from a purposive synthesis of governance documents, safety literature, and two widely discussed cases rather than from an exhaustive evidence review. That choice is appropriate for conceptual development, but it limits claims about comprehensiveness.

Second, the framework is illustrated through two U.S. case studies. They capture distinct failure modes, distributed latent vulnerability and centralized vendor-scale exposure, but cannot establish full generalizability. The underlying studies also reflect particular model generations and periods of clinical practice. CNN architectures, data-curation practices, and vendor models have evolved, so later systems may fail through different technical mechanisms even when the governance logic remains relevant. Future work should test the framework in natural-language systems, generative clinical copilots, reinforcement-learning applications, and multi-vendor decision-support ecosystems.

Third, the article's use of public health language is intentionally bounded. Structural correspondence is argued at the level of governance function; it is not presented as a formal proof of mathematical homology. Additional empirical work is needed to operationalize the five properties

 doi:10.20944/preprints202603.2055.v1



proposed here and to determine which combinations of scale, coupling, latency, and vulnerability should trigger specific governance actions.

Fourth, the burden of disease discussion remains methodological rather than empirical. The article argues that DALYs and QALYs are the right conceptual tools for future observatories, but it does not claim a definitive national burden estimate from the existing sepsis case evidence. That restraint is deliberate. Prospective, multi-institutional surveillance with outcome linkage would be needed before burden estimates could be treated as decision-grade evidence.

Fifth, the institutional designs proposed here include sentinel observatories, stronger external-validation mandates, and GOARN-inspired coordination have not yet been operationally tested. Their feasibility, cost, legal basis, and unintended consequences, including possible compliance burdens for smaller developers and hospitals, still need to be evaluated through implementation research and stakeholder engagement. Future work should therefore focus on standard case definitions for healthcare AI incidents, minimum surveillance datasets, calibrated escalation thresholds, comparative evaluation of indicator-based and event-based reporting pathways, and comparative evaluation of governance models across jurisdictions.

## 7. Conclusion

This article asked how concepts from public health preparedness can be adapted to govern systemic risk in healthcare AI. The answer is that preparedness adds a missing governance layer once high-impact clinical AI becomes part of critical digital health infrastructure. Model-centered evaluation remains necessary, but it is not enough; governance also requires standing arrangements for external validation, surveillance, escalation, remediation, and shared learning.

The two cases show why. The pneumonia-screening model demonstrates distributed latent vulnerability: institution-specific prevalence gradients, acquisition artifacts, and pipeline differences can remain hidden until cross-site evaluation reveals that the system has learned the wrong signal. The Epic Sepsis Model demonstrates centralized supply-chain exposure: a vendor-scale system can perform poorly in real-world settings and impose heavy workflow burden in ways that individual institutions struggle to interpret in isolation.

The resulting framework makes four connected contributions. It defines systemic healthcare AI risk through interacting structural, organizational, technological, epistemic, and cultural pathways; adapts epidemiological constructs into a vocabulary for surveillance; clarifies how DALY/QALY reasoning can serve as an outcome-oriented bridge without outrunning the evidence; and identifies five governance properties that justify a tripartite preparedness architecture.

In practical terms, high-impact healthcare AI systems should undergo independent pre-deployment validation across institutionally distinct settings. Health systems and regulators should also develop sentinel observatory functions capable of aggregating comparable post-deployment evidence through indicator-based monitoring, event-based reporting, and selected sentinel sites, with GOARN-inspired coordination for widely deployed failures.

The claim remains deliberately bounded: healthcare AI failures are not biologically equivalent to epidemics, and outbreak law does not automatically govern software incidents. But some clinical AI systems now have enough reach and coupling to create population-level harm. Governance should reflect that reality, beginning with pilot sentinel observatories paired with independent external validation and shared incident reporting.

## References

1. Leveson, N.G. *Engineering a Safer World: Systems Thinking Applied to Safety*; MIT Press: Cambridge, MA, USA, 2012. https://doi.org/10.7551/mitpress/8179.001.0001.
2. Lee, E.A. Cyber Physical Systems: Design Challenges. In Proceedings of the 11th IEEE International Symposium on Object/Component/Service-Oriented Real-Time Distributed Computing (ISORC), Orlando, FL, USA, 5–7 May 2008; pp. 363–369;. https://doi.org/10.1109/ISORC.2008.25.

