Vol. 39 (2026): Proceedings of 2026 International Conference on Life Sciences and Health Engineering
—— Session 3: Health for All Populations: Forum on Public Health and Clinical Health Sciences
(PHCHS-LSHE 2026)
E-ISSN: 3105-0522 P-ISSN: 3105-0514
DOI: https://doi.org/10.70267/PHCHS-LSHE-2026.2026.39.5765

Open-Source Data for Infectious Disease Early Warning: A Narrative Review

Zihua Fang*

School of AI and Advanced Computing, Xi'an Jiaotong-Liverpool University, Suzhou, China

**Corresponding author: Zihua Fang.*

Abstract

Before emerging signals enter routine reporting systems, publicly available online information may already provide relevant clues and offer a broader view of an outbreak, supporting earlier investigation while linking dispersed digital traces into a relatively continuous picture of how risk develops; research in this field is also moving from the monitoring of isolated signals towards multi-source integration and more adaptive analytical methods. However, online signals are influenced not only by disease activity but also by media platforms and the wider social environment, which may reduce data consistency and make model performance less stable across different periods and geographical settings. Open-source early-warning outputs should therefore be treated as preliminary evidence that requires professional interpretation and verification through existing surveillance channels; future work also needs to strengthen continuous calibration and evidence traceability, improve integration with public health workflows, and position open-source early warning as a supplementary risk-sensing function rather than a replacement for traditional infectious disease surveillance.

Keywords

open-source data, infectious diseases, early warning, digital epidemic intelligence, public health surveillance

1. Introduction

Infectious disease outbreaks can develop rapidly, leaving only a limited window for public health intervention. Traditional surveillance provides essential evidence through case reporting, clinical assessment, laboratory testing, and field investigation. However, such information usually becomes available only after the relevant events enter formal reporting channels, which can make early or geographically dispersed abnormalities difficult to identify before sufficient case-based evidence has accumulated. To understand potential risks at an earlier stage, complementary approaches are needed to detect relevant clues while a threat remains uncertain and continues to develop [1].

With the development of the digital information environment, publicly available online information has been used for information screening and event extraction in event-based surveillance [2]. News reports, public outbreak notices, social media activity, and search behaviour may contain signals related to emerging health events.

Compared with formally reported cases, these data are generated outside routine surveillance systems and can generally be collected more quickly [3, 4]. Such applications are associated with digital epidemic

Copyright © 2026 The Authors. All individual papers remain the copyright of their respective author(s) · Published by Zeus Academic Research Publishing Group

intelligence and event-based surveillance, which emphasise the use of rapidly available information to identify and assess potential public health threats. The resulting signals can help surveillance personnel direct their attention towards events that warrant further investigation [5].

Research on open-source early warning spans public health surveillance, epidemiology, and related fields. However, data sources, analytical methods, and practical applications are often discussed separately, which can make several issues difficult to understand, including how publicly available information is transformed into warning signals and how those signals should enter existing surveillance practice. This narrative review examines research on the use of open-source data for infectious disease early warning, with particular attention to data characteristics and processing pathways, as well as the relationship with traditional surveillance and the conditions that affect practical application, thereby providing a reference for understanding research in this field.

2. Research on Open-Source Data-Driven Early Warning for Infectious Diseases

2.1 Main Types of Open-Source Data

Table 1: Main Types of Open-Source Data and Their Surveillance Uses

Data Type	Content	Main Surveillance Uses
Online News and Publicly Available Outbreak Reports	News reports, publicly available webpages, health agency bulletins, ProMED-mail	Identifying clues related to disease events and extracting information such as time, location, disease name, and changes in case numbers
Social Media Posts	Public posts, expressions of symptoms, disease-related discussions, reposted information, and public interactions	Observing public attention to health issues, discussions of symptoms, and risk perceptions
Search Query Data	Changes in searches for disease names, symptom terms, location terms, and related health information	Observing public information-seeking behaviors and helping to identify disease-related attention and potential anomalous signals

The main types of open-source data include online news and publicly available outbreak reports, social media posts, and search query data; their principal characteristics and applications are summarised in Table 1.

