Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

CHAPTER

22

Artificial intelligence–enabled public health surveillance—from local detection to global epidemic monitoring and control

Daniel Zeng, Zhidong Cao and Daniel B. Neill

Abstract

Artificial intelligence (AI) techniques have been widely applied to infectious disease outbreak detection and early warning, trend prediction, and public health response modeling and assessment. Such public health surveillance and response tasks of major importance pose unique technical challenges such as data sparsity, lack of positive training samples, difficulty in developing baselines and quantifying the control measures, and interwoven dependencies between spatiotemporal elements and finer-grained risk analyses through contact and social networks. Traditional public health surveillance relies heavily on statistical techniques. Recent years have seen tremendous growth of AI-enabled methods, including but not limited to deep learning–based models, complementing statistical approaches. This chapter aims to provide a systematic review of these recent advances applying AI techniques to address public health surveillance and response challenges.

Keywords: AI-enabled public health surveillance; infectious disease surveillance; early warning; public health response

22.1 Introduction

In the recent decade, partially fueled by major advances in big data and raw computing power, artificial intelligence (AI) technology has entered an extraordinary phase of fast development and wide application. The techniques developed in traditional AI research areas such as computer vision, speech recognition, natural language processing, and robotics have found many innovative applications in an array of real-world settings, including medicine. The general methodological contributions from AI, such as a variety of recently developed deep

Artificial Intelligence in Medicine
DOI: https://doi.org/10.1016/B978-0-12-821259-2.00022-3

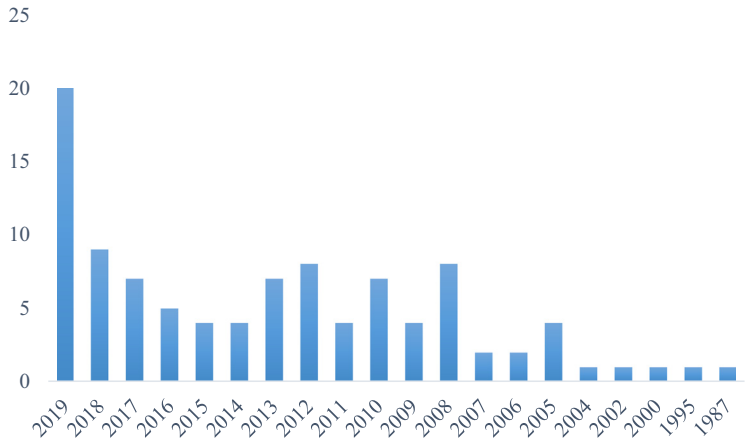

FIGURE 22.1 Numbers of published papers containing both "artificial intelligence" and "public health surveillance" as keywords from Web of Science, accessed in January 2020.

learning algorithms, have also been applied to a wide spectrum of fields. Public health surveillance is one such area that has benefited significantly from these recent AI advances. As shown in Fig. 22.1, the growing literature on the AI-enabled or -enhanced public health surveillance work illustrates the relevant research community's interest in applying AI techniques. In another example, as part of the research community's response to the COVID-19 pandemic, there was a specific call for developing AI-based public health solutions.

How can AI enhance existing public health surveillance and response approaches and enable new ones? The answer lies with the fundamental challenges facing public health surveillance and response. Public health surveillance is intrinsically data driven. Identifying early, accurate, and reliable signals of health anomalies and disease outbreaks from a heterogeneous collection of data sources has always been the main objective of public health surveillance. Technically, this translates into two distinct challenges: the data sourcing challenge and the analytics challenge. The first data sourcing challenge is concerned with determining easily operationalizable sources of data that contain useful signals. The second analytics challenge is concerned with developing effective computational frameworks to extract such signals. AI provides a range of methods and techniques to help tackle both challenges. Other major objectives of public health surveillance and response are to analyze and predict infectious disease trends through modeling the disease transmission dynamics and assess public health responses. Accomplishing these objectives entails domain knowledge– and context-rich predictions, fine-grained risk analytics through contact and social networks, quantification of responses and control measures, and assessment of these responses and measures in the presence of complex interactions and constraints. AI offers a suite of applicable modeling and analytics frameworks to address these complex considerations.

Fig. 22.2 summarizes the major ways through which AI enhances existing public health surveillance and response approaches and enables new ones.

First, AI opens the door to make use of a variety of novel or underexplored data sources for public health surveillance purposes, especially those not originally or intentionally designed to answer epidemiological questions. For instance, with the rapid development of the Internet and the Internet of Things applications, ubiquitous social and device

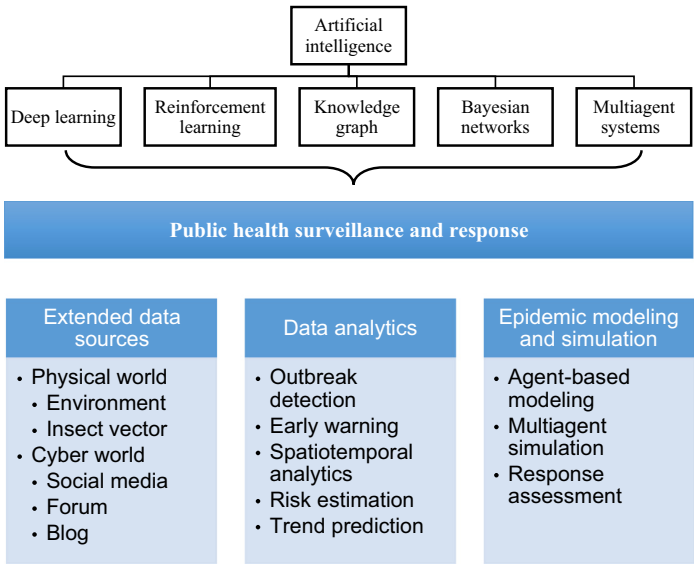

FIGURE 22.2 Enabling and enhancing public health surveillance and response approaches through AI applications.

sensing capabilities are becoming a reality, presenting significant surveillance potentials. A variety of open data, external to traditional public health surveillance systems, can be fruitfully exploited to enhance the surveillance capabilities.

Second, AI enhances the traditional suite of data analytics tools, which is mainly statistics based, to deal with the new surveillance data types that cover unstructured and semistructured text, images, and videos, in addition to structured information items. Dealing with unstructured data necessitates the use of AI methods such as natural language processing and image processing, which often include a deep learning component for automatic data-driven feature construction. Through the application of these methods, unstructured data can be converted into structured items through semantic labels and autofilled features.