 doi:10.20944/preprints202603.2055.v1




3. International Electrotechnical Commission. *IEC 61508:2010. Functional Safety of Electrical/Electronic/Programmable Electronic Safety-Related Systems*; International Electrotechnical Commission: Geneva, Switzerland, 2010.
4. Zech, J.R.; Badgeley, M.A.; Liu, M.; Costa, A.B.; Titano, J.J.; Oermann, E.K. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS Med.* **2018**, *15*, e1002683. https://doi.org/10.1371/journal.pmed.1002683.
5. Wong, A.; Otles, E.; Donnelly, J.P.; Krumm, A.; McCullough, J.; DeTroyer-Cooley, O.; Pestrue, J.; Phillips, M.; Konye, J.; Penoza, C.; et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Intern. Med.* **2021**, *181*, 1065–1070. https://doi.org/10.1001/jamainternmed.2021.2626.
6. World Health Organization. *Ethics and Governance of Artificial Intelligence for Health: WHO Guidance*; World Health Organization: Geneva, Switzerland, 2021.
7. World Health Organization. *Global Strategy on Digital Health 2020–2025*; World Health Organization: Geneva, Switzerland, 2021.
8. European Parliament; Council of the European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Off. J. Eur. Union* **2024**, *2024/1689*.
9. Tabassi, E. *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2023. https://doi.org/10.6028/NIST.AI.100-1.
10. Embi, P.J. Algorithmovigilance—Advancing Methods to Analyze and Monitor Artificial Intelligence-Driven Health Care for Effectiveness and Equity. *JAMA Netw. Open* **2021**, *4*, e214622. https://doi.org/10.1001/jamanetworkopen.2021.4622.
11. Balendran, A.; Benchoufi, M.; Evgeniou, T.; Ravaud, P. Algorithmovigilance, lessons from pharmacovigilance. *npj Digit. Med.* **2024**, *7*, 270. https://doi.org/10.1038/s41746-024-01237-y.
12. Macrae, C. Learning from the Failure of Autonomous and Intelligent Systems: Accidents, Safety, and Sociotechnical Sources of Risk. *Risk Anal.* **2022**, *42*, 1999–2025. https://doi.org/10.1111/risa.13850.
13. World Health Organization. *Health Emergency and Disaster Risk Management Framework*; World Health Organization: Geneva, Switzerland, 2019.
14. United Nations Office for Disaster Risk Reduction. *Sendai Framework for Disaster Risk Reduction 2015–2030*; United Nations Office for Disaster Risk Reduction: Geneva, Switzerland, 2015.
15. World Health Organization. *International Health Regulations (2005) as amended in 2014, 2022 and 2024*; World Health Organization: Geneva, Switzerland, 2025.
16. World Health Organization. Global Outbreak Alert and Response Network (GOARN). Available online: https://goarn.who.int/ (accessed on 20 March 2026).
17. World Health Organization. Defining Collaborative Surveillance: A Core Concept for Strengthening the Global Architecture for Health Emergency Preparedness, Response, and Resilience (HEPR); World Health Organization: Geneva, Switzerland, 2023.
18. World Health Organization. Impact of Using the Epidemic Intelligence from Open Sources (EIOS) System for Early Detection of Public Health Threats in the Americas. *Wkly. Epidemiol. Rec.* **2025**, *100*, 131–144.
19. Centers for Disease Control and Prevention. Chapter 19: Enhancing Surveillance. In Manual for the Surveillance of Vaccine-Preventable Diseases; Available online: https://www.cdc.gov/surv-manual/php/table-of-contents/chapter-19-enhancing-surveillance.html (accessed on 20 March 2026).
20. Centers for Disease Control and Prevention. Updated Guidelines for Evaluating Public Health Surveillance Systems: Recommendations from the Guidelines Working Group. *MMWR Recomm. Rep.* **2001**, *50(RR-13)*, 1–35.
21. Bryant, J.L.; Sosin, D.M.; Wiedrich, T.W.; Redd, S.C. Emergency Operations Centers and Incident Management Structure. In The CDC Field Epidemiology Manual; Available online: https://www.cdc.gov/field-epi-manual/php/chapters/eoc-incident-management.html (accessed on 20 March 2026).