Online news and publicly available outbreak reports usually provide event-level information. Such data may contain disease-related clues and describe when and where an event occurred. Some reports also record changes in case numbers and provide background information on affected populations or the course of an outbreak. Once organised by time and location, these reports can be used for event detection and spatial risk assessment. Bhatia et al. used ProMED and HealthMap data collected during the West African Ebola epidemic to retrospectively. Bhatia et al. used ProMED and HealthMap data collected during the West African Ebola epidemic estimate the risks of cross-border spread and case importation. Their results showed that organised open-source event data could support near-real-time risk mapping [6]. Social media posts mainly convey platform users' descriptions of symptoms and public responses to disease-related events. Such content has not entered formal reporting channels and more often represents personal experiences or opinions, so it is mainly used to supplement epidemic situational awareness. Previous studies have used social media to examine symptom-related expressions, levels of public attention, and attitudes towards vaccination [7]. Search query data reflect changes in population-level information-seeking behaviour. These data are usually presented as aggregated time series, making them particularly suitable for tracking temporal changes and supporting nowcasting, and they have been used for the surveillance or prediction of influenza, COVID-19, and dengue [4]. The ARGO model combined the frequency of Google searches with historical influenza-like illness data and used rolling estimation to continually adjust model weights, thereby improving real-time influenza estimation [8].

When considered along the same dimensions, the three data types do not fall into a strict hierarchy, but instead involve different trade-offs (Table 2). Social media can quickly reflect changes in online discussion, although posts sometimes express views on news stories or ordinary concerns and may not accurately represent personal experiences of illness or genuine health events [7]. Search queries can provide timely population-level trends, but their highly aggregated form makes them difficult to link to identifiable cases or specific events [9]. In an analysis of 29 digital surveillance studies, Liscano et al. found that the existing research was

Copyright © 2026 The Authors. All individual papers remain the copyright of their respective author(s) · Published by Zeus Academic Research Publishing Group

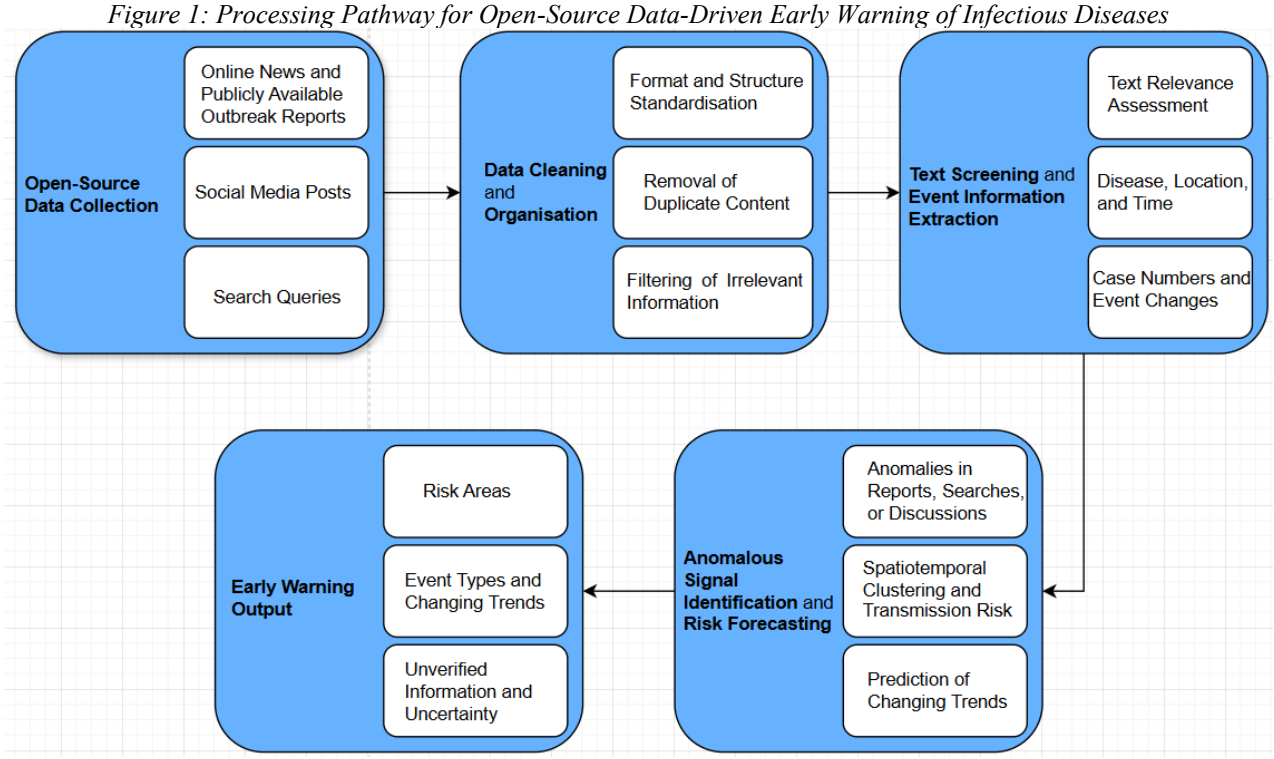
concentrated mainly on search engines such as Google and social media platforms such as X. Search data generally offered broader coverage, whereas social media provided richer contextual information but also contained more noise [10]. Both types of data can reflect activity across large populations, yet they often provide little usable information about where an event occurred. Online news and publicly available outbreak reports provide clearer event context and make it easier to trace the location and timing of an event, although they are generally limited to information available through public reporting channels [6]. The appearance of disease-related terms does not necessarily indicate a genuine health event, so user-generated text needs to be interpreted within its specific context [11]. Data-source selection should depend on the surveillance task and the type of information required, rather than on finding a source that is superior in every respect.