From an analytical standpoint, public health surveillance is concerned with timely and effectively assessing the risk of an epidemic, detecting abnormal changes in the spatiotemporal status of the epidemic, so as to provide early warning and predict the trend of the epidemic. AI methods, in particular, those based on machine learning, have long been applied to detect patterns, identify anomalies, and analyze trends and risks, from public health surveillance data streams. Such data streams often possess prominent temporal and spatial elements and need to be analyzed along with external social, economic, and environmental data. Traditional surveillance methods rely heavily on the use of statistical methods. As the data become increasingly complex, within the framework of these methods, statistical inferences become rather difficult. Moreover, statistical methods focus on conclusions at the macrolevel, whereas machine learning methods enable customized inferences aimed at characterizing local patterns.

Third, AI provides modeling frameworks to simulate complex setups and scenarios of infectious disease transmission and public health responses. The evolution of epidemics in time,

space, and people has a high degree of uncertainty and complexity. The infection transmission dynamics is nonlinear by nature, with chaotic characteristics and poor predictability. As such, the applicability of the models based on aggregated statistics and linear interactions is inherently limited. A subarea of AI, multiagent systems, offers modeling frameworks that allow for the study of the evolution of the epidemic under different conditions. These frameworks can also be used to quantitatively evaluate the effect of different interventions and control measures.

The remainder of this chapter reviews the application of AI in public health surveillance and response. Section 22.2 discusses AI-enhanced data analysis techniques for outbreak detection and early warning. Section 22.3 focuses on AI-enhanced prediction methods in support of surveillance and trend analysis tasks. In Section 22.4, we review AI-based simulation frameworks to characterize infectious disease transmission patterns and assess public health responses. In Section 22.5, we briefly summarize several Internet-based surveillance systems aimed at global epidemic monitoring. Section 22.6 concludes this chapter with a summary of the key findings and a discussion of challenges ahead.

22.2 Artificial intelligence–enhanced data analysis for outbreak detection and early warning

In order to improve the timeliness and accuracy of outbreak detection and early warning approaches, public health researchers continue to investigate and explore sensor data and indicators from the physical world covering health, environmental, societal, and economic aspects, among others. Significant efforts have been expended to make use of data from the cyberspace, such as keyword searches, blogs, and social networking posts. Fig. 22.3 illustrates these sources of data, along with the commonly used machine learning methods.

In the remainder of this section, we first discuss two sets of AI-enhanced data analysis techniques: one concerned with data from the physical world and the other from the cyberspace. Then we consider how machine learning techniques for text analysis and event detection can provide a "safety net" by identifying previously unseen disease outbreaks and other emerging events of interest to public health.

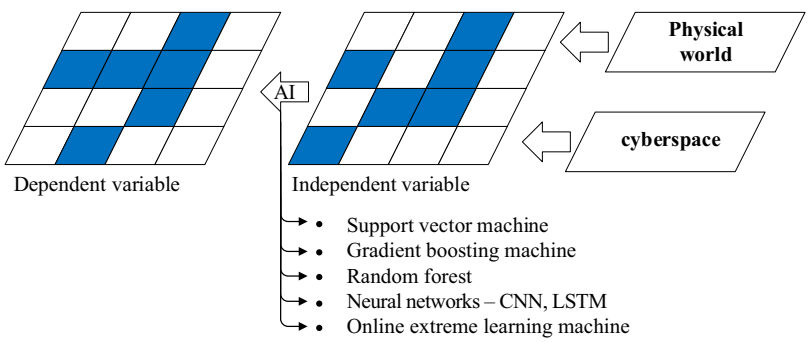

FIGURE 22.3 Sources of public health surveillance data and commonly used machine learning methods.

22.2.1 Analyzing data collected from the physical world

The spatiotemporal pattern of epidemiological risks is related to various factors such as climatic conditions, social and economic status, and vector transmissions. Assessing the outbreak risk of an infectious disease is important for early warning and effective resource deployment. Based on the data collected from the physical world, machine learning methods have been successfully applied to estimate the high-risk regions and outbreak periods.

Support vector machine (SVM), gradient boosting machine, and random forest (RF) were applied to simulate the global distribution of *Aedes aegypti* and *Aedes albopictus* to fight against mosquito-borne infectious diseases, for example, ZIKV, dengue, and chikungunya. Effectively killing the vector cuts off the disease's transmission path. Multidisciplinary datasets, such as occurrence records, social factors, and meteorological factors, were quantified to train the models. It is reported that RF obtained the highest AUC value, and the temperature suitability had the best discriminatory power among factors.[1] RF was also used to assess the risk of dengue transmission in Singapore with dengue, population, entomological, and environmental data. Random bootstrap samples were drawn from the data, and an unpruned decision tree was fitted to each bootstrap sample. The risk maps had high accuracy in that more than 80% of the observed risk ranks fell within the 80% prediction interval.[2] Another framework based on neural networks and online extreme learning machine (OLEM) estimated the distribution of kinds of water containers with the Aedes mosquito larvae in Recife, Brazil. Nine years of environmental and entomological data were used to train the OLEM model.[3]

Deep learning models have been applied to detect outbreaks of infectious diseases. A dynamic neural network model was developed to predict the outbreak risk of ZIKV in the Americas. This model utilized history epidemiological data, air travel volumes, vector distribution, and socioeconomic factors. The main feature of this modeling work is its flexibility. Decision-makers can easily modify the risk indicator, risk classification scheme, and prediction forecast window according to their own customized needs.[4] To examine emerging spatiotemporal hotspots of dengue fever at the township level in Taiwan, a deep AlexNet model was trained on sea surface temperature images and rainfall data by transfer learning. This transfer learning–based method overcame the overfitting problem due to the small dataset and yielded an accuracy of 100% on an eightfold cross-validation test dataset.[5]

The general trend of spatiotemporal analysis is that the data resolution and the number of exogenous variables are increasing. In this regard, the nonlinear fitting ability of machine learning, in particular, deep learning, models offers many advantages over classical statistical models.