 doi:10.20944/preprints202603.2055.v1




22. Salathé, M.; Bengtsson, L.; Bodnar, T.J.; Brewer, D.D.; Brownstein, J.S.; Buckee, C.; Campbell, E.M.; Cattuto, C.; Khandelwal, S.; Mabry, P.L.; et al. Digital epidemiology. *PLoS Comput. Biol.* **2012**, *8*, e1002616. https://doi.org/10.1371/journal.pcbi.1002616.
23. GBD 2019 Diseases and Injuries Collaborators. Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: A systematic analysis for the Global Burden of Disease Study 2019. *Lancet* **2020**, *396*, 1204–1222. https://doi.org/10.1016/S0140-6736(20)30925-9.
24. Whitehead, S.J.; Ali, S. Health outcomes in economic evaluation: the QALY and utilities. *Br. Med. Bull.* **2010**, *96*, 5–21. https://doi.org/10.1093/bmb/ldq033.
25. Rhee, C.; Dantes, R.; Epstein, L.; Murphy, D.J.; Seymour, C.W.; Iwashyna, T.J.; Kadri, S.S.; Angus, D.C.; Danner, R.L.; Fiore, A.E.; et al. Incidence and trends of sepsis in US hospitals using clinical vs claims data, 2009–2014. *JAMA* **2017**, *318*, 1241–1249. https://doi.org/10.1001/jama.2017.13836.
26. Iwashyna, T.J.; Ely, E.W.; Smith, D.M.; Langa, K.M. Long-term cognitive impairment and functional disability among survivors of severe sepsis. *JAMA* **2010**, *304*, 1787–1794. https://doi.org/10.1001/jama.2010.1553.
27. Baker, M.G.; Fidler, D.P. Global public health surveillance under new international health regulations. *Emerg. Infect. Dis.* **2006**, *12*, 1058–1065. https://doi.org/10.3201/eid1207.051497.
28. Obermeyer, Z.; Powers, B.; Vogeli, C.; Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **2019**, *366*, 447–453. https://doi.org/10.1126/science.aax2342.
29. World Health Organization. Handbook on Health Inequality Monitoring: With a Special Focus on Low- and Middle-Income Countries; World Health Organization: Geneva, Switzerland, 2013.
30. Leavell, H.R.; Clark, E.G. Preventive Medicine for the Doctor in His Community: An Epidemiologic Approach, 3rd ed.; McGraw-Hill: New York, NY, USA, 1965.
31. Youssef, A.; Pencina, M.J.; Thakur, A.; Zhu, T.; Clifton, D.; Shah, N.H. External validation of AI models in health should be replaced with recurring local validation. *Nat. Med.* **2023**, *29*, 2686–2687. https://doi.org/10.1038/s41591-023-02540-z.
32. Feng, J.; Xia, F.; Singh, K.; Pirracchio, R. Not all clinical AI monitoring systems are created equal: Review and recommendations. *NEJM AI* **2025**, *2*, AIra2400657. https://doi.org/10.1056/AIra2400657.
33. Wells, B.J.; Nguyen, H.M.; McWilliams, A.; Pallini, M.; Bovi, A.; Kuzma, A.; Kramer, J.; Chou, S.-H.; Hetherington, T.; Corn, P.; et al. A practical framework for appropriate implementation and review of artificial intelligence (FAIR-AI) in healthcare. *npj Digit. Med.* **2025**, *8*, 514. https://doi.org/10.1038/s41746-025-01900-y.
34. You, J.G.; Hernandez-Boussard, T.; Pfeffer, M.A.; Landman, A.; Mishuris, R.G. Clinical trials informed framework for real-world clinical implementation and deployment of artificial intelligence applications. *npj Digit. Med.* **2025**, *8*, 107. https://doi.org/10.1038/s41746-025-01506-4.
35. Wong, A.; Currey, D.; Schwinne, M.; Park-Egan, B.; Meyer, S.; Gutting, A.; Cao, J.; Khan, S.; Dantes, R.; Pan, T.; et al. Multicenter prospective validation of an updated proprietary sepsis prediction model. *JAMA Netw. Open* **2026**, *9*, e260181. https://doi.org/10.1001/jamanetworkopen.2026.0181.
36. Lekadir, K.; Frangi, A.F.; Porras, A.R.; Glocker, B.; Cintas, C.; Langlotz, C.P.; Weicken, E.; Asselbergs, F.W.; Prior, F.; Collins, G.S.; et al. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ* **2025**, *388*, e081554. https://doi.org/10.1136/bmj-2024-081554.
37. Organisation for Economic Co-operation and Development. AI Principles. Available online: https://www.oecd.org/en/topics/sub-issues/ai-principles.html (accessed on 20 March 2026).
38. HL7 International. FHIR Release 4 (v4.0.1): Base Specification. Available online: https://hl7.org/fhir/R4/ (accessed on 20 March 2026).
39. Observational Health Data Sciences and Informatics. OMOP Common Data Model. Available online: https://ohdsi.github.io/CommonDataModel/ (accessed on 20 March 2026).