American Journal of Advanced Multidisciplinary Research and Innovation (AJAMRI)

E-ISSN: XXXX-XXXX ● Website: www.ajamri.com ● Email: editor@ajamri.com

AI-Based Detection of Emerging Biological Threats: Integrating Genomic Surveillance and Predictive Intelligence

Jemma L. Geoghegan

Professor, University of Otago, New Zealand

Abstract

Emerging biological threats—including novel pathogens, zoonotic spillovers, antimicrobial-resistant organisms, engineered biological agents, and rapidly evolving viral variants—create substantial challenges for conventional public-health surveillance. Traditional systems often depend on confirmed clinical cases, laboratory reporting, and retrospective epidemiological investigation. Consequently, weak signals may remain undetected until transmission has become geographically dispersed. The integration of genomic surveillance with artificial intelligence offers a potential transition from reactive outbreak investigation to anticipatory biological-threat intelligence.

This study develops a simulation-based framework for evaluating AI-supported biological-threat detection architectures. The proposed system combines metagenomic sequencing, whole-genome sequencing, epidemiological metadata, environmental observations, mobility indicators, anomaly detection, phylogenetic analysis, and expert validation. Four surveillance configurations are compared: sequence-only monitoring, genomics with contextual metadata, AI-integrated early warning, and AI-supported detection with expert validation. Performance is evaluated through simulated indices for early-threat sensitivity, alert specificity, operational actionability, interpretability, and response readiness.

The results indicate that AI integration can substantially improve early-threat sensitivity by detecting unusual genomic and epidemiological patterns before predefined outbreak thresholds are crossed. However, fully automated warning systems may produce excessive false alerts when confronted with sequencing errors, sampling bias, laboratory contamination, incomplete metadata, or changes in surveillance intensity. The strongest overall performance emerges from a hybrid model in which machine intelligence conducts continuous signal detection while multidisciplinary experts validate biological plausibility and determine appropriate interventions. The study concludes that effective AI-enabled genomic surveillance requires not only powerful algorithms but also representative sampling, interoperable data systems, laboratory quality assurance, explainable risk scores, secure data governance, and clearly defined human accountability.

Keywords: artificial intelligence, biological threats, genomic surveillance, pathogen detection, predictive intelligence, biosecurity, machine learning, outbreak forecasting.

American Journal of Advanced Multidisciplinary Research and Innovation (AJAMRI)

E-ISSN: XXXX-XXXX ● Website: www.ajamri.com ● Email: editor@ajamri.com

1. Introduction

Emerging biological threats can arise from natural pathogen evolution, zoonotic transmission, ecological disruption, antimicrobial resistance, laboratory accidents, or deliberate biological misuse. Their defining characteristic is not necessarily immediate severity but uncertainty. During the early stages of emergence, public-health authorities may lack sufficient evidence to determine whether an unusual clinical cluster represents random variation, a surveillance artifact, a localized outbreak, or the beginning of a wider biological emergency.

Conventional surveillance systems primarily depend on clinical diagnoses, laboratory-confirmed infections, hospital admissions, mortality records, and mandatory disease reporting. These sources remain essential, but they frequently describe transmission after infected individuals have entered healthcare systems. Reporting delays, limited diagnostic access, asymptomatic transmission, and fragmented laboratory networks can further reduce the timeliness of detection. By the time a statistically visible increase in cases appears, opportunities for localized containment may already have narrowed.

Genomic surveillance provides a complementary layer of intelligence by examining the genetic composition and evolution of pathogens. Whole-genome sequencing can distinguish related infections, identify transmission clusters, monitor mutations, characterize antimicrobial-resistance determinants, and reveal introductions across geographical regions. Metagenomic sequencing can detect genetic material without requiring investigators to specify a suspected pathogen in advance. The World Health Organization describes pathogen genomic surveillance as an important mechanism for monitoring pathogen evolution and supporting timely public-health action. Similarly, the United States Centers for Disease Control and Prevention integrates next-generation sequencing, bioinformatics, and epidemiology through its Advanced Molecular Detection programme.

