World Journal of Advanced Research and Reviews

eISSN: 2581-9615 CODEN (USA): WJARAI

Cross Ref DOI: 10.30574/wjarr

Journal homepage: https://wjarr.com/

(Review Article)

Check for updates

PREDICTIVE CHRONIC DISEASE SURVEILLANCE FRAMEWORK: INTEGRATING MACHINE LEARNING, REAL-TIME DATA, AND PUBLIC HEALTH RESPONSE

Lord Amoateng Adjei [1], Franklin Adjei [2,*], Justice Manu [3], Kelvin McKeown Antwi Adjei [4], Augustine Afriyie [5] and Bernard Kwame Frempong [6]

[1] University of Energy and Natural Resources, Sunyani, Ghana.
[2] Division of Kinesiology and Health, College of Health Sciences, University of Wyoming, Laramie, WY, USA.
[3] Washington State Department of Ecology, Washington, USA.
[4] Kwame Nkrumah University of Science and Technology, Kumasi, Ghana.
[5] College of Basic and Applied Sciences, Middle Tennessee State University, Murfreesboro, TN, USA.
[6] Department of Biomedical Sciences, School of Medicine & Health Sciences, University of North Dakota, Grand Forks, ND, USA.
* Corresponding Author; Email: lordadjei509@gmail.com

ORCID Details
Lord Amoateng Adjei: https://orcid.org/0009-0009-1679-8575
Franklin Adjei: https://orcid.org/0009-0002-8158-1312
Justice Manu: https://orcid.org/0009-0003-2219-0554
Kelvin McKeown Antwi Adjei: https://orcid.org/0009-0006-0676-5006
Augustine Afriyie: https://orcid.org/0009-0008-1304-0403
Bernard Kwame Frempong: https://orcid.org/0009-0006-0676-5006

World Journal of Advanced Research and Reviews, 2026, 31(03), 772–782

Publication history: Received on 31 July 2026; revised on 05 September 2026; accepted on 08 September 2026

Article DOI: https://doi.org/10.30574/wjarr.2026.31.3.2320

ABSTRACT

Chronic disease surveillance is an essential part of public health because it helps governments and health organizations understand the burden of disease, identify risk factors, recognize disparities, and track changes in health outcomes. Most existing surveillance systems, however, are mainly concerned with describing what has already happened. The growing use of electronic health records, wearable devices, environmental sensors, administrative databases, and other continuously generated sources of information creates an opportunity to move surveillance toward a more forward-looking approach.

This article presents the Predictive Chronic Disease Surveillance Framework (PCDSF), a framework for using multiple data sources and machine learning to estimate changing population-level risk and to connect those estimates to public health action. The framework treats predictive surveillance as a continuous cycle: data are collected and combined, population risk is estimated, emerging patterns are identified, decisions are made, interventions are implemented, and the resulting outcomes are fed back into the system. This makes surveillance an ongoing learning process rather than a one-way reporting exercise.

The framework also makes an important distinction between predicting disease in an individual patient and predicting changes in risk across populations. Its purpose is not simply to identify who is likely to become ill, but to help public health organizations understand where risk is increasing, which groups may be most affected, and where preventive resources could have the greatest impact. Six propositions are developed around multimodal data integration, timely updating, predictive performance, intervention, equity, and continuous learning. The article also discusses privacy, interoperability, data quality, algorithmic bias, transparency, and organizational capacity as conditions that can determine whether predictive surveillance works in practice.

Keywords: Chronic Disease Surveillance; Machine Learning; Predictive Analytics; Artificial Intelligence; Digital Health; Population Health; Public Health Surveillance; Risk Prediction

1 INTRODUCTION

Noncommunicable diseases (NCDs), including cardiovascular disease, cancer, chronic respiratory diseases, and diabetes, continue to place a major burden on health systems and communities (Freihat et al., 2025; Ndubuisi, 2021). Their effects are seen not only in mortality, but also in disability, healthcare use, and economic costs (García-Morales et al., 2024). Preventing and managing these conditions requires more than knowing how many people are currently affected. Health systems also need to understand how risk is distributed, how it is changing, and which communities may need attention before the burden becomes more severe.

This is why public health surveillance remains central to chronic disease prevention. Surveillance-related activities are the systematic collection, analysis, sharing, and use of information to support public health policy and decision-making (Shah & Househ, 2024). The effective management of chronic disease also depends on reliable longitudinal data to support decisions at the facility, regional, and national levels (Hisashige, 2013). In the United States, the Centers for Disease Control and Prevention (CDC) operates several surveillance systems that track chronic diseases, behavioral risk factors, health indicators, and population trends (B. Lee et al., 2018).

These systems are extremely valuable, but much of conventional surveillance remains descriptive. In simple terms, it tells us what is happening, where it is happening, how common it is, and which populations are carrying the greatest burden. Those questions are still necessary. The limitation is that they are often most informative only after a change in disease burden has already become visible.

Digital health technologies present an opportunity to ask an additional question about what is likely to transpire next. For example, if surveillance shows that diabetes prevalence has already risen sharply in a community, health authorities can respond to an established problem. A predictive system could instead examine combinations of behavioral, clinical, environmental, and social indicators to identify communities whose risk appears to be increasing before the full increase in disease burden is observed. A similar approach could be used to identify changes in cardiovascular risk before they result in noticeable increases in hospitalizations or mortality.