Table 2: Comparative Characteristics of the Three Main Types of Open-Source Data

Comparison dimension	Online news and publicly available outbreak reports	Social media	Search queries
Timeliness	Subject to reporting, editing, and publication processes; signals may emerge later than user-generated online activity	Continuously generated by users and able to reflect changes in online discussion relatively quickly	Affected by the platform's update frequency, but able to reflect changes in information-seeking behaviour relatively quickly
Verifiability	Event sources, locations, and contextual information can usually be identified	Users' identities and health status, as well as the relationship between posts and actual events, are difficult to verify	Cannot be directly linked to individual cases or specific events
Spatial coverage and spatial precision	Provides relatively clear event locations, but coverage is limited to events that enter public reporting channels	Covers large populations, but most posts do not contain usable location information	Provides population-level trends across relatively broad areas, but spatial precision is constrained by the geographical units released by the platform
Sources of noise	Duplicate reporting, editorial selection, and uneven media attention	Reposts, discussions not based on personal experience, and semantic ambiguity	Media attention and information-seeking behaviour unrelated to actual illness

2.2 Data Processing Pathway and Implementation Methods

Figure 1: Processing Pathway for Open-Source Data-Driven Early Warning of Infectious Diseases



Copyright © 2026 The Authors. All individual papers remain the copyright of their respective author(s) · Published by Zeus Academic Research Publishing Group

After open-source data have been collected, the sources often differ in format, structure, and update frequency, and may also contain duplicate records or irrelevant content. They therefore cannot usually be converted directly into early-warning outputs. Data standardisation, deduplication, and preliminary screening are required first to provide a more consistent basis for subsequent analysis.

Text screening and event information extraction follow. Joshi et al. divided text-based epidemic intelligence into health mention classification and health event detection. The former screens large volumes of text for relevant content, whereas the latter identifies possible epidemic events from the temporal or spatial distribution of related texts [11]. Rule-based methods are a relatively conventional choice for these tasks. They rely on dictionaries and manually specified patterns, their decision criteria are comparatively transparent, and they require less annotated data, but they do not readily adapt to new forms of language. With suitable training data and feature design, machine-learning models can perform well, learning from labelled examples and offering greater flexibility. Deep-learning models require more data and computing resources, but can capture more complex linguistic context while reducing reliance on manually designed features. Their decision processes, however, are correspondingly harder to interpret [11].

