

AI-Based Genomic Surveillance: Strengthening Preparedness for Emerging Infectious Diseases

Carolyn Stein

Assistant Professor University of California, Berkeley

Abstract

The emergence and rapid spread of infectious diseases demonstrate the need for surveillance systems capable of detecting biological threats before they develop into large-scale public health emergencies. Conventional surveillance approaches often depend on clinical reporting, laboratory confirmation and retrospective epidemiological analysis, which can delay recognition of novel pathogens and emerging variants. AI-based genomic surveillance combines pathogen sequencing, bioinformatics, machine learning, epidemiological intelligence and real-time data integration to create more responsive approaches to infectious disease monitoring. This paper examines the role of artificial intelligence in strengthening genomic surveillance for emerging infectious diseases, with particular attention to pathogen detection, genomic classification, variant identification, mutation analysis, transmission monitoring, outbreak forecasting and genomic epidemiology. It proposes an integrated AI-enabled surveillance framework connecting clinical laboratories, sequencing platforms, genomic databases, epidemiological systems and public health authorities. Machine learning and deep learning can assist with sequence classification, anomaly detection and prediction of evolutionary patterns, while natural language processing can integrate information from scientific publications, clinical reports and other epidemiological sources. The paper also examines challenges involving genomic data quality, sampling bias, model interpretability, privacy, cybersecurity, infrastructure, interoperability and equitable access to sequencing technologies. The study argues that AI should complement rather than replace laboratory science and epidemiological expertise. A future-ready genomic surveillance ecosystem will require continuous sequencing, responsible AI, interoperable data infrastructure, transparent governance and rapid translation of genomic signals into public health action.

Keywords: Genomic Surveillance, Artificial Intelligence, Infectious Diseases, Pathogen Genomics, Machine Learning, Epidemiology, Variant Detection, Outbreak Prediction, Public Health Intelligence, Bioinformatics.

1. Introduction

Emerging infectious diseases can spread rapidly across geographical boundaries.

Globalisation, urbanisation, climate change, ecological disruption, population mobility and interactions between humans and animals can create conditions in which pathogens emerge and spread.

The COVID-19 pandemic demonstrated the importance of genomic surveillance in understanding:

- Viral evolution.
- New variants.
- Transmission patterns.
- Geographic spread.
- Potential changes in pathogen behaviour.

However, sequencing alone is not sufficient.

Modern genomic surveillance produces enormous quantities of biological data that require computational interpretation.

Artificial intelligence can assist by identifying patterns across genomic and epidemiological datasets.

The emerging model can therefore be represented as:

Pathogen Sampling → Sequencing → Genomic Data → AI Analysis → Epidemiological Intelligence → Public Health Response

2. Genomic Surveillance

Genomic surveillance involves systematically sequencing pathogens and analysing their genetic characteristics over time.

It can provide information about:

- Genetic variation.
- Evolution.
- Transmission.
- Emergence of new lineages.
- Geographic distribution.

Unlike traditional surveillance, which often focuses on clinical symptoms and case counts, genomic surveillance provides information about the pathogen itself.

3. Importance for Emerging Infectious Diseases

Emerging pathogens may initially circulate undetected.

Early genomic signals can help identify unusual genetic patterns before conventional surveillance systems recognise a major outbreak.

Potential benefits include:

- Earlier detection.
- Variant identification.

- Transmission tracking.
- Outbreak investigation.
- Vaccine and therapeutic monitoring.

4. From Genomic Data to Public Health Intelligence

A modern surveillance system can be divided into several stages:

1. Sample collection.
2. Laboratory processing.
3. Genome sequencing.
4. Quality control.
5. Bioinformatic analysis.
6. AI-assisted interpretation.
7. Epidemiological integration.
8. Public health decision-making.

Each stage contributes to overall surveillance quality.

5. Artificial Intelligence in Genomic Surveillance

AI can process complex genomic datasets much faster than conventional manual analysis.

Potential applications include:

- Sequence classification.
- Variant detection.
- Anomaly detection.
- Phylogenetic analysis.
- Mutation interpretation.
- Outbreak forecasting.
- Automated genomic reporting.

6. Machine Learning for Pathogen Classification

Machine-learning models can classify genomic sequences according to known pathogen groups or lineages.

Potential inputs include:

- Nucleotide sequences.
- Amino-acid sequences.
- Mutation profiles.
- Genomic signatures.

The model can then estimate the most likely classification.