The data needed for this type of surveillance are increasingly available. Electronic health records provide prospective or continuous clinical information (Torab-Miandoab et al., 2025). Wearable devices can generate frequent physiological and behavioral measurements. Environmental and geographic systems can describe pollution, temperature, neighborhood conditions, and the built environment (Brazil, 2022). Administrative systems provide information about healthcare use, while population surveys contribute behavioral and socioeconomic information. Wearable technologies, in particular, are increasingly used to monitor factors associated with cardiovascular, metabolic, respiratory, neurological, and other chronic conditions (Jafleh et al., 2024).

Machine learning adds another capability whereby these methods can work with large and varied datasets and can detect relationships that may be difficult to identify using simple, predefined indicators (Dritsas & Trigka, 2025; Lamichhane et al., 2025). Instead of analyzing each indicator individually at fixed intervals, predictive models can integrate multiple variables, assess their relationships to future outcomes, and revise these assessments as new data emerge (T. C. Lee et al., 2020). However, having more data or a more sophisticated algorithm does not automatically produce better public health surveillance. Much of the existing AI and machine-learning literature focuses on individual clinical prediction, for example, estimating whether a patient will develop a disease, experience a complication, be hospitalized, or respond to treatment. These applications are important, but public health surveillance has a broader focus and that is the health of populations.

An accurate model that predicts an individual's probability of developing diabetes is therefore not, by itself, a predictive surveillance system. For prediction to become surveillance, individual-level information needs to be transformed into an understanding of how risk is distributed across populations, locations, and time. More importantly, that information must be connected to organizations and processes capable of acting on it. This matters because prediction without action has limited public health value. An algorithm might identify a community where cardiovascular risk is rising, but that signal alone will not improve health conditions. Public health organizations still need to decide whether the signal is meaningful, determine an appropriate response, allocate resources, implement the response, and evaluate the outcome. Predictive surveillance should therefore be understood as a socio-technical system rather than simply a forecasting tool. It brings together data, algorithms, public health professionals, interventions, communities, and governance.

There are also important risks. AI systems can reproduce existing inequalities when their training data poorly represent certain populations or reflect historical differences in access to healthcare (Norori et al., 2021). Combining clinical, behavioral, environmental, geographic, and consumer-generated information can create additional privacy and governance challenges (Panteli et al., 2025). Model performance may also decline as populations, technologies, healthcare practices, or environmental conditions change (Sahiner et al., 2023). A model can therefore be statistically strong while still being difficult to interpret or used responsibly.

These challenges point to the need for a framework that connects machine learning and continuously generated data with the established goals of public health surveillance. The Predictive Chronic Disease Surveillance Framework (PCDSF) proposed in this article does this through six connected functions: multimodal data integration, dynamic risk estimation, population risk stratification, public health decision support, targeted intervention, and feedback-based learning.

The central argument is that machine learning creates public health value not simply by making accurate predictions, but by embedding those predictions into a continuous surveillance-to-action-to-learning cycle. The important question is therefore not only whether chronic disease can be predicted, but how predictive information can help public health systems stay proactive, target resources more effectively, and learn from the results.

2 THEORETICAL FOUNDATIONS

2.1 From descriptive surveillance to predictive surveillance

Public health surveillance has traditionally provided the information infrastructure needed to detect health problems, monitor trends, identify disparities, inform policy, allocate resources, and evaluate interventions (Soucie, 2012a). Chronic disease surveillance draws on sources such as surveys, disease registries, vital statistics, healthcare records, and administrative databases (Chow & Leo, 2017).

The CDC's chronic disease surveillance infrastructure illustrates this approach. Systems such as the Behavioral Risk Factor Surveillance System and Chronic Disease Indicators help public health authorities examine disease burden, behavioral risks, population trends, and program outcomes (Hacker et al., 2024). Predictive surveillance is not intended to replace these functions. Instead, it adds a forward-looking layer.

Three broad analytical perspectives can therefore be distinguished. Descriptive surveillance asks what is happening or what has happened. Explanatory and epidemiological analysis examines factors associated with those patterns. Predictive surveillance asks what population health state is likely to occur next, given what is known about the present and the past. This changes the time perspective of surveillance. A conventional indicator might show that obesity has increased in a particular area over five years. A predictive system could combine recent trends in obesity, behavior, healthcare access, socioeconomic conditions, and other relevant factors to estimate which populations may experience greater obesity-related disease burden in the future.

Such predictions should never be interpreted as certainty. Their purpose is to reduce uncertainty enough to give public health organizations more time and better information for preventive action.

2.2 Population health as the level of inference

One of the most important boundaries in the PCDSF is the difference between clinical prediction and population surveillance. Clinical prediction generally focuses on an individual and estimates the probability of a particular outcome within a defined period (Iwagami & Matsui, 2022). These models can support diagnosis, prognosis, treatment, and clinical decision-making.