22.2.2 Analyzing data from the cyberspace

Internet application usage data (e. g., keyword searches) and social media data are widely studied for rapid response to infectious disease outbreaks. Machine learning methods have been used for text classification and sentiment analysis from social media data for surveillance purposes. A social media–based early warning system for mosquito-borne disease in India was proposed.[6] Latent Dirichlet allocation–based topic modeling techniques were

applied to identify relevant topics related to symptoms, prevention, and public sentiments toward the disease. The real-time tracking of public sentiments provided an early warning mechanism. DEFENDER is a software system developed in the United Kingdom that integrates Twitter and news media for outbreak detection.[7] SVM and naive Bayes classifiers were used for disease-related text classification. The DBSCAN algorithm was utilized to cluster the geographic space and observe the movement behavior of Internet users. The second-generation system (SENTINEL) of DEFENDER further improved the text classification and denoising algorithm using CNN and LSTM networks.[8]

Twitter has proven its usefulness as a public health surveillance data source. The Twitter data feed is real time and can be obtained from a large number of users in different geographic regions. At the same time, interpreting Twitter data semantically can be challenging, given that tweets are often very short and full of incomplete and informal writing. Chen and Neill analyzed the heterogeneous network structure of Twitter using a nonparametric graph scan, and applied this approach to detection of hantavirus outbreaks in Chile.[9] Dai and Bikdash[10] studied influenza-related tweet classification as a surveillance tool. In their proposed hybrid classification approach, they combined artificially defined features with features automatically generated by supervised machine learning methods to separate tweets involving flu cases versus tweets that do not involve flue cases. Dai et al.[11] reported a clustering method based on word embedding for public health monitoring. This method learns semantically meaningful representational vectors from surrounding words. Based on the cluster similarity measures, tweets can in turn be classified as relevant or irrelevant to a certain topic (e.g., flu). Wang et al.[12] proposed a long- and short-term RNN structure to classify infectious disease-related tweets and showed that this deep learning model outperformed a range of standard machine learning models. Lampos et al.[13] used a neural network word embedding model trained on social media content on Twitter to determine the degree of semantic relevance of the text to infectious diseases. An "influenza infection" concept was developed and used to reduce false and potentially confusing features selected by previous commonly used methods. Edo-Osagie et al.[14] used an attention-based short text classification method to mine information on Twitter for public health monitoring. The goal of the algorithm was to automatically filter Twitter related to asthma syndrome. In addition, the algorithm contained a binary recurrent neural network architecture with an attention layer (ABRNN) that allows the network to weight words in Twitter based on perceived importance. Souza et al. develop new machine learning approaches to identify geographic hot-spots of dengue infection risk, using a large Twitter dataset from Brazil Souza et al.[15,16]

Recently, Shah and Dunn[17] proposed a machine learning method to generate a model to detect the magnitude of unexpected changes in terms of usage with spatiotemporal patterns from social media data streams. This work has direct public health relevance and can be used for health events represented by relatively infrequent terms.

In practical applications, social media–based outbreak detection and early warning methods are not without problems. The key challenges include the heterogeneity of the geospatial distribution of data, the demographic heterogeneity of online users, and the unavailability of data and language resources in underdeveloped regions. It was also noted that for such methods to work, supervised learning models would need a large amount of labeled data.[18] Special care has to be given to reduce biases when adopting these methods.

22.2.3 From syndromic to pre-syndromic disease surveillance: A safety net for public health

Over the past two decades, numerous techniques have been developed in the disease surveillance, statistics, and AI communities for *syndromic surveillance*: early detection of disease outbreaks by identifying emerging clusters of cases in space and time. A variety of public health data sources, such as hospital emergency department visits, over-the-counter medication sales[19], and more recently online data sources such as search queries[20,21] and social media[9], have been employed for this task, and spatial event detection approaches based on the spatial and subset scan statistics[22,23] have become increasingly widespread in public health practice. Such approaches typically classify cases to a set of known syndrome types (such as influenza-like illness or gastrointestinal illness) based on pre-established rules, then detect regions of space and time with significantly higher than expected case counts. Heuristic search methods such as simulated annealing and genetic algorithms[25,26], fast subset[24] scan approaches for exact optimization over subsets and machine learning approaches such as support vector machines[27] have been used to increase the flexibility of syndromic surveillance and to enable more accurate detection of irregularly-shaped spatial clusters.

However, these syndromic surveillance approaches, as well as other public health approaches such as notifiable disease reporting, are unable to detect newly emerging ("novel") outbreaks with previously unseen patterns of symptoms, or other unexpected events of relevance to public health. Such events would not be mapped to any of the existing syndrome categories, or would be lumped into a broader and less informative syndrome definition, thus diluting or entirely removing the outbreak signal. This necessitated the development of new machine learning approaches for "pre-syndromic" surveillance[28,29] that do not rely on existing syndrome categories, but instead analyze free text data such as emergency department chief complaints to identify emerging patterns of keywords. While early pre-syndromic surveillance approaches[30,31] treat each keyword in isolation, identifying any novel words or those that substantially increased in frequency, more recent machine learning approaches[32,34] develop novel variants of

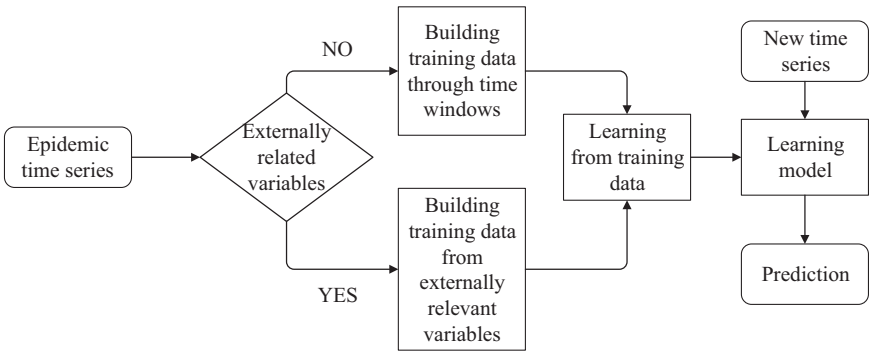

FIGURE 22.4 An AI-enhanced prediction framework for infectious disease time series.