7. Deep Learning for Sequence Analysis

Deep learning can identify complex patterns within biological sequences.

Architectures such as:

- Convolutional neural networks.
- Recurrent neural networks.
- Transformers.

can potentially analyse genomic patterns at large scale.

Transformer-based approaches are particularly promising because they can model relationships between distant positions in biological sequences.

8. Variant Detection

One of the most important applications is identifying emerging variants.

An AI system can compare new sequences against:

- Reference genomes.
- Existing lineages.
- Historical sequence collections.

It can then identify unusual combinations of mutations.

9. Mutation Surveillance

Pathogens continuously accumulate genetic changes.

Most mutations may have little practical significance.

However, certain mutations can influence:

- Host interaction.
- Transmission.
- Immune recognition.
- Drug susceptibility.

AI can help prioritise mutations that warrant further investigation.

Importantly, computational predictions should be experimentally and epidemiologically validated before being interpreted as evidence of altered biological behaviour.

10. Anomaly Detection

AI-based anomaly detection can identify genomic sequences that differ substantially from expected patterns.

A simplified process is:

Normal Genomic Distribution → New Sequence → Statistical/AI Comparison → Anomaly Signal

This can help identify potentially important changes.

11. Phylogenetic Analysis

Phylogenetics reconstructs evolutionary relationships among pathogen sequences.

AI and computational methods can help analyse large genomic datasets to determine:

- Genetic relationships.

- Evolutionary clusters.
- Possible transmission patterns.
- Emergence of new lineages.

12. Transmission Monitoring

Combining genomic and epidemiological data can provide a more detailed picture of transmission.

For example:

Genetic Similarity + Geographic Data + Time + Epidemiological Information

can help investigators understand how a pathogen is spreading.

Genomic similarity alone does not prove direct transmission.

13. Outbreak Detection

An intelligent surveillance platform can monitor genomic data continuously.

Potential warning signals include:

- Rapid emergence of a new lineage.
- Unexpected geographic expansion.
- Unusual mutation combinations.
- Sudden increases in genomic diversity.

Such signals can trigger further laboratory and epidemiological investigation.

14. Early-Warning Systems

A future genomic early-warning system could integrate multiple data streams.

Biological Data

- Pathogen genomes.
- Genomic mutations.
- Lineage distributions.

Epidemiological Data

- Case counts.
- Hospitalisation.
- Geographic distribution.

Environmental Data

- Wastewater signals.
- Climate variables.
- Animal surveillance.

Scientific Intelligence

- Research publications.
- Laboratory reports.
- Public health alerts.

15. Wastewater Genomic Surveillance

Wastewater can provide population-level information about pathogen circulation.

Sequencing wastewater samples can potentially identify:

- Pathogen presence.
- Variant diversity.
- Emerging lineages.

AI can help analyse complex mixtures of genomic signals.

16. One Health Surveillance

Many emerging infectious diseases involve interactions between:

- Humans.
- Animals.
- Wildlife.
- Ecosystems.

A One Health genomic surveillance framework can connect information across these domains.

This is important because pathogen emergence may occur outside traditional human healthcare systems.

17. Zoonotic Disease Monitoring

AI-based genomic surveillance can support monitoring of pathogens circulating in animal populations.

Potential applications include surveillance of:

- Livestock.
- Wildlife.
- Domestic animals.
- Vector populations.

Early identification of concerning genetic changes could support preventive investigation.

18. AI and Epidemiological Data

Genomic information becomes more valuable when combined with epidemiological data.

Relevant variables may include:

- Location.
- Date.
- Age group.

- Clinical outcomes.
- Travel history.
- Vaccination status.

Data integration can support more comprehensive outbreak intelligence.

19. Natural Language Processing

Important information about emerging diseases is not contained only in genomic databases.

Relevant information may appear in:

- Scientific papers.
- Clinical reports.
- Public health announcements.
- Laboratory reports.
- News and surveillance reports.

Natural language processing can help identify and structure relevant information.

20. Multimodal Disease Intelligence

The future surveillance system may integrate:

Genomics + Epidemiology + Clinical Data + Environmental Data + Scientific Literature

AI can analyse these heterogeneous information sources to identify relationships that may not be obvious from individual datasets.

21. Real-Time Genomic Surveillance

Traditional genomic analysis can involve delays between:

Sample Collection → Sequencing → Analysis → Reporting

AI-supported automation can shorten this pipeline.

A future system could provide near-real-time genomic intelligence while maintaining appropriate quality-control procedures.