Predictive surveillance requires an additional step. Individual observations are used to identify patterns in population risk. Predictions can be aggregated, mapped geographically, compared across demographic or socioeconomic groups, and examined over time to identify changes in the distribution of risk. The useful surveillance output is not simply that a particular patient has a high probability of developing a condition. It might instead show that a particular population in a particular area is experiencing a rapidly increasing concentration of risk and is projected to face a higher disease burden during a specified period. That kind of information has direct implications for public health planning because it identifies the population, location, direction of risk, and possible window for intervention. This population focus also changes how model performance should be judged. Individual-level measures such as discrimination remain important, but a surveillance system must also demonstrate that its predictions identify meaningful population patterns and improve decisions about resources or interventions.

2.3 Machine learning as an analytical mechanism

Machine learning is one possible way to implement predictive surveillance. It can handle large numbers of variables, model complex interactions, identify nonlinear relationships, and produce predictions from datasets that may be difficult to analyze using conventional methods (Aghaabbasi & Chalermpong, 2023).

Even so, machine learning is only one part of the surveillance architecture. An accurate algorithm does not automatically produce an effective surveillance system. Predictions depend on the quality and completeness of the training data, which populations are represented, how stable the relationships between predictors and outcomes are, and how the model is developed, validated, calibrated, and updated.

These issues are increasingly recognized in prediction-model research. TRIPOD+AI, for example, emphasizes transparent reporting of the development and evaluation of prediction models that use regression or machine-learning methods. (Collins et al., 2024) For surveillance, these requirements need to be combined with continuous monitoring, as the populations and environments that generate health data are constantly changing. The changes in healthcare access, clinical practice, population composition, environmental exposures, technology use, and data collection can change the relationship between predictors and outcomes.

2.4 The learning-system perspective

The final theoretical foundation is feedback. A simple data-driven process might be represented as data, analysis, decision, and intervention. The PCDSF instead treats surveillance as a cycle whereby data is collected, risk is predicted, population risk is assessed, decisions are made, interventions are implemented, outcomes are observed, and the new information is used to evaluate and update the system.

The reason for this cycle is straightforward in the sense that interventions change the conditions that the model is trying to predict. For example, suppose a system identifies neighborhoods where hypertension risk is rising, and public health agencies respond by expanding screening, community health-worker services, and preventive care. If those interventions work, future outcomes may be different from what the original model predicted. A system that ignores those changes may gradually become poorly calibrated. Predictive surveillance therefore needs to learn not only from new observations, but also from the effects of the actions that its predictions helped initiate. This feedback turns surveillance from a passive reporting system into an adaptive population health system.

3 THE PREDICTIVE CHRONIC DISEASE SURVEILLANCE FRAMEWORK

The PCDSF organizes predictive surveillance into six connected stages within a wider governance environment, and the process can be summarized as follows. A structured method for population data involves collecting information from multiple sources, preparing and validating the data, and using machine learning to identify and assess potential risks. This procedure enables dynamic risk categorization, supports public health initiatives, and facilitates targeted preventive measures. Ultimately, it provides insights into population outcomes and intervention impacts, offering valuable feedback to enhance data quality, models, and decision-making protocols at each stage.

Across all stages, privacy and security, interoperability, transparency, equity, bias monitoring, accountability, human oversight, and organizational capacity remain relevant. The framework is deliberately cyclical. The outcomes from both the population and the interventions return to the surveillance system so that predictions and decisions can be evaluated and improved.

3.1 Multimodal population data

The framework begins with a simple premise that chronic disease risk is shaped by many factors. Biology, behavior, healthcare access, social conditions, and the physical environment interact in ways that cannot always be captured by a single dataset (Eriksson et al., 2025).

The proposed data layer can therefore include electronic health records, disease registries, insurance or claims data, laboratory and pharmacy records, mortality data, behavioral surveillance, wearable devices, environmental sensors, geospatial information, census data, socioeconomic indicators, and other relevant sources. Each source contributes something different. Electronic health records provide clinical detail but mainly cover people who interact with healthcare systems. Surveys can provide information about behavior and social conditions but are often conducted periodically. Wearables can provide frequent measurements but may be more common among people who have access to certain technologies. Environmental and geographic data can describe exposures and neighborhood conditions, but they must be carefully linked to health outcomes.

Combining these sources can expand the range of information available for detecting emerging risk and reduce dependence on any one source. However, integration itself creates challenges. Poorly combined data can introduce duplication, inconsistent measurements, missing information, timing problems, and selection bias (Al-Sahab et al., 2023). For this reason, data integration in the PCDSF includes quality management. Before information is used for prediction, its source, completeness, timing, representativeness, and measurement consistency should be assessed.

3.2 Machine-learning risk estimation

The second stage converts integrated information into estimates of future chronic disease risk. The framework does not prescribe one particular algorithm. Logistic and survival models, decision trees, ensemble methods, neural networks, time-series approaches, and other techniques may all be appropriate depending on the problem, data, interpretability requirements, and intended use.