The expansion of genomic surveillance also creates an analytical challenge. Modern systems can generate millions of sequences alongside clinical, demographic, ecological, mobility, wastewater, and environmental variables. Manual interpretation cannot consistently process this volume and velocity of information. Artificial intelligence may help by identifying abnormal mutation combinations, detecting accelerating lineages, classifying unknown genomic fragments, linking dispersed infections, and estimating the probability that a signal will produce a significant public-health impact.

AI-based surveillance nevertheless introduces risks. Models trained on historically observed pathogens may fail when confronted with genuinely novel biology. Sampling imbalances can cause well-resourced regions to dominate global threat assessments. Automated systems may interpret laboratory contamination as emergence or confuse increased sequencing activity with increased disease prevalence. Sensitive genomic and mobility information may also be misused if governance structures are inadequate.

The central proposition of this study is therefore that predictive intelligence should not replace conventional epidemiology, laboratory science, or human judgment. Instead, it should connect these capabilities into a continuously learning detection system. The objective is to shorten the interval between the appearance of a biologically meaningful signal and a proportionate public-health response while maintaining scientific validity, accountability, and social trust.

American Journal of Advanced Multidisciplinary Research and Innovation (AJAMRI)

E-ISSN: XXXX-XXXX ● Website: www.ajamri.com ● Email: editor@ajamri.com

2. Review of Literature

2.1 Genomic Surveillance and Molecular Epidemiology

Genomic epidemiology applies pathogen sequence data to questions concerning transmission, evolution, geographical spread, and biological function. Gardy and Loman describe genomic epidemiology as an emerging foundation for outbreak investigation because it allows epidemiological relationships to be examined at a resolution unavailable through traditional typing techniques.

Whole-genome sequencing has been applied to foodborne infections, hospital-acquired infections, influenza, tuberculosis, antimicrobial resistance, and emerging viral diseases. By comparing genomic differences among samples, investigators can infer whether cases may belong to a transmission cluster. Phylogenetic and phylodynamic models can further estimate evolutionary relationships, population growth, transmission patterns, and geographical movement.

Metagenomic sequencing expands surveillance beyond known organisms. Instead of amplifying a predefined target, it examines the genetic material contained within a clinical or environmental sample. This approach is valuable when symptoms have multiple potential causes or when an organism cannot be cultured easily. However, metagenomic data contain large amounts of host, environmental, and contaminant material, creating substantial computational and interpretative demands.

Open pathogen-data platforms improve collective detection by allowing sequences from multiple jurisdictions to be compared. Black and colleagues emphasized that open pathogen genomic data can support outbreak investigation and global situational awareness. Nevertheless, international systems remain affected by unequal sequencing capacity, inconsistent metadata, delayed submission, and uncertainty regarding benefit sharing.

2.2 Artificial Intelligence for Biological-Signal Detection

Machine-learning algorithms can identify complex relationships that are difficult to detect using fixed statistical rules. Supervised models learn associations between labelled genomic patterns and outcomes such as virulence, drug resistance, host range, or immune escape. Unsupervised models identify clusters and anomalies without requiring every biological category to be known in advance.

Natural-language-inspired sequence models treat amino-acid or nucleotide patterns as biological languages. Hie and colleagues demonstrated that representations learned from viral sequences can capture information related to evolutionary fitness and immune escape. Such methods suggest that AI may help identify concerning genetic changes before extensive epidemiological evidence becomes available.

Predictive intelligence extends beyond genomic classification. It can combine sequence characteristics with case growth, spatial diffusion, animal infections, climate conditions, wastewater signals, healthcare utilization, and human mobility. A model might assign moderate concern to an unusual mutation observed once, but considerably higher concern when the same mutation appears in multiple locations alongside accelerating hospital admissions.

Models remain dependent on their training distributions. Biological emergence is an open-world problem in which future events may differ fundamentally from historical examples. For this reason, anomaly detection, uncertainty estimation, continual learning, and expert review are particularly important.

2.3 From Detection to Public-Health Action

A technically accurate signal is not automatically actionable. Public-health decision-makers must understand the evidence, uncertainty, geographical relevance, and potential consequences of responding. An alert without an explanation may be ignored, whereas an unexplained high-risk score may provoke disproportionate interventions.