22. Cloud and Edge Computing

Large genomic datasets require substantial computational resources.

Cloud infrastructure can support:

- Large-scale storage.
- Model training.
- Cross-regional analysis.

Edge computing can support local processing in laboratories and field environments where rapid analysis or limited connectivity makes local computation valuable.

23. Federated Learning

Genomic surveillance frequently involves multiple institutions.

Federated learning could allow organisations to train models collaboratively while retaining sensitive datasets locally.

Conceptually:

Laboratory A → Local Model Training

Laboratory B → Local Model Training

Laboratory C → Local Model Training

↓

Shared AI Model

This can support collaboration while reducing the need for centralising raw data.

24. Genomic Data Quality

AI performance depends heavily on data quality.

Potential sources of error include:

- Sequencing errors.
- Contamination.
- Incomplete genomes.
- Incorrect metadata.
- Sampling problems.

Quality-control pipelines must therefore operate before AI interpretation.

25. Sampling Bias

Genomic surveillance does not automatically provide a representative picture of pathogen circulation.

Sequencing may be more common in:

- Large hospitals.
- Wealthier regions.
- Research institutions.

Underrepresented populations may therefore remain invisible.

AI models trained on biased datasets can reproduce these limitations.

26. Global Health Equity

Advanced genomic surveillance requires:

- Sequencing equipment.
- Skilled personnel.
- Computational infrastructure.
- Data-management systems.

Resource limitations can create significant disparities between countries and regions.

Future surveillance strategies should therefore include capacity-building and equitable access to genomic technologies.

27. Privacy and Governance

Genomic surveillance can involve sensitive information.

Depending on the surveillance system, datasets may contain:

- Patient information.
- Geographic information.
- Clinical metadata.

Strong governance should address:

- Data minimisation.
- Access control.
- Anonymisation or appropriate pseudonymisation.
- Secure storage.
- Responsible data sharing.

28. Cybersecurity

Genomic databases are increasingly valuable digital infrastructure.

Potential threats include:

- Data theft.
- Ransomware.
- Unauthorised access.
- Data manipulation.

Cybersecurity should therefore be integrated into surveillance architecture from the beginning.

29. Explainable AI

Public health decisions can have significant consequences.

AI systems should therefore provide interpretable evidence where feasible.

Rather than:

“High-Risk Variant”

the system should ideally provide:

- Relevant genomic features.
- Confidence estimates.
- Comparison with known lineages.
- Supporting epidemiological evidence.

30. Human Oversight

AI should function as a decision-support mechanism.

A robust architecture is:

AI Signal → Expert Review → Laboratory Validation → Epidemiological Assessment → Public Health Action

This reduces the risk of acting on false positives or computational artefacts.

31. Predictive Modelling

AI can potentially forecast:

- Variant prevalence.
- Geographic spread.
- Mutation trajectories.
- Outbreak growth.

However, predictions are uncertain and should be interpreted probabilistically rather than as deterministic forecasts.

32. Evolutionary Forecasting

Predicting pathogen evolution is particularly challenging.

Evolution is influenced by:

- Mutation.
- Selection.
- Population structure.
- Transmission dynamics.
- Immune pressure.
- Ecological factors.

AI can identify statistical patterns but cannot guarantee which future mutations will emerge.

33. Vaccine and Therapeutic Surveillance

Genomic surveillance can monitor changes relevant to:

- Vaccine effectiveness.
- Antibody recognition.
- Therapeutic susceptibility.

AI can help prioritise genomic changes for laboratory testing.

It should not be assumed that a predicted change automatically produces clinical resistance.

34. Antimicrobial Resistance Surveillance

Genomic analysis can identify genetic determinants associated with antimicrobial resistance.

AI can assist in:

- Detecting resistance-associated genes.
- Classifying resistance patterns.
- Monitoring their geographic distribution.
- Identifying emerging resistance trends.

35. Hospital Surveillance

Hospitals can use genomic surveillance to investigate potential transmission clusters.

Potential applications include:

- Outbreak investigation.
- Infection-control support.
- Transmission-network analysis.

Genomic evidence should be combined with patient movement and epidemiological information.

36. Proposed AI-Based Genomic Surveillance Framework

The proposed architecture contains eight major components:

1. Sample Collection

Clinical, environmental and animal samples.

2. Sequencing Infrastructure

Generates pathogen genomic data.

3. Quality-Control Pipeline

Removes unreliable or contaminated data.

4. Genomic AI Engine

Performs classification and anomaly detection.