What matters most is that the prediction is clearly defined in time. A model designed to estimate five-year diabetes incidence addresses a different surveillance problem than one designed to estimate thirty-day cardiovascular hospitalization. The data, model, thresholds, and possible interventions will therefore differ. Prediction should also communicate uncertainty. Public health decisions often involve scarce resources (Ehmann et al., 2021), so decision-makers need to understand how reliable an estimate is. Probabilities and uncertainty intervals, where appropriate, can help prevent model outputs from being mistaken for facts about the future.

Evaluation should go beyond discrimination. Calibration is particularly important because consistent overprediction or underprediction in particular populations could result in resources being directed away from where they are actually needed.

3.3 Dynamic population risk stratification

Risk estimation becomes surveillance when individual predictions are turned into meaningful population patterns. The PCDSF therefore places population risk stratification between prediction and decision-making. The predicted risks can be grouped by geographic area, demographic characteristics, socioeconomic conditions, disease-risk profiles, or other units that matter for policy. The key is that stratification should be dynamic. The current level of risk matters, but so does the direction in which that risk is moving. Imagine two communities with the same predicted diabetes risk today. In one, risk has remained relatively stable for several years. In the other, risk has risen rapidly over the last six months. Treating them as identical could hide an important emerging problem.

The framework therefore considers three dimensions together: risk magnitude, risk distribution, and risk trajectory. The magnitude describes the expected level of burden. Distribution shows which groups are experiencing greater risk. Trajectory shows whether risk is rising, falling, or remaining stable. Together, these dimensions provide a more useful population risk profile than a single prediction score.

3.4 Public health decision support

A prediction has little practical value if decision-makers cannot understand or use it. The fourth stage of the framework therefore connects population risk profiles to public health decision-making. Depending on the setting, decision support could include dashboards, geographic hotspot maps, threshold alerts, scenario projections, resource-allocation tools, or periodic population-risk reports. The goal is not simply to display complicated model outputs, but to turn them into information that public health professionals can use.

Agencies also need processes for deciding what different signals mean and what action, if any, they justify. For instance, a rapid increase in cardiovascular risk might first trigger closer monitoring. A persistent high-risk pattern could support targeted screening, while high predicted burden combined with limited healthcare access might influence where community health workers or mobile clinics are deployed. Human judgment remains essential. WHO guidance emphasizes autonomy, transparency, accountability, inclusiveness, equity, and ongoing assessment when AI is used in health (WHO, 2021). Predictive systems should therefore support professional judgment rather than attempt to replace it.

3.5 Targeted preventive response

The fifth stage is where surveillance information is translated into action. The appropriate response will depend on the disease, the affected population, the risk factors involved, and the authority of the responding organization.

Possible responses include targeted screening, preventive services, health communication, community programs, environmental measures, changes in service capacity, and redistribution of public health resources. The framework does not suggest that every predicted increase in risk should automatically lead to intervention. False alarms can waste resources, and interventions themselves have costs and potential harms. Decisions should therefore consider the predicted burden, uncertainty, likely effectiveness of the intervention, available resources, possible harms, and equity. This leads to an important point where actionability connects predictive performance with public health value. A highly accurate prediction may have limited practical value if nothing can be done about the risk it identifies. In contrast, a moderately accurate prediction may be highly useful when it is connected to an intervention that is feasible, effective, and timely.

Evaluation should therefore consider the entire pathway from prediction to response, not just the algorithm's technical performance.

3.6 Feedback and adaptive learning

The final stage closes the surveillance loop. After an intervention is implemented, subsequent population outcomes should return to the system. These outcomes serve several purposes. First, they show whether the original predictions were accurate. Second, they provide evidence about whether the intervention changed the expected trajectory. Thirdly, they provide new information that can help determine whether the model remains calibrated as conditions change. This is important because predictive models can lose performance over time (Ivanescu et al., 2016). Changes in population composition, healthcare practice, disease patterns, data collection, and intervention environments can make relationships learned from historical data less useful (Nilsen et al., 2020).

Regulatory thinking around machine-learning-enabled health technologies also recognizes the need to monitor models after deployment and manage changes associated with retraining and performance (Rosenthal et al., 2025). Although public health surveillance systems are different from medical devices, the broader lesson is relevant: model performance should never be assumed to remain constant. The PCDSF therefore treats predictive surveillance as an adaptive learning process in which prediction, intervention, and evaluation continuously influence one another.

4 GOVERNANCE CONDITIONS

The six stages of the framework operate within a governance environment. Governance is not something that happens only at the end of the process but it influences every stage.

4.1 Privacy and data governance

Multimodal surveillance can require information from systems containing sensitive health, behavioral, socioeconomic, and geographic data (Soucie, 2012b). The benefits of linking these sources must be balanced against privacy risks. Data minimization should be part of system design. Information should not be collected simply because it is technically available. Each variable should have a clear connection to the surveillance objective and to the decisions the system is intended to support. Governance arrangements should also clearly define who can access information, why they can access it, how long data is retained, and how secondary uses are controlled.