Latent Dirichlet allocation (LDA) topic models[33] to identify newly emerging topics that cluster in space, in time, or among a subpopulation defined by observed demographic or behavioral features. When deployed in combination with existing syndromic surveillance and notifiable disease reporting systems, pre-syndromic surveillance provides a "safety net" for public health practitioners, calling their attention to newly emerging outbreaks and other events that they were not already looking for. Moreover, incorporating user feedback into the learned topic models[34] enables the system to better distinguish between relevant and irrelevant case clusters, and thus avoids overwhelming the user with false positives.

22.3 Artificial intelligence–enhanced prediction in support of public health surveillance

Time series of epidemiological data feature seasonality, nonstationarity, and sparsity. Predicting such time series has major public health implications and has attracted a lot of attention from the research and practitioner communities. Researchers have been proposing complex models for univariate prediction to extract useful patterns. In additional, efforts have been expended to develop multivariate prediction models. AI plays an important role in both research streams. Fig. 22.4 illustrates the overall AI-enhanced prediction framework.

22.3.1 Time series prediction based on dependent variables

Researchers have worked to extract long-term dependencies from one or more correlated incidence curves, without complex exogenous variables. A case in point is the "FluSight" task hosted by the US Centers for Disease Control and Prevention (CDC), which encourages seasonal influenza forecasting at the national and regional level using the weighted influenza-like illness (wILI) data.

The CNNRNN-Res model adopted RNNs to capture the long-term correlation in the wILI curves and CNNs to fuse curves of different states.[35] To avoid overfitting, this model utilized the residual links and dropout mechanism. CNNRNN-Res achieved better results than autoregressive methods and Gaussian process regression. To overcome the problem of data sparsity and improve model interpretability in influenza prediction, Adhikari et al.[36] designed a novel framework named EpiDeep. This framework consists of clustering/embedding, encoder, and decoder modules to learn meaningful embeddings of incidence curves in a continuous feature space and predicts peak intensity, peak time, onset week, and future incidences of wILI. The learned embeddings reveal the neighbor similarities, temporal similarities, intensity separation, and other patters in different flu seasons. Focusing on the problem of high-resolution ILI incidence forecasting, the DEFSI model used the SEIR model and multiagent simulation to obtain time series of incidence with high-spatial and -temporal resolution for model training. Results showed that the DEFSI model outperformed the baselines at the state level and for high-resolution forecasting at the county level.[37]

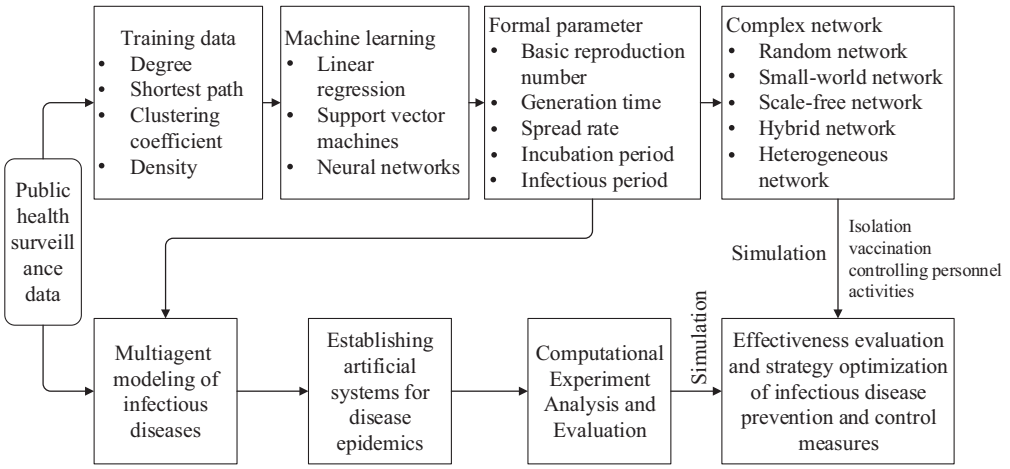

FIGURE 22.5 A simulation framework for analyzing infectious disease prevention and control strategies.

The abovementioned models focus on the time dependence of the incidence curves. In fact, these sequences of different regions also have spatial similarity. The graph neural network can be applied to capture the spatial correlation of different geographical scales. Li et al.[38] used a graph-structured recurrent neural network (GSRNN) to predict the influenza data provided by the US CDC. Nodes of the graph represent the Health and Human Services regions and edges adjacency between regions. This model, shown to deliver the state of the art performance, partitioned the nodes into two classes based on the influenza activity level, and the nodes and edges features were trained separately.

22.3.2 Time series prediction based on dependent and independent variables

Past research has also explored external variables that are highly correlated with the outbreak of infectious diseases to make multivariate predictions. The prediction performance of such methods has a lot to do with the selection of external variables.

Commonly used external variables are climate data (temperature, humidity, air quality), search trends (Google indexes, Baidu indexes), social media text data (Twitter, Sina), population migration, among others. In general, accurate and fine-grained multivariate data can improve the prediction accuracy. In order to measure the contribution of each variable to the prediction results, modelers will need to try different combinations of external variables for model training and prediction. Usually, the higher the prediction accuracy, the better the interpretation of these variables.

The LSTM model is the most commonly used prediction model. Chae et al.[39] investigated the LSTM and the DNN models to predict chicken pox, scarlet fever, and malaria in Korea. Daily Naver search frequency, number of Twitter mentions, and average daily temperature and humidity data were included as exogenous variables. The DNN model performed stably, and the LSTM model was more accurate when infectious disease was

spreading. It is critical to consider the time difference between clinical data and nonclinical data, and results indicated that a lag of 7 days was more suitable in this work. In another study a multichannel attention-based LSTM neural network was designed to forecast the real-time influenza-like illness rate (ILI%) in Guangzhou, China.[40] The external variables included medicine sales records, temperature records, and rainfall, among others. Because people of different ages have varying immunity to flu, the influenza-like cases were divided into five age groups. The approach trained the influenza- and climate-related channels separately and merged these features together in an ensuing step.