EventEpi combines several techniques within an operational workflow. It first assigns relevance scores to World Health Organization Disease Outbreak News and ProMED-mail articles, and then uses named entity recognition and information extraction to identify diseases and countries and to extract dates and confirmed case counts. The extracted information is stored in a structured database, reducing repeated screening and manual data entry [2].

The next stage is anomaly detection and risk prediction, in which event data and spatiotemporal distributions are analysed together to identify signals that may depart from normal patterns and to forecast changes in trends. Machine learning has become a major technical approach at this stage because it can integrate multiple predictors and represent nonlinear relationships [12, 13], although different algorithms are suited to different prediction tasks. Tree-based models, such as random forests and gradient boosting, are suitable for heterogeneous tabular data and can be used for risk classification or regional risk estimation. Temporal and spatial relationships usually need to be introduced through manually constructed features, however, and model performance may decline when the distribution of predictors changes [12]. Recurrent neural networks and long short-term memory networks can represent temporal dependencies and are better suited to sequence prediction when relatively large and consistent datasets are available, but the forecasting horizon needs to be controlled because errors may accumulate over longer periods. Graph neural networks are appropriate when disease risk is closely related to population movement or connections between regions; their reliability depends on how accurately the constructed network reflects actual transmission relationships [13]. The choice among these algorithms needs to take account of the amount of available data, the spatiotemporal structure of the task, interpretability requirements, and the intended form of the warning output.

Finally, the analytical results are organised into warning outputs. Rather than producing only a brief numerical value, the process gradually turns dispersed public information into risk clues that can support early warning, presenting risk areas, event types, and changing trends while marking information that has not yet been verified for further assessment and confirmation by public health personnel [5]. The overall processing pathway is illustrated in Figure 1.

The open-source data-driven hybrid modelling system developed by Zhang et al. provides a complete implementation of this pathway. It first applies semantic harmonisation to heterogeneous open and partner-shared signals, and then connects seasonality-adjusted baselines and anomaly detection with SEIR models incorporating population mobility and climate factors, as well as short-horizon sequence-learning models, to generate calibrated early-warning scores [14]. The results are delivered into existing public health workflows through operator-oriented dashboards and are fed back for recalibration against subsequent confirmation outcomes [14].

3. Relationship Between Open-Source Data and Traditional Infectious Disease Surveillance

Viewed across the surveillance process, open-source data-based early warning and traditional infectious disease surveillance differ in their emphasis and do not perform their tasks at the same stage (Table 3). Open-source early warning mainly searches publicly available information for unusual signals and places greater

Copyright © 2026 The Authors. All individual papers remain the copyright of their respective author(s) · Published by Zeus Academic Research Publishing Group

emphasis on identifying risks that warrant attention as early as possible, whereas traditional surveillance relies more heavily on conventional information that generally becomes available after patients seek care and cases are reported, using it to detect or confirm outbreaks and to continue monitoring disease trends.

Table 3: Comparison Between Traditional Infectious Disease Surveillance and Open-Source Data-Based Early Warning

Comparison Dimension	Traditional Infectious Disease Surveillance	Open-Source Data-Based Early Warning
Data Sources	Case reports, clinical records, laboratory data, and field investigations	Online news, social media, search queries, and digital epidemic intelligence platforms
Information Generation	Primarily structured data defined by case definitions, usually generated after diagnosis and reporting	Primarily unstructured information or behavioural signals, which may emerge earlier
Main Outputs	Case counts, disease trends, and investigation results	Anomalous clues and risk alerts
Information Certainty and Verification	Relatively high certainty, with relatively standard procedures	Relatively preliminary information that requires further verification and may include false-positive signals
Role	Detecting or confirming outbreaks, conducting continuous surveillance, and supporting formal responses	Providing alerts, screening clues, and prioritising investigations

In public health surveillance, the former pathway is generally close to event-based surveillance, while the latter more clearly reflects the features of indicator-based surveillance. Norzin et al. noted that event-based surveillance rapidly collects, assesses, and processes unstructured information to detect early signals of new public health threats, placing greater emphasis on finding relevant clues. Indicator-based surveillance, by contrast, relies on structured data that meet established case definitions and is mainly used to identify disease outbreaks and describe disease trends, with greater emphasis on standardised surveillance and continued observation; the two approaches can complement each other [5]. Related work by Magid et al. used the terms “informal digital surveillance” and “formal surveillance” in a distinction that broadly corresponds to open-source data-based early warning and traditional infectious disease surveillance. Informal digital surveillance can supplement formal surveillance during an outbreak, but its early-detection capacity still lacks sufficient prospective evidence and it cannot replace formal surveillance [3].