4.2 Equity and algorithmic bias

Predictive surveillance has the potential to expose health inequalities, but it can also reproduce them, thereby perpetuating them (Gupta et al., 2025). Historical health records may reflect unequal access to healthcare, differences in diagnosis, underrepresentation in research, structural inequalities, and gaps in data collection (Riley, 2012). An algorithm trained on those records can learn patterns that partly reflect an unequal system rather than underlying health need. Equity therefore needs to be considered throughout the surveillance process. Data representativeness should be assessed before modeling. Model performance should be compared across relevant population groups, and the consequences of resource allocation should also be examined. Digital data create another challenge. Wearables and other connected technologies can make continuous monitoring possible, but access to these technologies is not universal (Chen et al., 2025). If a system relies heavily on such data without accounting for unequal access, some populations may become much more visible to the surveillance system than others.

4.3 Transparency and accountability

Public health decision-makers need enough information about a predictive system to understand both its strengths and its limitations. Transparency does not mean that every user must understand the mathematics behind an algorithm. It does mean that the system's purpose, target population, input data, performance, limitations, and appropriate uses should be documented. TRIPOD+AI for example, provides a useful foundation for transparent reporting of prediction-model development and evaluation (Collins et al., 2024). In an operational surveillance system, transparency should also cover model updates and the decisions influenced by predictive outputs.

Ultimately, accountability must remain with people and institutions. Algorithms produce estimates; authorized public health professionals and organizations decide what those estimates mean and how they affect communities.

5 RESEARCH PROPOSITIONS

The framework yields six propositions that can be tested in future empirical research.

- Proposition 1: Multimodal integration and predictive capacity: Chronic disease surveillance systems that combine complementary clinical, behavioral, environmental, and social data should be better able to identify emerging population-level risk patterns than systems relying on a single data category, provided that the information is reliable, representative, and meaningfully integrated. This proposition follows from the multidimensional nature of chronic disease risk. Importantly, simply adding more datasets is not enough; additional information is useful only when its quality and integration are adequate.
- Proposition 2: Temporal updating and early detection: Surveillance systems that use more frequently updated data should identify meaningful changes in population chronic disease risk earlier than systems that depend mainly on periodic retrospective data. This can be tested by comparing how long different surveillance approaches take to detect a meaningful change in risk.
- Proposition 3: Dynamic stratification and decision usefulness: Predictions that account for risk magnitude, population distribution, and risk trajectory should provide greater value for public health resource allocation than predictions based solely on the current risk level. The reasoning is that the same risk level can have very different implications depending on whether it is stable, increasing, or declining.
- Proposition 4: Actionability as a mechanism linking prediction to outcomes: The relationship between predictive performance and improved population health should be stronger when predictions are linked to clearly defined, feasible public health interventions. This proposition reflects a central idea of the framework: prediction is a capability that supports action, not the final outcome itself.
- Proposition 5: Equity governance and distributional performance: Predictive surveillance systems that assess data representativeness, subgroup performance, and the distribution of interventions throughout development and deployment should produce more equitable public health decisions than systems that consider equity only after deployment. This proposition treats equity as an ongoing requirement rather than a final check.
- Proposition 6: Feedback and sustained predictive performance: Predictive surveillance systems that systematically incorporate population and intervention outcomes should maintain calibration and decision usefulness over time more effectively than systems whose models remain unchanged after deployment.

Together, these propositions move research beyond the narrow question of whether machine learning can predict chronic disease. They instead ask when predictive systems actually improve surveillance and how that improvement translates into public health value.

6 IMPLICATIONS FOR RESEARCH, PRACTICE, AND POLICY

6.1 Research implications

The PCDSF provides several opportunities for future research. Researchers can test specific relationships within the framework, such as whether combining different data sources improves early detection or whether including risk trajectories changes resource-allocation decisions. Research should also distinguish technical performance from decision performance. Sensitivity, specificity, discrimination, calibration, and prediction error remain important, but they should be complemented by questions about whether predictions changed decisions, whether resources were redirected, whether interventions happened earlier, and whether population outcomes improved. External validation is another important area. A model developed in one health system or region may perform differently in another(Yang et al., 2024). This matters even more when predictions influence population-level allocation of resources.

Finally, implementation research is needed. A technically strong system may still fail if an organization lacks suitable data infrastructure, trained personnel, analytical capacity, or authority to respond to the predictions (Sedlakova et al., 2023). Organizational readiness should therefore be treated as an important component of predictive surveillance research.

6.2 Implications for public health practice

For practitioners, the framework suggests that investing in algorithms without strengthening the underlying surveillance infrastructure is unlikely to deliver the expected benefits.

Organizations considering predictive surveillance need a reliable data architecture, interoperability, analytical capabilities, clear decision-making protocols, intervention capacity, and appropriate governance. The most useful question is not simply which machine-learning model to use. A better starting point is: which public health decision are we trying to improve, what information would improve that decision, and what action becomes possible if risk can be identified earlier?

This perspective can also prevent unnecessary use of AI. In some settings, conventional statistical approaches may provide all the predictive information needed. The PCDSF does not assume that a more complicated algorithm is automatically better. Complexity is justified only when it meaningfully improves surveillance or decision-making.

6.3 Policy implications

Predictive surveillance also raises important policy questions. As surveillance systems increasingly combine information across institutions, standards for interoperability and responsible data sharing will become more important. Policies should also establish expectations for validation, transparency, monitoring, accountability, and human oversight.