In addition to developing specific prediction models, there are also hybrid models that combine the prediction results of different methods in a weighted manner to produce better accuracy and improve robustness. A self-adaptive AI model (SAAIM) that predicts influenza activity in Chongqing, China, was developed by Su et al.[41] The multisource data include ILI%, weather data, Baidu search index, and Sina Weibo data of Chongqing. SAAIM hybrids the predictions of SARIMA and XGBoost in a Kalman filter, so the weights are self-adaptive. As for the contribution of different data to SAAIM, ablation experiments showed that ILI% 1 week prior to real time had the highest ranking score, suggesting that the ILI activity is highly autoregressive. Soliman et al.[42] developed a probabilistic forecasting of influenza in Dallas County using Bayesian model averaging (BMA). In this work the baseline models are feedforward neural networks, ARIMA, LASSO, and nonparametric multivariate adaptive regression splines (MARS) model. Influenza record, Google search, and atmospheric data are variables. The BMA model outperformed the individual methods in 1- and 2-week ahead forecasts.

AI-based infectious disease prediction is still in its infancy, and some key issues summarized by Viboud and Vespignani[43] for influenza prediction are also applicable to other AI-based infectious disease prediction: How do prediction capabilities scale with data accuracy and quantity? How should ensemble predictions be optimized? And how do prediction capabilities decrease with time horizon? Clearly, more research is needed to answer these questions.

22.4 Artificial intelligence–based infectious disease transmission modeling and response assessment

Modeling complex infectious disease transmission is key to public health emergency response. A basic framework for assessing infectious disease prevention and control strategies using simulation is shown in Fig. 22.5. Complex network models and agent-based computing methods are two widely used methods. Complex network models have low computational complexity and high abstract level. Agent-based methods have higher resolution and can flexibly reconstruct detailed plans and reproduce the complex process of disease transmission, which is necessary to assess major responses and interventions in real-world settings. Some studies[44] have combined these two models to capture multiple-level features on complex transmission systems.

22.4.1 Modeling disease transmission dynamics based on machine learning and complex networks

Complex network analysis[45,46] methods have been applied to study the spread of epidemics over typical network types such as small-world networks, scale-free networks, and community networks. Recently, researchers have attempted to combine AI and complex networks for infectious disease transmission modeling.

Tripathi et al.[47] used machine learning techniques to predict the controllability of diseases on complex networks. In their experiments the input of the training data was the complex network properties, including average degree, average shortest path length, clustering coefficient, density, diameter, and maximum degree. Their approach applied three machine learning methods, linear regression, SVM, and a neural network model, to predict an important parameter in disease transmission—basic reproduction number (R0), which determines whether the disease-free epidemic or an endemic state is asymptotically stable. SIR epidemic spreading models were used to simulate the disease spreading dynamics on four types of complex networks. The experimental results showed that the prediction of machine learning on complex networks was highly accurate.

Scarpino and Petri[48] adopted dynamic modeling approaches to study the predictability of infectious disease outbreaks. Permutation entropy, Markov chain simulations, and epidemic simulations were used in the experiment. The results indicated that both shifting model structures and social network heterogeneity are likely to lead to differences in the predictability of infectious diseases.

22.4.2 Modeling disease transmission dynamics based on multiagent modeling

Effective interventions are essential to curb the spread of infectious diseases. Assessing the effectiveness of such responses entails analysis of various "what-if" scenarios. Agent-based simulations provide an AI-based framework to carry out these assessment-related tasks.

Mei et al.[44] proposed a model that unified agent-based modeling and complex networks and applied complex agent networks to model infectious diseases. Rocha and Masuda[49] developed an individual-based approximation for the SIR epidemic model applicable to arbitrary dynamic networks. Großmann et al.[50] proposed the rejection-based simulation of non-Markovian agents on complex networks and demonstrated its efficacy on various models of epidemic spreading. Through multiagent simulation, Kuga and Tanimoto[51] found that a pandemic arises more easily in a scale-free network than in homogeneous networks.

To construct a multiagent simulation run, generating and simulating realistic and dynamic contact networks remains a major challenge. Considering spatiotemporal dynamics of influenza, Cliff et al.[52] proposed ACEMod (Australian Census-based Epidemic Model). This model employed a discrete-time and stochastic agent-based model to investigate complex outbreak scenarios at various spatiotemporal levels. In this model, each agent contained a set of attributes of an anonymous individual. The agents' distributions

TABLE 22.1 Examples of Internet-based surveillance systems.

Type	Example	Data source	Establishment
Moderated system	ProMED	Media reports, official reports, online summaries, local observers	1994
Partially moderated system	GPHIN	News sources	1998
Fully automated system	MedISys	ProMED, GPHIN	2004
	HealthMap	ProMED Mail, WHO, GeoSentinel, EuroSurveillance, Google News	2006
	PULS	Text-based news sites and social media resources	2007
	SENTINEL	Twitter data, news data, CDC materials	2019

GPHIN, Global Public Health Intelligence Network; *MedISys*, Medical Information System; *ProMED*, Program for Monitoring Emerging Disease; *PULS*, Pattern-based Understanding and Learning System.

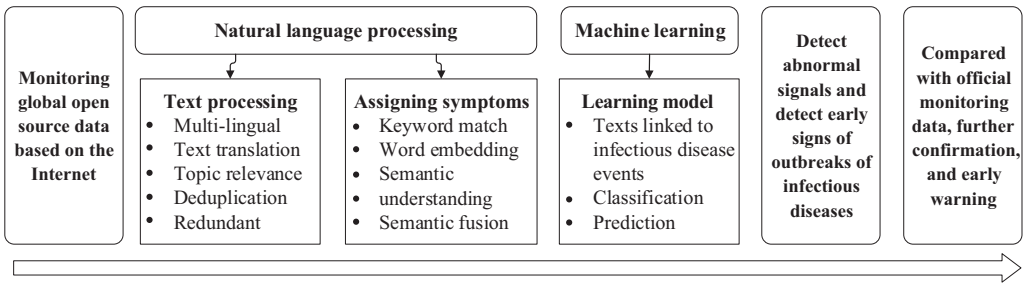

FIGURE 22.6 AI techniques in Internet-based global epidemic monitoring.

at multiple scales concurred with key demographic statistics from the 2006 Australian census.

The abovementioned studies were motivated by the needs of modeling infection and responses in large populations. In other settings, studying disease spread among occupants in confined spaces such as educational institutions was critically needed as well.