During the COVID-19 pandemic, Stolerman et al. combined public search activity, clinicians’ information-seeking activity, geolocated social media data, and epidemiological information to develop early-warning systems for 97 counties in the United States. Across 367 outbreak-onset events, the multi-source approach issued warnings one to six weeks in advance for 254 events (69%), helping to identify signs of increasing local disease activity before the changes were fully visible in case reports [15]. However, the study still relied on traditional data, particularly case reports, to determine when an outbreak began, and the system also generated 110 false alarms.

Hutchinson et al. used EPIWATCH to monitor online reports of rash and fever during the early phase of the 2022 mpox outbreak. The system identified 656 reports in 2022, compared with 75 during the corresponding period in 2021, and the marked increase suggested that an unusual pattern might be emerging [16]. However, the geographical distribution of these reports did not closely match that of confirmed mpox cases, and many of the reports concerned other diseases. The system could therefore indicate that unusual activity might be occurring, but it could not determine whether mpox was responsible. These cases show that open-source signals are better suited to identifying unusual activity and drawing surveillance personnel’s attention to changes at an earlier stage. They can raise questions requiring further investigation and help guide testing or field investigation, whereas formal surveillance provides clinical and laboratory evidence and determines whether an observed increase truly represents an outbreak; subsequent risk assessment and decisions on disease prevention and control also depend on the traditional surveillance system.

Copyright © 2026 The Authors. All individual papers remain the copyright of their respective author(s) · Published by Zeus Academic Research Publishing Group

4. Discussion

4.1 Application Value of Open-Source Data

The most direct value of open-source data lies in widening the observation window for infectious disease risk. Public reports, online discussions, and search activity may appear before the relevant event is fully reflected in confirmed case reports or routine surveillance [3]. When these signals are organised by time, location, and disease topic, surveillance personnel can examine whether they persist, whether they spread to new areas, and whether they change as the event develops. Early warning is therefore no longer limited to reacting to a single unusual item of information, but can follow the evolution of risk over time and give public health personnel more time to decide whether further investigation is needed. Open-source data can also broaden the range of observable health events, extending surveillance beyond confirmed diseases to potential risk manifestations that arise earlier and have not yet been clearly classified. The objects of surveillance may include not only diseases that have already been identified by name, but also symptom descriptions, syndromic patterns, and emerging events whose causes remain uncertain. Local reports may provide relatively early information on the time and location of an emerging event or a local cluster [6], while syndromic surveillance can indicate unusual patterns of rash or fever before the cause has been established [16]. Online information can also be collected across multiple regions, widening the spatial view of surveillance without being completely restricted by a single administrative reporting chain and bringing dispersed regional signals into a broader observation of disease activity. Digital public health surveillance studies have covered 54 countries and 26 types of platforms [9], supporting risk observation across different geographical settings. A wider spatial view can help identify where unusual activity is appearing and can also direct limited investigative resources towards areas that deserve greater attention.

Different data sources can also enrich the basis for judging the same risk by providing different forms of evidence. A multi-source influenza model performed better in real-time estimation than predictions based on any single source [17], while ARGO also demonstrated the value of dynamically combining search activity with historical surveillance data [8]. In county-level COVID-19 early warning, the multi-source method produced fewer false alarms than the single-source methods evaluated in the same study [15]. When several sources point to the same change, they can provide a stronger basis for assigning a higher priority to event investigation; however, multi-source integration does not always improve predictive performance. Yang et al. used Google Trends and Twitter data from the early phase of the pandemic in Canada to forecast COVID-19 cases. The LSTM model based on selected Google Trends symptom signals performed best, whereas adding Twitter data did not further improve the predictions [18]. Such differences can indicate where verification should begin and can also help determine whether an additional source has actually contributed complementary information.