Good Machine Learning Practice principles illustrate the broader move toward lifecycle governance of machine-learning-enabled health technologies, including attention to real-world performance and human-AI interaction. Similar principles can inform public health surveillance even when a particular system is outside medical-device regulation. Policy must also guard against predictive systems worsening existing inequalities. Identifying where future disease burden is likely to increase can help target resources, but only if allocation decisions take population needs and equity into account. Otherwise, communities with stronger digital infrastructure may generate more data and receive more algorithmic attention, while communities with greater needs but less digital visibility may be overlooked.

Limitations and Future Directions

The PCDSF is a conceptual framework, so its propositions still require empirical testing. The strength and relevance of the relationships described here are likely to vary across diseases, populations, health systems, and technological settings. Data availability is another limitation. Although the framework proposes integrating multiple sources, many public health agencies do not currently have access to all of them. Data systems can be fragmented, interoperability may be limited, and important information may be held by different public or private organizations (von Arnim et al., 2026). Predictive accuracy also cannot remove uncertainty. Chronic disease trajectories are influenced by interacting social, behavioral, environmental, biological, and policy factors (Cockerham et al., 2017; Rahelić et al., 2024). Predictions should therefore inform professional judgment rather than be treated as fixed forecasts.

The framework also does not prescribe a single machine-learning method. The appropriate approach depends on the surveillance objective, available data, interpretability requirements, and implementation setting. Future research can investigate which types of predictive methods work best for particular surveillance tasks.

The growing use of continuously generated digital data raises a further concern: digital inequality. People and communities differ in access to healthcare, broadband, smartphones, wearable devices, and other technologies that generate health information. Future research should examine how new data sources can be incorporated without systematically underrepresenting populations that are less digitally connected.

Future studies should test the PCDSF across different chronic diseases and institutional settings. Prospective studies are especially important because the proposed benefits of predictive surveillance depend on early detection, intervention, feedback, and model adaptation over time. Comparative studies could also examine whether predictive surveillance is more useful for conditions where effective preventive interventions already exist than for conditions where available responses are limited.

Finally, future work should examine how predictive information changes organizational behavior. The success of predictive surveillance ultimately depends not only on what an algorithm identifies, but also on whether institutions trust the information, understand its limitations, and respond to it appropriately.

7 CONCLUSION

The growing availability of continuous digital health data and advances in machine learning create an opportunity to move chronic disease surveillance beyond simply describing what has already happened. They enable a more forward-looking approach, allowing public health organizations to identify changing population risk patterns earlier and consider preventive action sooner. Achieving this shift, however, requires more than adding a prediction algorithm to an existing data system. The Predictive Chronic Disease Surveillance Framework brings together multimodal data, machine-learning risk estimation, dynamic population stratification, public health decision-making, targeted intervention, and feedback-based learning in one continuous cycle.

The framework's main argument is that predictive accuracy is only the beginning. Prediction becomes valuable to public health when it is actionable, equitable, properly governed, and connected to real interventions and subsequent learning. The framework integrates machine learning into the broader goals of public health surveillance, shifting the focus from simply predicting chronic disease to helping health systems intervene earlier, allocate resources more effectively, and continuously improve decisions based on observed outcomes.

STATEMENTS ON COMPLIANCE WITH ETHICAL STANDARDS

Acknowledgments

The authors acknowledge the researchers and scholars whose published work provided the foundation for this study and informed the perspectives presented in this manuscript.

Disclosure of conflict of interest

The authors declare no competing interests.

Data Availability

No original datasets were generated or analyzed for this theoretical manuscript. All evidence informing the framework was derived from published literature and publicly available sources.

REFERENCES

[1] Aghaabbasi, M., & Chalermpong, S. (2023). Machine learning techniques for evaluating the nonlinear link between built-environment characteristics and travel behaviors: A systematic review. *Travel Behaviour and Society*, *33*, 100640. https://doi.org/10.1016/j.tbs.2023.100640

[2] Al-Sahab, B., Leviton, A., Loddenkemper, T., Paneth, N., & Zhang, B. (2023). Biases in Electronic Health Records Data for Generating Real-World Evidence: An Overview. *Journal of Healthcare Informatics Research*, *8*(1), 121–139. https://doi.org/10.1007/s41666-023-00153-2

[3] Brazil, N. (2022). Environmental inequality in the neighborhood networks of urban mobility in US cities. *Proceedings of the National Academy of Sciences*, *119*(17), e2117776119. https://doi.org/10.1073/pnas.2117776119

[4] Chen, J., Jastrzebska-Perfect, P., Chai, P., Girayhan Say, M., Tu, J., Gao, W., Halperin, F., Korzenik, J., Huang, H.-W., Katabi, D., & Traverso, G. (2025). Barriers to translating continuous monitoring technologies for preventative medicine. *Nature Biomedical Engineering*, *9*(11), 1797–1815. https://doi.org/10.1038/s41551-025-01520-7

[5] Chow, A., & Leo, Y.-S. (2017). Surveillance of Disease: Overview. *International Encyclopedia of Public Health*, 124–138. https://doi.org/10.1016/B978-0-12-803678-5.00439-2