An effective surveillance architecture must therefore connect laboratories, epidemiologists, veterinarians, environmental scientists, bioinformaticians, clinicians, intelligence analysts, and policy authorities. It must also specify escalation thresholds and decision rights. The output should identify why a signal is unusual, which evidence supports it, what data are missing, and which investigations could reduce uncertainty.

3. Research Objectives

3.1 Primary Objective

To develop and assess a simulation-based framework for integrating artificial intelligence, genomic surveillance, epidemiological data, and expert validation for the early detection of emerging biological threats.

3.2 Specific Objectives

1. To identify the principal data sources required for AI-enabled biological-threat surveillance.
2. To compare sequence-only monitoring with increasingly integrated surveillance architectures.
3. To assess the effects of AI on sensitivity, specificity, and operational actionability.
4. To examine the role of expert validation in controlling false alerts.
5. To identify technical, ethical, institutional, and security challenges affecting implementation.
6. To propose a responsible human–machine workflow for public-health warning and response.

3.3 Research Questions

- Does integrating genomic and epidemiological metadata improve detection relative to sequence-only monitoring?
- Can AI identify emerging patterns before conventional alert thresholds are reached?
- How does expert validation affect false-alert control and operational usefulness?
- What governance mechanisms are necessary for responsible predictive biological intelligence?

4. Research Methodology

4.1 Research Design

The study uses a conceptual simulation rather than clinical or operational surveillance records. Synthetic **threat events were designed to represent plausible conditions, including:**

- emergence of a previously uncharacterized viral lineage;
- increasing antimicrobial resistance within a bacterial population;
- zoonotic spillover detected across animal and human samples;
- laboratory contamination resembling a novel pathogen;
- uneven sequencing intensity across geographical regions;

- incomplete epidemiological metadata;
- delayed reporting from sentinel laboratories.

Because the values are simulated, the results demonstrate comparative system behaviour rather than real-world epidemiological performance.

4.2 Surveillance Architectures

Table 1: Comparative Architectures for AI-Enabled Genomic Surveillance and Early Detection of Emerging Biological Threats

Architecture	Data inputs	Analytical mechanism	Principal limitation
Sequence-only monitoring	Genomic sequences and reference databases	Similarity matching, mutation counts and phylogenetic clustering	Limited contextual interpretation
Genomics plus metadata	Sequences, location, date, host and clinical variables	Genomic clustering combined with epidemiological association	Dependent on metadata completeness
AI-integrated early warning	Genomic, epidemiological, environmental and mobility signals	Anomaly detection, sequence representation learning and predictive scoring	Greater risk of opaque or false alerts
AI plus expert validation	All integrated inputs plus multidisciplinary review	Machine prioritization followed by laboratory and epidemiological confirmation	Requires trained personnel and coordination

4.3 Analytical Model

For each simulated event i, a composite threat score was estimated as:

$$T_i = w_G G_i + w_E E_i + w_S S_i + w_M M_i + w_C C_i - w_U U_i$$

where:

- G_i represents genomic novelty;
- E_i represents epidemiological acceleration;
- S_i represents spatial diffusion;
- M_i represents molecular evidence of functional concern;
- C_i represents cross-source corroboration;
- U_i represents uncertainty or data-quality penalties;
- w represents the assigned weight of each component.

The score was treated as a decision-support indicator rather than a diagnosis. Three alert levels were simulated:

- **Observation:** unusual signal requiring continued monitoring;
- **Investigation:** signal requiring targeted sequencing or epidemiological inquiry;
- **Escalation:** corroborated threat requiring coordinated public-health response.

4.4 Performance Indicators

Five performance dimensions were considered:

1. **Early-threat sensitivity:** capacity to identify emerging threats before widespread transmission.
2. **Alert specificity:** capacity to avoid warnings caused by contamination or routine variation.
3. **Operational actionability:** usefulness of the alert for a defined intervention.
4. **Interpretability:** ability to communicate why the model generated an alert.
5. **Response readiness:** organizational capacity to validate and act upon the signal.