Duan et al.[53] developed an agent-based model to simulate how epidemics spread in structured space. This model was used to evaluate the public health response policies used during an H1N1 outbreak. This simulation model captured spatial layouts, population distribution, social networks, and contact patterns. A hierarchical social contact network consistent with actual social relationships and characteristics was developed. An agent-based approach was used to simulate the behavior and actions of the students and the spread of epidemics in the hierarchical network. Ge et al.[54] also used an individual-based method to reconstruct an artificial university. High-resolution social interaction algorithms were designed and nonpharmaceutical interventions evaluated quantitatively. The framework of the virtual university was constructed with four components:

synthetic population, behavior schedule, social networks, and disease transmission model. To perform a more realistic simulation, Iwanaga et al.[55] focused on how to obtain the average infectivity ratio. Based on real epidemic data from the Japan Coast Guard Academy, they proposed a discrete-time epidemic model and investigated how to estimate the infectivity rate from the real data. After obtaining the infectivity ratio the authors simulated a seasonal influenza epidemic using the SEPIR model with multiagent simulation fed with estimated spatiotemporal parameters.

22.5 Internet-based surveillance systems for global epidemic monitoring

To develop an effective global epidemic monitoring approach, researchers have long used Internet-based methods. Internet-based disease surveillance serves as a real-time complementary approach to traditional indicator-based public health disease surveillance methods.[56] Typically, Internet-based surveillance systems use a variety of open-source Internet data, including online newswires, social media, and other Internet-based data streams to detect early warning signals of threats to public health.

There are three types of Internet-based surveillance systems: moderated, partially moderated, and fully automated (Table 22.1). The Program for Monitoring Emerging Diseases (ProMED)[57,58] is a moderated system. The Global Public Health Intelligence Network (GPHIN)[59,60] is a partially moderated system developed by the Canadian Government. On average, the GPHIN processes 3000 news reports every day. Both ProMED and GPHIN can function in multiple languages. Fully automated systems include the European Commission's Medical Information System (MedISys),[61] Pattern-based Understanding and Learning System (PULS),[62] and HealthMap.[63] SENTINEL[8] is a newly developed software system built upon recent developments in machine learning and data processing for real-time syndromic surveillance based on social media data. This system can detect disease outbreaks to provide situational awareness.

From a technical standpoint, in the application context of Internet-based global epidemic monitoring, AI techniques have played a major role in a sequence of data processing and analysis tasks. Fig. 22.6 shows a list of such tasks.

Text processing: Several important steps in text processing include translation, relevancy ranking, event extraction, and deduplication. GPHIN employs language-specific keywords and algorithms to extract relevant data from the Internet and news aggregator databases.[64] PULS employs language-specific linguistic analysis and ontologies and inference rules to extract relevant data.[62] Relevancy ranking is to assess the relevancy of the report according to the user's interest. Keyword recognition algorithms, Boolean combinations, and proximity searches are the more commonly used method in event extraction. Information extraction technology is the basis of PULS.

Assigning symptoms: The purpose of assigning symptoms is to determine which tweets show symptoms of illness. SENTINEL[8] used a keyword matching technique and enriched synonym lists by using the word embeddings trained on Twitter data. They generated a list of the 10 closest words in the embedding space by cosine similarity. They use Glove[65] and FastText[66] techniques to generate word embeddings.

Machine learning: Machine learning classifiers to identify those tweets and news articles that are genuinely health related.[67,68] The CNN and LSTM are applied in SENTINEL (see Section 22.2.2).

22.6 Conclusion

The core tasks of public health surveillance and response include infectious disease outbreak detection and early warning, trend prediction, and public health response modeling and assessment. Recent years have seen wide adoption of AI techniques in accomplishing these tasks in highly dynamic, complex, and data-rich environments.

In this chapter, we reviewed a collection of recent studies focusing on how to make use of sensor and social data from the physical world and cyberspace to improve the outbreak detection and early warning capabilities. We also discussed a set of methods aimed at modeling infection disease transmission and predicting various time series of epidemiological data. Several simulation frameworks were reviewed with the objective of modeling and assessing public health responses. A common theme cutting across these areas is that AI plays an essential role in this new generation of public health methods. With the AI technology itself, machine learning methods, including deep learning, and agent-based modeling, are among the most relevant, from a methodological standpoint.

Given the growing interests and research activities in the interdisciplinary area of AI and public health surveillance, despite the impressive set of accomplishments already achieved, one can almost surely conclude that this area is still in its early stage of development with a lot of potentials yet to be fulfilled.

To conclude this chapter, we briefly present several key challenges to be mindful about, while we as a community continue harnessing the power of rapidly developing AI technologies in the context of public health. At the present time, in the middle of the COVID-19 pandemic, there are a lot of ongoing discussions about developing public health big data for surveillance and response purposes. It is important to realize the limitations and potentially significant biases associated with public health big data. In particular, privacy protection, algorithm discrimination, and model interpretability must receive serious consideration to conform to social ethics and norms. As in the case of many other AI applications in the medical domain, AI-enabled and -enhanced public health surveillance and response hold real potentials with significant challenges remaining.

References

1. Ding FY, Fu JY, Jiang D, Hao MM, Lin G. Mapping the spatial distribution of *Aedes aegypti* and *Aedes albopictus*. *Acta Trop* 2018;**178**:155–62. Available from: https://doi.org/10.1016/j.actatropica.2017.11.020.
2. Ong J, Liu X, Rajarethinam J, Kok SY, Liang S, Tang CS, et al. Mapping dengue risk in Singapore using Random Forest. *PLoS Negl Trop Dis* 2018;**12**(6). Available from: https://doi.org/10.1371/journal.pntd.0006587. e0006587.
3. Rubio-Solis A, Musah A, Dos Santos PW, Massoni T, Birjovanu G, Kostkova, P. ZIKA virus: prediction of Aedes mosquito larvae occurrence in Recife (Brazil) using online extreme learning machine and neural networks. In: *Paper presented at the Proceedings of the ninth international conference on digital public health*; 2019.