4.2 Practical Challenges

One of the main gaps revealed by existing research is that publicly available information does not always correspond to conditions on the ground, which makes unusual signals difficult to interpret consistently as public health risks. News reports and social media are shaped by media attention and user behaviour, while the recommendation mechanisms used by some platforms may also influence search content. An observed increase may therefore reflect rising information interest rather than the spread of disease. Models built on historical associations may lose their original validity when a new disease emerges or the information environment changes. Research consequently needs to explain more clearly how these signals are generated and connect them with verifiable events, rather than treating online attention itself as evidence of an impending outbreak.

The operation of digital platforms represents a second source of instability, because researchers cannot fully control how data are generated, standardised, and returned. Franzén found that repeated Google Trends searches using the same conditions for the same geographical area and historical period could produce different curves when conducted on different dates [19]. Such changes alter the inputs available to a model and may make previously established thresholds or feature weights no longer applicable. ARGO uses rolling estimation to update continually the relationship between search activity and surveillance data [8], while the COVID-19 early-warning system developed by Stolerman et al. retrains its models after new outbreak events become

Copyright © 2026 The Authors. All individual papers remain the copyright of their respective author(s) · Published by Zeus Academic Research Publishing Group

available [15]. These approaches may improve model adaptability, but they cannot remove the dependence of the system on continued access to comparable platform data.

Patterns of internet use across countries and regions also affect data representativeness and model transferability. Internet access, language, commonly used platforms, and online health information-seeking behaviour all influence which populations and events become visible in digital data. A systematic review found that digital surveillance studies were concentrated in a small number of countries and that many did not examine demographic or geographical differences [9]. In the EPIWATCH study of mpox, online reports were concentrated mainly in parts of Asia, whereas the regions carrying the highest burden of confirmed cases were located elsewhere [16]. A model developed in one setting may therefore not be directly applicable to another, and data sources, language processing, and validation procedures need to be adapted to the local context.

Model reliability should also be assessed in relation to its intended public health use. Early signals are usually produced while information remains incomplete and therefore require further verification by trained personnel [5]. The county-level COVID-19 early-warning system, for example, did issue alerts in advance, but the number of false alarms could not be disregarded [15]. EventEpi likewise achieved relatively strong recall when screening articles, although its precision was not high enough to allow human review to be omitted [2]. In addition, much of the available evidence comes from retrospective studies, while only a small proportion of digital surveillance research has been deployed in routine public health settings [9]. Predictive accuracy alone is therefore an incomplete measure; evaluations should also consider missed events, the burden of verification, the interpretability of results, and whether an alert leads to public health action of practical value [13].

4.3 Future Development Directions

Future development should not focus only on increasing the scale of data or the complexity of models. The position of open-source early warning within the public health system needs to be clarified first. In its mature form, it should not operate as an automated platform that independently determines whether an outbreak is occurring, but as a risk-sensing layer at the front end of traditional surveillance. Its role is to widen the scope of information discovery, screen for changes that warrant investigation, and help limited verification resources give priority to high-risk events, while organising scattered clues into evidence that can be interpreted and verified.

At the technical level, the continued improvement and adaptability of early-warning models are important [12, 13]. Systems should move away from static models based on one-time training and fixed thresholds and towards dynamic tools that can be recalibrated continuously. Platform mechanisms and patterns of public attention will continue to change, so future models need to update anomaly baselines and adjust the weights assigned to different signals, while retaining the main evidence that triggered an alert and revising their judgements in light of professional review and subsequent surveillance findings. At the application level while retaining the main evidence that triggered an alert and revising their judgements in light of professional review and, future systems should establish a closed loop of “open-information discovery–multi-source evidence assessment–professional review–verification through traditional surveillance–feedback-based recalibration”. The hybrid early-warning system developed by Zhang et al. has connected open-source and partner-shared signals to existing public health workflows through semantic harmonisation, operator-oriented dashboards, and subsequent feedback recalibration [14]. On this basis, evaluation should shift from the accuracy of an individual model to whether the process as a whole is effective, including whether useful clues enter investigation earlier, whether the human workload is acceptable, and whether the results can be understood and used for action. When predictive tools enter public health practice, attention also needs to be paid to result interpretation, uncertainty communication, and user feedback [20]. The future direction, therefore, should not be the pursuit of complete automation, but the establishment of stable human–machine collaboration: algorithms manage the scale of information, professionals interpret the evidence, and traditional surveillance provides verification and response. Once incorporated into this working loop, open-source early warning can move from a technical prototype to a sustained public health capability.