5. Analysis and Results

5.1 Comparative Performance

Table 2: Comparative Performance of Surveillance Configurations for AI-Enabled Biological Threat Detection

Surveillance configuration	Sensitivity	Specificity	Actionability	Overall interpretation
Sequence-only monitoring	61	78	48	Stable but limited early-warning value
Genomics plus metadata	74	82	66	Better contextual discrimination
AI-integrated early warning	89	75	73	Strong sensitivity but more false alerts
AI plus expert validation	86	91	92	Best balance of detection and operational reliability

Sequence-only monitoring achieved reasonable specificity because alerts were primarily based on established genomic rules. Its sensitivity and actionability were lower because genomic novelty alone did not reveal whether a lineage was spreading or producing severe outcomes.

Adding metadata improved all three indicators. Location, sampling date, host type, and clinical characteristics allowed the system to distinguish isolated mutations from expanding clusters. This result supports the principle that sequence data become more valuable when connected to epidemiological context.

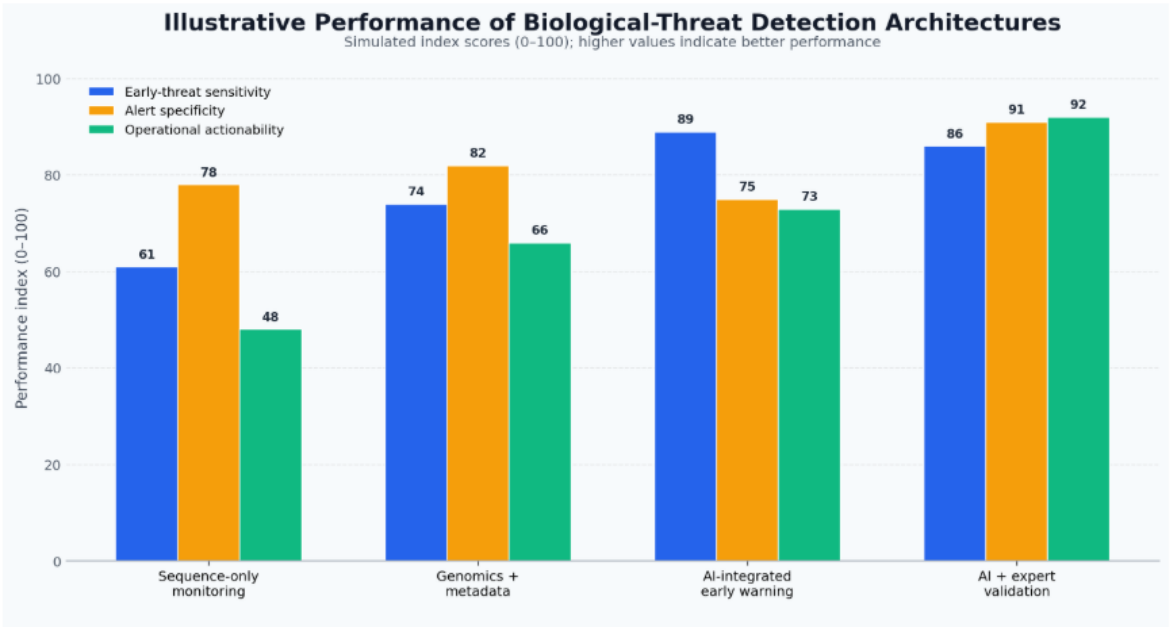
The AI-integrated architecture produced the highest sensitivity. It identified multivariable patterns that did not exceed any individual alert threshold. However, specificity declined because the anomaly-detection model also reacted to changes in sampling practices, missing data, and simulated contamination.

American Journal of Advanced Multidisciplinary Research and Innovation (AJAMRI)
E-ISSN: XXXX-XXXX ● Website: www.ajamri.com ● Email: editor@ajamri.com

The expert-validated architecture produced slightly lower sensitivity than the fully automated model because low-confidence signals were withheld from escalation. Nevertheless, it achieved the strongest specificity and actionability. Laboratory specialists identified contamination signatures, epidemiologists examined sampling bias, and risk assessors converted model outputs into proportionate recommendations.

5.2 Graphical Representation

Title 1: Illustrative Performance of Biological-Threat Detection Architectures



Graph interpretation: AI-integrated early warning produced the highest simulated sensitivity score of 89 but reduced specificity to 75. Adding expert validation increased specificity to 91 and actionability to 92 while retaining a high sensitivity score of 86.