[6] Cockerham, W. C., Hamby, B. W., & Oates, G. R. (2017). The Social Determinants of Chronic Disease. *American Journal of Preventive Medicine*, *52*(1 Suppl 1), S5–S12. https://doi.org/10.1016/j.amepre.2016.09.010

[7] Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., Ghassemi, M., Liu, X., Reitsma, J. B., van Smeden, M., Boulesteix, A.-L., Camaradou, J. C., Celi, L. A., Denaxas, S., Denniston, A. K., Glocker, B., Golub, R. M., Harvey, H., Heinze, G., … Logullo, P. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *The BMJ*, *385*, e078378. https://doi.org/10.1136/bmj-2023-078378

[8] Dritsas, E., & Trigka, M. (2025). Exploring the Intersection of Machine Learning and Big Data: A Survey. *Machine Learning and Knowledge Extraction*, *7*(1), 13. https://doi.org/10.3390/make7010013

[9] Ehmann, M. R., Zink, E. K., Levin, A. B., Suarez, J. I., Belcher, H. M. E., Daugherty Biddison, E. L., Doberman, D. J., D'Souza, K., Fine, D. M., Garibaldi, B. T., Gehrie, E. A., Golden, S. H., Gurses, A. P., Hill, P. M., Hughes, M. T., Kahn, J. P., Koch, C. G., Marx, J. J., Meisenberg, B. R., … Kachalia, A. (2021). Operational Recommendations for Scarce Resource Allocation in a Public Health Crisis. *Chest*, *159*(3), 1076–1083. https://doi.org/10.1016/j.chest.2020.09.246

[10] Eriksson, M., Sundberg, L. R., Santosa, A., Lindgren, H., Ng, N., & Lindvall, K. (2025). Health behavioural change – the influence of social-ecological factors and health identity. *International Journal of Qualitative Studies on Health and Well-Being*, *20*(1), 2458309. https://doi.org/10.1080/17482631.2025.2458309

[11] Freihat, O., Sipos, D., Aamir, M., & Kovacs, A. (2025). Global burden and future projections of non-communicable diseases (2000–2050): Progress toward SDG 3.4 and disparities across regions and risk factors. *PLOS One*, *20*(12), e0336036. https://doi.org/10.1371/journal.pone.0336036

[12] García-Morales, C., Heredia-Pi, I., Guerrero-López, C. M., Orozco, E., Ojeda-Arroyo, E., Nigenda, G., & Serván-Mori, E. (2024). Social and economic impacts of non-communicable diseases by gender and its correlates: A literature review. *International Journal for Equity in Health*, *23*, 274. https://doi.org/10.1186/s12939-024-02348-4

[13] Gupta, R., Sasaki, M., Taylor, S. L., Fan, S., Hoch, J. S., Zhang, Y., Crase, M., Tancredi, D., Adams, J. Y., & Ton, H. (2025). Developing and Applying the BE-FAIR Equity Framework to a Population Health Predictive Model: A Retrospective Observational Cohort Study. *Journal of General Internal Medicine*, *40*(11), 2537–2547. https://doi.org/10.1007/s11606-025-09462-1

[14] Hacker, K., Thomas, C. W., Zhao, G., Claxton, J. S., Eke, P., & Town, M. (2024). Social Determinants of Health and Health-Related Social Needs Among Adults With Chronic Diseases in the United States, Behavioral Risk Factor Surveillance System, 2022. *Preventing Chronic Disease*, *21*, E94. https://doi.org/10.5888/pcd21.240362

[15] Hisashige, A. (2013). The Effectiveness and Efficiency of Disease Management Programs for Patients with Chronic Diseases. *Global Journal of Health Science*, *5*(2), 27–48. https://doi.org/10.5539/gjhs.v5n2p27

[16] Ivanescu, A. E., Li, P., George, B., Brown, A. W., Keith, S. W., Raju, D., & Allison, D. B. (2016). The Importance of Prediction Model Validation and Assessment in Obesity and Nutrition Research. *International Journal of Obesity (2005)*, *40*(6), 887–894. https://doi.org/10.1038/ijo.2015.214

[17] Iwagami, M., & Matsui, H. (2022). Introduction to Clinical Prediction Models. *Annals of Clinical Epidemiology*, *4*(3), 72–80. https://doi.org/10.37737/ace.22010

[18] Jafleh, E. A., Alnaqbi, F. A., Almaeeni, H. A., Faqeeh, S., Alzaabi, M. A., & Al Zaman, K. (2024). The Role of Wearable Devices in Chronic Disease Monitoring and Patient Care: A Comprehensive Review. *Cureus*, *16*(9), e68921. https://doi.org/10.7759/cureus.68921

[19] Lamichhane, B. R., Isnan, M., & Horanont, T. (2025). Exploring machine learning trends in poverty mapping: A review and meta-analysis. *Science of Remote Sensing*, *11*, 100200. https://doi.org/10.1016/j.srs.2025.100200

[20] Lee, B., Martin, T., Khan, A., Fullerton, K., Duck, W., Kinley, T., Stoutenburg, S., Hall, J., Crum, M., Garcia, M. C., Iademarco, M. F., & Richards, C. L. (2018). Modernizing Centers for Disease Control and Prevention Informatics Using Surveillance Data Platform Shared Services. *Public Health Reports*, *133*(2), 130–135. https://doi.org/10.1177/0033354917751130