4. Akhtar M, Kraemer MU, Gardner LM. *A dynamic neural network model for predicting risk of Zika in real-time. bioRxiv, 466581*. 2019.
5. Anno S, Hara T, Kai H, Lee MA, Chang Y, Oyoshi K, et al. Spatiotemporal dengue fever hotspots associated with climatic factors in Taiwan including outbreak predictions based on machine-learning. *Geospat Health* 2019;**14**(2). Available from: https://doi.org/10.4081/gh.2019.771.
6. Jain VK, Kumar S. Effective surveillance and predictive mapping of mosquito-borne diseases using social media. *J Computat Sci* 2018;**25**:406–15. Available from: https://doi.org/10.1016/j.jocs.2017.07.003.
7. Thapen N, Simmie D, Hankin C, Gillard J. DEFENDER: detecting and forecasting epidemics using novel data-analytics for enhanced response. *PLoS One* 2016;**11**(5). Available from: https://doi.org/10.1371/journal.pone.0155417. ARTN e0155417.
8. Şerban O, Thapen N, Maginnis B, Hankin C, Foot V. Real-time processing of social media with SENTINEL: a syndromic surveillance system incorporating deep learning for health classification. *Inf Process Manage* 2019;**56**(3):1166–84. Available from: https://doi.org/10.1016/j.ipm.2018.04.011.
9. Chen F., Neill D.B. Non-parametric scan statistics for event detection and forecasting in heterogeneous social media graphs. In: *Proceedings of the 20th ACM SIGKDD conference on knowledge discovery and data mining*, 2014. p. 1166–1175.
10. Dai X, Bikdash M. Hybrid classification for tweets related to infection with influenza. In: *Paper presented at the SoutheastCon 2015*; 2015.
11. Dai X, Bikdash M, Meyer, B. From social media to public health surveillance: Word embedding based clustering method for Twitter classification. In: *Paper presented at the SoutheastCon 2017*; 2017.
12. Wang C-K, Singh O, Tang Z-L, Dai H-J. Using a recurrent neural network model for classification of tweets conveyed influenza-related information. In: *Paper presented at the proceedings of the international workshop on digital disease detection using social media 2017 (DDDSM-2017)*; 2017.
13. Lampos V, Zou B, Cox IJ. Enhancing feature selection using word embeddings: the case of flu surveillance. In: *Paper presented at the proceedings of the 26th international conference on World Wide Web*; 2017.
14. Edo-Osagie O, Lake I, Edeghere O, De La Iglesia B. Attention-based recurrent neural networks (RNNs) for short text classification: an application in public health monitoring. In: *Paper presented at the international work-conference on artificial neural networks*; 2019.
15. Souza RCSNP, Assuncao RM, Oliveira DM, Neill DB, Meira Jr W. *Where did I get dengue? Detecting spatial clusters of infection risk with social network data.. Spat. Spatiotemporal Epidemiol.* 2019;**29**:163–75.
16. Souza RCSNP, Assuncao RM, Neill DB, Meira W, Jr.. Detecting spatial clusters of disease infection risk using sparsely sampled social media mobility patterns. In: *Proc. 27th ACM SIGSPATIAL Intl. Conf.* on advances in geographic information systems, 2019b, p. 359–368.
17. Shah Z, Dunn AG. Event detection on Twitter by mapping unexpected changes in streaming data into a spatiotemporal lattice. In: *IEEE transactions on big data*; 2019.
18. Magumba MA, Nabende P, Mwebaze E. Design choices for automated disease surveillance in the social web. *Online J Public Health Inf* 2018;**10**(2): e214. Available from: https://doi.org/10.5210/ojphi.v10i2.9312.
19. Wagner MM, Tsui F-C, Espino J, et al. National Retail Data Monitor for public health surveillance. *Morb Mortal Wkly Rep* 2004;**53**(Supp):40–2.
20. Polgreen PM, Chen Y, Pennock DM, Nelson FD, Weinstein RA. Using internet searches for influenza surveillance. *Clin. Infect. Dis.* 2008;**47**(11):1443–8.
21. Ginsberg J, Mohebbi M, Patel R, et al. Detecting influenza epidemics using search engine query data. *Nature* 2009;**457**:1012–14.
22. Kulldorff M. A spatial scan statistic. *Commun Stat-Theor M* 1997;**26**(6):1481–96.
23. Kulldorff M. (Prospective time-periodic geographical disease surveillance using a scan statistic. *J R Stat Soc A* 2001;**164**:61–72.
24. Neill DB. Fast subset scan for spatial pattern detection. *J R Stat Soc B* 2012;**74**(2):337–60.
25. Duczmal L, Assuncao R. A simulated annealing strategy for the detection of arbitrary shaped spatial clusters. Comput. Statist. *Data Anal* 2004;**45**:269–86.
26. Duczmal L, Cancado A, Takahashi R, Bessegato L. A genetic algorithm for irregularly shaped scan statistics. *Comput. Statist. Data Anal.* 2007;**52**:43–52.
27. Fitzpatrick D, Ni Y, Neill DB. Support vector subset scan for spatial outbreak detection. *Online J. Public Health Inform* 2017;**9**(1):e021.