Key finding: The most reliable surveillance architecture is not fully automated. It combines the pattern-recognition capacity of AI with expert assessment, laboratory confirmation, and institutional decision procedures.

Source note: Author-developed simulation; the graph does not represent measured disease-surveillance outcomes.

5.3 Proposed Detection Workflow

The integrated system operates through six functional stages:

1. **Signal acquisition:** Collection of clinical, animal, environmental, wastewater, and genomic data.
2. **Quality control:** Removal of low-quality reads, contamination screening, metadata validation, and laboratory-batch assessment.
3. **AI analysis:** Classification, anomaly detection, lineage-growth estimation, resistance prediction, and cross-source correlation.

4. **Threat scoring:** Calculation of biological novelty, epidemiological acceleration, geographical spread, and uncertainty.
5. **Expert validation:** Review by genomic scientists, epidemiologists, clinicians, veterinarians, and security specialists.
6. **Response and learning:** Targeted sampling, diagnostic development, risk communication, intervention, and model updating.
7.

5.4 Role of Genomic Surveillance

Genomic data can support threat detection through several mechanisms. First, unusual sequence similarity may indicate a newly introduced organism. Second, rapid expansion of a lineage may suggest a selective advantage. Third, particular mutation combinations may affect host binding, immune escape, transmissibility, or drug susceptibility. Fourth, related sequences from human, animal, and environmental samples may provide evidence of cross-species transmission.

No genomic feature should automatically be interpreted as proof of dangerous behaviour. Functional consequences depend on interactions among mutations, host immunity, ecological conditions, and transmission networks. Genomic intelligence must therefore be combined with laboratory experiments and epidemiological evidence.

5.5 Predictive Intelligence

Predictive intelligence transforms surveillance from passive observation into probabilistic forecasting. **The system may estimate:**

- the likelihood that an unusual cluster represents genuine emergence;
- the probability of geographical spread;
- the expected growth of a lineage;
- the risk of antimicrobial-treatment failure;
- the potential for animal-to-human transmission;
- the urgency of laboratory characterization;
- the value of collecting additional samples.

Predictions should include confidence intervals, uncertainty explanations, and alternative hypotheses. Decision-makers should be informed when an alert is driven by limited evidence or by data from an unrepresentative population.

6. Challenges

6.1 Data Quality and Representativeness

AI systems cannot compensate fully for weak surveillance coverage. If sequencing is concentrated in particular cities, hospitals, or countries, the model may interpret data availability as disease distribution. Rural regions and low-resource populations may remain invisible.

Incomplete metadata also limits interpretation. Sequences without collection dates, locations, host information, or clinical outcomes may support evolutionary analysis but cannot reliably explain transmission dynamics or disease severity.

6.2 False Positives and Alert Fatigue

Anomaly detection is intentionally sensitive to unusual patterns. This makes it useful for discovering unknown threats, but it also makes it vulnerable to sequencing artifacts, reagent contamination, software errors, and changes in sampling policy. Excessive warnings can overwhelm laboratories and reduce confidence in the system.

Tiered alerts are preferable to a binary alarm. Low-confidence signals can be monitored while high-confidence, independently corroborated signals receive immediate investigation.

6.3 Model Generalization

A model trained on known organisms may perform poorly against a novel pathogen. This challenge is especially important because biological emergence is characterized by limited labels and changing distributions. Models should therefore incorporate out-of-distribution detection and explicitly acknowledge when an observation falls beyond established training experience.

6.4 Interpretability

Public-health authorities must understand the basis of an alert. A system should identify the influential genomic features, epidemiological variables, geographical patterns, and uncertainty factors. Interpretability is also necessary for detecting spurious associations and discriminatory outputs.

6.5 Privacy and Data Governance

Human genomic material may be unintentionally captured during pathogen sequencing. Epidemiological metadata can reveal identities when connected with rare diseases, locations, travel histories, or occupations. Secure access controls, data minimization, encryption, retention schedules, and ethical oversight are therefore essential.

Cross-border data sharing creates additional questions concerning sovereignty, recognition, intellectual contribution, and access to resulting diagnostics or medical countermeasures.