5. Epidemiological Intelligence Layer

Integrates genomic and epidemiological data.

6. Risk Assessment Engine

Prioritises signals for expert investigation.

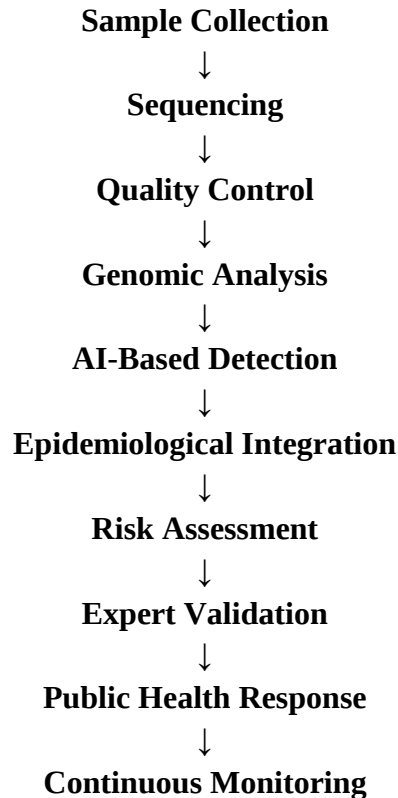
7. Public Health Dashboard

Provides interpretable surveillance information.

8. Response Coordination

Connects surveillance signals with appropriate public health processes.

37. Integrated Surveillance Architecture



38. AI-Based Variant Prioritisation

Not every newly detected mutation requires immediate action.

A prioritisation model can consider:

- Genetic novelty.
- Frequency increase.
- Geographic expansion.
- Epidemiological association.
- Functional evidence.
- Clinical evidence.

This can help focus scientific resources on the most relevant signals.

39. Surveillance Dashboard

A future surveillance platform could display:

- Newly detected lineages.
- Geographic distribution.
- Mutation frequency.
- Genomic clusters.
- Growth trends.
- Confidence estimates.

- Data-quality indicators.

The interface should clearly distinguish between:

Observed Evidence

And

AI-Generated Predictions

40. Interoperability

Genomic surveillance depends on data from multiple systems.

Interoperability is therefore essential across:

- Laboratories.
- Hospitals.
- Sequencing platforms.
- Public health agencies.
- Research institutions.

Standardised metadata and data-exchange mechanisms can improve integration.

41. Research Questions

1. How can AI improve early detection of emerging pathogen variants?
2. Which machine-learning architectures are most effective for genomic classification?
3. How can genomic and epidemiological data be integrated in real time?
4. How can sampling bias be identified and corrected?
5. Can federated learning support international genomic surveillance?
6. How can AI identify important mutations without excessive false positives?
7. How can wastewater and environmental genomics strengthen surveillance?
8. How can AI support antimicrobial-resistance monitoring?
9. How can genomic surveillance systems remain explainable and transparent?
10. What governance frameworks are needed for responsible AI-based genomic surveillance?

42. Future Research Directions

Future research should focus on:

- Foundation models for pathogen genomics.
- AI-assisted genomic anomaly detection.
- Multimodal outbreak intelligence.
- Federated genomic learning.
- Real-time wastewater genomics.
- One Health surveillance.
- Automated genomic quality control.
- AI-assisted antimicrobial-resistance surveillance.
- Secure international genomic data infrastructure.

- Explainable evolutionary modelling.

43. Strategic Recommendations

Public health and research organisations should:

1. Expand sequencing capacity.
2. Strengthen genomic data infrastructure.
3. Integrate genomic and epidemiological surveillance.
4. Establish robust data-quality standards.
5. Develop transparent AI models.
6. Maintain expert oversight.
7. Improve cybersecurity.
8. Address geographical and demographic sampling bias.
9. Invest in workforce training.
10. Build international genomic surveillance collaborations.

44. Discussion

AI-based genomic surveillance represents a transition from retrospective pathogen analysis toward continuous biological intelligence.

Traditional surveillance asks:

“Where is the disease spreading?”

Genomic surveillance adds:

“What is the pathogen doing genetically?”

AI adds another layer:

“Which genomic changes deserve attention, and what patterns are emerging?”

The integration of these capabilities can create a more responsive surveillance ecosystem.