28. Faigen Z, Deyneka L, Ising A, et al. Cross-disciplinary consultancy to bridge public health technical needs and analytic developers: asyndromic surveillance use case. *Online J. Public Health Inform* 2015;**7**(3):e228.
29. Nobles M, Lall R, Mathes R, Neill DB. Multidimensional semantic scan for pre-syndromic disease surveillance. *Online J. Public Health Inform* 2019;**11**(1):e255.
30. Lall R, Levin-Rector A, Mathes R, Weiss D. Detecting unanticipated increases in emergency department chief complaint keywords. *Online J Public Health Inform* 2014;**6**(1):e93.
31. Walsh A, Hamby St T, John TL. Identifying clusters of rare and novel words in emergency department chief complaints. *Online J Public Health Inform* 2014;**6**(1):e146.
32. Maurya A, Murray K, Liu Y, Dyer C, Cohen WW, Neill DB. Semantic scan: detecting subtle, spatially localized events in text streams. *arXiv preprint arXiv* 2016. 1602.04393.
33. Blei D, Ng A, Jordan M. Latent dirichlet allocation. *J Mach Learn* 2003;**3**:993–1022.
34. Nobles, M. Multidimensional semantic scan for pre-syndromic surveillance. Ph.D. thesis, H.J. Heinz III College, Carnegie Mellon University, 2019.
35. Wu YX, Yang YM, Nishiura H, Saitoh M. Deep learning for epidemiological predictions. In: ACM/*Sigir Proceedings* 2018; 2018. pp. 1085–1088. Available from: https://doi.org/10.1145/3209978.3210077.
36. Adhikari B, Xu X, Ramakrishnan N, Prakash BA. *EpiDeep*. In: *Paper presented at the proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining - KDD '19*; 2019.
37. Wang LJ, Chen JZ, Marathe M. DEFSI: Deep Learning Based Epidemic Forecasting with Synthetic Information. In: *Thirty-third AAAI conference on artificial intelligence/thirty-first innovative applications of artificial intelligence conference/ninth AAAI symposium on educational advances in artificial intelligence*; 2019. p. 9607–12.
38. Li Z, Luo X, Wang B, Bertozzi AL, Xin J. A study on graph-structured recurrent neural networks and sparsification with application to epidemic forecasting. In: *Paper presented at the world congress on global optimization*; 2019.
39. Chae S, Kwon S, Lee D. Predicting infectious disease using deep learning and big data. *Int J Env Res Public Health* 2018;**15**(8). Available from: https://doi.org/10.3390/ijerph15081596.
40. Zhu X, Fu B, Yang Y, Ma Y, Hao J, Chen S, et al. Attention-based recurrent neural network for influenza epidemic prediction. *BMC Bioinforma* 2019;**20**(18):1–10.
41. Su K, Xu L, Li G, Ruan X, Li X, Deng P, et al. Forecasting influenza activity using self-adaptive AI model and multi-source data in Chongqing, China. *EBioMedicine* 2019;**47**:284–92. Available from: https://doi.org/10.1016/j.ebiom.2019.08.024.
42. Soliman M, Lyubchich V, Gel YR. Complementing the power of deep learning with statistical model fusion: probabilistic forecasting of influenza in Dallas County, Texas, USA. *Epidemics* 2019;**28**:100345. Available from: https://doi.org/10.1016/j.epidem.2019.05.004.
43. Viboud C, Vespignani A. The future of influenza forecasts. *Proc Natl Acad Sci USA* 2019;**116**(8):2802–4. Available from: https://doi.org/10.1073/pnas.1822167116.
44. Mei S, Zarrabi N, Lees M, Sloot PM. Complex agent networks: an emerging approach for modeling complex systems. *Appl Soft Comput* 2015;**37**:311–21.
45. Nian F, Wang X. Efficient immunization strategies on complex networks. *J Theor Biol* 2010;**264**(1):77–83.
46. Ren G, Wang X. Epidemic spreading in time-varying community networks. *Chaos* 2014;**24**(2):023116.
47. Tripathi R, Reza A, Garg D. *Prediction of the disease controllability in a complex network using machine learning algorithms. arXiv preprint arXiv:1902.10224*. 2019.
48. Scarpino SV, Petri G. On the predictability of infectious disease outbreaks. *Nat Commun* 2019;**10**(1):898.
49. Rocha LE, Masuda N. Individual-based approach to epidemic processes on arbitrary dynamic contact networks. *Sci Rep* 2016;**6**:31456.
50. Großmann G, Bortolussi L, Wolf V. Rejection-based simulation of non-markovian agents on complex networks. In: *Paper presented at the international conference on complex networks and their applications*; 2019.
51. Kuga K, Tanimoto J. Impact of imperfect vaccination and defense against contagion on vaccination behavior in complex networks. *J Stat Mech: Theory Exp* 2018;**2018**(11):113402.
52. Cliff OM, Harding N, Piraveenan M, Erten EY, Gambhir M, Prokopenko M. Investigating spatiotemporal dynamics and synchrony of influenza epidemics in Australia: an agent-based modelling approach. *Simul Model Pract Theory* 2018;**87**:412–31.
53. Duan W, Cao Z, Wang Y, Zhu B, Zeng D, Wang F-Y, et al. An ACP approach to public health emergency management: using a campus outbreak of H1N1 influenza as a case study. *IEEE Trans Syst, Man, Cybernetics: Syst*. 2013;**43**(5):1028–41.

54. Ge Y, Chen B, Qiu X, Song H, Wang Y. A synthetic computational environment: To control the spread of respiratory infections in a virtual university. *Phys A: Stat Mech Appl* 2018;**492**:93–104.
55. Iwanaga S, Yoshida H, Kinjo S. Feasibility study on multi-agent simulations of a seasonal influenza epidemic in a closed space. In: *Paper presented at the symposium on intelligent and evolutionary systems*; 2019.
56. Bush GW. *Homeland security presidential directive/Hspd-12*. Office of the Press Secretary, White House; 2004.
57. Carrion M, Madoff LC. ProMED-mail: 22 years of digital surveillance of emerging infectious diseases. *Int Health* 2017;**9**(3):177–83.
58. Yu VL, Madoff LC. ProMED-mail: an early warning system for emerging diseases. *Clin Infect Dis* 2004;**39**(2):227–32.
59. M'ikanatha NM, Lynfield R, Van Beneden CA, De Valk H. *Infectious disease surveillance*. Wiley Online Library; 2013.
60. Mykhalovskiy E, Weir L. The global public health intelligence network and early warning outbreak detection. *Can J Public Health* 2006;**97**(1):42–4.
61. Rortais A, Belyaeva J, Gemo M, Van der Goot E, Linge JP. MedISys: An early-warning system for the detection of (re-) emerging food-and feed-borne hazards. *Food Res Int* 2010;**43**(5):1553–6.
62. Hartley DM, Nelson NP, Arthur R, Barboza P, Collier N, Lightfoot N, et al. An overview of internet biosurveillance. *Clin Microbiol Infect* 2013;**19**(11):1006–13.
63. Freifeld CC, Mandl KD, Reis BY, Brownstein JS. HealthMap: global infectious disease monitoring through automated classification and visualization of Internet media reports. *J Am Med Inform Assoc* 2008;**15**(2):150–7.
64. Mawudeku A, Blench M, Boily L, John RS, Andraghetti R, Ruben M. 31 The Global Public Health Intelligence Network. *Infect Dis Surveill* 2007;**457**.
65. Pennington J, Socher R, Manning C. Glove: global vectors for word representation. In: *Paper presented at the proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*; 2014.
66. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching word vectors with subword information. *Trans Assoc Comput Linguist* 2017;**5**:135–46.
67. Hawkins JB, Tuli G, Kluberg S, Harris J, Brownstein JS, Nsoesie E. A digital platform for local foodborne illness and outbreak surveillance. *Online J Public Health Inform* 2016;**8**(1).
68. Sarker A, Gonzalez G. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *J Biomed Inform* 2015;**53**:196–207.