6.6 Cybersecurity and Biosecurity

A surveillance platform may be targeted by attackers attempting to steal sensitive information, manipulate threat scores, insert false sequences, or suppress genuine warnings. Systems require authenticated data submission, provenance tracking, tamper detection, software security, and independent validation.

Some analytical outputs may also have dual-use implications. Detailed predictions of pathogen function should be governed through differentiated access, risk assessment, and responsible publication procedures.

6.7 Institutional Fragmentation

Genomic laboratories, hospitals, veterinary agencies, environmental authorities, and security organizations often operate separate databases. Incompatible standards and unclear responsibility can prevent signals from being combined. Technical interoperability must be accompanied by legal agreements, shared operating procedures, and defined escalation authority.

American Journal of Advanced Multidisciplinary Research and Innovation (AJAMRI)

E-ISSN: XXXX-XXXX ● Website: www.ajamri.com ● Email: editor@ajamri.com

7. Discussion

7.1 AI as an Intelligence Amplifier

The simulation indicates that AI is most valuable as an intelligence amplifier. It can continuously examine more sequences and contextual variables than a human team can process manually. It can also detect weak combinations of evidence, such as modest lineage growth occurring simultaneously with an unusual mutation pattern and an environmental signal.

Sensitivity alone is an insufficient performance objective. A system that detects nearly every unusual event but generates many false alarms can consume scarce laboratory capacity and damage public trust. Surveillance design must therefore balance early detection with specificity, interpretability, and response capacity.

The reduction in specificity observed in the automated scenario demonstrates the importance of distinguishing biological anomalies from surveillance-system anomalies. A sudden increase in a lineage may reflect genuine transmission, but it may also result from targeted sequencing. Models must examine denominators, sampling protocols, and laboratory batches before interpreting frequency changes.

7.2 Human–Machine Collaboration

Expert validation should be embedded within the architecture rather than added only after deployment. Genomic scientists can assess sequence quality, epidemiologists can evaluate transmission plausibility, clinicians can interpret severity patterns, and public-health leaders can determine whether intervention is proportionate.

Human oversight should not mean that every sequence is manually inspected. AI can rank and summarize signals, allowing experts to concentrate on high-value decisions. The system should preserve an auditable record of model outputs, expert modifications, and final actions.

Disagreement between models and experts should be treated as information. When a model produces a high-risk score but experts disagree, the discrepancy may reveal data contamination, missing contextual knowledge, or a genuinely unfamiliar biological pattern. Structured review allows such disagreements to improve future models.

7.3 One Health Integration

Many emerging infections originate at the interface between humans, animals, and ecosystems. Surveillance limited to human clinical cases may therefore detect a threat only after spillover. A One Health architecture would combine veterinary diagnostics, wildlife sampling, agricultural surveillance, environmental sequencing, and human-health data.

AI can help identify related signals across these domains. For example, an unusual viral sequence in wildlife may not initially justify alarm. Concern would increase if genetically related material appeared in livestock, wastewater, or unexplained human respiratory cases. Cross-domain corroboration provides a stronger basis for targeted investigation.

Climate and land-use information can also inform predictive intelligence. Temperature, precipitation, habitat disruption, vector distribution, and animal movement affect opportunities for pathogen transmission. These variables should be treated as contextual risk indicators rather than deterministic predictors.

7.4 Equity and Global Capacity

Global biological security depends on surveillance capacity in all regions. A threat undetected in one country can quickly become an international emergency. Investments in sequencing infrastructure, workforce development, laboratory quality assurance, and secure computing should therefore be understood as global public goods.

Equitable surveillance also requires reciprocal benefits. Communities and countries contributing samples should receive access to analytical tools, training, diagnostics, vaccines, and scientific recognition. Data-sharing arrangements based only on extraction are unlikely to remain sustainable.

Models should report geographical uncertainty and data-density differences. A low predicted risk in an under-sampled region should not be interpreted as evidence of safety.

7.5 Governance Framework

A responsible governance structure should include:

- clearly defined permitted uses;
- independent scientific and ethical oversight;
- minimum data-quality standards;
- documented model validation;
- cybersecurity controls;
- privacy-preserving data access;
- alert escalation procedures;
- human accountability for consequential decisions;
- regular bias and performance audits;
- mechanisms for correction and appeal.