5. Conclusion

Infectious disease control requires not only a timely response after cases have been confirmed, but also the early recognition of changes that warrant investigation before formal evidence has been fully established.

Copyright © 2026 The Authors. All individual papers remain the copyright of their respective author(s) · Published by Zeus Academic Research Publishing Group

Open-source data can turn information scattered across publicly available online sources into risk signals that can be followed over time, allowing public health personnel to understand the development of an event earlier and to judge which unusual changes deserve closer attention. Information from different sources can also complement one another, providing a stronger basis for prioritising investigations and subsequent verification. These data, however, are not equivalent to evidence of an outbreak; when the information environment or the operation of digital platforms changes, the meaning of online signals may change accordingly, so open-source early warning should not independently undertake outbreak confirmation, but should serve as a front-end mechanism for information discovery and risk signalling within traditional surveillance. Future work should likewise focus not only on improving model performance, but also on ensuring that warning outputs can be recalibrated continuously, explained clearly, and incorporated smoothly into professional review and formal verification; only when this connection is established can open-source early warning move from a technical tool used in research to a stable form of public health support.

References

[1] C. R. MacIntyre *et al.*, 'Artificial intelligence in public health: the potential of epidemic early warning systems', *J. Int. Med. Res.*, vol. 51, no. 3, p. 03000605231159335, Mar. 2023, doi: 10.1177/03000605231159335.

[2] A. Abbood, A. Ullrich, R. Busche, and S. Ghozzi, 'EventEpi—A natural language processing framework for event-based surveillance', *PLOS Comput. Biol.*, vol. 16, no. 11, p. e1008277, Nov. 2020, doi: 10.1371/journal.pcbi.1008277.

[3] A. Magid, A. Gesser-Edelsburg, and M. S. Green, 'The Role of Informal Digital Surveillance Systems Before, During and After Infectious Disease Outbreaks: A Critical Analysis', in *Defence Against Bioterrorism*, V. Radosavljevic, I. Banjari, and G. Belojevic, Eds, Dordrecht: Springer Netherlands, 2018, pp. 189–201. doi: 10.1007/978-94-024-1263-5_14.

[4] H. McClymont, S. B. Lambert, I. Barr, S. Vardoulakis, H. Bambrick, and W. Hu, 'Internet-based Surveillance Systems and Infectious Diseases Prediction: An Updated Review of the Last 10 Years and Lessons from the COVID-19 Pandemic', *J. Epidemiol. Glob. Health*, vol. 14, no. 3, pp. 645–657, Aug. 2024, doi: 10.1007/s44197-024-00272-y.

[5] T. Norzin *et al.*, 'Event-based surveillance: Providing early warning for communicable disease threats', *Can. Commun. Dis. Rep.*, vol. 49, no. 2–3, pp. 29–34, 2023, doi: 10.14745/ccdr.v49i23a01.

[6] S. Bhatia *et al.*, 'Using digital surveillance tools for near real-time mapping of the risk of infectious disease spread', *Npj Digit. Med.*, vol. 4, no. 1, p. 73, Apr. 2021, doi: 10.1038/s41746-021-00442-3.

[7] L. E. Charles-Smith *et al.*, 'Using Social Media for Actionable Disease Surveillance and Outbreak Management: A Systematic Literature Review', *PLOS ONE*, vol. 10, no. 10, p. e0139701, Oct. 2015, doi: 10.1371/journal.pone.0139701.

[8] S. Yang, M. Santillana, and S. C. Kou, 'Accurate estimation of influenza epidemics using Google search data via ARGO', *Proc. Natl. Acad. Sci.*, vol. 112, no. 47, pp. 14473–14478, Nov. 2015, doi: 10.1073/pnas.1515373112.