[21] Lee, T. C., Shah, N. U., Haack, A., & Baxter, S. L. (2020). Clinical Implementation of Predictive Models Embedded within Electronic Health Record Systems: A Systematic Review. *Informatics (MDPI)*, *7*(3), 25. https://doi.org/10.3390/informatics7030025

[22] Ndubuisi, N. E. (2021). Noncommunicable Diseases Prevention In Low- and Middle-Income Countries: An Overview of Health in All Policies (HiAP). *Inquiry: A Journal of Medical Care Organization, Provision and Financing*, *58*, 0046958020927885. https://doi.org/10.1177/0046958020927885

[23] Nilsen, P., Seing, I., Ericsson, C., Birken, S. A., & Schildmeijer, K. (2020). Characteristics of successful changes in health care organizations: An interview study with physicians, registered nurses and assistant nurses. *BMC Health Services Research*, *20*, 147. https://doi.org/10.1186/s12913-020-4999-8

[24] Norori, N., Hu, Q., Aellen, F. M., Faraci, F. D., & Tzovara, A. (2021). Addressing bias in big data and AI for health care: A call for open science. *Patterns*, *2*(10), 100347. https://doi.org/10.1016/j.patter.2021.100347

[25] Panteli, D., Adib, K., Buttigieg, S., Goiana-da-Silva, F., Ladewig, K., Azzopardi-Muscat, N., Figueras, J., Novillo-Ortiz, D., & McKee, M. (2025). Artificial intelligence in public health: Promises, challenges, and an agenda for policy makers and public health institutions. *The Lancet. Public Health*, *10*(5), e428–e432. https://doi.org/10.1016/S2468-2667(25)00036-2

[26] Rahelić, V., Perković, T., Romić, L., Perković, P., Klobučar, S., Pavić, E., & Rahelić, D. (2024). The Role of Behavioral Factors on Chronic Diseases—Practice and Knowledge Gaps. *Healthcare*, *12*(24), 2520. https://doi.org/10.3390/healthcare12242520

[27] Riley, W. J. (2012). Health Disparities: Gaps in Access, Quality and Affordability of Medical Care. *Transactions of the American Clinical and Climatological Association*, *123*, 167–174.

[28] Rosenthal, J. T., Beecy, A., & Sabuncu, M. R. (2025). Rethinking clinical trials for medical AI with dynamic deployments of adaptive systems. *NPJ Digital Medicine*, *8*, 252. https://doi.org/10.1038/s41746-025-01674-3

[29] Sahiner, B., Chen, W., Samala, R. K., & Petrick, N. (2023). Data drift in medical machine learning: Implications and potential remedies. *The British Journal of Radiology*, *96*(1150), 20220878. https://doi.org/10.1259/bjr.20220878

[30] Sedlakova, J., Daniore, P., Horn Wintsch, A., Wolf, M., Stanikic, M., Haag, C., Sieber, C., Schneider, G., Staub, K., Alois Ettlin, D., Grübner, O., Rinaldi, F., & von Wyl, V. (2023). Challenges and best practices for digital unstructured data enrichment in health research: A systematic narrative review. *PLOS Digital Health*, *2*(10), e0000347. https://doi.org/10.1371/journal.pdig.0000347

[31] Shah, H. A., & Househ, M. (2024). Purpose-oriented review of public health surveillance systems: Use of surveillance systems and recent advances. *BMJ Public Health*, *2*(1). https://doi.org/10.1136/bmjph-2023-000374

[32] Soucie, J. M. (2012a). Public Health Surveillance and Data Collection: General Principles and Impact on Hemophilia Care. *Hematology (Amsterdam, Netherlands)*, *17*(0 1), S144–S146. https://doi.org/10.1179/102453312X13336169156537

[33] Soucie, J. M. (2012b). Public Health Surveillance and Data Collection: General Principles and Impact on Hemophilia Care. *Hematology (Amsterdam, Netherlands)*, *17*(0 1), S144–S146. https://doi.org/10.1179/102453312X13336169156537

[34] Torab-Miandoab, A., Samad-Soltani, T., Jodati, A., Akbarzadeh, F., & Rezaei-Hachesu, P. (2025). The impact of electronic medical records on clinical documentation: A case study. *Journal of Education and Health Promotion*, *14*, 246. https://doi.org/10.4103/jehp.jehp_320_24

[35] von Arnim, G., Kohler, S., Hegselmann, S., Marschollek, M., Eils, R., & Wild, B. (2026). Rethinking healthcare data interoperability in the age of large language models. *Med*, *7*(6), 101146. https://doi.org/10.1016/j.medj.2026.101146

[36] Yang, J., Dung, N. T., Thach, P. N., Phong, N. T., Phu, V. D., Phu, K. D., Yen, L. M., Thy, D. B. X., Soltan, A. A. S., Thwaites, L., & Clifton, D. A. (2024). Generalizability assessment of AI models across hospitals in a low-middle and high income country. *Nature Communications*, *15*, 8270. https://doi.org/10.1038/s41467-024-52618-6