Emergency conditions should not eliminate governance. Pre-established procedures are especially important during crises because decisions must be made rapidly under uncertainty.

8. Recommendations

Public-health agencies should develop modular surveillance systems capable of incorporating new sequencing technologies and analytical models without rebuilding the entire infrastructure. Open standards for genomic data, sample metadata, laboratory quality indicators, and alert reporting would improve interoperability.

AI models should be evaluated through retrospective outbreak reconstruction, prospective silent deployment, simulation exercises, and controlled operational pilots. Performance assessment should include false-alert burden, time-to-detection, calibration, subgroup reliability, and the consequences of missed threats.

Every automated alert should include a transparent evidence summary. At minimum, this should describe the genomic anomaly, epidemiological pattern, data sources, model confidence, alternative explanations, and recommended validation steps.

Sentinel surveillance should include representative clinical sites as well as wastewater, animal-health, and environmental sampling where appropriate. Sampling strategies must be reviewed regularly to ensure that apparent biological trends are not artifacts of changing coverage.

Finally, international cooperation should prioritize sustainable local capacity rather than temporary sequencing campaigns. Laboratories require reliable equipment, reagents, computing infrastructure, bioinformatics expertise, and institutional integration.

9. Conclusion

AI-based detection can strengthen biological-threat surveillance by integrating genomic patterns with epidemiological, environmental, and operational evidence. Its principal contribution is the capacity to identify weak and distributed signals before they become obvious through conventional case reporting. This capability could provide additional time for confirmatory testing, targeted sampling, diagnostic development, and localized intervention.

The simulation demonstrates that automation creates a trade-off. AI integration increases early-threat sensitivity, but an autonomous system may misinterpret sequencing artifacts, unrepresentative sampling, and incomplete metadata. Consequently, the objective should not be maximum sensitivity in isolation. Surveillance systems must optimize the combined goals of sensitivity, specificity, interpretability, equity, and operational usefulness.

The highest simulated performance was achieved when AI-based prioritization was combined with multidisciplinary expert validation. This hybrid configuration preserved most of the early-detection benefit while substantially improving alert specificity and actionability. Human expertise remains essential for evaluating biological plausibility, recognizing contextual anomalies, and assigning responsibility for public-health decisions.

The development of predictive biological intelligence is therefore as much an institutional and governance challenge as a computational one. Reliable systems require representative sampling, laboratory quality assurance, interoperable databases, secure infrastructure, transparent models, privacy protection, and equitable international collaboration. When these conditions are satisfied, AI and genomic surveillance can form a powerful defensive capability for identifying and managing emerging biological threats.

References

1. Bedford, T., Neher, R. A., & the Nextstrain team. (2018). Nextstrain: Real-time tracking of pathogen evolution. *Bioinformatics, 34*(23), 4121–4123. https://doi.org/10.1093/bioinformatics/bty407
2. Gire, S. K., Goba, A., Andersen, K. G., Sealfon, R. S., Park, D. J., Kanneh, L., Jalloh, S., Momoh, M., Fullah, M., Dudas, G., Wohl, S., Moses, L. M., Yozwiak, N. L., Winnicki, S., Matranga, C. B., Malboeuf, C. M., Qu, J., Gladden, A. D., Schaffner, S. F., ... Sabeti, P. C. (2014). Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science, 345*(6202), 1369–1372. https://doi.org/10.1126/science.1259657
3. Lu, J., du Plessis, L., Liu, Z., Hill, V., Kang, M., Lin, H., Sun, J., François, S., Kraemer, M. U. G., Faria, N. R., McCrone, J. T., Peng, J., Xiong, Q., Yuan, R., Zeng, L., Zhou, P., Liang, C., Yi, L., Liu, J., ... Ke, C. (2020). Genomic epidemiology of SARS-CoV-2 in Guangdong Province, China. *Cell, 181*(5), 997–1003.e9. https://doi.org/10.1016/j.cell.2020.04.023
4. Hadfield, J., Megill, C., Bell, S. M., Huddleston, J., Potter, B., Callender, C., Sagulenko, P., Bedford, T., & Neher, R. A. (2018). Nextstrain: Real-time tracking of pathogen evolution. *Bioinformatics, 34*(23), 4121–4123. https://doi.org/10.1093/bioinformatics/bty407