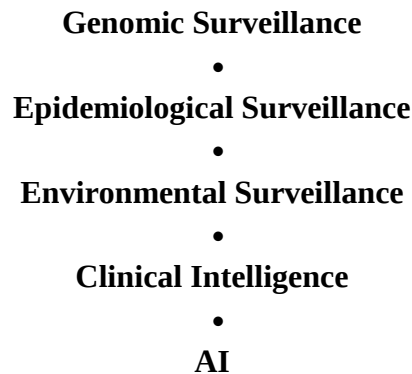
However, AI should not be treated as an autonomous pathogen-warning oracle.

The quality of the surveillance system remains dependent on:

- Sampling.
- Sequencing.
- Metadata.
- Laboratory validation.
- Epidemiological interpretation.
- Expert judgement.

45. Towards Predictive Pandemic Preparedness

Future preparedness systems may combine:



to create integrated early-warning systems.

Such systems could identify emerging signals earlier and support more targeted investigations.

46. From Detection to Response

The ultimate objective of genomic surveillance is not simply to generate genomic data.

Its value lies in translating information into appropriate action.

The complete pathway is:

Detection → Interpretation → Validation → Risk Assessment → Response

Reducing delays between these stages may substantially improve preparedness.

47. Conclusion

AI-based genomic surveillance provides a promising framework for strengthening preparedness against emerging infectious diseases.

By combining pathogen sequencing with machine learning, bioinformatics, epidemiological analysis, natural language processing and distributed data infrastructure, surveillance systems can move toward faster and more comprehensive detection of biological threats.

AI can support sequence classification, variant identification, anomaly detection, transmission analysis, antimicrobial-resistance surveillance and outbreak intelligence. However, computational predictions must be interpreted alongside laboratory and epidemiological evidence.

The future of genomic surveillance will depend on building systems that are not only technologically advanced but also:

- Reliable.
- Explainable.
- Secure.
- Interoperable.

- Equitable.
- Human-supervised.

The emerging paradigm can be summarised as:

Genomics + AI + Epidemiology + One Health → Intelligent Infectious-Disease Surveillance

With appropriate governance and investment, AI-enabled genomic surveillance could become an important component of global preparedness, helping public health systems detect, investigate and respond to emerging infectious threats more rapidly and effectively.

References

1. Gardy, J. L., & Loman, N. J. (2018). Towards a genomics-informed, real-time, global pathogen surveillance system. *Nature Reviews Genetics*, 19, 9–20.
2. Armstrong, G. L., MacCannell, D. R., Taylor, J., et al. (2019). Pathogen genomics in public health. *New England Journal of Medicine*, 381, 2569–2580.
3. Grubaugh, N. D., Ladner, J. T., Lemey, P., et al. (2019). Tracking virus outbreaks in the twenty-first century. *Nature Microbiology*, 4, 10–19.
4. Gardy, J. L., & Loman, N. J. (2018). Towards a genomics-informed, real-time, global pathogen surveillance system. *Nature Reviews Genetics*, 19, 9–20.
5. Khoury, M. J., Armstrong, G. L., Besser, J., et al. (2021). From public health genomics to precision public health: A 20-year journey. *Genetics in Medicine*, 23, 1–7.
6. Robishaw, J. D., et al. (2021). Genomic surveillance for emerging infectious diseases. *Nature Reviews Genetics*.
7. Lu, J., du Plessis, L., & others. (2020). Genomic epidemiology of SARS-CoV-2. *Nature*.
8. Hadfield, J., Megill, C., Bell, S. M., et al. (2018). Nextstrain: Real-time tracking of pathogen evolution. *Bioinformatics*, 34(23), 4121–4123.
9. Neher, R. A., & Bedford, T. (2018). Nextstrain: Real-time tracking of pathogen evolution. *Nature Methods*.
10. Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7. *Molecular Biology and Evolution*, 30(4), 772–780.
11. Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5), 1792–1797.
12. Suchard, M. A., Lemey, P., Baele, G., Ayres, D. L., Drummond, A. J., & Rambaut, A. (2018). Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evolution*, 4(1), vey016.
13. Rambaut, A., Holmes, E. C., O'Toole, Á., et al. (2020). A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nature Microbiology*, 5, 1403–1407.

14. Oude Munnink, B. B., Nieuwenhuijse, D. F., Stein, M., et al. (2020). Rapid SARS-CoV-2 whole-genome sequencing for informed outbreak response. *Nature Medicine*, 26, 1405–1410.
15. Tyson, J. R., James, P., Stoddart, D., et al. (2020). Improvements to the ARTIC network protocol enable rapid sequencing of SARS-CoV-2 genomes. *Nature Communications*, 11, 4859.