[9] Z. Shakeri Hossein Abad *et al.*, 'Digital public health surveillance: a systematic scoping review', *Npj Digit. Med.*, vol. 4, no. 1, p. 41, Mar. 2021, doi: 10.1038/s41746-021-00407-6.

[10] Y. Liscano, L. A. Anillo Arrieta, J. F. Montenegro, D. Prieto-Alvarado, and J. Ordoñez, 'Early Warning of Infectious Disease Outbreaks Using Social Media and Digital Data: A Scoping Review', *Int. J. Environ. Res. Public. Health*, vol. 22, no. 7, p. 1104, Jul. 2025, doi: 10.3390/ijerph22071104.

[11] A. Joshi, S. Karimi, R. Sparks, C. Paris, and C. R. Macintyre, 'Survey of Text-based Epidemic Intelligence: A Computational Linguistics Perspective', *ACM Comput Surv*, vol. 52, no. 6, p. 119:1-119:19, Oct. 2019, doi: 10.1145/3361141.

Copyright © 2026 The Authors. All individual papers remain the copyright of their respective author(s) · Published by Zeus Academic Research Publishing Group

[12] C. El Morr, D. Ozdemir, Y. Asdaah, A. Saab, Y. El-Lahib, and E. S. Sokhn, 'AI-based epidemic and pandemic early warning systems: A systematic scoping review', *Health Informatics J.*, vol. 30, no. 3, p. 14604582241275844, Jul. 2024, doi: 10.1177/14604582241275844.

[13] M. Liu, Y. Liu, and J. Liu, 'Machine Learning for Infectious Disease Risk Prediction: A Survey', *ACM Comput. Surv.*, vol. 57, no. 8, p. 212:1-212:39, Mar. 2025, doi: 10.1145/3719663.

[14] J. Zhang *et al.*, 'An Open-Source Data Driven Hybrid Modeling System for Infectious Disease Surveillance and Early Warning', *China CDC Wkly.*, vol. 8, no. 8, pp. 216–221, Feb. 2026, doi: 10.46234/ccdcw2026.036.

[15] L. M. Stolerman *et al.*, 'Using digital traces to build prospective and real-time county-level early warning systems to anticipate COVID-19 outbreaks in the United States', *Sci. Adv.*, vol. 9, no. 3, p. eabq0199, Jan. 2023, doi: 10.1126/sciadv.abq0199.

[16] D. Hutchinson, M. Kunasekaran, A. Quigley, A. Moa, and C. R. MacIntyre, 'Could it be monkeypox? Use of an AI-based epidemic early warning system to monitor rash and fever illness', *Public Health*, vol. 220, pp. 142–147, Jul. 2023, doi: 10.1016/j.puhe.2023.05.010.

[17] M. Santillana, A. T. Nguyen, M. Dredze, M. J. Paul, E. O. Nsoesie, and J. S. Brownstein, 'Combining Search, Social Media, and Traditional Data Sources to Improve Influenza Surveillance', *PLOS Comput. Biol.*, vol. 11, no. 10, p. e1004513, Oct. 2015, doi: 10.1371/journal.pcbi.1004513.

[18] Y. Yang, S.-F. Tsao, M. A. Basri, H. H. Chen, and Z. A. Butt, 'Digital Disease Surveillance for Emerging Infectious Diseases: An Early Warning System Using the Internet and Social Media Data for COVID-19 Forecasting in Canada', in *Caring is Sharing – Exploiting the Value in Data for Health and Innovation*, IOS Press, 2023, pp. 861–865. doi: 10.3233/SHTI230290.

[19] A. Franzén, 'Big data, big problems: Why scientists should refrain from using Google Trends', *Acta Sociol.*, vol. 66, no. 3, pp. 343–347, Aug. 2023, doi: 10.1177/00016993221151118.

[20] C. S. Lutz *et al.*, 'Applying infectious disease forecasting to public health: a path forward using influenza forecasting examples', *BMC Public Health*, vol. 19, no. 1, p. 1659, Dec. 2019, doi: 10.1186/s12889-019-7966-8.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.

Open Access

This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Copyright © 2026 The Authors. All individual papers remain the copyright of their respective author(s) · Published by Zeus Academic Research Publishing Group