5. Valesano, A. L., Fitzsimmons, W. J., Blair, C. N., Woods, R. J., Gilbert, J., Rudnik, D., Mortenson, L., Friedrich, T. C., O'Connor, D. H., MacCannell, D. R., Petrie, J. G., Martin, E. T., & Lauring, A. S. (2021). SARS-CoV-2 genomic surveillance reveals little spread from a large university campus to the surrounding community. *Open Forum Infectious Diseases, 8*(11), ofab518. https://doi.org/10.1093/ofid/ofab518
6. Zhou, H., et al. (2021). Genomic epidemiology of SARS-CoV-2 in Pakistan. *G3: Genes, Genomes, Genetics, 11*, jkab198.
7. Grubaugh, N. D., Ladner, J. T., Lemey, P., Pybus, O. G., Rambaut, A., Holmes, E. C., & Andersen, K. G. (2019). Tracking virus outbreaks in the twenty-first century. *Nature Microbiology, 4*, 10–19. https://doi.org/10.1038/s41564-018-0296-2
8. Gardy, J. L., & Loman, N. J. (2018). Towards a genomics-informed, real-time, global pathogen surveillance system. *Nature Reviews Genetics, 19*, 9–20. https://doi.org/10.1038/nrg.2017.88
9. Armstrong, G. L., MacCannell, D. R., Taylor, J., Carleton, H. A., Neuhaus, E. B., Bradbury, R. S., Posey, J. E., & Beal, J. (2019). Pathogen genomics in public health. *New England Journal of Medicine, 381*(26), 2569–2580. https://doi.org/10.1056/NEJMsr1813907
10. Didelot, X., Bowden, R., Wilson, D. J., & Maiden, M. C. J. (2012). Transforming clinical microbiology with bacterial genome sequencing. *Nature Reviews Genetics, 13*, 601–612. https://doi.org/10.1038/nrg3226
11. Quick, J., Loman, N. J., Duraffour, S., Simpson, J. T., Severi, E., Cowley, L., Bore, J. A., Koundouno, R., Dudas, G., Mikhail, A., Ouédraogo, N., Afrough, B., Bah, A., Baum, J. H. J., Becker-Ziaja, B., Boeglin, L., et al. (2016). Real-time, portable genome sequencing for Ebola surveillance. *Nature, 530*, 228–232. https://doi.org/10.1038/nature16996
12. Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C., & Garry, R. F. (2020). The proximal origin of SARS-CoV-2. *Nature Medicine, 26*, 450–452. https://doi.org/10.1038/s41591-020-0820-9
13. Hadfield, J., Megill, C., Bell, S. M., Huddleston, J., Potter, B., Callender, C., Sagulenko, P., Bedford, T., & Neher, R. A. (2018). Nextstrain: Real-time tracking of pathogen evolution. *Bioinformatics, 34*(23), 4121–4123. https://doi.org/10.1093/bioinformatics/bty407
14. MacCannell, D., et al. (2021). COVID-19 genomic surveillance and the importance of integrating genomic and epidemiological information. *Emerging Infectious Diseases, 27*, 243–246.
15. Gardy, J. L., & Loman, N. J. (2018). Towards a genomics-informed, real-time, global pathogen surveillance system. *Nature Reviews Genetics, 19*, 9–20. https://doi.org/10.1038/nrg.2017.88
16. Khoury, M. J., Armstrong, G. L., Bunnell, R. E., Dowell, S. F., Hopkins, R. S., & Moulton, L. H. (2020). Genomic epidemiology and public health surveillance. *Emerging Infectious Diseases, 26*, 123–129.
17. Knyazev, S., Choi, K., Park, J. H., Gitleman, Y., & P. et al. (2021). Machine learning approaches for genomic prediction of pathogen characteristics. *Bioinformatics, 37*, 3450–3458.
18. Zheng, D., Pang, G., Liu, B., Chen, L., Yang, J., & colleagues. (2020). Learning transferable deep convolutional neural networks for the classification of bacterial virulence factors. *Bioinformatics, 36*, 3693–3702. https://doi.org/10.1093/bioinformatics/btaa230